

Comparación entre MaxEnt y SVM como clasificadores multiclase. Caso de estudio: Río Amacuzac

Comparison between MaxEnt and SVM
as multiclass classifiers.
Case Study: Amacuzac River

**Yelena Méndez Torres^{1,*}, Carlos Alberto Mastachi Loza^{1,*},
Ricardo Arévalo Mejía¹, María de Lourdes Ruíz Gómez¹, Miguel Ángel Gómez Albores¹,
Humberto Mejía Mojica²**

Recibido: 7 de octubre de 2021 – Aceptado: 14 de marzo de 2022

RESUMEN

En los últimos años, gracias a satélites como el Sentinel 2 y la información que provee, la teledetección ha sido ampliamente utilizada en el estudio de los recursos naturales. Aunado a esta tendencia, se han desarrollado y

actualizado los algoritmos que se utilizan para procesar la información disponible. Una de las principales tareas a realizar con imágenes satelitales es la clasificación de características multiclase. Para llevar a cabo esta tarea, el algoritmo más utilizado hasta el momento ha sido la Máquina de Soporte de Vectores (SVM), sin embargo, trabajos recientes han empleado el

¹ Universidad Autónoma del Estado de México

² Universidad Autónoma del Estado de Morelos

* Autores de correspondencia:
yelenamendeztorres@gmail.com
camastachil@uaemex.mx



algoritmo de Máxima Entropía (MaxEnt) con resultados prometedores.

El objetivo de la presente investigación consiste en comparar el rendimiento de ambos algoritmos en la tarea de clasificar características como: río, vegetación y arena en imágenes satelitales del área del Río Amacuzac. Para poder comparar ambos algoritmos, se propone una metodología que consiste en combinar los resultados específicos del MaxEnt, para cada clase, en un mapa general. Se utilizaron los valores por defecto de los parámetros de ambos algoritmos y se compararon los resultados para los valores de Precisión General (*Overall Accuracy*), Precisión del Productor (*Producer's Accuracy*) y el Área bajo la Curva (AUC). Se llegó a la conclusión de que el algoritmo SVM logró clasificar con mayor precisión las características de interés anteriormente mencionadas para nuestro conjunto de datos en particular.

PALABRAS CLAVE: Teledetección; MaxEnt; SVM

ABSTRACT

The environmental impact caused by the growth and concentration of the population directly affects the amount of urban solid waste that is generated,

becoming one of the challenges on the world. A sanitary landfill is an infrastructure that involves engineering methods for the final disposal of MSW and special handling, to control, through compaction and additional infrastructure, environmental impacts. However, it has been observed that many of them function as controlled sites, causing serious contamination to groundwater due to leachate, product of complex biological and physicochemical processes that occur during storage and degradation. This study analyzed scientific articles from different countries. Numerous groundwater pollutants, as well as indicators of organic matter, have been detected in the various studies. Generally, it has been cited that groundwater contamination appears in the initial stage of the sanitary landfill after five years and reaches its maximum point of contamination. Groundwater contamination appears mainly at short distances from a sanitary landfill. This document describes the maximum concentrations of pollutants that have been reported, as well as their impact on health and the environment.

KEYWORDS: remote sensing; MaxEnt; SVM

INTRODUCCIÓN

La teledetección es cada vez más utilizada como soporte para los estudios ambientales, ya que ayuda a identificar y monitorear las superficies terrestres y las condiciones del ambiente. Una de las tareas dentro de la teledetección es la clasificación multiclase de las imágenes satelitales. La clasificación multiclase (MCC) consiste en la asignación de varias etiquetas (dentro de un conjunto predefinido) a ciertas regiones de una imagen en particular, lo que la hacen una tarea más compleja que la clasificación de una sola clase (OCC) Krawczyk et al., 2018).

Para facilitar la clasificación multiclase son utilizados algoritmos de clasificación automática que reducen el tiempo y la complejidad de la tarea. Uno de los algoritmos más comúnmente utilizado en la literatura es Máquina de Soporte Vectorial (SVM) (Cortes & Vapnik, 2006), sin embargo, algunos trabajos utilizan otros como el algoritmo *Random Forest*, KNN (*K-Nearest-Neighbor*) y *MaxEnt* (*Maximum Entropy*).

El MaxEnt⁴ (Phillips et al., 2006) es particularmente popular en el modelado de distribución de especies y de nichos ambientales. Este también ha sido aplicado en diferentes estudios que incluyen mapear cobertura de tierras urbanas, formaciones de matorrales y arbustos, así como estudios regionales y continentales (Mack & Waske, 2017). Su fácil utilización, debido a los parámetros predeterminados con los que trabaja, lo hacen un algoritmo muy accesible para todos. En algunos estudios se muestra cómo el *MaxEnt* obtiene mejores resultados que el SVM en tareas de clasificación de una clase. Sin embargo, no es común verlo utilizado como un algoritmo de clasificación multiclase. Varios estudios realizan comparaciones entre *MaxEnt* y SVM para encontrar el más idóneo a la hora de clasificar (tanto para una clase como para multiclase); incluso comparan estos con otros algoritmos (Mack & Waske, 2017; Pratiwi & Widodo 2017; Chen et al., 2017; Rahimian et al., 2019; Farrell et al., 2019), sin embargo, no se ha determinado la supremacía de un algoritmo sobre otro.

Inspirados en los buenos resultados de *MaxEnt* para problemas de una sola clase y en el poco uso hasta el momento como clasificador multiclase, en esta investigación lo comparamos con el algoritmo SVM, para contribuir a la estimación de su rendimiento en este tipo de tareas. En este trabajo en específico, se compara el rendimiento de ambos algoritmos

⁴ Este algoritmo se encuentra disponible de forma gratuita en <http://www.cs.princeton.edu/schapiro/maxent/>

en la tarea de clasificar características como: río, vegetación y arena en imágenes satelitales del área del Río Amacuzac. Además de los resultados desde el punto de vista comparativo para la tarea de clasificación multiclase, las imágenes etiquetadas servirán como entradas para un estudio más general que pretende determinar la dinámica espacio temporal del río Amacuzac y sus zonas riparias.

Para la selección del área de estudio se tuvieron en cuenta algunos de los puntos de interés seleccionados por Mejía-Mojica (2012). En estos se han encontrado alteraciones en el ecosistema dentro del río Amacuzac. Teniendo en cuenta los datos disponibles y el fin de determinar la dinámica espacial del río y sus zonas riparias en investigaciones posteriores, el área de estudio quedó limitada entre las estaciones hidrométricas 18232 y 18437 (Figura 1). Ubicada en la región hidrológica 18, pertenece a la cuenca del río Balsas; abarca una superficie aproximada de 430 km² y una longitud de 50 km.

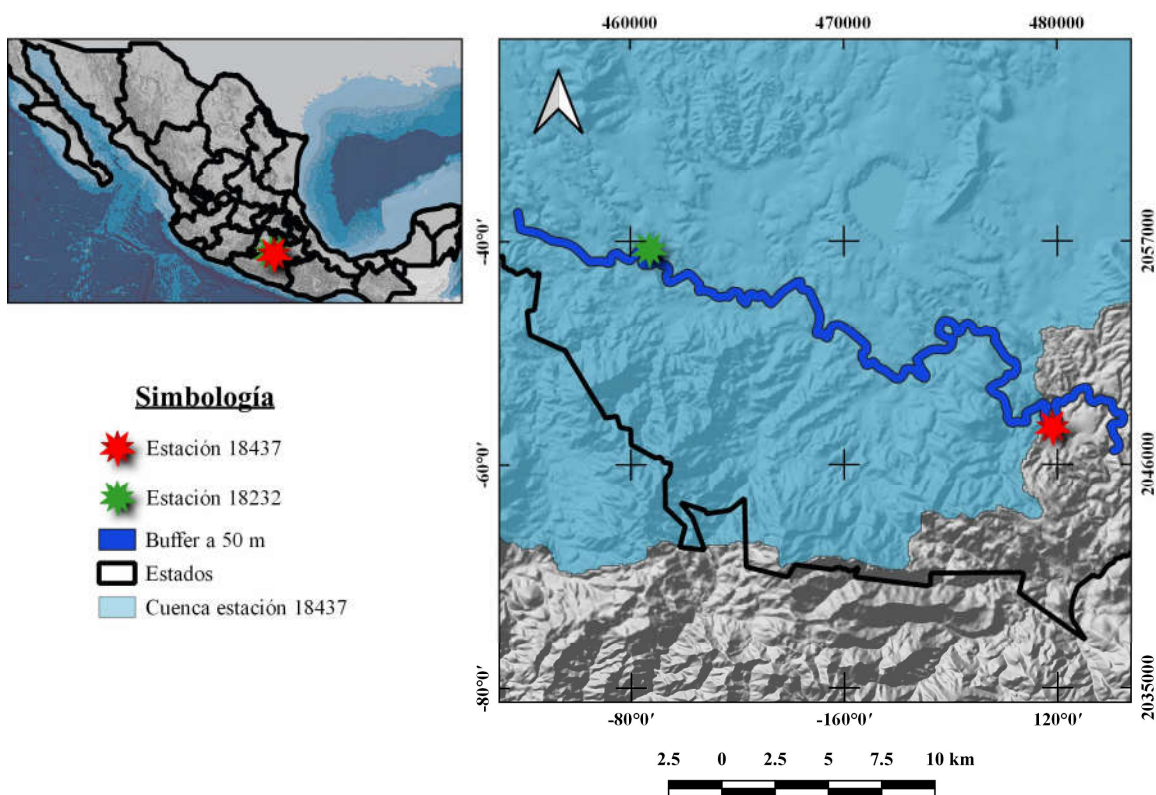


FIGURA 1.

DELIMITACIÓN DEL ÁREA DE ESTUDIO

El área comprende parte de los estados de Morelos, Guerrero y Estado de México. Geográficamente se localiza entre los paralelos 18°27' y 19°14' de latitud norte y los

meridianos $-98^{\circ}37'$ y $-100^{\circ}00'$ de longitud oeste. Se encuentra delimitada por las siguientes regiones y cuencas hidrológicas: al Norte por las Regiones Hidrológicas número 26 Pánuco y 12 Lerma-Santiago; al Sur por la cuenca hidrológica Río Bajo Atoyac; al Oeste por las cuencas hidrológicas Río Cutzamala y Río Medio Balsas; y al Este por la cuenca hidrológica Río Nexapa (DOF, 2011; Soares, 2008).

METODOLOGÍA

Procesamiento de la información

En el primer paso del proceso metodológico se obtuvieron las imágenes Sentinel 2⁵ a 10 m de resolución espacial del área de estudio entre un periodo de 2015-2021. Se realizó un filtro para descartar escenas con alta presencia de nubes sobre el área de estudio y aquellas en las que el satélite no tomó la imagen correctamente. Se procedió a la corrección atmosférica y geométrica de las imágenes satelitales con el programa SNAP (*Sentinels Application Platform*) versión 8.0 con el modelo *Sen2cor*³.

Se continuó el análisis con el Software de sistema de Información Geográfica QGIS versión 3.14. Primeramente se realizó un *buffer* a 50 m a partir del eje del río, con el cual se cubría el área de interés para la posterior investigación. Se procedió a la corrección de las nubes, fueron seleccionados los niveles de salida 2, 3, 4, 5, 6, y 8 como aquellos sin errores en la calidad. Fueron filtradas descartando aquellas que tenían una alta presencia de nubes sobre el área de estudio y aquellas en las que el satélite no tomó la imagen correctamente.

Con *Sen2Cor* se pasaron las imágenes del nivel de producto 1C de manera íntegra sin ser adulterado al 2A. Aquí se trabajó en la corrección atmosférica de los datos de entrada de Nivel 1C de la parte superior de la atmósfera.

Las salidas de Nivel 2A fueron: BOA (*Bottom-Of Atmosphere*), opcionalmente imágenes de reflectancia corregidas por terreno y nubes (cirros), mapas AOT (*Aerosol Optical Thickness*), WVP (*Water Vapour*), SLC (*Scene Classification*) e indicadores de calidad, incluidas las probabilidades de nubes y nieve. El producto de salida fue de imágenes JPEG, con tres resoluciones diferentes, 60, 20 y 10 m.

Finalmente se sumaron todas las imágenes de la banda SCL (*Scene Classification*), para determinar los píxeles en los que no se presentaba suficiente información (nubes). Se

⁵ Este algoritmo se encuentra disponible de forma gratuita en <http://www.cs.princeton.edu/schapiro/maxent/>

conformó una base de datos con áreas donde la omisión de información no fuera mayor a 5 píxeles contiguos. De esta forma, se aseguró que la disponibilidad de información fuese mayor al 95 % de la cobertura de la imagen satelital, creando así una máscara general para no cometer errores al momento de seleccionar los sitios de entrenamientos.

Para construir los sitios de entrenamiento, se determinó la cantidad de píxeles representativos de las imágenes, haciendo uso de la herramienta *Online Raosoft*⁶. Esta herramienta utiliza el tamaño total de una población para estimar el tamaño de muestra representativa. En nuestro caso, la población corresponde al total de píxeles de la imagen y la muestra a la cantidad de píxeles representativos. El valor obtenido representó el número mínimo de píxeles que se tomaron en cada imagen para determinar los sitios de entrenamiento.

Corridas de algoritmos

Se seleccionaron tres clases para los sitios de entrenamiento que se distribuyeron entre los píxeles anteriormente determinados. Las clases de interés fueron arena, vegetación y río. Para la primera se consideraron imágenes comprendidas entre los meses de noviembre a abril, periodo de lluvia. Para la vegetación y el río se seleccionaron imágenes entre los meses de mayo a octubre que corresponde al periodo de seca. De esta forma se garantizó que siempre los píxeles seleccionados correspondieran a la misma clase, aún con la dinámica temporal, pues para ello se consideró el caso más desfavorable.

Para el procesamiento con *MaxEnt* fueron introducidas las clases como datos de presencia (arena, vegetación y río) y como variables las bandas B2, B3, B4 Y B8 con una resolución de 10m y longitudes de ondas centrales respectivas de 492.4, 559.8, 664.6, 832.8. Cabe resaltar que el programa fue ejecutado con los valores predeterminados del mismo.

Como resultado de la ejecución del *MaxEnt* se obtuvieron mapas independientes para cada una de las clases. Los píxeles de cada uno de estos mapas están compuestos por valores de 0 a 1, que representan la probabilidad de encontrar la clase. Es decir, si en un pixel se encuentra el valor de 1, significa que es 100 % seguro que corresponde a la clase representada por esa imagen. Por el contrario, si se encuentra el valor de 0, significa que hay 0 % de probabilidad de encontrar ese pixel en la clase.

Como cada una de las imágenes de salida del *MaxEnt* representa una clase diferente, se determinó que se evaluaría en un mismo pixel cada una de las imágenes y se seleccionaría el de mayor probabilidad. Con estos píxeles de mayor probabilidad se creó el mapa final de clasificación.

⁶Disponible gratuitamente en <http://www.raosoft.com/samplesize.html>

Adicionalmente se utilizó el algoritmo de clasificación SVM de la caja de herramientas de Orfeo 6.7.0 en QGIS. Se le dio como entrada el compuesto de las bandas B2, B3, B4, B8 y los sitios de entrenamiento en forma de polígonos (Gualtieri, 2009). En este caso, el SVM se procesó con los valores predeterminados del módulo. Con la ejecución del algoritmo se obtuvo el mapa de clasificación para cada una de las imágenes.

Validación de resultados

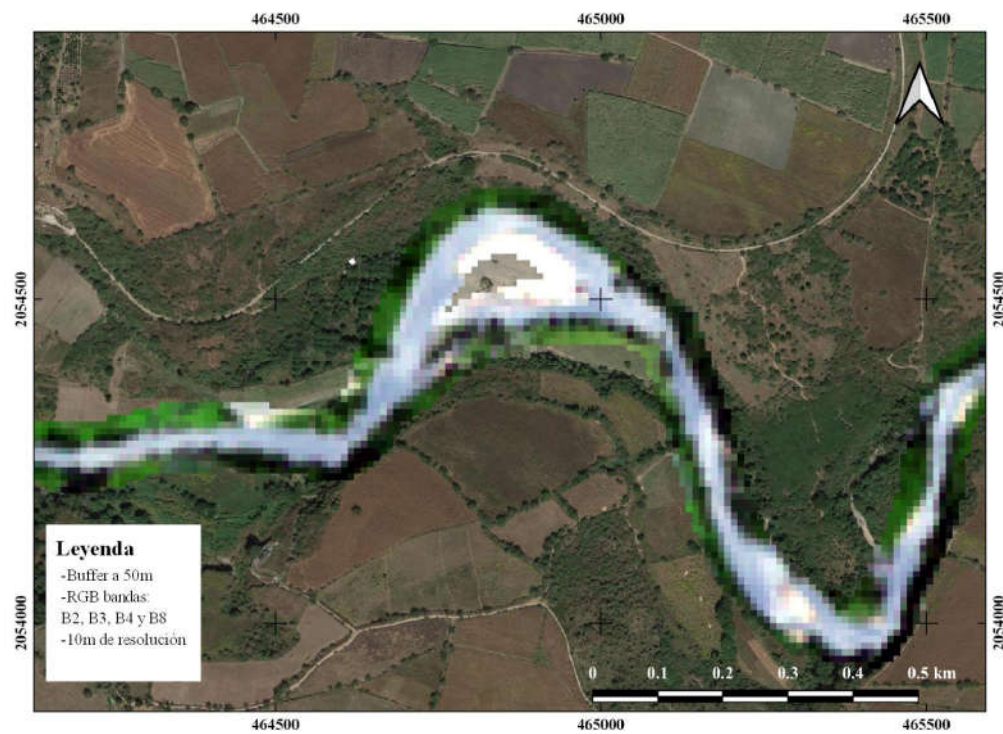
Para realizar la validación, se crearon puntos de control diferentes a los sitios de entrenamiento, pero siguiendo la metodología para determinar estos últimos. Se realizaron comparaciones entre los puntos seleccionados y cada una de las clasificaciones arrojadas por ambos algoritmos. A partir de éstas fueron creadas las Matrices de Confusión de la cual se obtuvieron los valores de *Overall Accuracy* (Stehman, 1997) y *Producer's Accuracy* (Congalton, 1991). Además, se crearon las curvas ROC y se obtuvieron los valores de AUC. Finalmente, con estos valores se logró determinar el algoritmo que clasificó mejor el área de estudio.

RESULTADOS

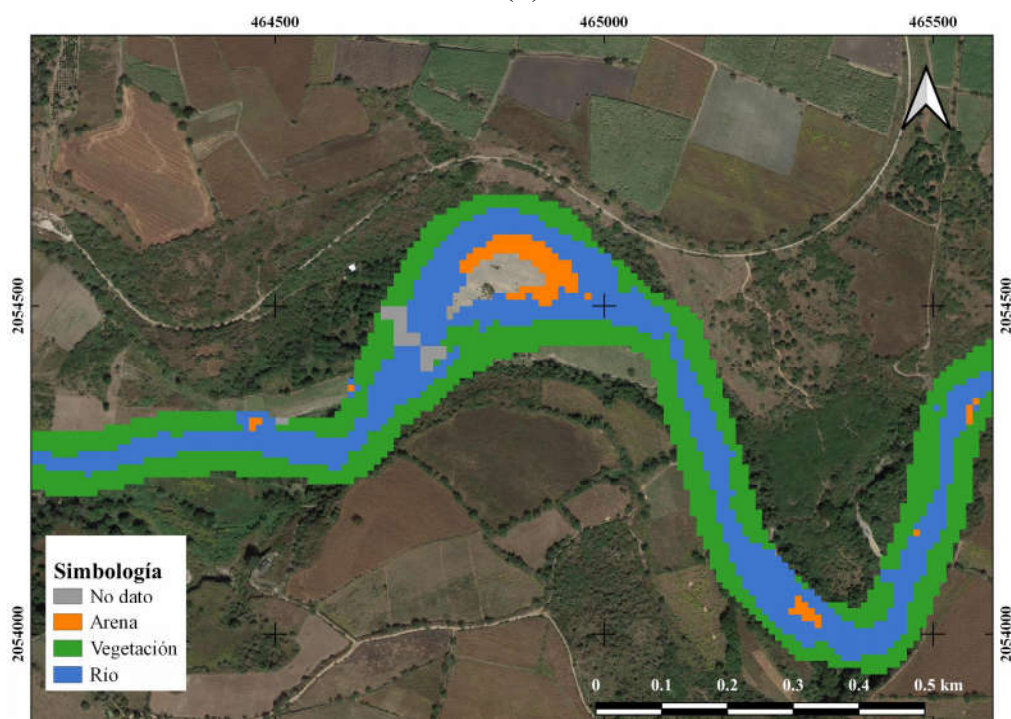
Clasificación de imágenes

Se descargó un total de 167 imágenes y luego del filtro quedaron 100 para procesar en el SNAP. Después, en la Figura 2 (a) fueron trabajadas en QGIS para realizar las correcciones geométricas y las correcciones en las nubes. El tamaño de muestra arrojado por el *Roasoft* fue de 384 píxeles, por lo que se tomaron valores por encima de este para determinar los sitios de entrenamientos.

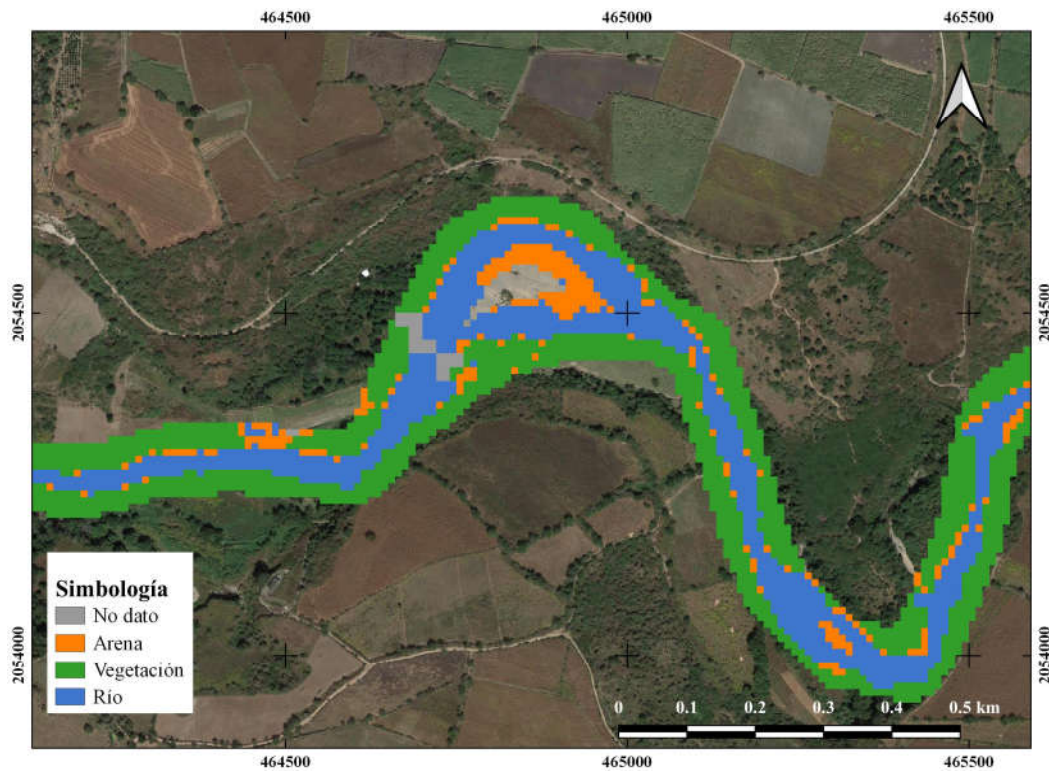
Fueron creados los sitios de entrenamiento para el *MaxEnt* y el SVM teniendo en cuenta las clases. Una vez corridos ambos algoritmos quedaron conformados los mapas de clasificación por cuenca. En la Figura 2 (b) se observa cómo el *MaxEnt* no fue capaz de identificar correctamente la arena, obteniendo mejores resultados para las otras clases. Por el contrario, en la Figura 2 (c), el SVM logró identificar un mayor grado de detalle en la arena, manteniendo resultados similares en cuanto al río y la vegetación.



(A)



(B)



(C)

FIGURA 2

(A) COMPOSICIÓN DE COLOR VERDADERO RGB BANDAS B2, B3, B4 Y B8. (B) CLASIFICACIÓN CON MAXENT,
(C) CLASIFICACIÓN CON SVM. IMAGEN 2017/10/14.

Validación de resultados

En la Figura 3 se pueden observar los resultados obtenidos a partir de matriz de confusión para ambos algoritmos. En estos se observa cómo el *MaxEnt* presenta valores de *Producer's Accuracy* inferiores al 70 % (línea naranja) para todas las clases excepto el río, mientras que el SVM también presenta valores por debajo de este umbral, pero en menor proporción. En cuanto al *Overall Accuracy* se observa que para el *MaxEnt* existen algunos valores por debajo del umbral establecido. Sin embargo, el SVM no presenta ningún valor por debajo de este umbral y la gran mayoría de sus valores están por encima del 80 %. Como resultado de la comparación de los dos algoritmos se determinó que el SVM presenta cierta superioridad ante *MaxEnt*.

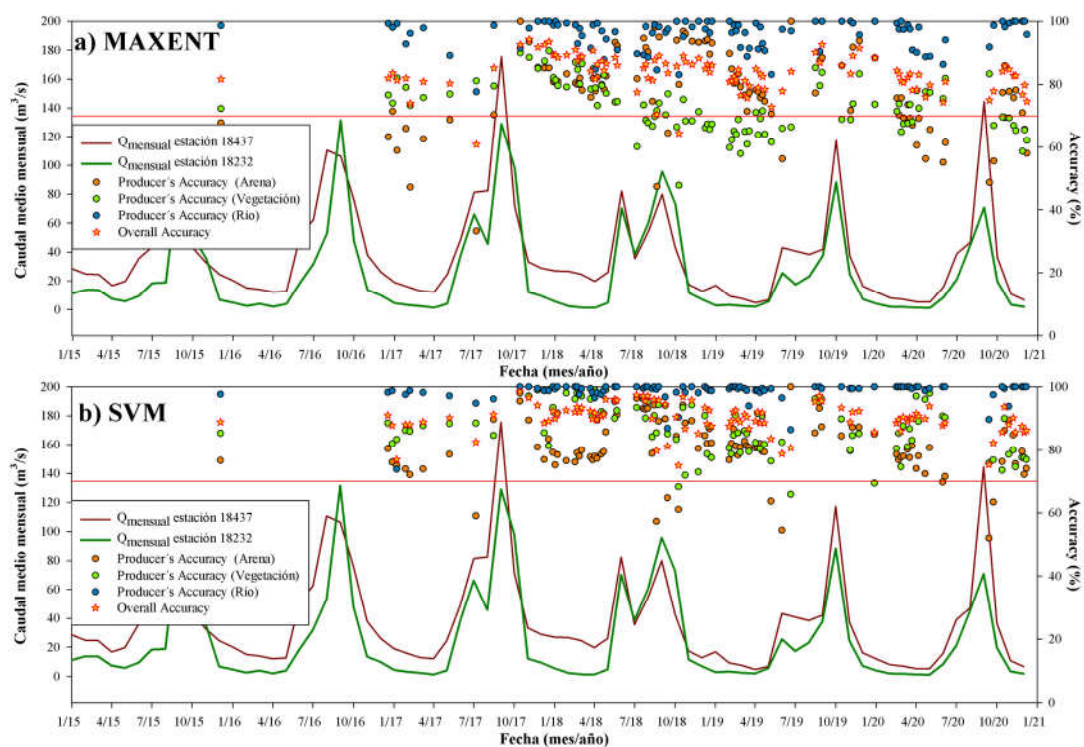


FIGURA 3.

PARÁMETROS DE CALIDAD DE CLASIFICACIÓN CON MAXENT OBTENIDOS A PARTIR DE MATRIZ DE CONFUSIÓN

Los valores arrojados por las curvas ROC, es decir los valores de AUC se muestran resumidos para ambos algoritmos en la Figura 4. En la misma se puede observar cómo la clasificación de la vegetación es la que mejor logran identificar los algoritmos, seguido por la del río y por último la de arena. Cabe resaltar que SVM presenta, para cada una de las clases, valores superiores a los reportados con el *MaxEnt*. Guiándonos por el valor límite escogido anteriormente se determinó que el algoritmo SVM muestra menos valores por debajo de éste, lo que representa una mejor clasificación. La comparación entre los algoritmos de clasificación *MaxEnt* y SVM permitió explorar el comportamiento del río desde dos puntos diferentes. Los resultados más cercanos a la realidad fueron obtenidos de la clasificación a partir del SVM.

En este estudio, las clasificaciones obtenidas por el *MaxEnt* con la metodología llevada a cabo por los autores, se presentó ligeramente inferior a la obtenida por el SVM. Cabe resaltar que en otros estudios (Li & Guo, 2010) que realizan comparaciones entre éstos, pero con una sola clase, el *MaxEnt* ha obtenido resultados más certeros. Sin embargo, esos resultados son de gran relevancia debido a los escasos estudios realizados respecto del tema,

además de que sienta las bases para determinar la dinámica espacial del río Amacuzac y sus zonas riparias.

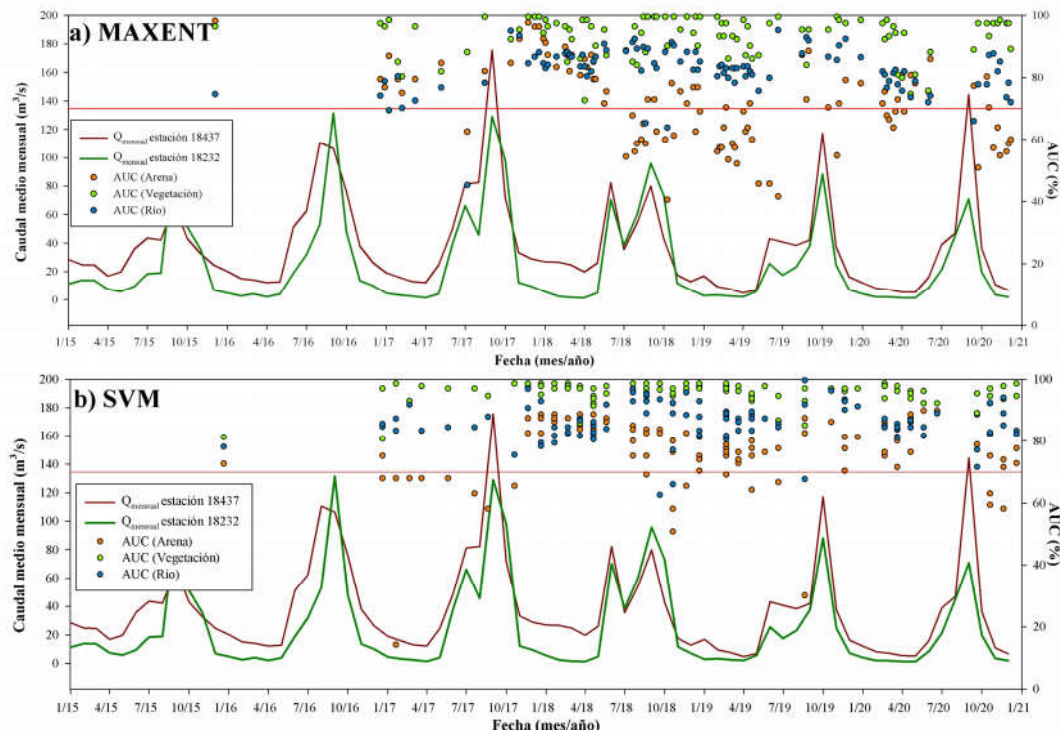


FIGURA 4

VALORES DE AUC DE LAS CLASIFICACIONES CON SVM Y MAXENT A PARTIR DE LA CURVA ROC.

CONCLUSIONES

Con el análisis de imágenes de satélites Sentinel 2 a 10m de resolución espacial, se logró obtener una buena precisión en la clasificación del área de estudio, logrando identificar cada una de las clases para la zona de interés.

Además, se comprobó que el *MaxEnt* puede ser utilizado como un clasificador multiclase con los parámetros predeterminados con los que trabaja, sin embargo, el SVM mostró una mayor precisión. Los resultados obtenidos se pueden deber únicamente a la metodología llevada a cabo, sin embargo, en futuras investigaciones se recomienda explorar los parámetros de entrada del *MaxEnt* y con ello quizás esperar resultados más acertados.

REFERENCIAS

- Chen, W., Pourghasemi, H.R., Kornejady, A. & Zhang N. (2017). Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. *Geoderma*, 305, 314–32. <http://doi.org/10.1016/j.geoderma.2017.06.020>
- Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, 37(1) 35–46. [https://doi.org/10.1016/0034-4257\(91\)90048-B](https://doi.org/10.1016/0034-4257(91)90048-B)
- Cortes, C., & Vapnik, V. (2006). Support-vector networks, *Machine learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- DOF. (2011). Acuerdo por el que se dan a conocer los estudios técnicos de aguas nacionales superficiales de la Región Hidrológica número 18 Balsas. Diario Oficial de la Federación, México. <https://dof.gob.mx>.
- Farrell, A., Wang, G., Rush, S. A., Martin, J.A., Belant, J. L., Butler, A. B., & Godwin, D. (2019). Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data. *Ecology and evolution*, 9(10), 5938–5949. <https://doi.org/10.1002/ece3.5177>
- Gualtieri, J. A. (2009). The support vector machine (SVM) algorithm for supervised classification of hyperspectral remote sensing data. En G. Camps-Valls y L. Bruzzone (Eds.), *Kernel methods for remote sensing data analysis*, 49–83, Singapore, John Wiley & Sons, Ltd.
- Krawczyk, B., Galar, M., Woźniak, M., Bustince, H. & Herrera, F. (2018). Dynamic ensemble selection for multi-class classification with one-class classifiers. *Pattern Recognition*, 83, 34–51. <https://doi.org/10.1016/j.patcog.2018.05.015>
- Li, W. & Guo, Q. (2010). A maximum entropy approach to one-class classification of remote sensing imagery. *International Journal of Remote Sensing*, 1(8), 227–2235. <https://doi.org/10.1080/01431161003702245>
- Mack, B. & Waske, B. (2017). In-depth comparisons of MaxEnt, biased SVM and one-class SVM for one-class classification of remote sensing data. *Remote sensing letters*, 8(3), 290–299. <http://doi.org/10.1080/2150704x.2016.1265689>
- Mejía-Mojica, H. (2012). Estructura y dinámica de una comunidad de peces bajo la influencia de un componente invasor.
- Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190 (3–4), 231–259. <https://doi.org/10.1016/j.ecolmodel.2005.03.026>

- Pratiwi, Y. & Widodo, E. (2017). Comparison of Maximum Entropy and Support Vector Machine Methods for Sentiment Analysis of Peralite Product through Twitter Social Network. *International Journal of Advances in Electronics and Computer Science*, 4(12), 10–14.
- Rahimian Boogar, A., Salehi, H., Pourghasemi, H. R. & Blaschke, T. (2019). Predicting habitat suitability and conserving *Juniperus* spp. habitat using SVM and maximum entropy machine learning techniques. *Water*, 11(10), 2049. <https://doi.org/10.3390/w11102049>
- Soares, D. (2008). Recursos hídricos de la cuenca del Amacuzac. *Inventio, La génesis de la cultura universitaria en Morelos*, 4(8). 15-23.
- Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1), 77–89. [https://doi.org/10.1016/S0034-4257\(97\)00083-7](https://doi.org/10.1016/S0034-4257(97)00083-7)