Clasificación de imágenes satelitales para la detección y cuantificación de inundaciones

Gustavo Dejean
Sebastián Steyerer
Nicolás Subotovsky
Fernando Véliz

Profesores: Mg. Federico Bayle - Dra. Carolina Ramos
Agosto de 2020

ÍNDICE

RESUMEN

El objetivo del presente trabajo fue calcular el área inundada a enero de 2019 en la provincia de Santa Fe - Argentina.

Se analizaron 6 rasters de 4 bandas cada uno, para cada una de las fechas de enero de 2018, noviembre de 2018 y enero 2020, junto con las capas de aguas continentales y la capa de evidencias de campo.

Se propuso realizar una clasificación en dos categorías agua - no agua, utilizando varios métodos que varían según los índices utilizados, los algoritmos empleados, la forma de generar la muestra de entrenamiento, el test realizado, la utilización o no de suavizado y el empleo de ensambles de modelos. Se usaron modelos predictivos como el Random Forest, Ranger, Árboles de Decisión, Bayes, KNN y otros modelos basados en matemática de raster.

Para mejorar la clasificación se usaron varios índices incorporandose el NDWI, NDVI, RVI y el NDTI. Asimismo se incorporaron las variaciones entre períodos de los dos primeros. Estos atributos serán utilizados como sustitutos o complementos de las variables originales de acuerdo a cada experimento.

Evaluada la superficie inundada resultante de cada modelo, se llegó a determinar una rango de superficie inundada de entre 818.567 y 2.337.118 hectáreas. Puede destacarse un rango muy amplio de hectáreas predichas de acuerdo a la metodología utilizada. Para la determinación final de las hectáreas inundadas se calculó el promedio ponderado de las hectáreas predichas por los modelos de acuerdo al kappa que arrojaron. El guarismo asciende a 1.806.624.

Se determinó un ranking de variables importantes para la clasificación. En el apartado final se discuten varias ventajas y desventajas de los modelos creados.

Palabras claves: raster , GIS, modelos predictivos, inundación, índices climáticos.

INTRODUCCIÓN

En enero de 2019, en el nordeste de Argentina, se produjo una inundación. Los departamentos más afectados fueron San Martín, Belgrano, 9 de Julio y Vera, de la provincia de Santa Fe, donde desde mediados de diciembre de 2018, cuando se iniciaron las lluvias, se acumularon unos 600 milímetros, a los que se sumaron las masas de agua que bajaron desde el sur de Santiago del Estero y Chaco, provincias que también fueron afectadas por inundaciones.

Las localidades más afectadas fueron: El Nochero, Santa Margarita, Villa Minetti y San Bernardo. Una zona delimitada de la ruta 291 para el norte, entre la 35 y la 95, y algo más hacía el este.

El presente estudio se realizó sobre 18 rasters del satélite Sentinel 2 que cubren el área bajo análisis. Se cuenta con información de 4 bandas (B2, B3, B4 y B8). En total cubren una superficie de 7.218.038 hs. de la provincia de Santa Fe, Argentina.

El detalle de cada raster es:

Raster	Superficie en ha
s2-20190100000000000-0000000000.tif	2.250.216
s2-20190100000000000-0000016384.tif	2.301.521
s2-20190100000000000-0000032768.tif	619.056
s2-20190100000016384-0000000000.tif	899.622
s2-20190100000016384-0000016384.tif	923.635
s2-20190100000016384-0000032768.tif	223.988

Tabla [1]: Distribución de hectáreas por raster.

Para el análisis propuesto, la información de los rasters fue complementada con la capa de aguas continentales provistas por el Instituto Geográfico Nacional (IGN[1]) y por la capa de evidencias de campo. La superficie perteneciente a la capa de aguas continentales es de: 969.225 ha y la pertenecientes a evidencias de campo es de 509.194,5 ha. Ambas capas se usaron para el entrenamiento de los modelos. En la figura 1, se muestra la zona de estudio y las tres capas mencionadas.

[1] https://www.ign.gob.ar/

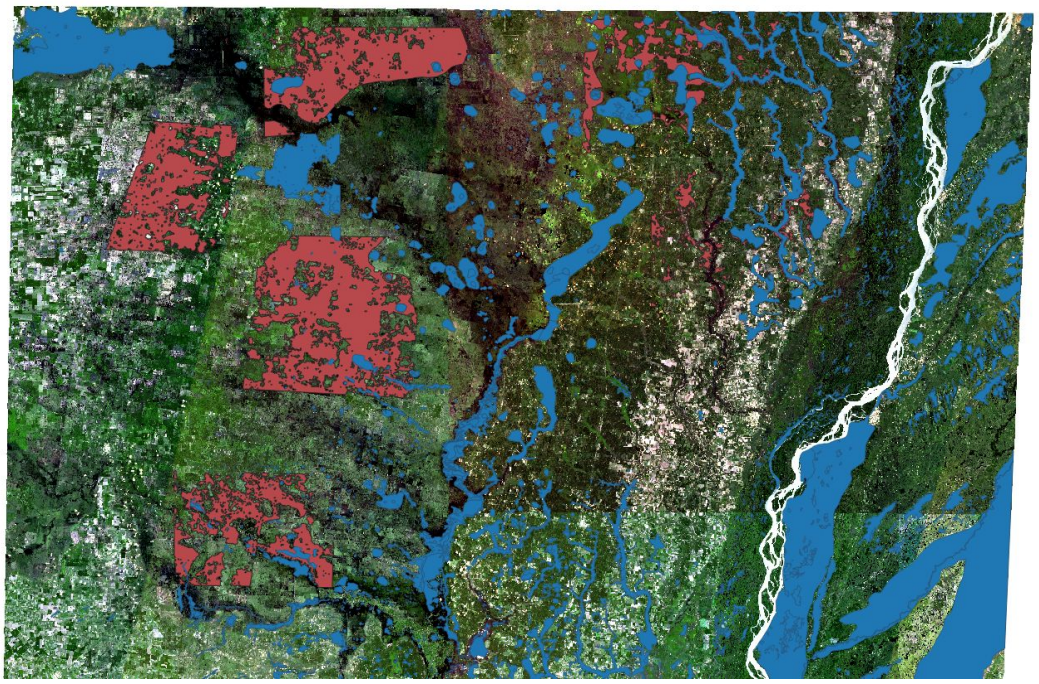


Figura [1]: Imagen completa de la zona de estudio, corresponde a la unión de 6 raster. En azul se observa la capa de aguas continentales, en rojo la capa de evidencias de campo.

Conjunto de datos

Para realizar la clasificación de las imágenes raster, se usaron 4 bandas del satélite Sentinel 2: NIR- Near InfraRed (o infrarrojo cercano) y las bandas BGR (Blue, Green y Red). En total son 6 rasters de enero de 2019, otros 6 rasters de noviembre de 2018 y otros seis de enero de 2018. A lo anterior se le suma una capa vectorial de aguas continentales obtenidas del IGN y otra capa vectorial de evidencias de campo usadas para los experimentos.

Software utilizado

- QGis
- R y RStudio
- Orfeo Toolbox

MARCO TEÓRICO

Índices utilizados

Para mejorar la clasificación se obtuvieron los siguientes índices:

- NDWI[2] (Método McFeeters, 1996): Los potenciales valores obtenidos a partir del NDWI oscilan entre -1 y 1 cuyos valores describirán superficies de agua y vegetación con contenido en agua o zonas terrestres y con ausencia de humedad. Este índice resalta las masas de agua. El cálculo del índice se realizó según [1].

$$\textbf{NDWI} = (\text{Green} - \text{NIR}) / (\text{Green} + \text{NIR}) \qquad [1]$$

- NDTI: Normalized difference turbidity index. Registra diferencias en la claridad del agua. El cálculo del índice se realizó según [2].

$$\textbf{NDTI} = (\text{Red} - \text{Green}) / (\text{Red} + \text{Green}) \qquad [2]$$

- NDVI: Este índice resalta la vegetación. Resultará de utilidad teniendo en cuenta que para definir el estado de la vegetación se tiene en cuenta el agua en estas. El cálculo del índice se realizó según [3].

$$\textbf{NDVI (Sentinel 2)} = (\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red}) \qquad [3]$$

- RVI[3]: Ratio vegetation index. Jordan [4] propuso en 1969 uno de los primeros VI llamados Ratio Vegetation Index (RVI), que se basa en el principio de que las hojas absorben relativamente más rojo que la luz infrarroja. El cálculo del índice se realizó según [4].

$$\textbf{RVI (Sentinel 2)} = \text{Red} / \text{NIR} \qquad [4]$$

Otros índices

Si bien Sentinel-2 posee 13 bandas en total, solo se trabajó con las 4 bandas provistas (Red, Green, Blue y NIR). Esto limitó la posibilidad de utilizar otros índices útiles para calcular concentraciones de agua, como por ejemplo, NDWI propuesto por McFeeters, que requiere la banda SWIR.

[2] http://www.gisandbeers.com/calculo-del-indice-ndwi-diferencial-de-agua-normalizado
[3] https://www.hindawi.com/journals/js/2017/1353691/

Modelos predictivos

Algunos de los modelos predictivos utilizados fueron Random Forest y Ranger. Dada la naturaleza del proceso de bagging, resulta posible estimar de forma directa el test error sin necesidad de recurrir a cross-validation o a un test set independiente. Sin entrar en demostraciones matemáticas, el hecho de que los árboles se ajusten de forma repetida empleando muestras generadas por bootstrapping conlleva que, en promedio, cada ajuste usa solo aproximadamente dos tercios de las observaciones originales. Al tercio restante se le llama out-of-bag (OOB). Si para cada árbol ajustado en el proceso de bagging se registran las observaciones empleadas, se puede predecir la respuesta de la observación I_i haciendo uso de aquellos árboles en los que esa observación ha sido excluida (OOB) y promediando (la moda en el caso de los árboles de clasificación). Siguiendo este proceso, se pueden obtener las predicciones para las n observaciones y con ellas calcular el OOB-mean square error (para regresión) o el OOB-classification error (para árboles de clasificación). Como la variable respuesta de cada observación se predice empleando únicamente los árboles en cuyo ajuste no participó dicha observación, el OOB-error sirve como estimación del test-error. De hecho, si el número de árboles es suficientemente alto, el OOB-error es prácticamente equivalente al leave-one-out cross-validation error. Esta es una ventaja añadida de los métodos de bagging, ya que evita tener que recurrir al proceso de cross-validation (computacionalmente costoso) para la optimización de los hiper parámetros[4].

[4] Joaquín Amat Rodrigo j.amatrodrigo@gmail.com; Febrero, 2017

METODOLOGÍA Y DESARROLLO

Previo al desarrollo de los experimentos, se procedió a analizar el comportamiento de las variables durante el período de análisis.

Análisis exploratorio

Índices

- NDWI

En las Figuras [2] y [3] podemos observar el comportamiento de este índice en las fechas de enero 2018 y enero 2019 (fecha de inundación). Se puede destacar con facilidad las zonas específicas que se vieron sometidas a exceso hídrico. En la Figura [2] tenemos un flujo de agua estable y zonas aleatorias con niveles medios del índice que se corresponden con los humedales propios de la región.

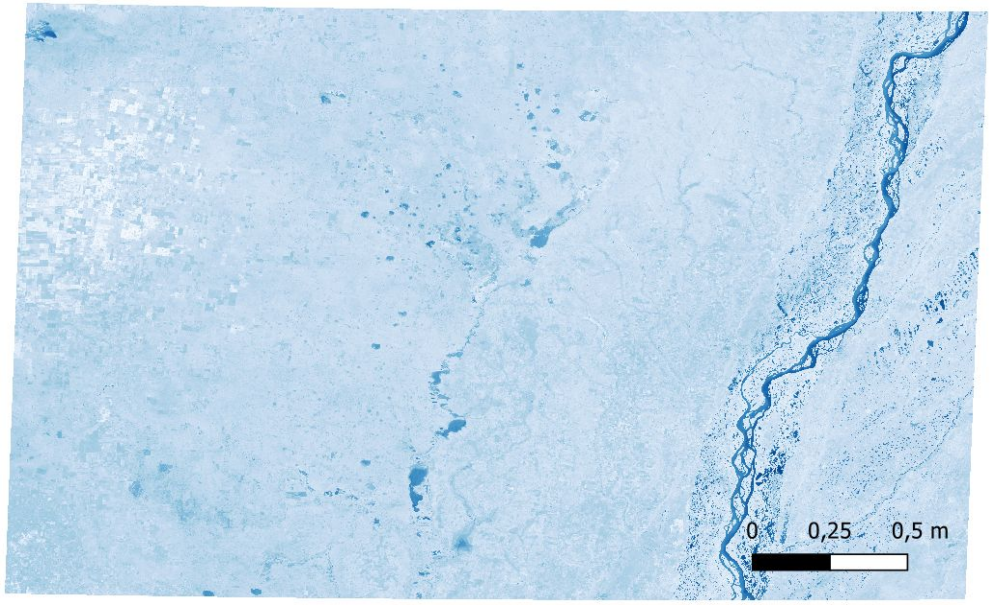


Figura [2]: Índice NDWI para la zona bajo análisis en enero 2018.

Ocurrida la inundación (ver Figura[3]), es posible observar un cambio brusco en la intensidad del índice NDWI. Esto se encuentra reflejado en los colores azules más intensos en la figura. Esta variación implica una mayor presencia de agua en la zona.

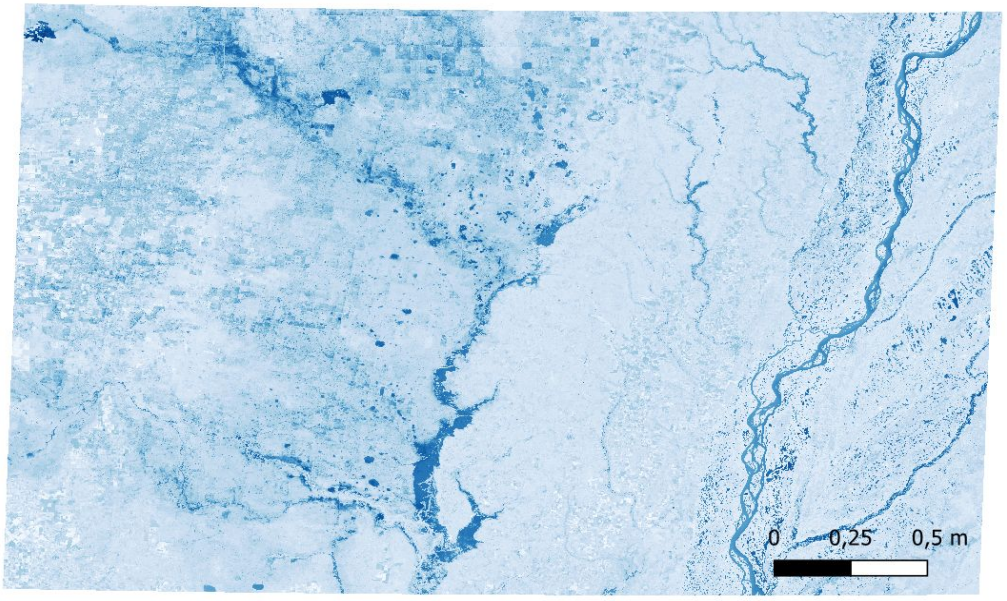


Figura [3]: Índice NDWI para la zona bajo análisis en enero 2019.

Este cambio radical puede apreciarse con mayor notoriedad en la zona centro y norte de la región bajo análisis. Se considera que esta variable será de gran utilidad para determinar la cantidad de hectáreas inundadas.

- NDTI

Esta variable fue incorporada con el fin de poder distinguir zonas con agua estable y áreas inundadas. Si bien, como veremos más adelante, esta es una variable utilizada para modelar, no será utilizada con el fin original con la que fue mentada.

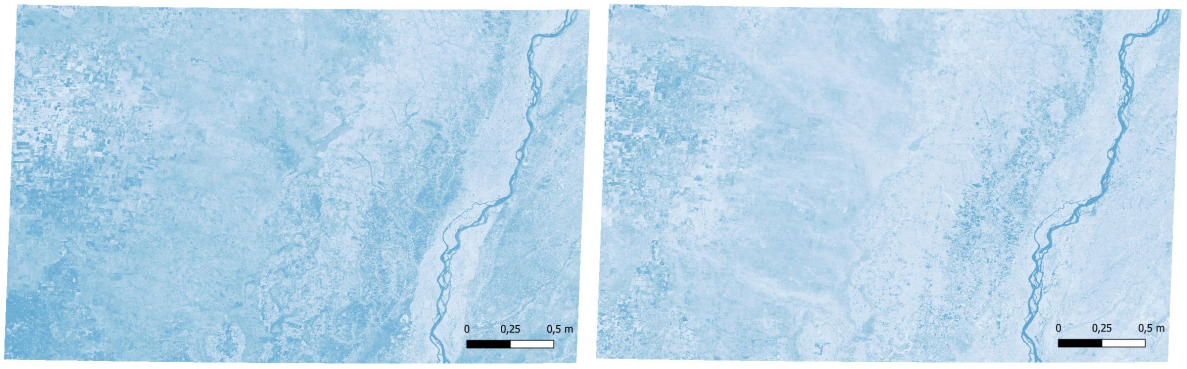


Figura [4]: Evolución del índice NDTI entre enero 2018 y enero 2019

Se observan variaciones entre un período y el otro. Aun así, no se puede afirmar a priori que sea una variable relevante para el problema.

- NDVI

Observando la evolución de este índice entre los periodos de Noviembre 2018 y Enero 2019, es posible destacar con claridad las zonas afectadas por la inundación.

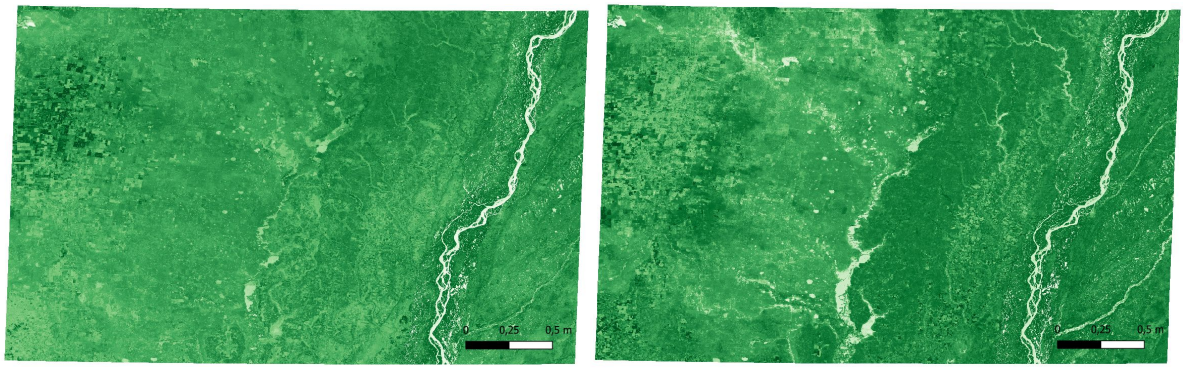


Figura [5]: Evolución del índice NDVI entre enero 2018 y enero 2019

Tanto este índice como el ya mencionado NDWI reflejan, en sí, el mismo fenómeno, la presencia de agua en el píxel evaluado. Estos índices están fuertemente correlacionados ver Figura [9].

- RVI

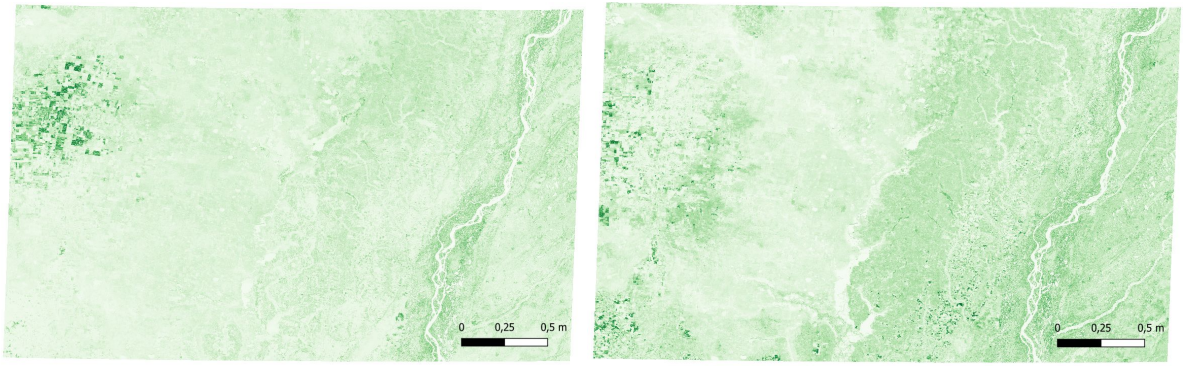


Figura [6]: Evolución del índice RVI entre Noviembre 2018 y Enero 2019

En principio, parecería que este índice logra diferenciar relativamente bien zonas con exceso hídrico (colores menos intensos).

Análisis exploratorio multivariado

Del análisis de los gráficos de dispersión se puede afirmar que tienen una fuerte correlación lineal entre las variables BGR por un lado y por el otro están altamente correlacionadas las variables NIR, NDVI, NDWI y RIV. Por otro lado, la variable NDTI está correlacionada con casi todas las variables, pero principalmente con las variables R y B. Todo esto se puede ver en la Figura [7]:

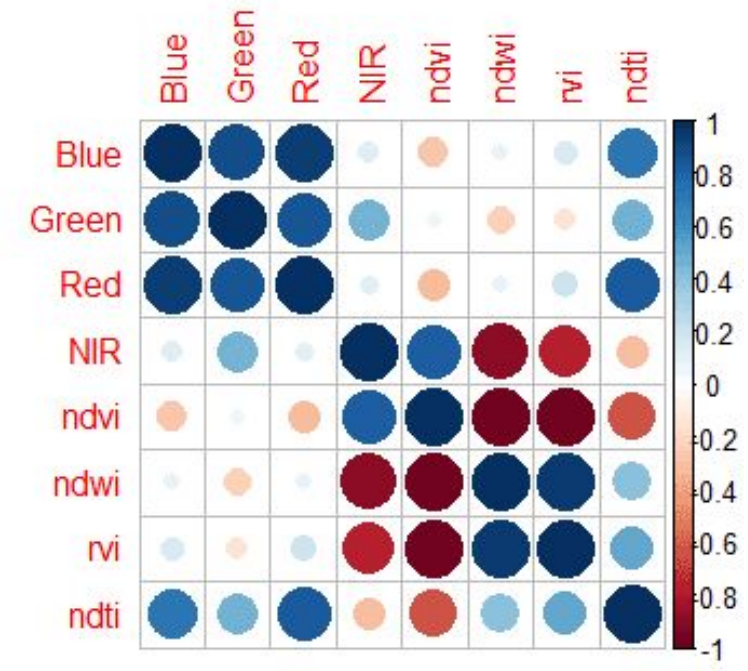


Figura [7]: Correlación entre variables. Se puede apreciar la correlación existente entre las variables representadas por colores, siendo las de tono azul las que poseen una correlación positiva y rojo siendo una correlación negativa.

Las variables BGR están fuertemente correlacionadas de forma positiva entre sí. Por otro lado, los índices también presentan fuertes correlaciones pero en este caso no es unidireccional. La excepción a esta regla resulta ser el índice NDTI que presenta correlaciones débiles con los índices pero correlaciones fuertes y positivas con las bandas BGR.

Analizando la distribución de las variables, en el caso de los histogramas de las variable NDWI, la cola alargada que tiene hacia los valores grandes nos muestra la presencia de aguas. Análogamente, la cola alargada del histograma de la variable NDVI hacia los valores bajos (< 0) nos muestra la presencia de agua.

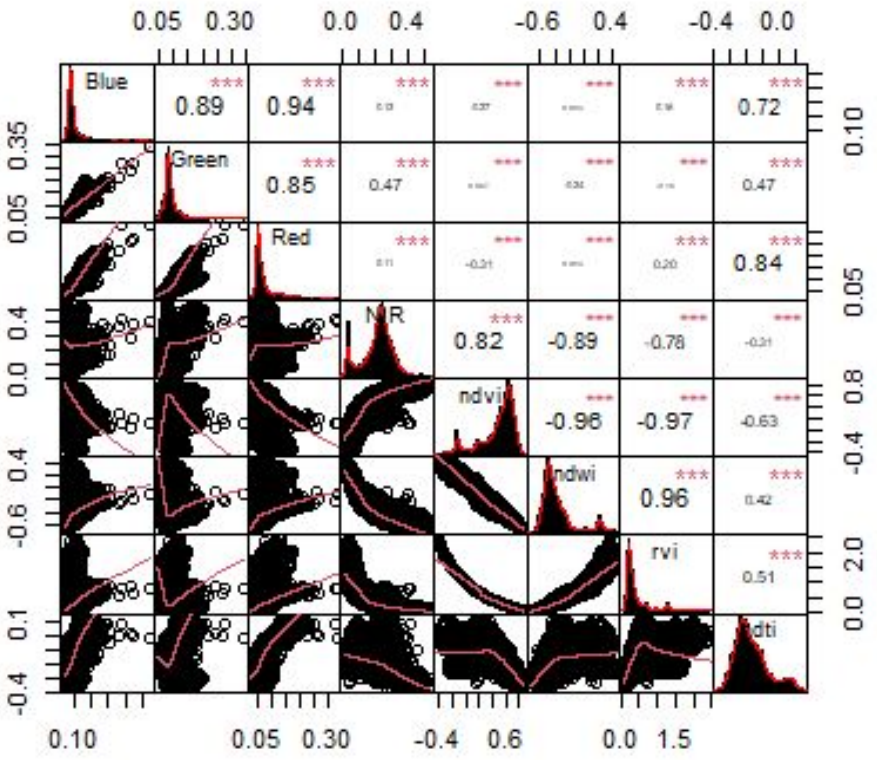


Figura [8]. Histogramas de cada banda / indice y R^2 de las correlaciones.

Se observan fuertes correlaciones lineales entre las variables. En particular nos resulta interesante destacar la relación entre NDWI y NDVI ya que son las variables que mejor representan el problema. En la Figura [8], se representa la colinealidad entre los dos índices. Se puede observar cómo se separan las zonas de agua - no agua en la figura, siendo el color rojo la zona seca y la verde el agua. El gráfico representa a 20.000 puntos (pixels) de la muestra de entrenamiento donde se precisa cuál es el agua y las zonas secas.

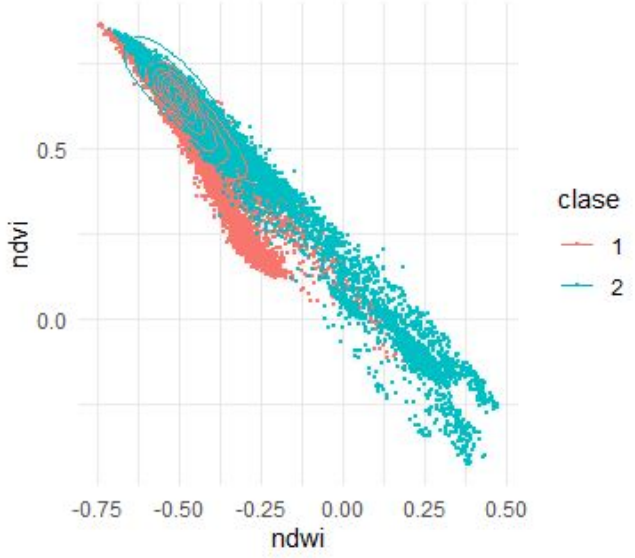


Figura [9] Distribución del índice NDWI y NDVI para una muestra de 20.000 puntos con un R^2 = 0.96. la clase 2 representa agua.

Se observa la fuerte correlación entre los índices NDVI y NDWI: el color verde indica las muestras de agua y las rojas las de no agua. Los datos fueron obtenidos de la muestra de entrenamiento del experimento 1 donde se tomaron unos 40 polígonos agua - no agua. La correlación entre los índices NDVI y NDWI es de un $R^2 = 0.96$.

Análisis de Componentes Principales

Del análisis exploratorio de las variables, se desprende la existencia de colinealidad. El test KMO da 0.59 e indica la adecuación de los datos para realizar el estudio de CP.

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = entrenamiento_df[, -1]) Overall MSA = 0.59

En la matriz de correlaciones, se puede observar la existencia de una gran correlación entre las variables. Esto en concordancia con lo que ya se determinó en el análisis exploratorio. Se determina que las dos primeras CP explican el 93% del problema. En la Figura [10] se muestra el biplot donde se observa cómo se contraponen los índices NDWI y RVI contra los índices NIR y NDVI. Los puntos que caen en el primer cuadrante, PC2 positivo y PC1 negativo, indicarían agua. Por el contrario, el cuadrante opuesto, sería zona seca. Por claridad se incluyó en el biplot a una muestra aleatoria pequeña.

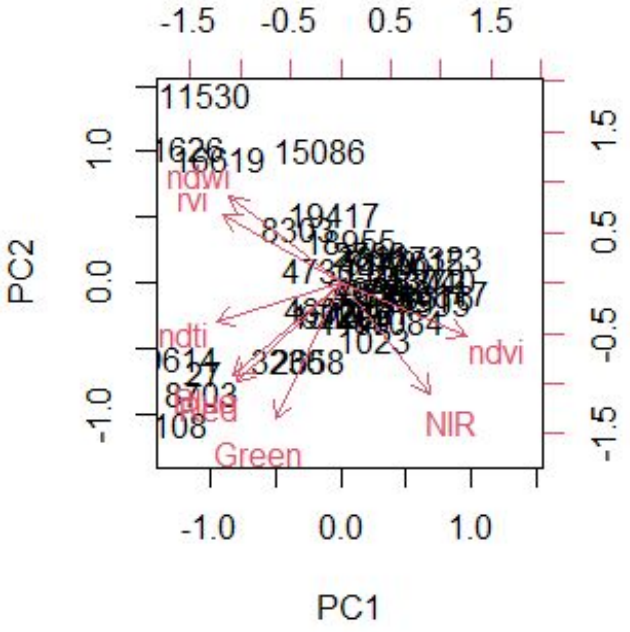


Figura [10]: Gráfico de componentes principales sobre una muestra aleatoria de puntos.

Experimentos

A fin de estimar la cantidad de hectáreas inundadas en enero 2019, se realizaron una serie de experimentos que varían según los índices utilizados, los algoritmos empleados, la forma de generar la muestra de entrenamiento, y de test, la utilización de suavizado, y el empleo de ensambles de modelos. Sin embargo los pasos comunes a todos los experimentos fueron:

Paso 1: armado de capas - visualización de imágenes - análisis de los histogramas

Se armaron las siguientes capas:

- aguas continentales
- polígono de referencia
- registro de evidencia de campo (RE) con áreas inundadas.
- 2018-01
- 2018-11
- 2019-01

Visualización del área de estudio.

Paso 2: cortar

Recortar la capa de aguas-continentales al polígono de referencia.
Los stack raster de las tres capas 2018-01, 2018-11, 2019-01 ya estaban ajustados al polígono de referencia así que se dejaron tal cual como estaban.

Paso 3: cálculo de índices y su visualización; análisis visual de los histogramas

Se obtuvieron los siguientes índices: NDTI, RVI, NDVI y NDWI.

Paso 4: análisis multivariado y de CP

A continuación se detallan cada uno de los experimentos:

Experimento 1 - Variación del NDWI

Para una primera aproximación del problema planteado, se propone una estimación netamente analítica. Partiendo de las Figuras [2] y [3] se calculó la diferencia entre estos dos rasters. De esta manera se pretende destacar las zonas que hayan tenido un cambio alto en este valor entre los períodos enero 2018 y enero 2019. En la Figura [11] se puede observar los resultados.

Diferencias negativas ($NDWI_{2019} < NDWI_{2018}$) se destacan con colores más oscuros. En cambio, variaciones positivas se reflejan con colores claros. De esta forma es posible

distinguir zonas que no presenten cambio (o presenten cambio negativos) de zonas con cambios positivos.

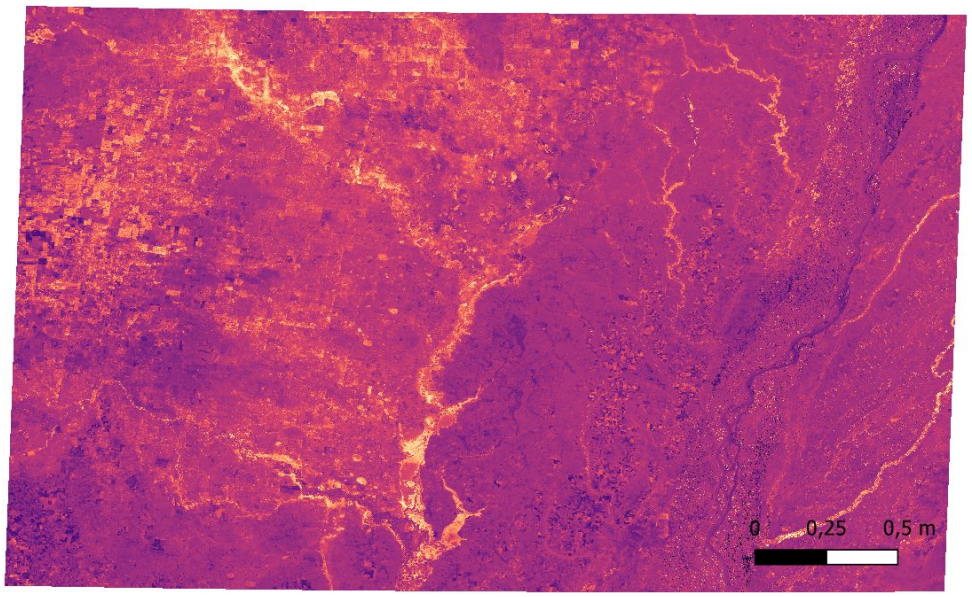


Figura [11]. Diferencia del índice NDWI entre enero 2019 y enero 2018.

Como era de esperarse, toda la zona inundada refleja cambios abruptos en el índice NDWI producto de una presencia mayor de agua. Partiendo de esto, se establecieron una serie de reglas para definir qué será considerado inundación:

- $NDWI_{2018}$ entre -1 y -0.7 con una variación entre períodos mayor a 0.8.
- $NDWI_{2018}$ entre -0.7 y -0.4 con una variación entre períodos mayor a 0.5.
- $NDWI_{2018}$ entre -0.4 y -0.1 con una variación entre períodos mayor a 0.25.
- $NDWI_{2018}$ entre -0.1 y -0.1 con una variación entre períodos mayor a 0.05.

El objetivo es captar, de las zonas que no presentan agua en enero 2018 ($NDWI_{2018} < 0$), aquellas que sufrieron una variación tal que en enero 2019 si se encontraban inundadas. Aplicando las reglas expuestas con anterioridad se llega a la siguiente clasificación:

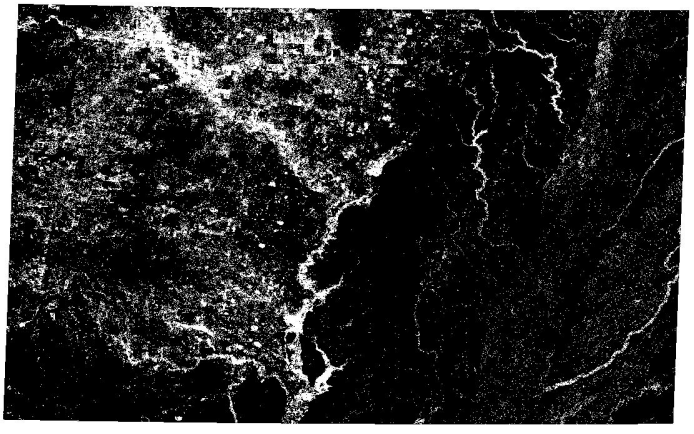


Figura [12]. Clasificación de píxeles en función de las reglas establecidas en el Experimento 1 - Variación del NDWI.

Se contabilizan 818.567 ha inundadas. A priori y en comparación con los resultados de los próximos modelos, se podría asumir que esta metodología de estimación podría tender a subestimar las hectáreas inundadas.

Experimento 2 - Ranger

El modelo de clasificación utilizado fue Ranger, que es una implementación rápida de Random Forest (Breiman 2001) y que fue desarrollado por Leo Breiman y Adele Cutler (2001). Los pasos fueron los siguientes:

Resumen del experimento

Se crea un conjunto de entrenamiento con polígonos pequeños de agua - no agua. Se crea un modelo predictivo usando el raster mas grande y central de la zona. Se obtiene la importancia de las variables, el OOB, la curva ROC y la matriz de confusión entre el conjunto de entrenamiento y un conjunto de test. Se predice con el modelo obtenido la superficie inundada de cada raster.

Detalle de cada paso:

1. Trabajando con la capa 2019, se demarcan unos 40 polígonos pequeños, mitad de cada clase aproximadamente.

   Asignamos clase = 1 a los polígonos de las capas del área-inundada y de aguas continentales y clase = 0 a los polígonos sobre áreas secas (cercanas a la capa RE) y sobre otras áreas, según zonas donde se visualizaban secas en las imágenes de los índices NDVI y NDWI.

2. Con los polígonos obtenidos en 1, creamos un modelo Ranger. Aquí se afinaron los parámetros del algoritmo y las variables de entrada.

3. Se determina un ranking de la importancia de cada una de las variables, y en forma iterativa se van eliminando los índices por no representar ninguna mejora, aun eliminando las variables importantes. El riesgo de overfiting lo atenuamos eligiendo solamente 20.000 puntos, diez mil de cada clase, en forma aleatoria sobre aquellos 40 polígonos. La eliminación de variables no importantes, favoreció la velocidad de los procesamientos.

4. Se analiza la curva ROC y el error OOB en forma conjunta y se vuelve al paso 2 en forma iterativa. Se determina el punto de corte.

5. Para eliminar el posible overfiting que tenga el OOB-error determinado en el paso 4, se testeo también analizando la matriz de confusión obtenida a partir del conjunto de entrenamiento (prueba ingenua), separamos al conjunto de entrenamiento en dos: train y test (75 - 25) y se obtuvo la matriz de confusión, resultando levemente más

bajo que el OOB como es de esperar. El valor obtenido se puede tomar como el error del método.

6. Aplicamos el métodos de clasificación Ranger creado en paso 2 y afinado en los pasos 3, realizamos las predicciones sobre los seis raster separadamente.

7. Calculamos el área inundada para cada raster (ver Tabla [2])

8. El resultado final es la suma de los seis resultados del paso anterior menos la superficie de aguas continentales.

Resultado segmentados:

raster	sup en ha	porcentaje inundado %	sup inundada
s2-2019010000000000-0000000000.tif	2.250.216	40.31	927.775
s2-2019010000000000-0000016384.tif	2.301.521	39.11	900.227
s2-2019010000000000-0000032768.tif	619.056	66.47	392.938
s2-2019010000016384-0000000000.tif	899.622	29.77	267.793
s2-2019010000016384-0000016384.tif	923.635	48.03	443.612
s2-2019010000016384-0000032768.tif	223.988	61.06	136.767

Tabla [2]: superficies inundadas por raster

Total de aguas = 3.069.112 ha
Total de aguas continentales = 969.225 ha
Superficie inundada = Total de aguas - Total de aguas continentales
Superficie inundada total = 2.099.887 ha

Otros resultados intermedios:

En el entrenamiento con 1000 arboles. 2 variables y con 20000 puntos se obtiene un OOB prediction error (Brier s.): 0,1197072.

La curva ROC se muestra en la figura [13], la determinación del punto de corte óptimo es: 0.4953228. La curva ROC es similar tanto en el conjunto de entrenamiento como en el conjunto de prueba dando valores de hasta del 93% si se aumenta la cantidad de puntos. La división tomada para el testing fue de 0.75 - 0.25. El porcentaje de aciertos de la matriz de confusión (MC) es: 84,62%.

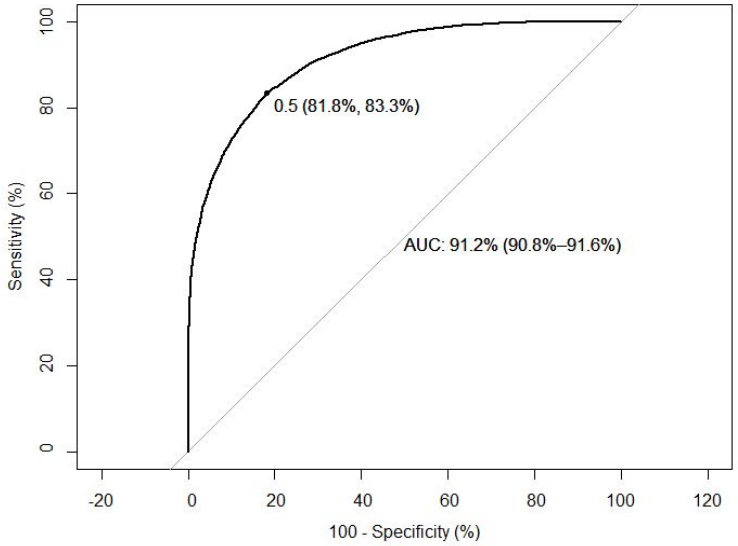


Figura [13] - Curva ROC del experimento 2. El área bajo la curva ROC es del 91.31% y el intervalo de confianza es del 95% es CI: 0.9084-0.9168.

Importancia de las variables:

Red:	1529.128
Blue:	1249.125
NIR:	1147.888
Green:	1096.673

Experimento 3 - Ensambles

Para este experimento se trabajó con los 6 rasters en forma agregada para cada uno de los años. Se unieron los rasters de menor dimensión, se generaron nuevas variables y se unieron estas variables a la información de las bandas blue, green, red y NIR. Este proceso llevo aproximadamente unas 14hs netas[5].

Features

Las variables generadas fueron los índices: NDVI, NDTI, NDWI y RVI. Todas estas definidas en apartados anteriores. Asimismo, para el raster de Enero 2019, se generaron variables de diferencia de índices respecto a las mediciones de Noviembre 2018 para los atributos NDVI y NDWI. Fueron seleccionadas estas variables por ser consideradas las que mejor podían lograr captar el fenómeno puesto en consideración.

[5] No se tienen en cuenta problemas de conexión ni pruebas intermedias.

Para el modelado únicamente serán utilizados los cuatro índices generados más las dos variaciones registradas. De esta forma, se elimina del análisis la información original.

Información de entrenamiento

Respecto a la información de campo, se incorporaron a la información suministrada como "inundado" 110 polígonos a lo largo de todo el raster. Se utiliza el índice NDWI para marcar zonas donde el índice regístre valores tendientes a -1. Asimismo, se incorporaron nuevos polígonos del tipo "inundado" donde se observó zonas tendientes a 1. En Figura [14] se destaca la distribución de los mismos:

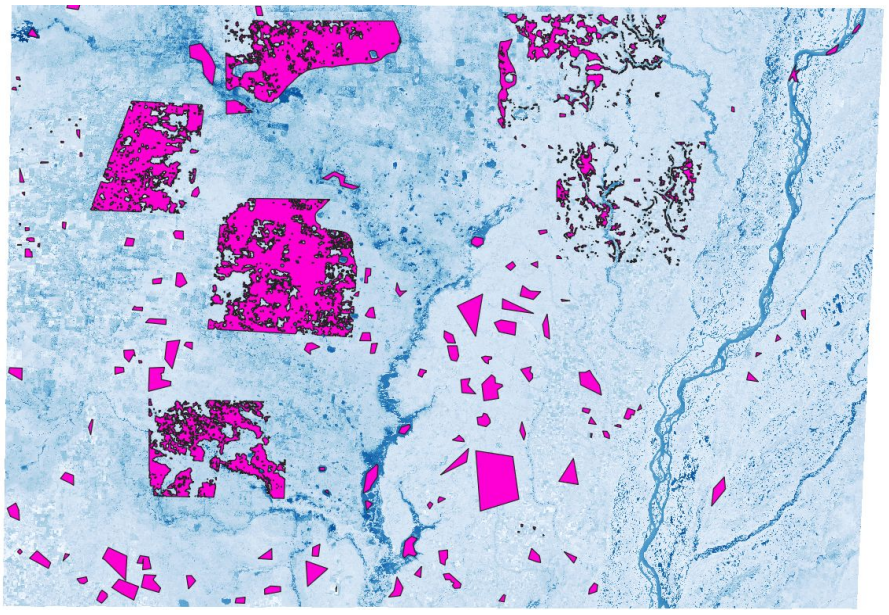


Figura [14]: Polígonos de muestreo.

Tomando los polígonos generados, se procedió a hacer el preprocesamiento de la información. Los tiempos netos de procesado se exponen en Tabla [3]:

Proceso	Tiempo
`otbcli_PolygonClassStatistics`	1hs
`otbcli_SampleSelection`	2hs
`otbcli_SampleExtraction`	5hs
`otbcli_ComputeImagesStatistics`	3min

Tabla [3] Tiempo de preprocesamiento para clasificación

Dado lo desequilibrado de las muestras tomadas se optó como estrategia de muestreo el undersampling. Como muestra de entrenamiento nos quedaremos con 24 millones de puntos distribuidos 50-50 entre las clases agua y no agua. Teniendo en cuenta la cantidad de datos con los que se entrena el modelo y su dispersión en el espacio, se opta por no separar entre entrenamiento y validación. Esta decisión puede corromper las medidas de performance

del modelo propuestos pero no afectarán al modelo final. Asimismo reduce el tiempo de procesamiento permitiendo que sea posible realizar el experimento.

Modelo

Se propone realizar un ensamble por votación de 3 modelos. Fueron seleccionados para el experimento: Bayes, Árboles de Decisión y Shark kmeans. Sus parámetros fueron:

- Bayes: sin parámetros.
- Árboles de Decisión: profundidad máxima del árbol (*classifier.dt.max*) 20.
- KNN: número de vecinos 5 (*-classifier.knn.k*).

Resultados parciales

Las métricas individuales fueron:

Modelo	Accuracy	Kappa
Bayes	92%	83.5%
Árboles de Decisión	95%	89%
KNN	89%	80.5%

Tabla [4] Métricas individuales.

Para evaluar las predicción se optó por armar un score. Se sumarizan las predicciones, de esta forma, tendremos un score alto para píxeles con 100% de predicciones del tipo agua y valores bajos para píxeles donde nunca se predijo el tipo agua. El resultado lo podemos observar en la Figura [15]:

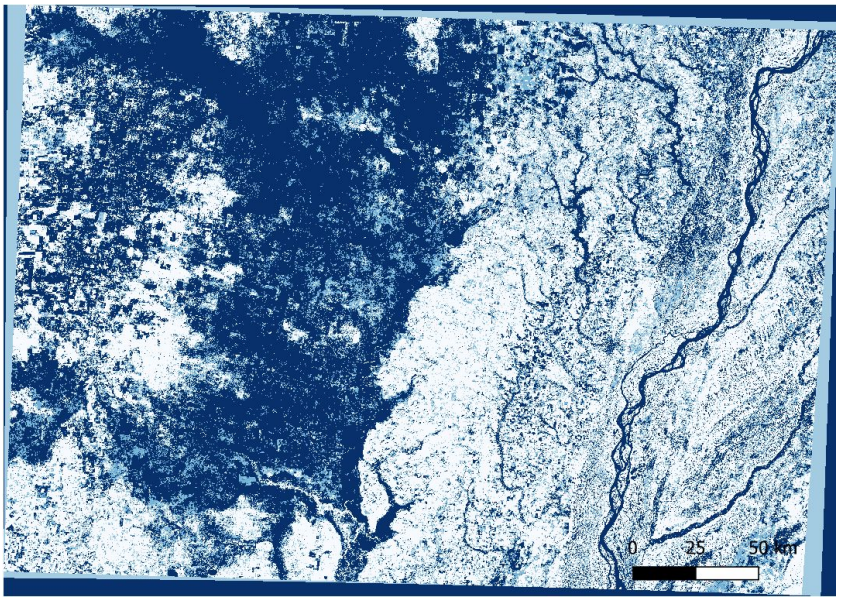


Figura [15]: Score del ensamble de modelos.

En color más intenso tenemos los píxeles que tienden a tres y en color más claro los que tienden a cero. Podemos observar que, de acuerdo a las métricas individuales, el score logra definir bastante bien las dos clases. Hay pocos puntos "indecisos". Como puede observarse en Tabla [5] más del 78% de los píxeles es clasificado por mayoría absoluta (ya sea "agua" o como "no agua").

Score	ha	% ha
3	3.360.119	42,9%
2	826.999	10,5%
1	882.632	11,3%
0	2.769.420	35,3%

Tabla [5] Distribución de score Experimento 3.

Para definir la cantidad de píxeles agua se aplica la metodología de votación. Es decir, píxeles con score mayor o igual a 2 serán consideradas de esta categoría. En base a los resultados, se estima una cantidad de 4.187.118 ha (53.4%).

Por último, se corrige este valor por la cantidad de hectáreas consideradas como aguas en el período noviembre 2018. Esto nos da un total de 2.337.118 ha que podrían ser contabilizadas como inundadas.

Experimento 4 - Varios con Random Forest

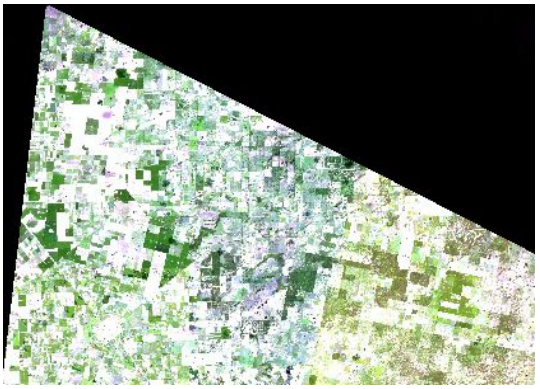
Efectuamos una serie de cuatro experimentos utilizando casi en su totalidad el software Orfeo Toolbox (OTB). Estos experimentos tuvieron un grado creciente de complejidad, y pueden sintetizarse de la siguiente forma:

Modelos

4.1) Tomamos como verdad de campo uno de los polígonos de "agua" (id 3) y dibujamos un polígono de "no agua" en una región que, tanto por inspección visual como por índice NDVI, se alejaba de las características de las zonas con agua (colores verdosos e índice NDVI negativo). De estos puntos tomamos una muestra de 10.000 puntos por clase (20.000 puntos en total). Es decir, se trató de una técnica de undersampling con una distribución en partes iguales para ambas clases. Este modelo fue puesto a prueba en un conjunto de validación independiente de 14 millones de puntos provenientes de los mismos polígonos (ver figuras [16] a [19]).

El algoritmo utilizado fue Random Forest, en su implementación "Shark Random Forest" de OTB. Se mantuvieron los hiperparámetros por default:

a) nbtrees=100
b) nodesize=25
c) mtry=2
d) oobr=0.66



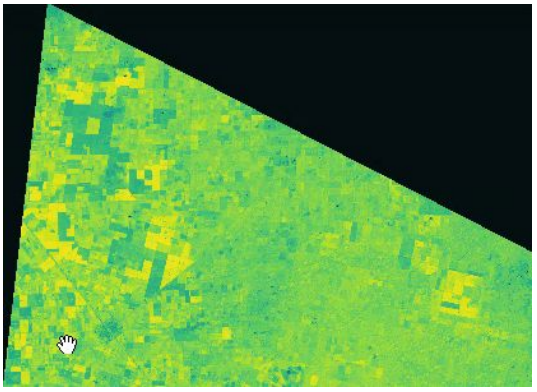
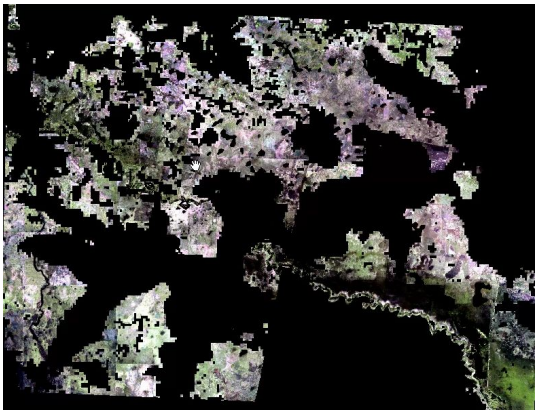


Figura [16]. Izquierda. Clase 0 del conjunto de entrenamiento (original)
Figura [17]. Derecha. Clase 0 del conjunto de entrenamiento (NDVI)



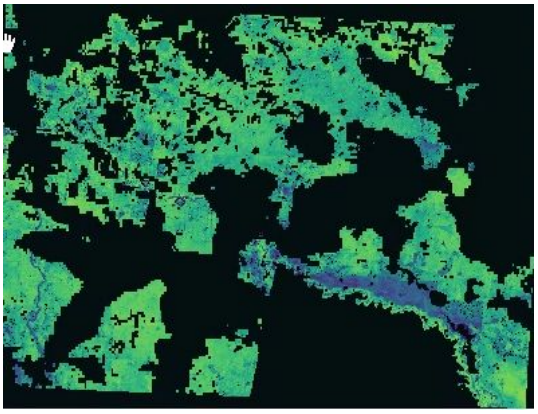


Figura [18]. Izquierda. Clase 1 del conjunto de entrenamiento (original)
Figura [19]. Derecha. Clase 1 del conjunto de entrenamiento (NDVI)

4.2) Se repitió el experimento anterior, esta vez a partir de un conjunto de entrenamiento de aproximadamente 3,5 millones de puntos por clase (7 millones en total), es decir, unas 350 veces más grande que en el experimento 1. Se utilizó el mismo conjunto de validación y el mismo algoritmo con idénticos hiperparámetros. Con este experimento nos propusimos identificar el efecto de incrementar el conjunto de entrenamiento sobre las métricas finales.

4.3) Tomamos en este caso un conjunto de entrenamiento de 4 millones de puntos por clase (8 millones en total), ligeramente mayor al anterior. La diferencia radicó, en este caso, en variar el conjunto de test: en vez de tomar puntos de los mismos polígonos de entrenamiento, se tomaron 2 millones de puntos de otros polígonos: el polígono "id 4" de

“agua” y 4 nuevos polígonos de “no agua” dibujados siguiendo la misma metodología del experimento 1 (ver figuras [20] a [23]).

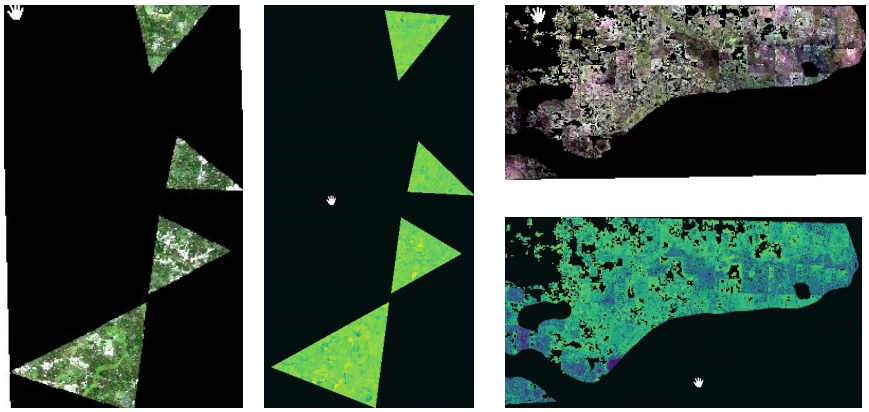


Figura [20]. Izquierda. Clase 0 del conjunto de test (original)
Figura [21]. Centro. Clase 0 del conjunto de test (NDVI)
Figura [22]. Derecha arriba. Clase 1 del conjunto de test (original)
Figura [23]. Derecha abajo. Clase 1 del conjunto de test (NDVI)

4.4) Con este experimento buscamos aislar el efecto de testear sobre polígonos diferentes a los utilizados para entrenar. El experimento replica las características del experimento 3, en cuanto a cantidad de puntos muestreados, y polígonos utilizados para extraer los puntos de entrenamiento y test. La diferencia, y cuyo efecto quisimos comprobar, radicó en efectuar un suavizado a los puntos previamente a entrenar el modelo.

Para efectuar este suavizado apelamos al algoritmo “Mean Shift Smoothing”. Este algoritmo plantea, en su primera iteración, que el valor filtrado, para cualquier pixel input dado, corresponde a la firma espectral promedio de los píxeles vecinos que, al mismo tiempo, están espacialmente más cercanos con respecto al parámetro de radio espacial (spatialr) y con una firma expectral que tiene una distancia euclídeana al píxel input menor que el parámetro de rango radial (ranger). Esto es, píxeles cercanos tanto en espacio como en firma espectral. En iteraciones posteriores el proceso se repetirá considerando que la firma del pixel corresponde a la firma espectral promedio computada en la iteración previa, y que la posición del pixel corresponde a la posición promedio de píxeles utilizados para computar la firma promedio. El algoritmo se detiene luego de alcanzado el máximo número de iteraciones (maxiter) o cuando la posición y la firma espectral del píxel no cambia significativamente entre iteraciones.

Podemos apreciar los efectos del suavizado en los siguientes extractos:

Figura [24]: Extracto de raster enero 2019, sin suavizado.

Figura [25]: Extracto raster de enero 2019, con suavizado.

Resultados parciales

Experimento	Kappa	Predicción clase "agua" - 201901 (ha) (A)	Aguas continentales (B)	Inundaciones (ha) - Estimación 1 (A-B)	Predicción clase "agua" - 201811 (ha) (C)	Inundaciones (ha) - Estimación 2 (A-C)
4.1	0.83	3.099.969	969.225	2.130.744	1.694.661	1.405.308
4.2	0.90	3.579.015	969.225	2.609.790	1.931.252	1.647.763
4.3	0.84	3.534.831	969.225	2.565.606	1.927.406	1.607.425
4.4	0.91	3.698.598	969.225	2.729.373	1.892.494	1.806.104

Tabla [6]: Resultados parciales experimento 4.

Respecto al conjunto de experimentos realizados con Shark Random Forest, observamos, a priori, un resultado muy bueno, en términos de kappa, en el primer intento. A pesar de haberse efectuado un entrenamiento de tan solo 10.000 puntos, y en solamente un

polígono por cada clase, el algoritmo logra un desempeño muy respetable (0,83). El desempeño mejora (a 0,90), como era esperable, al incrementar el conjunto de entrenamiento.

El experimento 3, si bien decae ligeramente en términos de kappa (0,84), nos muestra de forma más fehaciente el comportamiento esperable del modelo en regiones diferentes del sector analizado. Por último, observamos una mejoría en kappa (0,91) al efectuar un suavizado de los datos previo al entrenamiento del modelo. Nuevamente remarcamos, esta mejoría nos habla de un modelo robusto que mantiene una buena performance incluso en regiones (y no solo puntos) distintos a aquellas utilizadas para entrenar.

En vistas de predecir la cantidad de hectáreas inundadas, no sería correcto utilizar la predicción de agua total por los modelos para enero 2019, por cuanto la clase "agua" no diferencia aguas estables de aguas provenientes de inundaciones. Por ello, establecimos dos áreas que sirvan de contrapeso a esta medición inicial. Por un lado, la capa de aguas continentales, que arroja una superficie de 969.225 ha de aguas estables. A partir de la inspección visual es que albergamos dudas sobre la veracidad de las áreas de agua representadas en esta capa. Es por ello que, como segundo contrapeso, aplicamos los mismos modelos a noviembre 2018, y contabilizamos la superficie de agua predicha, a fin de tener un indicador "proxy" de aguas estables. Apreciamos una variabilidad notable entre estas estimaciones: mientras que la capa de aguas continentales ubica a las aguas estables en aproximadamente 1 millón de ha, las predicciones de los modelos para noviembre 2018 las ubican entre 1,7 y 1,9 millones de ha. Como consecuencia de esto, las estimaciones de los diferentes modelos oscilan entre 2,1 y 2,7 millones de ha, a partir de la estimación 1, y entre 1,4 y 1,8 millones, para la estimación 2.

Experimento 5 - Contraste entre predicciones Nov-2018 y Ene-2019

Modelo

Partiendo de la premisa que los algoritmos entrenados son principalmente 'detectores de agua', se propone, usando el mismo modelo, realizar predicciones para la capa de noviembre de 2018 y la de enero 2019, y calcular la diferencia de las probabilidades de los modelos anteriores. El objetivo sería calcular de forma más precisa qué áreas fueron efectivamente 'inundadas', y diferenciarlas mejor de las áreas que son ríos o lagunas (que en este estudio se considera que no pueden ser inundadas si ya estaban cubiertas de agua en su estado normal). Se trabajó con probabilidades en lugar de con predicciones finales, para poder así analizar mejor las variaciones en base al punto de corte elegido, y ver que zonas contrastan más (que zonas se estiman como "más probablemente inundadas"). Evaluando el histograma de la variable calculada, se ven 2 picos principales: alrededor del 0, que representan los valores que permanecen igual en ambas muestras (ya sea cubiertos de agua, o secos), y alrededor del máximo negativo (que en este contexto representan los valores que más probablemente sufrieron inundaciones).

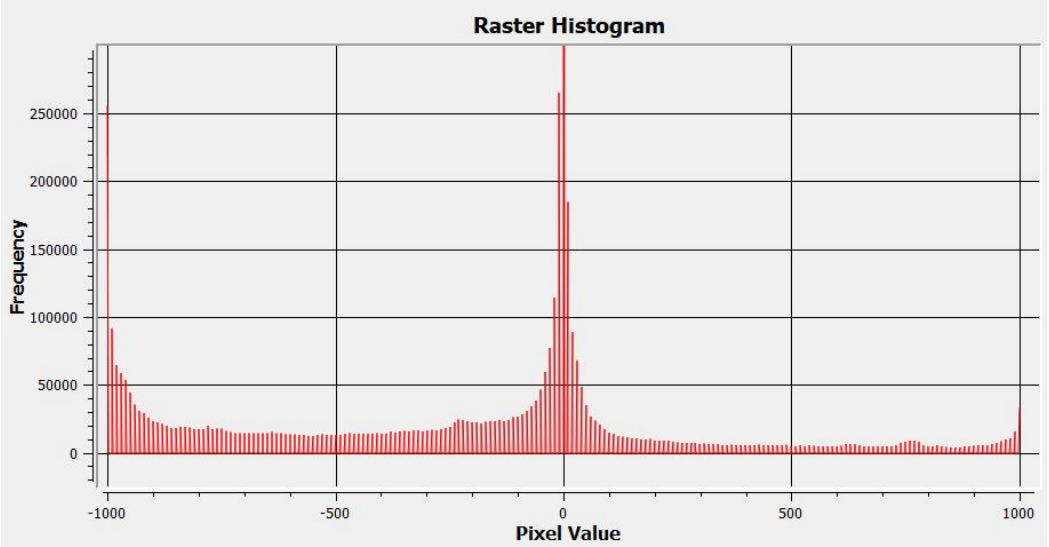


Figura [26]: Histograma de diferencia de probabilidades.

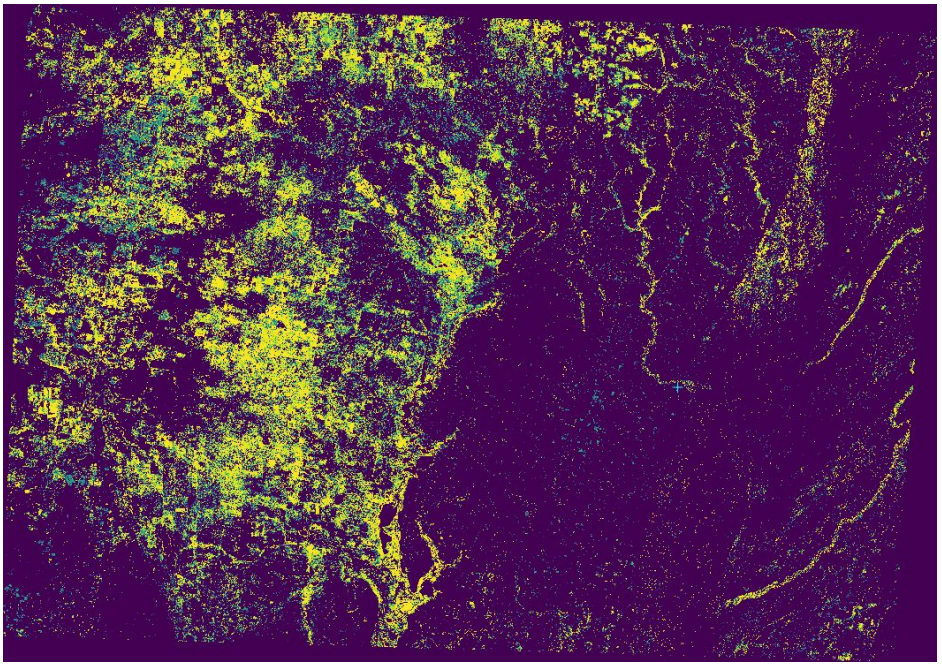


Figura [27]: Mapa de inundaciones; en amarillo las zonas que más aumentó el agua, en azul las zonas que permanecieron igual.

Resultados parciales

Del histograma anterior, y ajustando por área de pixel de estimación, se llegan a los siguientes valores de hectáreas inundadas, con sus respectivos 'márgenes de confianza' (calculados en este caso como diferencia de probabilidades calculadas por el modelo, entre las predicciones de noviembre y enero).

margen de confianza	> 80%	> 60%	> 40%
Hectáreas inundadas	1.212.000	1.633.000	2.000.000

Tabla [7]: Resultados parciales experimento 5.

RESULTADOS Y CONCLUSIONES

A modo de resumen de los resultados se exponen en Tabla [8] las hectáreas inundadas predecidas por cada modelo y la medida de ajuste kappa.

Nro de experimento	Modelo	Superficie inundada predicha (ha)	Kappa
1	Diferencia NDWI	818.567	-
2	Ranger	2.099.887	0,67
3	Ensamble	2.337.118	0,85[6]
4.1	Random Forest	1.405.308	0,83
4.2	Random Forest	1.647.763	0,90
4.3	Random Forest	1.607.425	0,84
4.4	Random Forest	1.806.104	0,91
5	Diferencial de experimento 4	1.633.000[7]	-

Tabla [8]: Resumen de los resultados de los experimentos realizados.

Como podemos observar, los experimentos realizados arrojaron un rango de hectáreas inundadas bastante amplio. Los resultados varían entre 818.567 y 2.337.118 hectáreas. Puede destacarse un rango muy amplio de hectáreas predichas de acuerdo a la metodología utilizada. Como predicciones extremas tenemos el experimento 1 y el experimento 3. Entendemos que el experimento 1 desarrollado en base a la diferencia de NDWI podría estar subestimando las hectáreas inundadas por las reglas que se fijaron. El objetivo fue captar variaciones que lleven, de las observaciones que registraban valores menores a cero en enero 2018, a valores cercanos a cero para enero 2019. Probablemente flexibilizando un poco las condiciones se tendrían valores más homogéneos a los otros experimentos.

Sobre el experimento 2, tiene la ventaja de su simpleza, usa un conjunto de polígonos en el raster central para entrenar y probar un modelo. Una vez creado y afinado hace la predicción con el mismo modelo en todos los rasters. Otras ventajas, son que da el error OOB y el AUC como medidas del error de los predichos y útiles para la comparación con otros métodos, que proporcionen dichas medidas, también como resultado secundario nos da la importancia de las variables y que las mismas se puede usar en otros experimentos. La desventaja es que el valor kappa no es bueno en comparación a otros experimentos. Otra desventaja propia de todos los modelos predictivos derivados de RF, es que por ser un ensamble de árboles, no da idea de la clasificación al especialista.

[6] Promedio del kappa de los modelos ensamblados. Se toma sobre train dada la cantidad de datos en entrenamiento.

[7] Con 'margen de confianza' > 60%

Los modelos asociados a los experimentos 4.3 y 4.4 sugieren un comportamiento robusto, por cuanto fueron entrenados con una cantidad significativa de puntos (8 millones) y testeados en otros sectores del mapa, con muy buen desempeño medido por kappa. El modelo aplicado al experimento 4.4, además, minimiza las probabilidades de *overfitting,* en tanto fue entrenado a partir de datos suavizados. Esta hipótesis es validada por el kappa superior generado.

El experimento 5 tiene de ventaja que calcula en forma diferencial el área inundada, evitando así depender de otras fuentes de información que pudieran estar desactualizadas; y también evitando errores que podrían surgir de estimar incorrectamente como 'seca' áreas que son previamente calculadas como 'agua'. Además permite calcular con cierto tipo de confianza; aunque teniendo en cuenta que esta 'confianza' está atada a lo bien (o mal) que el modelo base calculó las probabilidades en un principio.

En general se puede observar como los modelos suelen ser comparables en sus predicciones en cuanto al orden de magnitud predecida de hectáreas inundadas; pero aún así encontramos diferencias considerables de modelo a modelo. Estas predicciones podrían concordar más (lo cual asemejamos a 'estimar mejor') si se tuviera más información: ya sea las bandas faltantes de los rasters (por e.j., SWIR para calcular otros índices de agua), una verdad de campo de áreas inundadas más grande, o incluso también verdad de campo de áreas no inundadas (estas tuvieron que ser seleccionadas apelando a una inspección visual de los rasters). Además, al tratarse de zonas de humedales, donde no queda claro si el agua presente es por inundación o su estado 'normal', arroja mayor confusión a cualquier modelo predictivo que se entrene.

Para la determinación final de las hectáreas inundadas se calculó el promedio ponderado de las hectáreas predichas por los modelos de acuerdo al kappa que arrojaron. El guarismo asciende a 1.806.624 ha.

BIBLIOGRAFÍA

[1] CF Jordan, "Derivación del índice de área foliar a partir de la calidad de la luz en el suelo del bosque" , Ecology , vol. 50, no. 4, págs. 663–666, 1969.