DIGITAL AGRI

MASTER EN TRANSFORMACIÓN DIGITAL DEL SECTOR AGROALIMENTARIO Y FORESTAL

TRABAJO FIN DE MÁSTER

Clasificación supervisada de la cobertura del suelo en fincas ganaderas mediante el uso de imágenes Sentinel-2 y técnicas de inteligencia artificial

Alumno: **Diego Varona Renuncio**

Directores:

Francisco Javier Mesas Carrascosa

Fernando Pérez Porras

Fecha: Córdoba, 20 de septiembre de 2022

Índice de Figuras

Índice de Tablas

Clasificación supervisada de la cobertura del suelo en fincas ganaderas mediante el uso de imágenes Sentinel-2 y técnicas de inteligencia artificial

Alumno: Diego Varona Renuncio

Directores: Francisco Javier Mesas Carrascosa

Fernando Pérez Porras

Resumen

Los mapas de cobertura del suelo son de gran utilidad para la monitorización y manejo de los ecosistemas. El objetivo general de este Trabajo Final de Máster (TFM) es el desarrollo de modelos supervisados de inteligencia artificial alimentados con imágenes de Sentinel-2 para la clasificación de coberturas del suelo en fincas ganaderas. Se han obtenido las imágenes disponibles de un año completo en las fincas del estudio y se ha propuesto un procedimiento basado en el cálculo de variables estadísticas de las series temporales que incluyen las diferentes bandas e índices espectrales con una resolución de 10 m. Posteriormente, se ha realizado una selección de variables para la alimentación de los diferentes modelos predictivos. La precisión de los modelos obtenidos, representada por el coeficiente de correlación de Matthews (MCC), ha sido superior a 0,87 en todos los casos, incluyendo los modelos individuales por finca y el modelo conjunto. Por otro lado, el modelo XGBoost se plantea como la alternativa más adecuada para la implementación a larga escala de estos modelos, ya que permite reducir significativamente el número de variables a calcular, manteniendo una precisión similar al resto de los modelos. Una de las posibles mejoras de este procedimiento podría ser la inclusión de variables agroclimáticas adicionales al modelo o la creación de diferentes modelos en función de las características del entorno.

Palabras clave: datos remotos, cobertura del suelo, inteligencia artificial, ganadería extensiva.

1 Introducción

Los pastizales son los ecosistemas más extendidos a nivel mundial, ocupando más de un 30% de la masa terrestre del planeta (Latham et al., 2014). Estos ecosistemas tienen un rol fundamental en la regulación del ciclo de carbono (Scurlock & Hall, 1998), el mantenimiento de la biodiversidad (Tilman & Downing, 1994) y la provisión de alimentos de alta calidad. Además, desde un punto de vista agronómico, los pastizales aportan al ganado extensivo el alimento más económico, que es el pasto.

En este contexto, actualmente existe una gran demanda de mapas de usos o coberturas de suelo para la monitorización y el manejo de los ecosistemas a nivel global, regional o local. Para la generación de estos mapas, una de las fuentes de información más empleada son las imágenes de satélite dada su amplia cobertura geográfica y temporal. Sin embargo, la resolución espacio-temporal y el tiempo de procesamiento de éstas plantea todavía numerosos retos (Talukdar et al., 2020). A escala global son numerosos los programas de observación de la Tierra operando en la actualidad, destacando en los últimos años el programa Copernicus de la Agencia Espacial Europea (ESA), ofreciendo distintas misiones para la monitorización de tierra, océanos y atmósfera. En el caso de la monitorización de la cobertura del suelo, Copernicus obtiene datos mediante las misiones Sentinel-1 (radar) y Sentinel-2 (multiespectral), cada una con 2 satélites hoy día, ofreciendo datos con una resolución espacial de diez metros cada cinco días.

Las imágenes multiespectrales de Sentinel-2 pueden ser de gran utilidad para la monitorización de pastizales, obteniendo información relacionada con la productividad de los ecosistemas y la biodiversidad de especies (Ali et al, 2016; Schwieder et al., 2020). Sin embargo, en muchas ocasiones, las fincas ganaderas contienen múltiples coberturas del suelo, como pasto, arbolado o zonas improductivas, que deben ser previamente identificadas para que el usuario pueda obtener información exclusivamente de las zonas de interés para así optimizar la toma de decisiones. Para ello, en las últimas décadas se ha desarrollado un creciente interés en los métodos de clasificación automática de imágenes, especialmente en aplicaciones relacionadas con la observación de la tierra tanto empleando sensores pasivos como activos. Estos métodos permiten la explotación de la información espacial y espectral tanto de una única imagen como de un conjunto

de ellas a través de series temporales (Rußwurm & Körner, 2018; Baamonde et al, 2019). Ante este escenario de trabajo, las técnicas basadas en inteligencia artificial están permitiendo simplificar el proceso de análisis y extracción de la gran cantidad de datos contenidos en estas imágenes para realizar la clasificación. Recientemente, la aplicación de modelos de "machine learning" sobre imágenes satelitales para el mapeo de coberturas del suelo ha captado una gran atención. Estas técnicas están divididas en técnicas supervisadas o no supervisadas, dependiendo de si requieren un conjunto de muestras previamente etiquetadas. Los modelos supervisados incluyen algoritmos como las máquinas de vector de soporte (SVM), árboles de decisión (DT), random forest (RF) o redes neuronales (NN), mientras que los modelos no supervisados incluyen algunos algoritmos como la clusterización K-medias o la clusterización por propagación de afinidades (Maxwell et al., 2018).

El objetivo general de este Trabajo Fin de Máster es la aplicación de un modelo de inteligencia artificial alimentado con datos multitemporales de Sentinel-2 para la clasificación supervisada de las diferentes coberturas del suelo presentes en fincas ganaderas.

2 Materiales y métodos

2.1 Zona de estudio

Los datos utilizados en este estudio corresponden al período discurrido entre septiembre de 2020 y septiembre de 2021. La zona de estudio comprende cuatro fincas situadas en diferentes zonas de la geografía española (Figura 1): Cabañas (Riofrío, Ávila), Fuente del Perro (Pedroches, Córdoba), Vall d'en Bas (La Vall d'en Bas, Gerona) y Llodio (Llodio, Álava).

Figura 1 Localización de las parcelas de estudio.

Debido a la dispersión geográfica, cada parcela se encuentra en un contexto climático diferente. En la Tabla 1 se indica tanto la superficie de cada una de las fincas ganaderas como los valores climatológicos más representativos durante el período de estudio. La estructura de la vegetación de estas fincas es muy diversa. La finca "Cabañas" destaca por la ausencia de arbolado y la presencia de pasto verde durante los meses desde marzo hasta junio (primavera) y desde octubre hasta diciembre (otoño). La finca "Fuente del Perro" está dominada por el ecosistema de la dehesa, un sistema agrosilvopastoral caracterizado por la presencia de arbolado disperso y un estrato inferior de pasto y matorrales. En este caso, dadas las condiciones climáticas, la presencia de pasto verde durante el año es más limitada, especialmente en los meses de verano. Si bien las fincas "Llodio" y "Vall d'en Bas" se encuentran en entornos agroclimáticos diferentes, la

estructura de la vegetación es similar, ya que están dominadas por arbolado concentrado en bosques y presencia de pasto verde durante prácticamente todo el año.

Tabla 1 Superficie y variables meteorológicas recogidas en el año 2021 en las fincas de estudio.

Finca	Superficie [Has]	Temperatura media (ºC)	Temperatura mínima (ºC)	Temperatura máxima (ºC)	Precipitación acumulada
Cabañas	139	12,1	-10,7	38,8	381,0
Fuente del Perro	207	16,5	-2,3	44,6	374,4
Vall d'en Bas	707	12,9	-8,7	38,0	948,2
Llodio	1007	14,9	-1,9	33,4	1292,8

2.2 Metodología

En la Figura 2, se muestra esquemáticamente el flujo de trabajo diseñado para alcanzar los objetivos del estudio. En primer lugar, se ha llevado a cabo una fase de preprocesamiento de los datos, cuyo objetivo es preparar los datos obtenidos de forma que estos contengan el formato adecuado para alimentar los modelos predictivos. Posteriormente, en la fase de modelado o procesamiento de los datos, se han aplicado diversos algoritmos de inteligencia artificial para la clasificación de coberturas del suelo y se ha evaluado su precisión mediante métricas avanzadas. Se ha utilizado el software QGIS 3.22.3 (QGIS Development Team, 2022) para el procesamiento de la información geográfica y el software Python (Python Software Foundation, 2019) para el preprocesamiento de datos y el desarrollo de los modelos predictivos.

2.2.1 Adquisición de datos

El programa Copernicus de la Agencia Espacial Europea (ESA) tiene por objeto agrupar fuentes de datos obtenidos por sensores embarcados en plataformas espaciales y terrestres para proporcionar una visión global del «estado de salud» de la Tierra. Dentro del programa Copernicus, la misión Sentinel-2 está destinada a la observación terrestre e incluye 2 satélites idénticos, Sentinel-2A y Sentinel-2B, ofreciendo datos a modo de imágenes multiespectrales con 13 bandas abarcando el espectro visible, infrarrojo cercano e infrarrojo de onda corta, con una resolución temporal de 5 días y una

resolución espacial de 10 m, 20 m y 60 (https://sentinel.esa.int/web/sentinel/missions/sentinel-2). Además, las imágenes Sentinel-2 incluyen una banda de clasificación de escenas (Scene Classification Layer, SCL) con una resolución espacial de 20 m consistente en un mapa de clasificación de píxeles que incluye 4 clases diferentes de nubes y otras 6 clases que incluyen: sombra, sombra de nubes, vegetación, suelo, agua y nieve.

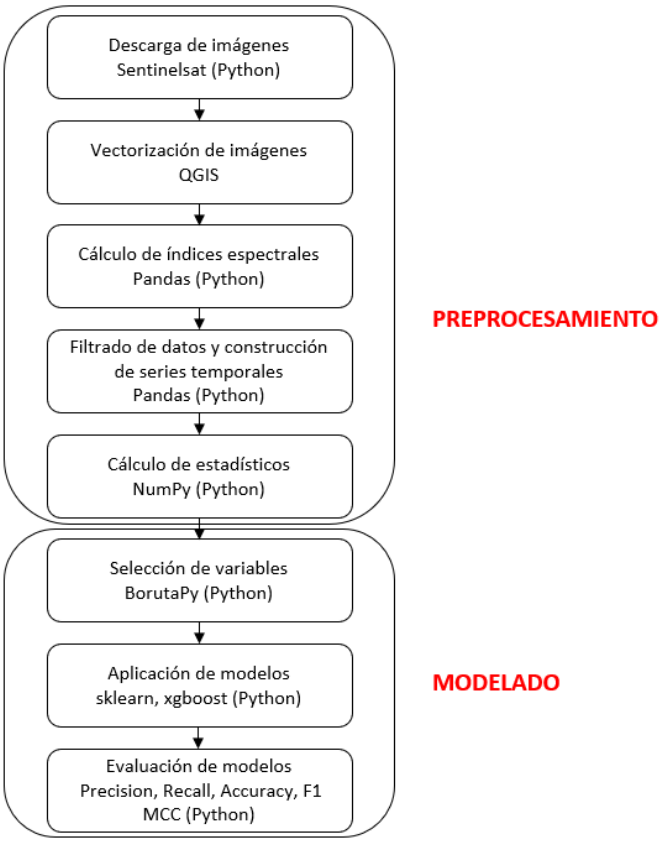

Figura 2 Figura esquemática representando el flujo de trabajo para el preprocesamiento y modelado de los datos.

Dados los objetivos de este trabajo, se han utilizado únicamente las bandas de cada imagen con una resolución espacial de 10 m correspondientes con 3 bandas del espectro visible: azul (B2), verde (B3) y rojo (B4); además de una banda en el infrarrojo cercano (B8). Además, se ha utilizado también la banda SCL de cada escena con una resolución de 20 m, con el objetivo de realizar un filtrado de aquellos píxeles cuyos datos no son válidos, ya sea por presencia de nubes, sombras, etc. Dichas imágenes se encuentran disponibles en abierto y han sido obtenidas a través del servicio Copernicus Open Access Hub (https://scihub.copernicus.eu/) y la librería de Python Sentinelsat

(https://sentinelsat.readthedocs.io/en/stable/). Las imágenes se descargaron para ser procesadas localmente, siendo de interés aquellas que cubrían alguna de las cuatro fincas objeto del estudio durante el período transcurrido entre septiembre de 2020 y septiembre de 2021, previamente filtradas a partir de los metadatos, descartando aquellas con una proporción de nubes superior al 30%. La Tabla 2 recoge el total de imágenes empleadas en este trabajo para cada una de las fincas.

Tabla 2 Número de imágenes Sentinel-2 con una proporción de nubes inferior al 30% en cada una de las fincas.

Finca	Número de imágenes
Cabañas	38
Fuente del Perro	80
Vall d'en Bas	55
Llodio	30

2.2.2 Preprocesamiento de los datos

En primer lugar, se utilizó el software QGIS para extraer la información de las imágenes ráster y construir las series temporales. Para ello, se realizó una vectorización de cada banda de cada una de las imágenes de la serie temporal, creando una capa vectorial en forma de cuadrícula coincidente con la malla de píxeles de Sentinel-2. Cada capa vectorial contó con un identificador único que incluía el día de obtención de la imagen y el tipo de banda. Posteriormente, estas capas vectoriales fueron transformadas en un archivo en formato CSV para su posterior manipulación como conjunto de datos (dataset) utilizando el lenguaje de programación Python versión 3.8. A continuación, se calcularon los índices espectrales NDVI, GNDVI, EVI, GRVI y NDWI (Tabla 3):

Tabla 3 Índices espectrales.

Índice	Formulación
Índice de vegetación de diferencia normalizada (NDVI)	$NDVI = \frac{B8 - B4}{B8 + B4}$
Índice de vegetación de diferencia normalizada verde (GNDVI)	$GNDVI = \frac{B8 - B3}{B8 + B3}$

Índice de vegetación mejorado (EVI)	$EVI = 2{,}5 \times \frac{B8 - B4}{B8 + 6 \times B4 - 7{,}5 \times B2 + 1}$
Índice de vegetación rojo-verde (GRVI)	$GRVI = \frac{B3 - B4}{B3 + B4}$
Índice de agua de diferencia normalizada (NDWI)	$NDWI = \frac{B3 - B8}{B3 + B4}$

A continuación, se eliminaron aquellos píxeles cuyo valor correspondiente a la banda SLC no correspondía con alguna de las siguientes clases: vegetación, sin vegetación o agua. De esta forma, se eliminó la posible influencia de las nubes para la posterior clasificación. Finalmente, se agruparon todos los datos de bandas e índices de vegetación obtenidas durante el período de estudio correspondientes a cada píxel para construir las series temporales (Figura 3).

id	Date	B02	B03	B04	B08	NDVI	GNDVI	EVI	GRVI	NDWI
1	2020-09-03	722.0	1038.0	1506.0	2542.0	0.255929	0.420112	0.420182	-0.183962	-0.591195
1	2020-09-08	952.0	1278.0	1840.0	2872.0	0.219015	0.384096	0.380924	-0.180244	-0.511225
1	2020-09-18	457.0	673.0	1040.0	1960.0	0.306667	0.488796	0.481827	-0.214244	-0.751313
1	2020-09-28	630.0	882.0	1328.0	2288.0	0.265487	0.443533	0.433839	-0.201810	-0.636199
1	2020-10-03	344.0	495.0	653.0	1236.0	0.308629	0.428076	0.566019	-0.137631	-0.645470
1	2020-11-17	634.0	902.0	1256.0	2250.0	0.283514	0.427665	0.493839	-0.164041	-0.624652
1	2020-11-22	688.0	1014.0	1426.0	2388.0	0.252229	0.403880	0.415730	-0.168852	-0.563115
1	2020-12-12	499.0	737.0	928.0	1856.0	0.333333	0.431546	0.630007	-0.114715	-0.672072
1	2021-01-06	452.0	607.0	979.0	2066.0	0.356979	0.545829	0.597122	-0.234552	-0.919924
1	2021-03-07	731.0	962.0	1392.0	2362.0	0.258391	0.421179	0.463450	-0.182668	-0.594732
1	2021-03-17	943.0	1302.0	2010.0	3284.0	0.240650	0.432185	0.385011	-0.213768	-0.598430
1	2021-03-22	871.0	1282.0	2004.0	3195.0	0.229083	0.427295	0.342734	-0.219720	-0.582167
1	2021-04-06	962.0	1348.0	2076.0	3422.0	0.244816	0.434801	0.388389	-0.212617	-0.605724
1	2021-04-16	1094.0	1496.0	2256.0	3768.0	0.250996	0.431611	0.415385	-0.202559	-0.605544
1	2021-05-06	762.0	1064.0	1636.0	2758.0	0.255348	0.443223	0.408892	-0.211852	-0.627407
1	2021-05-31	883.0	1256.0	1916.0	3000.0	0.220504	0.409774	0.344149	-0.208071	-0.549811
1	2021-06-10	974.0	1418.0	2098.0	3306.0	0.223538	0.399661	0.351572	-0.193402	-0.536974
1	2021-06-15	1106.0	1458.0	2200.0	3514.0	0.229961	0.413516	0.390143	-0.202843	-0.562056
1	2021-06-25	1024.0	1378.0	1920.0	3018.0	0.222357	0.373066	0.400204	-0.164342	-0.497271
1	2021-06-30	926.0	1318.0	1940.0	3042.0	0.221196	0.395413	0.356035	-0.190915	-0.529159
1	2021-07-05	924.0	1274.0	1926.0	2952.0	0.210332	0.397066	0.338435	-0.203750	-0.524375
1	2021-07-10	943.0	1332.0	1958.0	3042.0	0.216800	0.390947	0.351104	-0.190274	-0.519757
1	2021-07-15	1072.0	1436.0	2150.0	3266.0	0.206056	0.389196	0.343300	-0.199108	-0.510318
1	2021-07-20	1042.0	1544.0	2260.0	3448.0	0.208129	0.381410	0.323037	-0.188223	-0.500526
1	2021-07-25	1208.0	1632.0	2430.0	3602.0	0.194297	0.376385	0.321166	-0.196455	-0.484983
1	2021-07-30	1074.0	1490.0	2242.0	3274.0	0.187092	0.374475	0.297509	-0.201501	-0.478028
1	2021-08-09	935.0	1368.0	2004.0	3098.0	0.214426	0.387371	0.337217	-0.188612	-0.513049
1	2021-08-19	936.0	1298.0	1926.0	2944.0	0.209035	0.388025	0.340195	-0.194789	-0.510546
1	2021-08-24	1018.0	1340.0	1986.0	2998.0	0.203050	0.382204	0.347527	-0.194227	-0.498497
1	2021-08-29	938.0	1290.0	1922.0	2908.0	0.204141	0.385422	0.332838	-0.196762	-0.503736

Figura 3 Serie temporal con las 4 bandas de Sentinel-2 y los 5 índices espectrales derivados.

Después, se procedió al cálculo de unas variables estadísticas (Tabla 4) para cada una de estas series temporales. Estos estadísticos se obtuvieron para cada una de las bandas

(B2, B3, B4 y B8) y cada uno de los índices espectrales (NDVI, GNDVI, EVI, GRVI y NDWI), generando 407 variables por cada píxel.

Tabla 4 Variables estadísticas calculadas para cada serie temporal de bandas o índices espectrales (IE).

Variable	Definición
mean	Media
median	Mediana
std	Desviación estándar
var	Varianza
max	Máximo
min	Mínimo
mean_diff	Diferencia media entre 2 bandas/IE
std_diff	Desviación estándar de la diferencia entre 2 bandas/IE
mad_{}	Media de las desviaciones absolutas respecto a un punto {media, mediana y moda}
p2p	Amplitud entre picos de la serie temporal
amp	Diferencia entre la máxima amplitud y la media
s2e	Valor desde el inicio hasta el final de la serie
ZCR	Número de veces que cambia de signo la serie temporal
MCR	Número de veces que la serie temporal cruza el valor medio
line_cross	Número de veces que 2 series temporales se cruzan
norm	Normal vectorial de la serie temporal
skew	Asimetría de la distribución de probabilidad
kurt	Medida de valores atípicos
perc_{}	Percentil {25%, 50% y 75%}
IQR	Rango intercuartílico
corr	Correlación de Pearson entre 2 series temporales
cov	Covarianza entre 2 series temporales
ecc_{}	Vector de aceleración entre 2 series temporales

2.2.3 Procesamiento de los datos - Modelado

2.2.3.1 Selección de variables

El número de variables disponibles para la introducción en los modelos fue de 407, sin embargo, muchas de estas tendrán una importancia residual en el desarrollo posterior de los modelos predictivos y pueden aumentar el ruido y el tiempo de computación, resultando importante realizar una selección de variables previa al desarrollo de los mismos. Actualmente existen numerosos algoritmos para la selección de variables, sin embargo, muchos de ellos requieren de la determinación de un valor umbral de importancia, lo que convierte la selección de variables en un proceso relativamente arbitrario. Sin embargo, otros algoritmos como Boruta, no necesitan de esta selección. Boruta es un algoritmo envolvente de selección de variables automático, capaz de trabajar con cualquier método de clasificación que produzca una medida de importancia variable, como son los modelos RandomForest, XGBoost o Extra Trees Classifier. Este algoritmo realiza una búsqueda descendente de variables relevantes comparando la importancia de las variables originales con la importancia alcanzable al azar, estimada mediante sus copias permutadas y eliminando progresivamente las variables irrelevantes para estabilizar esa prueba. La aplicación de este algoritmo se realizó mediante la librería BorutaPy (https://github.com/scikit-learn-contrib/boruta_py).

2.2.3.2 Modelos predictivos

Existen una gran variedad de modelos de inteligencia artificial o "machine learning" para tareas de clasificación tanto supervisada como no supervisada. Desde modelos sencillos como la regresión logística, k-vecinos más cercanos o árboles de decisión, hasta modelos más complejos como las redes neuronales o las máquinas de vectores de soporte. En este trabajo, se evaluaron los modelos Random Forest (RF), XGBoost (XGB) y Extra Trees Classifier (ETC), implementados con la ayuda de las librerías sklearn (https://scikit-learn.org/stable/) y xgboost (https://xgboost.readthedocs.io/en/stable/).

Al trabajar con métodos de clasificación supervisada es necesario contar con un conjunto de datos para el entrenamiento. Utilizando el software QGIS, se realizó una fotointerpretación de la ortofotografía aérea de máxima actualidad del plan PNOA, identificando 3 clases (arbolado, pasto e improductivo) en un determinado número de

píxeles de cada finca (Tabla 5). Posteriormente, el conjunto de etiquetas generado se dividió en 2 grupos, uno destinado al entrenamiento de los modelos (75%), y el resto a su validación a partir de una muestra externa (25%). Como puede observarse, el número de etiquetas por clase está desbalanceado, lo que puede afectar a la precisión de las predicciones, sin embargo, se han aplicado modelos no paramétricos y métricas de evaluación que disminuyen el efecto de la presencia de clases desbalanceadas.

Tabla 5 Número de etiquetas obtenidas en cada una de las fincas ganaderas.

Finca	Etiquetas arbolado	Etiquetas pasto	Etiquetas improductivo	Etiquetas totales
Cabañas	101	655	244	1000
Fuente del Perro	480	390	130	1000
Vall d'en Bas	741	172	87	1000
Llodio	664	235	131	1000

2.2.3.3 Evaluación de los modelos

Existen numerosas métricas de evaluación de modelos predictivos de clasificación. A continuación, se muestra las métricas utilizadas para la evaluación (Tabla 6).

Tabla 6 Definición de las métricas de evaluación de los modelos.

Métrica	Formulación
Precision	$Precision = \frac{TP}{TP + FP}$
Accuracy	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Recall	$Recall = \frac{TP}{TP + FN}$
F1 score	$F1 = 2 \times \frac{precision \times recall}{precision + recall}$
TP: número de verdaderos positivos - TN: número de verdaderos negativos. FP: número de falsos negativos - FN: número de falsos negativos.	

Dado que las muestras de las 3 clases introducidas no están balanceadas, ciertas métricas de evaluación, como "accuracy" y "F1-score" pueden no estar evaluando los modelos de una forma adecuada porque no son capaces de considerar la ratio entre los elementos positivos y negativos (Boughorbel et al., 2017). Con coeficiente de correlación de Matthews (MCC), sin embargo, para obtener una evaluación positiva con el coeficiente de correlación de Matthews, el modelo debe realizar predicciones correctas tanto en la mayoría de los casos negativos, como en la mayoría de los casos positivos, independientemente de las ratios en el conjunto de datos (Chicco & Jurman, 2020). En clasificaciones con más de 2 clases, el rango del coeficiente de correlación de Matthews varía entre un valor mínimo entre -1 y 0, dependiendo de la distribución de los datos y un valor máximo de +1.

El coeficiente de correlación de Matthews (MCC) se define mediante la siguiente ecuación:

$$MCC = \frac{cs - \vec{t} \cdot \vec{p}}{\sqrt{s^2 - \vec{p} \cdot \vec{p}}\sqrt{s^2 - \vec{t} \cdot \vec{t}}}$$

Donde:

t: número de veces que la clase ha sido predicha correctamente.

p: número de veces que la clase ha sido predicha.

s: número total de muestras.

p: número de muestras predichas correctamente.

3 Resultados

En este apartado se presentan los resultados obtenidos por los diferentes modelos aplicados para la clasificación de coberturas del suelo en fincas ganaderas. Por un lado, se han desarrollado modelos individuales para cada una de las fincas ganaderas, y, por otro lado, se ha desarrollado un modelo conjunto para todas las fincas ganaderas, con el objetivo de analizar la capacidad de generalización de estos.

3.1.1 Modelos individuales para cada finca ganadera

3.1.1.1 Selección de variables

Una vez aplicado el algoritmo Boruta con cada uno de los modelos de base mencionados en el apartado de modelos predictivos, se han obtenido automáticamente el número de variables seleccionadas por cada modelo. En la Figura 4, puede observarse que el modelo XGBoost selecciona un número significativamente menor de variables, con una media de 24 variables selecciónadas, respecto a los modelos Random Forest, con una media 125 variables seleccionadas, y Extra Trees Classifier, con una media de 260 variables seleccionadas. Esta observación es consistente con los resultados obtenidos en todas las fincas incluidas en el estudio y sugiere, que, a igualdad de precisión de los diferentes modelos, el modelo XGBoost sería el más adecuado para su implementación a larga escala.

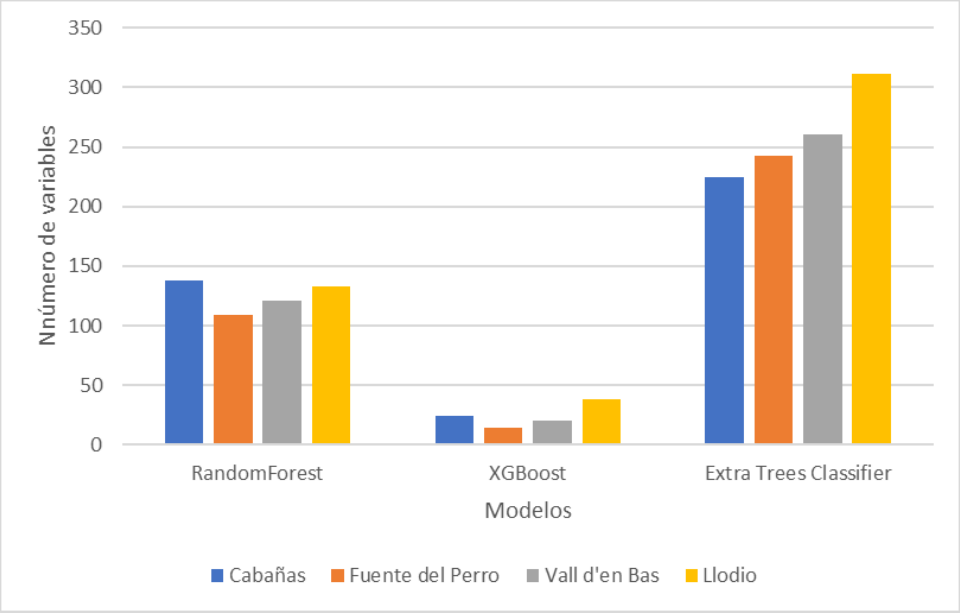

Figura 4 Número de variables seleccionadas por el algoritmo Boruta por tipo de modelo y finca ganadera.

Respecto a la tipología de variables seleccionadas por el algoritmo Boruta, no se ha observado ningún patrón relevante que pudiera indicar la importancia de una banda, índice espectral o estadístico para la clasificación de coberturas del suelo. De hecho, en los cuatro modelos XGBoost (uno para cada finca) ninguna de las variables calculadas se repite como variable seleccionada en los 4 modelos. Esto puede deberse a la heterogeneidad de las zonas de estudio seleccionada.

3.1.1.2 Evaluación de los modelos

Todos los modelos aplicados han obtenido una evaluación positiva, siendo el coeficiente de correlación de Matthews obtenido en las 4 fincas superior a 0,9, excepto en el caso del modelo XGBoost en la finca "Fuente del Perro", cuyo valor es de 0,87. Además, el coeficiente de correlación de Matthews es inferior en los 3 modelos desarrollados en la finca "Fuente del Perro". Estos resultados son consistentes con los aportados por las métricas adicionales (Tabla 7), donde se puede observar que "precision", "recall" y "F1" presentan valores inferiores con el modelo XGBoost, especialmente en la clase "improductivo", lo que se debe a que el número de muestras por clase está desbalanceado (improductivo: 130, arbolado: 480 y pasto: 390). Aunque este desequilibrio es natural en fincas ganaderas, debido a la escasa presencia de edificios, carreteras u otros elementos no productivos, una de las soluciones sería generar datos sintéticos de la clase "improductivo" para equilibrar los datos. Los modelos desarrollados en "Vall d'en Bas" y "Llodio" muestran una gran precisión independientemente del tipo de modelo utilizado. En estos casos, la presencia de arbolado denso en forma de bosque y zonas aisladas de pasto pueden facilitar la clasificación de los píxeles. Sin embargo, la precisión de los modelos desarrollados en las 4 fincas sigue siendo aceptable, como puede observarse en la Figura 5. Estos resultados sugieren que la clasificación supervisada de píxeles de una única finca puede alcanzar una gran precisión, aunque existen varias desventajas. Por un lado, un modelo individualizado presentará un alto sobreajuste y su capacidad de generalización será limitiada. Por otro lado, el etiquetado manual de píxeles requiere de una gran cantidad de tiempo y lo convierte en un proceso no escalable con una mayor cantidad de fincas y superficies. Esto justifica el desarrollo de un modelo general para la clasificación de píxeles de las 4 fincas ganaderas.

Tabla 7 Resumen de métricas (P: precision, A: accuracy, R: recall, F1) para cada una de las clases y modelos predictivos (RF: RandomForest, XGB: XGBoost, ETC: Extra Trees Classifier) en la finca "Fuente del Perro".

	Arbolado				Pasto				Improductivo			
	P	A	R	F1	P	A	R	F1	P	A	R	F1
RF	0,98	0,97	0,96	0,97	0,90	0,95	0,97	0,93	0,90	0,96	0,75	0,82
XGB	0,97	0,97	0,96	0,96	0,87	0,93	0,96	0,91	0,84	0,94	0,67	0,75
ETC	0,98	0,98	0,99	0,98	0,94	0,96	0,97	0,95	0,90	0,96	0,75	0,82

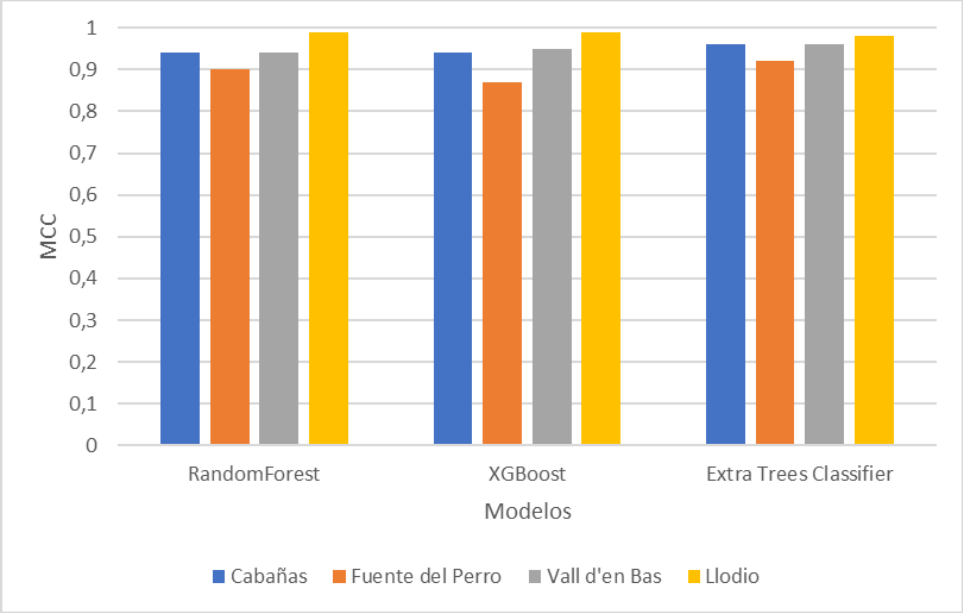

Figura 5 Coeficiente de correlación de Matthews (MCC) por tipo de modelo y finca ganadera.

3.1.2 Modelo general para las 4 fincas ganaderas

3.1.2.1 Selección de variables

Una vez aplicado el algoritmo Boruta con cada uno de los modelos de base mencionados en el apartado de modelos predictivos, se han obtenido automáticamente el número de variables seleccionada. En la Tabla 6, puede observarse el mismo comportamiento respecto a los modelos individuales, donde el modelo XGBoost selecciona un número menor de variables, lo que sugiere que es más eficiente a la hora de reducir la información necesaria manteniendo una alta precisión.

Tabla 8 Número de variables seleccionadas por tipo de modelo.

RandomForest	XGBoost	Extra Trees Classifier
281	33	362

3.1.2.2 Evaluación de los modelos

Los resultados mostraron que la precisión del modelo general para la clasificación supervisada de cobertura del suelo es muy alta independientemente del tipo de modelo seleccionado (Tabla 7). A igualdad de precisión, el modelo más adecuado para su implementación a gran escala sería el modelo XGBoost, ya que permitió reducir el número de variables estadísticas a calcular en un 91% respecto a Extra Trees Classifier y un 88% respecto a RandomForest.

Tabla 9 Coeficiente de correlación de Matthews por tipo de modelo.

RandomForest	XGBoost	Extra Trees Classifier
0.94	0.95	0.95

En la Tabla 8 se muestra un resumen de las métricas de evaluación del modelo general para cada uno de los modelos predictivos. La clase mayoritaria (arbolado), que cuenta con un número total de 1.986 etiquetas, cuenta con la evaluación más positiva, lo que puede deberse a este hecho. La clase "pasto" cuenta con 1.452 etiquetas. La clase "improductivo" cuenta con un número significativamente menor de etiquetas, con 592. Si bien el "recall", que indica la capacidad para el modelo para detectar muestras positivas, presenta valores inferiores a 0,90, la evaluación de los modelos sigue siendo muy positiva.

Tabla 10 Resumen de métricas (P: precision, A: accuracy, R: recall, F1) para cada una de las clases y modelos predictivos (RF: RandomForest, XGB: XGBoost, ETC: Extra Trees Classifier) en los modelos generales.

	Arbolado				Pasto				Improductivo			
	P	A	R	F1	P	A	R	F1	P	A	R	F1
RF	0,98	0,98	0,98	0,98	0,94	0,97	0,98	0,96	0,98	0,98	0,89	0,93
XGB	0,99	0,98	0,98	0,98	0,94	0,97	0,98	0,96	0,98	0,95	0,90	0,94
ETC	0,98	0,98	0,99	0,98	0,95	0,98	0,98	0,96	0,99	0,98	0,88	0,93

Finalmente, en las Figuras 6, 7, 8 y 9, se muestra una representación visual de las predicciones con el modelo XGBoost en las parcelas "Cabañas" y "Vall d'en Bas". El recuadro amarillo encuadra la zona que está ampliada en la imagen posterior. Además de la evaluación positiva de las métricas utilizadas, la evaluación visual de las predicciones también es positiva.

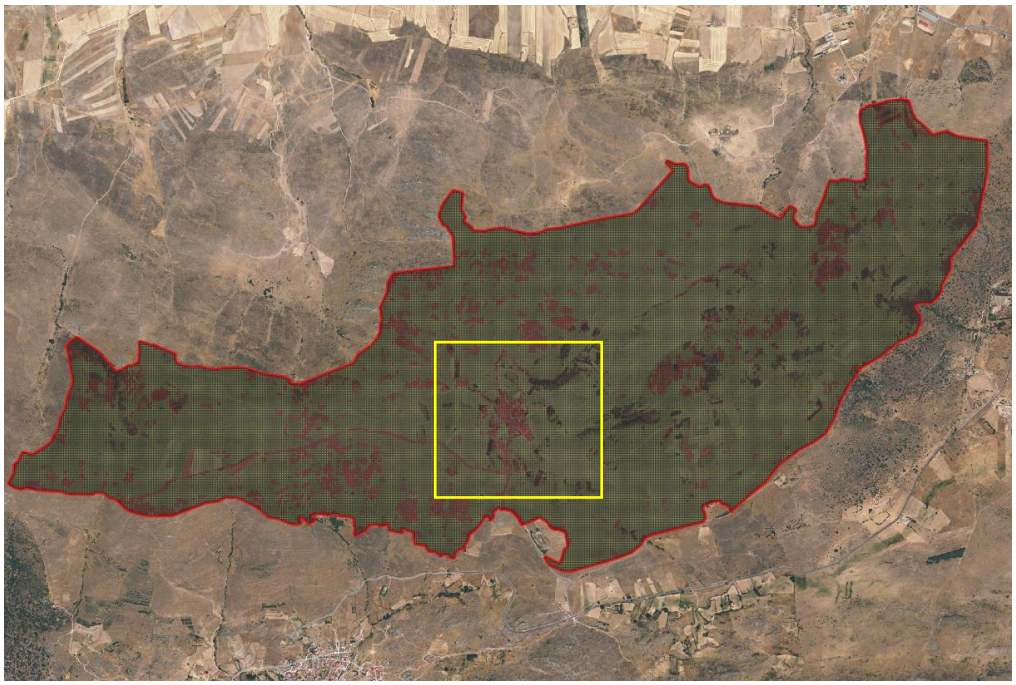

Figura 6 Representación visual de la predicción del modelo general en la finca "Cabañas" con XGBoost (rojo: improductivo, verde: pasto, marrón: arbolado).

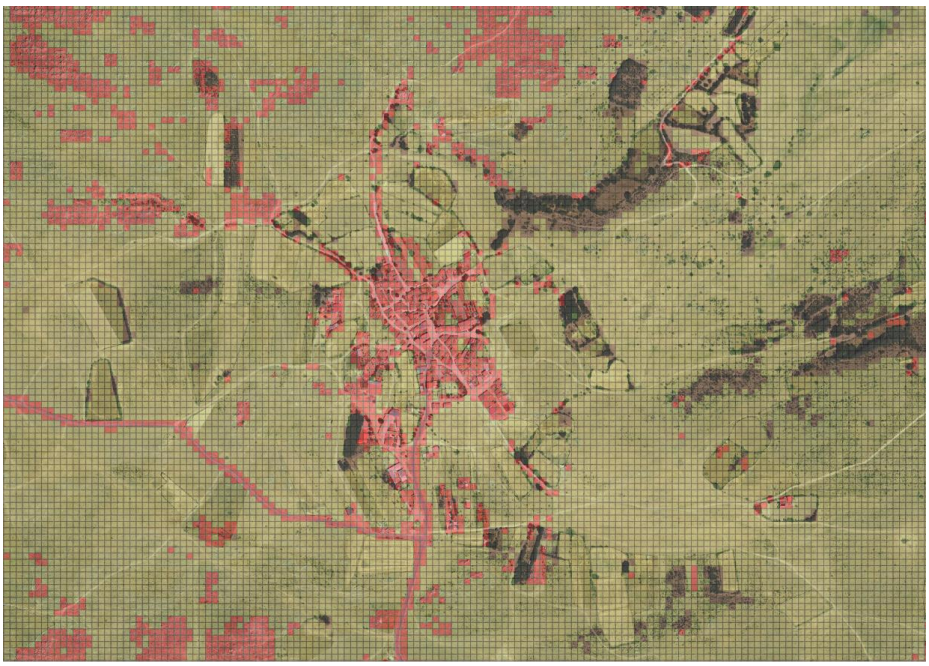

Figura 7 Representación visual de la predicción del modelo general en la finca "Cabañas" con XGBoost en un mayor nivel de detalle (rojo: improductivo, verde: pasto, marrón: arbolado).

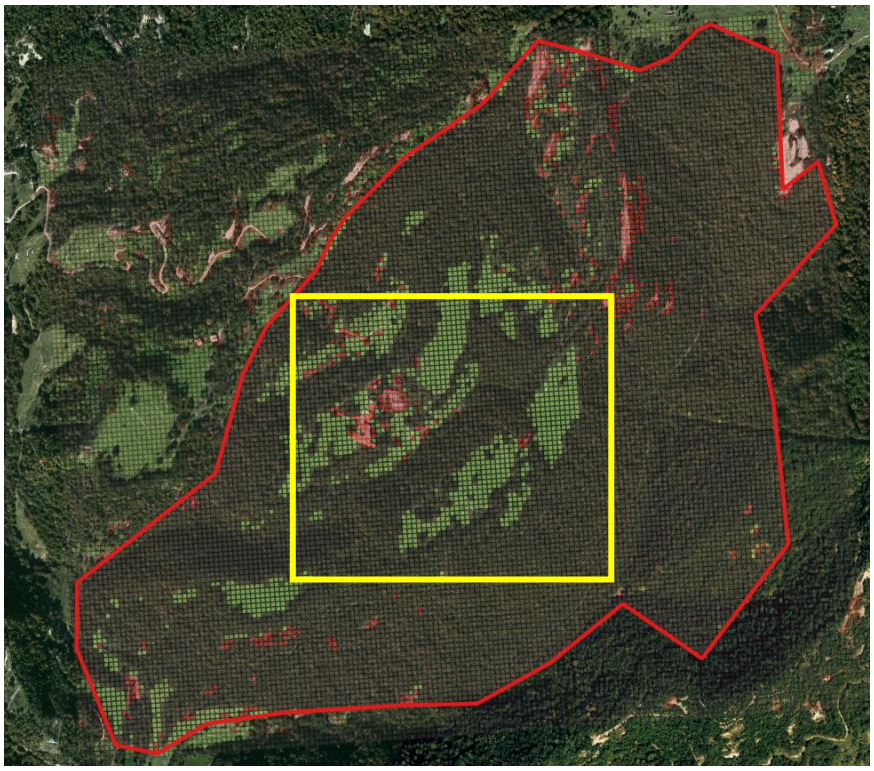

Figura 8 Representación visual de la predicción del modelo general en la finca "Vall d'en Bas" con XGBoost (rojo: improductivo, verde: pasto, marrón: arbolado).

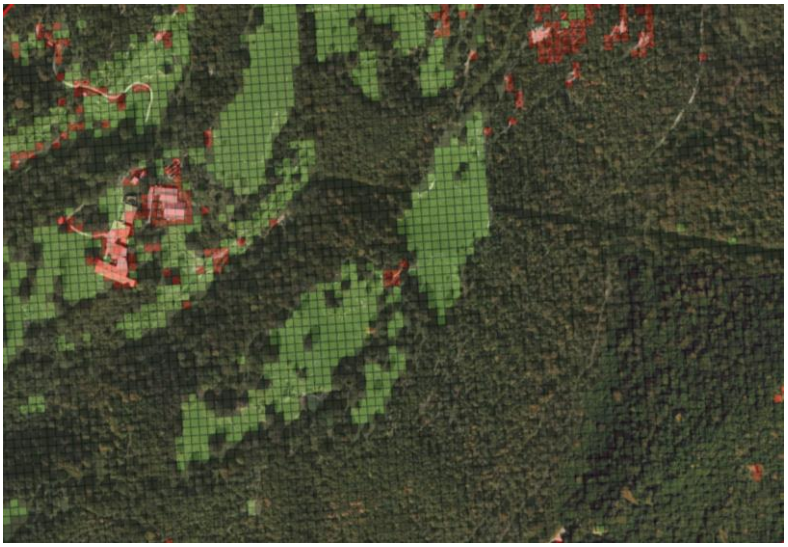

Figura 9 Representación visual de la predicción del modelo general en la finca "Vall d'en Bas" con XGBoost en un mayor nivel de detalle (rojo: improductivo, verde: pasto, marrón: arbolado).

4 CONCLUSIONES

Este Trabajo Fin de Máster ha demostrado que la utilización de datos multitemporales de Sentinel-2 junto con técnicas avanzadas de inteligencia artificial pueden ayudar a predecir de forma automática la clasificación de coberturas del suelo en fincas ganaderas con una gran precisión. Si bien existe una importante variabilidad en la disponibilidad de imágenes Sentinel-2 con un porcentaje de nubes inferior al 30% en las diferentes fincas (Tabla 2), se ha observado que los resultados de la clasificación han sido positivos en todos los casos. La adicción de otras fuentes de datos, como datos meteorológicos o de otros sensores remotos, como Sentinel-1, podrían ser de utilidad a la hora de implementar este procedimiento a gran escala, una vez la variabilidad de los datos aumente, si bien algunos autores han observado que la inclusión de datos Sentinel-1 no mejora significativamente la precisión de los modelos (Conlon & Rustowicz, 2022). Respecto a los diferentes tipos de modelos testeados, existen mínimas diferencias en la precisión, sin embargo, el modelo XGBoost se posiciona como la mejor alternativa, ya que aplicando la técnica Boruta para la selección previa de variables, reduce significativamente el número de variables seleccionadas, lo que disminuye también el tiempo de computación de las mismas. El cálculo de variables estadísticas que describen las series temporales se presenta también como una gran ventaja, ya que permite la aplicación de modelos de inteligencia artificial más fáciles de implementar en comparación con aquellos modelos que toman directamente como entrada las series temporales en bruto. La validación externa de los modelos se ha realizado con píxeles que pertenecen a las fincas analizadas, por lo que muchos de esos píxeles pueden presentar una gran similitud, en términos espectrales, a aquellos que forman parte del grupo de entrenamiento, por lo que, en una fase posterior, sería importante aplicar el modelo desarrollado con los datos de una finca que no haya sido utilizada previamente en el desarrollo del modelo para evitar un posible sobreajuste del modelo de clasificación. Por último, el desequilibrio de los datos por clases, con un número significativamente inferior de datos de clase "improductivo", ha provocado una reducción de la precisión en la predicción de los modelos. Si bien se han planteado estrategias para limitar su influencia, como el uso de modelos predictivos no

paramétricos y métricas avanzadas, otra de las posibles soluciones podría ser la creación de datos sintéticos para equilibrar las muestras por clase.

5 REFERENCIAS

Ali, I., Cawkwell, F., Dwyer, E., Barrett, B., & Green, S. (2016). Satellite remote sensing of grasslands: from observation to management. Journal of Plant Ecology, 9(6), 649-671.

Baamonde, S., Cabana, M., Sillero, N., Penedo, M. G., Naveira, H., & Novo, J. (2019). Fully automatic multi-temporal land cover classification using Sentinel-2 image data. Procedia Computer Science, 159, 650-657.

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS one, 12(6), e0177678.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 108-122.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), 1-13.

Conlon, T. M., & Rustowicz, R. M. (2022). SEN12TS: Fusing Sentinel-1 Backscatter and InSAR Timeseries with Multispectral Sentinel-2 Imagery for Deep Learning.

Latham, J., Cumani, R., Rosati, I., & Bloise, M. (2014). Global Land Cover SHARE (GLC-SHARE): Database Beta-Release Version 1.0-2014.

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing, 39(9), 2784-2817.

Python Software Foundation. (2019). Python Language Reference, version 3.8. Available at https://www.python.org/

QGIS Development Team. (2022). QGIS Geographic Information System 3.22.3. QGIS Association website: https://www.qgis.org

Rußwurm, M. & Körner, M. (2018). Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders. ISPRS International Journal of Geo-Information, 7(4), 129.

Schwieder, M., Buddeberg, M., Kowalski, K., Pfoch, K., Bartsch, J., Bach, H., ... & Hostert, P. (2020). Estimating grassland parameters from Sentinel-2: A model comparison study. PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 88(5), 379-390.

Scurlock, J. M. O., & Hall, D. O. (1998). The global carbon sink: a grassland perspective. Global Change Biology, 4(2), 229-233.

Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y. A., & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations — A review. Remote Sensing, 12(7), 1135.

Tilman, D., & Downing, J. A. (1994). Biodiversity and stability in grasslands. Nature, 367(6461), 363-365.