Data science, citizen science and smart official statistics

Published Version:
Carfagna, E., Di Fonzo, G., Jona Lasinio, G., Canas Rodrigues, P. (2025). Data science, citizen science and smart official statistics. Cham : Springer [10.1007/978-3-031-64346-0_29].

Availability:
This version is available at: https://hdl.handle.net/11585/1014490 since: 2025-04-17

Published:

DOI: http://doi.org/10.1007/978-3-031-64346-0_29

(Article begins on next page)

Data science, citizen science and smart official statistics

Elisabetta Carfagna[1], Gianrico Di Fonzo[2], Giovanna Jona Lasinio[3], Paulo Canas Rodrigues[4]

Abstract. The discussion on advantages, disadvantages, limitations, and requirements of using alternative data sources integrated with probability sample surveys informs the debate in national and international statistical systems worldwide. The temptation to replace rigorous and costly data collection approaches with "smarter" ones is increasing. However, evaluating the reliability of statistics produced by elaborating alternative data sources is mandatory. In this work, we analyze the relationship between data science, new data sources, machine learning, citizen science and smart statistics, focusing on satellite data. We show that elaborating satellite data through parametric and machine learning classifiers does not always provide accurate statistics in complex landscapes, and machine learning classifiers do not systematically outperform parametric classifiers. Moreover, data collected by probabilistic samples play a crucial role. They should not be replaced by data collected by citizens without clear and strict guidelines in case statistics have to be produced.
Keywords: Data science, Citizen science, Sentinel satellites, Machine learning

1 Introduction

In today's digital era, the accessibility and analysis of extensive datasets from various sources, including administrative registers, satellites, social media, and digital platforms, have become increasingly feasible. [7] underscores the abundance of data in modern society, generated rapidly from diverse sources such as mobile phones, social media interactions, electronic transactions, and satellite imagery. These data sources present new opportunities for collection, processing, and storage, offering cost-effective alternatives to traditional statistical approaches involving surveys and questionnaires. However, this paper delves into critical questions regarding the reliability of statistics produced through data science applied to big data, exploring the relationship between data science, big data, new data sources, machine learning, and smart statistics. Additionally, the paper investigates whether machine learning classifiers outperform parametric models and explores the potential of integrating machine learning with citizen-collected data to replace probability sample surveys.

[1] University of Bologna, Department of Statistical Sciences, elisabetta.carfagna@unibo.it
[2] Sapienza University Roma and Italian Health Ministry, gianrico.difonzo@uniroma1.it
3 Sapienza University Roma, giovanna.jonalasinio@uniroma1.it
4 Department of Statistics, Federal University of Bahia, Salvador, BA, Brazil, paulocanas@gmail.com

2 The data set

A wide array of big data types, including administrative and unstructured data, are harnessed to generate statistics. However, unstructured data do not readily identify statistical units for integration with probability sample data, making its use for statistical purposes complex [8]. Relying solely on data science methods for processing such data can introduce significant biases that are difficult to quantify and mitigate. Satellite data, increasingly vital for analyzing agricultural and environmental phenomena, are based on raster-shaped pixels that may not align perfectly with territorial parcels, leading to complexities in classification tasks. This study uses a multitemporal satellite dataset covering a region north of the Tuscany Region, Italy, to evaluate its potential contribution to land cover estimation.

The dataset comprises six Sentinel 1 and Sentinel 2 images, six 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)), and a Digital elevation model derived from satellite data. For the same area, the Italian Ministry of Agriculture kindly provided real ground data collected in 2016 on 574 geo-referenced points selected using a probabilistic sampling design in the framework of the AGRIT project. Information about land use cropping patterns, and farm management (soil cover, tillage practices, ground cover technique, irrigation, and presence of fences) was collected on the 574 points selected according to a two-phase probability area sampling strategy: a regular grid with 500 meters side was overlaid on the territory in the first phase. The points at the cross of the grid were the first phase sample (aligned systematic sample in two dimensions) and were photo-interpreted on orthorectified aerial photos. Based on the photo interpretation, the points were attributed to the following land use strata: arable land, permanent crops, forage, scattered trees, forest and others. An additional stratification criterion was considered: low, medium, and high slope; thus, the intersection of the two stratification criteria generated the adopted stratification. The second phase sample (AGRIT sample) was a subset of the first phase, randomly selected according to specific sampling rates. The study area and the sample points are shown in Fig. 1.

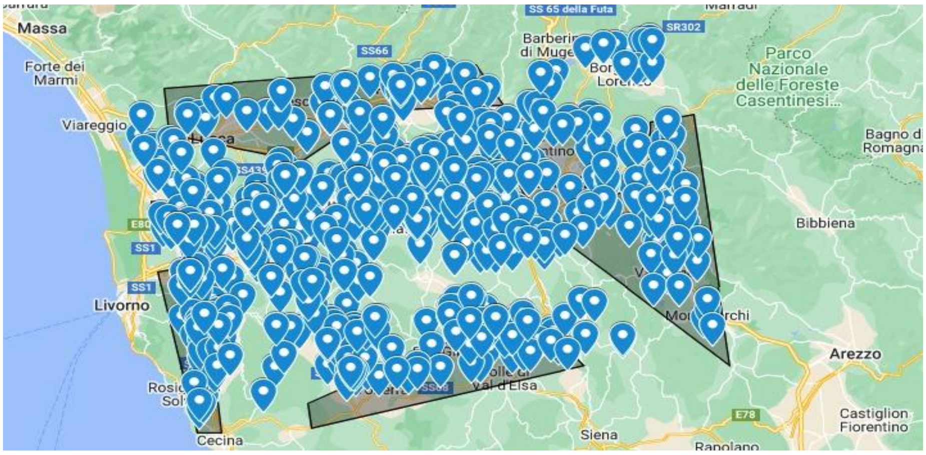

Fig. 1 Spatial distribution of the 574 second-phase sample points on which information about land use and farm management is collected. Dark areas contain a nonrandomly selected subset of the ground data (177 out) close to cities and coastal areas chosen to simulate citizen-provided data.

3 Performance of parametric and machine learning classifiers

[2] and [4] offer a comprehensive overview of the strengths and weaknesses associated with algorithms utilized for classifying satellite images. Linear classification methods are commonly employed for this purpose, with the parametric classification problem being addressed by partitioning the input space into regions labelled according to classification criteria [3]. In our study, we implemented the penalized logistic classifier, a classical parametric model, alongside three supervised machine learning classifiers [5]. The penalized logistic model employs likelihood maximization coupled with a lasso penalty term to manage a large number of explicative variables, selecting a subset of predictors while discarding the remainder. Additionally, we considered machine learning techniques such as bagging, random forest, and boosting. We implemented multiple splits in training and test sets, maintaining consistent proportions (1000 simulations). Specifically, 80% of the sample was allocated for training the classifiers (459 points), while 20% was reserved for testing (115 points). Boosting emerges as the most accurate classifier when ground data are incorporated among the explanatory variables, while random forest outperforms others when only remote sensing data are utilized. Penalized logistic multinomial regression ranks second in median accuracy, both with all explanatory variables and with satellite-derived variables alone. However, bagging exhibits the highest dispersion and lowest accuracy values among the classifiers assessed.

The achieved accuracies are notably lower when solely satellite-derived variables are considered. These findings contrast with prevailing trends observed in machine learning applications on remote sensing data, as reported in a 2019 review by [6]. These authors conducted a comprehensive review of machine/deep learning applications across various remote sensing datasets, particularly emphasizing publicly available benchmark datasets with extremely high resolution (pixel size < 10 meters). However, such findings may not fully reflect practical applications in complex landscapes, where satellite data classification serves as a proxy for crop acreage estimation, complementing ground data at the estimator level rather than offering a reliable estimate of crop acreage itself.

Table 1 Median accuracy and Kappa value from the experiment. Classifiers are ordered according to their accuracy with all explanatory variables and satellite explanatory variables only

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
Penalized logistic multinomial regression	0.836	0.767	Penalized logistic 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

4 Citizen science versus probability sample surveys

In response to the growing trend of utilizing citizen-provided data and minimizing reliance on expert-led ground data collection efforts, we deliberately selected a non-random subset of 177 points from the original 574-point dataset. These points were strategically chosen to mimic the spatial distribution of data collected spontaneously by citizens, particularly focusing on areas near cities and coastal regions. From this subset, 142 points were allocated for training the classifiers, while the remaining 35 points were reserved for testing their accuracy. Additionally, to assess the impact of sampling methodology, we selected a stratified random subsample of 177 points from the original dataset. Due to the reduced sample size, the dispersion of accuracy across different training and test set splits is notably larger. This comparison enables an evaluation of classifier performance under varying sampling approaches, shedding light on the effectiveness of utilizing citizen-provided data instead of traditional ground data collection methods. The overall median accuracies and Kappa indexes of the classifiers with the probabilistic subsample and with the citizen sample are very similar (see Tab 2). According to these results, a purposive sample of data collected by citizens seems to work as well as a stratified random sample for training and testing the classifier. This result can be misleading, since the purposive sample selection can generate a quite homogeneous sample and thus can facilitate the classification of the satellite images on sample units. However, when the trained classifier is applied to the whole study area, the result can be strongly biased.

Table 2 Comparison of median accuracies and Kappa values for the various classifiers, using the subset of points close to cities and along the coast and the subset of points selected according to stratified random sampling.

	Stratified random subsample		**Citizen science subsample**	
Classifier	**Accuracy**	**Kappa**	**Accuracy**	**Kappa**
Boosting	0.77	0.66	0.77	0.63
Penalized logistic multinomial regression	0.76	0.64	0.79	0.65
Random forest	0.72	0.59	0.74	0.55

In order to produce land use statistics, the classification of satellite data cannot be considered as the only data source, since pixel counting is known to be a biased estimator, particularly because of mixed pixels. Therefore, ground data are essential both for training and testing classifiers and for estimating the acreage of various land uses. Then, these estimates are improved combining ground-based estimates with classified data, treating the latter as auxiliary variables in a calibration or regression estimator [1].

To assess the possibility to replace the probability sample with citizen data, we compared the acreage estimates of different crops obtained from the non-probability sample with estimates derived from the entire AGRIT sample, consisting of the above mentioned 574 points selected via stratified random sampling. The expansion factor for the stratified random sample was determined based on the photo-interpreted systematic sample (29,658 points), and the same factor was applied to the estimates derived from the citizen subsample. The comparison presented in Table 3 reveals significant discrepancies in acreage estimates. Sunflower acreage is notably underestimated, while overestimations are observed for winter cereals, olive groves, and vineyards. It is important to note that in practical applications, expanding data to the entire population poses challenges, especially when collected by citizens without specialized skills and lacking a probability sample selection scheme. Consequently, the observed underestimations and overestimations may be even more pronounced in real-world scenarios.

Table 3 Difference in square kilometers and per cent between the acreage estimate of the various land uses based on the non-probabilistic sample design and the probabilistic sample design (AGRIT sample)

Vegetation Class	Area estimated with citizen subsample (ha)	Area estimated with AGRIT sample (ha)	Area estimated citizen-AGRIT	Relative difference
Other	148,654	152,364	-3,709	-2.43
Olive groves	52,896	47,043	5,853	12.44
Vineyard	50,577	37,368	13,209	35.35
Winter cereals	47,488	42,877	4,611	10.75
Sunflowers	5,285	9,07	-3,785	-41.73

5 Concluding remarks

In this paper we have shown that the classification of Sentinel satellite data does not provide high accuracies in complex landscapes. We have also noticed that machine learning classifiers do not systematically outperform parametric classifiers. In fact, boosting is the most accurate classifier with the entire data set and all explanatory variables, random forest is the most accurate classifier with satellite explanatory variables only, and the penalized logistic multinomial regression has the second highest median accuracy both with all explanatory variables and with the explanatory variables derived from satellite data only. Furthermore, ground data collected on probability samples play a vital role in estimating land use acreage. In fact, the classification based on a purposive sample can be as accurate as the one based on a probability sample but can be affected by the bias of the non-probability sample. Thus, the land use estimates based on regression or calibration estimators with the classification as auxiliary variables are affected by the bias of the non-probability sample. It is important to note that expanding data to the entire population when collected by citizens who lack specialized skills and do not follow a probability sample selection scheme is problematic; consequently, observed discrepancies may be exacerbated in real-world applications.

References

1. Carfagna, E., Gallego, F.J. (2005) Using remote sensing for agricultural statistics. International Statistical Review, 73: 389-404.
2. 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.
3. Friedman J., Hastie T., Tibshirani R. (2001) The elements of statistical learning, volume 1. Springer series in statistics New York, NY, USA.
4. Gómez, C., White, J.C. & Wulder, M.A. (2016) Optical Remotely Sensed Time Series Data for Land Cover Classification: A Review. ISPRS Journal of Photogrammetry and Remote Sensing, 116: 55–72.
5. 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
6. 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, Pages 166-177, ISSN 0924-2716
7. Pratesi M. (2023) Letter from the President, The Survey Statistician, 2023, Vol. 88, 4.
8. 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