An automatic Sentinel-2 forest types classification over the Roncal Valley, Navarre (Spain)

A. Fernandez-Carrillo[*a], D. de la Fuente[a], F.W. Rivas-Gonzalez[a], A. Franco-Nieto[a]
[a]Remote Sensing Services and Exploitation Platforms Division, GMV, C/ Isaac Newton 11, 28760 Tres Cantos (Madrid), Spain

ABSTRACT

Forests cover 36.5% of Spanish land. Natural and man-induced disturbances are causing important changes in species distribution. As Spanish National Forest Inventory is updated every 10 years, a more recurrent periodic data source providing information on species distribution is needed in order to predict changes in forest area and composition. Remote Sensing meets this demand, as it provides periodic and spatially continuous data on forest status. In this context, MySustainableForest (MSF) H2020 project aims at providing remote sensing-based geo-information services through a web service platform.

One of MSF products is a classification of main forest types, whose preliminary development was tested over a 950 km^2 area located in Northern Spain. A Random Forest model was trained with data delineated with the help of local forest data. The output was validated using stratified k-fold cross-validation. Validation metrics were computed from the confusion matrix for each class separately and for the total set of classes.

Although overall metrics were high (OA = 95%; DC = 85.1%), they varied significantly for different classes (e.g., *Fagus sylvatica* was classified with higher accuracy than *Pinus nigra*, which was mainly confused with other *Pinus* species), showing that species with higher seasonal variations were easier to identify. Random Forest feature importance ranking showed that bands in the near-infrared (NIR) and shortwave-infrared (SWIR) wavelengths were essential to discriminate forest species, since they explained more than 40% of the variations alone and 82% in combination with Red wavelength.

Keywords: Sentinel-2, forest types, forest classification, Random Forest, Mediterranean-Alpine forest, forest management

1. INTRODUCTION

According to the Spanish National Forest Inventory, forests cover around 18.5 million ha, which implies 36.5 % of Spanish land. Forests constitute not only a source of wood and other economic goods, but they also provide natural (e.g., climate change mitigation through carbon sequestration) and aesthetic values[1,2]. Although forested area is currently growing in Spain, natural and man-induced disturbances increasingly threaten forests[3,4], thus inducing important changes in species distribution.

Field-based traditional inventories are costly and each update requires several years to be completed. These issues are even more acute in mountain areas. As Spanish National Forest Inventory is updated every 10 years, a more recurrent data source providing information on species distribution is needed in order to predict changes in forest area and composition. Remote Sensing (RS) meets this demand, as it provides periodic and spatially continuous data on forest status.

MySustainableForest (MSF)[5] in an EU-H2020 project which is developing remote sensing-derived geo-information services for integrated forest management, at pre-commercial stage, to be provided through a web service platform. The objective is to support forest stakeholders' decision-making processes and operations based on Earth Observation data. Services have been produced for different areas in Europe and across a variety of management protocols. It engages relevant European public and private forest stakeholders. It is expected that end-users will optimize their operations up to 10% and participating companies' headcounts increase up to 20%. MSF portfolio. MSF includes services on forest site characterisation, wood characterisation, biomass and CO_2 stocking, forest condition, ecosystem vulnerabilities and forestry accounting. This study presents the development of the Main Forest Types product for a study area located in the Pyrenees.

[*]aafernandez@gmv.com; phone: +34 91 807 21 00

MSF Main Forest Types is an optical-based dominant species forest map at 10 m spatial resolution based on Sentinel-2 (S2) multi-spectral data.

RS-based forest types classifications have been carried out using different sensors and in different geographic regions. In the last two years, the use of S2 images for this purpose has notably increased[6–9]. Among the algorithms used, Random Forest (RF)[10] is clearly the most common. The multi-temporal availability of S2 data makes it suitable to distinguish forest species with different phenology, which would be otherwise complex using a single-date image. S2 classification studies achieve accuracies in the range from 80 to 90%[6–9].

The aim of the study is to explore the extent to which S2 data can provide information about forest species composition, complementing field-based inventories. The specific objectives are: i) assess the effectiveness of Random Forest algorithm to classify main forest types in an area characterised by different altitudinal zones, and ii) explore the importance of different wavelengths in forest types discrimination. The results of this study may be used by forest managers to implement new processes that allow improving forest inventories.

2. STUDY AREA

The Roncal Valley (Figure 1) is located in the Northeast of Navarre, Spain, in the Western Pyrenees, forming a natural border between Spain and France. Valley heights range from 629 to 2,428 m. The Larra massif, one of the most extensive karst in Europe, is to the Northeast. The Esca River flows north-south, 52 km down the valley.

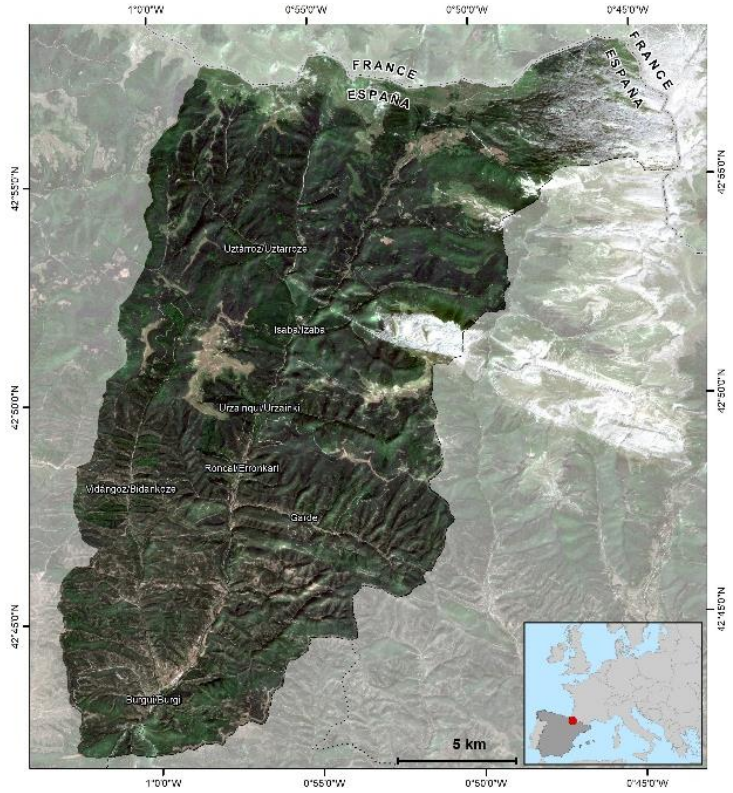

Figure 1: Roncal Valley location

Subalpine bioclimatic domain is present in the high peaks while Sub-Mediterranean conditions prevail in the low areas. Vegetation coverage is complex due to height gradient, as well as to Atlantic and Mediterranean influences. The main forest systems in the Roncal Valley[11] are:

- Subalpine *Pinus uncinata* forests, found at altitudes between 1,600 and 2,400 m. Most *P. uncinata* dominate the Larra massif.
- Montane *Abies alba* forests, found at altitudes between 1,200 and 1,600 m. *A. Alba* is commonly found in association with *Fagus sylvatica*, conditioned by terrain aspect.

- Montane *Fagus sylvatica* forests, found at altitudes between 1,200 and 1,700 m. They are in limestone lands and rainy orientations. Silver firs appears frequently as an accompanying specie.
- Montane *Pinus sylvestris* var. *pyrenaica* forests are found at altitudes between 700 and 1,600 m.
- Montane *Quercus pubescens* forests are found at altitudes between 700 and 1,300 meters. It is very commonly found in mixed forests with *F. sylvatica*.
- *Pinus nigra* forests are the result of reforestation carried out to protect the soil from erosion in those places where native forest had disappeared.

The Roncal Valley is mainly covered by natural forests with remarkable difficulties to be managed, due to steep slopes and a dense bush understorey that drastically prevent access for the acquisition of field data, construction of infrastructures and mechanisation of silvicultural treatments. Remote Sensing can help to monitor the state of forests in this area, facilitating sustainable forest management to stakeholders.

3. METHODS

3.1 Image acquisition and preprocessing

Two Sentinel-2 Level-2A (orthorectified bottom-of-atmosphere reflectance) were used, one per season: winter (t0) and summer (t1), with less than 10% cloud cover. Image preprocessing involved resampling bands to 10m and clipping to the AOI.

The following vegetation indices were generated:

- Normalized Difference Vegetation Index (NDVI, Eq. 1), a spectral index of plant greenness or photosynthetic activity.

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \tag{1}$$

- Soil Adjusted Vegetation Index (SAVI, Eq. 2)[12], a modification of the NDVI to correct the influence of soil brightness when vegetative cover is low.

$$SAVI = \frac{(NIR - Red)}{(NIR + Red + L)}(1 + L) \tag{2}$$

where *L* is a canopy background adjustment factor, the standard value is L=0.5.

Texture indices were generated from NDVI:

- Homogeneity (Eq. 3)[13] is high when grey-level co-occurrence matrix concentrates along the diagonal. This occurs when the image is locally homogeneous.

$$Homogeneity = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} P(i,j)(i-j)^2 \tag{3}$$

where *n* is the number of grey levels and *P(i, j)* defines the entries of the grey-level co-occurrence matrix

- Entropy (Eq. 4)[13] is high when pixels within a fixed window around the pixel are dissimilar.

$$Entropy = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} P(i,j)\log P(i,j) \tag{4}$$

S2 bands, vegetation and texture indices for each season were stacked and clipped with a forest mask generated in an earlier stage of the MSF project. This forest mask (i.e., forest/no-forest binary classification) was created using K-means clustering algorithm to segment the images in clusters with similar features. Clusters were subsequently assigned to a forest class depending on its overlap with the forest classes of ancillary land cover data.

A stack was build containing the bands in visible, NIR and SWIR wavelength, one band in the Red-Edge wavelength (RE2) (Table 1), and the vegetation and texture indices for both seasons.

Table 1. Sentinel-2 MSI bands

Sentinel-2 Bands	Central wavelength (nm)*	Spatial resolution (m)
Band 01 - Coastal aerosol	442.7	60
Band 02 - Blue	492.4	10
Band 03 - Green	559.8	10
Band 04 - Red	664.6	10
Band 05 - Red Edge 1	704.1	20
Band 06 - Red Edge 2	740.5	20
Band 07 - Red Edge 3	782.8	20
Band 08 - NIR	832.8	10
Band 8A - Narrow NIR	864.7	20
Band 09 - Water vapour	945.1	60
Band 10 - Cirrus	1373.5	60
Band 11 - SWIR 1	1613.7	20
Band 12 - SWIR 2	2202.4	20

* S2A Central wavelength (nm)

3.2 Feature selection

Selection of the inputs to perform forest types classification was carried out using RF feature importance analysis. RF is an ensemble classifier based on randomized decision trees[10]. At each node, the best split is selected using Gini impurity, which measures the probability of classifying a randomly selected sample into an incorrect class if the sample was labelled according the distribution of data in that node (i.e., Gini impurity measures how pure node outputs are). Feature importance is computed from the Gini impurity decrease when each feature is removed from the split. The impurity decrease of each feature is obtained by averaging decrease values across trees[10,14].

Features with an importance lower than 4% were excluded for further analysis. Of the initial 20 features (i.e., 10 bands for each season), 11 bands were finally selected (Table 2). The reduction of features allowed decreasing the execution time of the RF classifier, removing variables with less information which may add noise to the model.

Table 2: Feature importance

Band	Importance
Summer SAVI	12.47
Summer Entropy	9.84
Winter Entropy	9.11
Summer SWIR1	8.56
Winter SWIR2	6.91
Summer SWIR2	6.81
Summer RE2	6.72
Summer NDVI	6.41
Winter SWIR1	6.28
Summer NIR	5.13
Winter SAVI	4.07

3.3 Main Forest types classification

RF was trained using areas delimited in the field by forest experts in the framework of the project. Training areas belong to the dominant species listed in the study area section, namely: *P. uncinata, P. sylvestris, P. nigra, F. sylvatica, Q. pubescens* and *A. alba*. The number of RF trees was set to 150 and class weights were proportional to class sizes. A filter was applied to refine the classification, removing all the polygons with less than 0.1 ha (i.e., 10 pixels), which define the minimum mapping unit of the Main Forest Types product.

Stratified K-fold cross-validation was carried out to assess the classification accuracy. The dataset was split in 10 stratified folds, each one containing 10% of the data. In each iteration, a confusion matrix was built with one of the folds as ground truth and the rest as training data. The final metrics were computed by pooling all confusion matrices into one.

Validation metrics computed from the confusion matrix were Overall Accuracy (OA), Dice similarity Coefficient (DC), Omission Error (OE) and Commission Error (CE)[15]. These metrics were also computed for each class separately.

4. RESULTS

4.1 Feature importance

Random Forest feature importance ranking (Figure 2) showed that bands in the near-infrared (NIR and RE) and shortwave-infrared (SWIR) wavelengths were essential to discriminate forest species. These wavelengths account for 40.4% of the total initial variables and 49.1% of the final features selected for the model. In combination with Red wavelength (i.e., NDVI, SAVI and Entropy derived from NDVI), NIR and SWIR account for 82.3% of the initial variables and 100% of the final ones.

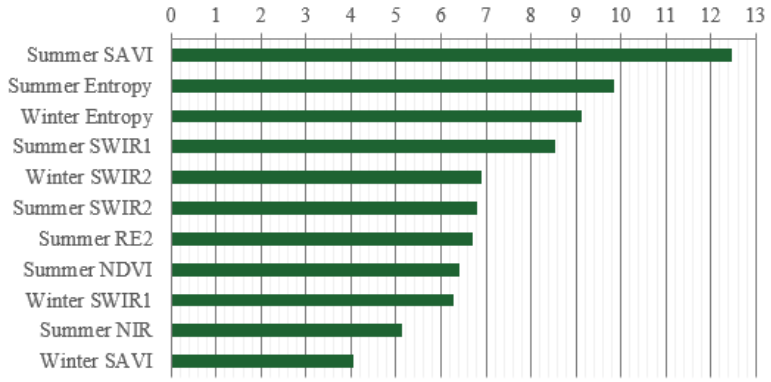

Figure 2. Importance of features selected for the final model (only features with more than 4% of importance).

4.2 Main Forest Types product

A total forested area of 423.32 km^2 in the Roncal Valley was classified. 50 % of the area was classified as genus *Pinus*, with *P. nigra* being the most frequent (30%) followed by *P. sylvestris* and *P. uncinata*. *A. alba* represented 19 % of the area, 17 % was covered by *Q. pubescens* and 13 % of *F. sylvatica* (Figure 3 and Figure 4).

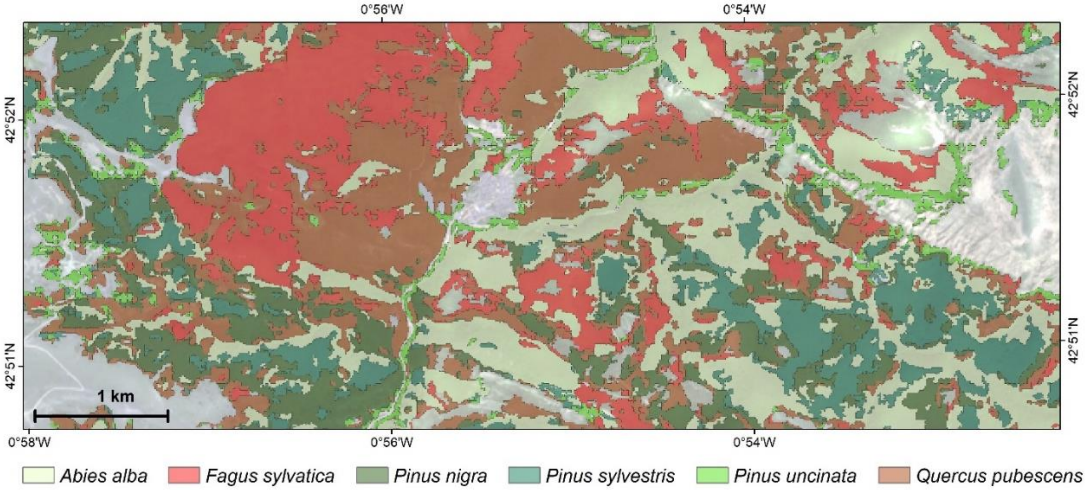

Figure 3: Detail of MSF Main Forest Types product (dominant species) in Roncal Valley.

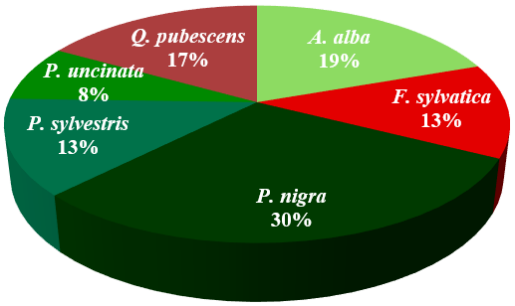

Figure 4: Dominant species per area in Roncal Valley.

4.3 Accuracy assessment

The confusion matrix (Table 3) and accuracy assessment (Table 4) show the high accuracy of the product (OA = 95 %; DC = 85.1 %; OE = 0.3% and CE = 4.9%).

Table 3: Confusion matrix for dominant species.

Ground Truth	**Classification**						
	A. alba	*F. sylvatica*	*P. nigra*	*P. sylvestris*	*P. uncinata*	*Q. pubescens*	TOTAL
A. alba	**1031**	117	123	137	4	100	1512
F. sylvatica	32	**1859**	0	0	0	17	1908
P. nigra	66	0	**1039**	188	42	62	1397
P. sylvestris	64	0	146	**1294**	52	1	1557
P. uncinata	0	0	12	9	**940**	5	966
Q. pubescens	45	69	35	0	7	**1434**	1590
TOTAL	1238	2045	1355	1628	1045	1619	**8930**

The accuracy assessment grouped by species showed some significant differences among forest types. *F sylvatica* (OA = 97.4%; DC = 94.1%) and *P. uncinata* (OA = 98.5%; DC = 93.5%) yielded the best accuracies. Although OA was higher for *P. uncinata*, DC was higher in *F. sylvatica*, since errors were more balanced (relB = -6.7%). The high accuracy of *F. sylvatyca* and *Q. pubescens* (OA = 96.2%; DC = 89.4%) is explained by seasonal changes in deciduous forests compared to conifer forests.

A. alba showed the lowest accuracy (OA = 92.3%; DC = 75%) and highest errors (CE = 6.3%; OE = 16.7%) compared to other species. These values, together with the high relative bias (relB = 22.1%) reveal that confusion between *A. alba* and other species was high.

Accuracy in species of the genus *Pinus* was variable. For all pines, OA was higher than 90%, being *P. uncinata* the best (OA = 98.5%) and *P. nigra* the worst (OA = 92.5%). Differences were higher in DC, ranging from 93.5% in *P. uncinata* to 75.5% in *P. nigra*, since it penalizes the false positives and true negatives found. The three pine species were confused due to their similar spectral response. Confusion was stronger between *P. nigra* and *P. sylvestris* (Table 3), because they share the same altitudinal zone.

Q. pubescens was the species with the lowest bias (relB = -1.8%).

Table 4: Accuracy assessment per dominant species (%).

Species	**OA**	**DC**	**CE**	**OE**	**relB**
Pinus uncinata	98.5	93.5	0.3	10.0	-7.6
Fagus sylvatica	97.4	94.1	0.7	9.1	-6.7
Quercus pubescens	96.2	89.4	2.1	11.4	-1.8
Pinus sylvestris	93.3	81.3	3.6	20.5	-4.4
Pinus nigra	92.5	75.5	4.7	23.3	3.1
Abies alba	92.3	75.0	6.3	16.7	22.1
All species	**95**	**85.1**	**0.3**	**14.9**	**-**

5. DISCUSSION AND CONCLUSIONS

In this study, the ability of Random Forest to classify main forest types in a complex mountain environment was assessed. MSF Main Forest Types algorithm classified six forest types (dominant species) with an overall accuracy of 95% and a Dice coefficient of 85%. All species were classified with an OA higher than 92% and a DC higher than 75%, although validation metrics varied significantly for different species. *Fagus sylvatica* was classified with higher accuracy than *Abies alba* or *Pinus nigra*, the latter being confused with other *Pinus* species, showing that species with higher seasonal variations were easier to identify.

Random Forest feature importance revealed that Red, NIR and SWIR wavelengths were essential to distinguish main forest types, due to the optical properties of vegetation. One key aspect was the importance given to NDVI Entropy. This may be explained by the structural properties of the forests (i.e., density and height of trees), which reflect on more homogeneous masses (e.g., *F. sylvatica*) and more heterogeneous ones (e.g., *P. uncinata*).

Classification errors were mainly caused by confusion between species of the same genus (*Pinus*) and confusion in mixed forests. The relatively low accuracy of *A. alba* might be due to the bioclimatic characteristics of the Roncal Valley. As *A. alba* is at the climatic limit of its distribution, it grows only in slopes oriented to the North, competing with *F. sylvatica*. Hence, data used to train the model might contain pixels of these mixed forests, leading to confusion between species.

One of the shortcomings of this study is its high dependence on inputs: forest mask and training areas. The methods proposed could be improved with: i) better adjustment of the training datasets; ii) improvement of forest mask; iii) considering factors affecting the distribution of dominant species, such as aspect or altitudinal zones; iv) adjusting input S2 data to each dominant species phenology. Hence, deeper research is needed to improve the generalization of this type of methods for different local conditions and forest types.

ACKWOLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 776045.

The authors are grateful to the Forest Owners Association of Navarra (Foresna-Zurgaia) for providing data about the study area.

REFERENCES

[1] Trumbore, S., Brando, P. and Hartmann, H., "Forest health and global change" (2016).

[2] Schröter, D., Cramer, W., Leemans, R., Prentice, I. C., Araújo, M. B., Arnell, N. W., Bondeau, A., Bugmann, H., Carter, T. R., Gracia, C. A., De La Vega-Leinert, A. C., Erhard, M., Ewert, F., Glendining, M., House, J. I., Kankaanpää, S., Klein, R. J. T., Lavorel, S., Lindner, M., et al., "Ecology: Ecosystem service supply and vulnerability to global change in Europe," Science (80-.). **310**(5752), 1333–1337 (2005).

[3] van Lierop, P., Lindquist, E., Sathyapala, S. and Franceschini, G., "Global forest area disturbance from fire, insect pests, diseases and severe weather events," For. Ecol. Manage. **352**, 78–88 (2015).

[4] Senf, C., Pflugmacher, D., Zhiqiang, Y., Sebald, J., Knorn, J., Neumann, M., Hostert, P. and Seidl, R., "Canopy mortality has doubled in Europe's temperate forests over the last three decades," Nat. Commun. **9**(1), 4978 (2018).

[5] MySustainableForest, "My Sustainable Forest. Earth Observation services for silviculture," 2018, <https://mysustainableforest.com/> (29 May 2019).

[6] Puletti, N., Chianucci, F. and Castaldi, C., "Use of Sentinel-2 for forest classification in Mediterranean environments," Ann. Silvic. Res. (2017).

[7] Cheng, K. and Wang, J., "Forest Type Classification Based on Integrated Spectral-Spatial-Temporal Features and Random Forest Algorithm—A Case Study in the Qinling Mountains," For. **10**(7) (2019).

[8] Liu, Y., Gong, W., Hu, X. and Gong, J., "Forest Type Identification with Random Forest Using Sentinel-1A, Sentinel-2A, Multi-Temporal Landsat-8 and DEM Data," Remote Sens. **10**(6) (2018).

[9] Hościło, A. and Lewandowska, A., "Mapping Forest Type and Tree Species on a Regional Scale Using Multi-Temporal Sentinel-2 Data," Remote Sens. **11**(8) (2019).

[10] Breiman, L., “Random Forests,” Mach. Learn. **45**(1), 5–32 (2001).
[11] Gobierno de Navarra, “Plan Forestal de Navarra” (1999).
[12] Huete, A. R., “A soil-adjusted vegetation index (SAVI),” Remote Sens. Environ. **25**(3), 295–309 (1988).
[13] Haralick, R. M., “Statistical and structural approaches to texture,” Proc. IEEE **67**(5), 786–804 (1979).
[14] Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W. and Hamprecht, F. A., “A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data,” BMC Bioinformatics **10**(1), 213 (2009).
[15] Fernandez-Carrillo, A., Belenguer-Plomer, M. A., Chuvieco, E. and Tanase, M. A., “Effects of sample size on burned areas accuracy estimates in the Amazon Basin,” Proc. SPIE - Int. Soc. Opt. Eng. **10790** (2018).