Data Science Structure and the role of Statistics

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

Abstract
Data science has emerged as a very strong, visible, and publicly recognized label for problem-solving using ever-growing, large datasets and new data sources such as administrative registers, satellites and aircrafts, webcams, data voluntarily provided by the internet users, data harvested from the web and so on. The applications of data science tools range from earth observation to official statistics. The discussion on advantages, disadvantages, limitations, and requirements of the use of alternative data sources integrated with probability sample surveys is informing the debate in national and international statistical systems all over the world. Of course, the temptation of replacing the traditional data collection approach with "smarter" ones is strong. In this paper we address the evaluation of the reliability of statistics produced through the elaboration of big data, focusing on their structure. We analyse the relationship between data science, new data sources, machine learning, citizen science and smart statistics, focusing on satellite data. We show that the elaboration of satellite data through parametric and machine learning classifiers does not provide accurate statistics in complex landscapes and machine learning classifiers do not systematically outperform parametric classifiers. Moreover, data collected on a probability sample play a crucial role and should not be replaced by data collected by citizen without clear and strict guidelines, in case statistics have to be produced.

Keywords: Data science, Citizen science, Sentinel satellites, Machine learning

1. Introduction

Large datasets, new data sources, administrative registers, satellites and aircrafts, webcams, data voluntarily provided by the internet users, data harvested from the web and so on are becoming easier to access and elaborate. According to Pratesi (2023): "In our digital era, data are everywhere: new sources, such as mobile phones, social media interactions, electronic commercial transactions, sensor networks, smart meters, GPS tracking devices, or satellite images, produce new information at an incredible speed. Digital technologies offer new opportunities for data collection, processing, storage."
These kinds of data are cheaper respect to the traditional statistical approach that implies delineating the aim of the survey, developing the questionnaire to be used, recruiting trained personnel that ask questions or make measurements (according to the kind of phenomenon to be surveyed) on a probability sample of statistical units.
Of course, the temptation of replacing the traditional data collection approach with "smarter" ones is strong. In this paper we address some crucial questions: "How reliable statistics produced through data science applied to big data are?", "How can this reliability be measured?", "Which is the impact of the structure of the data?", "Which is the relationship between data science, big data, new data sources, machine learning, and smart statistics?", "Do machine learning classifiers outperform parametric models?", "Can machine learning applied to big data integrated with data collected by citizens replace probability sample surveys?"
In paragraph 2 the dataset used for addressing these questions is described. It includes probability sample data collected on the ground by the Italian Ministry of Agriculture

[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

(MiPAAF) in the north of the Tuscany Region, Italy and Sentinel satellites data from the Copernicus project of the European Space Agency.
In paragraph 3 the performance of one parametric and three machine learning classifiers is evaluated on a real dataset in a complex landscape, adopting supervised classification of Sentinel satellites data, using the ground data as training and test sets. Paragraph 4 faces the impact on the accuracy of the classification and on statistical estimates of the use of data collected on points close to cities and coastal areas, which simulate a probable spatial distribution on the ground of data collected and spontaneously provided by citizens. Finally, some conclusions are drawn.

2. Data structure and statistics

Nowadays, a variety of different kinds of big data can be exploited for producing statistics, ranging from administrative data to completely unstructured data. Administrative data are generally structured data; for example, the ones collected by municipalities are based on the same statistical unit of the population censuses. Although collected for administrative and not statistical purposes and affected by under and over coverage, administrative data can be used for producing statistics, provided that traditional probability sample data are collected for estimating the level of under and over coverage and for collecting information non included in the administrative registers.
Other kinds of non-conventional big data used for describing social or physical phenomena are unstructured data and do not allow identifying statistical units to be integrated with probability sample data. Relying on the elaboration of these data through data science methods regardless of the mentioned issues can generate strong biases hard to measure and remove. An important example is offered by satellite data, which are gaining increasing importance to analyse agricultural and agri-environmental phenomena and monitor SDGs (Olofsson *et al*. 2011). Satellite data are based on raster shaped elementary units called pixels which are not consistent with the parcels of the territory and can be smaller or larger then the parcels, according to the kind of satellite data and the complexity of the territory under consideration. Moreover, the big amount of information collected by satellite data is just a proxy of the needed one for agricultural and agri-environmental monitoring and the classification of these data requires ground data for training the classifier and for testing the classification (Carfagna and Di Fonzo, 2021).
Besides the level of structure of the data, another important aspect is the use of big data; namely, if the big data are analysed for producing official statistics, particular attention has to be devoted to the characteristics of the data, their quality, and the level of bias they can generate (Tillé *et al.* (2022)).
In this paper, we use a set of multitemporal satellite data covering an area in the north of the Tuscany Region, Italy, for assessing the contribution they can give to land cover estimation: 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)), and a Digital elevation model derived from satellite data.
For the same area, the Italian Ministry of Agriculture kindly provided real data collected on the ground on points selected by a probability sampling survey. 574 geo-referenced points in the north of the Tuscany Region, Italy, on which the land use was assessed on the ground by the Italian Ministry of Agricultural Food and Forestry Policies in 2016, in the framework of the AGRIT project, aimed at producing estimates of acreage and yield for the main Italian crops and estimating some agri-environmental parameters. For these points, the Ministry collected in situ information about land use: agricultural land use and cropping patterns, and farm

management: soil cover, tillage practices, ground cover technique, irrigation, presence of fences.
The AGRIT project collected ground information on un-clustered sampling points adopting 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 photointerpretation, the points were attributed to the following land use strata: arable land, permanent crops, forage, scattered trees, forest and other. 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 one, randomly selected according to the sampling rates described in table 1.

Table 1 Sampling rate of the second phase sample points in the different strata

	Arable land	Permanent crops	Forage	Scattered trees	Forest	Other
Low slope (0-7.5 %)	8.25 %	8.25 %	5.00 %	8.25 %	0.00 %	0.00 %
Medium slope (7.5-15 %)	8.25 %	14.00 %	5.00 %	8.25 %	0.00 %	0.00 %
High slope (higher than 15 %)	4.50 %	14.00 %	3.00 %	4.50 %	0.00 %	0.00 %

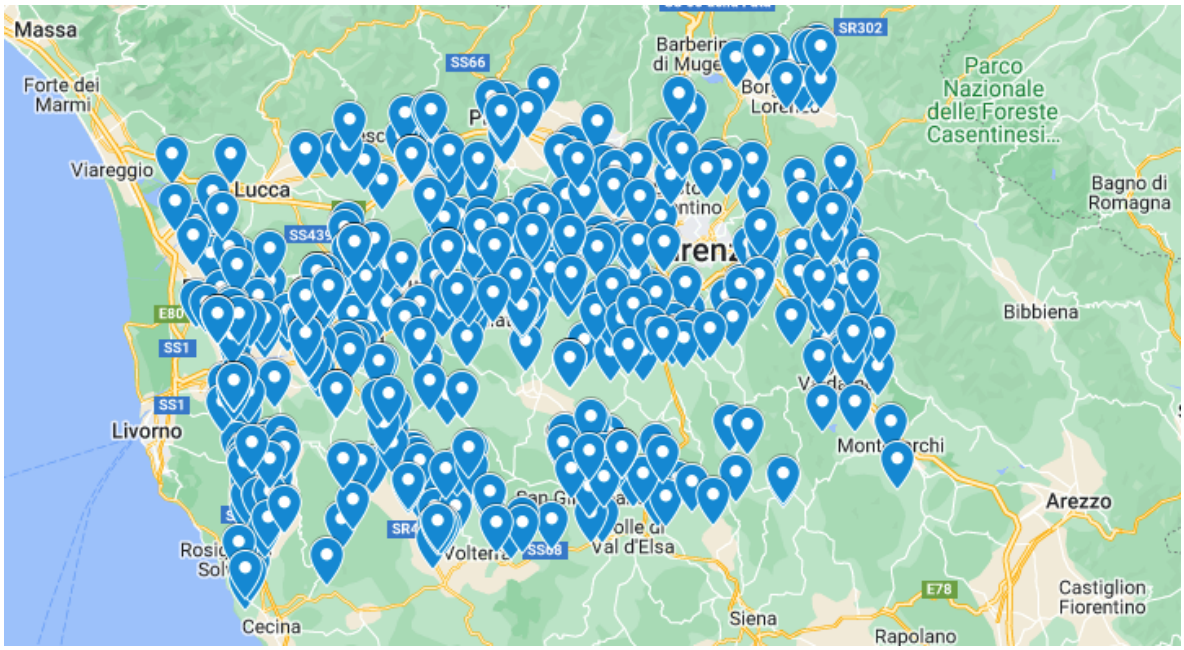

Figure 1 shows the uneven spatial distribution of the 574 second phase sample points on which information about land use and farm management is collected. The uneven spatial distribution of the points is due to the differentiated sampling rate in the various strata described in table 1.

3 Performance of classifiers on a real dataset in a complex landscape

An overview of the strengths and weaknesses of algorithms used for classifying satellite images is provided by Gómez et al. (2016) and Defourny (2017).

Linear classification methods are used for classifying remote sensing data. Since the predictor G(x) takes values in a discrete set G, the parametric classification problem can be solved by dividing the input space into a collection of regions labelled according to the classification (Friedman et al. 2001).

We applied the penalized logistic classifier from the class of classical parametric models and three supervised machine learning classifiers (Hamza et al. 2005). The penalized logistic model is estimated using the maximization of the likelihood function combined with a lasso penalty term to deal with a large number of explicative variables. This model retains a subset of the predictors and discards the rest.

The machine learning techniques we considered are bagging, random forest and boosting.

Each classifier has been repeated for different training and test sets, maintaining the same proportions (1000 simulations): 80% of the sample used for training the classifiers (459 points) and 20% of the sample used for testing (115 points).

Figure 2 shows the distribution of the overall accuracy levels corresponding to the different sample selections for random forest, boosting, bagging and penalized logistic multinomial regression. The overall accuracy is computed taking into consideration few classes, among which a group of crops due to the limited presence of some minor crops: vineyards; olive groves; sunflower; winter cereals, other crops.

The first set of box plots on the lefthand side of the figure shows the accuracies obtained with all explicative variables, that is including the variables concerning land use and farm management. Clearly, this kind of information is available only on the sample points; thus, the distribution of the accuracy achievable in an operational project in a complex landscape like the one we have taken into consideration is probably similar to those shown in the righthand side of the figure.

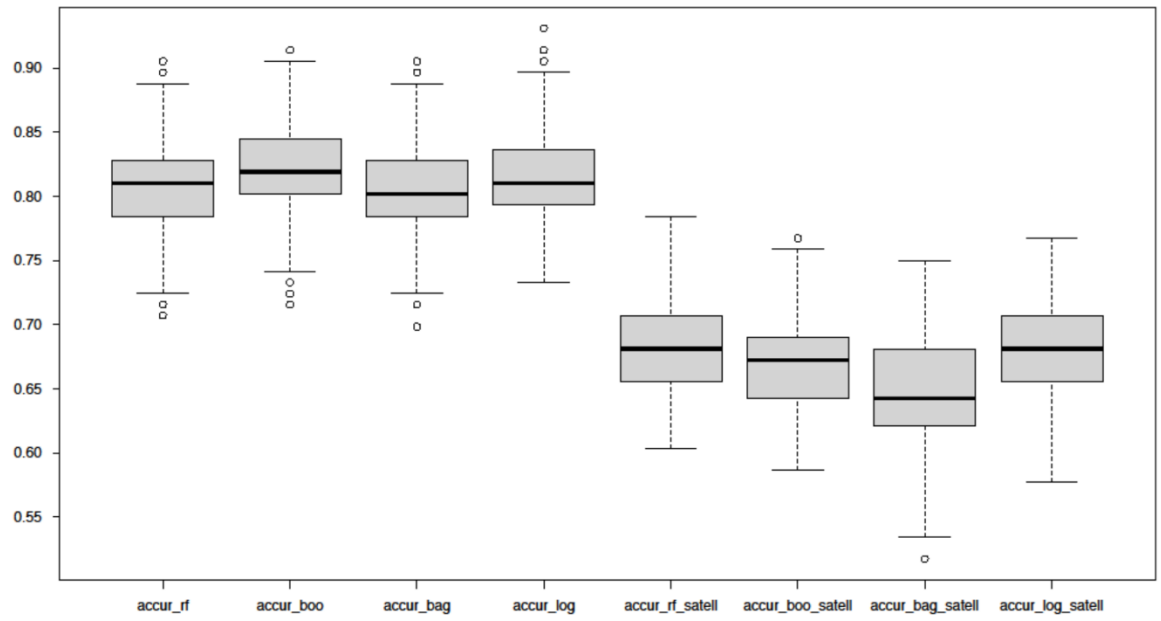

Figure 2. Distribution of accuracy of random forest, boosting, bagging and penalized logistic multinomial regression with all explicative variables and with only explicative variables derived from satellite data

When ground data are included in the set of explanatory variables, boosting results to be the most accurate classifier and random forest is the most accurate classifier when only remote sensing data are taken into consideration. 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. The accuracy dispersion is similar for the different classifiers, except for bagging which shows the highest dispersion and the lowest accuracy values.
The accuracies obtained are low when only variables generated by satellite data are taken into consideration. These results are not in line with the ones reached in most applications of machine learning on remote sensing data according to a review made in 2019 (Lei et al., 2019). These authors made a review concerning applications of machine/deep learning to various kinds of remote sensing data and most land use/cover applications concerned publicly available benchmark image datasets with extremely high resolution (pixel size much smaller than 10 meters) and did not focus on practical applications in complex landscapes, in which the classification of satellite data has to be considered as a proxy of the crops acreage to be combined with ground data at the estimator level and not as a reliable estimate of the crop acreage.

Table 2 Median accuracy and Kappa value for the various classifiers with all explicative variables and with only explicative variables derived from satellite data. The classifiers are ordered according to their accuracy

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 Probability sample surveys replaced by citizen science
Given the pressure to reduce as much as possible ground data collection and to use data spontaneously provided by citizens instead of planning ground data collections conducted by experts, out of the 574 points, we have selected a non-random subset of 177 points close to cities and coastal areas, to simulate a probable spatial distribution on the ground of data collected and spontaneously provided by citizens.

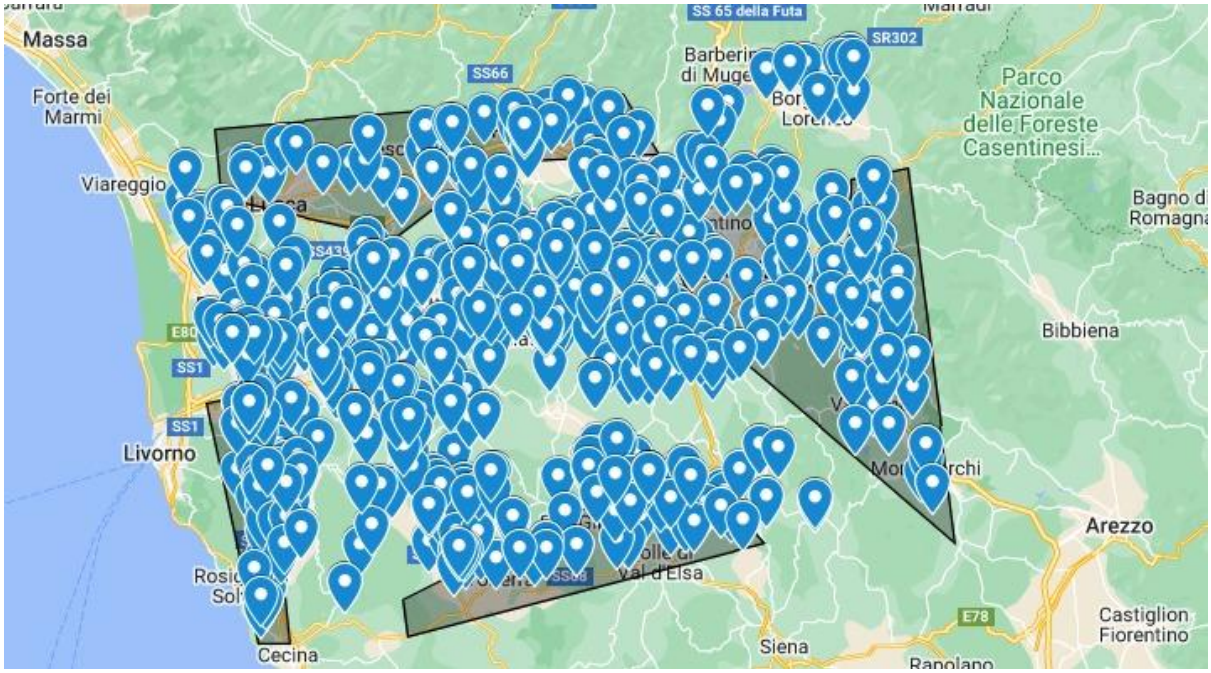

Figure 3 Spatial distribution of all sample points and of the nonrandomly selected subset of the ground data (177 out of the 574 points on darker areas) close to cities and coastal areas selected to simulate a probable spatial distribution on the ground of data collected and spontaneously provided by citizens

Out of the 177 points, 142 points were used for training the classifiers and 35 for testing their accuracy. We also selected a stratified random subsample of 177 points out of the 574 to compare the accuracy of the classifiers when the probability and non-probability subsets are used. Given the smaller number of points taken into consideration, the dispersion of the accuracy with different splits of the dataset into training and test sets is obviously larger, as shown in figure 4.

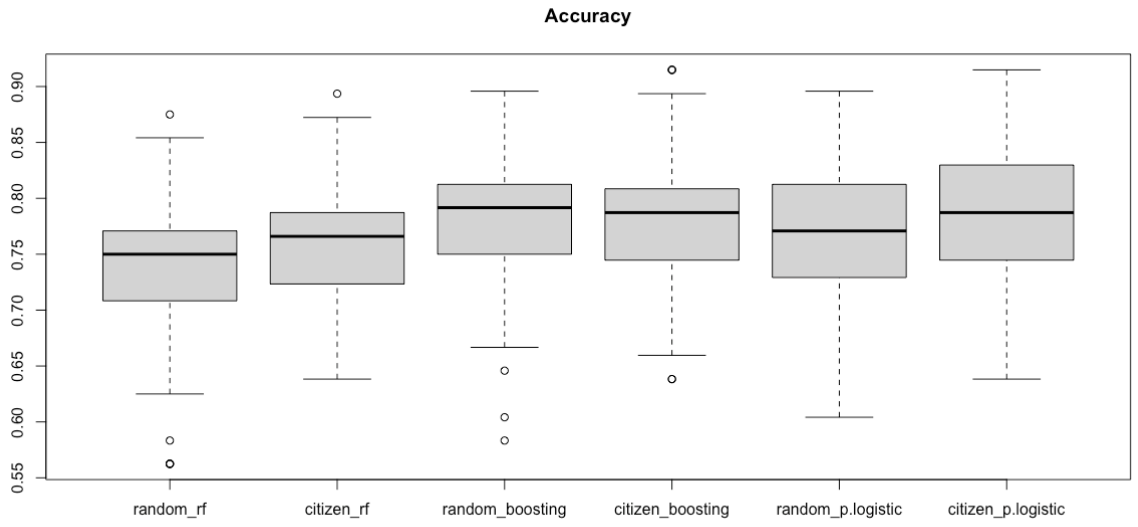

Figure 4. Distribution of the accuracy of random forest, boosting, and penalized logistic multinomial regression with different splits of the dataset into training and test sets, with the subset of 177 points selected according to stratified random sampling and the subset of 177 points close to cities and along the coast

The overall median accuracies and Kappa indexes of the classifiers with the probabilistic subsample and with the citizen sample are very similar. According to these results, a purposive sample of data collected by citizens without specific skills seems to work as well as a stratified random sample for training and testing the classifier.

Table 3 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 subsample		
Median accuracy of classifiers	Accuracy	Kappa	Median accuracy of classifiers	Accuracy	Kappa
Boosting	0.77	0.66	Boosting	0.77	0.63
Penalized logistic multinomial regression	0.76	0.64	Penalized logistic multinomial regression	0.79	0.65
Random forest	0.72	0.59	Random forest	0.74	0.55

Ground data are not used only for training and testing the classifiers; in fact, in most cases, the acreage of the different land uses must be estimated. Since pixel counting is well known to be a biased estimator, the estimate of the acreage is obtained by combining the estimate based on ground data and classified data treated as an auxiliary variable in a calibration or regression estimator (Carfagna and Gallego, 2005). Thus, we have compared the estimate of the acreage of the various crops based on the non-probability sample and the estimates obtained with the whole AGRIT sample, that is the 574 points selected according to stratified random sampling showed in Figure 1.
The expansion factor for the stratified random sample is based on the photo-interpreted systematic sample (29,658 points) (the first phase sample of the AGRIT project) and the same expansion factor has been adopted for the estimates based on the citizen subsample.
The comparison of the estimates in table 4 highlights a considerable underestimate of the acreage of sunflowers and overestimates of the acreage of winter cereals, olive groves and vineyards. Consider that in real applications, no information is available for expanding the data to the entire population, since no phase one sample of the AGRIT project is available if data are collected by citizens without specific skills and no probability sample selection scheme, thus the underestimates and the overestimates could be even higher.

Table 4 Difference in square kilometres and per cent between the acreage estimate of the various land uses based on the non-probability sample and on the probability sample (AGRIT sample)

	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	-2.43
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	9,070	-3,785	-41.73

5 Concluding remarks

In this paper we have taken into consideration big data (Sentinel satellite data) with a structure different from the one of statistical units (pixels instead of parcels) and we have shown that the elaboration of this kind of data through data science methods like parametric and machine learning classifiers does not provide accurate statistics 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.

We have analysed the impact of non-random selection of a subset of points, mainly points close to cities and coastal areas, to simulate a probable spatial distribution on the ground of data collected by citizens. If the acreage of the main land uses is estimated based on this nonprobability subsample, the results are considerably different from the ones obtained with the whole probability sample; thus, ground data collected on a probability sample play a crucial role and should not be replaced by data collected by citizen without clear and strict guidelines, in case statistics have to be produced.

References

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

Carfagna E. and Di Fonzo G. (2021) Land cover/use analysis and modelling in Postiglione, P., Benedetti, R., & Piersimoni, F. (Eds.). Spatial Econometric Methods in Agricultural Economics Using R (1st ed.). CRC Press. https://doi.org/10.1201/9780429155628, ISBN: 978-0-429-15562-8 (ebk) pp. 88-107

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 Photogrammetry 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, Pages 166-177, ISSN 0924-2716
Olofsson, P., Kuemmerle, T., Griffiths, P., Knorn, J., Baccini, A., Gancz, V., et al. (2011) Carbon implications of forest restitution in post-socialist Romania. Environmental Research Letters, 6, 045202.
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