scientific data

OPEN

DATA DESCRIPTOR

Generating high-resolution land use and land cover maps for the greater Mariño watershed in 2019 with machine learning

Améline Vallet[1,2] ✉, Stéphane Dupuy[3], Matthieu Verlynde[1,2] & Raffaele Gaetano[3]

Land Use and Land Cover (LULC) maps are important tools for environmental planning and social-ecological modeling, as they provide critical information for evaluating risks, managing natural resources, and facilitating effective decision-making. This study aimed to generate a very high spatial resolution (0.5 m) and detailed (21 classes) LULC map for the greater Mariño watershed (Peru) in 2019, using the MORINGA processing chain. This new method for LULC mapping consisted in a supervised object-based LULC classification, using the random forest algorithm along with multi-sensor satellite imagery from which spectral and textural predictors were derived (a very high spatial resolution Pléiades image and a time serie of high spatial resolution Sentinel-2 images). The random forest classifier showed a very good performance and the LULC map was further improved through additional post-treatment steps that included cross-checking with external GIS data sources and manual correction using photointerpretation, resulting in a more accurate and reliable map. The final LULC provides new information for environmental management and monitoring in the greater Mariño watershed. With this study we contribute to the efforts to develop standardized and replicable methodologies for high-resolution and high-accuracy LULC mapping, which is crucial for informed decision-making and conservation strategies.

Background & Summary

Land Use and Land Cover (LULC) play a key role in environmental planning, management and monitoring. Accurate LULC information is key for evaluating potential risks to ecosystems and biodiversity, ensuring food security, mitigating natural hazards, and facilitating effective urban planning. LULC maps are often used as an indicator or a proxy of natural and economic processes in environmental modeling. For instance, they are used as inputs in models aiming to map population distribution[1,2], poverty or income[3,4], ecosystem services (carbon storage, water yield, etc.)[5,6], ecological accounting[7].

Over the last decades, remote sensing and satellite products have revolutionized the detection and mapping of LULC, as they provide a spatially extensive, multi-temporal and time saving source of information about LULC[8]. Earlier LULC mapping studies have intensively used medium and low-resolution earth observation satellites, such as LANDSAT (MSS and TM), ASTER, MODIS, SPOT, but with important limitations. First, they often lead to confusion between land-cover types because of a limited number of spectral bands to distinguish them. Second, they poorly captured changes in vegetation overtime, because of low return frequencies. And finally, they showed a limited ability to capture fine details and small-scale features on the Earth's surface because of their rough spatial resolution[9–11]. New satellites, such as Pléiades, Landsat 9, Sentinel-2, with high return frequencies of multitemporal products, large multispectral sensors and very high-resolution imagery address the above-mentioned limitations and offer new opportunities to LULC mapping[12,13].

The methods used for classifying LULC from remote sensing products have also considerably evolved in the recent years, with machine learning algorithm driving the latest developments in LULC mapping. Techniques

[1]Université Paris-Saclay, CNRS, AgroParisTech, Ecologie Systématique et Evolution, 91190, Gif-sur-Yvette, France. [2]Université Paris-Saclay, AgroParisTech, CNRS, Ecole des Ponts ParisTech, Cirad, EHESS, UMR CIRED, 94130, Nogent-sur-Marne, France. [3]TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, 34398, Montpellier, France. ✉e-mail: ameline.vallet@agroparistech.fr

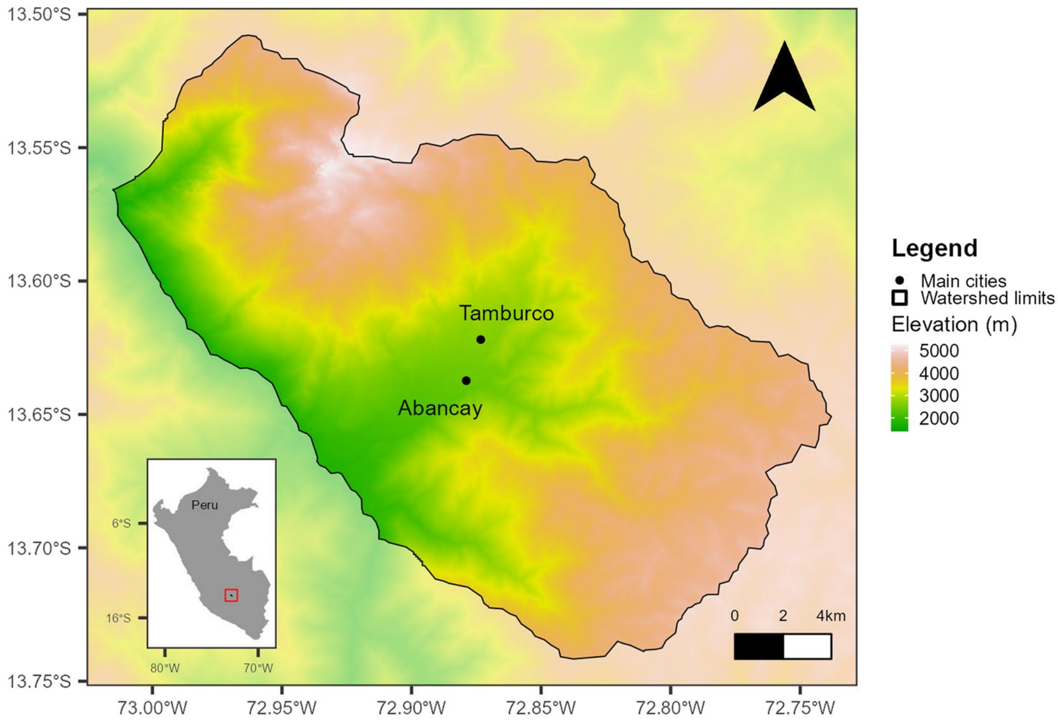

Fig. 1 Map of the greater Mariño watershed.

such as Random Forest, Support Vector Machine, and Artificial Neural Networks were found to significantly improve the accuracy and efficiency of traditional approaches, that historically relied on manual interpretation of satellite imagery or simple spectral analysis[8,14,15]. Machine learning algorithms are very flexible regarding input data, which enables them to process multisource remote products - including LiDAR, radar, and hyperspectral imagery - of varying resolution and spectra. In addition, they allow a full automation of the classification process and enable efficient analysis of large volumes of data.

Recently published high resolution global LULC datasets are making use of new remote sensing products and advanced machine learning classification algorithms. For instance, WorldCover, launched in 2020 by the European Space Agency, is an open-access global land cover map at 10 m resolution, including 11 classes, based on both Sentinel-1 and Sentinel-2 images[16]. Other similar initiatives include GlobeLand30[17], ESRI 2020 global LULC map[18] or Google Earth Engine Dynamic World NRT[19]. While these global datasets have the advantage of providing new information about countries with limited data until now (e.g. South America, Africa), they often contain limited number of LULC classes, and show varying levels of accuracy, strongly depending on ecological biomes[19,20]. Indeed, the main challenges to LULC mapping consist in the detection of specific ecosystems, such as wetlands or mangroves and the detection of small-scale features, such as agro-forest mosaics, urban areas, dispersed settlements. Integrating multiple sources of remote sensing products, at different time periods to capture changes in vegetation, with precise *in-situ* data is often mentioned as the way to improve their detection[9,21,22].

The aim of this study is to apply a new method, the MORINGA processing chain, to generate a high resolution and detailed (21 LULC classes) LULC map for the greater Mariño watershed (Peru) in 2019, using the most recent remote sensing imagery (Sentinel-2 and Pléiades) and a random forest algorithm. The greater Mariño watershed is an important area for biodiversity conservation and water management in the Andes, and accurate LULC mapping is crucial for informed decision-making about natural resources. Identifying changes in LULC over time, will allow for more effective management and conservation efforts, and will facilitate better management and conservation strategies. With this study we also contribute to the efforts to develop standardized and replicable methodologies for high resolution, and high accuracy LULC mapping.

Material and Methods

Study site. The greater Mariño watershed stretches over 403 km^2 along the eastern slopes of the Southern Peruvian Andes, in the region of Apurimac, Peru (Fig. 1). The local climate is dry and hot in the interandean valleys and cold and humid on the highlands. Annual precipitations are also highly variable, with a dry season (June to August) characterized by lower rainfalls in contrast with the wet season (December to march)[23]. The elevation varies from 1614 to 5180 m, with very diverse landscapes and ecosystems: dry forests, glaciers, wetlands (*bofedales*) and more than a dozen of high-elevation lakes. Approximately 70000 people live in the watershed, mostly in two urban areas, Abancay and Tamburco. Agriculture at high and mid elevations is subsistence oriented, whereas at low elevations both crop and livestock farming are commercially oriented and more intensive. There are also tourism activities in the Ampay Forest Sanctuary, which protects 36 km^2 of land. Like other mountain social-ecological systems, the greater Mariño watershed provides important but vulnerable ecosystem

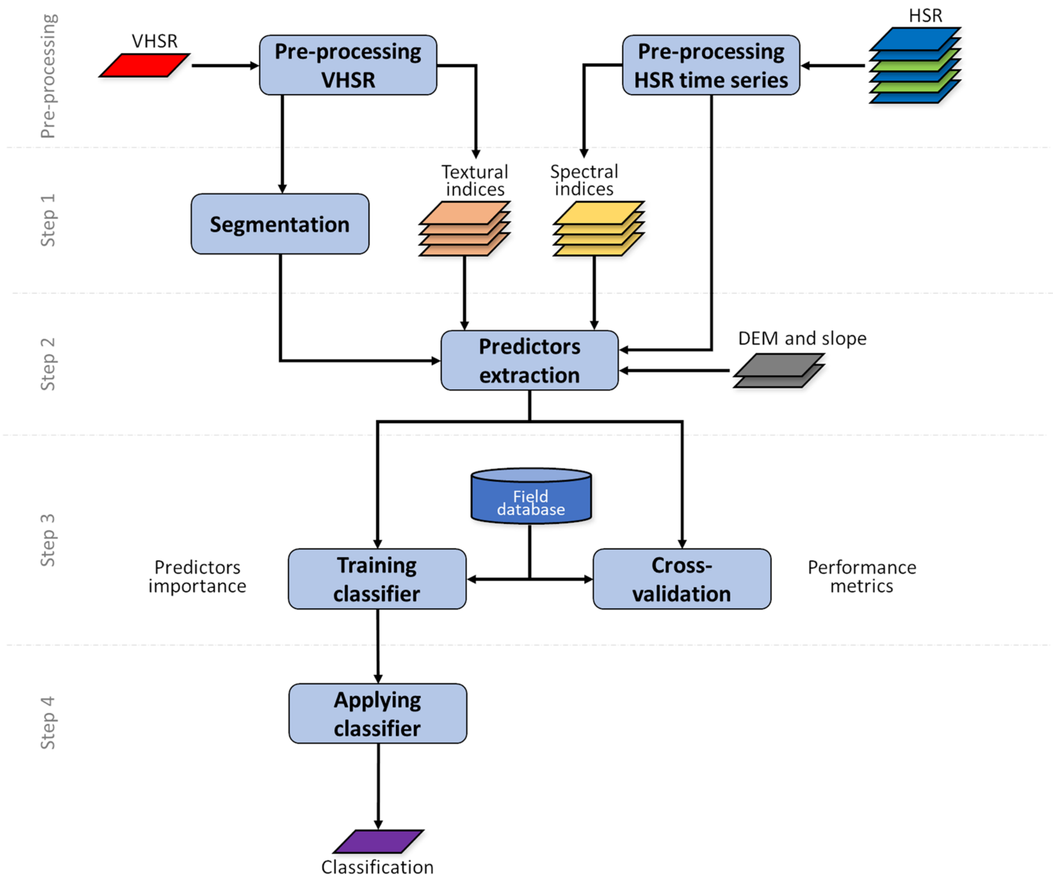

Fig. 2 Overview of the MORINGA processing chain.

services that contribute substantially to people good quality of life in the area. Some landscape planning instrument oriented toward biodiversity and ecosystem conservation have been implemented in the past, or are under implementation. These include, for example, the creation of a protected area (the Ampay National Sanctuary), a payment for hydrological services, and several nature-based solution programs, such as reforestation schemes or wetland restoration projects[24].

Overview of the MORINGA processing chain. The LULC classification was produced thanks to the MORINGA processing chain, a supervised object-based LULC classification methodology/technique using multi-sensor satellite imagery[25]. It has been applied recently to several tropical agrosystems of the world, including La Réunion island[12], Madagascar[26,27], Senegal[28], Haiti[29,30]. The MORINGA chain is composed of four steps (1) segmentation of a Very High Spatial Resolution (VHSR) satellite image (such as Spot 6/7 or Pléiades); (2) object level extraction of spectral and textural predictors derived from several High Spatial Resolution (HSR) satellite images (such as Sentinel-2, or Landsat 8) at different dates, along with the VHSR satellite image and other remote sensing products (such as DEM); (3) training and validation of a random forest classifier using a field database (possibly at different levels of a LULC nomenclature); (4) application of the classifier to the whole study area to map LULC (Fig. 2). The pre-processing of satellite images, so that they can be used at steps 1 and 2, is also part of the MORINGA processing chain.

The MORINGA processing chain is compiled within a Python 3.8 environment and relies mainly on the GDAL/OGR library and the Orfeo ToolBox (OTB) version 7.2 (https://www.orfeo-toolbox.org). It is complemented with custom modules for specific steps (e.g., for computing reasons the calculation of object-based statistics at step 2 makes use an ad-hoc C++ module, "obiatools", whose source code is available at https://gitlab.irstea.fr/raffaele.gaetano/obiatools). Some pre-processing steps are also performed out of the Python under QGIS (e.g. slope calculation). The source code of the Moringa processing chain is available at https://gitlab.irstea.fr/raffaele.gaetano/moringa. The implementation of these different steps in the greater Mariño watershed, as well as the satellite images used are described more in detail in the following sections.

Field database and land-use land-cover nomenclature. Fieldwork was carried out in May and June 2016, which corresponds to the beginning of winter and the dry season (i.e. the end of the peak of the growing season), and in agricultural areas, to the beginning of harvest. Sampling sites were selected through a mix of systematic sampling (points distributed in all the study area to capture the altitudinal gradient effect) and stratified sampling (to ensure that sufficient observations are collected for each LULC class). Some sampling sites were located outside the greater Mariño watershed - while maintaining a close proximity - in order to sample specific

Level 1	Level 2	Level 3	Code	Description
Agricultural areas	Agricultural areas	Sugar cane	1	Agricultural areas dedicated to the cultivation of sugar cane, a tall perennial grass used primarily for sugar and spirit (*aguardiente*) production.
		Pasture, fallow and feed	2	Agricultural areas used for grazing livestock (pasture), left to rest and regenerate naturally (fallow), and fields planted with grass specifically for animal (feed).
		Crop and alfalfa	3	Agricultural areas used for growing various crops (vegetables, corn, Andean tubers, cereals), including alfalfa, a perennial forage crop commonly used for livestock feed due to its high protein content.
		Fruit crop	4	Orchards or plantations dedicated to growing fruit-trees such as avocado, tangerine and lemon.
Natural spaces and forest plantations	Woodlands	Polylepis mountain forest	5	High-elevation forests dominated by *Polylepis* spp. trees (*queña* in Quechua), known for their resilience to harsh mountain climates.
		Podocarpus glomeratus mountain forest	6	Forests of mountainous regions dominated by *Podocarpus glomeratus* (*intimpa* in Quechua), a coniferous tree species adapted to cool and high-elevation environments.
		Dry forest	7	Found in deep inter-Andean valleys on steep, rocky slopes. Composed of an ephemeral herbaceous layer, shrubs, succulents (like *Browningia hertlingiana*, *Opundia* sp. or *Corryocactus* sp), and deciduous trees adapted to dry climate (such as *Schinus molle or Eriotheca* sp. called *pati locally*).
		Other tree vegetation	8	Mixed woodlands made of natural or planted trees, with no dominant species, and including for instance *Escallonia* spp. (known locally as chachacomo), or *Vallea stipularis* (called *Chuyllur*) in mixed forests found at mid and high elevations, *Juglans neotropica* and *Fuchsia boliviana* along rivers in the valley.
		Pine plantation	9	Monoculture forests planted with pine trees, usually for timber production and carbon sequestration.
		Eucalyptus plantation	10	Forest planted with eucalyptus trees (mainly in monoculture), which grows rapidly and are adaptable to a range of climatic conditions. Used for timber, fuelwood and carbon sequestration.
	Shrubs and natural grasslands	Mixed shrubland	11	Shrubs found at medium to high-elevation areas, adapted to cold and often windy condition (such as *Lupinus* spp., *Baccharis* spp., *Berberis* spp., *Rubus* spp.)
		Dry shrubland and semi-arid steppe	12	Lowland areas covered with drought-resistant shrubs species (such as A*gave americana, Jatropa spp. or Dodonaea viscosa* called locally *chamana*), typically found in arid and semi-arid inter-Andean valleys.
		High-elevation grassland	13	Grasslands found at high elevation and composed of dense tussocks and hardy vegetation (*pajonal*) or low grasses (*césped*). It includes various species of *Festuca* spp., *Calamagrostis* spp., as well as *Stipa ichu*.
	Rocks and natural bare soils	Rock and natural bare soil	14	Exposed rocky surfaces and areas with little to no vegetation, including bare soil regions naturally with no significant plant cover.
		Beach and riverine rock	15	Riverside areas dominated by sandy beaches, pebbles, and rocks, influenced by water dynamics.
Water and glacier	Glacier	Glacier	16	Masses of ice found at high elevation, formed from compacted snow over many years.
	Water	Wetland	17	High-elevation water-saturated soils, extremely rich in organic matter, formed in flat areas around ponds or streams (know as *bofedales* locally). Their specialized vegetation (including for instance *Distichia muscoides*, *Lachemilla pinnata* or *Werneria* spp.) is adapted to humid and cold environments, and is vital for biodiversity, water regulation, and support traditional livestock grazing.
		Lake	18	Inland water bodies found in high-elevation areas, providing critical habitats for wild aquatic species and aquaculture (trout farming).
		River network	19	Rivers and streams constituting the greater Mariño watershed.
Impervious areas	Impervious areas	Built-up area	20	Residential buildings in both rural and urban settings, as well as commercial, industrial, and institutional buildings. These areas are characterized by significant human and economic activities.
		Road	21	Paved roads, urban streets, and gravel roads, which are essential for transportation and connectivity.

Table 1. Description of the land-use and land-cover classes.

LULC classes which are scarce in the study area (e.g. Pine plantations, *Polylepis* sp. forests). At each sampling site, we recorded GPS coordinates, took pictures in the direction of the four cardinal points, and registered the vegetation and species observed. Each sampled site was then digitalized into a polygon delimiting a plot with homogeneous LULC inside, which was classified into one of the categories presented in Table 1 (level 3). This nomenclature is aligned with other LULC maps provided at national[31,32] or regional[33,34] scale. The VHSR image was used for the delineation of the polygon, based on photointerpretation. 1698 polygons composed the final field database, covering a total of 16.75 km^2 (Table 2, Table SM 2).

Satellites images and their pre-processing. *Topography*. TanDEM-X Digital Elevation Model (DEM) was obtained thanks to the European Space Agency (ESA), through its scientific research support program. TanDEM-X is part of ESA Third Party Missions Programme, that comprises 50 satellites dedicated to earth observation (https://earth.esa.int/eogateway/missions/terrasar-x-and-tandem-x). TanDEM-X is almost identical to its twin, TerraSAR-X, with which they fly on close formation to produce high accuracy and resolution elevation models (12 m spatial resolution), thanks to a powerful radar system: Synthetic Aperture Radar (SAR)[35]. Pixels with no data were filled with mean elevation of the study area, using OTB *BandMathX* application.

Very high spatial resolution (VHSR). We used three Pléiades images (of different sizes) acquired on the 7th of octobre 2019 (i.e. at the end of the dry season) simultaneously for both the panchromatic and the multispectral mode, at a spatial resolution of 0.5 and 2 m respectively (Table SM 1). These images are distributed commercially by AIRBUS Defense and Space at primary geometric processing level and a basic radiometric processing (12-bit native). The access to the Pléiades images was funded and facilitated by DINAMIS, a French institutional data hub that provides an access to high and very high resolution optical and radar data (https://dinamis.data-terra.org). DINAMIS is part of the Data Terra national research infrastructure, whose main mission is to develop an integrated platform for Earth system data, services and products (https://www.data-terra.org).

Level 1	Level 2	Level 3	Code	F1-score Moringa (cross-validation mean and sd are presented between brackets)	F1-score post-treatment	Reference database		
						Number of polygons (% of total number of polygons in the dataset)	Total surface in km² (% of total surface in dataset)	Average size of polygons (m²)
Agricultural areas	Agricultural areas	Sugar cane	1	96.4% [80.8 ± 12.4]	96.4%	23 (1.4%)	0.03 (0.2%)	1 582
		Pasture, fallow and feed	2	99% [89.9 ± 3.6]	99.1%	245 (14.4%)	0.69 (4.1%)	2 839
		Crop and alfalfa	3	98% [89.1 ± 2.9]	98.1%	353 (20.8%)	0.67 (4.0%)	1 906
		Fruit crop	4	91.6% [76.3 ± 4.1]	89.5%	52 (3.1%)	0.09 (0.5%)	1 737
Natural spaces and forest plantations	Woodlands	Polylepis mountain forest	5	99.4% [84.9 ± 3.4]	99.4%	23 (1.4%)	0.16 (0.9%)	7 010
		Podocarpus glomeratus mountain forest	6	99.3% [92.3 ± 6.7]	99.3%	26 (1.5%)	0.86 (5.1%)	33 186
		Dry forest	7	99.1% [88.7 ± 6.4]	99.2%	56 (3.3%)	1.26 (7.5%)	22 651
		Other tree vegetation	8	97.9% [85.3 ± 5.9]	98.1%	320 (18.8%)	1.93 (11.5%)	6 035
		Pine plantation	9	97.4% [74.6 ± 18.0]	97.7%	40 (2.4%)	0.17 (1.0%)	4 311
		Eucalyptus plantation	10	98.9% [79.7 ± 10.9]	99.4%	82 (4.8%)	0.99 (5.9%)	12 174
	Shrubs and natural grasslands	Mixed shrubland	11	98.2% [86.7 ± 8.0]	97.6%	83 (4.9%)	0.76 (4.5%)	9 202
		Dry shrubland and semi-arid steppe	12	99.3% [87.2 ± 6.7]	98.7%	46 (2.7%)	1.13 (6.7%)	24 574
		High-elevation grassland	13	99.8% [98.0 ± 1.2]	99.7%	53 (3.1%)	4.51 (26.8%)	85 018
	Rocks and natural bare soils	Rock and natural bare soil	14	99.4% [92.1 ± 7.0]	99.2%	44 (2.6%)	1.09 (6.5%)	24 890
		Beach and riverine rock	15	99.0% [83.0 ± 20.1]	86.0%	14 (0.8%)	0.03 (0.1%)	2 183
Water and glacier	Glacier	Glacier	16	100.0% [99.9 ± 0.1]	100.0%	6 (0.4%)	0.26 (1.5%)	44 271
	Water	Wetland	17	99.7% [95.5 ± 3.9]	99.5%	58 (3.4%)	0.68 (4.0%)	11 730
		Lake	18	99.8% [98.7 ± 1.1]	99.9%	35 (2.1%)	0.57 (3.4%)	16 559
		River network	19	99.3% [98.3 ± 1.4]	99.5%	19 (1.1%)	0.02 (0.1%)	1 176
Impervious areas	Impervious areas	Built-up area	20	99.7% [99.0 ± 1.1]	99.8%	120 (7.1%)	0.78 (4.7%)	6 571
		Road	21	—	—	—	—	—
Total						1,698 (100%)	16.75 (100%)	9 866
Performance metrics			OOB	96.086%	—			
			Overall accuracy	98.6% [89.0 ± 2.9]	97.8%			
			Cohen's K index	0.990 [0.910 ± 0.019]	0.989			
			Pontius' Q index	0.0048 [0.0513 ± 0.0191]	0.0042			
			Pontius' A index	0.0039 [0.0273 ± 0.0122]	0.0053			

Table 2. Three level LULC nomenclature. For each LULC class, the characteristics of the field database and the performance of the MORINGA chain before/after post-treatment is presented. Performance is assessed with the Out-Of-Bag error, the cross-validation, and by comparing the whole field database to the LULC classification before/after post-treatment.

Pre-processing consisted in (1) the calculation of Top Of Atmosphere (TOA) reflectance, by correcting distributed images for sensor calibration and radiation incidence, and (2) the orthorectification of images using TanDEM-X DEM (with OTB *OrthoRectification* application). The three pre-processed tiles of Pléiades panchromatic and multispectral images were then mosaicked, and finally, the two resulting mosaics were pansharpened using the Bayesian fusion algorithm (OTB *Pansharpening* application), to obtain a multispectral image at 0.5 m spatial resolution. Pléiades multispectral image at 2 m resolution was then no longer used in the processing chain (only the pan-sharpened image at 0.5 m resolution is used).

High spatial resolution (HSR). We also used a time series of 333 Sentinel-2 images, acquired between the 1st of January 2018 and the 30th of October 2019 to capture the vegetation dynamics all along the year before the acquisition date of the Pleiades image (Table SM 1). Sentinel-2 images are provided by two satellites (Sentinel-2 A and B), deployed by the European Space Agency (ESA) and the Copernicus program. The time span between the acquisition by either satellite is five days. The images were downloaded free of charge through the PEPS platform (https://peps.cnes.fr) at level 1 C (i.e. orthorectified TOA reflectance). The Sen2Cor (https://step.esa.int/main/snap-supported-plugins/sen2cor/) atmospheric correction processor for Sentinel-2 images

Indices		Description	Source images	Reference
Textural indices				
Haralick energy for four radius sizes (1, 5, 11 et 21)		Texture uniformity	VHSR	[37]
Haralick variance for four radius sizes (1, 5, 11 et 21)		Texture heterogeneity		
Haralick inertia for four radius sizes (1, 5, 11 et 21)		Intensity contrast between a pixel and its neighborhood		
Haralick correlation for four radius sizes (1, 5, 11 et 21)		Correlation of a pixel with its neighborhood		
Spectral indices				
Vegetation	Normalized Difference Vegetation Index - NDVI	Standardized index displaying relative biomass as a proxy of vegetation greenness. NDVI = (NIR-Red)/(NIR + Red)	VHSR and HSR	[69]
	Ratio vegetation index - RVI	Index quantifying vegetation greenness, more sensitive to stressed or sparse vegetation than NDVI. RVI = NIR/Red	VHSR	[70]
	Normalized Difference Red Edge Index - NDRE	A variant of NDVI, that uses Red Edge band instead of visible red. It better detects vegetative stress and is less sensitive to saturation in the presence of dense vegetation. NDRE = (NIR-RedEdge)/(NIR + RedEdge)	HSR	[71]
Water	Normalized Difference Water Index - NDWI	Measures the presence and amount of water in vegetation canopies or water bodies. NDWI = (Green - NIR)/(Green + NIR)	VHSR and HSR	[72]
	Short Wave Normalized Difference Vegetation Index - SWNDVI	Another measures of the presence and amount of water in vegetation canopies or water bodies. SWNDVI = (NIR-MIR)/(NIR + MIR)	HSR	[73]
	Modified Normalized Difference Water Index - MNDWI	Modified version of the NDWI that improves water detection and reduce the influence of urban features and bare soil MNDWI = (Green - MIR)/(Green + MIR)	HSR	[74]
Soil	Brightness index - BI2	The relative brightness of pixels, with an enhanced contrast between bright and dark pixels. $BI2 = \sqrt{(Red^2 + Green^2 + NIR^2)/3}$	VHSR	[75]
	Brightness index - BI	The Euclidean norm of the surface reflectances (except aerosols bands, which are not pertinent for vegetation mapping). $BI = \sqrt{\sum_i Band_i^2}$	HSR	[76]

Table 3. Textural and spectral indices computed from VHSR and HSR images. Green (Pléiades B2, Sentinel-2 B3); Red (Pléiades B3, Sentinel-2 B4); RedEdge (Sentinel-2 B5); NIR (Pléiades B4, Sentinel-2 B8); MIR (Sentinel-2 B11).

allowed to obtain a level 2 A Bottom-Of-Atmosphere (BOA) reflectance product from distributed level 1 C images, as well cloud, cloud shadows and snow masks[36].

Two Sentinel-2 tiles (T18LYL and T18LYK) were necessary to cover the whole study area: they were mosaicked to generate a time series of Sentinel-2 mosaics at different dates. Although already orthorectified, Sentinel-2 images were also readjusted to the VHSR Pleiades image using OTB *HomologousPointsExtraction* application with red band (Pléiades band 1, Sentinel-2 band 3) as a reference (step 2 of MORINGA processing chain). To eliminate clouds, we created synthetic images every 20 days (gapfilling processing, *ImageTimeSeriesGapFilling* OTB application). The final Sentinel-2 time series is thus composed of 22 synthetic images, from the 25th of July 2018 and 5th of October 2019.

LULC classification with the MORINGA processing chain. *Predictors calculation: topographic, textural and spectral indices.* Several indices were calculated from the VHSR and HSR images (Table 3), to be later used as predictors in the classification model. Following previous studies, four textural indices developed by Haralick[37] were calculated using the panchromatic Pléiades image[27,38,39]. Textures are important for detecting landscape patterns, such as tree or crop rows, easily detectable in the VHSR image. Textural indices were computed thanks to *HaralickTextureExtraction* OTB application. Four sizes of sliding window were used for each index, with radius values of 1 (i.e. a sliding window of 3 × 3 pixels), 5 (11 × 11 pixels), 11 (23 × 23 pixels) et 21 (43 × 43 pixels) (Table 3).

Nine spectral indices were also calculated from Pléiades pansharpened image and from the Sentinel-2 time series of synthetic images (Table 3), using OTB *RadiometricIndices* application. Sentinel-2 sensor delivers 13 spectral bands, ranging from 10 to 60 m resolution, but only the 10 bands with a resolution of 20 m or less were exploited in this study (i.e. three 60 m resolution bands were discarded), as direct predictors in the classification model, but also to compute 6 spectral indices that are commonly used to characterize and classify LULC (Table 3).

Finally, slope was calculated from TanDEM-X DEM with QGIS and used as a predictor in addition to elevation. To classify LULC, we therefore used a total of 352 Sentinel-2 derived predictors (= 22 dates * 10 bands + 22 dates * 6 spectral indices), 20 Pléiades derived predictors (= 4 spectral indices + 4 textural indices * 4 radius) and 2 TanDEM-X derived predictors (elevation and slope).

Object detection by segmentation of the VHSR image. For the segmentation of Pléiades pansharpened mosaic, we used a method proposed by Baatz and Schäpe[40] and implemented in OTB *LargeScaleGenericRegionMerging* remote application, available at https://gitlab.irstea.fr/remi.cresson/LSGRM[41]. Various tuning tests were performed on different sub-regions of the study area before selecting the following values (tested values are indicated between brackets):

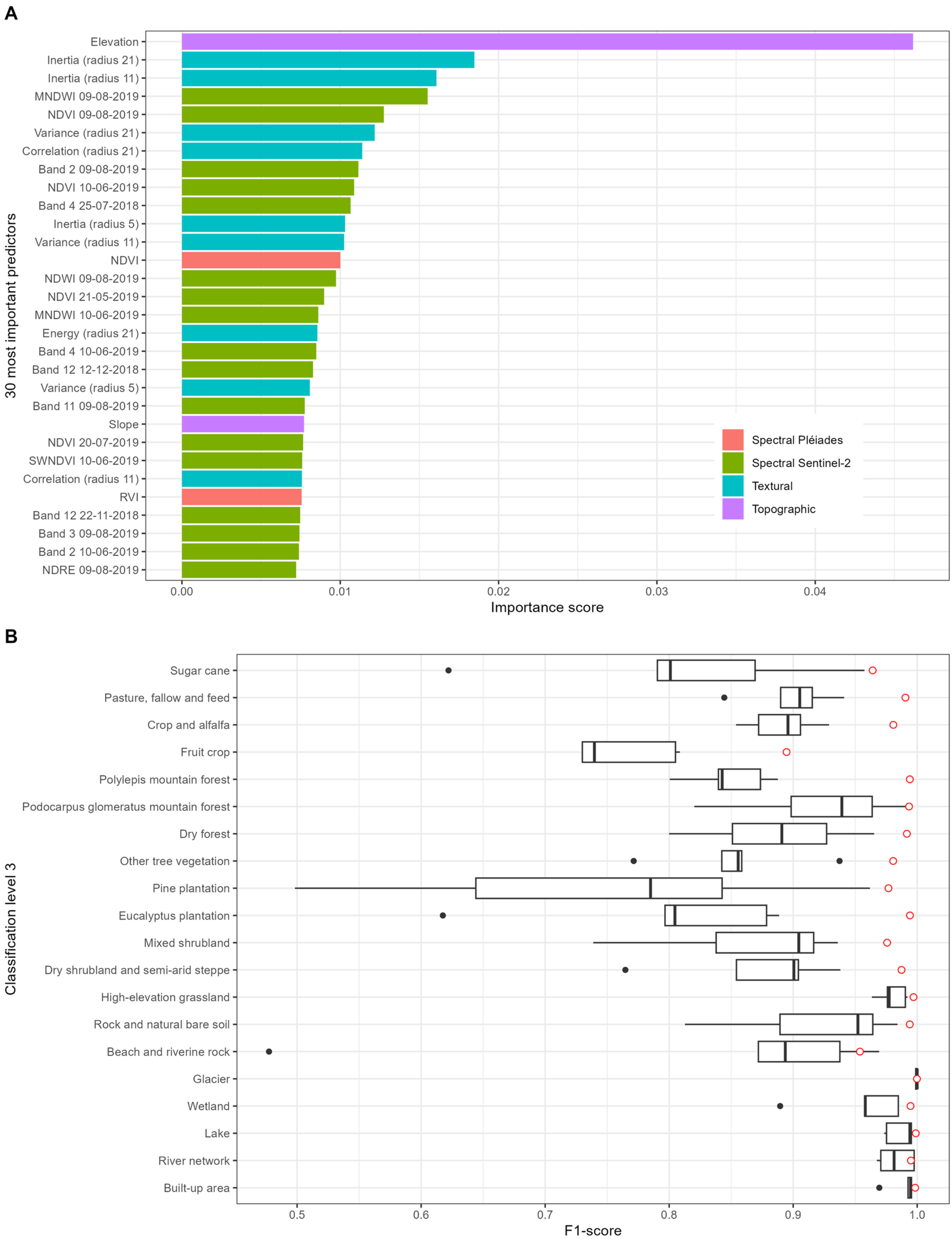

Fig. 3 Interpretation and validation of the random forest classifier. (**A**) Importance scores of the selected predictors for the level 3 classification. (**B**) Boxplots of the F1-scores obtained during the cross-validation (level 3 of the LULC classification). For comparison, the F1-scores obtained by comparing the MORINGA classification to the full field database are represented in red.

- scale parameter: 350 [70–450]
- weight parameter on the shape: 0,3 [0.1–0.8]
- weight parameter on compactness: 0,7 [0.5–0.7]

This segmentation step partitions the image into homogenous objects and extracts their contours. The geometries delimited in the image were exported as a shapefile, and for each constitutive element (i.e. for each object of the segmentation), we extracted the mean values of each of the 374 textural, spectral and topographic predictors presented in the previous section. The segmentation was then intersected with the polygons of the field database for which LULC was recorded/identified, and for each element of the intersection the mean values of predictors were also extracted, which composed the training dataset of the classification algorithm (35 392 training elements). Extractions were made thanks to the C++ "obiatools" module.

Random forest training. The random forest algorithm was used to classify LULC from the training dataset produced at the previous step. This algorithm is based on an ensemble of classification or regression decision trees, each created using random subsets of predictors and training data, whose predictions are combined by majority voting or averaging[42,43]. Over the last two decades, the use of random forest for remote sensing applications has received an increasing attention due to its capacity to handle large datasets (of observations and predictors) and missing data, its processing speed, and high accuracy[8,14]. Applications focused for instance on mapping LULC[27,44], vegetation biomass[45,46], urban areas[47,48] and habitat quality and health[49–51].

One random forest model was trained at the level 3 of the LULC nomenclature, using OTB *TrainVectorClassifier* application, and the following tuning options:

- Maximum depth of the tree: 25
- Minimum number of samples in each node: 10 (OTB default value)
- Cluster possible values of a categorical variable into K < = cat clusters to find a suboptimal split: 10 (OTB default value)
- Size of the randomly selected subset of features at each tree node: square root of the total number of predictors (OTB default value, in this application: $\sqrt{374} = 19.34$)
- Maximum number of trees in the forest: 800
- Sufficient accuracy (OOB error): 0.01 (OTB default value)

All observations in the training dataset whose size was greater than $25 m^2$ were used for training the classifier, which was then applied to each element of the segmentation for which we extracted predictors values, in order to generate a level 3 LULC classification.

Predictors importance (also called variable or feature importance) was calculated in order to highlight which predictors contributed more to the classification, and were the most influential. Predictor importance is commonly used as a tool for interpretating machine learning algorithms and explaining how particular predictions are made[52]. Predictors importance were calculated using Python module *scikit-learn*, and a random forest model-specific importance score based on mean accumulation of impurity decrease (https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html).

Elevation showed the highest importance, then followed by two textural indices (Haralick contrast with radius of 21 and 11), and two vegetation and water spectral indices derived from Sentinel-2 HSR images in August 2019 (Fig. 3A). Slope also appeared as an important predictor, which suggests that considering topography is crucial for LULC classification in areas of high relief such as the Andes. Half of the 16 textural indexes were among the most important predictors, which also indicates that Pléaides-derived textures drove the LULC classification and explained a large amount of our training dataset variance. Finally, several Sentinel-2 spectral indices and bands at different dates were among the most important predictors, which underlines the importance of considering time series of multispectral images for characterizing vegetation dynamics during the classification process.

Post-processing procedure. The post-processing of the LULC classification produced by the MORINGA chain consisted in four steps: (1) conversion to raster; (2) smoothing by majority filter; (3) cross-checking with GIS data and (4) manual correction by photo-interpretation. All post-processing operations were conducted at the finest level of the classification (level 3), and then scaled-up thanks to the nested structure of the nomenclature.

First, the vectorial classification obtained with the MORINGA chain was converted to a raster format, at the resolution of Pléiades' pansharpened and panchromatic images (0.5 m). The resolution of the Pleiades image was preferred over that of the Sentinel images (10 and 20 m), as the Pleiades image is the one used for the construction of the field database (polygon delineation based on Pléiades image photointerpretation), and segmentation, which are two crucial steps for the supervised classification. As the object were identified at a 0.5 m resolution, it is essential to convert the MORINGA classification into a raster at the same resolution to ensure their integrity. Indeed, the 0.5 m resolution allowed to preserve the isolated landscape features identified during segmentation (such as rural buildings, or roads): they would be merged with neighboring LULC classes with a rasterization at lower resolution.

Second, a majority filter resampling was used to remove isolated pixels and smooth out the classification contours, with OTB *ClassificationMapRegularization* application and a radius of 3 (corresponding to a 7×7 pixels sliding windows). This smoothing only removed objects whose size was inferior to $1.75 m^2$ (in comparison the size of a residential house in rural areas is approximately $10 m^2$), and therefore did not alter the identification of the isolated landscape features mentioned above.

Then, we cross-checked the LULC classification with external data sources to detect unexpected behavior of the MORINGA classifier. For each LULC class of the nomenclature at level 3, specific GIS references, all accessible in open-access, were identified (Table SM 3) and intersected with the classification to highlight potential errors. All disagreements between the classification and the reference GIS data were systematically inspected and eventually corrected manually by photo-interpretation of the Pléiades image, using the *Thematic Raster Editor*

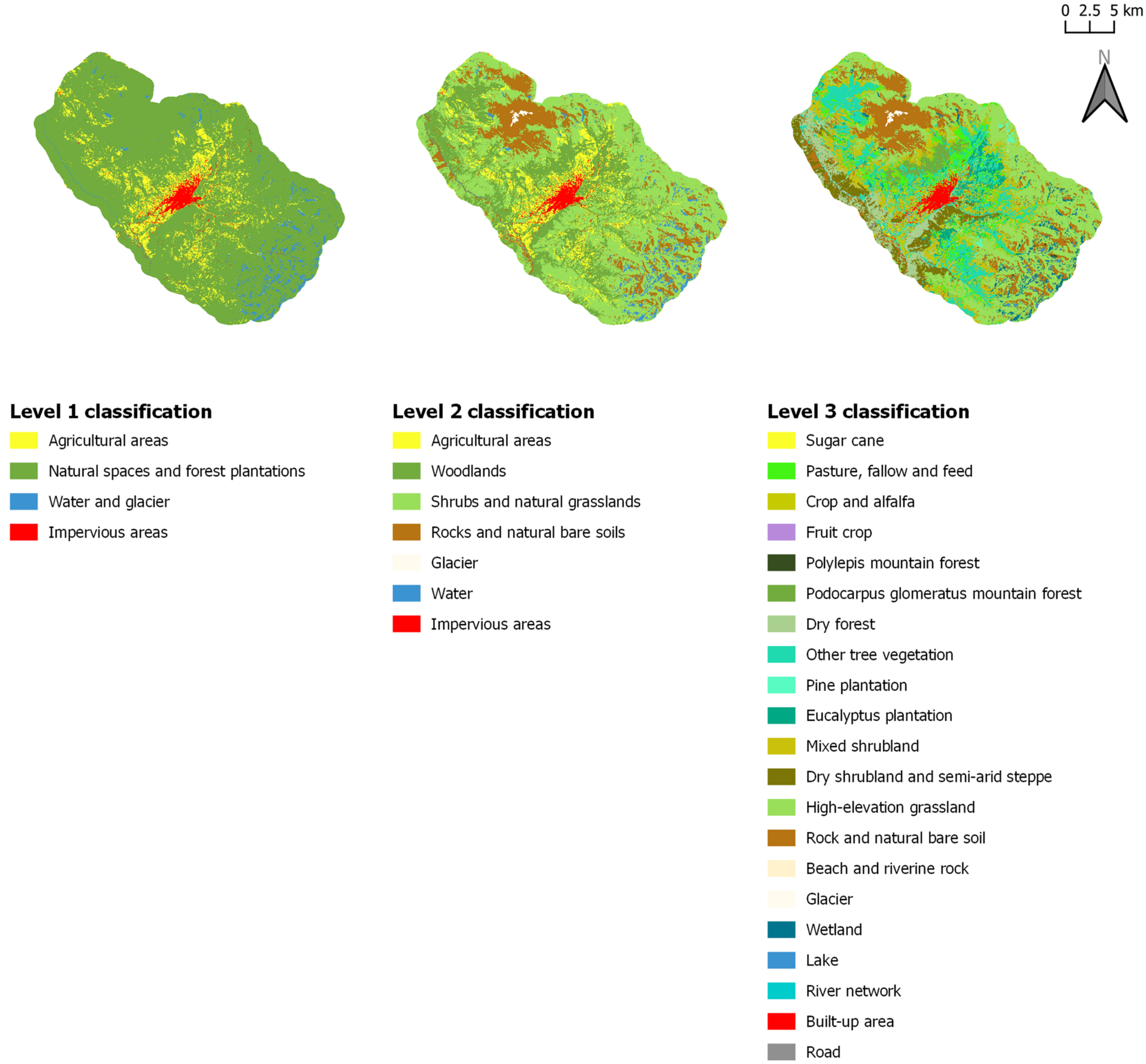

Fig. 4 Three-level classification of land-use land-cover in the greater Mariño watershed (Peru) in 2019.

(ThRasE) a QGIS Python plugin that allows flexible and fast raster editing (https://plugins.qgis.org/plugins/ThRasE/). For instance, crops and pastures classes were compared to the map of agricultural areas (https://siea.midagri.gob.pe/portal/informativos/superficie-agricola-peruana) developed by the Peruvian ministry in charge of Agriculture[53], and water bodies to the Global Surface Water Explorer[54]. Other data sources were provided by the European Commission, the Peruvian Ministry of Agrarian Development and irrigation, the Ministry of the Environment of Peru and the OpenStreetMap community.

Finally, the classification was carefully screened using the tile-by-tile navigation option of ThRasE (tile size of approximately 4 km^2), and the Pléiades image as a reference (with true and false color composites to highlight vegetation areas). All the classification errors detected were manually corrected. The road network LULC category was added at this stage, by combining elements of the classification from different LULC classes (built-up areas mainly, but also other land use classes at lower percentage). OpenStreetMap data was used to confirm the location of photo-interpreted roads[55].

Vectorization. The post-treated classification raster was converted to a vector database using the *Raster To Polygon* conversion tool from ArcGIS Pro, with the polygon simplification option activated to smooth contours. The *Repair Geometry* tool was then applied to inspect polygons for geometry problems and repair them, with the "Delete Features with Null Geometry" option set off.

Data Records

The final LULC classification (Fig. 4) and its description is available at the Recherche Data Gouv repository under the CC BY 4.0 license, in both raster and vector format (https://doi.org/10.57745/DDP1ZR)[56]. The raster format is only provided for the level 3 of the LUCL nomenclature at 0.5 m resolution, but the three nomenclature

Post-treatment / Moringa output	Sugar cane	Pasture, fallow and feed	Crop and alfalfa	Fruit crop	Polylepis mountain forest	Podocarpus glomeratus mountain forest	Dry forest	Other tree vegetation	Pine plantation	Eucalyptus plantation	Mixed shrubland	Dry shrubland and semi-arid steppe	High-elevation grassland	Rock and natural bare soil	Beach and riverine rock	Glacier	Wetland	Lake	River network	Built-up area	Road
Sugar cane	—																				
Pasture, fallow and feed		—				1%		1%					1%				1%				
Crop and alfalfa			—									15%		2%							
Fruit crop				—				1%													
Polylepis mountain forest					—																
Podocarpus glomeratus mountain forest						—		3%													
Dry forest							—	1%			1%										
Other tree vegetation								—		8%											1%
Pine plantation								1%	—												
Eucalyptus plantation								4%		—	2%										
Mixed shrubland											—	1%									
Dry shrubland and semi-arid steppe											10%	—									
High-elevation grassland													—	3%							
Rock and natural bare soil														—							
Beach and riverine rock														5%	—						
Glacier																—					
Wetland								1%	1%		7%		7%	1%			—				
Lake														5%				—			
River network														1%					—		
Built-up area												1%		3%						—	1%
No data																					

Table 4. Corrections applied to the MORINGA classification (rows) during post-treatment (columns). Cells on the diagonal correspond to the 91.5% of the study area to which no changes were applied during post-processing (they are indicated with "-"). The percentages represent the share of each LULC class in the corrections applied during post-processing (i.e. they were calculated considering only pixels whose LULC class was manually changed). Only values > 1% are indicated for readability.

levels are provided in separate layers of the geopackage file (Table SM 5). The field database used to train the random forest is accessible at the same repository and under the same license; this dataset contains LULC observations at the three nomenclature levels, in a geopackage file (Table SM 6). All three datasets are delivered in the local UTM projection (WGS 84 UTM 18 S, EPSG code 32718).

Technical Validation

Random forest cross-validation and performance metrics. In the random forest algorithm, the subset of training data left out from each tree (also called Out-Of-Bag -OOB- observations) can be used for assessing the prediction error rate, yielding the so-called OOB error, a measure of the classifier performance. Random forests can therefore be trained and validated using all available observations. However, as some noted, this approach can lead to a biased estimation of performance, because of overfit and because it does not consider the size of training observations[57,58]. In this study we therefore decided to implement, in addition to OOB error, a second approach for estimating the random forest classifier performance, based on cross-validation.

Cross-validation is a procedure to estimate classification performance, where the training dataset is split into K separate folds. For each fold k, a random forest model is trained on the K-1 other folds (i.e. excluding k fold data), and then applied to the k fold data, to assess its performance, taking into account the size of training observations. It is worth noting that the K models developed during the cross-validation procedure are slightly different from the overall model fitted using all observations from the training dataset, as they are trained with only a subset of the data: the objective is not to generate final predictions (i.e. final LULC classification), but to evaluate the quality of the classification model[58]. In this study we implemented a 5-fold cross-validation, and we estimated the quality of the classification in each fold using the following performance metrics (that were calculated on training observations weighted by their surface, Fig. 3B). The same metrics were also calculated before and after the post-processing, considering all observations available in the training database (Table 1).

- F1 score, a harmonic mean of the precision and recall, ranging from 0 to 1, computed for each LULC class separately[59].
- An overall accuracy score, computed as the average of each LULC accuracy score (corresponding to the total surface of correctly classified objects divided by the total surface of training observations)[59].
- Cohen's kappa, which reflects level of agreement between the proposed classification, and a random one[60].
- Pontius' quantity disagreement (Q, which measures the differences in the proportion of area or quantity of each LULC class), allocation disagreement (A, which measures the measures the differences in the spatial arrangement or allocation of each LULC class) and total disagreement (D, calculated as the average of Q and A). Pontius metrics have been proposed to address some of the limitations of Cohen's kappa, by explicitly considering the spatial allocation of LULC classes, distinguishing between false positives and false negatives, and not assuming that the disagreement is due to chance[61].

We used Python module sklearn.metrics (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.htm) to calculate F1 score, accuracy and Cohen's kappa during cross-validation. Pontius' metrics (A, Q, and D), were manually calculated for each fold validation observation, by generating a corresponding confusion matrix for, using OTB *ComputeConfusionMatrix* application. The same application was used to calculate all performance matrix before and after post-treatment, considering all available observations from the training database.

Corrections applied to the classification during post-processing. Corrections were applied to 8.5% of the study area (a map locating the exact changes is provided in Figure SM 1). The most frequent error was crops confounded with dry shrublands and semi-arid steppes of the valley (15% of the area corrected during post-processing) (Table 4). Mixed shrublands, found at higher elevation and grasslands were often misclassified into wetlands (14% of corrections), and eucalyptus plantations confounded with other types of woodlands (8% of the corrected surface). Frequent confusions were also observed between types of shrublands (11% of total corrections).

Some LULC classes, that did not cover large portions of the study area, showed higher levels of post-treatment corrections (Table SM 4). For instance, 77% of the areas classified as beach and riverine rocks by the MORINGA were confounded with rocks and natural bare soils. And 57% of the area classified as lakes were indeed rocks and natural bare soils. The confusion between surface water and bare soils can be explained by relief and shadow effects, as observed in other publications[62–64]. The presence of clouds on the Pléiades image affected the quality of the segmentation in small areas of the study site: the contours of the objects affected by clouds were corrected manually during this post-processing stage. The confusion between riverine rocks and bare soils is due to the close resemblance of their multispectral signal and suggest that other topographic parameters could be added to the MORINGA predictors, such as distance to river network, to improve the distinction between these LULC classes.

Final classification validation. The overall accuracy (i.e. the arithmetic mean of F1-scores from each LULC class) and Cohen's K index showed a very high agreement between the post-processing map and the training database. The final level of disagreement quantity obtained after post-processing (Pontius Q), was of 0.0042, while the allocation disagreement (Pontius A) was of 0.0053 (Table 1). This means that most of the disagreement (approx. 60%) is explained by the precise location of the different LULC classes in the maps (Pontius A), and not each LULC class relative importance (Pontius Q). Pontius total disagreement (D) disagreement) was very low, which confirm the strong agreement between the post-processing map and the training database.

The slight decrease of overall accuracy and Cohen's K index observed after post-processing can be explained by changes in F1-score in two LULC classes (Table 1): "Beach and riverine rock" and "Fruit crop". Fruit crops are among the most complicated classes of LULC to detect, along with wetlands, small-scale fields, and urban areas, that machine learning algorithms typically tend to misidentify[65–68]. For "Beach and riverine rock", the change in accuracy can be explained by an error in the training database, where a polygon of 5451m2 was wrongly classified as "Beach and riverine rock" instead of "Rock and natural bare soil", among the 14 polygons identified as "Beach and riverine rock" areas in the training database (Table SM 2).

Code availability

The source code of the Moringa processing chain is available at https://gitlab.irstea.fr/raffaele.gaetano/moringa.git. It is complemented with custom modules for specific steps (e.g., for computing reasons the calculation of object-based statistics at step 2 makes use an ad-hoc C++ module, "obiatools") whose source code is available at https://gitlab.irstea.fr/raffaele.gaetano/obiatools.

Received: 2 February 2024; Accepted: 5 August 2024;
Published online: 23 August 2024

References

1. Sorichetta, A. *et al*. High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Sci. Data* **2**, 150045 (2015).
2. Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLOS ONE* **10**, e0107042 (2015).
3. Bosco, C. *et al*. Exploring the high-resolution mapping of gender-disaggregated development indicators. *J. R. Soc. Interface* **14**, 20160825 (2017).
4. Steele, J. E. *et al*. Mapping poverty using mobile phone and satellite data. *J. R. Soc. Interface* **14**, 20160690 (2017).
5. Cabral, P., Feger, C., Levrel, H., Chambolle, M. & Basque, D. Assessing the impact of land-cover changes on ecosystem services: A first step toward integrative planning in Bordeaux, France. *Ecosyst. Serv.* **22**, 318–327 (2016).
6. Vallet, A. *et al*. Dynamics of Ecosystem Services during Forest Transitions in Reventazón, Costa Rica. *PLOS ONE* **11**, e0158615 (2016).
7. Chen, Y., Vardon, M., Keith, H., Van Dijk, A. & Doran, B. Linking ecosystem accounting to environmental planning and management: Opportunities and barriers using a case study from the Australian Capital Territory. *Environ. Sci. Policy* **142**, 206–219 (2023).
8. Talukdar, S. *et al*. Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. *Remote Sens.* **12**, 1135 (2020).
9. Ban, Y., Gong, P. & Giri, C. Global land cover mapping using Earth observation satellite data: Recent progresses and challenges. *ISPRS J. Photogramm. Remote Sens.* **103**, 1–6 (2015).
10. Guo, H., Fu, W. & Liu, G. *Scientific Satellite and Moon-Based Earth Observation for Global Change*. (Springer Singapore, 2019).
11. Kramer, H. J. *Observation of the Earth and Its Environment: Survey of Missions and Sensors*. (Springer Science & Business Media, 2002).
12. Dupuy, S., Gaetano, R. & Le Mézo, L. Mapping land cover on Reunion Island in 2017 using satellite imagery and geospatial ground data. *Data Brief* **28**, 104934 (2020).
13. Wang, D. *et al*. Evaluating the Performance of Sentinel-2, Landsat 8 and Pléiades-1 in Mapping Mangrove Extent and Species. *Remote Sens.* **10**, 1468 (2018).
14. Belgiu, M. & Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **114**, 24–31 (2016).
15. Wang, J., Bretz, M., Dewan, M. A. A. & Delavar, M. A. Machine learning in modelling land-use and land cover-change (LULCC): Current status, challenges and prospects. *Sci. Total Environ.* **822**, 153559 (2022).
16. Zanaga, D. *et al*. ESA WorldCover 10 m 2020 v100. *Zenodo* https://doi.org/10.5281/zenodo.5571936 (2021).
17. Chen, J., Cao, X., Peng, S. & Ren, H. Analysis and Applications of GlobeLand30: A Review. *ISPRS Int. J. Geo-Inf.* **6**, 230 (2017).
18. Karra, K. *et al*. Global land use/land cover with Sentinel 2 and deep learning. in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS* 4704–4707 https://doi.org/10.1109/IGARSS47720.2021.9553499 (2021).
19. Brown, C. F. *et al*. Dynamic World, Near real-time global 10 m land use land cover mapping. *Sci. Data* **9**, 251 (2022).
20. Venter, Z. S., Barton, D. N., Chakraborty, T., Simensen, T. & Singh, G. Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover. *Remote Sens.* **14**, 4101 (2022).
21. Szantoi, Z. *et al*. Addressing the need for improved land cover map products for policy support. *Environ. Sci. Policy* **112**, 28–35 (2020).
22. Zhang, C. & Li, X. Land Use and Land Cover Mapping in the Era of Big Data. *Land* **11**, 1692 (2022).
23. SENAMHI. *Caracterización Climática de Las Regiones Apurímac y Cusco*. (2012).
24. SUNASS. *Documento de orientación para la implementación de los Merese Hídricos*. https://www.sunass.gob.pe/sunass-te-informa/publicaciones/documento-orientacion-implementacion-merese-hidricos/ (2021).
25. Gaetano, R. *et al*. The MORINGA processing chain: Automatic object-based land cover classification of tropical agrosystems using multi-sensor satellite imagery. https://agritrop.cirad.fr/594650/ (2019).
26. Dupuy, S., Defrise, L., Gaetano, R., Andriamanga, V. & Rasoamalala, E. Land cover maps of Antananarivo (capital of Madagascar) produced by processing multisource satellite imagery and geospatial reference data. *Data Brief* **31**, 105952 (2020).
27. Dupuy, S. *et al*. Analyzing Urban Agriculture's Contribution to a Southern City's Resilience through Land Cover Mapping: The Case of Antananarivo, Capital of Madagascar. *Remote Sens.* **12**, 1962 (2020).
28. Jolivot, A. Cartographie de l'occupation du sol de la zone des Niayes (Sénégal) en 2018 (1.5 m de résolution). *CIRAD Dataverse* https://doi.org/10.18167/DVN1/KJAS6S (2021).
29. Dupuy, S., Lelong, C. & Gaetano, R. *Rapport méthodologique: Cartographie de l'occupation du sol sur le site des NIPPES à Haïti*. https://agritrop.cirad.fr/597938/ (2021).
30. Gaetano, R., Dupuy, S. & Lelong, C. Nippes - Haïti - 2020, Land Cover Map at high spatial resolution. CIRAD Dataverse https://doi.org/10.18167/DVN1/ZAN2WN (2021).
31. MINAM. *Mapa nacional de ecosistemas del Perú - Memoria descriptiva*. https://geoservidor.minam.gob.pe/wp-content/uploads/2017/06/MEMORIA-DESCRIPTIVA-MAPA-DE-ECOSISTEMAS.pdf (2019).
32. MINAM. *Mapa nacional de cobertura vegetal: Memoria descriptiva*. http://www.minam.gob.pe/patrimonio-natural/wp-content/uploads/sites/6/2013/10/MAPA-NACIONAL-DE-COBERTURA-VEGETAL-FINAL.compressed.pdf (2015).
33. Cuadros Loayza, J. A., Peña Caytuiro, R. & Valenzuela Trujillo, J. J. *Memoria Descriptiva de La Cobertura y Uso de La Tierra Del Proceso de Meso Zonificación Ecológica Económica de La Región Apurímac*. 201 http://sigrid.cenepred.gob.pe/docs/PARA%20PUBLICAR/OTROS/Estudio_de_cobertura_y_uso_de_la_tierra_del_proceso_de_meso_ZEE_de_la_region_Apurimac.pdf (2016).

34. UE-Prodesarrollo Apurímac. *Caracterización Ecológica Económica de La Microcuenca Mariño*. (2010).
35. Wessel, B. *TanDEM-X Ground Segment – DEM Products Specification Document*. https://tandemx-science.dlr.de/ (2018).
36. Main-Knorn, M. *et al*. Sen2Cor for Sentinel-2. in *Image and Signal Processing for Remote Sensing XXIII* (eds. Bruzzone, L., Bovolo, F. & Benediktsson, J. A.) 3 https://doi.org/10.1117/12.2278218 (SPIE, Warsaw, Poland, 2017).
37. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3**, 610–621 (1973).
38. Beguet, B., Chehata, N., Boukir, S. & Guyon, D. Classification of forest structure using very high resolution Pleiades image texture. in *2014 IEEE Geoscience and Remote Sensing Symposium* 2324–2327 https://doi.org/10.1109/IGARSS.2014.6946936 (2014).
39. Rajendran, G. B., Kumarasamy, U. M., Zarro, C., Divakarachari, P. B. & Ullo, S. L. Land-Use and Land-Cover Classification Using a Human Group-Based Particle Swarm Optimization Algorithm with an LSTM Classifier on Hybrid Pre-Processing Remote-Sensing Images. *Remote Sens.* **12**, 4135 (2020).
40. Baatz, M. & Schäpe, A. Multiresolution Segmentation: an optimization approach for high quality multi-scale image segmentation. in (2000).
41. Lassalle, P., Inglada, J., Michel, J., Grizonnet, M. & Malik, J. Large scale region-merging segmentation using the local mutual best fitting concept. in *2014 IEEE Geoscience and Remote Sensing Symposium* 4887–4890 https://doi.org/10.1109/IGARSS.2014.6947590 (2014).
42. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. https://doi.org/10.1007/978-1-4614-6849-3 (Springer New York, New York, NY, 2013).
43. Lantz, B. *Machine Learning with R: Learn How to Use R to Apply Powerful Machine Learning Methods and Gain an Insight into Real-World Applications*. (Packt Publ, Birmingham, 2013).
44. Gislason, P. O., Benediktsson, J. A. & Sveinsson, J. R. Random Forests for land cover classification. *Pattern Recognit. Lett.* **27**, 294–300 (2006).
45. Baccini, A., Laporte, N., Goetz, S. J., Sun, M. & Dong, H. A first map of tropical Africa's above-ground biomass derived from satellite imagery. *Environ. Res. Lett.* **3**, 045011 (2008).
46. Karlson, M. *et al*. Mapping Tree Canopy Cover and Aboveground Biomass in Sudano-Sahelian Woodlands Using Landsat 8 and Random Forest. *Remote Sens.* **7**, 10017–10041 (2015).
47. Deng, C. & Wu, C. The use of single-date MODIS imagery for estimating large-scale urban impervious surface fraction with spectral mixture analysis and machine learning techniques. *ISPRS J. Photogramm. Remote Sens.* **86**, 100–110 (2013).
48. Xia, N., Cheng, L. & Li, M. Mapping Urban Areas Using a Combination of Remote Sensing and Geolocation Data. *Remote Sens.* **11**, 1470 (2019).
49. Fraser, B. T. & Congalton, R. G. Monitoring Fine-Scale Forest Health Using Unmanned Aerial Systems (UAS) Multispectral Models. *Remote Sens.* **13**, 4873 (2021).
50. Ozigis, M. S., Kaduk, J. D. & Jarvis, C. H. Mapping terrestrial oil spill impact using machine learning random forest and Landsat 8 OLI imagery: a case site within the Niger Delta region of Nigeria. *Environ. Sci. Pollut. Res.* **26**, 3621–3635 (2019).
51. Wang, H., Zhao, Y., Pu, R. & Zhang, Z. Mapping Robinia Pseudoacacia Forest Health Conditions by Using Combined Spectral, Spatial, and Textural Information Extracted from IKONOS Imagery and Random Forest Classifier. *Remote Sens.* **7**, 9020–9044 (2015).
52. Molnar, C. *Interpretable Machine Learning*. (2020).
53. Livia Alejandro, L. *et al*. *Atlas de la superficie agrícola del Perú*. https://repositorio.ana.gob.pe/handle/20.500.12543/4895 (2021).
54. Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A. S. High-resolution mapping of global surface water and its long-term changes. *Nature* **540**, 418–422 (2016).
55. OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org (2022).
56. Vallet, A. High resolution land use and land cover map for the greater Mariño watershed in 2019. *Recherche Data Gouv* https://doi.org/10.57745/DDP1ZR (2024).
57. Janitza, S. & Hornung, R. On the overestimation of random forest's out-of-bag error. *PLOS ONE* **13**, e0201904 (2018).
58. Makowski, D., Brun, F., Doutart, E., Duyme, F. & Jabri, M. E. *Data science pour l'agriculture et l'environnement: Méthodes et applications avec R et Python*. (Ellipses, 2021).
59. Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **17**, 168–192 (2020).
60. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).
61. Pontius, R. G. & Millones, M. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int. J. Remote Sens.* **32**, 4407–4429 (2011).
62. Ji, L., Gong, P., Geng, X. & Zhao, Y. Improving the Accuracy of the Water Surface Cover Type in the 30 m FROM-GLC Product. *Remote Sens.* **7**, 13507–13527 (2015).
63. Myeong, S., Nowak, D. J., Hopkins, P. F. & Brock, R. H. Urban cover mapping using digital, high-spatial resolution aerial imagery. *Urban Ecosyst.* **5**, 243–256 (2001).
64. Van de Voorde, T., De Genst, W. & Canters, F. Improving Pixel-based VHR Land-cover Classifications of Urban Areas with Post-classification Techniques. *Photogramm. Eng. Remote Sens.* **73** (2007).
65. Ozesmi, S. L. & Bauer, M. E. Satellite remote sensing of wetlands. *Wetl. Ecol. Manag.* **10**, 381–402 (2002).
66. Rapinel, S. *et al*. National wetland mapping using remote-sensing-derived environmental variables, archive field data, and artificial intelligence. *Heliyon* **9**, e13482 (2023).
67. Yang, X. *et al*. Detection and characterization of coastal tidal wetland change in the northeastern US using Landsat time series. *Remote Sens. Environ.* **276**, 113047 (2022).
68. Zhou, X.-X. *et al*. Research on remote sensing classification of fruit trees based on Sentinel-2 multi-temporal imageries. *Sci. Rep.* **12**, 11549 (2022).
69. Rouse, J., Rh, H., Ja, S. & Dw, D. Monitoring vegetation systems in the great plains with ERTS. (1974).
70. Pearson, R. L. & Miller, L. D. Remote Mapping of Standing Crop Biomass for Estimation of the Productivity of the Shortgrass Prairie. 1355 (1972).
71. Barnes, E. M. *et al*. Coincident detection of crop water stress, nitrogen status and canopy density using ground-based multispectral data. in *Proceedings of the 5th International Conference on Precision Agriculture and other resource management July 16-19, 2000, Bloomington, MN USA* (2000).
72. McFeeters, S. K. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *Int. J. Remote Sens.* **17**, 1425–1432 (1996).
73. Gao, B. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **58**, 257–266 (1996).
74. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **27**, 3025–3033 (2006).
75. Escadafal, R. Remote sensing of arid soil surface color with Landsat thematic mapper. *Adv. Space Res.* **9**, 159–163 (1989).
76. Inglada, J. *et al*. Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery. *Remote Sens.* **7**, 12356–12379 (2015).

Acknowledgements
The authors are grateful to DINAMIS for providing them with access to Pléiades images. They also thank the European Space Agency for granting them free access to TanDEM-X images. This study was funded by CLAND and MSH Paris Saclay (grant 20-EM-06). Without the support and resources provided by these organizations, this research would not have been possible. The authors are also thankful to Yésica Quispe Conde and Jaime J. Valenzuela Trujillo for their help with the organization of fieldwork and their useful feedback on the LULC nomenclature and classification.

Author contributions
A.V. supervised the study and collected the field data. A.V. and S.D. designed the methodology and processed the data. M.V. processed the data post-processing, validation and visualization. A.V. prepared the manuscript, with contributions from all co-authors.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03750-x.

Correspondence and requests for materials should be addressed to A.V.