

Clasificación de imágenes satelitales para la detección de cultivos y cuantificación de áreas sembradas en Argentina

Juan Ignacio Vázquez Broqua
juanivazquez@gmail.com

Fernando Antoni
fernandoantoni98@gmail.com

Virginia Diaz Villa
virginia.diazvilla@gmail.com

Santiago Guizzardi
santiagoguizzardi@gmail.com

Joan Alberto Cerretani
joancerretani@gmail.com

Abstract—

En este trabajo se propone estimar la cantidad de hectáreas sembradas con maíz y soja para la campaña 20/21 en los partidos de Roque Saenz Peña (Córdoba), General Roca (Córdoba) y General Villegas (Buenos Aires). Para esto se utilizaron técnicas de clasificación por píxel utilizando un modelo de Random Forest y SVM sobre distintos conjuntos de entrenamiento. Los mejores resultados se obtuvieron utilizando ensambles de modelos por fecha con información de todas las bandas o aplicando PCA a las mismas. También se realizó una comparación en la estimación de áreas por píxel y a través del método de segmentación de imagen, que presentó una mejora en la performance del modelo.

Index Terms—teledetección, cultivos, satelitales

I. INTRODUCCIÓN

La clasificación de imágenes satelitales es un procedimiento útil para extraer información sobre temáticas ambientales y socioeconómicas [1]. La obtención de mapas temáticos de tipos de cultivos puede ser utilizada para beneficio de los productores y de la industria agrícola [2].

Existen diversas maneras de clasificar las diferentes metodologías de clasificación. Entre las técnicas más utilizadas se destacan los modelos de redes neuronales y máquinas de soporte vectorial (SVM).

Otra forma de distinguir estas técnicas es si la clasificación es basada en pixeles o en objetos [3]. La clasificación basada en pixeles realiza una clasificación para cada píxel de la imagen, mientras que la clasificación por objeto se realiza sobre una segmentación de la imagen; de esta forma todos los pixels dentro del polígono de la segmentación reciben la misma clasificación.

Por otro lado, la clasificación puede realizarse a partir de multi-datos, es decir utilizar diferentes sensores que miden diferentes características del terreno (imagen RGB, imagen infrarroja, etc.), así como también el uso de datos multi-temporales, es decir, utilizar información temporal (por ejem-

plo, la evolución de los cultivos) para realizar una clasificación mas precisa.

En este trabajo se propone estimar la cantidad de hectáreas sembradas con maíz y soja para la campaña 20/21 en los partidos de Roque Saenz Peña (Córdoba), General Roca (Córdoba) y General Villegas (Buenos Aires), utilizando distintos conjuntos de entrenamiento, clasificadores y metodologías de estimación de áreas.

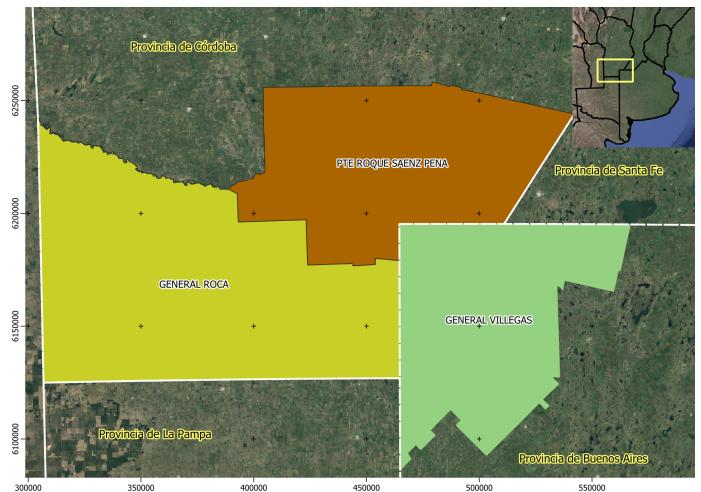


Fig. 1. Partidos de Roque Saenz Peña (Córdoba), General Roca (Córdoba) y General Villegas (Buenos Aires).

II. MÉTODOS

A. Estructura de los datos

Se obtuvieron imágenes del satélite Sentinel-2 correspondientes a seis fechas: Octubre, Noviembre y Diciembre del 2020 y en Enero, Febrero y Marzo del 2021. Cada imagen está compuesta por 10 bandas (Tabla I).

canal	nombre	descripción
1	B2	Azul
2	B3	Verde
3	B4	Rojo
4	B5	Borde rojo 1
5	B6	Borde rojo 2
6	B7	Borde rojo 3
7	B8	NIR
8	B8A	Borde rojo 4
9	B11	SWIR 1
10	B12	SWIR 2

TABLE I

Por otro lado, se contó con información de campo de 466 lotes donde cada punto identifica el tipo de cultivo: soja, maíz, girasol, alfalfa y cobertura natural (Fig. 2 y 3). Tanto los valores de verdad de campo como la imágenes originales fueron re-proyectados al sistema de referencia de coordenadas UTM Zona 20S (EPSG:32720).

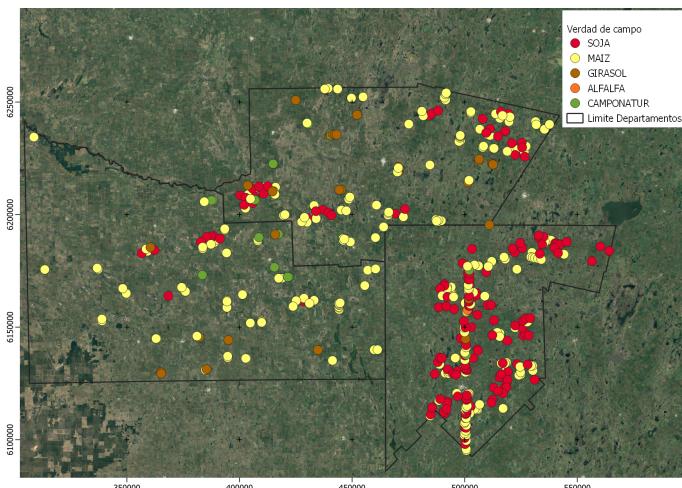


Fig. 2. Puntos de "verdad de campo".

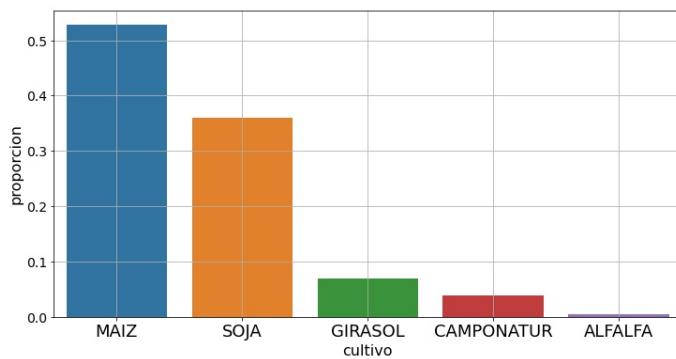


Fig. 3. Proporción de cultivos entre los datos de "verdad de campo".

Alrededor de cada punto se construyó un buffer de 40 m, de manera de garantizar que cada área buffer se encontrara dentro de los límites de los lotes. A partir de este procedimiento se pudo incrementar las cantidad de observaciones de

entrenamiento de 466 a 7675 sin perdida de generalidad (Tabla II).

Cultivo	Id	cant Obs	cant Pixels
SOJA	1	134	2203
MAIZ	2	107	1754
MAIZ	3	139	2278
SOJA	4	34	563
GIRASOL	5	32	539
ALFALFA	10	2	30
CAMPO NATURAL	20	18	308

TABLE II

B. Pre-Procesamiento

En primer lugar se calcularon dos índices de vegetación, el NDVI o índice de vegetación de diferencia normalizada y el EVI o índice de vegetación mejorado. Ambos índices son ampliamente utilizados en estudios de vegetación. Ambos índices se basan en las diferencias en reflectancia en las bandas del infrarrojo cercano (IR) y rojo (R). Los tejidos fotosintéticos absorben la mayor parte de la radiación que se encuentra en el rango del visible, mientras que reflejan o trasmiten la radiación de longitudes de onda mayores (infrarrojo). La informaciónpectral proveniente de estas regiones permite caracterizar la vegetación debido al patrón característico de absorción y reflexión de la luz en estas zonas del espectro, llamado firmaespectral [4] [5] [6] [7] [8] [9] [10]. Cualquier factor que afecte el estado fisiológico o la capacidad fotosintética de las plantas afectará el patrón de reflexión en dichas longitudes de onda.

$$NDVI = \frac{NIR - Red}{NIR + Red} = \frac{B8 - B4}{B8 + B4}$$

$$EVI = 2.5 \frac{B8 - B4}{(B8 + 6 \cdot B4 - 7.5 \cdot B2) + 1}$$

Por otro lado, se realizó un Análisis de Componentes Principales (PCA) de las 12 capas de NDVI y EVI y se seleccionaron las primeras 6 componentes.



Fig. 4. NDVI para Noviembre del 2020.

Luego de los pre-procesamientos descriptos se generaron los siguientes conjuntos de entrenamientos:

- 1) NDVI para las 6 fechas
- 2) EVI para las 6 fechas
- 3) NDVI + EVI para las 6 fechas
- 4) PCA (6 CP) sobre NDVI + EVI para las 6 fechas
- 5 a 10) NDVI para cada fecha
- 11 a 16) PCA para imágenes originales (10 canales) para cada fecha
- 17 a 22) Imagen original de cada fecha

C. Clasificación

En todos los casos se utilizó un modelo random forest (RF) con una profundidad máxima de 5 y utilizando 150 árboles. Además, para el conjunto de entrenamiento NDVI+EVI se utilizó también un modelo SVM con un kernel del tipo lineal.

- M1: RF NDVI para las 6 fechas
- M2: RF EVI para las 6 fechas
- M3: RF NDVI + EVI para las 6 fechas
- M4: RF PCA (6 CP) sobre NDVI + EVI para las 6 fechas
- M5: SVM NDVI + EVI para las 6 fechas
- M6 a M11: RF NDVI para cada fecha
- M12 a M17: RF PCA de imágenes originales para cada fecha
- M18 a M23: RF Imágen orginal de cada fecha

A partir de los resultados obtenidos, se realizaron diferentes ensambles:

- ME1: Modelos 1 a 4.
- ME2: Modelos 6 a 11 (NDVI por fecha).
- ME3: Modelos 12 a 17 (PCA por fecha).
- ME4: Modelos 18 a 23 (imágenes originales/fecha).
- ME5: Tres modelos con mayor índice Kappa de los modelos 18 a 23 (M20, M21 y M22).

Para los modelos M1 a M23 se realizó la predicción por píxel, a excepción del M3 donde también se realizó la estimación a través de la segmentación. Para los modelos de ensambles (ME1 a ME5) en todos los casos se estimó la superficie de cada cultivo por segmentación.

D. Cálculo de áreas

Para realizar la estimación de área de cada cultivo por departamento se utilizaron 2 técnicas. Por un lado, se estimó el área a partir de la suma de píxeles de cada cultivo por departamento. Para esto, se obtuvo una máscara de cada cultivo y departamento de la que se estimó la cantidad total de píxeles. Para obtener el área de cada cultivo en hectáreas (ha) el total de píxeles se multiplicó por un factor de 0.04.

Por otro lado, se realizó una segmentación de la imagen de NDVI de 6 fechas, utilizando un radio espacial de 5 píxeles, un rango espectral de 0.1 y se estableció el tamaño mínimo de los segmentos en 5 píxeles. A cada segmento (lote) se le asignó el cultivo utilizando la capa resultante del modelo 3 (RF NDVI+EVI). Para ello, se filtraron los segmentos menores a 1 ha y se estimó la moda (categoría mayoritaria) de los píxeles para cada segmento.

III. RESULTADOS

Para comparar los modelos se utilizaron como datos de referencias la cantidad de hectáreas cultivadas estimadas por la bolsa de cereales (Estimaciones 1) y por el Ministerio de Agroindustria, Ganadería y Pesca (Estimaciones 2) [11].

En la figura 5 se pueden observar los resultados de los modelos M1, M2, M3, M4 y el ensamble de estos cuatro (ME1).

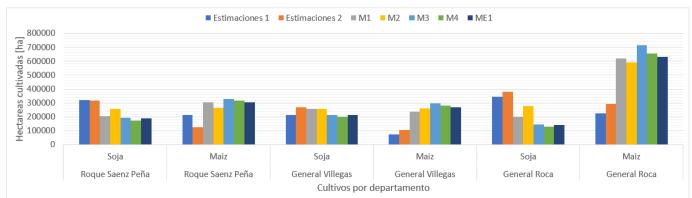


Fig. 5. Cantidad de hectáreas. Modelos generados a partir de los índices.

En la figura 6 se presentan los resultados de los modelos M3 (Random Forest) y M5 (SVM). Ambos modelos utilizan el mismo dataset: NDVI y EVI para los seis fechas. Nuevamente, la cantidad de hectáreas estimadas fue similar entre ambos modelos.

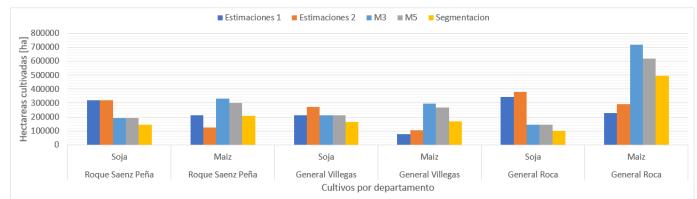


Fig. 6. Cantidad de hectáreas. Modelos Random Forest y Support Vector Machine.

Dado estos resultados obtenidos se optó por cambiar de estrategia. Se crearon modelos por fechas para tener en cuenta la variabilidad temporal de los cultivos. Además, en lugar de utilizar los índices NDVI y EVI, se utilizaron las imágenes originales y las bandas resultantes del PCA realizado sobre estas como datos de entrada.

En la figura 7 se muestran los resultados de los modelos M3, ME2, ME3, ME4 y ME5. En este caso, la superficie estimada entre cultivos y departamentos varió entre modelos. Se observa que los modelos ME3, ME4 y ME5 presentaron mejores resultados, particularmente para la soja en los departamentos de Gral. Villegas y Gral. Roca. Estos modelos corresponden a los ensambles realizados sobre modelos entrenados por fecha. Para profundizar la comparación entre estos modelos se presentan los valores de Kappa obtenidos en cada uno de los modelos individuales (Tabla III). Se observa que, en general, los resultados obtenidos fueron mejores cuando se utilizaron las imágenes obtenidas a partir del PCA para cada fecha (M12 a M17). Además, se observa que los mayores valores del índice Kappa se obtuvieron, en ambos casos, para el mes de Febrero ($K = 0.82$ y 0.8 , respectivamente).

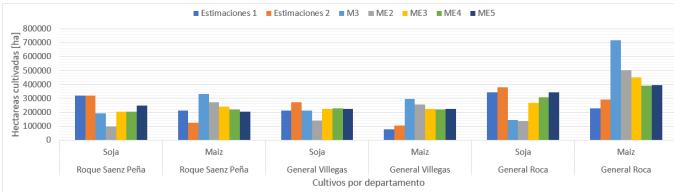


Fig. 7. Cantidad de hectáreas. Modelos de ensambles.

Nombre	Kappa
M12	0.51
M13	0.56
M14	0.64
M15	0.70
M16	0.82
M17	0.78
M18	0.49
M19	0.49
M20	0.61
M21	0.64
M22	0.80
M23	0.72

TABLE III

Por último, se presentan los resultados obtenidos a partir del ensamble ME5 (ensamble de modelos M20, M21 y M22) (Tabla IV, figuras 8 y 9).

Deptó	Cultivo	Hectáreas
Roque Saenz Peña	Soja	246374
Roque Saenz Peña	Maiz	202524
General Villegas	Soja	223180
General Villegas	Maiz	222332
General Roca	Soja	345438
General Roca	Maiz	395252

TABLE IV

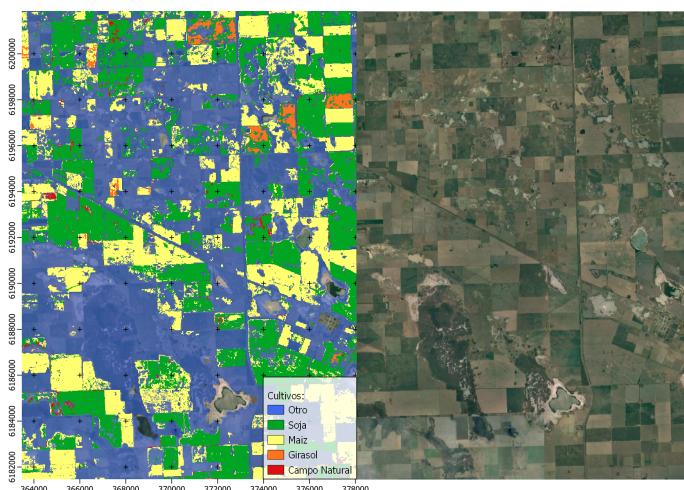


Fig. 8. Predicciones.

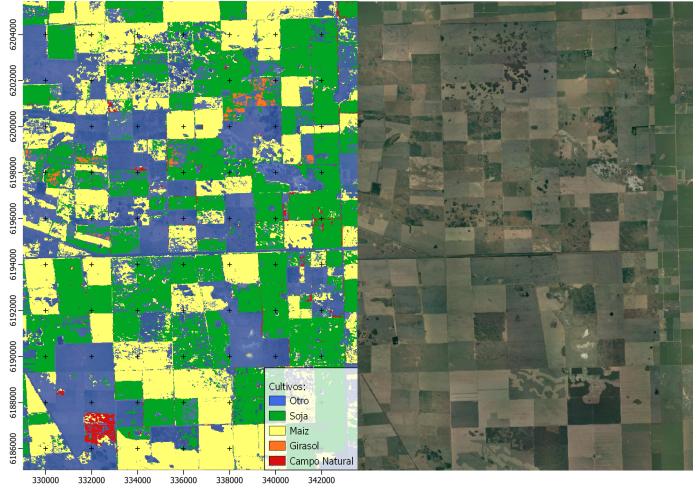


Fig. 9. Predicciones.

IV. DISCUSIÓN Y CONCLUSIONES

Para este trabajo se utilizaron diferentes técnicas y modelos para realizar la predicción de hectáreas sembradas con maíz y soja para 3 departamentos de la provincia de Córdoba y Buenos Aires de Argentina.

En general, cuando se entrenaron los modelos utilizando todas las fechas, se observó que los modelos tendieron a subestimar las hectáreas sembradas con soja respecto a lo informado por la bolsa de cereales del ministerio de Agroindustria, Ganadería y Pesca. Por el contrario, se observó una sobre-estimación de las hectáreas sembradas con maíz. A su vez, no se observaron diferencias en los resultados obtenidos al utilizar distintos clasificadores como random forest o SVM. Por otro lado, se obtuvieron mejores resultados cuando se entrenaron modelos por fecha. En particular, los mayores valores de Kappa se observaron en el mes de febrero y utilizando segmentación y ensambles de modelos.

REFERENCES

- [1] J. Li, Z. Zhang, X. Jin, J. Chen, S. Zhang, Z. He, S. Li, Z. He, H. Zhang, and H. Xiao, "Exploring the socioeconomic and ecological consequences of cash crop cultivation for policy implications," *Land Use Policy*, vol. 76, pp. 46–57, July 2018.
- [2] J. Brinkhoff, J. Vardanega, and A. Robson, "Land cover classification of nine perennial crops using sentinel-1 and 2 data," *Remote Sensing*, vol. 12, p. 96, 12 2019.
- [3] A. Dervisoglu, B. Bilgilioğlu, and N. Yağmur, "Comparison of pixel-based and object-based classification methods in determination of wetland coastline," *International Journal of Environment and Geoinformatics*, vol. 6, pp. 327–332, 12 2019.
- [4] E. O. Box, B. N. Holben, and V. Kalb, "Accuracy of the avhrr vegetation index as a predictor of biomass, primary productivity and net co 2 flux," *Vegetatio*, vol. 80, no. 2, pp. 71–89, 1989.
- [5] I. C. Burke, T. G. Kittel, W. K. Lauenroth, P. Snook, C. Yonker, and W. Parton, "Regional analysis of the central great plains," *BioScience*, vol. 41, no. 10, pp. 685–692, 1991.
- [6] S. N. Goward, B. Markham, D. G. Dye, W. Dulaney, and J. Yang, "Normalized difference vegetation index measurements from the advanced very high resolution radiometer," *Remote Sensing of Environment*, vol. 35, pp. 257–277, Feb. 1991.
- [7] S. D. Prince, "Satellite remote sensing of primary production: comparison of results for sahelian grasslands 1981–1988," *International Journal of remote sensing*, vol. 12, no. 6, pp. 1301–1311, 1991.

- [8] M. S. Moran and J. Irons, "New imaging sensor technologies suitable for agricultural management," *Aspects of Applied Biology*, vol. 60, pp. 1–10, 2000.
- [9] J. Paruelo, J. Guerschman, G. Piñeiro, E. Jobbág, Verón, G. Baldi, and S. Baeza, "Cambios en el uso de la tierra en argentina y uruguay: marcos conceptuales para su análisis," *Agrociencia*, vol. 10, pp. 47–61, 01 2006.
- [10] P. Santos and A. J. Negri, "A comparison of the normalized difference vegetation index and rainfall for the amazon and northeastern brazil," *Journal of Applied Meteorology*, vol. 36, pp. 958–965, July 1997.
- [11] MAGyP, "bolsa de cereales del ministerio de agroindustria, ganadería y pesca."