Detección de outliers en muestras de entrenamiento generadas mediante interpretación visual

Santiago Banchero[1,3], Santiago Verón[1,4], Mariana Petek[1,2],

Sofia Sarrailhe[1,2], Diego de Abelleyra[1]

[1] Instituto de Clima y Agua (INTA), [2] Facultad de Agronomía (UBA)
[3] Universidad Nacional de Luján,[4] CONICET

{banchero.santiago, veron.santiago, deabelleyra.diego}@inta.gob.ar

{mpetek, ssarrailhe}@agro.uba.ar

Resumen
Las clasificaciones supervisadas son procesos extremadamente sensibles a la calidad de las muestras utilizadas. La presencia de outliers en las muestras de entrenamiento suele ser una fuente de error muy frecuente. El objetivo de este trabajo es presentar una metodología de detección de outliers con Isolation Forest, en muestras recolectadas mediante interpretación visual de imágenes satelitales generadas por el Proyecto MapBiomas Pampa Trinacional. Isolation Forest, el algoritmo no supervisado utilizado puede detectar anomalías directamente basándose en el concepto de aislamiento sin utilizar ninguna métrica. La metodología consiste en la identificación de outliers (preparación de muestras, modelado y definición del umbral) y la validación del método. El modelado permite etiquetar de manera automática cada muestra como outlier o normal a partir del score. Se logró verificar los píxeles de la muestra señalada como outlier y tipificar el error en 6 categorías. Los resultados muestran una cantidad decreciente de outliers a lo largo del periodo analizado. Los años con mayor cantidad de outliers tienen una correspondencia con los años de menor disponibilidad de imágenes para la construcción de los mosaicos y contribuciones importantes del tipo Error del Mosaico. La clase con mayor porcentaje de error fue Bosque cerrado (14.7%) y los tipos de errores con mayor proporción fueron Clase Mal Asignada (20.39%) y Borde (19.57%). La metodología propuesta permitió el mejoramiento de muestras obtenidas mediante interpretación visual de imágenes satelitales de manera automática con un 80% de acierto.

Keywords: Machine Learning, Isolation Forest, Sensores Remotos, Detección de anomalías.

1 Introducción

La importancia relativa de las muestras de entrenamiento en la calidad de las clasificaciones supervisadas ha aumentado a la par de la disminución en las limitaciones debidas a la capacidad de cómputo y a la disponibilidad de imágenes satelitales [**1**]. La reciente implementación de plataformas basadas en la nube no sólo permitió el acceso a múltiples catálogos de imágenes sino también a avanzados algoritmos de clasificación. De hecho, plataformas como Google Earth Engine [**2**]

facilitaron también la generación de las muestras de entrenamiento al proveer una alternativa a los tradicionales - y costosos - relevamientos a campo con Sistema de Posicionamiento Global (GPS por sus siglas en inglés). A pesar de su practicidad y economía, la colecta de muestras de entrenamiento a partir de interpretación visual de imágenes de alta resolución por múltiples colaboradores involucra múltiples desafíos. En particular, la falta de homogeneidad de criterio en cuanto a la traducción de la leyenda en atributos espaciales, temporales y espectrales por parte de los colaboradores, la identificación de los límites de las muestras y la calidad del espacio de atributos aumentan la probabilidad de incluir muestras que no son representativas de la clase que se intenta caracterizar.

En una tarea de clasificación supervisada, los outliers -es decir, muestras que poseen valores inusitados del espacio de atributos para la clase a la que pertenecen- suelen ser una fuente de error para dicha clase que conducen a resultados no óptimos. Estos─outliers pueden presentarse por problemas en la generación de la verdad terrestre pero también pueden provenir de mediciones defectuosas. Los datos provenientes de sensoramiento remoto suelen estar contaminados con ruido o errores que requieren ser identificados y luego corregidos o eliminados. Por lo tanto, existe una gran demanda de herramientas de detección de anomalías eficaces y genéricas que requieran una participación mínima de los expertos en la materia y que, al mismo tiempo, tengan la capacidad de adaptarse a diversos conjuntos de datos.

En aprendizaje automático existen diferentes abordajes para la detección de outliers a través de técnicas supervisadas y no supervisadas. Los primeros requieren que los datos ya estén rotulados con los posibles valores atípicos y esto los vuelve menos flexibles en problemas donde la variedad y el volumen son un desafío. Por este motivo, las aproximaciones no supervisadas habilitan soluciones que se adaptan mejor a problemas donde las anomalías son complejas o desconocidas aún.

En este trabajo se utilizó el algoritmo no supervisado de detección de anomalías Isolation Forest (iForest) [**3**] que puede identificar los valores atípicos directamente basándose en el concepto de aislamiento sin utilizar ninguna métrica. Si bien existen otros algoritmos de detección de anomalías que operan a partir de cálculo de distancia [**4**] o basados en densidad [**5**], iForest reduce significativamente el costo de cómputo permitiendo escalar a grandes volúmenes de datos.

El objetivo de este trabajo es presentar una metodología robusta de detección de outliers con iForest, en muestras recolectadas por interpretación visual de imágenes satelitales. Para ello capitalizamos un conjunto de muestras basadas en interpretación visual recolectadas por 8 colaboradores en el marco de la iniciativa MapBiomas Pampa Sudamericano para el área de Argentina [**6**].

1.1 Trabajos relacionados

El mejoramiento de la calidad de las muestras a través de la detección de anomalías es una tarea que en ciencia de datos en general y en teledetección ha recibido poca atención. En el área de mapeo de cobertura y uso del suelo varias iniciativas han propuesto distintos procesos de eliminación de outliers como a través de búsquedas iterativas y definiendo umbrales de probabilidad sobre las bandas espectrales [**7**][**8**] o

realizando una limpieza morfológica de muestras de entrenamiento para reducir el impacto de los píxeles de referencia mal etiquetados [**9**]. Otros métodos como [**10**] utilizan un abordaje de identificación de anomalías no supervisado analizando a través de un criterio de agrupamiento tanto efectos espaciales como temporales de un entorno de vecindad.

2 Materiales y Métodos

2.1 Área de estudio

El área de estudio comprende el territorio argentino de la Región Pampeana e incluye parte de la región del Espinal y el Delta del Paraná. En conjunto esta área cubre una superficie de 634.777 km^2 e involucra 705 millones de píxeles Landsat.

Fig. 1. Área de estudio comprendida por la región Pampa Argentina, Delta y Espinal.

2.2 Conjunto de datos

Se utilizaron muestras recolectadas por 8 colaboradores con experiencia previa en el marco del proyecto MapBiomas Pampas Trinacional [**11**]. Este proyecto tiene como objetivo generar una serie de mapas de cobertura y uso del suelo a partir de clasificaciones supervisadas en forma anual. Dicha serie se compone de un mapa de cobertura por año entre 2000 y 2019 para toda la ecorregión de Pampas, incluyendo Uruguay, Brasil y Argentina. La leyenda de cada mapa incluye Formación forestal natural, Plantación forestal, Zona pantanosa y pastizal inundable, Pastizal, Área agropecuaria, Área no vegetada y Cuerpos de agua. Para ello, el protocolo de muestreo del proyecto describe cómo realizar la interpretación visual de imágenes

Landsat, de mosaicos de muy alta resolución de acceso libre como Google Maps, Google Earth y Bing Maps y también de series temporales del índice de vegetación normalizado (NDVI) para detectar patrones temporales.

El conjunto de datos consistió en 4215 muestras divididas en 10 clases de cobertura del suelo: bosque cerrado, bosque abierto, plantaciones forestales, área húmeda natural no forestal, pastizal, pastura, agricultura, agricultura-pastura, área no vegetada, y ríos, lagos y océanos. Cada una de las muestras estuvo caracterizada por 107 variables de entrada, incluyendo las bandas originales de Landsat (filtradas por calidad), la información fraccional y la información textural, así como también reductores para generar características temporales tales como: medidas de tendencia central por períodos y dispersiones [**11**].

2.3 Isolation Forest

Isolation Forest (iForest) es un método de detección de anomalías que utiliza árboles binarios y el concepto de aislamiento para encontrar valores atípicos [**3**]. Una estructura de árbol es construida desde la raíz hasta las hojas dividiendo los datos en cada nodo a partir de un atributo y un umbral de separación, ambos escogidos al azar. Cada árbol crece hasta que cada dato se aísla en una hoja (Figura 2). La longitud del camino de una instancia, llamada profundidad de aislamiento, se define como la cantidad de aristas que la instancia atraviesa desde la raíz del árbol binario hasta un nodo hoja. Por lo tanto, en un árbol binario generado aleatoriamente en el que las instancias se dividen recursivamente, las instancias anómalas tienen una rápida llegada a los nodos hoja, mientras que las instancias normales requieren muchas más particiones para llegar finalmente a los nodos hoja. En conjunto, las instancias anómalas tienen una longitud media de camino notablemente más corta que la de las instancias normales para un conjunto de árboles binarios.

La puntuación de anomalía asignada a una instancia puede calcularse utilizando la profundidad de aislamiento media en los árboles binarios. Un iForest es un conjunto de árboles de aislamiento q = {Q_1, Q_2, ..., Q_q} y la longitud del camino para un ejemplo x se denota como $h_i(x)$ para un caso $x \in X$ en el árbol de aislamiento Q_i, el promedio de la longitud de caminos sobre todos los árboles q se calcula como:

$$E(h(x)) = \frac{1}{q}\sum_{i=1}^{q} h_i(x)$$

De esta manera, el score de anomalía $s \in (0, 1]$ para cada caso x es calculado de la siguiente forma:

$$s(x) = 2^{-\frac{E(h(x))}{c(M)}}$$

Donde M denota el tamaño de la submuestra y $c(M) = 2H(M - 1) - 2(M - 1)/M$. Aquí $H(M - 1)$ es el número armónico calculado como $\ln(M) + 0.5772156649$ (constante de Euler).

En nuestro trabajo cada caso *x* corresponde a un punto recolectado e interpretado como una de las clases enumeradas en la Sección 2.2 y los valores de todas las bandas del mosaico de un año. La parametrización del algoritmo iForest fue configurada para ajustar 500 árboles y un tamaño de muestra del 25% de los datos disponibles para entrenar. El proceso fue implementado con la librería Python Scikit-learn [**12**] y se ajustaron un total de 200 ajustes para los 20 años (2000 a 2019) y las 10 clases de cobertura del suelo.

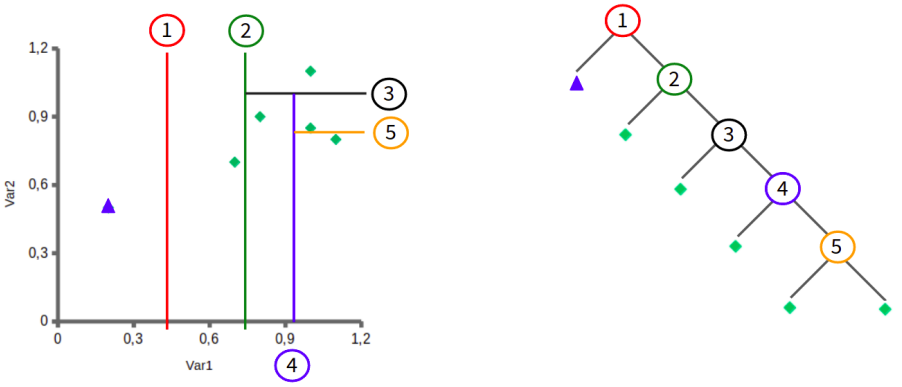

Fig. 2: Esquema que ilustra cómo es el proceso de construcción de un IForest. Las instancias de datos anómalas (triángulo azul) se pueden aislar de los datos normales (rombos verdes) mediante la partición recursiva del conjunto de datos. Los valores atípicos se ubican más cerca de la raíz del árbol.

2.4 Metodología

El esquema de la metodología propuesta se resume en la Figura 3, donde se presentan dos flujos concatenados. El primero muestra la identificación de outliers con la preparación de muestras, modelado y definición del umbral de outlier. El segundo, describe los pasos seguidos para la validación del método.

La identificación de valores atípicos comienza integrando los datos relevados por interpretación visual con la información del espacio de atributos de los mosaicos Landsat de cada año. Esto genera un conjunto de datos con valores de píxeles contenidos en los 4215 polígonos muestreados por 20 años con 107 valores posibles para cada píxel provenientes de los mosaicos.

La etapa de modelado con iForest fue realizada siguiendo un enfoque de búsqueda de outliers de contexto, donde el conjunto de datos fue dividido por año y por clase para correr el algoritmo. Esta segmentación de los datos permitió identificar anomalías presentes en cada una de las coberturas relevadas considerando la variación temporal y los posibles errores de interpretación de la clase relevada. Posteriormente para cada una de estas corridas (200 ajustes en total) se obtuvo un *score* de outliers para cada modelo - es decir, el resultado de una corrida de iForest con sus 500 árboles - que es utilizado para etiquetar cada muestra como outlier o normal.

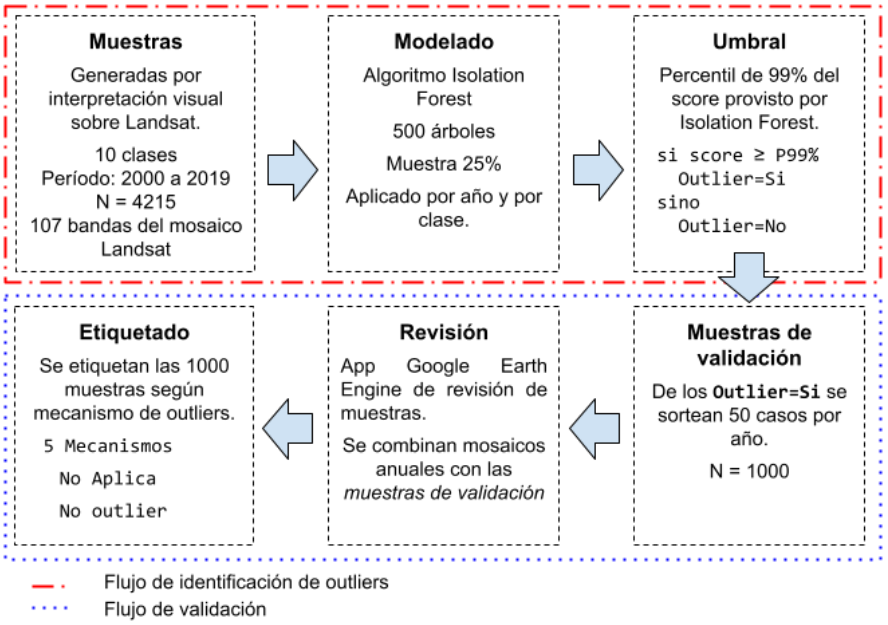

Fig. 3. Esquema metodológico del proceso de obtención y validación de outliers propuesto.

La definición del umbral es un proceso que va acompañado de una exploración de las distribuciones que siguen esos *scores*. Estas tienen un patrón de sesgo muy marcado a derecha y con la presencia de outliers se acentúa aún más (Figura 4). En este trabajo se utilizó un criterio empírico y conservador para definir el umbral de corte tomando, para todas las clases, el valor del percentil del 99%. Todos los valores que superaron ese percentil fueron etiquetados como outliers.

El proceso de validación consistió en la generación de una muestra representativa de los casos etiquetados como outliers y tratar de identificar el mecanismo que originó ese valor atípico a través de una verificación manual. En este sentido se sortearon 50 muestras por año para el período 2000 a 2019 y de esas 1000 muestras resultantes se realizó la revisión por parte de 5 intérpretes experimentados.

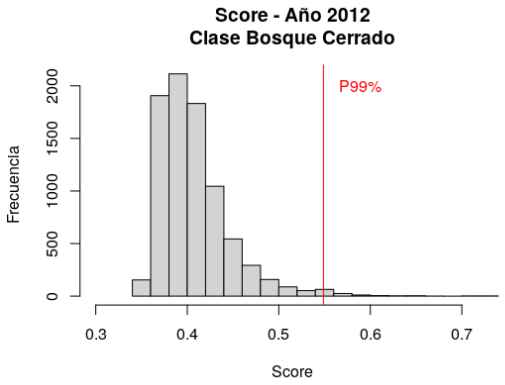

Fig. 4. Histograma del score para el año 2012 de la clase Bosque Cerrado donde se muestra la línea vertical roja correspondiente al percentil del 99%. Todos los valores a derecha de esa línea fueron etiquetados como *outlier=SI* para ese año y esa clase.

Tabla 1. Tipificación de los diferentes mecanismos de outliers identificados.

Identificador	Descripción
Clase mal asignada	Corresponde a una muestra que fue etiquetada de manera equivocada por un intérprete.
Borde	Son muestras cuya geometría fue dibujada en el límite de la clase a relevar. Ejemplo: un polígono dibujado sobre una cortina forestal.
Mezcla con área húmeda	Corresponde a una muestra donde la señal de la imagen satelital se ve influenciada por la aparición de agua de manera no habitual. Ejemplo: un lote agropecuario con una zona baja.
Error del mosaico	Asociado a la presencia de nubes, sombras, entre otros en una o varias imágenes satelitales.
Mezcla de clases	Ocurre cuando una muestra presenta más de una cobertura, ya sea en el tiempo o en el espacio. Ejemplo: la playa de una laguna, en algunos años es agua y en otros es suelo desnudo.
No es un outlier	Muestras clasificadas como outliers que a juicio del intérprete no deberían considerarse como un valor atípico.
No aplica	Corresponde a muestras con valores altos del score donde no se puede tipificar el error. Ejemplo, casos donde los valores atípicos deben estar en alguna de las 107 bandas no utilizadas en el falso color compuesto.

A través de una aplicación desarrollada en Google Earth Engine [**2**] para la revisión, se integraron los mosaicos Landsat de los 20 años y las muestras de outliers sorteadas (Figura 5). La herramienta permitió revisar dicha muestra y verificar si efectivamente se trataba de outliers o no.. De esta manera, para cada año un intérprete evaluó el tipo de outlier según la tipificación de la Tabla 1 para las 50 muestras. El proceso se complementó con imágenes de muy alta resolución disponibles en Google Earth Pro [**13**] y el mosaico de Bing Maps [**14**]. En algunos casos, también se utilizó la verificación con series temporales del índice de vegetación normalizado (NDVI), calculado sobre imágenes Landsat, para complementar la tipificación.

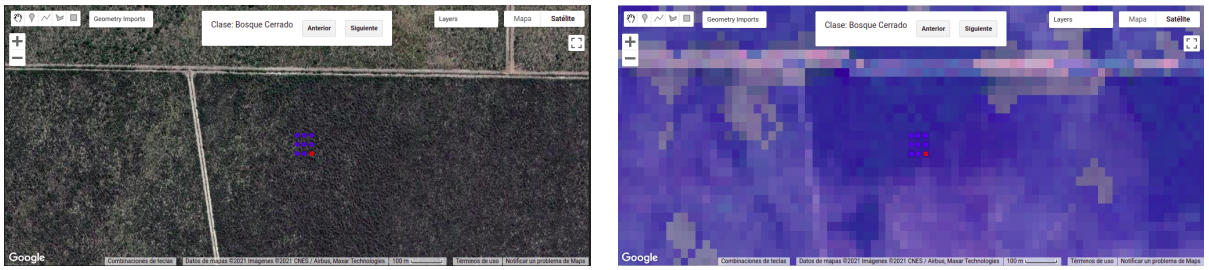

Fig. 5: Aplicación de recolección de muestras de validación. Puntos rojos corresponden a muestras identificadas como outlier por iForest mientras que los azules son valores esperados. La imagen de la izquierda permite visualizar la muestra sobre la imagen de Google Maps de muy alta resolución y en la derecha con el mosaico Landsat de un año en particular.

3 Resultados y Discusión

La metodología propuesta permitió la identificación de muestras atípicas y relacionarlas, analizar su dinámica temporal y relacionarlas con las clases de cobertura del suelo. La cantidad de outliers correctamente identificados por el método iForest disminuyó a lo largo del periodo analizado (Figura 6). Los años con mayor presencia de outliers son 2003, 2009 y 2012, estos reúnen 40 o más casos verificados sobre los 50 relevados para cada año. Esos errores tienen una correspondencia con los años de menor disponibilidad de imágenes Landsat 5 y 7 para la construcción de los mosaicos. Además, en esos años se observan contribuciones importantes del tipo Error del Mosaico (Tabla 1), especialmente en el año 2012 donde el efecto de este mecanismo de outlier explicó aproximadamente el 50% de los casos detectados. El año 2007, que contiene una cantidad de outliers levemente menor que los casos anteriores, también presentó una merma en la calidad de los mosaicos -estimada a partir del número de escenas con datos de buena calidad- y una proporción importante de errores originados por mosaicos.

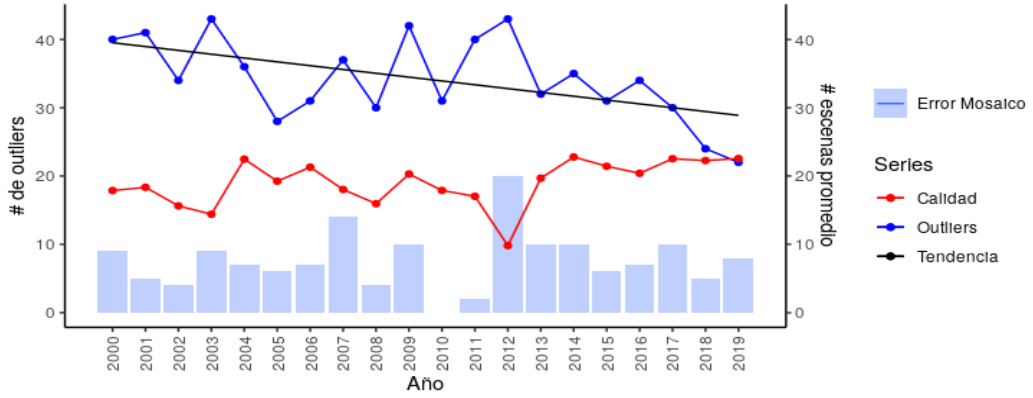

Fig. 6. Evolución temporal de la cantidad de outliers (serie azul), calidad de los mosaicos Landsat medida en disponibilidad de escenas por año para los puntos evaluados (serie roja) y las barras de cantidad de outliers originados por error del mosaico.

Hacia el final del período de estudio, año 2013 en adelante, la calidad de los mosaicos se estabilizó y la cantidad de outliers comenzó a disminuir, en especial los dos últimos años. Aún así los errores detectados del tipo Error del Mosaico no disminuyeron significativamente y la presencia de este mecanismo aportó entre el 10 y el 20 porciento de los outliers de esos años.

Además de la variación temporal se analizó la relación entre coberturas del suelo y los tipos de error (Tabla 2). Los resultados derivados de ese cruce mostraron que la clase con mayor porcentaje de error fue Bosque cerrado (14.7%). Esta clase también registró el tipo de error más frecuente, es decir muestras tomadas en los bordes, que alcanzó el 6.49% de los casos.

El tipo de error que se identificó en mayor proporción fue Clase Mal Asignada, donde se encontraron el 20.39% de las muestras erróneas. Las clases que contribuyeron en mayor medida a este error fueron Áreas húmedas no forestales, Pastizales y Pasturas. También hubo un aporte importante de ríos, lagos y océanos.

Otro error que se encontró con mucha frecuencia fue el generado por muestras ubicadas en los bordes de las coberturas. El 19.57% de las muestras fueron afectadas por este tipo de error originado principalmente en parches de bosques cerrados y abiertos.

Los errores de mosaicos fueron encontrados con mucha frecuencia (15.52% de las muestras) y estuvieron relacionados en su mayoría con la clase Agricultura (4.56%). En el resto de las clases este tipo de error se distribuye de manera uniforme con valores de 0.4% a 2%. En la clase de Áreas no vegetadas no se encontraron casos de error de mosaico. Ante la escasez de escenas las imputaciones de esos huecos no presentan artefactos distinguibles a través de interpretación visual, al menos en la configuración RGB de falso color compuesto que se utilizó en este trabajo.

Tabla 2: Porcentajes de casos por clase y por tipo de los 986 que fueron verificados. Los tipos se corresponden con: No aplica (1), Clase mal asignada (2), Borde (3), Mezcla con área húmeda (4), Problema en mosaico (5), Mezcla de clases (6) y No es outlier (7).

	Tipo							
Clase	1	2	3	4	5	6	7	Total
Bosque Cerrado	0.81	2.84	**6.49**	0.10	1.93	0.30	2.23	**14.70**
Bosque Abierto	0.20	1.22	5.17	0.10	1.32	0.61	2.03	10.65
Plantaciones Forestales	0.51	1.12	2.13	0.00	0.91	0.10	1.62	6.39
Área Húmeda Natural No Forestal	0.61	**3.45**	1.62	0.10	1.42	**3.65**	**3.25**	14.10
Pastizal	0.91	**3.45**	0.41	**2.43**	1.42	0.20	1.72	10.54
Pastura	1.83	2.64	0.20	1.52	2.23	0.10	2.74	11.26
Agricultura	**2.54**	0.91	0.81	1.93	**4.56**	0.10	2.84	13.69
Agricultura - Pastura	1.22	0.30	0.20	0.10	0.41	0.20	0.91	3.34
Área No Vegetada	0.41	1.01	1.32	0.00	0.00	0.61	0.81	4.16
Ríos Lagos y Océanos	1.42	**3.45**	1.22	0.51	1.32	1.22	2.03	11.17
Total	10.46	**20.39**	19.57	6.79	15.52	7.09	20.18	

El tipo de error Mezcla de clases generó el 7.09% de los outliers totales y a pesar de ser uno de los mecanismos menos frecuentes cabe destacar que su principal ocurrencia es en la clase Área húmeda no forestal (3.65%). Esto tiene sentido debido a la dinámica compleja de los humedales donde la variación temporal en la presencia de agua puede generar mucha heterogeneidad. Por otro lado, el tipo de error por Efecto de mezcla con área húmeda fue el que se encontró en menor frecuencia (6.79%) y sus principales interacciones fueron con las clases Pastizales (2.43%) y Agricultura (1.93%).

4 Conclusiones

La identificación de outliers para el mejoramiento del conjunto de muestras obtenidas con procesos de interpretación visual de imágenes satelitales es posible con métodos automáticos como Isolation Forest. La metodología propuesta permitió identificar anomalías de manera automática con una tasa de acierto del 80% aproximadamente. A partir del proceso de validación de los resultados obtenidos se pudo concluir que los errores de mosaico explican algunos casos de los outliers presentes pero no es la principal fuente de error. El principal generador de errores se atribuyó a confusiones en la interpretación de las clases relevadas y a efectos de borde. Estos problemas se asocian a distracciones del intérprete, tiempo de asimilación de los criterios de

relevamiento por parte de los intérpretes, falta de pautas claras en los protocolos de relevamiento, entre otros factores.

Futuros trabajos podrán evaluar el desempeño de la metodología propuesta en comparación con otros algoritmos en términos de capacidad de identificación de outliers y del tiempo de procesamiento. A su vez, un aspecto central no abordado en este trabajo es cuantificar en qué medida la eliminación de las muestras de entrenamiento anómalas genera una mejora en la exactitud de la clasificación de uso y cobertura del suelo. Por último, se podría evaluar cómo funciona el método en otras regiones como el Gran Chaco Americano u otras iniciativas de MapBiomas Global.

Referencias

1. Fritz, S., McCallum, I,, Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., & Obersteiner, M. Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. Remote Sensing, 1, 345-354. (2009).
2. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment. (2017).
3. Liu F. T., Ting K. M., Zhou H.: Isolation-based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data 6(1), 1556-4681 (2012)
4. Knorr, E. M., & Ng, R. T. Algorithms for mining distance-based outliers in large datasets. In VLDB (Vol. 98, pp. 392-403). (1998)
5. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104). (2000)
6. Proyecto MapBiomas Pampa Sudamericano - Colección 1 de la Serie Anual de Mapas de Cobertura y Uso del Suelo del Pampa Sudamericano, https://pampa.mapbiomas.org/, last accessed 2021/07/6
7. Waldner, F., Canto, G. S., & Defourny, P. Automated annual cropland mapping using knowledge-based temporal features. ISPRS Journal of Photogrammetry and Remote Sensing, 110, 1-13. (2015)
8. Radoux, J., & Defourny, P. Automated image-to-map discrepancy detection using iterative trimming. Photogrammetric Engineering & Remote Sensing, 76(2), 173-181. (2010)
9. Radoux, J., Lamarche, C., Van Bogaert, E., Bontemps, S., Brockmann, C., & Defourny, P. Automated training sample extraction for global land cover mapping. Remote Sensing, 6(5), 3965-3987. (2014)
10. Liu, Q., Klucik, R., Chen, C., Grant, G., Gallaher, D., Lv, Q., & Shang, L. Unsupervised detection of contextual anomaly in remotely sensed data. Remote Sensing of Environment, 202, 75-87. (2017).
11. Proyecto MapBiomas Pampa Sudamericano ATBD, https://pampa.mapbiomas.org/es/atbd---know-each-step, last accessed 2021/07/18
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830 (2011)
13. Google Earth Pro, http://www.earth.google.com, last accessed 2021/07/18
14. Bing Maps, https://www.bing.com/maps, last accessed 2021/07/18