EVALUACIÓN DE LOS MODOS FUNCIONALES DEL MÉTODO DE MÁQUINA DE SOPORTE VECTORIAL PARA LA CLASIFICACIÓN DEL USO DE SUELO EN LA CUENCA CACHI, REGIÓN AYACUCHO, 2019

Wilmer E. Moncada Sosa, Alex M. Pereda Medina, Lidia J. Verde Rodríguez
Unidad de Investigación e Innovación de Ingeniería de Minas, Geología y Civil
Programa: Física Aplicada- Área de Biofísica
E-mail: wilmer.moncada@unsch.edu.pe

RESUMEN

La presente investigación evalúa los modos funcionales del método de Máquina de Soporte Vectorial (MSV) para la clasificación del uso de suelo en la cuenca Cachi de la Región Ayacucho, durante el año 2019. La corrección de las imágenes Sentinel 2 se realizaron con la herramienta Sen2Cor del software SNAP. Los polígonos de clasificación de las zonas de uso de suelo se hizo con ayuda de la firma espectral de cada píxel según el valor de reflectancia y longitud de onda de cada banda. El reconocimiento de los distintos tipos de uso de suelo se logró mediante la aplicación de la MSV, para lo cual se requiere un conjunto de datos de entrenamiento, los que se etiquetan como clases para después construir un modelo que prediga una nueva muestra. Para ello, se utilizó el paquete Orfeo Toolbox en QGIS, siendo el modo funcional lineal el algoritmo más óptimo para la clasificación. Los resultados muestran que los suelos agrícolas son las áreas de mayor cobertura ocupando un área de 89246.83 ha equivalente al 24.66 % del área total de la cuenca Cachi, seguido del suelo desnudo con 85298.82 ha equivalente a 23.57 %, las áreas de suelo con vegetación tienen un porcentaje de cobertura de 19.89 %, las áreas de menor cobertura son la clase nieve con 0.09 % y la clase agua con 0.32 %.

Palabras clave: Sentinel 2. Máquina de Soporte Vectorial. Uso de Suelo. Orfeo.

EVALUATION OF THE FUNCTIONAL MODES OF THE VECTOR SUPPORT MACHINE METHOD FOR LAND USE CLASSIFICATION IN THE CACHI BASIN, AYACUCHO REGION, 2019

ABSTRACT

This research evaluates the functional modes of the Vector Support Machine (VSM) method for land use classification in the Cachi basin of the Ayacucho Region, during 2019. The correction of the Sentinel 2 images was done with the Sen2Cor tool of the SNAP software. The classification polygons of the land use zones were made with the help of the spectral signature of each pixel according to the reflectance value and wavelength of each band. The recognition of the different types of land use was achieved through the application of the MSV, for which a set of training data is required, which are labelled as classes in order to later build a model that predicts a new sample. For this purpose, the Orfeo Toolbox package in QGIS was used, being the linear functional mode the most optimal algorithm for classification. The results show that agricultural soils are the areas of greatest coverage occupying an area of 89246.83 ha equivalent to 24.66 % of the total area of the Cachi basin, followed by bare soil with 85298.82 ha equivalent to 23.57 %, vegetated soil areas have a percentage of coverage of 19.89 %, the areas of least coverage are the snow class with 0.09 % and the water class with 0.32 %.

Keywords: Sentinel 2. Vector Support Machine. Use of Soil. Orpheus.

INTRODUCCIÓN

En la actualidad, la cuenca del rio Cachi en la Región Ayacucho, a experimentado una disminución en la intensidad de lluvias e incremento en la temperatura máxima, repercutiendo directamente en la vegetación, humedad del suelo y uso de suelo, los que a su vez impactan negativamente sobre la calidad de suelo (erosión hídrica y de suelos, degradación, pérdida de vegetación) por aumento de la velocidad de los vientos, por nombrar algunos (Moncada et al., 2015). La cuenca Cachi no es ajena a la intervención de la mano del hombre y al cambio global, por lo que es de suma importancia monitorear la cobertura de espacial de los cuerpos de agua para la agricultura, consumo humano de centros poblados y actividades agrícolas-ganaderas productivas (Henao, 1988). Una de las herramientas de mejor uso y de fácil acceso son las imágenes Sentinel 2 que resolución media y mejor aún son de libre disponibilidad (European Space Agency, 2017). En el presente trabajo de investigación se da respuesta a la pregunta ¿En qué medida los modos funcionales del método de máquina de soporte vectorial permiten la clasificación del uso de suelo

en la cuenca Cachi, en la Región Ayacucho, durante el periodo 2019?, de esta manera se plantea una metodología que consiste en aprovechar los distintos modos funcionales con el uso del método de máquina de soporte vectorial para la clasificación del uso de suelo, cuyos datos son mapeados por medio de un kernel Gaussiano u otro tipo de kernel a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases (Bentancourt, 2005).

La investigación tiene como propósito evaluar los modos funcionales del método de máquina de soporte vectorial para la clasificación del uso de suelo en la cuenca Cachi de la Región Ayacucho, durante el periodo 2019. El reconocimiento de distintos tipos de coberturas de uso de suelo como la función de decisión que se mueven hacia la línea punteada de su propio lado, dando la certeza de que es posible encontrar un conjunto infinito de hiperplanos que clasifiquen correctamente los datos de entrenamiento. Sin embargo, es claro que la precisión de clasificación al generalizar será directamente afectada por la posición de las funciones de decisión (Castellón, 2015). El método de máquina de soporte vectorial a diferencia de otros métodos de clasificación considera esta desventaja y encuentra la función de decisión de tal forma que la distancia entre los datos de entrenamiento es maximizada. Esta función de decisión es llamada función de decisión óptima o hiperplano de decisión óptima (Cristianini et al., 2000). En una máquina de soporte vectorial, el hiperplano óptimo es determinado para maximizar su habilidad de generalización. Pero, si los datos de entrenamiento no son linealmente separables, el clasificador obtenido puede no tener una alta habilidad de generalización, aun cuando los hiperplanos sean determinados óptimamente, para maximizar el espacio entre clases, el espacio de entrada original es transformado dentro de un espacio altamente dimensional llamado “espacio de características” (Cervantes, 2009). Se espera que el método utilizado sea una alternativa rápida, repetible, evaluable para clasificar las clases de uso de suelo en la cuenca Cachi de la Región Ayacucho, para lo cual se requiere comprender "las acciones, actividades e intervenciones que realizan las personas sobre un determinado tipo de superficie para producir, modificarla o mantenerla". Abarca la gestión y modificación del medio ambiente natural para convertirlo en terreno agrícola: campos cultivables, pastizales; o asentamientos humanos. El término uso del suelo también se utiliza para referirse a los distintos usos del terreno en zonificaciones (Guttenberg, 1959).

MATERIALES Y MÉTODOS

La figura 1 muestra la ubicación geográfica de la cuenca Cachi en la Región Ayacucho, su principal afluente el rio Cachi se forma por el aporte de los ríos Apacheta, Choccoro y Chicllarazo que dan lugar al rio Vinchos, este a su vez se junta con el río Paccha para formar el río Cachi. El río Cachi recibe el caudal de los ríos Chillico, Huatatas y Yucaes. Tiene un área de 361873.8435 ha. La posición de su centroide en UTM es latitud 575770.284 y longitud 8537945.565, con una altitud mínima media de 2434 ms.n.m y una altitud máxima media de 4728 ms.n.m.

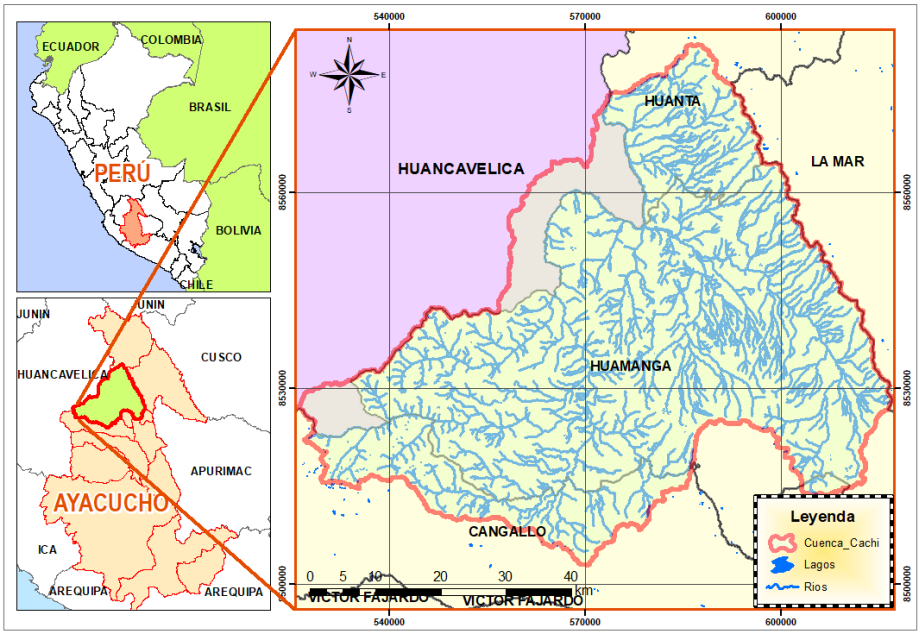


Figura 1: Ubicación Geográfica de la cuenca Cachi, Región Ayacucho.

Las imágenes ópticas Sentinel 2 se procesan con el software SNAP, para la corrección atmosférica de las imágenes se hace uso de la herramienta Sen2Cor, la cual permite generar valores de reflectancia entre 0 y 1. Este rango de

valores permite la detección de cambios y el establecimiento de un valor de reflectancia para la longitud de onda de cada banda de la imagen de satélite Sentinel 2 con resolución de 10, 20 y 60 m por píxel, resampleadas a 10 m.
En primer término, la metodología involucra la aplicación del método de clasificación máquina de soporte vectorial, el cual contiene distintos modos funcionales. Para ello se considera un conjunto S de puntos (x_i, y_i) etiquetados para entrenamiento. Los datos son linealmente separables si existen diferentes hiperplanos que pueden realizar la separación, la línea que separa el espacio de entrada es definida por la ecuación $w \cdot z + b = 0$, donde w define el hiperplano de separación óptimo y b es el sesgo (Cristianini et al., 2000). En la mayoría de las clases, no únicamente se traslapan o interceptan los datos al generar un hiperplano de separación, sino que la separación genuina de estos datos está dada por hiper superficies no lineales. Una característica del enfoque presentado anteriormente radica en que éste, puede ser fácilmente extendido para crear cotas de decisión no lineal (Vapnik, 1999a). El motivo de tal extensión es que una máquina de soporte vectorial puede crear una hiper superficie de decisión no lineal, capaz de clasificar datos separables no linealmente, el análisis previo puede ser generalizado introduciendo algunas variables no-negativas $\xi_i \geq 0$ de tal modo que la ecuación anterior se modifica como $y\,(w \cdot z + b) \geq 1 - \xi_i$.

En una máquina de soporte vectorial, el hiperplano óptimo es determinado para maximizar su habilidad de generalización, pero si los datos de entrenamiento no son linealmente separables, el clasificador obtenido puede no tener una alta habilidad de generalización, por lo tanto, la regla de decisión puede ser evaluada usando productos punto $f(x) = \sum_{i=1}^{l} \alpha_i y_i \langle \Phi(x_i) \cdot \Phi(x) \rangle + b$, siempre y cuando se tiene una forma de capturar el producto $\langle \Phi(x_i) \cdot \Phi(x) \rangle$ en el espacio de características, directamente como una función de los puntos de entrada originales, esto hace posible unir los dos pasos necesarios para construir una máquina de aprendizaje no-lineal. A este método de cómputo directo se le llama función kernel (Vapnik, 1999b).

Con las imágenes Sentinel 2, se pretende realizar la clasificación de las diferentes clases de uso de suelo en la cuenca Cachi, para lo cual se tomará en cuenta algunos tipos de uso de suelo tales como cuerpos de agua, suelo desnudo, suelo con vegetación suelo agrícola, zonas urbanas, zonas rurales, bofedales, vías asfaltadas, nieve, sombras y zonas sin clasificación. Estas muestras de zonas se almacenan en un archivo con formato shape como regiones de interés, las mismas que son requeridas para el entrenamiento de la máquina de soporte vectorial y validación del producto. La metodología propone la aplicación del paquete Orfeo Toolbox con el software QGIS, primero aplicando el algoritmo “Compute Images Statistics” para la creación de los archivos XML. Segundo, aplicando el algoritmo “Train SVM Image Classification”, con la aplicación de este algoritmo se crean los archivos Matriz de confusión en formato csv y el modelo de salida en formato txt. 7. Tercero, se aplica el algoritmo “Create Image Classification” en donde se ingresan la información construida con la aplicación de los dos algoritmos anteriores, de esta manera se crea un archivo raster en formato tif con las áreas debidamente clasificadas.

RESULTADOS Y DISCUSIÓN

1. El algoritmo MSV, tiene varios componentes de entrada como los polígonos de clases y una sola componente de salida que es la imagen clasificada, donde los componentes de entrada se pueden agrupar en tres categorías distintas: Componentes de entrada relacionada con la imagen de satélite Sentinel 2 a clasificar. El algoritmo MSV tiene dos entradas de este tipo, la primera es la aplicación de lectura de la imagen y la segunda es la aplicación de lectura del conjunto de muestras (Polígonos de clases para el entrenamiento) los cuales son seleccionados desde la propia imagen de satélite previamente preprocesada, tomando en cuenta varias muestras de cada clase para diferentes clases. Dichas muestras están formadas por pixeles representativos que componen el denominado conjunto de entrenamiento (tomadas tomando en cuenta la firma espectral de cada píxel), sobre el que se basa el proceso de clasificación supervisada. Por lo tanto, en este método, el conocimiento que se posee sobre el área de estudio determina la calidad, tanto del conjunto de entrenamiento como de la tasa de acierto obtenida por parte del proceso de clasificación. Este algoritmo por ser de tipo supervisado compara cada píxel de la imagen con estas firmas elegidas y, a continuación, cada píxel es etiquetado en la clase a la que más se asemeja espectralmente obtenido previamente a partir de pixeles que pertenecen a las clases elegidas para el proceso de clasificación. Estos dos componentes se encuentran almacenados dentro una base de datos espacial y proporcionan al algoritmo MSV todos los elementos necesarios para realizar una clasificación de la imagen de tipo supervisada. Componentes de entrada del algoritmo MSV relacionados con los demás paquetes de procesamiento digital de imágenes con SNAP, para de esta manera producir puntos de comparación y evaluación de las diferentes clasificaciones. El algoritmo MSV utiliza los componentes de entrada y muestras, pero se debe aclarar que estos no se encuentran almacenados en una base de datos espacial. Componentes de entrada del algoritmo MSV relacionados con el tipo clasificación realizada que para este caso se utilizó como temática de clasificación los usos del suelo y como modelo piloto de la herramienta el modelo linealmente

separable el cual se apoya en la generación de hiperplanos de separación para realizar la clasificación. Algunos de estos componentes relacionados con MSV deben ser configurados por el analista antes de llevar a cabo el proceso de clasificación, como el tipo de separación de hiperplanos (lineal, no lineal) o la temática a trabajar dentro de la clasificación, personalizando la configuración del algoritmo MSV para cada imagen concreta.

2. Se propone un algoritmo de clasificación de imágenes supervisado, basado en el algoritmo de clasificación máquinas de soporte vectorial. El algoritmo se implementó directamente sobre una base de datos espaciales aprovechando las propiedades del ráster definidos en las imágenes, con el propósito de implementar un proceso de clasificación supervisado espectral. Por lo tanto, en esta etapa se aplicó el algoritmo planteado en la etapa de desarrollo sobre una imagen satelital Sentinel 2 para la cuenca Cachi perteneciente a la Región Ayacucho para lo cual se definió como temática para la clasificación las distintas clases de uso de suelo, se analizó el comportamiento del algoritmo comparándolo con los paquetes de procesamiento digital de imágenes con el software libre SNAP, y se establecieron las ventajas y desventajas de cada uno de los algoritmos y se comparó a través de cada matriz de confusión y coeficiente Kappa el rendimiento de cada clasificador mediante el software QGIS. En esta última fase se asigna la clase a las diferentes clases de uso de suelo que se van a reconocer. Una vez descritos los objetos según todas las características, es necesaria la asignación de los objetos a una de las clases de la leyenda.

3. Entrenamiento: Para el entrenamiento de los datos se requiere el vector de características de cada una de las coberturas a clasificar, siguiendo un determinado proceso. Para los datos de entrada que pertenecen a una cierta clase, sus respectivos vectores de características se colocan en una matriz en forma de columnas. Cabe mencionar que se deben colocar primero todos los datos que pertenecen a una clase y después todos los que pertenecen a la siguiente clase. Para cada dato de entrada x_i le corresponde una salida y_i por lo que tenemos una pareja (x_i, y_i). En este sentido se debe indicar al método de entrenamiento cuál es la salida y_i que le corresponde a cada entrada x_i. La figura 02 muestra los polígonos de clases de entrenamiento utilizados durante el proceso de clasificación:

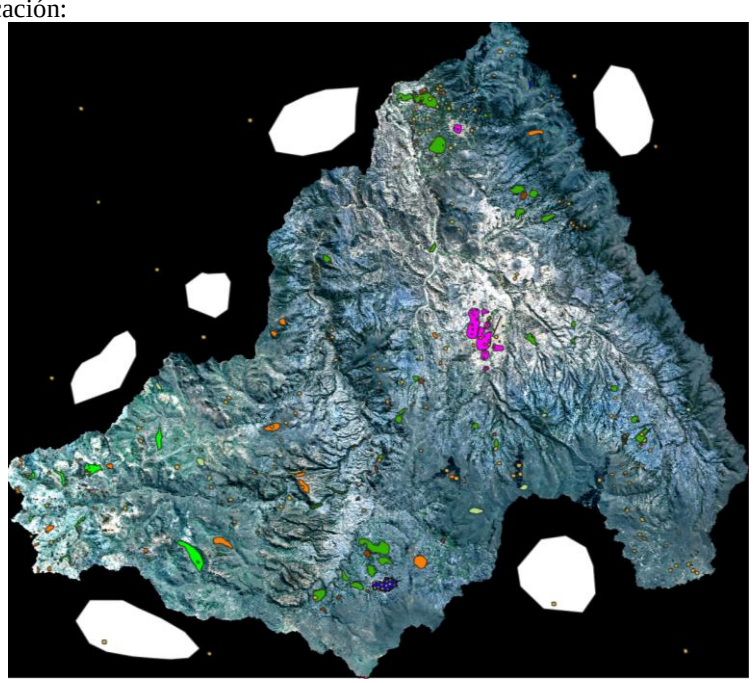
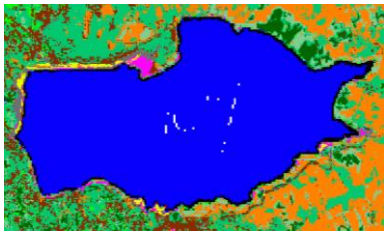


Figura 2: polígonos de clases de uso de suelo en la cuenca Cachi utilizados para el entrenamiento durante el proceso de clasificación.

4. Clasificación: los datos deben estar bien entrenados, es decir, si el vector de características de cada cobertura efectivamente es diferente a los demás, la clasificación se hará de manera correcta, siendo el modo funcional lineal el que mejor resultados muestra en el momento de realizar la clasificación de uso de suelo en la cuenca Cachi. El clasificador lineal empleado obtiene la línea (para 2 dimensiones o el hiperplano para un mayor número de dimensiones) que separe limpiamente las dos clases maximizando la distancia a la frontera de los ejemplos más próximos a la misma. El algoritmo es muy eficiente incluso para cientos de dimensiones, ya que el separador lineal puede tener únicamente en cuenta los puntos más próximos y descartar los más lejanos a la frontera. Para efecto de la clasificación del uso de suelo en la cuenca Cachi se adoptaron las categorías de cobertura de la. Los distintos niveles y clases de este sistema de clasificación de coberturas de la Tierra se adaptan a las necesidades de identificación de categorías de uso. El cuadro 01, muestra las características de las

12 clases de uso de suelo en la cuenca Cachi, clasificadas en la imagen de satélite Sentinel 2, cuyas bandas han sido resampleadas a 10 m.

Cuadro 1. Clases de uso de suelo en la cuenca Cachi, mediante el método de máquina de soporte vectorial con imágenes de satélite Sentinel 2.

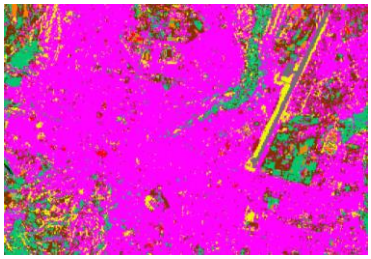

Clase 1: Cuerpos de agua

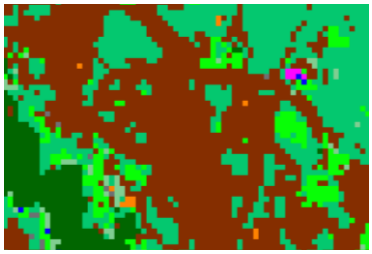

Clase 2: Suelo Desnudo

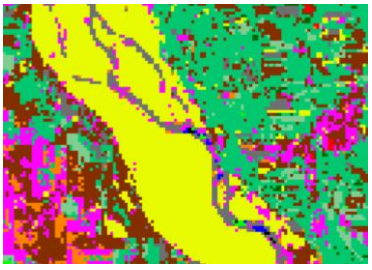

Clase 3: Suelo con Vegetación

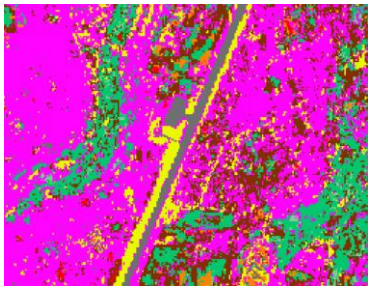

Clase 4: Suelo Agrícola

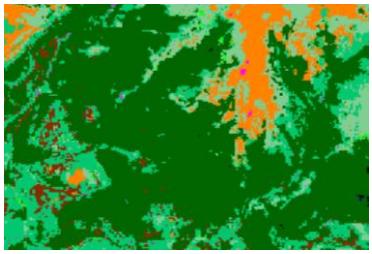

Clase 5: Bofedal

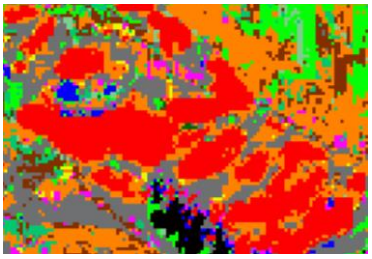

Clase 6: Zona Urbana

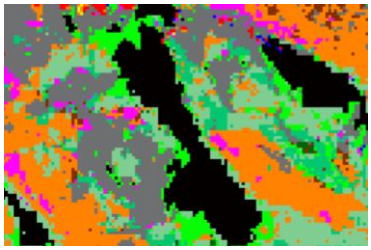

Clase 7: Zona Rural


Clase 8: Arena de Rio


Clase 9: Vía Asfaltada


Clase 10: Bosque Denso


Clase 11: Nieve


Clase 12: Sombra

5. Validación de la clasificación de imágenes: existen dos posibilidades, evaluar una estimación teórica del error en función de las características del algoritmo de clasificación o analizar una serie de pruebas de validación obtenidas del mismo modo que las áreas de entrenamiento. Aquí se ha implementado el segundo modo, ya que permite obtener una estimación más realista de los errores mientras la muestra de pixeles para la estimación del error sea lo suficientemente grande y representativa. Para la evaluación de los errores se utiliza una matriz de confusión de clases ya que, con este tipo de análisis, se obtiene, no sólo una caracterización del error cometido, sino también una medida sobre la adecuación de la clasificación considerada a la realidad y de los parámetros utilizados para caracterizarlas.

6. La figura 03 muestra el mapa de clasificación del uso de suelo en la cuenca Cachi mediante el método de máquina de soporte vectorial con imágenes de satélite Sentinel 2.

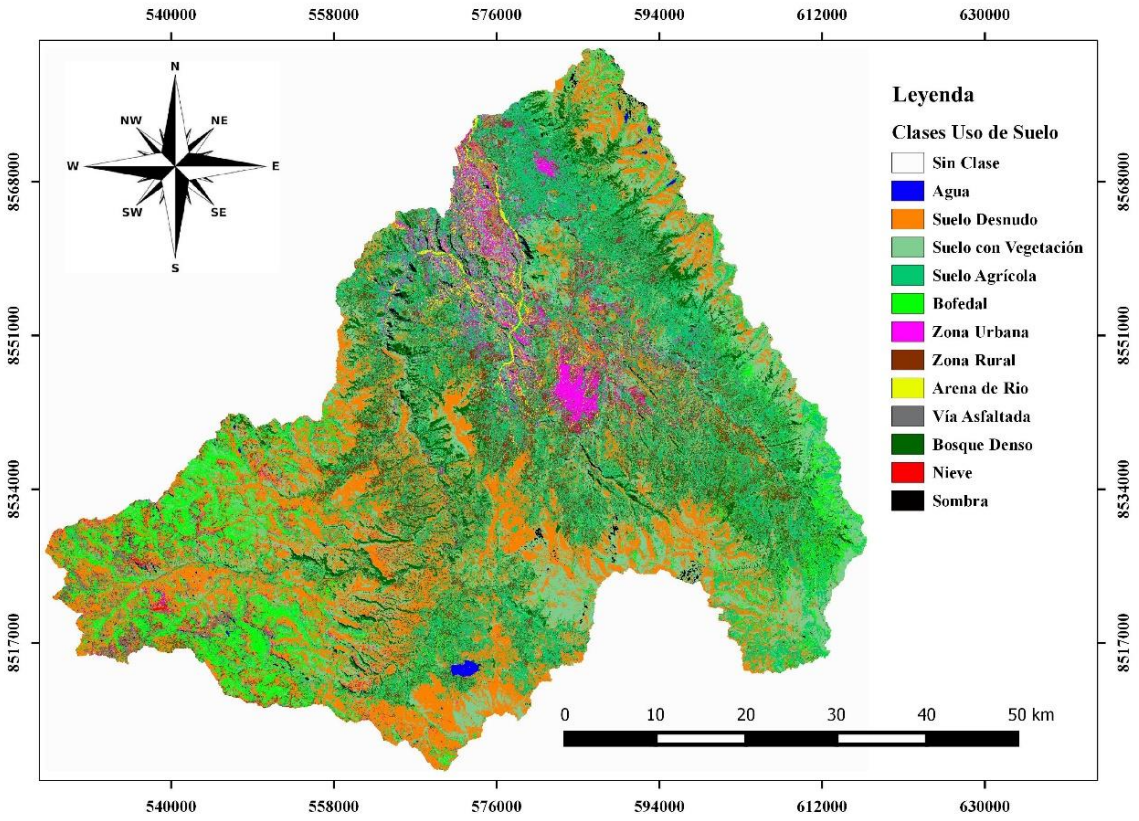


Figura 3. Mapa de uso de suelo en la cuenca Cachi mediante el método de máquina de soporte vectorial con imágenes de satélite Sentinel 2, durante el año 2019.

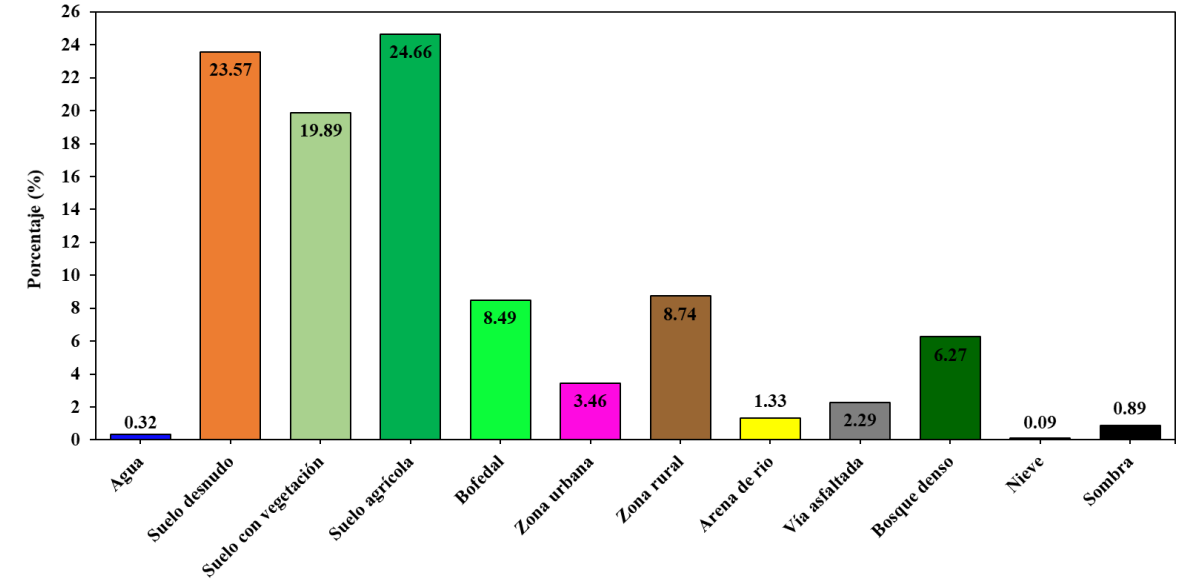
7. La tabla 1 muestra el número de píxeles clasificados para cada clase de uso de suelo en la cuenca Cachi mediante el método de máquina de soporte vectorial en imágenes de satélite Sentinel 2, el cual es multiplicado por el área de cada píxel que es 100 m^2 y dividido entre 10000 para llevar el área en unidades de hectáreas, luego considerando que el área de la cuenca Cachi es 361873.8435 ha se expresa la cobertura espacial de cada clase de uso de suelo en la cuenca Cachi en valores de porcentaje.

Tabla 4.1: Cobertura espacial de las clases de uso de suelo en la cuenca Cachi mediante el método de máquina de soporte vectorial en imágenes de satélite Sentinel 2, durante el año 2019

Número	Clase	Número de Píxeles	Área (ha)	%
0	Sin Clase	36512157	365121.57	
1	Agua	116134	1161.34	0.32
2	Suelo desnudo	8529882	85298.82	23.57
3	Suelo con vegetación	7197716	71977.16	19.89
4	Suelo agrícola	8924683	89246.83	24.66
5	Bofedal	3072559	30725.59	8.49
6	Zona urbana	1252806	12528.06	3.46
7	Zona rural	3162368	31623.68	8.74
8	Arena de rio	480863	4808.63	1.33
9	Vía asfaltada	827710	8277.1	2.29
10	Bosque denso	2268916	22689.16	6.27
11	Nieve	33433	334.33	0.09
12	Sombra	320405	3204.05	0.89

El gráfico 1 muestra el porcentaje de cobertura espacial de uso de suelo en la cuenca Cachi mediante el método de máquina de soporte vectorial en imágenes de satélite Sentinel 2 durante el año 2019.

Gráfico 01: Porcentaje de cobertura de uso de suelo en la cuenca Cachi mediante el método de máquina de soporte vectorial en imágenes de satélite Sentinel 2, durante el año 2019



Se observa que los suelos agrícolas son las áreas de mayor cobertura en la cuenca Cachi ocupando un área de 89246.83 ha equivalente al 24.66 % del área total de la cuenca Cachi, seguido del suelo desnudo con un área de 85298.82 ha equivalente a 23.57 %, las áreas de suelo con vegetación tienen un porcentaje de cobertura de 19.89 %, las áreas de menor cobertura en la cuenca Cachi son la nieve con un 0.09 % y el agua con un 0.32 %.

AGRADECIMIENTOS

A las personas que en forma desinteresada aportaron en el avance del presente trabajo de investigación, al Laboratorio de Teledetección y Energías Renovables LABTELER de la Escuela Profesional de Ciencias Físico Matemáticas de la Universidad Nacional de San Cristóbal de Huamanga por permitirnos hacer uso de sus ambientes y equipos. En especial a la Oficina General de Investigación e Innovación de la Universidad Nacional de San Cristóbal de Huamanga por haber hecho posible su realización, quienes han aportado económicamente en su ejecución sin lo cual no hubiera sido posible su desarrollo.

REFERENCIAS BIBLIOGRÁFICAS

Bentancourt, G., 2005. Máquinas de Soporte Vectorial. Sci. Tech. XI, 67–72.

Castellón, J., 2015. Análisis comparativo entre ENVI y Orfeo Toolbox SVM. ResearchGate 9. http://dx.doi.org/10.13140/RG.2.1.1991.1844

Cervantes, J., 2009. Clasificación de grandes conjuntos de datos vía Máquinas de Vectores Soporte y aplicaciones en sistemas biológicos (Tesis Doctoral). Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México.

Cristianini, N., Shawe-Taylor, J., Shawe-Taylor, D. of C.S.R.H.J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge. University Press.

European Space Agency, 2017. Sentinel-2 [WWW Document]. Eur. Space Agency. URL http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-2 (accessed 11.8.17).

Guttenberg, A.Z., 1959. A Multiple Land Use Classification System. J. Am. Inst. Plann. 25, 143–150. https://doi.org/10.1080/01944365908978322

Henao, J., 1988. Introducción al manejo de cuencas hidrográficas. Universidad Santo Tomas, Bogota.
Moncada, W., Pereda, A., Aldana, C., Masias, M., Jimenez, J., 2015. Cuantificación hidrográfica de la cuenca del rio Cachi-Ayacucho, mediante imágenes satelitales. Inst. Investig. Científica E Innov. Tecnológica UNSCH II.
Vapnik, V., 1999a. An overview of statistical learning theory. IEEE Trans. Neural Netw. 10, 12. https://doi.org/10.1109 / 72.788640
Vapnik, V., 1999b. The Nature of Statistical Learning Theory, Segunda. ed. Springer, New York.