Funded by
The European Union

Mapping growing stock volume integrating Sentinel-2 and National Forest Inventory data

A National Small Area Estimation approach for Italy

Gherardo Chirici (gherardo.chirici@unifi.it)
Elia Vangi (elia.vangi@unifi.it)

SPACE IT UP!

NextGenCarbon
Monitoring Earth's Carbon Balance

FORWARDS

UNIVERSITÀ DEGLI STUDI FIRENZE
SCHOOL OF FORESTRY
SINCE 1869

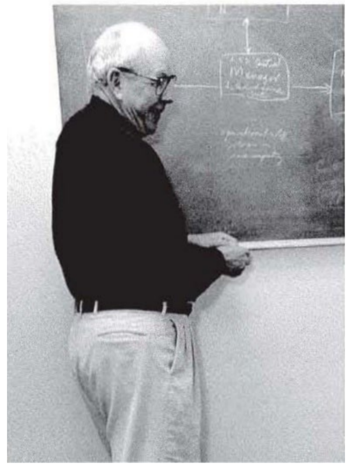
geo LAB
Laboratory of Forest Geomatics

National Forest Inventories are based on **sampling theory**

In sampling, a part of a population is selected and used to obtain estimates of characteristics of that population

To **infer** from the sample to the population sampling must be random (this is a practical way to know the probability of extraction of the sample from the population)

Dr. Daniel Goodman Horvitz
1921 - 2008



Dr. Donovan J. Thompson
1919 - 1991

A GENERALIZATION OF SAMPLING WITHOUT REPLACEMENT FROM A FINITE UNIVERSE*

D. G. Horvitz† and D. J. Thompson
Iowa State College

This paper presents a general technique for the treatment of samples drawn without replacement from finite universes when unequal selection probabilities are used. Two sampling schemes are discussed in connection with the problem of determining optimum selection probabilities according to the information available in a supplementary variable. Admittedly, these two schemes have limited application. They should prove useful, however, for the first stage of sampling with multi-stage designs, since both permit unbiased estimation of the sampling variance without resorting to additional assumptions.

INTRODUCTION

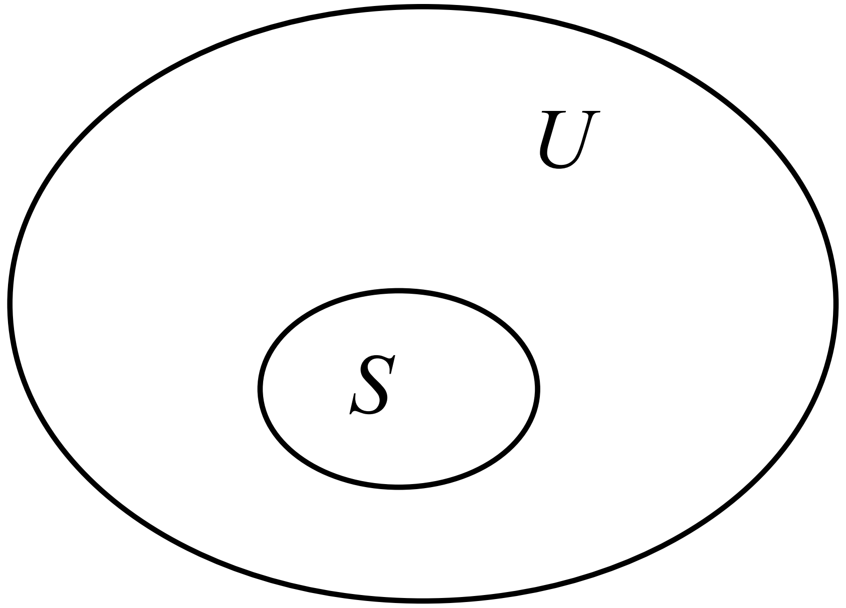
When sampling a finite universe in which we can identify the individual elements, we are free to assign in a completely arbitrary manner the probability of selecting an element on any particular draw. By appropriate assignment of the selection probabilities it is possible to reduce considerably the sampling variances of unbiased sample estimates over those obtained when sampling with equal probabilities throughout.

The possibility of using unequal probabilities for selecting the sample elements from the universe as a means of increasing precision perhaps received its first impetus for applied sampling from Hansen and Hurwitz [2] in 1943. They introduced the selection of primary units (in a subsampling scheme) with probabilities proportionate to some measure of their size and presented the appropriate theory. Their sampling scheme was confined (when sampling without replacement) to samples of one primary unit per stratum, however, the theory not having been extended beyond this point. More recently, Midzuno [6] has generalized the Hansen and Hurwitz approach to sampling a combination of n elements of the universe with probability proportionate to some measure of size of the combination. Madow [5] has made some contributions to the theory of the systematic selection of several clusters with probability proportionate to a measure of size.

* Journal Paper No. J2139 of the Iowa Agricultural Experiment Station, Ames, Iowa, Project 1005. Presented to the Institute of Mathematical Statistics, March 17, 1951.
† Now at the University of Pittsburgh.

663

Simple Random Sampling: the problem of clustering



U = universe (population)
S = sample

HT estimators

$$n = \sum_{j=1}^{N} \pi_j$$

HT total estimator:

$$\hat{T}_{HT} = \sum_{j \in S} \frac{y_j}{\pi_j}$$

$$\pi_j = \frac{n}{N}$$

if N is known

with variance:

$$V(\hat{T}_{HT}) = \sum_{j=1}^{N} \sum_{h>j} (\pi_j \pi_h - \pi_{jh}) \left(\frac{y_j}{\pi_j} - \frac{y_h}{\pi_h} \right)^2$$

HT average estimator:

N known

$$\hat{\mu}_{HT} = \frac{\hat{T}_{HT}}{N}$$

$$\hat{\mu}_{HT} = \frac{\hat{T}_{HT}}{\hat{N}_{HT}} \qquad \hat{N}_{HT} = \sum_{J \in S} \frac{1}{\pi_j}$$

N unknown

With variance:

$$V(\hat{\mu}_{HT}) = \frac{V(\hat{T}_{HT})}{N^2}$$

$$V(\hat{\mu}_{HT}) = \frac{V(\hat{T}_{HT})}{\hat{N}_{HT}^2}$$

Tabella 1.3.1 - Valori totali e per unità di superficie del volume del fusto e dei rami grossi per le categorie inventariali Boschi alti, Impianti di arboricoltura da legno e Aree temporaneamente prive di soprassuolo e per la macrocategoria Bosco

Distretto territoriale	Boschi alti				Impianti di arboricoltura da legno				Aree temp. prive di soprassuolo				Totale Bosco			
	Volume (m^3)	ES (%)	Volume (m^3 ha^{-1})	ES (%)	Volume (m^3)	ES (%)	Volume (m^3 ha^{-1})	ES (%)	Volume (m^3)	ES (%)	Volume (m^3 ha^{-1})	ES (%)	Volume (m^3)	ES (%)	Volume (m^3 ha^{-1})	ES (%)
Piemonte	126 821 547	3.1	151.0	2.9	2 947 269	15.7	103.2	12.7	7 613	60.2	3.3	45.9	129 776 430	3.0	149.1	2.8
Valle d'Aosta	15 334 302	7.6	156.0	6.9	0	-	0	-	0	-	0	-	15 334 302	7.6	155.8	6.9
Lombardia	105 423 629	4.2	182.4	3.9	2 613 095	19.6	97.4	17.8	0	-	0	-	108 036 723	4.1	178.3	3.9
Alto Adige	104 721 523	4.6	315.0	4.3	0	-	0	-	467 004	31.4	109.6	14.7	105 188 527	4.6	312.4	4.3
Trentino	105 715 538	4.8	283.5	4.6	0	-	0	-	61 011	97.3	24.2	89.8	105 776 549	4.7	281.8	4.6
Veneto	80 931 420	4.6	204.7	4.2	260 012	45.4	124.4	25.6	4 529	100.0	13.4	-	81 195 960	4.6	204.1	4.2
Friuli V.G.	67 066 949	5.4	212.1	5.1	763 052	26.8	100.3	18.7	0	-	0	-	67 830 001	5.3	209.5	5.0
Liguria	49 379 829	4.5	147.3	4.3	39 233	100.0	107.1	-	19 730	56.8	5.7	49.0	49 438 791	4.5	145.8	4.3
Emilia Romagna	71 063 339	3.9	128.7	3.7	1 274 428	20.6	130.8	13.1	356	72.6	0.3	53.7	72 338 122	3.9	128.4	3.6
Toscana	130 873 621	3.1	129.9	3.0	1 042 114	43.2	189.7	34.7	40 250	66.0	15.6	60.5	131 955 985	3.1	129.9	2.9
Umbria	29 142 004	4.8	79.2	4.6	112 665	55.0	33.3	44.6	0	-	0	-	29 254 669	4.8	78.7	4.6
Marche	24 231 008	6.6	83.5	6.3	62 614	66.5	51.6	41.1	0	-	0	-	24 293 622	6.6	83.4	6.3
Lazio	57 249 600	4.7	107.0	4.4	180 483	66.2	105.9	46.5	80 552	46.9	11.1	41.4	57 510 635	4.6	105.7	4.4
Abruzzo	50 404 587	4.6	129.5	4.4	87 051	71.8	77.5	42.2	1 193	100.0	1.0	-	50 492 831	4.6	129.0	4.4
Molise	14 523 394	9.0	110.5	8.5	106 992	93.1	120.0	54.1	5 598	100.0	22.4	-	14 635 984	8.9	110.4	8.4
Campania	42 353 904	6.0	111.5	5.7	112 595	48.3	97.4	42.7	36 194	59.6	11.2	49.3	42 502 693	6.0	110.6	5.7
Puglia	12 046 337	11.0	84.2	10.5	108 303	95.5	123.5	66.4	5 844	100.0	3.0	-	12 160 485	10.9	83.4	10.4
Basilicata	27 415 389	6.9	106.3	6.5	230 731	46.0	123.8	11.1	15 086	73.0	4.6	64.8	27 661 206	6.9	105.1	6.4
Calabria	86 990 394	4.7	190.0	4.3	706 558	54.8	267.7	39.2	270 501	57.0	35.5	53.1	87 967 454	4.7	187.9	4.3
Sicilia	23 125 002	6.7	91.2	6.1	56 190	95.7	49.4	76.4	1 605	86.9	1.1	70.6	23 182 797	6.7	90.5	6.1
Sardegna	31 286 179	5.8	57.1	5.4	1 543 109	22.2	60.4	20.0	53 445	56.1	5.6	53.1	32 882 733	5.6	56.4	5.2
Italia	1 256 099 493	1.1	146.4	1.0	12 246 493	8.7	100.2	7.4	1 070 512	21.4	19.8	20.0	1 269 416 499	1.1	144.9	1.0

UNIVERSITÀ DEGLI STUDI FIRENZE
SCHOOL OF FORESTRY
SINCE 1869

In NFIs Remote sensing contributed:

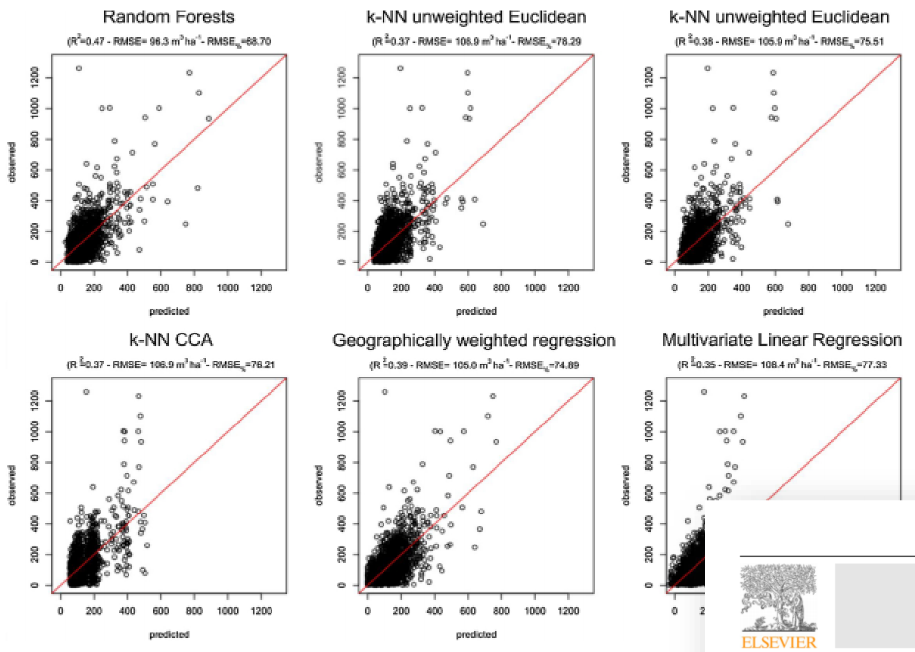
- To the production of more timely, cost efficient, and precise **traditional inventory estimates**

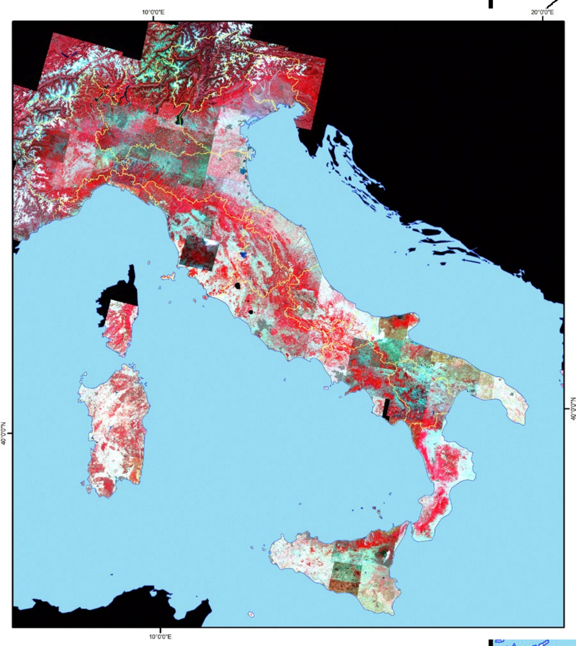
- To derive **new spatial products** (maps, small area estimates)

Technologies that are now on the horizon have the potential to alter radically the ways in which trees are measured,estimates are produced, and products are delivered.

The **lack** of standardized, spatially exhaustive **open access datasets**, as well as community consensus on methods and best practices limits the broader uptake and operationalization of these approaches

Mapping forest variables integrating NFI with EO





Int J Appl Earth Obs Geoinformation 84 (2020) 101959

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

journal homepage: www.elsevier.com/locate/jag

ELSEVIER

Wall-to-wall spatial prediction of growing stock volume based on Italian National Forest Inventory plots and remotely sensed data

Gherardo Chirici[a], Francesca Giannetti[a], Ronald E. McRoberts[b,c], Davide Travaglini[a], Matteo Pecchi[a], Fabio Maselli[d], Marta Chiesi[d], Piermaria Corona[e]

[a] Dipartimento di Scienze e Tecnologie Agrarie, Alimentari, Ambientali e Forestali, Università degli Studi di Firenze, 50145, Firenze, Italy
[b] Department of Forest Resources, University of Minnesota, Saint Paul, MN, 55108, USA
[c] Northern Research Station, U.S. Forest Service, Saint Paul, MN, 55108, USA
[d] CNR-IBE, Via Madonna del Piano, 10, 50019, Sesto Fiorentino, FI, Italy
[e] CREA, Research Centre for Forestry and Wood, viale Santa Margherita 80, 52100, Arezzo, Italy

ARTICLE INFO

Keywords:
National Forest Inventory
Spatial estimation
Growing stock
Landsat
Italy
Growing stock volume

ABSTRACT

Spatial predictions of forest variables are required for supporting modern national and sub-national forest planning strategies, especially in the framework of a climate change scenario. Nowadays methods for constructing wall-to-wall maps and calculating small-area estimates of forest parameters are becoming essential components of most advanced National Forest Inventory (NFI) programs. Such methods are based on the assumption of a relationship between the forest variables and predictor variables that are available for the entire forest area. Many commonly used predictors are based on data obtained from active or passive remote sensing technologies. Italy has almost 40% of its land area covered by forests. Because of the great diversity of Italian forests with respect to composition, structure and management and underlying climatic, morphological and soil conditions, a relevant question is whether methods successfully used in less complex temperate and boreal forests may be applied successfully at country level in Italy.

For a study area of more than 48,657 km^2 in central Italy of which 43% is covered by forest, the study presents the results of a test regarding wall-to-wall, spatially explicit estimation of forest growing stock volume (GSV) based on field measurement of 1350 plots during the last Italian NFI. For the same area, we used potential predictor variables that are available across the whole of Italy: cloud-free mosaics of multispectral optical satellite imagery (Landsat 5 TM), microwave sensor data (JAXA PALSAR), a canopy height model (CHM) from satellite LiDAR, and auxiliary variables from climate, temperature and precipitation maps, soil maps, and a digital terrain model.

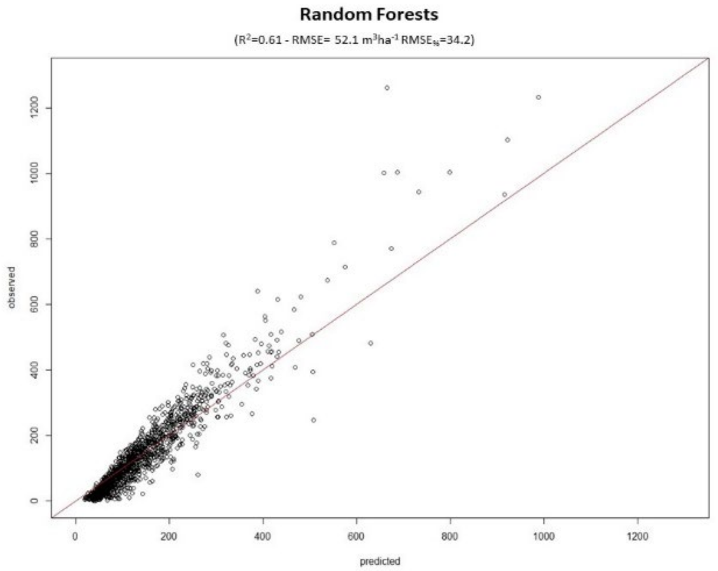
Two non-parametric (random forests and k-NN) and two parametric (multiple linear regression and geographically weighted regression) prediction methods were tested to produce wall-to-wall map of growing stock volume at 23-m resolution. Pixel level predictions were used to produce small-area, province-level model-assisted estimates. The performances of all the methods were compared in terms of percent root mean-square error using a leave-one-out procedure and an independent dataset was used for validation. Results were comparable to those available for other ecological regions using similar predictors, but random forests produced the most accurate results with a pixel level R^2 = 0.69 and $RMSE_{\%}$ = 37.2% against the independent validation dataset. Model-assisted estimates were more precise than the original design-based estimates provided by the NFI.

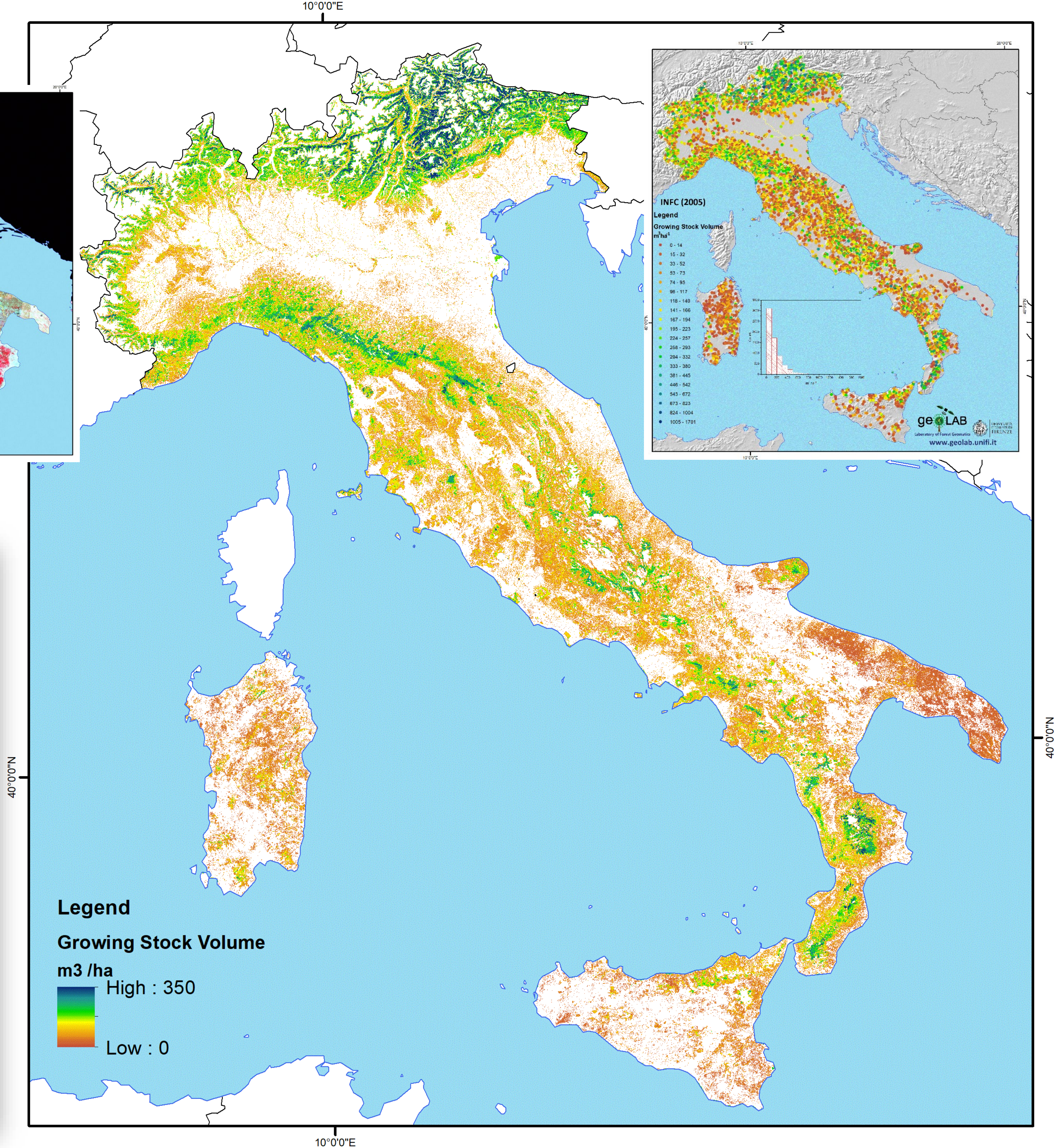
1. Introduction

Forest data are essential for multiple purposes including international and national forest monitoring programs, reporting and assessing forest resource distribution (e.g. Kyoto protocol) (Corona et al., 2011; FAO, 2010), monitoring biodiversity (Chirici et al., 2012; FOREST EUROPE, 2015), improving restoration programs (FAO and UNCCD, 2015; Smith et al., 2016) and managing at local scales to improve decision-making processes, silvicultural measures, harvesting and conservation activities.

Usually, in the context of international and national programs, this type of data is collected using sample-based National Forest Inventories (NFIs) that are designed to provide aggregated estimates of forest parameters such as forest area, growing stock volume, biomass, increments at national and regional levels (Brosofske et al., 2014; Kangas et al., 2018). These aggregated statistics are essential to support decision-making processes and to develop strategies over large areas only, because they just provide limited explicit geographic spatial detail, such as large sub-national regions. In these traditional NFIs, remote sensing is used for purposes such as initial stratification of sampling units

https://doi.org/10.1016/j.jag.2019.101959
Received 17 May 2019; Received in revised form 8 August 2019; Accepted 2 September 2019
Available online 03 October 2019



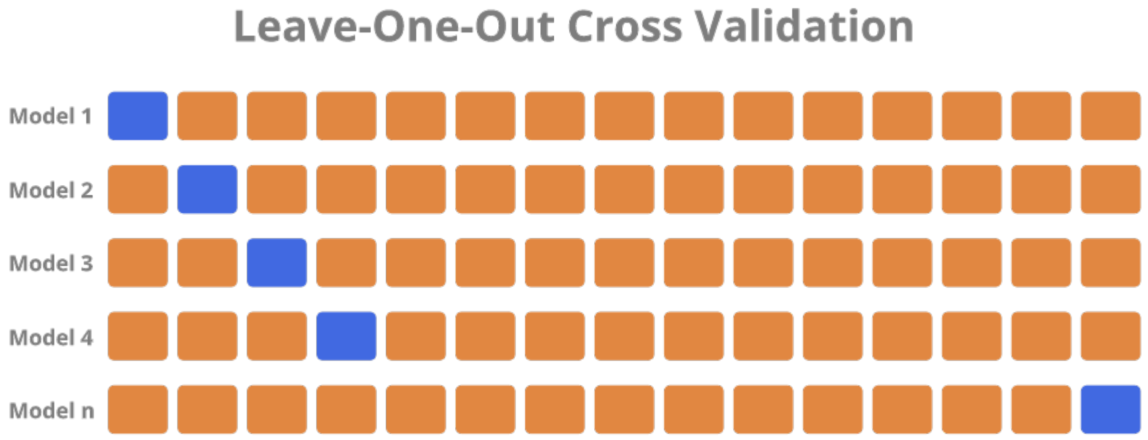


Performance evaluation

The performance of the model was evaluated through a leave-one-out cross validation (or Jack-knife) procedure through the most common metrics used for continuous outputs

RMSE (root mean square error)

R^2 (correlation coefficient)



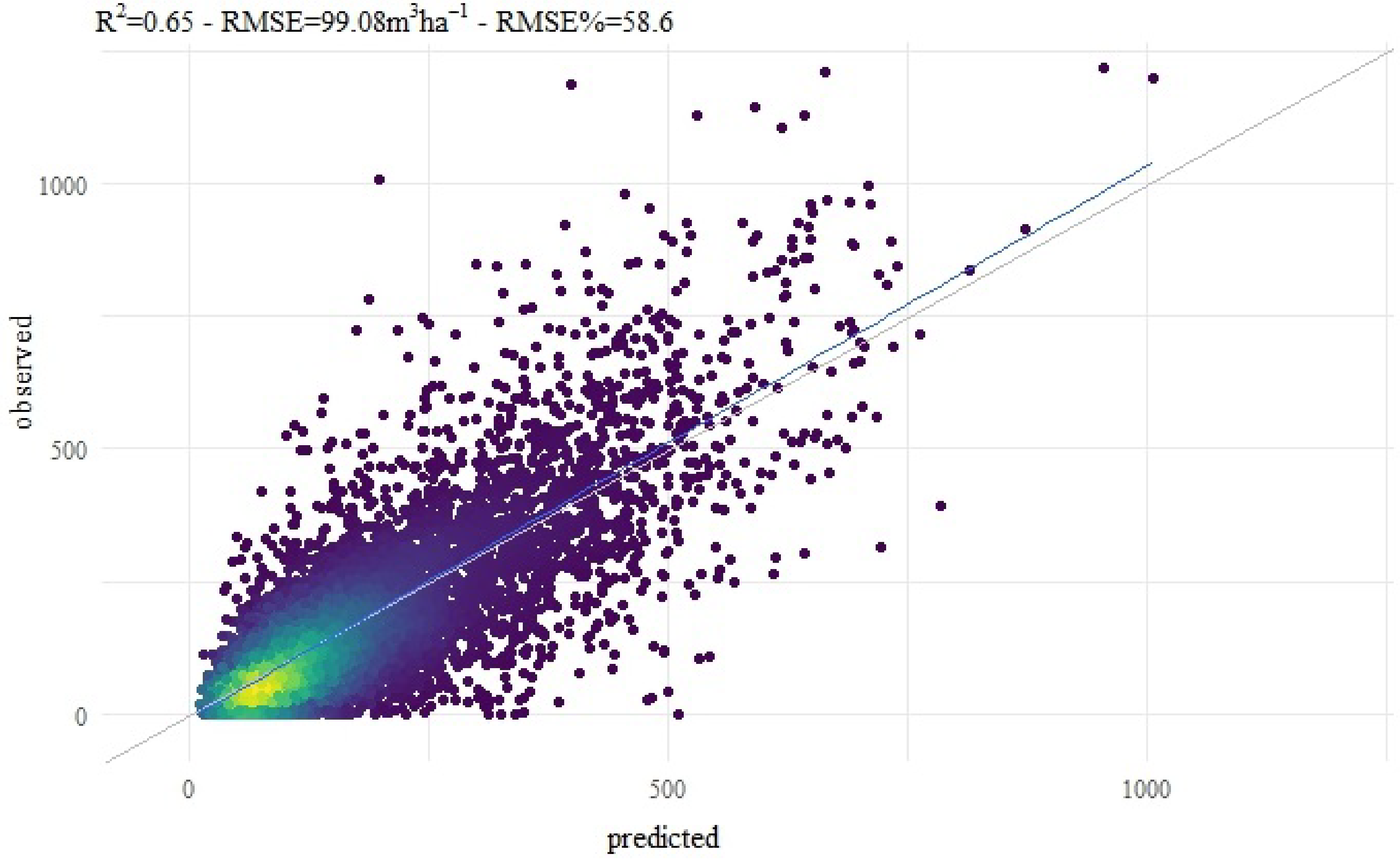
$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

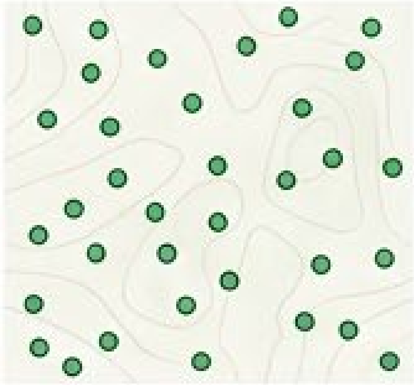
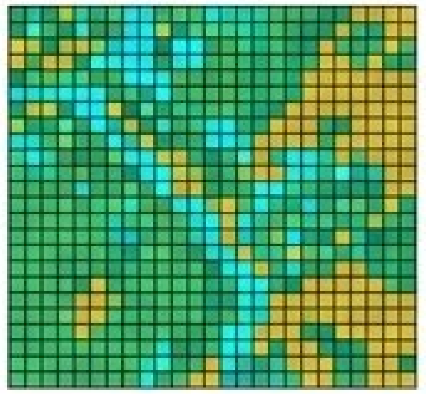
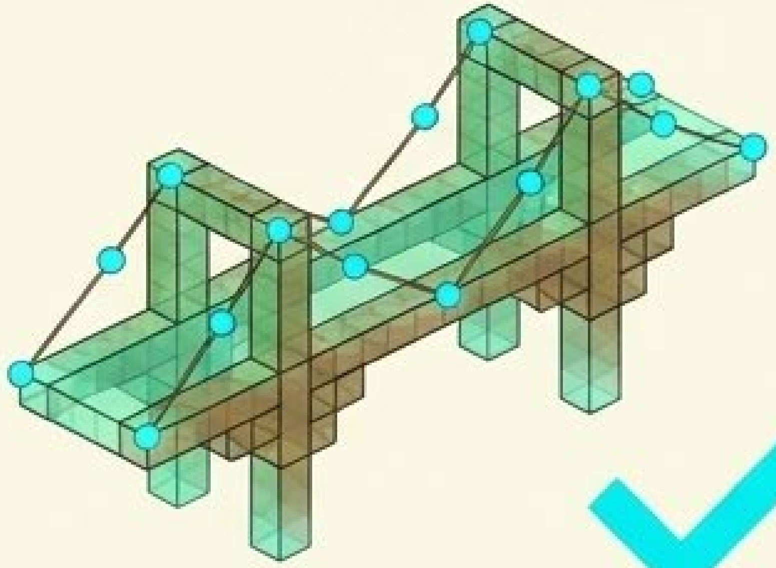
Where,

$\hat{y}$ – *predicted value of y*

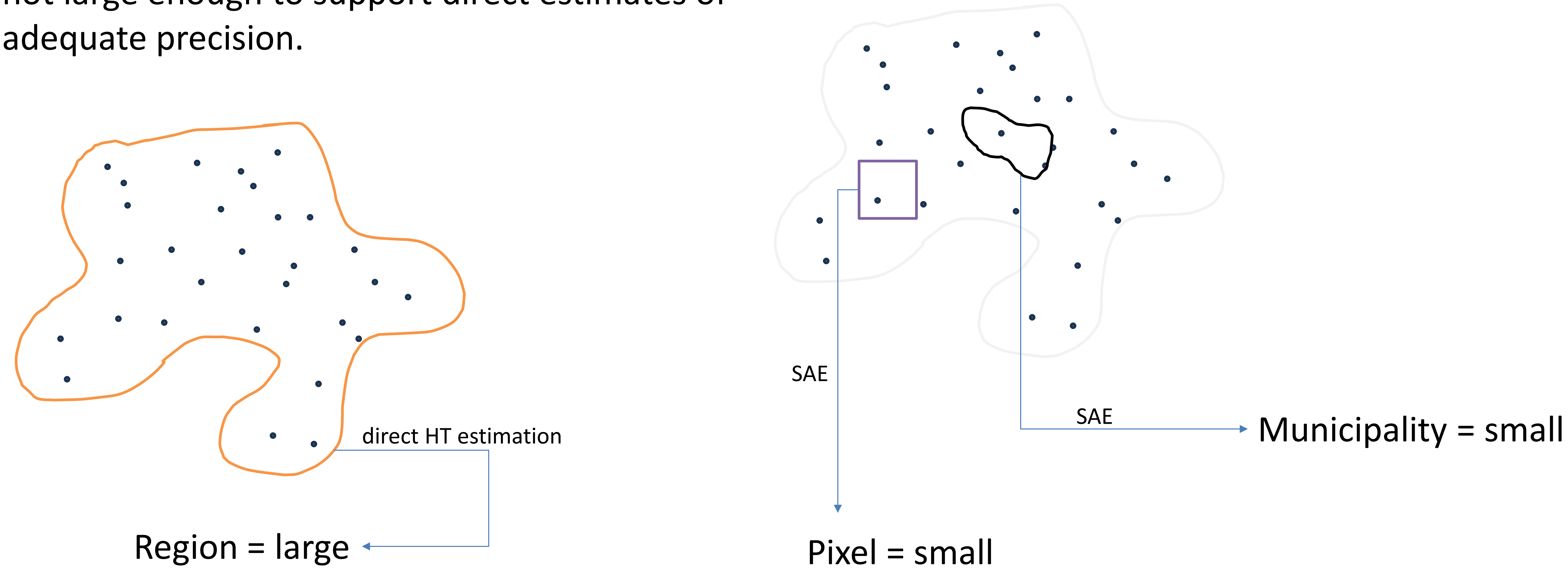
$\bar{y}$ – *mean value of y*



The Methodological Chasm

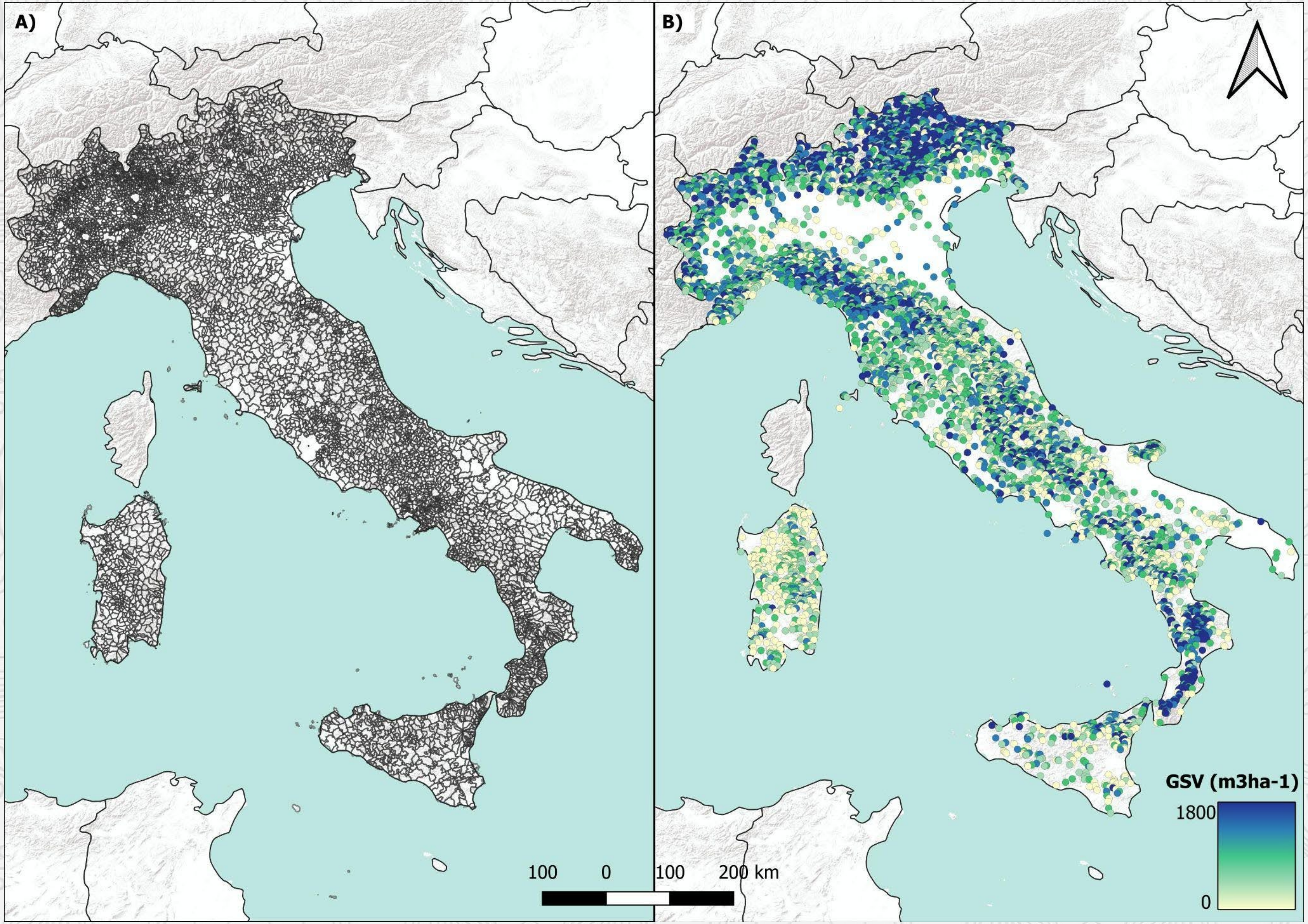
Traditional NFI	Naïve Pixel Aggregation (Predictive Maps)	**Small Area Estimation (SAE)**
		
High statistical rigor & unbiased.	**Highly detailed** spatial information.	
Limitation: Cannot estimate for small domains/municipalities due to critically low plot density.	**Limitation**: Aggregating pixel predictions overlooks spatial model bias and totally fails to account for sampling error.	Borrows strength across areas, across areas, utilizes auxiliary variables, provides unbiased estimates, and explicitly quantifies uncertainty at the municipal level.

A domain (area) is regarded as **large** (or major) if the domain-specific sample is large enough to yield “direct estimates” of adequate precision. A domain is regarded as “**small**” if the domain-specific sample is not large enough to support direct estimates of adequate precision.



2. STUDY AREA

- 6,174 Italian NFI plots
- Across 11mln ha of forest
- Measured from 2018 to 2020
- 7896 municipality units
- 38 km^2 mean area covered by a municipality
- 37% of municipalities with at least NFI sampling plot



3. DATA

Ground Truth

Italian NFI Data

- 2015 reference data
- Growing Stock Volume (GSV) computed from 6,174 plot-level field measurements

Remote Sensing

- Sentinel-2 Surface Reflectance
- Sept 2023–Aug 2024
- 10 original bands + 7 Vegetation Indices
- Cloud probability masked at 65%

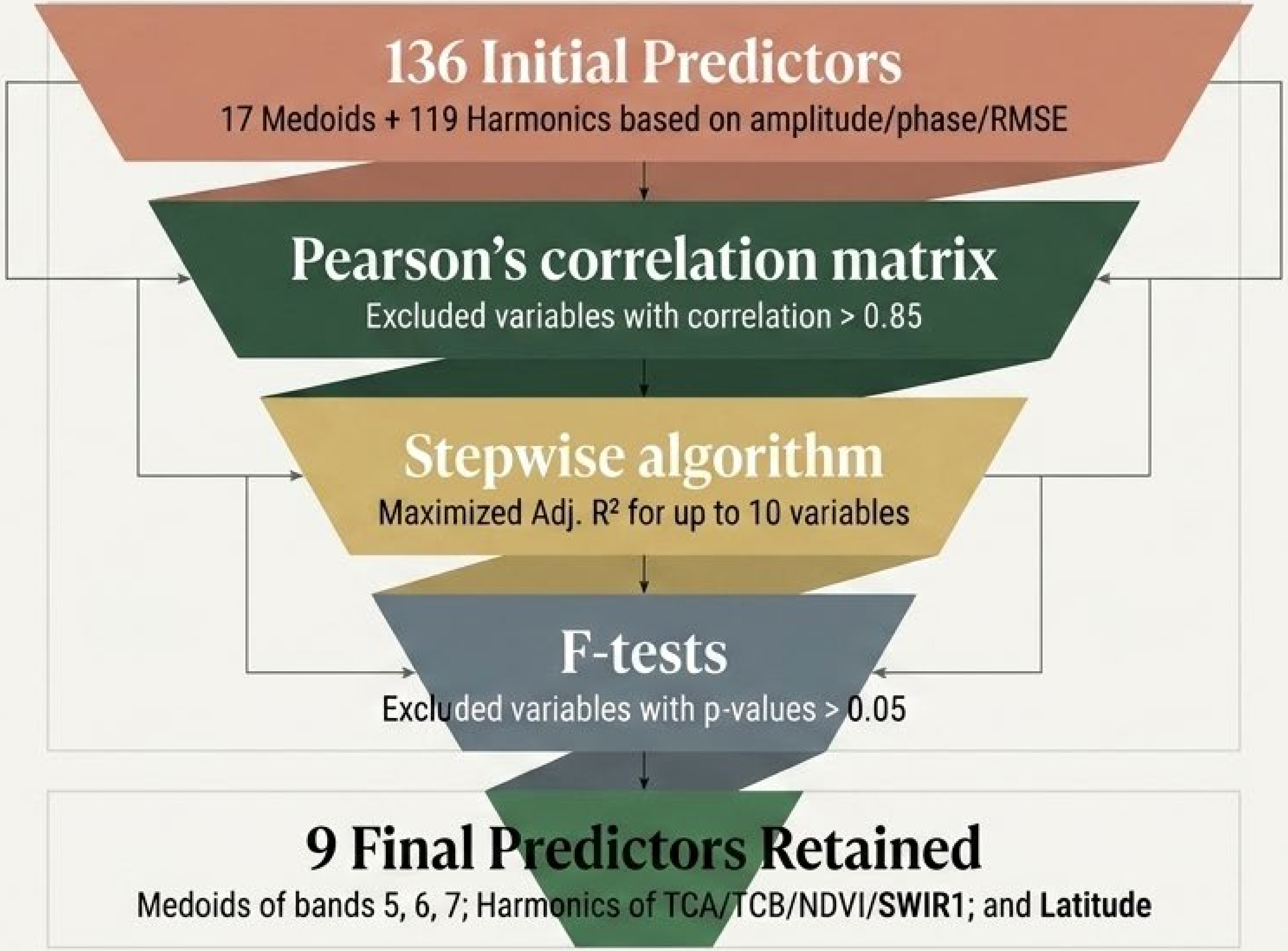
NDVI, NBR, EVI, TCB, TCW, TCG, TCA

Spatial Constraint

- National Forest Mask
- 2020 update, 70x70m resolution

To Small Area Estimation Architecture

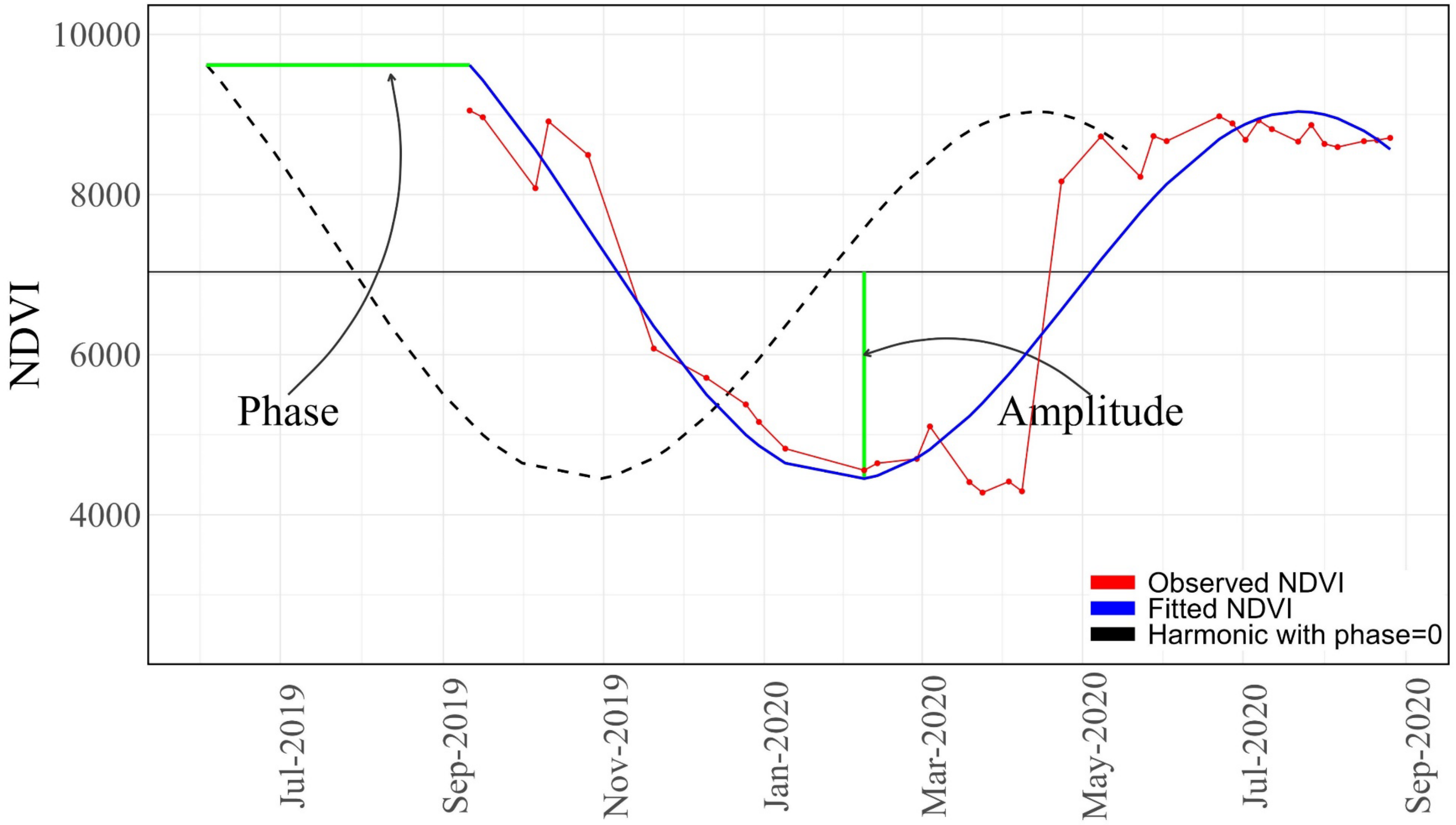
Predictor Selection Funnel

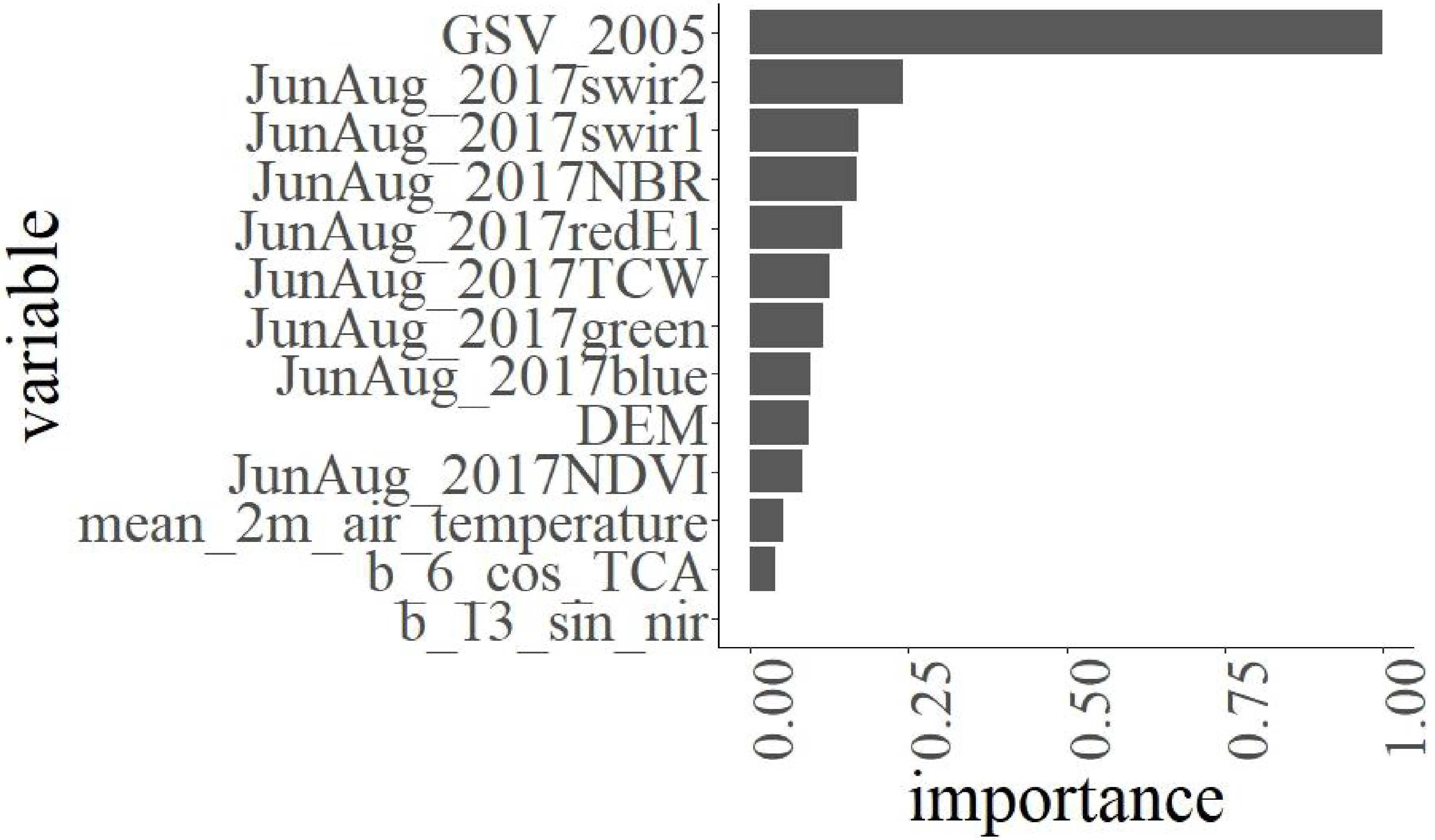


Harmonics from Spectral trajectories

The most common spectral indices (NDVI, EVI, NBR, TCB-W-G-A) and 91 harmonic metrics, derived from time trends (2017-2019) for each pixel, were calculated from the Sentinel-2 composites.

Although RF is known to be insensitive to the number of variables and the autocorrelation between them, a variable selection procedure was applied to reduce computational effort and simplify the final model.

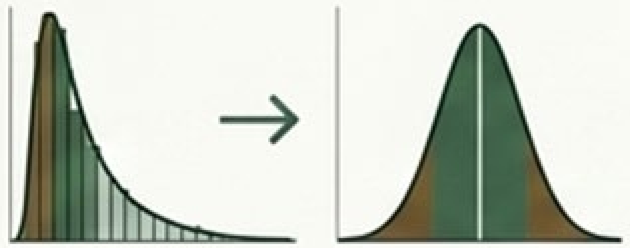




4. METHODS

The implementation of the EBP method is based on the theory described in Molina and Rao (2010) and Rao and Molina (2015). The underlying model is a unit-level mixed model, also known as

Nested error linear regression (NER) model

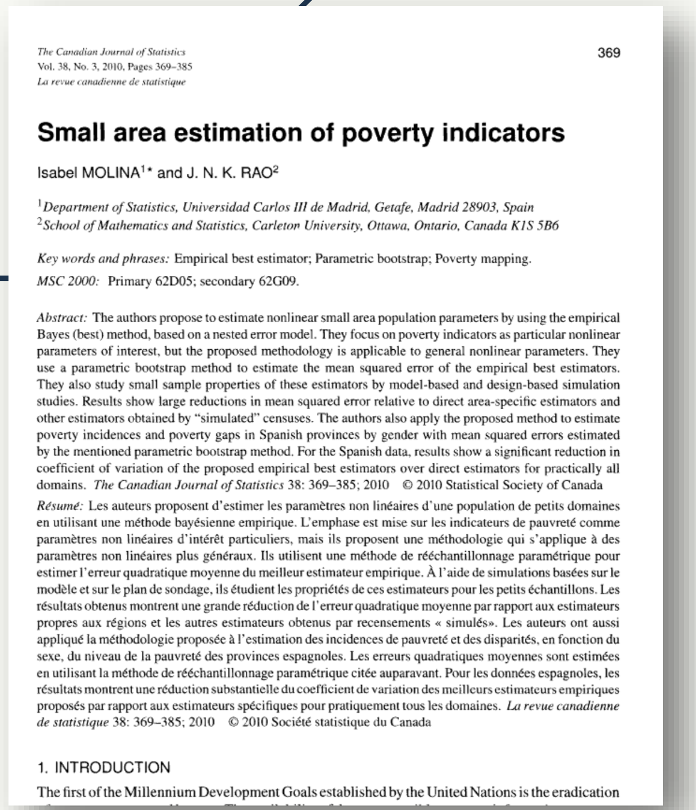


Box-Cox transformation: GSV is inherently right-skewed. Transformation relaxes the model's assumption

Fixed-effect: Predictive power of RS covariates

$$T(y_{ij}) = \mathbf{x}_{ij}^T\beta + u_i + e_{ij}$$

Domain random effect: The «between» variance effect. Captures area-specific heterogeneity across municipalities



The Canadian Journal of Statistics
Vol. 38, No. 3, 2010, Pages 369–385
La revue canadienne de statistique

369

Small area estimation of poverty indicators

Isabel MOLINA[1]* and J. N. K. RAO[2]

[1] Department of Statistics, Universidad Carlos III de Madrid, Getafe, Madrid 28903, Spain
[2] School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6

Key words and phrases: Empirical best estimator; Parametric bootstrap; Poverty mapping.

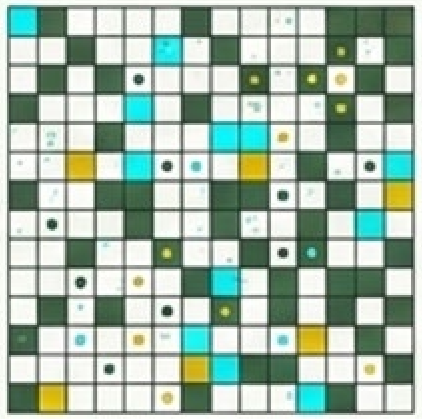
MSC 2000: Primary 62D05; secondary 62G09.

Abstract: The authors propose to estimate nonlinear small area population parameters by using the empirical Bayes (best) method, based on a nested error model. They focus on poverty indicators as particular nonlinear parameters of interest, but the proposed methodology is applicable to general nonlinear parameters. They use a parametric bootstrap method to estimate the mean squared error of the empirical best estimators. They also study small sample properties of these estimators by model-based and design-based simulation studies. Results show large reductions in mean squared error relative to direct area-specific estimators and other estimators obtained by "simulated" censuses. The authors also apply the proposed method to estimate poverty incidences and poverty gaps in Spanish provinces by gender with mean squared errors estimated by the mentioned parametric bootstrap method. For the Spanish data, results show a significant reduction in coefficient of variation of the proposed empirical best estimators over direct estimators for practically all domains. *The Canadian Journal of Statistics* 38: 369–385; 2010 © 2010 Statistical Society of Canada

Résumé: Les auteurs proposent d'estimer les paramètres non linéaires d'une population de petits domaines en utilisant une méthode bayésienne empirique. L'emphase est mise sur les indicateurs de pauvreté comme paramètres non linéaires d'intérêt particuliers, mais ils proposent une méthodologie qui s'applique à des paramètres non linéaires plus généraux. Ils utilisent une méthode de rééchantillonnage paramétrique pour estimer l'erreur quadratique moyenne du meilleur estimateur empirique. À l'aide de simulations basées sur le modèle et sur le plan de sondage, ils étudient les propriétés de ces estimateurs pour les petits échantillons. Les résultats obtenus montrent une grande réduction de l'erreur quadratique moyenne par rapport aux estimateurs propres aux régions et les autres estimateurs obtenus par recensements « simulés». Les auteurs ont aussi appliqué la méthodologie proposée à l'estimation des incidences de pauvreté et des disparités, en fonction du sexe, du niveau de la pauvreté des provinces espagnoles. Les erreurs quadratiques moyennes sont estimées en utilisant la méthode de rééchantillonnage paramétrique citée auparavant. Pour les données espagnoles, les résultats montrent une réduction substantielle du coefficient de variation des meilleurs estimateurs empiriques proposés par rapport aux estimateurs spécifiques pour pratiquement tous les domaines. *La revue canadienne de statistique* 38: 369–385; 2010 © 2010 Société statistique du Canada

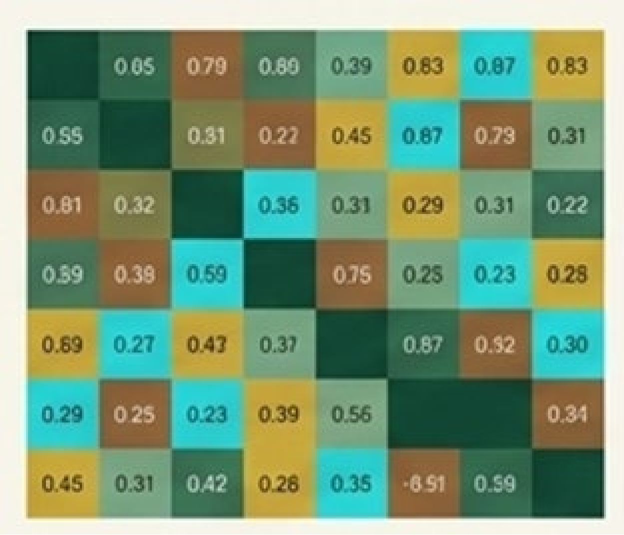
1. INTRODUCTION

The first of the Millennium Development Goals established by the United Nations is the eradication



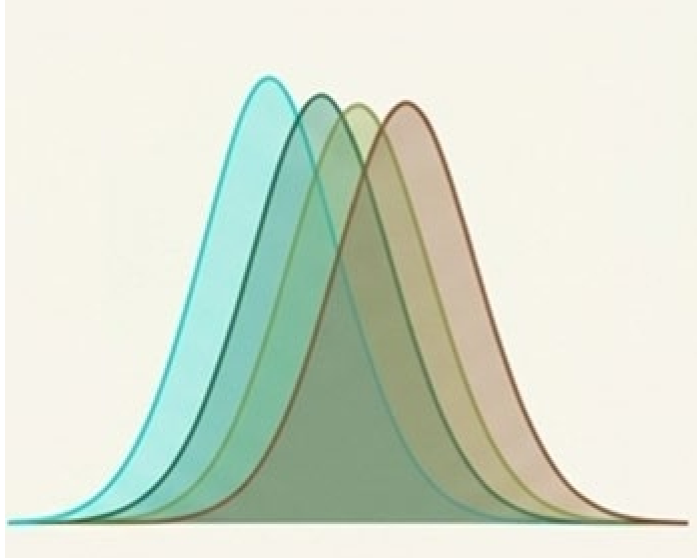
Unit-level error: The «within» variance effect. Captures pixel-level noise

4. METHODS: estimating the EBP mean



Use training data to estimate model coefficient (β) and variance component ($\sigma_u{}^2$ and $\sigma_e{}^2$) by maximizing the restricted loglikelihood (REML). Estimate the shrinkage factor $\widehat{\gamma}_i$ for computing the conditional expectation of random effect $\widehat{u}_i = \widehat{\gamma}_i(\bar{y}_i - x_{ij}{}^T\beta)$

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$$



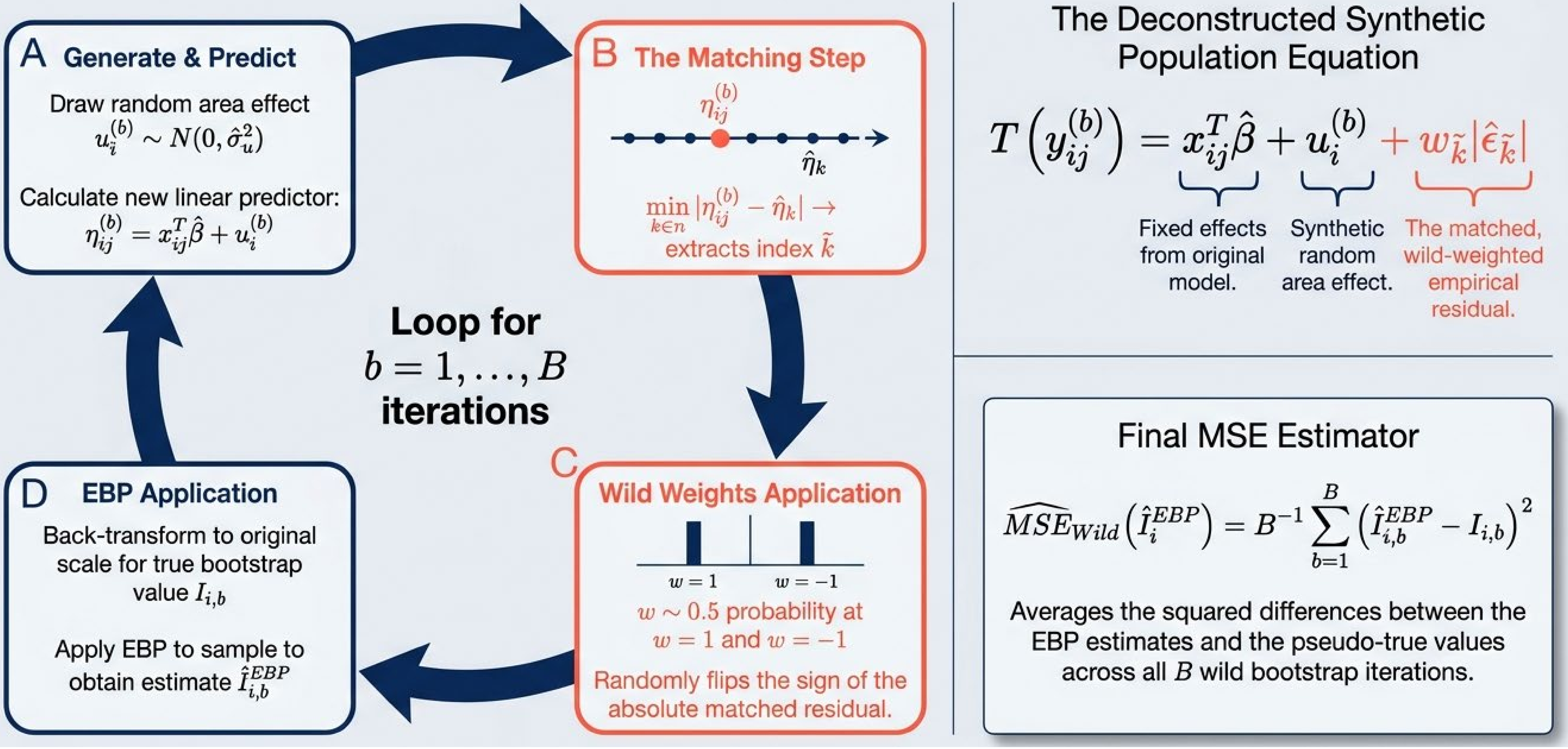
Generate a synthetic in-sample and out-of-sample population using the NER model with Monte Carlo (MC) Simulation, with the conditional expectation of $\widehat{u}_i$. Back-transform the data

Compute the indicator for each domain:

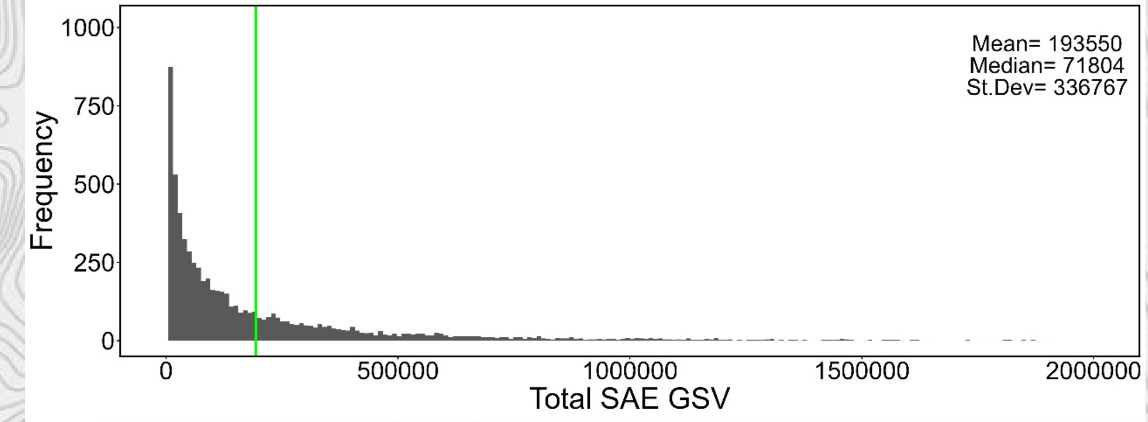
$$\theta_d = \frac{1}{N_d}\sum_{i=1}^{N_d} y_{di}$$

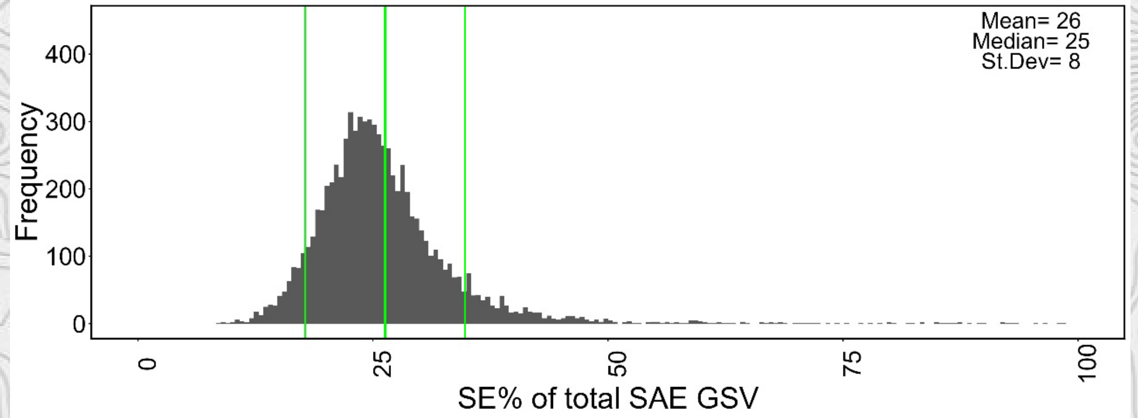
as the mean of all MC simulations

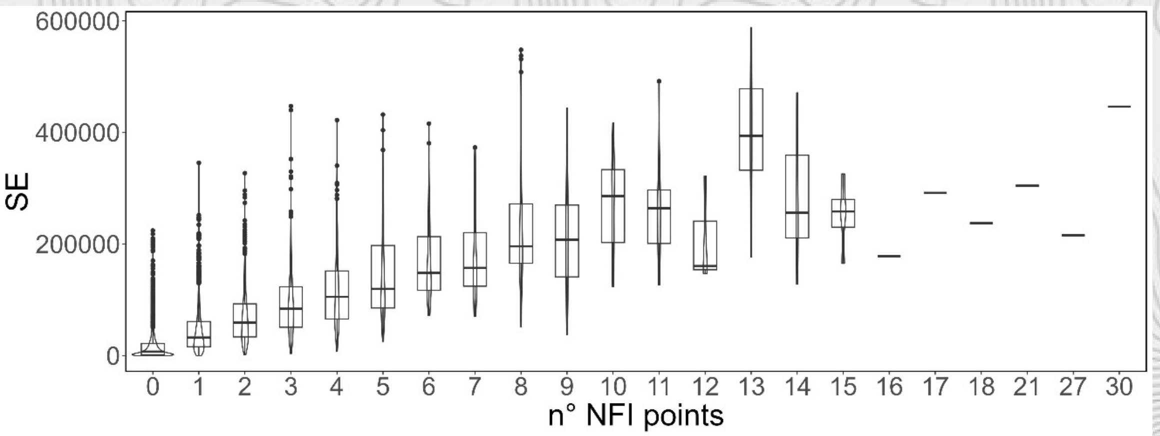
The Semi-Parametric Wild Bootstrap Loop and MSE Evaluation

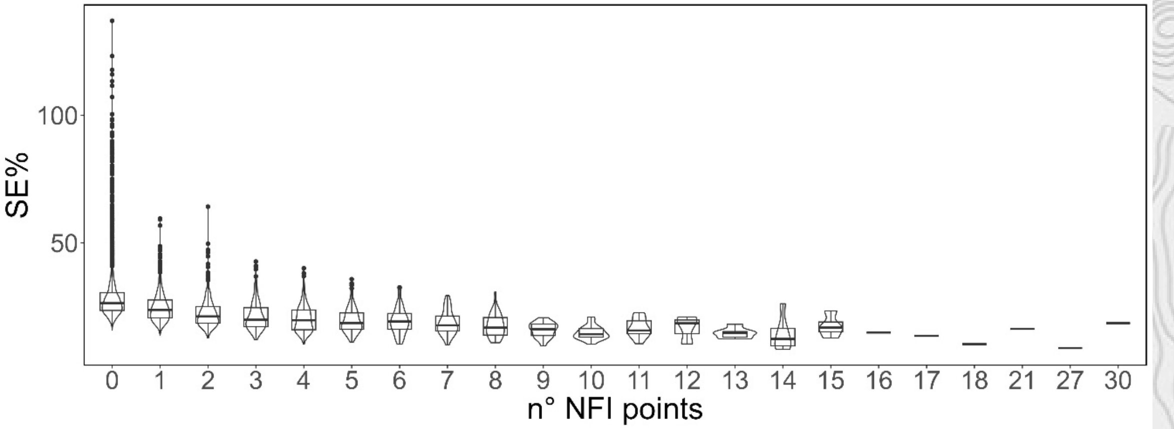


5. RESULTS

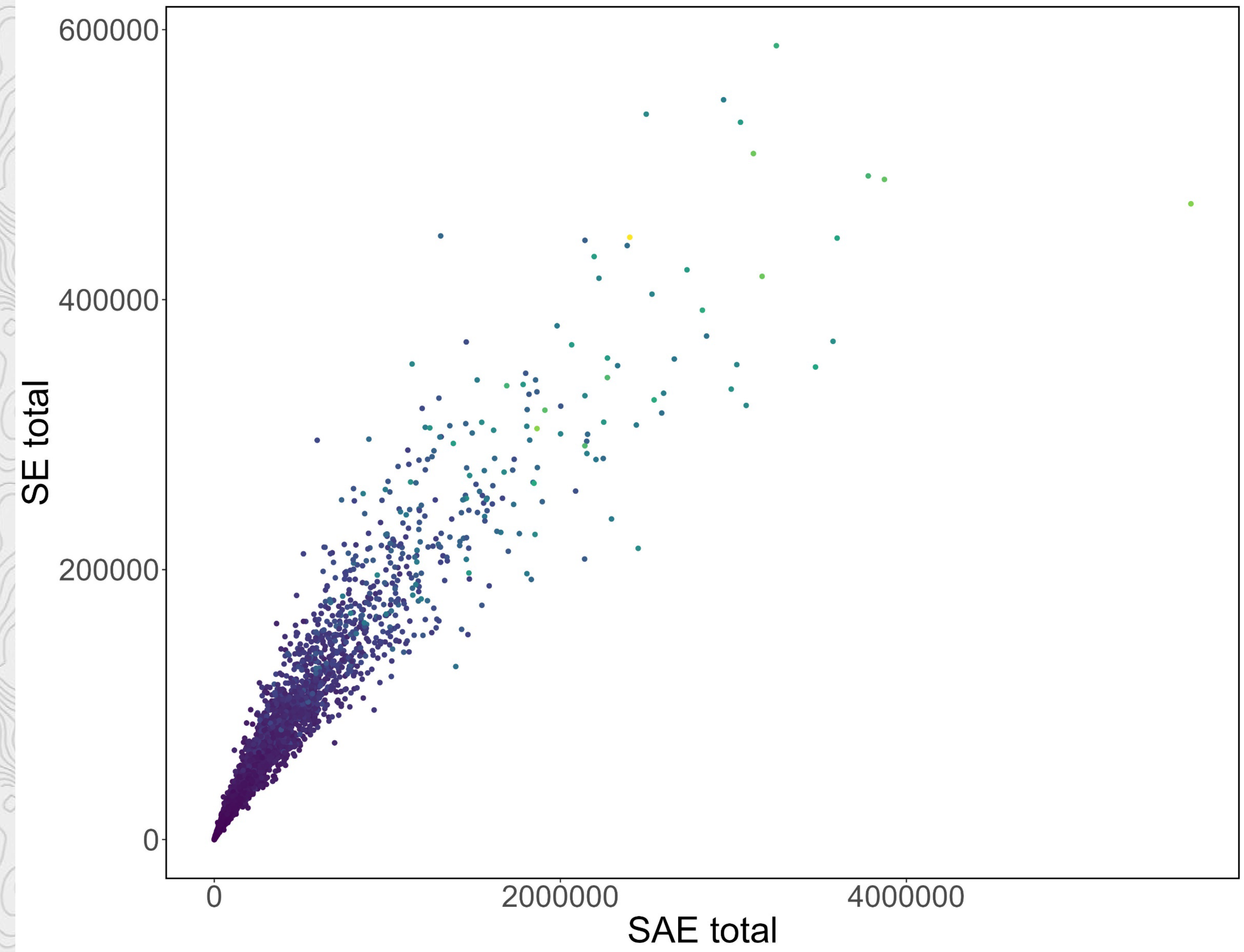




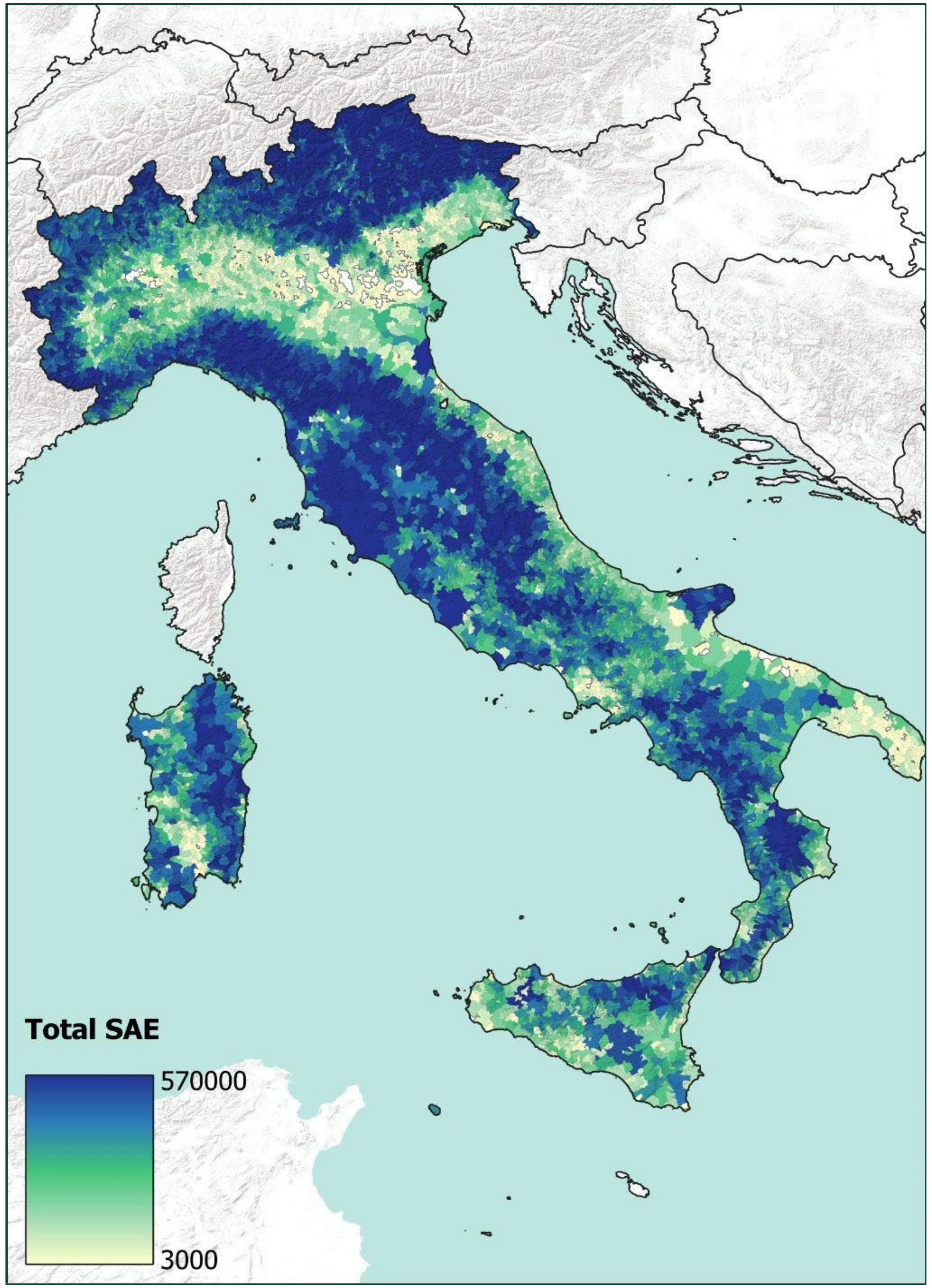




R^2 of 0.4 and an RMSE of 80 m^3 ha-¹. The optimal λ and shift parameters for the Box-Cox transformation resulted in λ = 0.39 and shift = 1



5. RESULTS

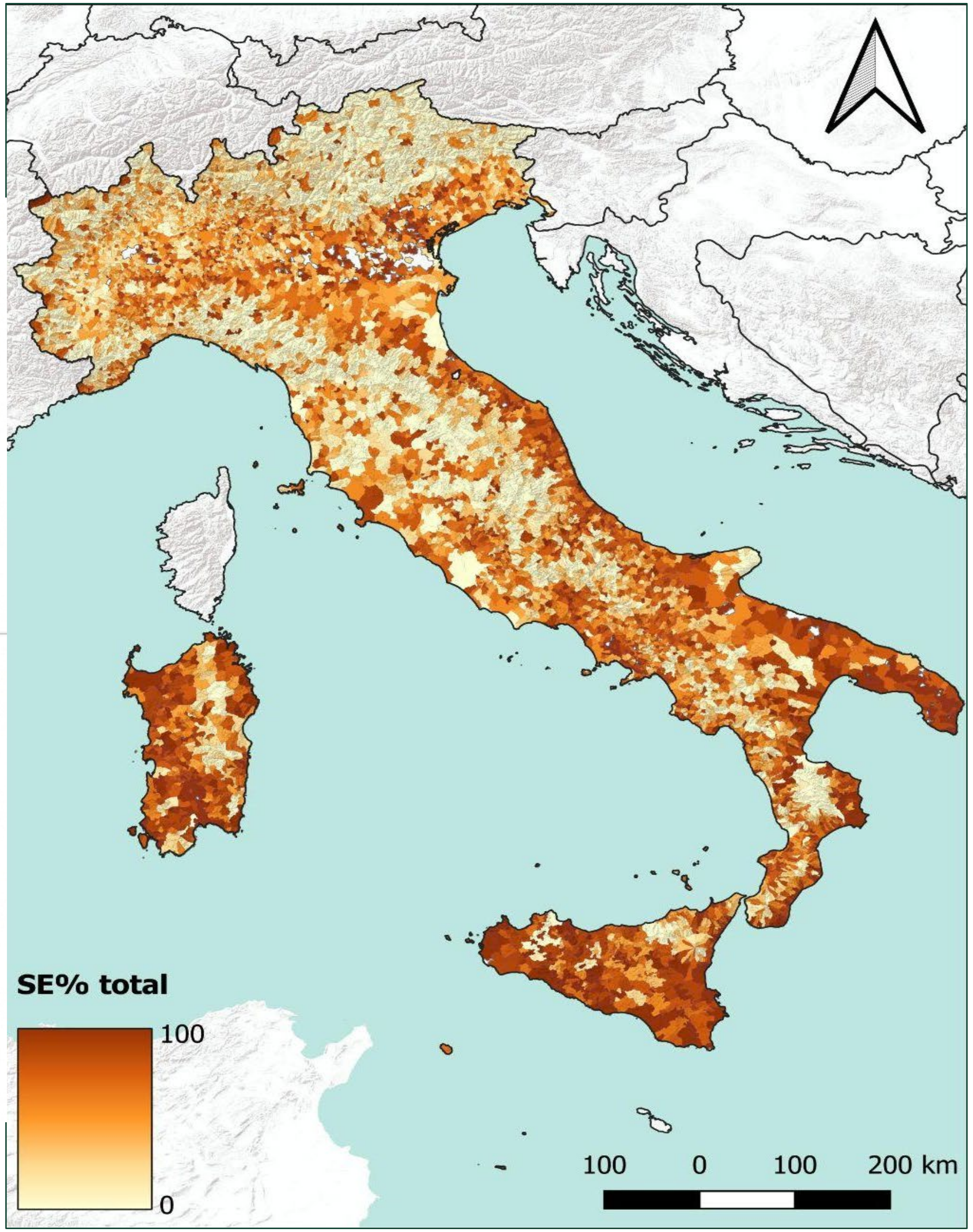


Extremes of Total Volume

Ranges from 15 m³ (Sestu) to an immense 5,643,856 m³ (San Giovanni in Fiore).

Extremes of Mean Volume

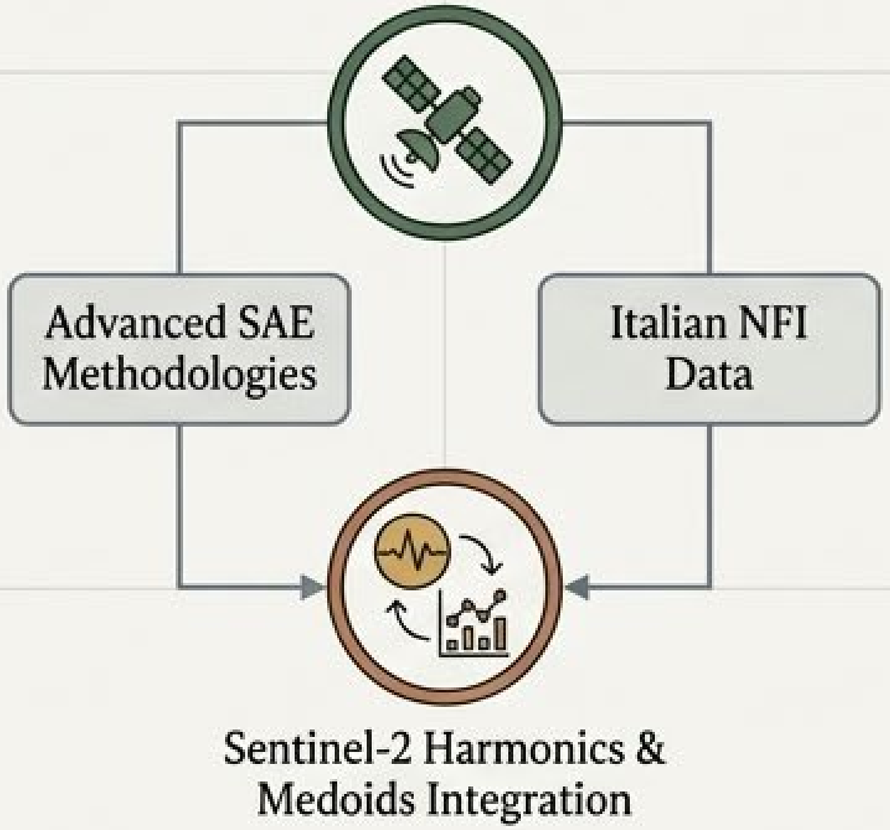
Averages range from 15 m³ ha⁻¹ (Pachino) to 449 m³ ha⁻¹ (Pagazzano).



Methodological Success

Methodological Success

- Successfully bridged advanced SAE methodologies with the Italian NFI using Sentinel-2 harmonics and medoids.



Statistical Rigor

- Quantified uncertainty at the municipality level—a first-of-its-kind achievement for Italy.

Operational Blueprint

- Provides a reproducible framework for transparent international climate reporting, local risk-informed forest management, and decentralized decision-making.

NEXT STEPS

Empirical Best Predictor (EBP)

The Traditional Standard

Core Model:	Nested-Error Linear Regression Model.
Mechanism:	Uses explicit linear equations and variable selection. Relies on data transformation (Box-Cox) to force normality and homoscedasticity.
Status:	The design-consistent benchmark in official statistics.

Mixed-Effects Random Forest (MERF)

The ML Challenger

Core Model:	Non-parametric Machine Learning + Mixed Models.
Mechanism:	Uses decision tree ensembles. Handles complex, non-linear interactions without explicit specification or data transformation.
Status:	The flexible, data-driven alternative emerging in socio-economics and healthcare.

Thanks for your attention

Gherardo Chirici (gherardo.chirici@unifi.it)
Elia Vangi (elia.vangi@unifi.it)

Funded by
The European Union