Teledetección y Cambio Global: Retos y Oportunidades para un Crecimiento Azul. XX Congreso de la Asociación Española de Teledetección
(Eds. Isabel Caballero, Gabriel Navarro, Luis Barbero y Jesús Gómez-Enri). 2024, Cádiz. ISBN: 978-84-9828-941-1

Predicción del rendimiento del cultivo de arroz mediante imágenes Sentinel-2 y el algoritmo *Random Forest*

Javier A. Quille-Mamani [1], Luis A. Ruiz [1], Juan Pedro Carbonell-Rivera [1], Lia Ramos-Fernández [2]

[1] Grupo de Cartografía GeoAmbiental y Teledetección, Universitat Politècnica de València, Camí de Vera s/n, 46022 Valencia, España.

[2] Departamento de Recursos Hídricos, Universidad Nacional Agraria La Molina, Lima 15024, Perú.

Resumen: La predicción del rendimiento de los cultivos es crucial para la gestión y planificación de la política alimentaria. Este estudio tiene el objetivo de generar y evaluar modelos de rendimiento del cultivo de arroz en el norte de Perú utilizando imágenes Sentinel-2, aplicando el algoritmo *Random Forest* (RF). Para ello, se seleccionaron 12 fechas de imágenes Sentinel-2 utilizando la plataforma *Google Earth Engine* (GEE), en las que se evaluaron 37 parcelas de arrozal. Se calcularon ocho índices de vegetación: índice de vegetación mejorado (EVI), índice de clorofila verde (GCI), índice de vegetación ajustado al suelo modificado 2 (MSAVI2), índice de estrés hídrico (MSI), índice de vegetación de diferencia de agua (NDWI), índice de proporción de pigmento clorofila normalizado (NPCI) e índice de vegetación ajustado al suelo (SAVI). Se generaron modelos para cada fecha, calculándose el coeficiente de determinación (R^2). El valor máximo se obtuvo para la fecha 7 (R^2 = 0.79). Aplicándose el algoritmo de RF y validación cruzada (*leave one out*), los mejores resultados se obtuvieron con el NDWI y MSI, con un R^2 de 0.57, error medio cuadrático (RMSE) de 1.51 t/ha y error medio absoluto (MAE) de 1.15 t/ha. El modelo obtenido por RF con validación cruzada proporciona una fiabilidad notable para predecir el rendimiento antes de la cosecha.

Palabras clave: aprendizaje automático, teledetección, imágenes multiespectrales, índice de vegetación.

Rice crop yield prediction using Sentinel-2 imagery and the Random Forest algorithm

Abstract: *Crop yield prediction is crucial for food policy management and planning. This study aims to generate and evaluate rice crop yield models in northern Peru using Sentinel-2 images and applying the Random Forest (RF) algorithm. Twelve dates of Sentinel-2 images were selected for 37 rice crop plots, working on the Google Earth Engine (GEE) platform. Eight vegetation indices were calculated: enhanced vegetation index (EVI), green chlorophyll index (GCI), modified soil adjusted vegetation index 2 (MSAVI2), water stress index (MSI), difference water difference vegetation index (NDWI), normalized chlorophyll pigment ratio index (NPCI) and soil adjusted vegetation index (SAVI). Models were generated for each date and the coefficient of determination (R^2) was calculated. The maximum value was obtained for date 7 (R^2 = 0.79). Applying the RF and cross-validation (leave one out) algorithm, the best results were obtained with the NDWI and, MSI with an R^2 of 0.57, root mean square error (RMSE) of 1.51 t/ha and mean absolute error (MAE) of 1.152 t/ha. The model obtained by RF with cross-validation is more plausible and provides reliability for predicting pre-harvest yield.*

Keywords: *machine learning, remote sensing, multispectral imagery, vegetation index.*

1. INTRODUCCIÓN

El arroz (*Oryza sativa* L.), es considerado uno de los más importantes cultivos para el mundo, lo consume más del 50% de la población mundial (Choudhary *et al.*, 2022; Ge *et al.*, 2021). El 20% de la energía alimentaria consumida por el ser humano es proporcionada por el cultivo de arroz. La predicción del rendimiento antes de la cosecha es esencial para garantizar la seguridad alimentaria y optimizar la producción, promoviendo la sostenibilidad en la agricultura (Htun *et al.*, 2023). El cultivo de arroz en la costa sur del Perú supone la mayor actividad agrícola, teniendo especial importancia económica y social, al representar el 28% de la producción del país, de ahí la importancia de predecir el rendimiento ante diferentes escenarios climatológicos.

La teledetección es una herramienta importante para el monitoreo de los cultivos. La plataforma *Google Earth Engine* (GEE) facilita el uso de imágenes satelitales, disponiendo entre su catálogo de las imágenes Sentinel-2. En anteriores estudios, las bandas espectrales de estas imágenes y los índices de vegetación (IV) derivados de ellas se han utilizado para la predicción de rendimiento (Franch *et al.*, 2021).

Las diversas técnicas de aprendizaje automático, entre las cuales destaca el algoritmo *Random Forest* (RF), son frecuentemente empleadas en investigaciones debido a su robustez frente a la colinealidad y a la distribución no normal de las variables introducidas. Este algoritmo se ha empleado para la predicción del rendimiento en diversos cultivos, habiendo demostrado su eficacia en investigaciones previas sobre maíz (Avestisyan y Cvetoanova, 2019), soja (Crusiol *et al.*, 2022) y arroz (Kim *et al.*, 2019).

En este estudio presentamos un enfoque de predicción del rendimiento del cultivo de arroz utilizando imágenes Sentinel-2, aplicando el algoritmo RF.

Citar como: Quille-Mamani, J.A., Ruiz, L.A., Carbonell-Rivera, J.P., Ramos-Fernández, L. (2024). Predicción del rendimiento del cultivo de arroz mediante imágenes Sentinel-2 y el algoritmo Random Forest. En *Teledetección y Cambio Global: Retos y Oportunidades para un Crecimiento Azul, Actas del XX Congreso de la Asociación Española de Teledetección*, pp. 365-368. 2024, Cádiz. Disponible en http://www.aet.org.es/?q=congresos

2. MATERIAL Y MÉTODOS

2.1. Zona de estudio

La zona de estudio pertenece al distrito de Ferreñafe, región de Lambayeque, Perú (79°47′09.73″ W; 6°35′36.68″ S; 46 m.s.n.m.). Esta zona recibe una precipitación media anual de 22 mm, teniendo temperaturas medias anuales con una mínima promedio de 15.4 °C y una máxima promedio de 28.8 °C. La recolección de los datos de rendimiento se llevó a cabo en 37 parcelas de arroz (Tabla 1), con dimensiones de 0.5 a 2 ha, la variedad predominante es INIA508-Tinajones y el periodo vegetativo fue desde el 22 de diciembre de 2021 al 15 de junio de 2022.

Tabla 1. *Datos de área y rendimiento de los campos de arroz en la zona de estudio.*

N°	área (ha)	Rendimiento (t/ha)	N°	área (ha)	Rendimiento (t/ha)
1	1.98	10.41	21	0.99	13.26
2	1.74	9.66	22	1.38	14.31
3	1.58	9.72	23	1.49	12.29
4	1.11	10.45	24	0.79	15.95
5	0.96	10.51	25	0.95	13.88
6	0.76	10.90	26	1.02	10.61
7	0.75	10.75	27	1.42	7.36
8	1.16	9.91	28	1.11	7.87
9	1.13	11.06	29	0.80	6.28
10	0.90	10.18	30	1.58	4.98
11	0.78	11.53	31	0.66	5.06
12	1.13	10.71	32	0.99	7.57
13	1.03	10.81	33	1.06	11.97
14	1.15	8.69	34	0.74	9.01
15	0.64	11.63	35	1.24	12.28
16	1.14	8.11	36	0.82	10.35
17	1.11	10.00	37	1.99	8.81
18	1.03	12.04			
19	0.85	10.21			
20	0.77	11.91			

2.2. Datos de campo y procesado de imágenes

Los datos de rendimiento real se obtuvieron para cada parcela en el momento de la cosecha. A partir de las imágenes Sentinel-2, con una resolución de 10 m, se calcularon ocho IV: índice de vegetación mejorado (EVI), índice de clorofila verde (GCI), índice de vegetación ajustado al suelo modificado 2 (MSAVI2), índice de estrés hídrico (MSI), índice de vegetación de diferencia de agua (NDWI), índice de proporción de pigmento clorofila normalizado (NPCI) e índice de vegetación ajustado al suelo (SAVI) (Tabla 2). Las imágenes se filtraron, de forma que tuvieran un porcentaje menor al 30% de nubosidad, y se consideraron los valores máximos de los índices en intervalos de 15 días para asegurar la disponibilidad de imágenes sin nubosidad.

Tabla 2. *Índices de vegetación aplicados para la evaluación del rendimiento de arroz.*

Índice de veget.	Fórmula	Referencia
EVI	2.5 x (NIR - R) / (NIR + 6xR --7.5 x B + 1)	Choudhary *et al.*, 2020
GCI	NIR / G −- 1	Wang *et al.*, 2023
GNDVI	(NIR - G) / (NIR + G)	Htun *et al.*, 2023
MSAVI2	0.5 x (2 x NIR + 1) - (((2 x NIR)2 - 8 x (NIR - R))1/2)	Misra *et al.*, 2020
MSI	SWIR1/NIR	Avetisyan & Cvetanova, 2019
NDVI	(NIR - R) / (NIR + R)	Choudhary *et al.*, 2020
NDWI	(NIR - SWIR2) / (NIR + SWIR2)	Htun *et al.*, 2023
NPCI	(R - B) / (R + B)	Wang *et al.*, 2023
SAVI	(NIR - R) x (1 + L) / (NIR + R + L)*	Htun *et al.*, 2023

*L con un valor de 0.5.

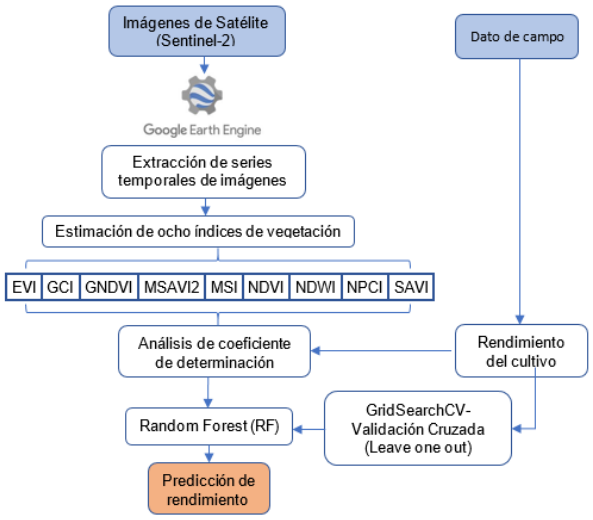

Figura 1. *Diagrama de flujo de la metodología general.*

Se obtuvieron datos para 12 fechas con intervalos de 15 días, donde se estableció una relación entre los IV y el rendimiento en cada fecha correspondiente. Para estimar el rendimiento se seleccionó el valor máximo de los IV en cada fecha específica. Además, se empleó el algoritmo RF, para identificar las variables más importantes y generar la predicción del rendimiento (Figura 1). Todos los cálculos fueron llevados a cabo en el lenguaje de programación Python (versión 3.1.1) utilizando la librería Scikit learn (versión 1.2.2).

2.3. Análisis estadístico

El análisis de datos, para determinar la importancia y la permutación para una reducción dimensional de las variables de predicción de rendimiento, se realizó mediante RF y se evaluó con validación cruzada *Leave One Out* (*LOOCV)* (Wong, 2015). Posteriormente se evaluando las muestras de aprendizaje obteniendo su coeficiente de

Tabla 3. *Coeficiente de determinación entre el rendimiento y los índices de vegetación en diferentes fechas del cultivo de arroz.*

Fechas de estudio	EVI	GCI	GNDVI	MSAVI2	MSI	NDVI	NDWI	NPCI	SAVI	COM_IV
Fecha_1	0.01	0.18	0.00	0.01	0.05	0.02	0.00	0.00	0.01	0.42
Fecha_2	0.01	0.02	0.00	0.01	0.03	0.02	0.01	0.02	0.00	0.64
Fecha_3	0.00	0.00	0.04	0.00	0.01	0.00	0.00	0.06	0.00	0.55
Fecha_4	0.01	0.01	0.00	0.01	0.00	0.01	0.06	0.14	0.01	0.41
Fecha_5	0.06	0.01	0.07	0.04	0.00	0.02	0.00	0.00	0.05	0.53
Fecha_6	0.28	0.00	0.44	0.28	0.32	0.27	0.18	0.29	0.28	0.74
Fecha_7	0.36	0.04	0.39	0.37	**0.69**	0.30	0.40	0.24	0.36	**0.79**
Fecha_8	0.43	0.17	0.39	0.45	0.59	0.39	0.33	0.17	0.43	0.73
Fecha_9	0.37	0.29	0.41	0.39	0.37	0.39	0.41	0.00	0.39	0.68
Fecha_10	0.25	0.17	0.27	0.27	0.02	0.17	0.08	0.34	0.25	0.79
Fecha_11	0.07	0.00	0.08	0.07	0.05	0.10	0.16	0.00	0.08	0.65
Fecha_12	0.43	0.24	0.37	0.42	0.30	0.51	0.38	0.04	0.44	0.68

determinación (R^2), la raíz del error cuadrático medio (RMSE) y el error medio absoluto (MAE).

3. RESULTADOS Y DISCUSIÓN

Tras un análisis con las bandas espectrales e IV, no se encontraron diferencias significativas en los coeficientes de determinación. La Tabla 3 muestra los coeficientes de determinación de los modelos obtenidos con los IV individualmente y la combinación de IV en relación con el rendimiento en cada fecha correspondiente. Se observa que el MSI obtiene un R^2 (0.69) en la fecha 7, superior al resto de IV y fechas. Además, en la misma fecha 7, se obtiene un R^2 de 0.79 mediante la combinación de índices de vegetación (COM_IV).

En la figura 2, se presenta la optimización de los parámetros de RF y la reducción dimensional de las variables. En la figura 2a se ilustra la elección del número de árboles con respecto al error OOB (*Out of bag*) con un valor de 141. En la figura 2b se muestra la selección del número máximo de variables, indicando un valor de dos. Ambos casos fueron evaluados mediante la validación cruzada de *leave one out* (LOOCV). Las figuras 2 (c y d) muestran la importancia de las variables y la permutación a través de RF, respectivamente. En ambos escenarios se resaltan dos variables más significativas, MSI y NDWI.

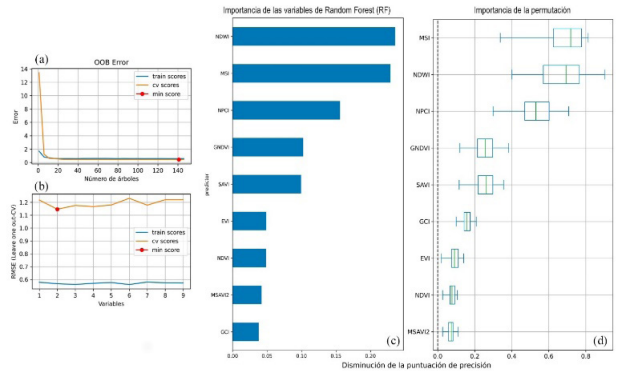

Figura 2*. Random Forest para la estimación de la fecha 7, a. error OOB, b. Selección del número de variables utilizando validación cruzada (leave one out) basada en RMSE mínimo, c. importancia de las variables de Random Forest, d. importancia de la permutación de la variable predictiva.*

El modelo generado con RF utilizando LOOCV obtuvo un test-RMSE de 2.33 t/ha. En la Figura 3, se presenta una comparación entre el rendimiento real y el predicho mediante el algoritmo RF. Los resultados muestran un R^2 de 0.57, RMSE de 1.51 t/ha y MAE de 1.15 t/ha. Estos resultados son coherentes con Wan *et al.* (2020), que reportaron resultados similares en predicciones por etapas fenológicas, obteniendo un rango de R^2 de 0.24 a 0.64 durante la etapa de crecimiento, R^2 de 0.53 durante el macollamiento, R^2 de 0.75 en floración y un R^2 de 0.53 en el llenado de grano. Sin embargo, estudios de Franch *et al.* (2021), exhiben valores más elevados, con un R^2 de 0.69 en el estado de floración y 0.67 en el estado de llenado de grano.

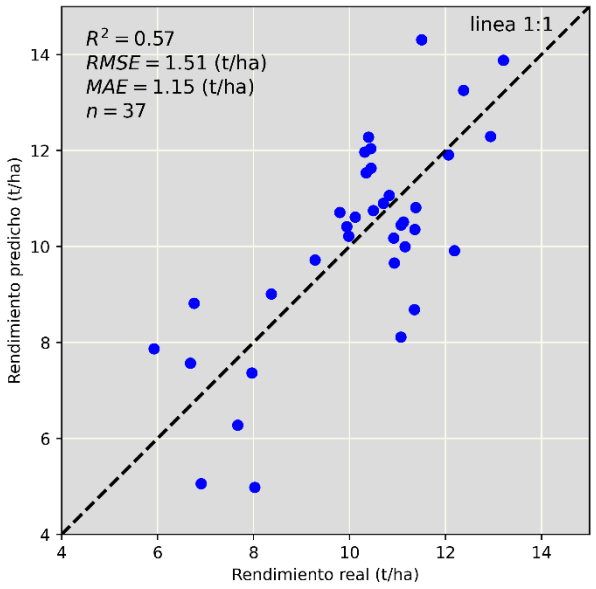

Figura 3. *Comparación del rendimiento real vs el predicho con Random Forest.*

4. CONCLUSIONES

En este estudio, se estimó el rendimiento del cultivo de arroz en una zona árida, utilizando datos del satélite Sentinel-2 mediante IV y la aplicación del algoritmo de RF, analizándose la importancia de las variables predictoras. Además, se muestra que la regresión de RF y la validación cruzada proporcionan una buena fiabilidad en la predicción de rendimiento, incluso con un reducido conjunto de 37 parcelas. Se sugiere la aplicabilidad de esta metodología en diferentes contextos y con mayor cantidad de parcelas agrícolas, además de la inclusión de más variables (climatológicas y fenológicas) para mejorar la precisión en el ámbito de la agricultura.

5. AGRADECIMIENTOS

Beca de generación del bicentenario del gobierno peruano para la realización de la tesis (Programa Nacional de Becas y Crédito Educativo (Pronabec)del Ministerio de Educación (Minedu) del Perú).

6. BIBLIOGRAFÍA

Avetisyan, D., Cvetanova, G. (2019). Water Status Assessment in Maize and Sunflower Crops Using Sentinel-2 Multispectral Data. *Space Ecol. Saf*, 152-157.

Choudhary, K., Shi, W., Dong, Y., Paringer, R. (2022). Random Forest for rice yield mapping and prediction using Sentinel-2 data with Google Earth Engine. *Advances in Space Research*, *70*(8), 2443-2457. https://doi.org/10.1016/j.asr.2022.06.073

Crusiol, L.G.T., Sun, L., Sibaldelli, R.N.R., Junior, V.F., Furlaneti, W.X., Chen, R., Sun, Z., Wuyun, D., Chen, Z., Nanni, M.R., Furlanetto, R.H. (2022). Strategies for monitoring within-field soybean yield using Sentinel-2 Vis-NIR-SWIR spectral bands and machine learning regression methods. *Precision Agriculture*, *23*(3), 1093-1123. https://doi.org/10.1007/s11119-022-09876-5

Franch, B., Bautista, A.S., Fita, D., Rubio, C., Tarrazó-Serrano, D., Sánchez, A., Skakun, S., Vermote, E., Becker-Reshef, I., Uris, A. (2021). Within-field rice yield estimation based on Sentinel-2 satellite data. *Remote Sensing*, *13*(20), 4095. https://doi.org/10.3390/rs13204095

Htun, A.M., Shamsuzzoha, M., Ahamed, T. (2023). Rice yield prediction model using normalized vegetation and water indices from Sentinel-2A satellite imagery datasets. *Asia-Pacific Journal of Regional Science*, 1-29. https://doi.org/10.1007/s41685-023-00299-2

Kim, J., Lee, J., Sang, W., Shin, P., Cho, H., Seo, M. (2019). Rice yield prediction in South Korea by using random forest. *Korean Journal of Agricultural and Forest Meteorology*, *21*(2), 75-84.

Misra, G., Cawkwell, F., Wingler, A. (2020). Status of phenological research using Sentinel-2 data: A review. *Remote Sensing*, *12*(17), 2760. https://doi.org/10.3390/rs12172760

Wan, L., Cen, H., Zhu, J., Zhang, J., Zhu, Y., Sun, D., Du, X., Zhai, L., Weng, H., Li, Y., Li, X. (2020). Grain yield prediction of rice using multi-temporal UAV-based RGB and multispectral images and model transfer–a case study of small farmlands in the South of China. *Agricultural and Forest Meteorology*, *291*, 108096. https://doi.org/10.1016/j.agrformet.2020.108096

Wang, X., Blesh, J., Rao, P., Paliwal, A., Umashaanker, M., Jain, M. (2023). Mapping cover crop species in southeastern Michigan using Sentinel-2 satellite data and Google Earth Engine. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1035502

Wong, T.T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition, 48*, 2839-2846. https://doi.org/10.1016/j.patcog.2015.03.009