Revista
GEOGRÁFICA
VENEZOLANA

Volumen
66(2) 2025
julio-diciembre

ISSNe 2244-8853 ISSNp 1012-1617
Recibido: septiembre, 2024/ Aceptado: enero, 2025
pp. 290-308 / https://doi.org/10.53766/RGV/

Mapping a rice region in South America using

Geo Big Data and Sentinel 2

Mapeo de uma región arrocera en Sudamérica utilizando Geo Big Data y Sentinel 2

Mapeamento de uma região de arroz na América do Sul usando Geo Big Data e Sentinel 2

Giancarlo Alciaturi[1], María del Pilar García-Rodríguez[2] y Virginia Fernández[3]

[1] Universidad Complutense de Madrid, Programa de Doctorado en Geografía
[2] Universidad Complutense de Madrid, Departamento de Geografía
Madrid, España
[3] Universidad de La República, Departamento de Geografía
Montevideo, Uruguay
galciatu@ucm.es; mpgarcia@ucm.es; vivi@fcien.edu.uy

Alciaturi: https://orcid.org/0000-0003-1687-9593
García: https://orcid.org/0000-0002-7237-2335
Fernández: https://orcid.org/0000-0003-2891-1896

Abstract

Geo Big Data and Sentinel-2 have been extensively employed to map rice paddies and general land use/land cover. Focusing on the environmental value and relevance of rice production in Cuenca de la Laguna Merín in Uruguay, the research aimed to 1) map rice paddies and other land use/land cover classes, 2) compare the capabilities of Random Forest and Support Vector Machine for classifying two different Sentinel-2 time series stacks, and 3) identify the most important features according to Random Forest. In addition to quoted imagery and classifiers, the materials include Google Earth Engine, GEEMAP, and Python's Scikit-learn GridSearchCV. The main methods comprised hyperparameter tuning, supervised classification, and accuracy assessment. Quoted assessment revealed that the four maps performed well. The feature importance analysis highlighted the Near-Infrared and Shortwave Infrared as the most relevant features. Future research should focus on integrating diverse data sources and comparing different time series than those employed here.
KEYWORDS: Uruguay; laguna Merín; Land use/Land Cover; rice paddies.

Resumen

Geo Big Data y Sentinel-2 son eficientes para cartografiar arrozales y otras categorías de uso y cobertura del suelo. Dada la relevancia ambiental de la cuenca de la Laguna Merín y su rol en la producción arrocera del Uruguay, con este trabajo se pretendió: 1) mapear los arrozales y clases generales de uso y cobertura del suelo; 2) comparar el desempeño de *Random Forest* y *Support Vector Machine* para clasificar dos juegos temporales Sentinel-2, y 3) identificar las bandas más importantes según *Random Forest*. Los materiales incluyen las imágenes y clasificadores mencionados, *Google Earth Engine*, GEEMAP, y *GridSearchCV* de *Python*. Como métodos, destacan el ajuste de hiperparámetros, la clasificación supervisada, y el cálculo de métricas de precisión. Estas últimas sugieren que los cuatro mapas aportan resultados óptimos. Las bandas infrarrojas cercano y de onda corta son las más relevantes para clasificar. Futuras iniciativas deben enfocarse en integrar imágenes de sensores diversos y utilizar series temporales distintas a las aquí empleadas.
PALABRAS CLAVE: Uruguay; laguna Merín; uso/cobertura del suelo; arrozales.

Resumo

Geo Big Data e Sentinel-2 são amplamente reconhecidos para mapear arrozais e outras categorias de uso e cobertura do solo. Dada a relevância ambiental da bacia da Lagoa Mirim e seu papel na produção de arroz no Uruguai, este trabalho teve como objetivos: 1) mapear os arrozais e classes gerais de uso e cobertura do solo, 2) comparar o desempenho dos algoritmos Random Forest e Support Vector Machine na classificação de duas séries temporais de Sentinel-2, e 3) identificar as bandas mais importantes segundo Random Forest. Os materiais incluem as imagens e classificadores mencionados, Google Earth Engine, GEEMAP e GridSearchCV do Python. Os métodos utilizados incluem ajuste de hiperparâmetros, classificação supervisionada e cálculo de métricas de precisão. Estas últimas sugerem que os quatro mapas fornecem resultados ótimos. As bandas de infravermelho próximo e infravermelho de onda curta são as mais relevantes para a classificação. Futuras iniciativas devem se concentrar em integrar imagens de sensores diversos e utilizar séries temporais diferentes das aqui empregadas.
PALAVRAS-CHAVE: Uruguay; lagoa Merín; uso/cobertura do solo; arrozais.

1. Introduction

Rice production is vital for Uruguay's domestic economy and global food security. It is considered the most widely cultivated crop, grown in over one hundred countries, and consumed by at least half of the world's population (Food and Agriculture Organization [FAO], 2004). Uruguay is one of the most export-oriented country globally, with approximately 95% of its total production sold in foreign markets (Pittelkow *et al*., 2016). The local production model, known as 'irrigated lowland', facilitates the extensive use of agricultural machinery and allows for more land-extensive production compared to traditional methods used in various ecosystems such as irrigated upland, rainfed lowland, rainfed upland, and deepwater/floating ecosystems (Bray, 1986).

Uruguay's cultivation is limited to a single season, typically from late October for establishment to March–April for maturity. The Cuenca de la Laguna Merín (CLM) has historically been a leading region in Uruguay, boasting the largest surface area, production, and workforce, as reported by the Ministerio de Ganadería, Agricultura, y Pesca (MGAP, 2020). These conditions render this region the most representative of the country.

By identifying the spatial distribution of Rice Paddies (RP) and other general Land Use / Land Cover (LULC) classes through corresponding cartography, policymakers and agricultural experts can enhance their understanding of the rural environments and develop effective strategies for sustainability.

LULC mapping could be appropriately developed by Geo Big Data (GeoBD) and Sentinel 2 (S2). GeoBD processes remote sensing data while considering velocity, veracity, volume, and value characteristics. This allows for handling massive, diverse, multi-temporal, multi-scalar, and complex data (Zhu, 2019) through tools such as Earth Observation Data Cube online portals (EOD) and Analysis-Ready Data (ARD). EODs are a solution for storing, organising, managing, and analysing remote sensing data in a previously impossible way (Giuliani *et al*., 2017) because they overcome restrictions connected to traditional local processing and data distribution methods. ARD is an excellent way to carry out remote sensing projects, as users can focus on analysis and denote inputs with the highest scientific standards and level of processing required for direct use in assessing LULC (Dwyer *et al*., 2018). ARD must satisfy conditions such as geometric and radiometric consistency and be organised into a specific format that supports stacking along the time dimension to create a time series or dense mosaics using all existing pixel values. GeoBD provides robust capabilities for linking time series and current machine learning classifiers for cartography, focusing on broad LULC (Simón-Sanchez *et al*., 2022) or RP (Huang & Zhang, 2022).

S2 is the multispectral freely accessible option offering the highest spatial (up to 10 m) and temporal resolutions. The revisit frequency of each S2 satellite is 10 days, and the combined constellation revisit is 5 days.

The current scientific literature has contributed to LULC mapping in Uruguay (e.g. Stanimirova *et al*., 2022; Zarza *et al*., 2022; Alciaturi *et al*., 2023). However, there is a noticeable lack of research on mapping rice crops. Therefore, this study aims to accomplish the following objectives: 1) map RP and other LUC classes for the agricultural period known as 'Zafra' from 2019 to 2020; 2) compare the classification capabilities of Random Forest (RF) and Support Vector Machine (SVM) for two S2 layer stacks; and 3) identify the most important features according to the RF.

Due to various limitations, SVM has constraints for identifying features importance. The availability of high-resolution imagery from Planet Group for cartography validation drove the decision to focus on the 2019-2020 season. Moreover, this time frame marked the most recent period at the beginning of the study.

Based on the resources available on Scopus or the Web of Science, this research project is considered pioneering in RP mapping for Uruguay.

2. The case study´s geographic context

The study area, located between 31° 49' 48" S—34° 26' 37" S and 53° 10' 51" ° W—55° 21' 35" W, covers 27,892 km^2. This surface is part of Uruguay and Brazil's 62,250 km^2 trans boundary watershed (also known as Cuenca de la Laguna Merín), (FIGURE 1). The predominant ecosystem, known as 'pampa', consists of diverse herbaceous communities and wetlands. The region could be divided into three landscape units based on meters above sea level: mountain ranges (150 – 517), hills (50 – 150), and lowlands (under 50), (Achkar *et al.*, 2012). From a farming perspective, the soils in the study area are generally poorly drained and offer prospects for mechanization in certain zones.

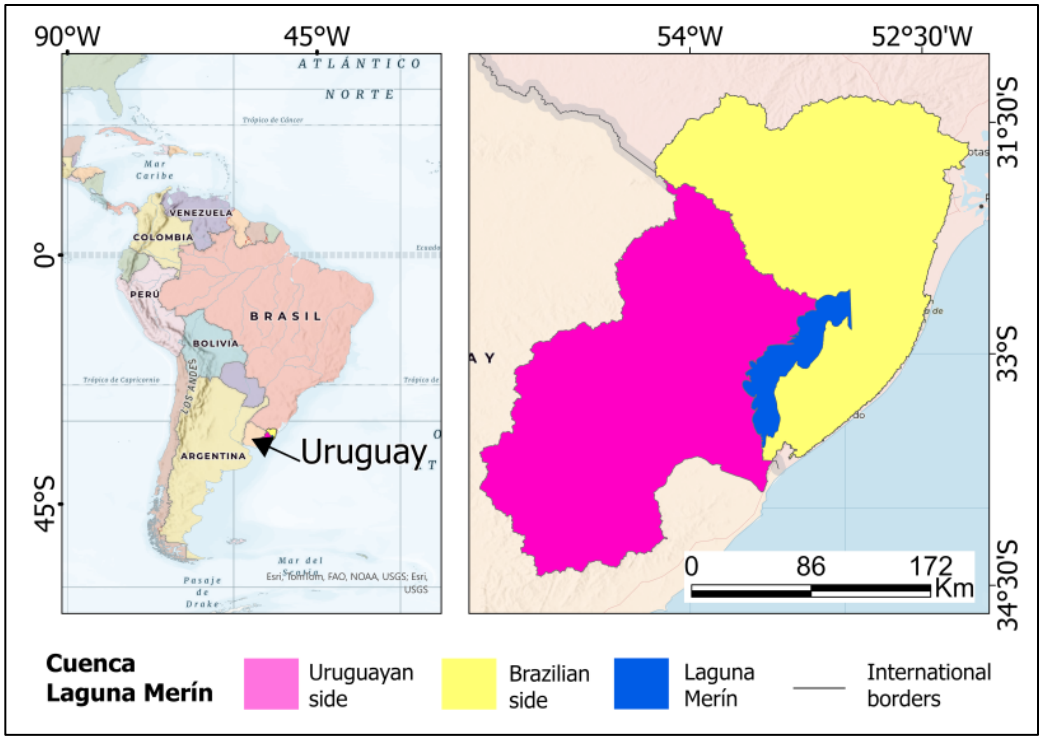

FIGURE 1. Study area's regional and local context

Lowlands are the most representative landscape with low drainage rates, leading to extensive flooded areas suitable for wetlands. The main reason for rice farming in CLM is the optimal water volume and flow from Laguna Merín or streams from an extensive network (Frank, 2022). Certain social factors positively impact rice activities, including government policies, partnerships between farmers and millers, and the development of local infrastructure such as roads and electricity (Zorrilla, 2015). Furthermore, the region also supports livestock, soybean cultivation, and small-scale fisheries.

3. Materials

This section focuses on the time series database (TSD), the classifiers, the software, the training samples, and the validation dataset.

3.1 Time series database

S2L2A is the Analysis ARD used to construct the TSD. Products are geometrically rectified and provide bottom-of-atmosphere reflectance. This reflectance is calculated by correcting the scattering of air molecules (Rayleigh scattering), the effects of atmospheric gases, such as oxygen, ozone, and water vapour, and the absorption and scattering due to aerosol particles. The usage of

TSD aims to differentiate between classes with different phenological cycles, with a specific focus on rice paddies (RP). It also enables the classification of other summer crops (OSC), bare land and its transitions (BLT), water bodies (WA), natural grasslands, livestock, and post-agricultural fields (NGPA), seasonally flooded vegetation (SFV), native forest and commercial afforestation (NFC), and built-up areas (BU). The TSD consists of two-layer stacks. One layer contains optical bands and the Normalised Vegetation Index (NDVI), while the other layer includes quoted inputs along with the Enhanced Vegetation Index (EVI) and Land Surface Water Index (LSWI). These quoted Indices have been widely used for RP (Zhao *et al*., 2021) and LUC mapping (Tobar-Díaz *et al*., 2023). The TSD construction required outlining the RP cultivation, spectral and time filtering, temporal composites and index computing, and layer stacking.

3.1.1 Outlining the RP cultivation

An outline of the process of cultivating RP in CLM is presented based on local experts. The cultivation process comprises five stages: planting, germination, vegetative reproduction, senescence-maturation, and harvesting (FIGURE 2).

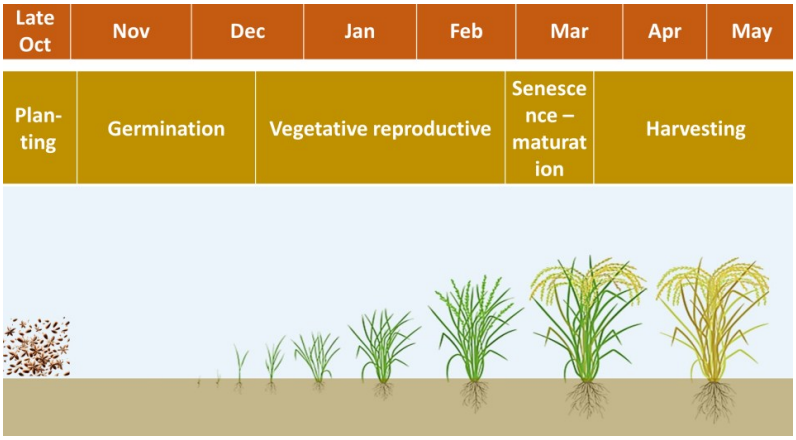

FIGURE 2. RP evolution for CLM. Source: adapted from Kuenzer & Knauer (2013)

RP cultivation exhibits unique temporal and spectral features that enable its identification from other LULC classes. During planting to early vegetative growth (November to December), bare soil or shallow water layers with limited vegetation cover were observed. Subsequently, in January, signs of vegetative growth became apparent as the plants developed a dense canopy. The peak vegetative growth stage was witnessed in February when the rice plants displayed a vigorous and healthy canopy. Following the postharvest period (typically between March and May), the vegetation is cleared, leaving the ground resembling bare soil. This outline guided the following spectral and time filtering.

3.1.2 Spectral and time filtering

Spectral filtering is limited to B2, B3, B4, B8, B11, and B12. The absorption of leaf pigments, such as chlorophyll a and b and carotenoids, significantly impacts the visible part of incoming radiation. These pigments are closely linked to the plant's physiological status. In this way, Van Niel & McVicar (2004) determined that red reflectance starts at 10% during emergence, decreases to 2% at flowering and gradually increases to 16%–18% at maturity due to the loss of green brightness by leaves and stems and the yellowness of the rice grains. In addition, Blackburn (1998) stated that near-infrared (NIR) reflectance changes over time according to biomass, increasing from a minimum of 15% during early tillering to a

maximum of 50% during heading. Finally, Short Wave Infrared (SWIR) enhances substrate discrimination due to its water absorption properties (Casanova *et al*., 1998).

The time filtering was limited to acquisitions from 10/1/2019 to 5/20/2020, with a cloud cover of 10% or less. The quoted imagery encompasses rice season and a few post-harvest days. This timeframe also lends itself to mapping the remaining LULC classes.

3.1.3 Temporal composites and Indices computing

Temporal composites (TC) are an effective way to map large areas by combining pixels based on statistical measures such as mean, median, minimum, and maximum across matching bands within a specific period (Meng *et al*., 2023). These calculations help fill in gaps in satellite data and reduce data anomalies (Carrasco *et al*., 2022). A critical advantage is their possibility to avoid clouds and shadows.

This study calculated TC by taking the median value across filtered scenes grouped by specific acquisition dates and RP stages (TABLE 1). Each TC is represented visually using the RGB B8/B11/B4 combination (FIGURE 3).

TABLE 1. S2 temporal composites

Groups	Acquisition dates	RP stage
S2nov	11/04/2019; 11/16/2019	Germination / Vegetative reproductive*
S2deca	11/28/2019; 12/10/2019	Vegetative reproductive
S2decb	12/22/2019; 01/03/2020	Vegetative reproductive
S2jan	01/15/2020; 01/27/2020	Vegetative reproductive
S2feb	02/08/2020; 02/20/2020	Vegetative reproductive
S2mar	03/03/2020; 03/15/2020	Senescence maturation
S2may	05/16/2020; 05/18/2020	Post agricultural fields

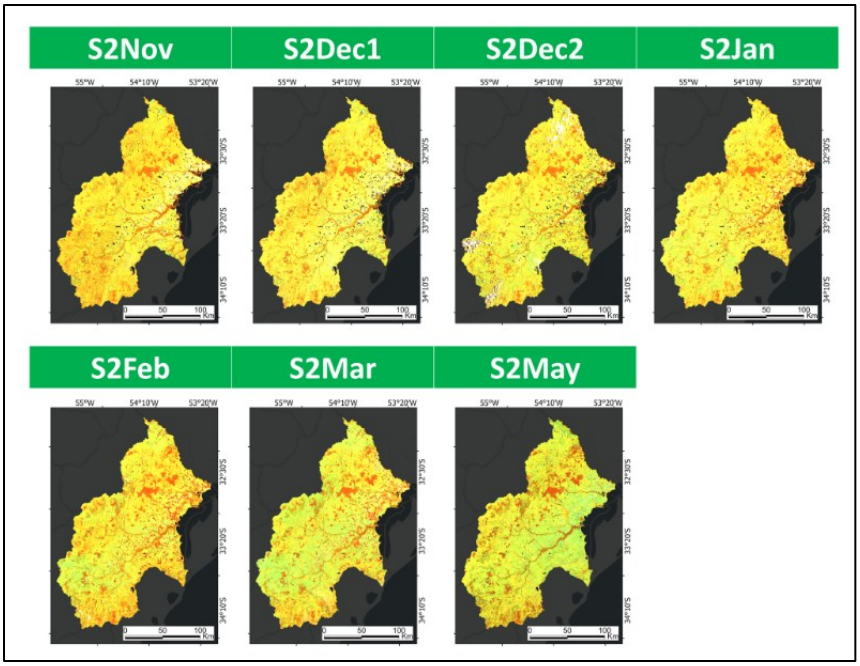

FIGURE 3. S2 composites

The NDVI, EVI, and LSWI calculations used various spectral bands from each time group. The results are displayed in FIGURE 4.

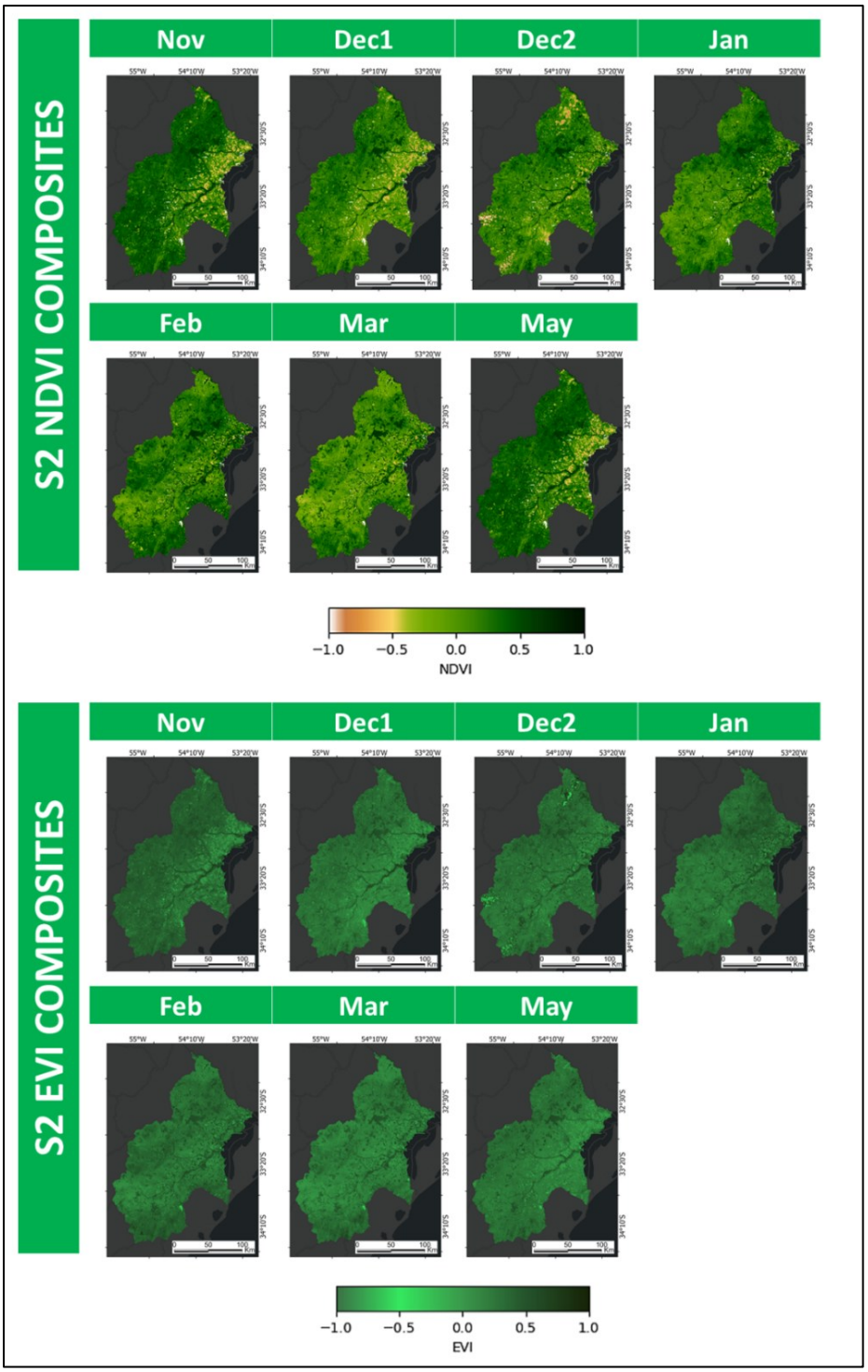

FIGURE 4. S2 Indices

3.1.4 Layer stacking

In FIGURE 5, layer stacking is depicted. The first stack, TSD_1, was formed using spectral bands from each time group and NDVI. The second stack, TSD_2, was created by adding the previous inputs and respective EVI and LSWI composites.

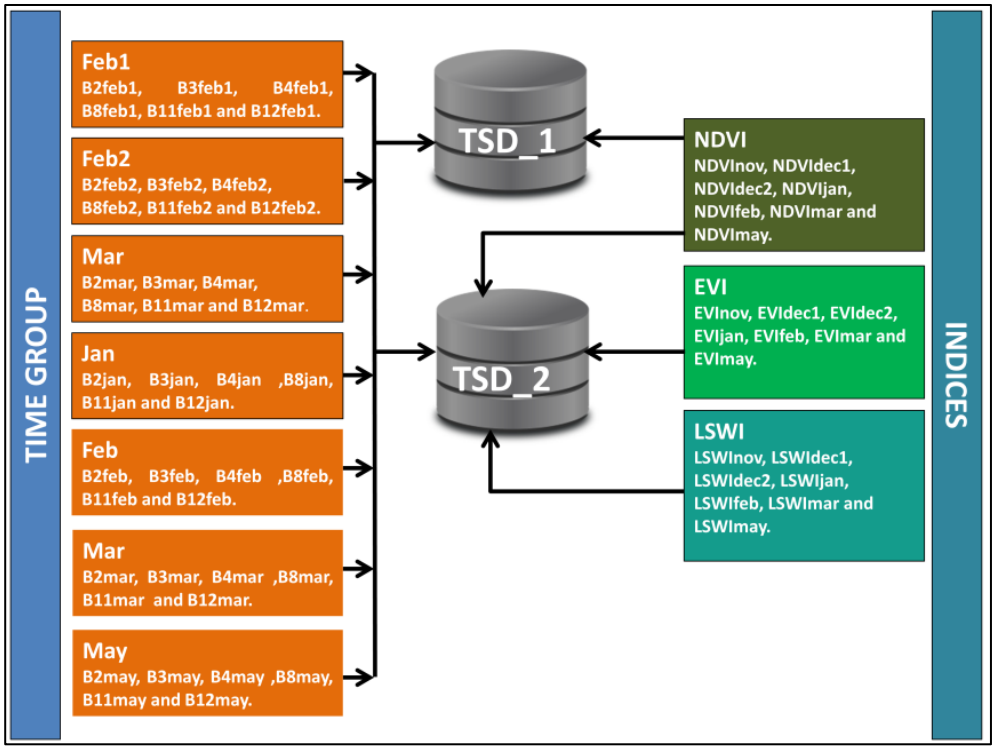

FIGURE 5. Layer stacking

3.2 The classifiers: Random Forest and Support Vector Machine

RF is a nonparametric statistical method used for classification tasks within a flexible framework. One of its main advantages is integrating data from different scales and sources (Ramo & Chuvieco, 2017). RF can identify the most prominent features by computing the impurity among decision trees using the mean decrease in Gini and the mean decrease in accuracy (Dunne *et al*., 2023). The algorithm has three primary HPS that must be set before training: the number of trees (specifies how many decision trees will be built in the forest), variables per split (determines how many variables will be considered when making a split at each node), and minimum leaf population (sets the minimum number of samples required to be in a leaf node).

SVM is a supervised, nonparametric learning technique that constructs hyperplanes or sets of hyperplanes in a high-dimensional space. The key hyperparameters to set for SVM are the kernel type, which is a mathematical function used to transform input data into a higher-dimensional space, and the cost, which controls the balance between maximizing the margin and minimizing classification errors on the training set (Adugna *et al*., 2022). One area for improvement of SVM is its inability to assess feature importance effectively, as it focuses on finding the best hyperplane to separate classes rather than individual feature weighting.

3.3 The software: GEE, GEEMAP & Python Scikit-Learn

GEE is a platform for retrieving, managing, and analysing large volumes of remote sensing data. Python was selected as the primary programming language for GEEMAP due to its compatibility with Scikit-learn's GridSearchCV. The last offers a wide range of capabilities, including assessing the importance of RF features and hyperparameter tuning (HPT) to identify the most effective combination of HPS (model) that enhance the performance of RF and SVM. The models are determined through k-fold cross-validation (Marcot & Hanea, 2021).

3.4 The training dataset

Experts' opinions and visual inspection of the S2 composites support the creation of a training dataset by digitising various representative features of RP, OSC, BLT, WA, NGPA, SFV, NFC, and BU. TABLE 2 shows the number of pixels per class.

TABLE 2. Sampling pixels per class

Class	RP	OSC	BLT	WA	NGPA	SFV	NFC	BU
Pixels	2000	1960	1000	400	3000	500	1000	200

3.5 The validation database

It was crucial to create a thorough validation database using information from four different sources: field surveys, visual analysis of high-resolution imagery, the Dynamic World project (Brown *et al*., 2022), and vector archives provided by experts in the rice industry. Essential matters are outlined below: a) Field surveys conducted on 10/2/2020 were instrumental in identifying different classes within the LUC, mainly RP, during a vigorous growth phase; b) polygons were digitised using imagery provided by the Planet consortium, specifically those acquired by the Dove constellation. The visual interpretation was carried out using the February monthly true-colour composites, which are consistent with RP's high vegetative stage; and c) the Dynamic World project enabled the creation of time LUC maps based on predefined dates. This allows for the creation of samples (excluding RP) that were previously impossible to obtain.

This integrated approach optimised the creation of a validation dataset across all CLM, addressing limitations associated with incomplete acquisitions from a single source. TABLE 3 displays pixels per class. Additionally, the official rice surface statistics from MGAP (2020) support validation.

TABLE 3. Validation pixels per class

Class	RP	OSC	BLT	WA	NGPA	SFV	NFC	BU
Pixels	4404	2828	338	196	6164	836	1778	120

4. Methods

The methods include class sampling, feature importance calculation, hyperparameters tuning and model creation, classification, and accuracy assessment.

4.1. The class sampling

TSD_1 and TSD_2 stacks, along with the training dataset, were used to produce TSD_1sam and TSD_2sam files. These files contain representative spectral and index values from each LULC class based on each layer stack.

4.2 Hyperparameter tuning and models creation

The HPT uses TSD_1sam and TSD_2sam files to determine the best models for classifying each layer stack according to RF and SVM. The first

step in HPT is to create a parameter grid as a dictionary of potential HPS and the values that need optimisation. TABLE 4 presents the HPS suggested for RF and SVM, broadening the alternatives proposed by Belgiu and Drăguţ (2016) and Shetty (2019).

TABLE 4. Alternatives of RF for optimising

RF		SVM	
HPS	Alternatives	HPS	Alternatives
Number of trees	50, 162, 275, 387, 500	Kernel	Linear, polynomial, sigmoid, Radial Basis Function
Variables per split	2,4,6	Cost	2, 5, 10, 15, 20, 35, 30
Min leaf population	2,4,6	Gamma	0.1,1,2,3

The datasets TSD_1sam and TSD_2sam were split into two subsets for model creation and testing purposes. The model creation subset comprised 70% of the data, while the remaining 30% was designated for testing the performance of each model. To enhance the robustness of the evaluation, K-Fold Partitioning was utilized. This method involves dividing the dataset into K subsets (or 'folds') and training the model K times, each time using a different fold as the test set and the remaining folds as the training set. The results of these K tests are then averaged to obtain a general measure of the model's performance. For the RF models, an exhaustive search for hyperparameters was conducted using 10-fold cross-validation, evaluating 540 parameter combinations, and performing 5,400 fits. Similarly, for the SVM process, there were five-fold cross-validations for each of the 112 parameter combinations, leading to a total of 560 fits. The final model was chosen based on the best-performing model configuration, also known as HPS.

4.3 Feature importance calculation

Feature importance is calculated based on Gini importance, which measures the relative importance of each feature in the model. It is determined from the decrease in Gini impurity resulting from splitting a node on a particular feature. Features with higher Gini importance are more influential in making predictions within the RF model.

4.4 Supervised classification

RF and SVM were optimised using models to classify TSD_1 and TSD_2.

4.5 Accuracy assessment

Validation was aided by extensively used statistics, such as overall accuracy (OA), user accuracy (UA), producer accuracy (PA) and Kappa coefficient (Kappa).

5. Results & discussions

The results and discussions are presented in four sections: model creation; maps, LUC surface estimations and accuracy statistics; efficiency of optical imagery and classifiers, and RF importance features.

5.1 Model creation

Based on each layer stack and classifier, a model was developed with corresponding optimal HPS values as detailed in TABLE 5.

TABLE 5. Layer stack, classifier, model names, and optimal hyperparameters

Layer stack	Classifier	Model name	Optimal hyperparameters
TSD _1	RF	S2tsRFmap	Number of trees = 162; Variables per split = 4; Min leaf population = 2
	SVM	S2tsSVMmap	C=2; Kernel=Linear
TSD _2	RF	S2tsRFmap2	Number of trees = 275; Variables per split = 4; Min leaf population = 2
	SVM	S2tsSVMmap2	C=2; Kernel=Linear

5.2 Maps, LUC surface estimations, and accuracy statistics

Four different maps were generated (FIGURE 6): the S2tsRFmap, S2tsSVMmap, S2tsRFmap2, and S2tsSVMmap2.

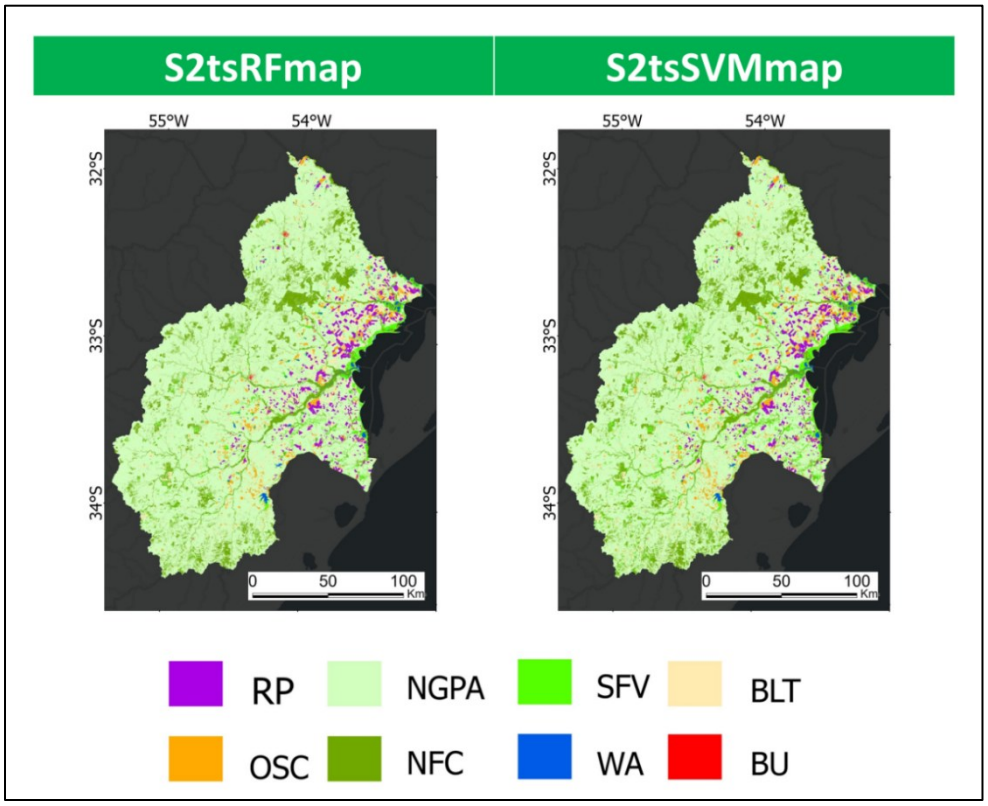

FIGURE 6. The maps

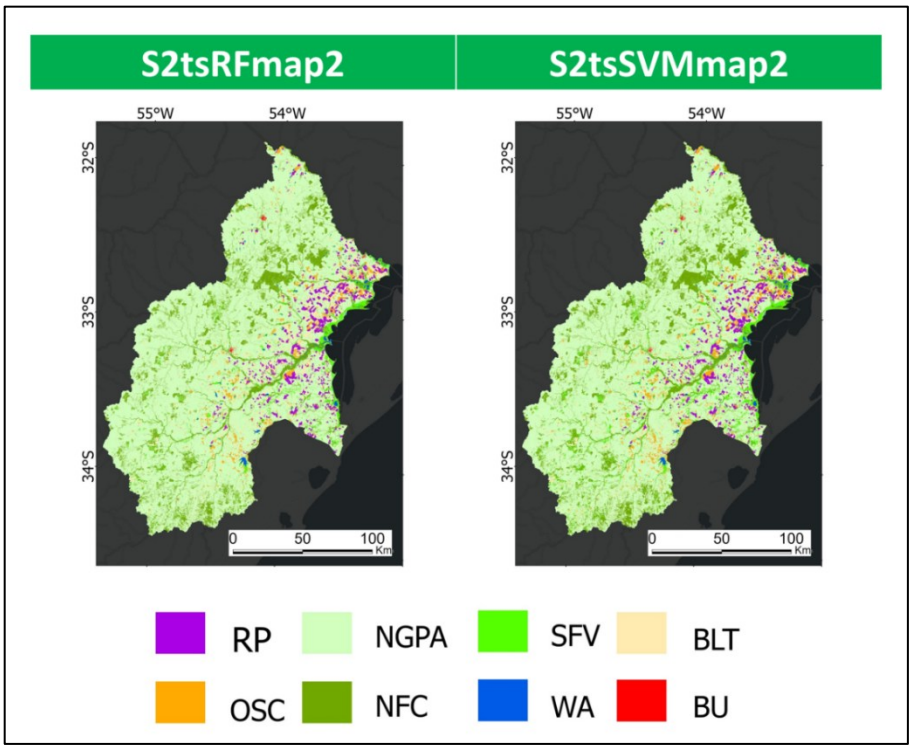

FIGURE 6. The maps (continued)

TABLE 6 provides an estimation of the surface area for each class. The RP category covers a small but consistent percentage of the CLM, ranging from 3.09% to 3.14%. On the other hand, NGPA dominates the landscape, as it covers over 70% of the area. NFC also covers a significant surface area, indicating the importance of commercial forestation activities. SFV shows inconsistency in classifier performance. BLT covers a range of 3.23% to 4.66%, which suggests that some areas are undergoing transitional phases or lack vegetation cover. Lastly, BU covers a small portion and comprises only 0.06% to 0.09%.

TABLE 6. Surface per class and map

Class	S2tsRFmap		S2tsSVMmap		S2tsRFmap2		S2tsSVMmap2	
	km^2	%	km^2	%	Km^2	%	Km^2	%
WA	175.64	0.61	140.32	0.49	169.08	0.59	133.17	0.46
NGPA	21432.95	74.47	20813	72.32	21669.49	75.29	20814.29	72.32
NFC	3311.44	11.51	3532.64	12.27	3232.51	11.23	3531.82	12.27
SFV	957.36	3.33	1522.75	5.29	953.31	3.31	1523.90	5.29
BLT	1341	4.66	929.98	3.23	1190.27	4.14	931.10	3.24
OSC	644.66	2.24	919.58	3.2	655.36	2.28	919.27	3.19
RP	890.52	3.09	904.31	3.14	887.72	3.08	905.09	3.14
BU	26.56	0.09	18.13	0.06	22.27	0.08	22.23	0.08
Total	28780.12	100	28780.7	100	28780.01	100.00	28780.86	100.00

The study utilised confusion matrices (TABLE 7) to compute accuracy metrics such as OA, Kappa, and estimators specific to each class. The findings indicated that all models performed well, with S2tsRFmap registering an OA accuracy of 91.5% and a Kappa of 0.887, S2tsSVMmap recording an OA accuracy of 92.6% and a Kappa of 0.901, S2tsRFmap2 achieving an OA accuracy of 92.4% and a Kappa of 0.899, and S2tsSVMmap2 is exhibiting an OA accuracy of 92.7% and a Kappa of 0.903.

TABLE 7. Confusion matrix for each map

S2tsRFmap									
	WA	NGPA	NFC	SFV	BLT	OSC	RP	BU	Total
WA	185	0	0	8	0	0	3	0	196
NGPA	0	5849	20	83	172	27	13	0	6164
NFC	0	181	1537	56	0	4	0	0	1778
SFV	16	4	57	736	0	10	13	0	836
BLT	0	56	0	0	282	0	0	0	338
OSC	0	143	0	0	280	2405	0	0	2828
RP	0	29	28	75	12	108	4152	0	4404
BU	0	5	0	0	13	0	0	102	120
Total	201	6267	1642	958	759	2554	4181	102	16664
OA			91,5%		Kappa			0.887	

TABLE 7. Confusion matrix for each map (continued)

S2tsSVMmap									
	WA	NGPA	NFC	SFV	BLT	OSC	RP	BU	Total
WA	180	0	0	14	0	0	2	0	196
NGPA	0	5877	18	114	70	55	14	0	6148
NFC	0	132	1521	140	1	2	0	0	1796
SFV	0	0	0	784	0	0	52	0	836
BLT	0	94	0	0	236	4	0	0	334
OSC	0	154	0	0	131	2539	0	0	2824
RP	0	12	28	97	0	80	4195	0	4412
BU	0	7	0	0	13	2	0	100	122
Total	180	6276	1567	1149	451	2682	4263	100	16668
OA			92,6 %		Kappa			0.901	
S2tsRFmap2									
	WA	NGPA	NFC	SFV	BLT	OSC	RP	BU	Total
WA	180	0	0	12	0	0	4	0	196
NGPA	0	5911	3	80	110	16	20	0	6140
NFC	0	163	1553	66	0	4	0	0	1786
SFV	32	8	11	749	0	0	24	0	824
BLT	0	85	0	0	247	0	0	0	332

OSC	0	142	0	1	262	2429	2	0	2836
RP	0	34	2	70	4	94	4208	0	4412
BU	0	6	0	0	13	0	0	103	122
TOTAL	212	6349	1569	978	636	2543	4258	103	16648
OA			92,4 %		Kappa			0.899	
S2tsSVMmap2									
	WA	NGPA	NFC	SFV	BLT	OSC	RP	BU	Total
WA	174	0	0	17	0	0	3	0	194
NGPA	0	5868	7	111	81	55	15	0	6137
NFC	0	113	1561	122	2	4	2	0	1804
SFV	0	1	0	779	0	0	52	0	832
BLT	0	104	0	0	226	4	0	0	334
OSC	0	175	0	0	143	2561	0	0	2879
RP	0	12	22	82	0	71	4210	0	4397
BU	0	4	0	0	13	0	0	101	118
Total	174	6277	1590	1111	465	2695	4282	101	16695
OA			92,7 %		Kappa			0.903	

TABLE 8 shows the accuracy statistics per class for all four maps. Most classes achieve optimal producer and user accuracy, while 'BLT' has poor precision scores. In summary, consistent performance is revealed in almost all classes, and challenges persist in accurately classifying 'BLT'
.

TABLE 8. The accuracy statistics per class

Class	**Maps**							
	S2tsRFmap		S2tsSVMmap		S2tsRFmap2		S2tsSVMmap2	
	PA	UA	PA	UA	PA	UA	PA	UA
WA	0.92	0.94	1	0.92	0.85	0.92	1	0.9
NGPA	0.93	0.95	0.94	0.96	0.93	0.96	0.93	0.96
NFC	0.94	0.86	0.97	0.85	0.99	0.87	0.98	0.87
SFV	0.77	0.88	0.68	0.94	0.77	0.91	0.7	0.94
BLT	0.37	0.83	0.52	0.71	0.39	0.74	0.49	0.68
OSC	0.94	0.85	0.95	0.9	0.96	0.86	0.95	0.89
RP	0.99	0.94	0.98	0.95	0.99	0.95	0.98	0.96
BU	1	0.85	1	0.82	1	0.84	1	0.86

5.3 Efficiency of optical imagery and classifiers

The RP surface estimated for the four maps aligns with the 1,007 km^2 reported by MGAP (2020) for the Uruguayan Easter region. This is particularly noteworthy when considering that CLM covers primarily, but not exclusively, the territory belonging to the region above.

The outcomes are consistent with recent studies that reached optimal accuracy for classifying RP or other LUC through optical imagery and RF or SVM. As prominent examples, (Zhang *et al*., 2020) reached 88.57% overall in an SVM employed for map RP in the Banan District and Zhongxian County of Southwestern China. De Abreu *et al*. (2021) identified RP for Rio Grande do Sul in Brazil at 96.5% OA. İnalpulat (2023) mapped various LUC classes for Çanakkale Province in Türkiye and found that RF could identify RP areas with 96% OA. Wei *et al*. (2022) created RP maps for China from 2014 to 2019, with Kappa coefficients ranging from 0.67–0.80. This research and other evidence support using classifiers such as RF or SVM as trustworthy alternatives for RP and general LUC mapping.

Although many classes demonstrated optimal user and producer accuracy performance, some issues require further attention. One is differentiating between herbaceous (natural and artificial) and post-agricultural land, which can be difficult due to similarities in spectral behaviour. Reinerman *et al*. (2020) findings support this statement.

It is crucial to differentiate between natural and cultivated forests. However, determining whether a forest is young or sparse can be challenging, as these forests can be mistaken for native trees. Fassnacht *et al*. (2016) came to similar conclusions. Additionally, it is important to thoroughly sample harvests of OSC (like soybeans or corn), which will likely lead to a more detailed classification.

It is important to focus on improving and accurately documenting the mapping of SFV over an extensive time series, as the analysed period only covers one rice season, which may not be representative of SFV patterns. The classification of SFV posed challenges due to the presence of various landscape areas, including shrubs, swampy woodlands, palm groves, and wooded prairie, which are associated with marshy, lacustrine, artificial, riverine, and other systems. Meeting the criteria outlined by Sahour *et al*. (2022) was difficult due to the presence of fuzzy boundaries and transition zones between wetlands and adjacent uplands, as well as the existence of elements causing variations in water spectral properties, making accurate mapping a challenge.

The low accuracy of BLT mapping is attributed to its resemblance to other substrates, such as sparse vegetation, leading to classification errors, particularly in classes with overlapping spectral characteristics.

5.4 The feature importance according to the RF

TSD_1 consists of 49 features. The top 20 belong to various moments during the 2019–2020 season, explaining 76% of the total. This underscores the importance and robustness of the time-series approach. Notably, features from May contribute about 27%, primarily due to their ability to highlight agricultural postharvest substrate evolution while retaining the spectral values of stable covers.

Elements from January, February and March, accounting for approximately 34% of the total, correspond with substantial phenological variations in dynamic features. The NDVI was highly ranked, and its importance summarised about 26%. This is because NDVI is capable of monitoring RP and OSC in multiple stages and because it has firmly known capabilities of correlating with the status of a broad array of vegetation properties for large-scale monitoring (Huang *et al*., 2021), such as different kinds of pastures (Edirisinghe *et al*., 2011) or forests with distinctive characteristics (Huete *et al*., 2002). SWIR has significant, as it provides about 31% importance. This value is due to the band's capabilities of differentiating water content and spongy mesophyll structure in crops and other distinctive vegetation in the study area. Visible region contributions are determined to be about 12% and mostly belong to the red band.

TSD_2 consists of 63 features. The top 20 features contributed almost 70% of the dataset's overall importance. Like TSD_1, several coincidences are observed here. The top concerns from the dataset are present in different periods throughout the 2019–2020 season. May remains the period with the most importance, accounting for 23% of the dataset. Moreover, the broad significance of January, February and

March reaches almost 29%. These results support earlier statements explaining the relevance of the time patterns observed in TSD_1. At this top, the Indices reached an importance of almost 39%, distributed, respectively, 19.2% (EVI), 11. 16% (NDVI) and 8.3% (LSWI). EVI's relative significance is explained by its capacity to identify vegetation structure variation, making it helpful in monitoring seasonal variations (Zhang *et al*., 2023) like those analysed in this research. The arguments previously made for NDVI are also valid in this circumstance. LSWI's importance may be explained by its accuracy in monitoring land surface water changes and mapping irrigated and flooded regions, such as those in the study area.

6. Conclusions

The methodology has proven to be effective in identifying most classes. It is trustworthy and aligned with the research objectives in the context of GeoBD attributes such as velocity, veracity, volume, and value. The four models exhibited remarkable accuracy on both training and test datasets, indicating that they have effectively learned from the training and can generalize to other suitable data. The similar accuracy of the four maps in global and per-class statistics provides robust evidence of the optimal performance of the RF and SVM classifiers. However, using these classifiers did not result in significant differences in accuracy. These results can be attributed to the adequate addressing of the HPT process, effectively avoiding typical constraints, such as methodological uncertainty, a common issue when implementing RF or SVM for LUC classifications.

The maps generated through a comprehensive analysis may serve as a valuable tool for addressing various questions related to rice production and other landscape environmental concerns associated with LUC in CLM. Since the availability of the S2 was optimal for this research, future initiatives should determine if other less dense time series, different from those utilised for this research or mono-temporal layer stacks, perform optimally in classifying LUC over the complex landscape that is CLM. Also, future research should employ diverse data sources, classify other crops besides RP, and test state of the art deep-learning approaches. Due to the study area's extensive surface, finding agricultural or other thematic detailed class validation datasets different from field campaigns or high-resolution imagery is challenging. Therefore, citizen science initiatives should be enforced as a potential source for achieving more comprehensive field sampling over time and space.

7. Acknowledgments

The authors thank the Planet Group for providing the high-resolution imagery and extend sincere thanks to Yannet Interian, José Rodó, and Gonzalo Carracelas.

8. References quoted

ACHKAR, M.; DOMÍNGUEZ, A. y F. PESCE. 2012. *Cuenca de la Laguna Merín – Uruguay: Aportes para la discusión ciudadana*. Redes – Amigos de la Tierra. Montevideo, Uruguay.

ADUGNA, T.; XU, W. & J. FAN. 2022. "Comparison of random forest and support vector machine classifiers for regional land cover mapping using coarse resolution FY-3C images". *Remote Sensing*, 14(3): 574.

ALCIATURI, G.; UMPIÉRREZ, R.; AGUDELO, F.; PANZL, R. y V. FERNÁNDEZ. 2023. Una propuesta para cartografiar el uso/cobertura de suelo mediante el Geo Big Data. Uruguay, año agrícola 2021-2022. *XVIII Conferencia Iberoamericana de Sistemas de Información Geográfica*. pp. 85-92. Cáceres, España. (Extended abstract).

BELGIU, M. & L. DRĂGUŢ. 2016. "Random forest in remote sensing: a review of applications and future directions". *ISPRS Journal of Photogrammetry and Remote Sensing*, 114: 24-31.

BLACKBURN, G. 1998. "Spectral indices for estimating photosynthetic pigment concentrations: A test using senescent tree leaves". *International Journal of Remote Sensing*, 19(4): 657- 675.

BRAY, F. 1986. *The rice economies: Technology & development in Asian societies*. Wiley. New York, United States of America.

BROWN, C.; BRUMBY, S.; GUZDER-WILLIAMS, B.; BIRCH, T.; BROOKS, S.; MAZARIELLO, J.;... & A. TAIT. 2022. "Dynamic World, Near real-time global 10 m land use land cover mapping". *Scientific data*, 9(1): 251.

CARRASCO, L.; FUJITA, G.; KITO, K. & T. MIYASHITA. 2022. "Historical mapping of rice fields in Japan using phenology and temporally aggregated Landsat images in Google Earth Engine". *ISPRS Journal of Photogrammetry and Remote Sensing*, 191: 277-289.

CASANOVA, D.; EPEMA, G. & J. GOUDRIAAN. 1998. "Monitoring rice reflectance at field level for estimating biomass and LAI". *Field Crops Research*, 55(1-2): 83-92.

DE ABREU, J.; LIMA, A.; DALAGNOL, R. & L. SOARES. 2021. Mapping irrigated rice using MSI/Sentinel-2 time series of vegetation indices and Random. *XXII Brazilian Symposium on Geoinformatics*, pp. 37-45. São José dos Campos, Brazil. (Extended abstract).

DUNNE, R.; REGUANT, R.; RAMARAO-MILNE, P.; SZUL, P.; SNG, L., LUNDBERG, M. & D. BAUER. 2023. "Thresholding Gini variable importance with a single-trained random forest: An empirical Bayes approach". *Computational and Structural Biotechnology Journal,* 21: 4354-4360.

DWYER, J.; ROY, D.; SAUER, B.; JENKERSON, C.; ZHANG, H. & L. LYMBURNER. 2018. "Analysis Ready Data: enabling analysis of the Landsat archive". *Remote Sensing*, 10(9): 1363.

EDIRISINGHE, A.; HILL, M.; DONALD, G. & M. HYDER. 2011. "Quantitative mapping of pasture biomass using satellite imagery". *International Journal of Remote Sensing*, 32(10): 2699-2724.

FASSNACHT, F.; LATIFI, H.; STEREŃCZAK, K.; MODZELEWSKA, A.; LEFSKY,M.; WASER, L.; ... & A. GHOSH. 2016. "Review of studies on tree species classification from remotely sensed data". *Remote Sensing of Environment*, 186: 64-87.

FOOD AND AGRICULTURE ORGANIZATION (FAO). 2004. "The international year of rice". Disponible en: https://www.fao.org/3/J1706e/J1706e00.htm. [Consulta: May, 2021].

FRANK, N. 2022. "Approach to labor mobility in the rice complex of Laguna Merín (Uruguay) using location-allocation techniques with Flowmap". *GeoFocus*, (29): 35-58.

GIULIANI, G.; CHATENOUX, B.; DE BONO, A.; RODILA, D.; RICHARD, J.; ALLENBACH, K.;... & P. PEDUZZI. 2017. "Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD)". *Big Earth Data*, 1(1-2): 100-117.

HUANG, S.; TANG, L.; HUPY, J.; WANG, Y. & G. SHAO. 2021. "A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing". *Journal of Forestry Research*, 32(1): 1-6.

HUANG, C. & C. ZHANG. 2022. "Time-series remote sensing of rice paddy expansion in the Yellow River Delta: towards sustainable ecological conservation in the context of water scarcity". *Remote Sensing in Ecology and Conservation*, 9(4): 454-468.

HUETE, A.; DIDAN, K.; MIURA, T.; RODRIGUEZ, E.; GAO, X. & L. FERREIRA. 2002. "Overview of the radiometric and biophysical performance of the MODIS vegetation indices". *Remote Sensing of Environment*, 83(1-2): 195-213.

İNALPULAT, M. 2023. "Comparison of Different Supervised Classification Algorithms for Mapping Paddy Rice Areas Using Landsat 9 Imageries". *Turkish Journal and Nature and Science*, 12: 52-59.

KUENZER, C. & K. KNAUER. 2013. "Remote sensing of rice crop areas". *International Journal of Remote Sensing*, 34(6): 2101-2139.

MARCOT, B. & A. HANEA. 2021. "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?". *Computational Statistics*, 36(3): 2009-2031.

MENG, X.; XIE, S.; SUN, L.; LIU, L. & Y. HAN. 2023. "Evaluation of temporal compositing algorithms for annual land cover classification using Landsat time series data". *International Journal of Digital Earth,* 16(1): 2574-2598.

MINISTERIO DE GANADERÍA, AGRICULTURA y PESCA (MGAP). 2020. *Encuesta de arroz. Zafra 2019 – 2020*. Informe anual. Montevideo, Uruguay.

PITTELKOW, C.; ZORRILLA, G.; TERRA, J.; RICCETTO, S.; MACEDO, I.; BONILLA, C. & A. ROEL. 2016. "Sustainability of rice intensification in Uruguay from 1993 to 2013". *Global Food Security*, 9: 10-18.

RAMO, R. & E. CHUVIECO. 2017. "Developing a Random Forest Algorithm for MODIS Global Burned Area Classification". *Remote Sensing*, 9(11): 1193.

REINERMANN, S.; ASSAM, S. & C. KUENZER. 2020. "Remote Sensing of Grassland Production and Management: a Review". *Remote Sensing*, 12(12): 1949.

SAHOUR, H.; KEMINK, K. & J. O'CONNELL. 2022. "Integrating SAR and Optical Remote Sensing for Conservation-Targeted Wetlands Mapping". *Remote Sensing*, 14(1): 159.

SHETTY, S. 2019. *Analysis of Machine Learning Classifiers for LULC Classification on Google Earth Engine. Faculty of Geo-Information Science and Earth Observation*. University of Twente. Enschede. The Netherlands. Master dissertation.

SIMÓN-SANCHEZ, A.; GONZALEZ-PIQUERAS, J.; DE LA OSSA, L. & A. CALERA. 2022. "Convolutional Neural Networks for Agricultural Land Use Classification from Sentinel-2 Image Time Series". *Remote Sensing*, 14(21): 5373.

STANIMIROVA, R.; GRAESSER, J.; OLOFSSON, P. & M. FRIEDL. 2022. "Widespread changes in 21st-century vegetation cover in Argentina, Paraguay, and Uruguay". *Remote Sensing of Environment*, 282: 113277.

TOBAR-DÍAZ, R.; GAO, Y.; MAS, J. y V. CAMBRÓN-SANDOVAL. 2023. "Clasificación de uso y cobertura del suelo a través de algoritmos de aprendizaje automático: revisión bibliográfica". *Revista de Teledetección*, 62: 1-19.

VAN NIEL, T. & T. McVICAR. 2004. "Current and potential uses of optical remote sensing in rice-based irrigation systems: a review". *Australian Journal of Agricultural Research*, 55(2): 155.

WEI, J.; CUI, Y.; LUO, W. & Y. LUO. 2022. "Mapping Paddy Rice Distribution and Cropping Intensity in China from 2014 to 2019 with Landsat Images, Effective Flood Signals, and Google Earth Engine". *Remote Sensing*, 14(3): 759.

ZARZA, R.; CAL, A.; FORMOSO, D.; MEDINA, S.; REY, D. & L. CARRASCO-LETELIER. 2022. "First delimitation and land-use assessment of the riparian zones at Uruguayan Pampa". *Ecological Informatics*, 71: 101781.

ZHANG, W.; LIU, H.; WU, W.; ZHAN, L. & J. WEI. 2020. "Mapping rice paddy based on Machine Learning with Sentinel-2 multi-temporal data: model comparison and transferability". *Remote Sensing*, 12(10): 1620.

ZHANG, T.; SONG, J.; FAN, Y.; LIU, Y.; YU, S., GUO, D. &...K. GUO. 2023. "Vegetation Index Research on the Basis of Tree-Ring Data: Current Status and Prospects". *Forests*, 14(10): 2016.

ZHAO, R.; LI, Y. & M. MA. 2021. "Mapping paddy rice with satellite remote sensing: a review". *Sustainability*, 13(2): 503.

ZHU, Y. 2019. "Geospatial semantics, ontology and knowledge graphs for Big Earth Data". *Big Earth Data*, 3(3): 187-190.

ZORILLA, G. 2015. "Uruguayan rice: the secrets of a success story". *Rice Today*, 14: 18-19.

Lugar y fecha de finalización del artículo:
Montevideo, Uruguay; septiembre, 2024