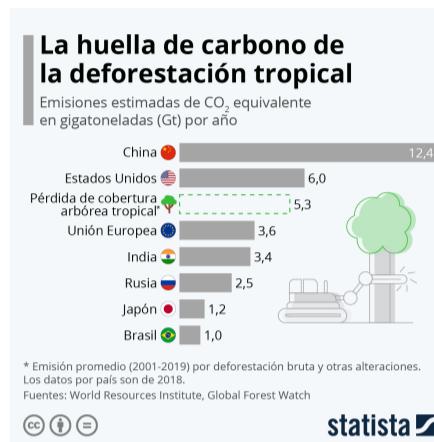


MEMORIA - EDA MONITORIZACIÓN DE LA DEFORESTACIÓN DE LOS BOSQUES AMAZÓNICOS. CASO DE ESTUDIO EN PERÚ

1. CONTEXTO

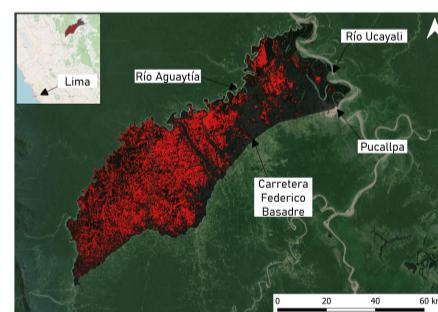
- La **deforestación**, además de la pérdida de biodiversidad y servicios ecosistémicos locales y regionales, supone **una de las principales fuentes de emisión de gases de efecto invernadero (GEI)**. Desde 1970, las emisiones acumuladas de CO_2 procedentes de la deforestación y cambio de usos del suelo han incrementado un 40%; y en 2010 ya suponían el 24% del total de las emisiones de GEI; más que las emisiones procedentes del sector industrial (21%) o del transporte (14%)
- **Si la deforestación tropical fuera un país, tendría la tercera mayor huella de carbono del mundo.** Según datos e imágenes de Global Forest Watch, analizados en una investigación recientemente publicada por la revista Nature, la pérdida de cobertura arbórea tropical provocó un promedio de emisiones anuales equivalentes a 5,3 gigatoneladas entre 2001 y 2019. Esta cifra la sitúa en tercer lugar, después de China y Estados Unidos, si se excluyen los efectos del cambio de uso del suelo y la silvicultura.



- En el contexto actual de cambio climático, y sin una reducción eficiente y sostenida de las emisiones de GEI, es muy importante contar con **sistemas automáticos de monitorización** que faciliten y fomenten la implementación de **programas para la conservación de bosques**, como el programa REED+ de las Naciones Unidas o los proyectos de captura y absorción de CO_2 en bosques (**carbon farming**). Estos sistemas deben ser transparentes, sólidos, robustos y fiables para evitar prácticas de **green washing** que cuestionan la implantación de este tipo de soluciones climáticas. En este sentido, los **sistemas basados en datos de teledetección (imágenes satelitales ópticas, radas, LiDAR, etc) y modelos de inteligencia artificial** para la identificación de patrones de pérdida o ganancia de biomasa forestal, y estimación de balances netos de CO_2 , se consideran una herramienta con gran potencial (**Monitoring of forests through remote sensing**).

CASO DE ESTUDIO:

- **Perú contiene el 16% de los bosques amazónicos del mundo.** Dentro del país, una de las regiones más afectadas por la deforestación es el departamento de Ucayali. En esta región el Gobierno de Perú a reportado una **pérdida de más de 540.000 ha en el periodo 2001 - 2021**. Se ha seleccionado un área de estudio en esta región, en la provincia de Padre Abad, en la que la expansión de las actividades ganaderas y agroindustriales han provocado importantes pérdidas de cobertura vegetal. En la siguiente imagen se muestra el área de estudio. En color rojo se resaltan las áreas afectadas por la deforestación para el periodo 2001-2021 según datos del **Gobierno de Perú**.



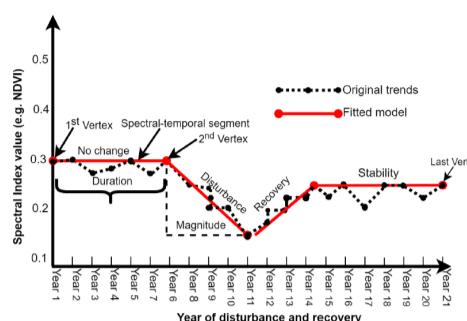
2. HIPÓTESIS

1. Patrón de deforestación: "La deforestación en el caso de estudio sigue el mismo patrón de deforestación de los bosques amazónicos a nivel regional y global"
2. Identificación robusta: "Los modelos LandTrendr y CCDC permiten identificar y caracterizar de manera precisa y robusta la pérdida de cobertura vegetal en el área de estudio."
3. MRV y algoritmos: "Los algoritmos LandTrendr y CCDC tienen el potencial de constituir la base para sistemas de monitoreo, reporte y verificación (MRV) de emisiones y captura de CO_2 , debido a su precisión en la identificación de cambios."
4. Comparación de datos espectrales: "El uso de series temporales de imágenes procesadas con LandTrendr y CCDC mejora la identificación de patrones de deforestación en comparación con métodos tradicionales basados en análisis de imágenes individuales."

3. DATOS

RECOLECCIÓN DE DATOS

1. **Dataset con la información espectral:** Datos obtenidos al aplicar el algoritmo LandTrendr para la segmentación de una serie temporal de las imágenes de alta resolución Landsat (30m) desde el año 1985 al año 2022, en la plataforma *Google Earth Engine*. El modelo LandTrendr ajusta los valores espectrales de la serie temporal de imágenes en un modelo de segmentos lineales que capturan tanto los cambios bruscos de la señal espectral (*Spectral Index value*), como las tendencias graduales, descartando aquellos valores que son producto de ruido de fondo (nubosidad, ruido electrónico o digital del sensor, etc). En la siguiente imagen se muestra el modelo conceptual. Se puede consultar más información sobre el modelo y su implementación en GEE en este enlace: [LT-GEE Guide](#).



- Para el estudio de deforestación se ha seleccionado, para cada píxel, el segmento que representa pérdida de valor espectral (pendiente negativa) de mayor magnitud (mayor diferencia del valor espectral entre los vértices del segmento), denominado ***Loss Big-Delta**, entre los años 2001 y 2021. Es importante señalar que aunque no se haya producido un evento de deforestación en un píxel, si se puede obtener un segmento de pendiente negativa para ese píxel, con una pendiente suave y bajo valor de magnitud, relacionado con degradación gradual de la cobertura forestal, sin que haya una pérdida total de la cobertura. Igualmente esos segmentos pueden estar relacionados con la presencia de ruido que el modelo no ha sido capaz de distinguir y descartar.
- El resultado se ha exportado en un archivo imagen de formato Geotiff desde GEE, con 6 bandas. En cada banda se guardan los siguientes atributos que caracterizan el segmento:
 - Banda 1 - Año del cambio o perturbación (yod):** Año del cambio que representa el segmento.
 - Banda 2 - Magnitud:** Diferencia del valor espectral de los dos vértices del segmento.
 - Banda 3 - Duración:** Diferencia de los años de los dos vértices del segmento
 - Banda 4 - Valor espectral previo al cambio:** valor del índice espectral en el primer vértice del segmento (preval)
 - Banda 5 - Tasa:** Calculada como *magnitud/duración*
 - Banda 6 - ratio DSNR:** ratio que mide la proporción de señal espectral respecto del ruido de fondo
- El valor espectral sobre el que se ha realizado la segmentación ha sido el **índice TCW (Tasseled Cap Wetness)** que es muy sensible a los cambios en la cobertura vegetal. A continuación se muestra su fórmula de cálculo:

$$TCW = 0.0315 \cdot B + 0.2021 \cdot G + 0.3102 \cdot R + 0.1594 \cdot NIR - 0.6806 \cdot SWIR1 - 0.6109 \cdot SWIR2$$

NIR : banda infrarrojo cercano
 SWIR1 : banda infrarrojo de onda corta 1
 SWIR2 : banda infrarrojo de onda corta 2
 B : banda azul
 G : banda verde
 R : banda rojo

2. **Dataset de referencia:** Para identificar las áreas deforestadas de las no deforestadas se ha generado una máscara booleana, en formato Geotiff, para el periodo 2001-2021, a partir de los datos del producto denominado *Hansen Global Forest Change v1.11 (2000-2023)*, disponible en el catálogo de datos de GEE ([GFC](#)). Este producto, identifica a nivel global, las pérdidas y ganancias de cobertura forestal.

LIMPIEZA Y PREPARACIÓN DE DATOS

- Visualización y estadísticas básicas de la información espectral contenida en los archivos imagen con formato geotiff, y creación de un dataframe a partir de la información contenida en cada una de las bandas (un dataframe para la información espectral y otro para la información de referencia). Se ha creado una columna para la información contenida en cada una de las bandas
- Unión de los dos dataframes a partir de la información de localización de los píxeles en los ejes x e y. Establecer como índice del df las coordenadas x,y para cada registro. Cada registro corresponde con un píxel
- Comprobación de coincidencia de coordenadas geográficas de cada píxel en los dos dataframes, para validar que las dos imágenes de donde proceden los datos están co-registradas (alineadas geográficamente).
- Análisis de valores NaN: El % de NaN en el df resultante es muy alto porque corresponden con los píxeles de la imagen que están fuera del área de estudio (valores enmascarados) y que al exportar la imagen con los resultados desde GEE, como tipo float, se registran como valores NaN. Se eliminan porque corresponde con registros que están fuera del área de estudio. Después de eliminar los NaN se cuenta con más de millón y medio de registros para el análisis
- No se identifican valores duplicados
- Se renombran las columnas para claridad y consistencia
- Se normalizan, a valores entre 0 y 1, los valores de las variables magnitud, preval y tasa para que facilitar la comparabilidad entre variables
- Se convierten los valores de yod, duración y clasificación a tipo int
- Se crea una columna, tipo str, con la descripción de los valores de clasificación, para facilitar la comparación posterior entre grupos

- Finalmente se analiza la cardinalidad de las variables.

En la siguiente tabla se resumen las principales características de las variables del dataframe final que se usa como base del análisis.

TABLA DE VARIABLES:

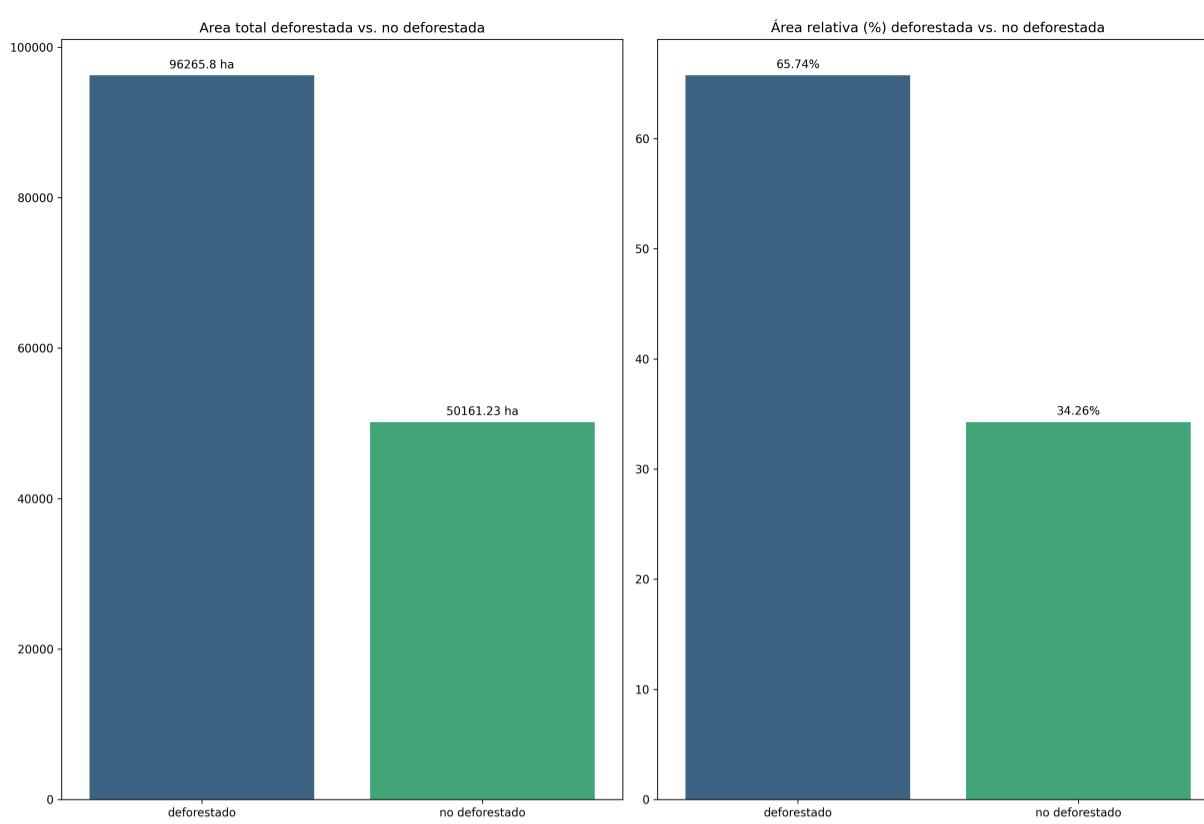
Columna/Variable	Descripción	Tipo_de_Variable	Rol en el EDA	Nota
x	Coordenada x de localización del píxel en la imagen	Numérica discreta	Define posición en x del píxel dentro de la imagen. Para que se pueda considerar como índice de los registros, habría que combinarla con los valores de y	
y	Coordenada y de localización del píxel en la imagen	Numérica discreta	Define posición en y del píxel dentro de la imagen. Para que se pueda considerar como índice de los registros, habría que combinarla con los valores de x	
long	Coordenada longitud de localización del píxel en la imagen	Numérica discreta	Define posición del píxel con coordenadas geográficas de longitud dentro del ROI	
lat	Coordenada latitud de localización del píxel en la imagen	Numérica discreta	Define posición del píxel con coordenadas geográficas de latitud dentro del ROI	
clasificacion	Variable con los datos de referencia que indica si ha habido deforestación en el píxel	Categórica/Binaria	Variable directora del análisis	valor 1 = no deforestación, valor 2 = deforestación
clasificacion_desc	Variable con los datos de referencia, donde se identifican las dos clases (deforestado y no deforestado)	Categórica/Binaria	Variable directora del análisis	facilita la presentación de resultados en análisis bivariantes y multivariantes
magnitud	Variable normalizada que representa la magnitud del cambio (negativo) en la señal espectral del píxel. La señal espectral se mide como el índice TCW que es sensible a los cambios en coberturas vegetales	Numérica continua	Esencial para identificar y caracterizar los cambios que podrían estar relacionados con eventos de deforestación	Normalizada entre 0 y 1
yod	Variable que indica el año en	Numérica discreta	Importante para identificar los	Convertida a tipo int

	el que se ha producido un cambio negativo en la señal espectral		patrones temporales de deforestación del ROI (área de estudio)	
duración	Variable que indica la duración del segmento que define al cambio negativo en la señal espectral	Numérica discreta	Esencial para identificar y caracterizar los cambios que podrían estar relacionados con eventos de deforestación (eventos de corta duración)	Convertida a tipo int
tasa	Variable normalizada que indica la tasa de cambio del segmento que se ajusta al cambio negativo en la señal espectral. Se calcula como magnitud/duración	Numérica continua	Importante para caracterizar los eventos de pérdida de cobertura vegetal y posibles causas de los mismos. Ej: Un evento de deforestación por incendio generalmente tiene tasas más altas que eventos de degradación de la cobertura vegetal graduales generados por tala ilegal	Normalizada entre 0 y 1
preval	Variable normalizada que representa el valor de la señal espectral (reflectividad) antes de producirse el cambio	Numérica continua	Indicador del verdor previo al evento de pérdida de cobertura vegetal. Permite conocer el estado de la cobertura vegetal previa al evento	Normalizada entre 0 y 1
ratio DSNR	Variable que representa la proporción de señal espectral frente al ruido espectral presente en la señal	Numérica continua	Importante para caracterizar la fiabilidad del modelo de segmentación. Valores bajos indican que la segmentación está fuertemente influenciada por el ruido espectral vs. cambios de señal espectral relacionados con eventos de pérdida de cobertura vegetal	

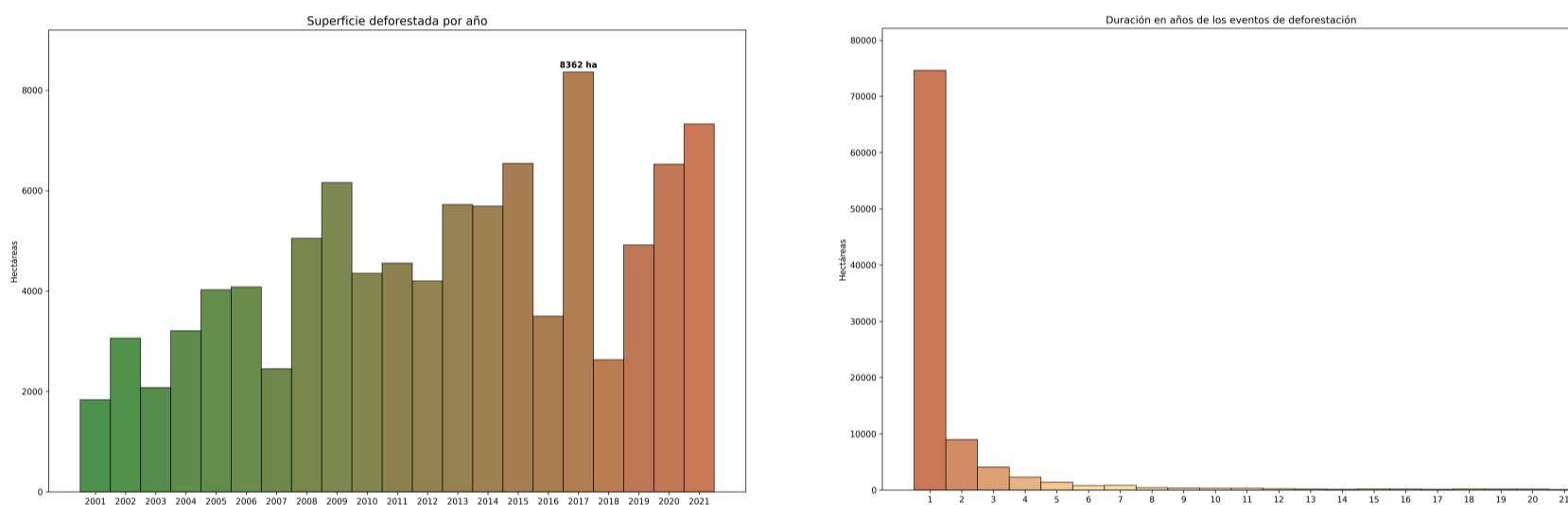
4. ANÁLISIS REALIZADO Y PRINCIPALES HALLAZGOS

Análisis Univariante

- Más del 65% del área de estudio ha sido deforestada en el periodo 2001-2021. La superficie deforestada es de casi 100.000 ha. Esta superficie equivale aproximadamente a 134,800 campos de fútbol, o la mitad de la superficie de la isla de Tenerife. La superficie deforestada equivale a más del 18% de la superficie deforestada en el departamento de Ucayali.



- La mayor parte de los eventos de pérdida de cobertura vegetal tienen una duración de 1 año, y la mayor superficie deforestada se ha identificado en el año 2017, con una pérdida de casi 8.400 ha. A pesar de que en el año 2018 parece que se produjo una ralentización en la deforestación del área de estudio, se volvió a reactivar a partir del año 2019, con un crecimiento continuado desde ese año hasta el final del periodo analizado



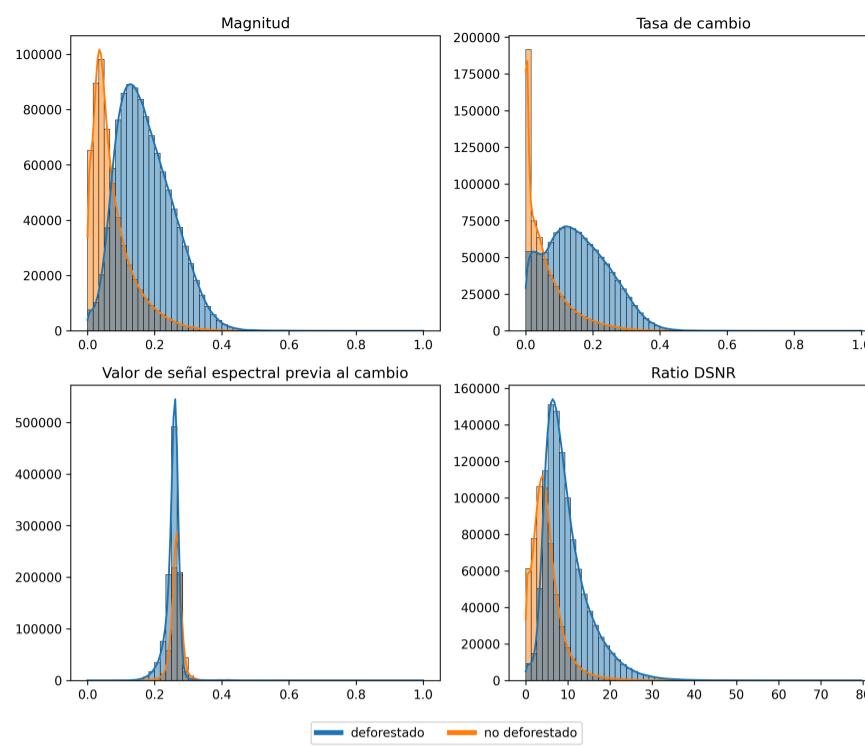
- Distribución de las variables numéricas continuas:** Las variables numéricas como `magnitud`, `tasa` y `dsnr` muestran una alta variabilidad

Variable	std	mean	CV
magitud	0.08	0.17	63.94
tasa	0.09	0.15	79.28
preval	0.02	0.25	8.32
dsnr	5.48	9.78	66.46

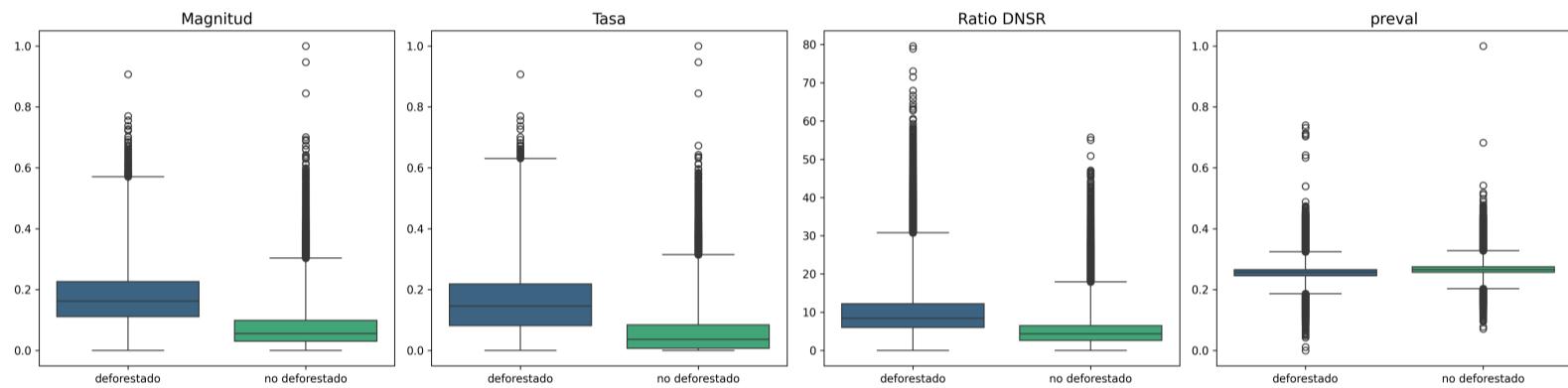
Análisis Bivariante

- Comparación entre Deforestado y No Deforestado:**

- Las variables `magnitud`, `tasa`, `dsnr` muestran diferencias significativas entre las áreas deforestadas y no deforestadas. Por el contrario la variable `preval` muestra un gran solapamiento entre las distribuciones de áreas deforestadas y no deforestadas. Las áreas deforestadas tienden a tener valores más altos en `magnitud`, `tasa`, `dsnr`. Los resultados de la **prueba U de Mann-Whitney** confirman diferencias estadísticamente significativas para estas variables



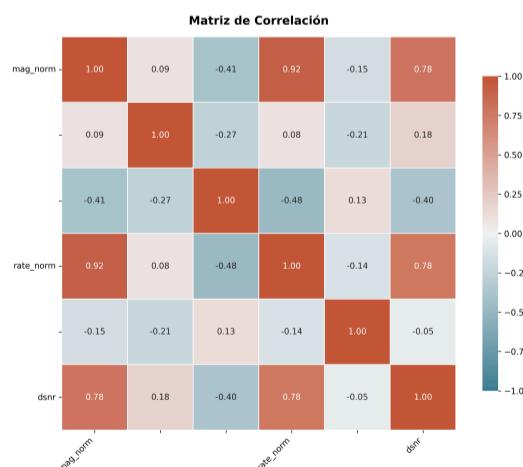
- La variabilidad de `magnitud`, `tasa` y `dsnr` se reduce significativamente cuando se consideran únicamente las áreas deforestadas. No obstante siguen presentando alta presencia de **outliers** que podrían estar relacionados con eventos extremos de deforestación. Los outliers presentes en la distribución de estas variables para las áreas clasificadas como no deforestadas podrían estar apuntando a posibles errores de clasificación (errores de omisión: píxeles donde ha habido deforestación pero que no han sido identificados en el dataset de referencia).



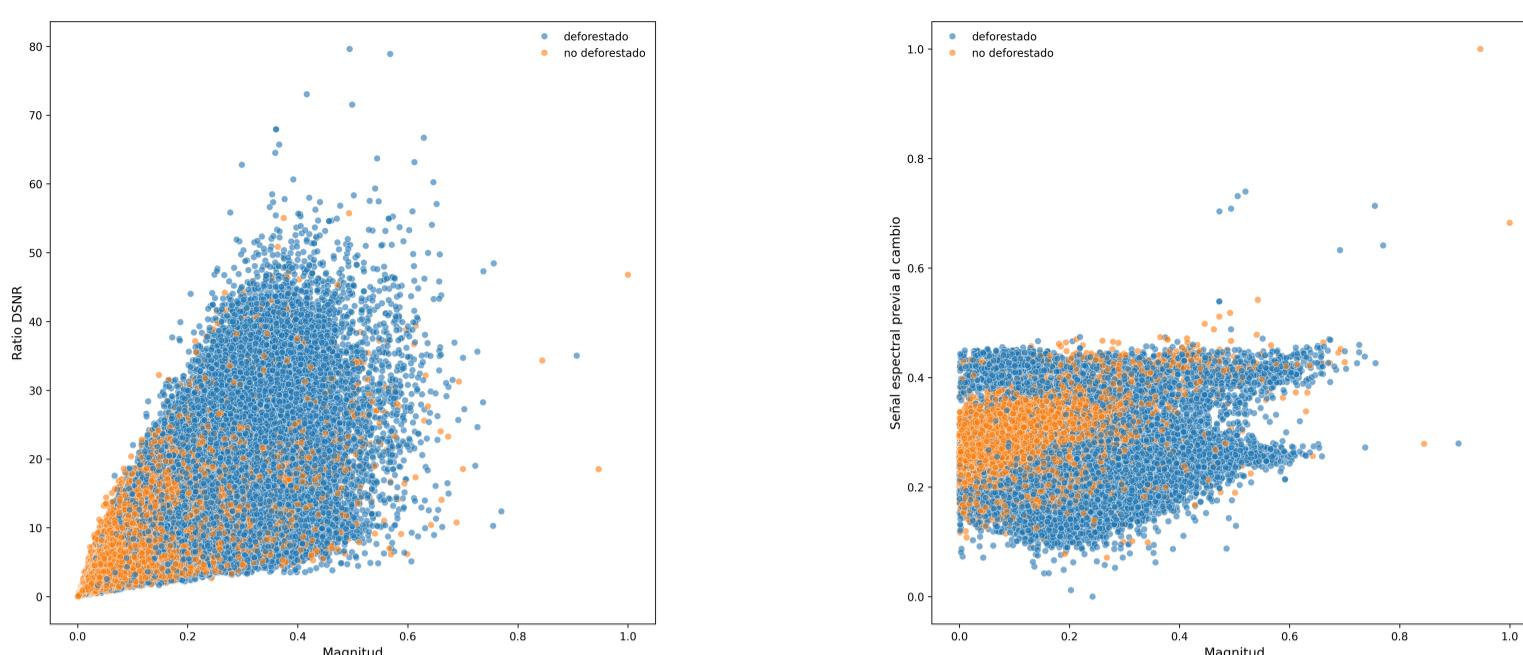
- La alta presencia de **outliers*** para la variable `preval`, tanto de valores mínimos como máximos, indican una alta fragmentación del bosque amazónico en el área de estudio, y presencia de un mosaico de coberturas del suelo con reductos de bosque amazónico, suelo desnudo, pastizales y cultivos en diferentes fases fenológicas. Los reductos de bosque amazónico son los que presentan valores de reflectividad extremos altos, y las áreas modificadas, valores de reflectividad extremos bajos. Valores altos en áreas deforestadas apuntan a posibles errores de comisión (píxeles clasificados como deforestados pero que no han sido deforestados). Por el contrario, valores extremos bajos en áreas no deforestadas indican potenciales errores de omisión.

Análisis Multivariante

- Correlación:** Existe una alta correlación entre `magnitud` y `tasa`, debido a que la variable `tasa` se deriva de `magnitud`. También hay una correlación significativa con la variable `dsnr`.



- Patrones de Agrupación:** Se identifican patrones claros de agrupación cuando los datos se representan en el espacio bidimensional formado por `magnitud` y `dsnr`, así como `magnitud` y `preval`. Sin embargo, las fronteras de decisión son complejas.



5. CONCLUSIONES Y RECOMENDACIONES BASADAS EN LOS HALLAZGOS

1. La deforestación en el área de estudio sigue un patrón consistente con el observado en otras regiones amazónicas y a nivel global. Más del 65% del área ha sido deforestada en el periodo 2001-2021, con un claro pico de deforestación en 2017. Este comportamiento sugiere que los factores de presión sobre la cobertura vegetal (como actividades agroindustriales y ganaderas) están en línea con las tendencias observadas en la región amazónica. Por tanto, el área es adecuada para continuar con el análisis temporal de deforestación y el desarrollo de modelos de identificación de eventos de pérdida de cobertura vegetal, así como para la predicción de balances de biomasa y CO_2 asociados a dichas pérdidas y ganancias. Además, el Gobierno de Perú está llevando a cabo estudios de seguimiento que podrían ser una valiosa fuente de datos de campo para el entrenamiento de estos modelos.
2. La aplicación del modelo LandTrendr sobre series temporales proporciona una ventaja significativa en comparación con enfoques tradicionales basados en imágenes individuales. Este modelo permite identificar y caracterizar, de manera precisa y robusta, los eventos de pérdida de cobertura vegetal. Las métricas derivadas, como la magnitud, tasa, duración y ratio DS/NR, presentan diferencias significativas entre áreas deforestadas y no deforestadas, validando su capacidad para capturar los cambios espectrales asociados a la deforestación. Esto posiciona a LandTrendr como una base sólida para la creación de sistemas de MRV y sistemas de alerta temprana para la pérdida de cobertura vegetal.
3. Sin embargo, se observa una importante dispersión de los valores y grandes zonas de solapamiento en las distribuciones de estas variables. Estas zonas de confusión pueden complicar la definición de fronteras de decisión durante el entrenamiento del algoritmo de clasificación y dificultar la predicción precisa de variables como biomasa o toneladas de CO_2 . Para reducir estas zonas de confusión, se recomienda:
 - Considerar transformaciones logarítmicas para incrementar la separación entre clases (deforestado y no deforestado).
 - Evaluar el impacto de los outliers durante el entrenamiento, analizando la mejora del rendimiento del modelo al eliminar valores extremos por encima de un determinado umbral.
 - Definir clases o grados de pérdida de cobertura vegetal a partir de la variable magnitud, lo que podría ayudar a refinar la clasificación.
4. La variable preval muestra un importante solapamiento entre las clases, por lo que se recomienda no incluirla en los análisis. En su lugar, se podría utilizar información de la señal espectral posterior al evento de cambio, lo que podría añadir valor predictivo adicional. Sería útil evaluar si la incorporación de esta nueva variable incrementa la separabilidad de las clases o patrones en un espacio multidimensional.
5. Es importante evitar incluir variables con alta correlación con la variable magnitud, como la tasa, para minimizar problemas de multicolinealidad durante el desarrollo del modelo. Esto ayudará a garantizar una mayor estabilidad y precisión en las predicciones.
6. Se recomienda incorporar datos derivados de otros modelos de segmentación de series temporales, como CCDC, para analizar la coherencia en la caracterización espectro-temporal de las áreas clasificadas como deforestadas y no deforestadas. Esto permitirá una comparación más robusta entre modelos y una mayor confiabilidad en los resultados obtenidos.