4th IFC and Bank of Italy Workshop on “Data Science in Central Banking”

Data science: the role of statisticians

Elisabetta Carfagna, University of Bologna
Chair of the Special Interest Group on Data Science
International Statistical Institute (ISI)

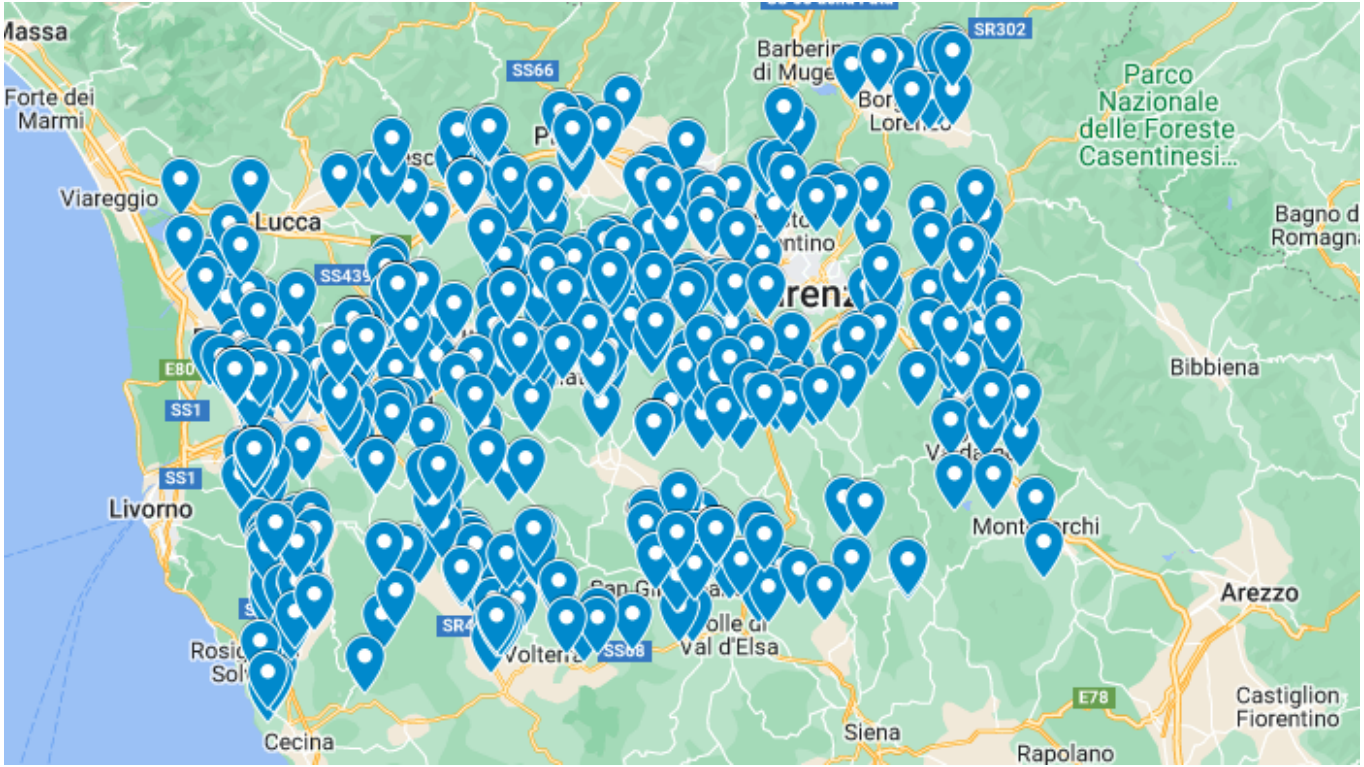
Alternative data sources and statistics

- **Non-probability data:**
  - Large datasets, new data sources, administrative registers, satellites and aircrafts, webcams, data voluntarily provided by the internet users, data harvested from the web
- **Questions**
  - How reliable are statistics produced through data science applied to big data?
  - Do machine learning classifiers outperform parametric models?
  - Which is the role of statisticians?

Data elaborated for estimating acreage of crops

- Satellite data:
  - 6 images from Sentinel 1 and Sentinel 2
  - 6 vegetation indexes for each of the Sentinel 2 images:
  - Normalized Difference Vegetation Index (NDVI)
  - Green Normalized Vegetation Index (GNDVI)
  - Two-band Enhanced Vegetation Index (EVI2)
  - Normalized Difference Water Index (NDWI)
  - Chlorophyll Red-Edge (ClRed-edge)
  - Soil-Adjusted Vegetation Index (SAVI)
- Digital elevation model

Area sampling frame - Un-clustered point sampling



Training and test data: ground data collected by Italian Ministry of Agricultural Food and Forestry Policies (MiPAAF) in 2016 (AGRIT project) 574 geo-referenced points in the north of Tuscany Region

Information collected on the ground on the sample of 574 geo-referenced points

- Land use
  - Agricultural land use
  - Cropping patterns
- Farm management
  - Soil cover
  - Tillage practices
  - Ground cover technique
  - Presence and kind of irrigation
  - Presence of fences

Esempio di 781 - SIEPI E FILARI inseriti nella matrice agricola

Esempio di 781 - SIEPI E FILARI pertinenti alla rete stradale principale: da non rilevare

Comparison of the performance of classifiers

- Regularized multinomial regression - penalized logistic classifier estimated using the maximization of the likelihood function combined with a lasso penalty term to deal with many explicative variables

$$Pr(G = k|X = x) = \frac{\exp\left(\beta_{k0} + b_k^T x\right)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + b_l^T x)}, k = 1, \dots, K-1$$

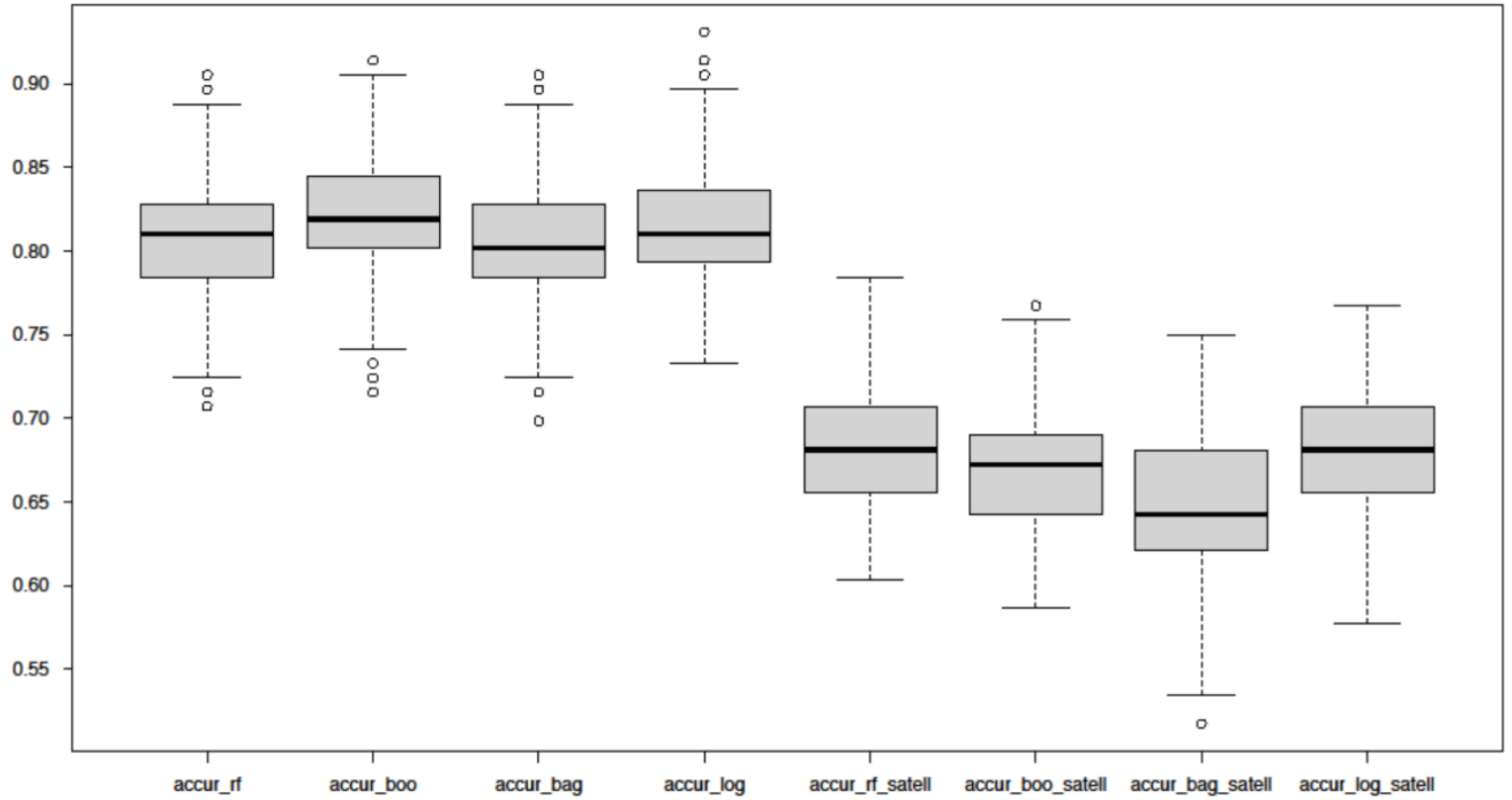
- 3 machine learning classifiers:
  - random forest
  - boosting
  - bagging

Each classifier has been repeated for different training and test sets, maintaining the same proportions of training and test sets (1000 simulations):

- 80% of the sample used for training the classifiers (459 points)
- 20% of the sample used for testing (115 points)

Distribution of accuracy of Random Forest, Boosting, Bagging and Regularized multinomial regression

- with all explicative variables
- with only explicative variables derived from satellite data

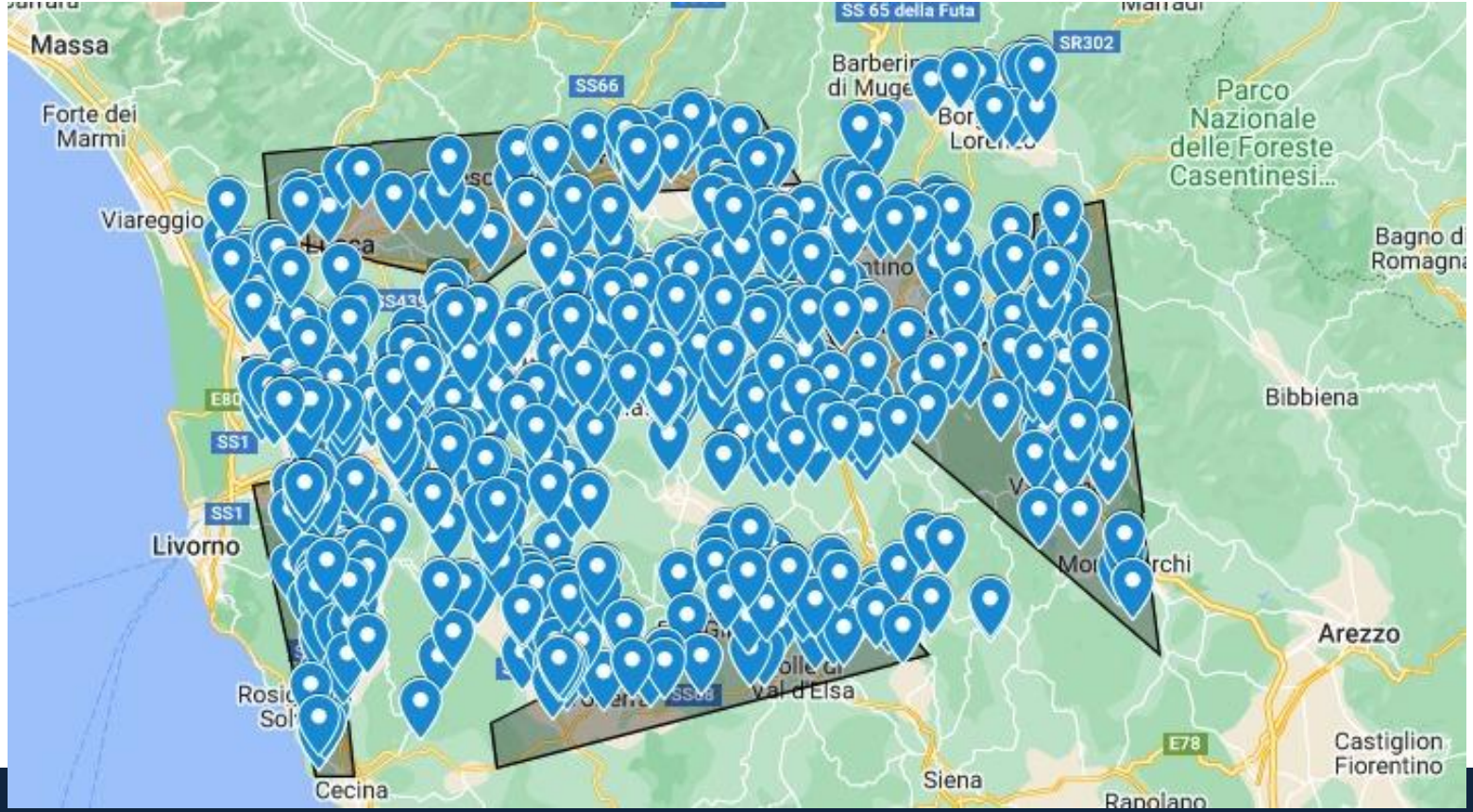


Median accuracy and Kappa value for the various classifiers

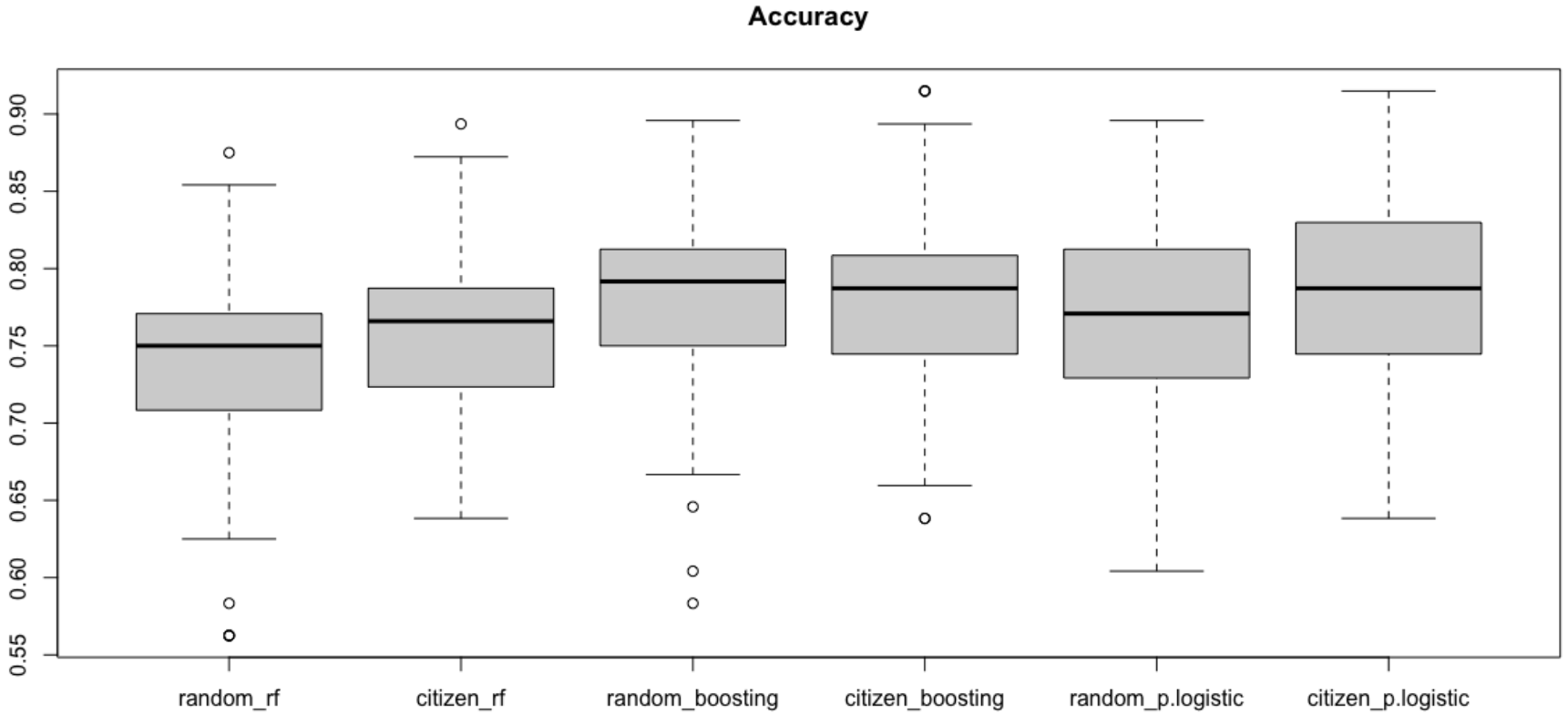
Vineyards; olive groves; sunflower; winter cereals, other crops

All explanatory variables			Satellite explanatory variables only		
Median accuracy of classifiers	Accuracy	Kappa	Median accuracy of classifiers	Accuracy	Kappa
Boosting	0.845	0.777	Random forest	0.691	0.526
Regularized multinomial regression	0.836	0.767	Regularized multinomial regression	0.689	0.521
Random forest	0.819	0.731	Boosting	0.677	0.528
Bagging	0.812	0.713	Bagging	0.652	0.495

Subset of ground data - 177 out of the 574 points close to cities and coastal areas selected to simulate a probable spatial distribution of data collected by citizens (citizen science)



Distribution of accuracy of Random Forest, Boosting, and Regularized multinomial regression with 177 sample points out of the 574: stratified random subsample and citizens subsample - 142 points (80%) for training – 35 points



Median accuracy and Kappa value for the various classifiers with 177 sample points: stratified random subsample and citizen subsample

Vineyards; olive groves; sunflower; winter cereals, other crops

Stratified random subsample			Citizen subsample		
Median accuracy of classifiers	Accuracy	Kappa	Median accuracy of classifiers	Accuracy	Kappa
Boosting	0.77	0.66	Boosting	0.77	0.63
Regularized multinomial regression	0.76	0.64	Regularized multinomial regression	0.79	0.65
Random forest	0.72	0.59	Random forest	0.74	0.55

Comparisons of area estimates based on: 1) citizen subsample 2) entire AGRIT sample

Total study area 741,450 ha

	Area estimate with citizen subsample (ha)	Area estimate with AGRIT sample (ha)	Area estimate citizen-AGRIT	Relative difference %
Other	148,654	152,364	-3,709	-2
Olive groves	52,896	47,043	5,853	12
Vineyard	50,577	37,368	13,209	35
Winter cereals	47,488	42,877	4,611	11
Sunflowers	5,285	9,070	-3,785	-42

Expansion factor based on photo-interpreted systematic sample (29,658 points) has been adopted also for estimates based on citizen subsample

Role of statisticians in data science

- **Population and variables of interest - under and over coverage**
- **Data collection methodology - training and test data**
- **Data quality - sampling and non sampling errors**
- **Quantify uncertainty - Classification accuracy - small classes**
- **Classifiers - assumptions - interpretation**
- **Models for combining proxies with real data - bias-variance tradeoffs**
- **Results analysis**

ISI welcomes IFC sessions at ISI regional conference, Malta, June 3-6, 2026

Thank you elisabetta.carfagna@unibo.it

Main references

Carfagna, E., Gallego, F.J. (2005) Using remote sensing for agricultural statistics. International Statistical Review, 73: 389-404.

Defourny, P. (2017) Land cover mapping and monitoring. In: J. Delincé (ed.), Handbook on Remote Sensing for Agricultural Statistics (Chapter 2). Handbook of the Global Strategy to improve Agricultural and Rural Statistics (GSARS): Rome.

Friedman J., Hastie T., Tibshirani R. (2001) The elements of statistical learning, volume 1. Springer series in statistics New York, NY, USA.

Gómez, C., White, J.C. & Wulder, M.A. (2016) Optical Remotely Sensed Time Series Data for Land Cover Classification: A Review. ISPRS Journal of Photo-grammetry and Remote Sensing, 116: 55–72.

Hamza M. and Larocque D. An empirical comparison of ensemble methods based on classification trees. Journal of Statistical Computation and Simulation, 75(8):629–643, 2005

Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, Brian Alan Johnson, (2019) Deep learning in remote sensing applications: A meta-analysis and review, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 152, 2019, Pag-es 166-177, ISSN 0924-2716

Pratesi M. (2023) Letter from the President, The Survey Statistician, 2023, Vol. 88, 4.

Tillé Y., Debusschere, M., Luomaranta, H., Axelson, M., Elvers, E., Holmberg, A. & Valliant, R. (2022) Some Thoughts on Official Statistics and its Future (with discussion), Journal of Official Statistics,38(2) 557-598. https://doi.org/10.2478/jos-2022-0026