# APLICACIÓN DE UN ALGORITMO DE CLUSTERING A IMÁGENES SATELITALES PARA EL ANÁLISIS DE BIOMASA EN EL DEPARTAMENTO DE NARIÑO

**ALISON GIOVANNA BASTIDAS CHITÁN**

**ANDREA LORENA BRAVO SUÁREZ**

**UNIVERSIDAD DE NARIÑO**
**FACULTAD DE INGENIERÍA**
**DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA**
**SAN JUAN DE PASTO**
**2017**

# APLICACIÓN DE UN ALGORITMO DE CLUSTERING A IMÁGENES SATELITALES PARA EL ANÁLISIS DE BIOMASA EN EL DEPARTAMENTO DE NARIÑO

**ALISON GIOVANNA BASTIDAS CHITÁN**

**ANDREA LORENA BRAVO SUÁREZ**

**TRABAJO DE GRADO PARA OPTAR POR EL TÍTULO DE INGENIERO ELECTRÓNICO**

**DIRECTOR**

**PhD. ANDRÉS DARIO PANTOJA BUCHELI**

**INGENIERO ELECTRÓNICO**

**UNIVERSIDAD DE NARIÑO**
**FACULTAD DE INGENIERÍA**
**DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA**
**SAN JUAN DE PASTO**
**2017**

## NOTA DE RESPONSABILIDAD

“Las ideas y conclusiones aportadas en el siguiente trabajo son responsabilidad exclusiva del autor”.

Acuerdo 1. Artículo 324. Octubre 11 de 1966, emanado del honorable Consejo Directivo de la Universidad de Nariño.

**NOTA DE ACEPTACIÓN:**

_____
_____
_____
_____
_____
_____

_____
Firma del presidente del jurado

_____
Firma del jurado

_____
Firma del jurado

San Juan de Pasto, 24 de noviembre de 2017

## RESUMEN

La información proporcionada por la reflectancia en diferentes bandas de las imágenes satelitales es analizada para obtener características relevantes en amplias áreas de territorio. En este trabajo se evalúan distintas técnicas de clustering para la determinación de tipos de cobertura vegetal en el departamento de Nariño a partir del procesamiento de una base de datos de imágenes libres, con el fin de aportar con información relevante de tipos de biomasa presentes en la región.

En el desarrollo de la investigación se realizan etapas de pre-procesamiento, procesamiento, análisis de componentes principales, agrupamiento de pixeles, validación y caracterización de tipos de cobertura. Se comparan algoritmos de clustering como K-Means, EM e Isodata mediante índices de desempeño en una zona característica y se extrapola los mejores resultados a un mapa de clasificación en todo el departamento teniendo en cuenta la información de un mapa teórico de la región.

A partir de los resultados obtenidos se evidencia que el algoritmo K-Means es robusto, arrojó los mejores resultados con los datos trabajados y se pudo adaptar con facilidad a los requerimientos establecidos. En cuanto a la caracterización obtenida se destaca que el 48% de superficie territorial del departamento corresponde a Bosque. De esta manera, las principales contribuciones de este trabajo son; un mapa de caracterización de biomasa en el departamento de Nariño, la adaptación de algoritmos de clustering al caso de estudio y el aporte al análisis de oportunidades energéticas en la región.

## ABSTRACT

The information provided by the reflectance in different bands of the satellite images is analyzed to obtain relevant characteristics in broad areas of territory. In this paper different clustering techniques are evaluated for the determination of types of plant cover in the department of Nariño from the processing of a database of free images, in order to provide with relevant information of types of biomass present in the region.

In the development of the research, stages of pre-processing, processing, principal component analysis, grouping of pixels, validation and characterization of types of coverage are performed. Clustering algorithms such as K-Means, EM and Isodata are compared by performance indices in a characteristic area and the best results are extrapolated to a classification map in the whole department taking into account the information of a theoretical map of the region.

Based on the results obtained, it is evident that the K-Means algorithm is robust, it showed the best results with the worked data and it was able to adapt easily to the established requirements. Regarding the characterization obtained, it is highlighted that 48% of the territorial area of the department corresponds to Bosque. In this way, the main contributions of this work are; a map of biomass characterization in the department of Nariño, the adaptation of clustering algorithms to the case of study and the contribution to the analysis of energy opportunities in the region.

## TABLA DE CONTENIDO

## LISTA DE TABLAS

## LISTA DE FIGURAS

## GLOSARIO

**BASES DE DATOS:** información que incluye coordenadas geográficas, índices, valores de reflectancia, resultados de las clasificaciones, entre otros aspectos, que se almacenan en tablas en el servidor del grupo de investigación GIIEE. La base de datos fundamental contiene los pixeles que representan el departamento de Nariño (34’202.925 pixeles de resolución 30m × 30m), y cada pixel posee información de latitud, longitud y reflectancias para sus 7 bandas espectrales.

**BIOMASA:** conjunto de materia orgánica renovable de origen vegetal, animal o procedente de la transformación natural o artificial de la misma.

**CENTROIDE:** en este trabajo, centroide hace referencia a la media representativa de un cluster.

**CLUSTER:** en este trabajo, cluster hace referencia a un grupo de pixeles o clase de cobertura natural.

**CLUSTERING:** es un proceso que tiene como finalidad agrupar individuos que tienen un alto grado de semejanza y a su vez son muy diferentes respecto a los de otro grupo.

**EM:** es el acrónimo del algoritmo de clustering Expectation Maximization o Esperanza-Maximización.

**EPSG:** es un repositorio de parámetros geodésicos que contiene información sobre sistemas de referencia antiguos y modernos, proyecciones cartográficas y elipsoides de todo el mundo.

**FIRMA ESPECTRAL:** es la radiación reflejada en función de la longitud de onda.

**GK:** es el acrónimo del algoritmo de clustering Gustafson Kessel.

**IMÁGENES SATELITALES:** son la representación visual de la información capturada por un sensor montado en un satélite. Se consideran el insumo raíz del proyecto, se organizan como la fuente de información para consolidar las bases de datos y se usan como entrada a la ejecución de los algoritmos.

**ISODATA:** es el acrónimo del algoritmo de clustering Iterative Self-Organizing Data Analysis Techniques.

**MAPA DE COBERTURAS DE CORPONARIÑO (2015):** mapa del departamento de Nariño proporcionado por Corponariño, que identifica diferentes tipos de cobertura en la región.

**MATLAB:** herramienta de codificación y procesamiento a través de la que se ejecutan los algoritmos de clustering. Con los scripts desarrollados se analizan y se almacenan los resultados necesarios dentro de la investigación, requiriéndose la consulta de la documentación en su portal web, MathWorks.

**MSI:** es el Índice de Estrés Hídrico.

**NDVI:** es el Índice de Vegetación de Diferencia Normalizada.

**NIR:** es la región espectral del infrarrojo cercano.

**PostgreSQL:** sistema de gestión de bases de datos relacional orientado a objetos y libre. Principalmente se usa para ejecutar consultas que facilitan el uso y análisis de las bases de datos empleadas en la investigación.

**QGIS:** sistema de información geográfica donde se grafican los resultados obtenidos. Especialmente se usa para cargar un archivo *.csv* que contiene información de latitud, longitud y la clase a la que pertenece cada pixel.

**RADIANCIA:** medida radiométrica que describe la cantidad de luz que pasa a través o es emitida de un área particular.

**REFLECTANCIA:** relación entre la potencia electromagnética incidente con respecto a la reflejada en una interface o superficie.

**RESOLUCIÓN ESPACIAL:** es el tamaño de un píxel que corresponde con áreas cuadradas.

**RESOLUCIÓN ESPECTRAL:** es la amplitud de la longitud de onda de las diferentes frecuencias grabadas.

**RESOLUCIÓN RADIOMÉTRICA:** es la capacidad del sensor para distinguir diferentes intensidades de radiación.

**RESOLUCIÓN TEMPORAL:** es la frecuencia con la que el satélite sobrevuela una zona.

**SERVIDOR:** herramienta facilitada por el grupo de investigación GIIEE de la Universidad de Nariño, en el que se almacenan las bases de datos y el resultado final del presente trabajo de grado.

**SWIR:** infrarrojo de onda corta.

**TIRS:** hace referencia al sensor térmico infrarrojo en Landsat.

# INTRODUCCIÓN

La problemática ambiental actual involucra los recursos energéticos, su principal influencia radica en el consumo de energía proveniente de fuentes no renovables [1] y el desaprovechamiento de ciertos recursos naturales útiles para la generación de energía eléctrica. La caracterización de estos constituye el principal aporte para la proposición de alternativas en la creciente necesidad de diversificación de la matriz energética de los países [2]. En cuanto a la biomasa vegetal, la estimación de su potencial energético hace necesario estudiar y determinar los diferentes tipos de cobertura presentes en regiones amplias, que implican grandes cantidades de información, generalmente provenientes de imágenes satelitales, que requieren de técnicas computacionales eficientes para su procesamiento. Sin embargo, usualmente la determinación de estas clases implica el procesamiento de imágenes sobre áreas pequeñas, se presenta en [3].

La caracterización de las coberturas naturales es un problema abierto y altamente dependiente de los lugares de aplicación, puesto que la respuesta espectral útil de las fotografías para la determinación de los tipos de biomasa (i.e., valores de reflectancia), puede ser afectada por la evolución del estado fenológico de la vegetación, el clima presente cuando se capturan las fotos y la evolución temporal de los terrenos [4]. En este sentido, la presente investigación se desarrolla con el fin de contribuir en el análisis del potencial energético de biomasa del departamento de Nariño, que cuenta con abundantes recursos que no se encuentran debidamente clasificados, estudiados, analizados y aprovechados.

Diferentes métodos de agrupamiento de datos en clases congruentes (clustering) se usan como alternativa dentro del análisis de características de imágenes. En particular, los métodos con análisis no supervisado son relevantes en estas aplicaciones puesto que no requieren un etiquetado previo en el conjunto de individuos y el proceso de clasificación es más flexible [5]. Entre los algoritmos de clustering más utilizados en estudios similares se encuentran K-Means e Isodata [6], [7], en donde se clasifica el uso del suelo en diferentes regiones utilizando las propiedades de imágenes QuickBird y MODIS, respectivamente. Por su parte, los autores en [8] proponen la clasificación automática de cubiertas terrestres usando Region Growing y K-Means, mostrando la adaptabilidad de estos métodos a los requerimientos establecidos. Usando también este último método para obtener los mejores resultados, en [9] se identifican unidades boscosas utilizando imágenes Landsat. Por último, en [10] se aborda el problema de la segmentación de imágenes satelitales multiespectrales a partir de la aplicación del algoritmo Expectation-Maximization (EM), cuyo rendimiento se evalúa por medio del tiempo de cálculo y una medida de calidad de cluster.

Teniendo en cuenta que el resultado de los diferentes algoritmos de clustering depende en gran medida de la calidad de los datos de entrada, en este trabajo se propone una metodología para analizar imágenes en diferentes bandas de

Landsat 8, permitiendo la adaptación del clustering al caso de estudio. Debido a la extensión territorial de Nariño, los métodos se aplican inicialmente a una base de datos pre-procesada de reflectancias en un área representativa del departamento (municipio de Tumaco), determinando el método con mejor desempeño de acuerdo a los índices de cohesión y separación e indicador de comparación visual de las clases identificadas con mapas de referencia. Una vez definido el método más apropiado, se aplica a la base de datos de todo el departamento con el fin de establecer un mapa de alta resolución de clasificación de tipos de cobertura natural. El resultado final se contrasta con un mapa oficial de coberturas vegetales para determinar los tipos finales de biomasa, que pueden interpretarse para el estudio de su aprovechamiento energético en la región. Adicionalmente, se contribuye al planteamiento y desarrollo de nuevos trabajos en el área y especialmente, se constituye una base para continuar el análisis de tipos de biomasa en la región.

Los resultados más relevantes de la investigación se presentaron en el Congreso Internacional Multimedia 2017, llevado a cabo en el municipio de Cajicá-Cundinamarca, durante los días 28 y 29 de septiembre del año en curso. Se logró una gran aceptación por parte de los asistentes al evento, se reconoció el trabajo realizado destacando la originalidad del mismo y la complacencia de los resultados.

El documento se organiza de la siguiente forma: se presenta el marco teórico en la siguiente sección como base fundamental para el desarrollo del trabajo, a continuación las herramientas y métodos de procesamiento empleados, posteriormente se muestra el análisis de los resultados obtenidos, las principales conclusiones de la investigación y finalmente se menciona las recomendaciones y el trabajo futuro.

# 1. MARCO TEÓRICO

## 1.1. BIOMASA

### 1.1.1. Concepto de biomasa

Hace referencia a toda materia orgánica originada de forma inmediata en un proceso biológico, espontáneo o provocado, utilizable como fuente de energía [11]. También se define como el conjunto de materia orgánica renovable de origen vegetal, animal o procedente de la transformación natural o artificial de la misma [12]. Es comúnmente utilizada para fines de producción: energía térmica y eléctrica, como también para biocombustibles líquidos y gaseosos.

### 1.1.2. Tipos de biomasa

De acuerdo con [12] la biomasa se clasifica como:

- **Natural:** producida en ecosistemas naturales.

- **Residual:** que comprende:

    - Residuos forestales: en este grupo encajan los residuos como leña, madera, resina, corcho, etc.
    - Residuos agrícolas: están los restos de podas, rastrojos de cultivos. En Nariño, por ejemplo, es el caso de la rama de arveja, la rama de zanahoria, pajas de cereales de grano y cañote de maíz, entre otros.
    - Residuos de industrias forestales: representa los aserraderos, fábricas de pasta y papel. En este grupo caben residuos como cortezas, ramas y hojas, etc.
    - Residuos de industrias agrícolas: comprende los bagazos, orujos, cáscaras, vinazas, huecos, etc.
    - Residuos biodegradables: hace referencia a los purines, estiércoles, lodos de depuradoras, huesos, sebos, etc.

- **Cultivos energéticos:** son cultivos específicos dedicados exclusivamente a la producción de energía. Tienen una elevada habilidad de sobrevivir a condiciones adversas de crecimiento.

    - Especies leñosas en turnos de 3-4 años y con 10.000 pies/Ha.

- Especies herbáceas.
- Cultivos para producir etanol (trigo, maíz, pataca, sorgo azucarero).
- Cultivos para producir biodiesel (colza, girasol, lino, oleaginoso).

❖ **Excedentes agrícolas:** sirven para completar los cultivos no alimentarios y sustituir parcialmente los biocarburantes y los combustibles fósiles. Entre ellos se encuentran: aceite de algodón, aceite de soja y aceite de cártamo.

El departamento de Nariño, debido a su posición geográfica, posee grandes riquezas naturales caracterizadas por su gran diversidad y complejidad biológica. Según [13], el departamento posee 3’326.800 hectáreas, donde un 74% corresponde a usos no agropecuarios, el 3% a pastos y sabanas y el 23% restante a cultivos transitorios y permanentes.

Además, con base en la extensión del departamento [14], un 8% de su territorio pertenece al Pie de Monte de la Amazonía, una de las grandes reservas de biodiversidad del mundo, el 52% corresponde a la Llanura del Pacífico o Chocó Biogeográfico, que presenta condiciones excepcionales en diversidad de comunidades y especies, y el 40% restante, pertenece a la Zona Andina, en donde se destacan los páramos y volcanes, aspectos que posicionan a Nariño como una de las regiones más diversas de Colombia y el mundo. En la Figura 1 se muestra un mapa correspondiente a las zonas ambientales de Nariño.

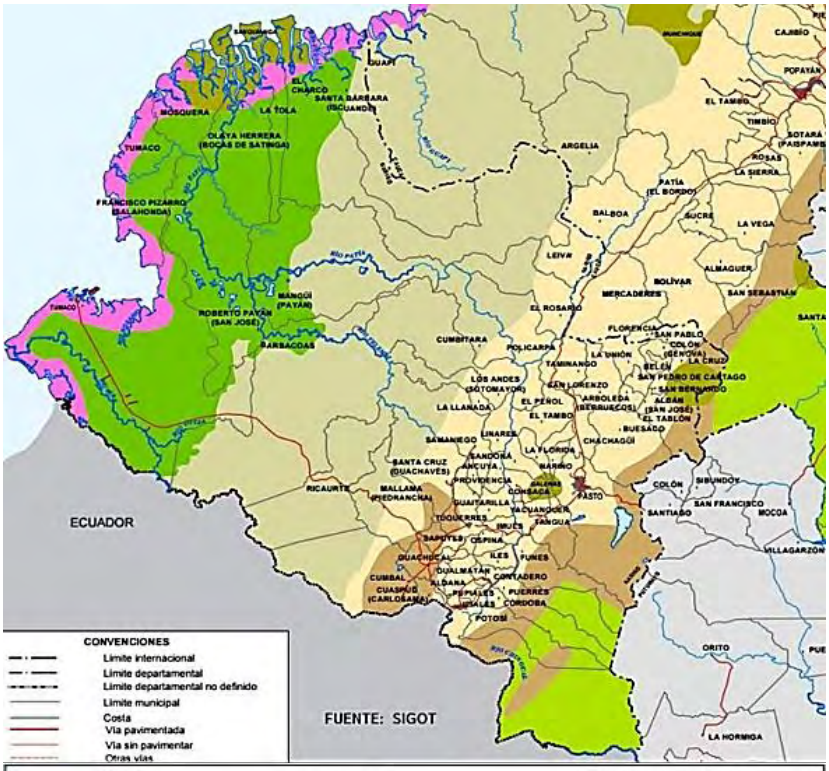
Corponariño, como entidad pública del departamento de Nariño, que ejerce el rol de autoridad ambiental, administrando los recursos naturales renovables y protegiendo el medio ambiente, facilita un mapa de clasificación, en el que se detallan diferentes tipos de biomasa. El mapa plantea dos clasificaciones, una general denominada *corpocorin* y una extendida llamada *cobertura*, que comprenden 13 y 63 clases, respectivamente. En las Figuras 2 y 3 se puede apreciar su visualización.

## 1.2. IMÁGENES SATELITALES

### 1.2.1. Teledetección

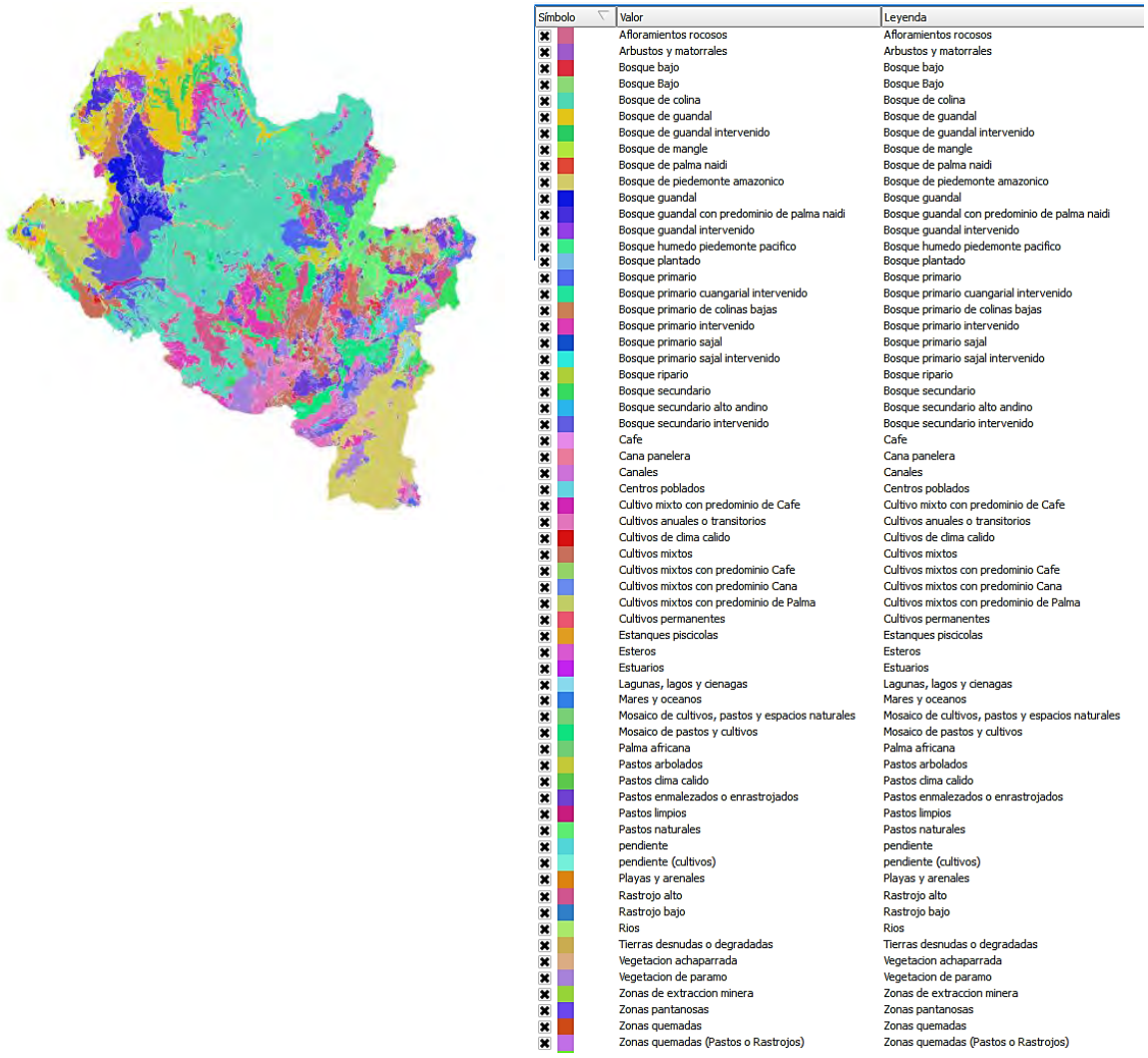
La teledetección es la detección a distancia de información producida en la superficie, gracias a la relación sensor-cobertura expresada a través de la radiación electromagnética. Cabe destacar que los conceptos de emisión, reflexión y emisión-reflexión, son determinados por el flujo de energía que se produce en función de la transmisión de energía térmica y conductividad espectral, propias de cada elemento de la naturaleza [15].

Los sensores de teledetección son instrumentos que transforman la radiación electromagnética en información perceptible y analizable. El sensor capta la energía proveniente del sol que ha sufrido interacciones fundamentales, y que deben ser entendidas para interpretar de manera correcta los datos captados.




| Nombre | Caracterización | Importancia - Amenazas |
|---|---|---|
| Zona Amazónica | Bosque natural zona de protección. | Alta importancia por biodiversidad - Selva amazónica - Presión por cultivos. |
| Zona de Alta Demanda Ambiental | Se localiza la mayor cantidad de Bosque Andino y Alto Andino, Páramos pero con la mayor presión productiva. - Amenaza alta por erosión y volcanismo. | Se localizan los principales centros urbanos con alta dependencia de agua con fuente de páramos, Ecosistemas de bosque con buena conservación, problemas por cultivos ilicitos. |
| Corredor Pacifico y Piedemonte | Región con alto potencial de biodiversidad y caracterizada por alta precipitación. | Áreas de bosques conservados, aprovechamiento de recursos forestales. Alta biodiversidad. |
| Zona Pacifico de Alta Intervención | Zona de ecosistemas fragmentados y altamente intervenidos con extrema sobrexplotacion forestal. | Zona de altos recursos forestales y explotación minera que requiere atención prioritaria por defoerstación extrema y deterioro por mineria. Amenaza alta por inundaciones y sismicidad. |
| Ecosistema de Páramos | Ecosistema estratégico con alta biodiversidad y fuente abastecedora de agua. | Interés ecológico con diversidad biótica y ambiental, zonas especiales con potencial de almacenamiento, regulación y nacimiento de recursos hídricos. Amenaza media a alta por volcanismo. |
| Ecosistema de Manglares | Alta productividad, hábitat de especies animales acuáticas y terrestres y recursos de madera, filtrador natural biológico y adicionalmente destino ecoturístico. | Protección de áreas de costa, hábitat de especies de interfase marina - terrestre. |
| Áreas Naturales Protegidas - ANP | Sistema de Parques Nacionales Naturales reglamentadas para conservación de ecosistemas. | Reserva natural de bosques, biodiversidad. |

**Figura 1.** En la parte superior, zonas ambientales de Nariño y en la parte inferior, caracterización. **Fuente:** SIGOT [14].



| Símbolo | Valor | Leyenda |
|---|---|---|
| | Afloramientos rocosos | Afloramientos rocosos |
| | Arbustos y matorrales | Arbustos y matorrales |
| | Bosque bajo | Bosque bajo |
| | Bosque Bajo | Bosque Bajo |
| | Bosque de colina | Bosque de colina |
| | Bosque de guandal | Bosque de guandal |
| | Bosque de guandal intervenido | Bosque de guandal intervenido |
| | Bosque de mangle | Bosque de mangle |
| | Bosque de palma naidi | Bosque de palma naidi |
| | Bosque de piedemonte amazonico | Bosque de piedemonte amazonico |
| | Bosque guandal | Bosque guandal |
| | Bosque guandal con predominio de palma naidi | Bosque guandal con predominio de palma naidi |
| | Bosque guandal intervenido | Bosque guandal intervenido |
| | Bosque humedo piedemonte pacifico | Bosque humedo piedemonte pacifico |
| | Bosque plantado | Bosque plantado |
| | Bosque primario | Bosque primario |
| | Bosque primario cuangarial intervenido | Bosque primario cuangarial intervenido |
| | Bosque primario de colinas bajas | Bosque primario de colinas bajas |
| | Bosque primario intervenido | Bosque primario intervenido |
| | Bosque primario sajal | Bosque primario sajal |
| | Bosque primario sajal intervenido | Bosque primario sajal intervenido |
| | Bosque ripario | Bosque ripario |
| | Bosque secundario | Bosque secundario |
| | Bosque secundario alto andino | Bosque secundario alto andino |
| | Bosque secundario intervenido | Bosque secundario intervenido |
| | Cafe | Cafe |
| | Cana panelera | Cana panelera |
| | Canales | Canales |
| | Centros poblados | Centros poblados |
| | Cultivo mixto con predominio de Cafe | Cultivo mixto con predominio de Cafe |
| | Cultivos anuales o transitorios | Cultivos anuales o transitorios |
| | Cultivos de clima calido | Cultivos de clima calido |
| | Cultivos mixtos | Cultivos mixtos |
| | Cultivos mixtos con predominio Cafe | Cultivos mixtos con predominio Cafe |
| | Cultivos mixtos con predominio Cana | Cultivos mixtos con predominio Cana |
| | Cultivos mixtos con predominio de Palma | Cultivos mixtos con predominio de Palma |
| | Cultivos permanentes | Cultivos permanentes |
| | Estanques piscicolas | Estanques piscicolas |
| | Esteros | Esteros |
| | Estuarios | Estuarios |
| | Lagunas, lagos y cienagas | Lagunas, lagos y cienagas |
| | Mares y oceanos | Mares y oceanos |
| | Mosaico de cultivos, pastos y espacios naturales | Mosaico de cultivos, pastos y espacios naturales |
| | Mosaico de pastos y cultivos | Mosaico de pastos y cultivos |
| | Palma africana | Palma africana |
| | Pastos arbolados | Pastos arbolados |
| | Pastos clima calido | Pastos clima calido |
| | Pastos enmalezados o enrastrojados | Pastos enmalezados o enrastrojados |
| | Pastos limpios | Pastos limpios |
| | Pastos naturales | Pastos naturales |
| | pendiente | pendiente |
| | pendiente (cultivos) | pendiente (cultivos) |
| | Playas y arenales | Playas y arenales |
| | Rastrojo alto | Rastrojo alto |
| | Rastrojo bajo | Rastrojo bajo |
| | Rios | Rios |
| | Tierras desnudas o degradadas | Tierras desnudas o degradadas |
| | Vegetacion achaparrada | Vegetacion achaparrada |
| | Vegetacion de paramo | Vegetacion de paramo |
| | Zonas de extraccion minera | Zonas de extraccion minera |
| | Zonas pantanosas | Zonas pantanosas |
| | Zonas quemadas | Zonas quemadas |
| | Zonas quemadas (Pastos o Rastrojos) | Zonas quemadas (Pastos o Rastrojos) |

**Figura 2.** A la izquierda, mapa de *cobertura* del departamento de Nariño y a la derecha, su caracterización. **Fuente:** Corponariño 2015.

Una de las aplicaciones de la información recogida por la teledetección remota es la prospección de minerales, detectar o monitorizar el uso de tierras, deforestación, el estado de salud de plantas indígenas y cultivos, incluyendo zonas enteras de cultivo o bosques, a través de plataformas multi-espectrales simultáneas como Landsat, que ha estado en uso desde los años 70 [16]. Estos proyectos toman imágenes en múltiples longitudes de onda del espectro electromagnético con sensores ubicados normalmente en satélites de observación terrestre. Landsat ha sido el único sistema de satélite diseñado y operado para observar repetidas veces la cubierta de la tierra con una resolución moderada.

Lo realmente importante es que mientras que el objeto o fenómeno en cuestión (el estado) no se van a medir de manera directa, existen otras variables que se

detectan y miden (la observación), que están intrínsecamente relacionadas con el objeto de interés a través de un modelo. De esta manera, la calidad de la información recogida a distancia depende de sus resoluciones espacial, espectral, radiométrica y temporal.

- **Resolución espacial:** es el tamaño de un píxel, que corresponde con áreas cuadradas cuyo lado varía de 1 a 1000 metros y que se guarda en una imagen en mapa de bits.

- **Resolución espectral:** es la amplitud de la longitud de onda de las diferentes frecuencias grabadas. Por ejemplo, el último proyecto Landsat, "Landsat 8", comprende 11 bandas diferentes incluyendo varias del espectro infrarrojo; en total adquiere desde los 0,43 µm a los 12,51 µm [17].

- **Resolución radiométrica:** es la capacidad del sensor para distinguir diferentes intensidades de radiación. Normalmente comprende de 8 a 14 bits, correspondientes a los 256 niveles de una escala de grises, y puede llegar a 16.384 intensidades de color en cada banda.

- **Resolución temporal:** es la frecuencia con la que el satélite sobrevuela una zona.


**Figura 3.** A la izquierda, mapa de *corpocorin* del departamento de Nariño y a la derecha, su caracterización. **Fuente:** Corponariño 2015.

Los instrumentos en Landsat 8, Operational Land Imager (OLI) y el sensor térmico infrarrojo (TIRS) muestran avances representativos en la tecnología de sensores remotos y en su rendimiento. OLI y TIRS miden la superficie terrestre en el visible, infrarrojo cercano, infrarrojo de onda corta, e infrarrojo térmico. OLI con una resolución espacial entre 15 y 30 metros, cubre amplias zonas de la tierra, mientras que proporciona una resolución suficiente como para distinguir las características tales como centros urbanos, granjas, bosques y otros tipos de

cubiertas del suelo. TIRS, por su lado, presenta una resolución espacial de 100 metros, diseñada para captar información acerca de la temperatura de la superficie terrestre a través de dos bandas del infrarrojo térmico (banda 10 y banda 11), permite distinguir entre la temperatura de la superficie terrestre y la temperatura atmosférica.

Cabe resaltar que además de las 7 bandas multiespectrales del sensor ETM+ a bordo del anterior Landsat 7 (Figura 4), OLI agrega dos nuevas bandas espectrales, un nuevo canal en azul visible, "costera", (banda 1), diseñado específicamente para observar la calidad del agua en lagos someros y zonas costeras y para detectar aerosoles, y una banda en el infrarrojo de onda corta "cirros" (banda 9), para determinar la presencia de nubes, fundamentalmente cirrus. Ver Tabla I expuesta en [18]. Estas nuevas bandas, ayudan a los científicos a medir la calidad del agua y facilitan la detección de nubes altas y delgadas que previamente han sido difíciles de observar en las imágenes Landsat.

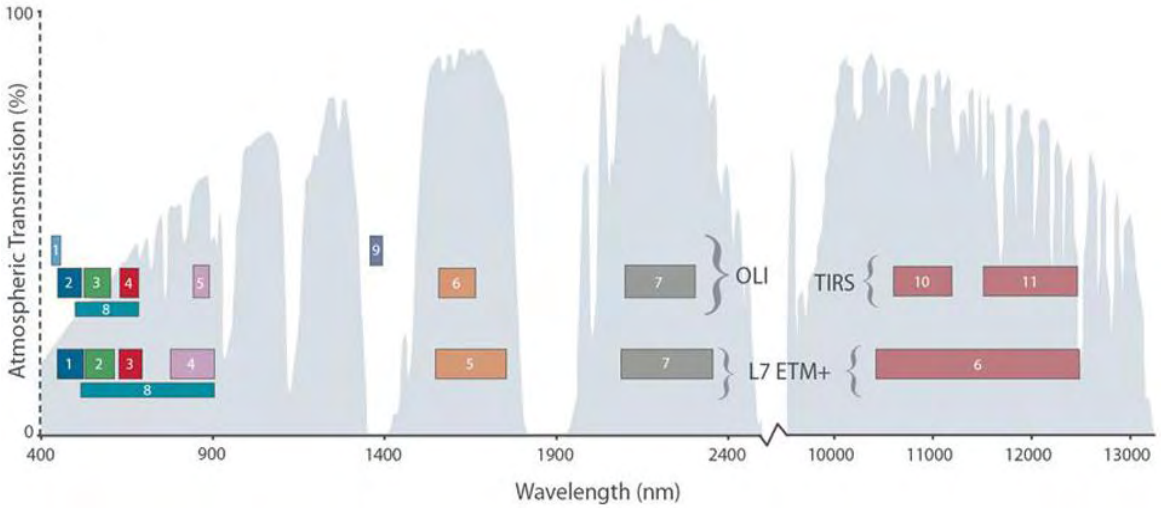



**Figura 4.** Ancho de banda para los sensores OLI y TIRS en Landsat 8 en comparación con ETM+ en Landsat 7. Los valores de transmisión atmosférica para este gráfico se calculan usando MODTRAN (transmisión atmosférica de resolución moderada) para una atmósfera brumosa de latitudes medias durante el verano (alrededor de 5 km de visibilidad). (USGS. 2015). **Fuente:** [18].

### 1.2.2. Concepto de imagen satelital

La imagen satelital o imagen de satélite se puede definir como la representación visual de la información capturada por un sensor montado en un satélite. De esta manera, la imagen satelital es entendida como una matriz digital de puntos, que se encuentra en formato raster o imagen de mapa de bits y que por tanto se compone de pixeles, donde para cada pixel se registra un valor de reflectancia. Dentro del

campo de aplicación, se han convertido en una herramienta eficaz en el estudio del clima, los océanos, los vientos y la vegetación [19].

TABLA I
BANDAS DISPONIBLES EN LANDSAT 8

| | Bandas | Longitud de onda (micrómetros) | Resolución (metros) |
|---|---|---|---|
| **Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS)** | Banda 1 – Aerosol costero | 0,43 – 0,45 | 30 |
| | Banda 2 – Azul | 0,45 – 0,51 | 30 |
| | Banda 3 – Verde | 0,53 – 0,59 | 30 |
| | Banda 4 – Rojo | 0,64 – 0,67 | 30 |
| | Banda 5 – Infrarrojo cercano (NIR) | 0,85 – 0,88 | 30 |
| | Banda 6 – SWIR 1 | 1,57 – 1,65 | 30 |
| **Febrero 11, 2013** | Banda 7 – SWIR 2 | 2,11 – 2,29 | 30 |
| | Banda 8 – Pancromático | 0,50 – 0,68 | 15 |
| | Banda 9 – Cirrus | 1,36 – 1,38 | 30 |
| | Banda 10 – Infrarrojo térmico (TIRS) 1 | 10,60 – 11,19 | 100 |
| | Banda 11 – Infrarrojo térmico (TIRS) 2 | 11,50 – 12,51 | 100 |

Las imágenes por lo general tienen una resolución de 30 metros; es decir cada pixel representa un área de 30 × 30 metros correspondientes a $900\mathrm{m}^2$. Las imágenes en mención son multiespectrales, conteniendo información de muchas bandas del espectro electromagnético. Según [20], los satélites multiespectrales de hoy en día, miden la reflectancia simultáneamente para entre tres y catorce bandas.

- **Descripción de productos de Landsat 8:** para esta investigación se usan datos predefinidos en formato Geographical Tagged Image File Format (GeoTIFF) y se obtienen a través de la página web *earthexplorer* [21]. Adicionalmente, los datos contienen una corrección topográfica por el desplazamiento del terreno debido al relieve.

  Los productos usados se encuentran en formato de niveles digitales enteros (DN) con una resolución radiométrica de 16 bits. Estos se pueden convertir a valores de reflectancia en el techo de la atmósfera (TOA) - (bandas 1-9) o radiación (Bandas 1-11), con factores de escala previstas en los productos. Las imágenes satelitales que se descargan para el trabajo son libres, contienen archivos para cada una de las bandas espectrales de Landsat 8 e incluyen información adicional sobre el tipo de proyección, sistema de coordenadas, y todo lo necesario para que la imagen pueda ser automáticamente posicionada en un sistema de referencia espacial.

Uno de los archivos, identificado con la etiqueta _BQA.TIF, contiene un valor decimal que representa las combinaciones de bits de relleno de la superficie, la atmósfera y las condiciones de sensores que pueden afectar a la utilidad general de un píxel dado. Adicionalmente, el archivo _MTL.txt contiene la información específica del producto, como los nombres de los archivos, constantes térmicas, entre otros detalles. Este archivo es fundamental para llevar a cabo ciertos procesos como la corrección radiométrica en imágenes de Landsat 8.

### 1.2.3. Reflectancia

La reflectividad mide la relación entre la amplitud del campo electromagnético reflejado respecto a la amplitud del campo incidente, mientras que la reflectancia se refiere a la relación entre la potencia electromagnética incidente con respecto a la potencia que es reflejada en una interface o superficie. Por lo tanto, la magnitud de la reflectancia es el cuadrado de la magnitud de la reflectividad. La reflectancia es siempre un número real positivo.

Algunos materiales pueden ser identificados por su espectro de reflectancia, por lo que es común corregir una imagen a la reflectancia como un primer paso hacia la localización o identificación de características en una imagen satelital. Estos valores de reflectancia incluyen contribuciones de nubes, aerosoles y gases atmosféricos. Existen muchas formas de obtener imágenes de reflectancia a partir de imágenes de radiancia incluyendo correcciones atmosféricas basadas en diferentes modelos.

En la identificación de objetos y procesos en la superficie terrestre, lo que interesa es la reflectividad de los objetos respecto a las diferentes longitudes de onda. Cada tipo de material, suelo, vegetación, agua, etc., reflejará la radiación incidente de forma diferente, lo que permite distinguirlo de los demás si se mide la radiación reflejada. A partir de medidas de laboratorio se ha obtenido la reflectividad para distintas cubiertas terrestres en diferentes longitudes de onda. El gráfico que muestra la reflectividad en porcentaje para cada longitud de onda se conoce como signatura o firma espectral (ver ejemplo en la Figura 5), y constituye una marca de identidad de los objetos, lo que facilita distinguir entre suelo y vegetación, e incluso entre diferentes tipos de suelo o diferentes tipos de vegetación.

El análisis para distinguir entre un tipo de cobertura y otro, dependiendo de la firma espectral, tiene muchos aspectos que dificultan el procesamiento directo, ya que las variables distorsionan las medidas obtenidas por el satélite. Para disminuir dicho error, es necesario llevar a cabo correcciones en las imágenes satelitales para el posterior procesamiento de los resultados.

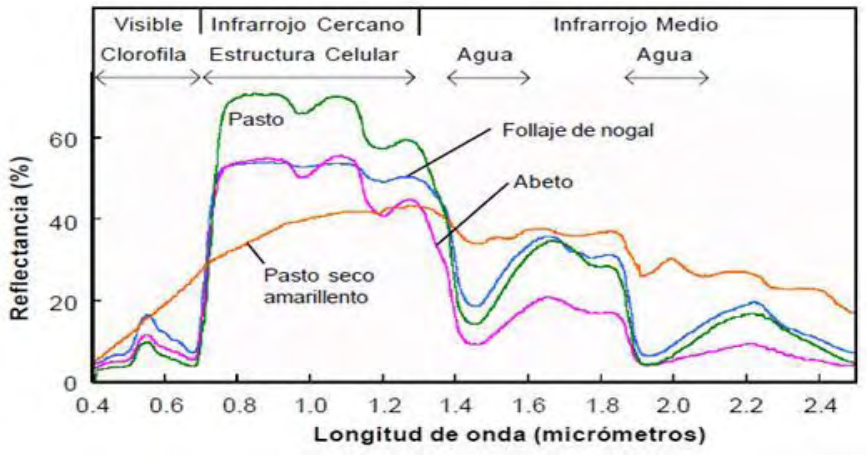



**Figura 5.** Firma o signatura espectral: gráfico en el que se compara la reflectancia espectral versus la longitud de onda para diferentes materiales. **Fuente:** [22].

## 1.3. CLUSTERING

### 1.3.1. Concepto de clustering

El clustering es un proceso que tiene como finalidad agrupar individuos o datos en clases o clusters, en donde los datos que pertenecen a un mismo grupo tienen un alto grado de semejanza y a su vez son muy diferentes respecto a los de otro cluster. Dentro de esta investigación, un pixel de la imagen satelital es considerado como un individuo y la reflectancia en cada una de sus bandas espectrales son las características que se analizan para determinar a qué grupo o cluster pertenece, descubriendo distribución de patrones y correlaciones entre los atributos. Como resultado de la aplicación de clustering se obtiene para cada individuo la clase a la que pertenece y diferentes centroides, que son puntos equidistantes de los objetos pertenecientes a un mismo cluster. En la Figura 6 se puede apreciar el resultado del proceso de clustering para un conjunto determinado de datos.

A diferencia de la clasificación, el clustering no requiere clases predefinidas o conjuntos de entrenamiento. Por esta razón, en [23] se aclara que el clustering es un ejemplo de aprendizaje por observación, mientras que la clasificación es una técnica de aprendizaje por ejemplos. Para el desarrollo de este trabajo no se contó con un conjunto de datos de entrenamiento para llevar a cabo un método supervisado, por lo que se aplica clustering como técnica de agrupamiento. Existen diferentes métodos de clasificación en función de la metodología utilizada para la caracterización espectral de distintas cubiertas. En el caso de la

clasificación no supervisada, se realiza la caracterización espectral, agrupando pixeles similares de la imagen.

El algoritmo de clustering hace una medición de similitud entre pixeles, basada en la distancia entre las diferentes características o bandas espectrales que componen la imagen. Según [6], los algoritmos más utilizados son k-means e isodata, basados en la búsqueda automática de grupos de valores digitales homogéneos. En este caso, el intérprete tiene que determinar el grado de correspondencia entre las clases obtenidas y las buscadas en la clasificación (clases informacionales), pudiendo no coincidir unas con otras.

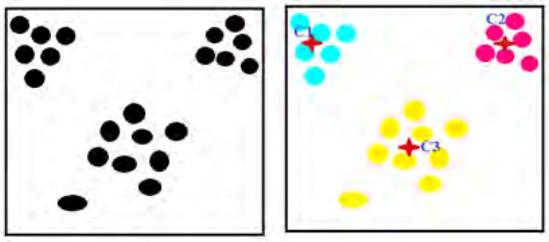



**Figura 6.** Resultado de la aplicación de clustering. C1, C2, y C3, son los centroides de los 3 grupos encontrados a partir de los datos de entrada.

Por su parte, en el método de clasificación supervisado se tiene conocimiento previo de los tipos de grupos existentes a clasificar, lo que permite definir en la imagen una serie de localizaciones denominadas áreas de entrenamiento. Además, se debe realizar una correcta selección de muestras o inicialización de centroides, de lo contrario los resultados pueden ser erróneos.

El resultado de la clasificación dependerá no solo del algoritmo seleccionado, sino también de las necesidades y requerimientos del estudio. Es por esto que se debe elegir un algoritmo de asignación acorde a los fines de investigación.

### 1.3.2. Clasificación de clustering

A pesar de la gran cantidad de técnicas de agrupamiento existentes en la literatura, todas pueden ser clasificadas en alguno de los siguientes cuatro tipos [24]:

- **Algoritmos de agrupamiento particionales:** son aquellos que obtienen como resultado una única partición de los datos iniciales en lugar de una estructura de agrupamiento con varios niveles de particiones. Asigna a un conjunto de objetos *K* grupos sin estructura jerárquica, siendo *K* un número entero menor que el número total de objetos. Este tipo de algoritmos son muy eficientes en aquellas aplicaciones con conjuntos de datos de gran

dimensionalidad, aunque presentan el problema de que es necesario escoger el número de grupos deseados a priori. Los algoritmos particionales por lo general producen grupos optimizando una función criterio definida. Se ejecuta varias veces el algoritmo con distintos puntos de entrada y la mejor configuración obtenida de entre todas las ejecuciones es usada como salida del algoritmo.

- **Algoritmos de agrupamiento jerárquicos:** organizan los datos en estructuras jerárquicas de acuerdo a la matriz de proximidades (matriz de similaridades/distancias de orden $N \times N$ siendo $N$ el número de individuos). Los resultados de estos algoritmos son, por lo general, mostrados en un árbol binario o en un dendograma (gráfico que permite visualizar el proceso de agrupamiento de los clusters, sus proximidades, y ayuda a decidir cuántos grupos formar).

- **Algoritmos de agrupamiento probabilísticos:** desde el punto de vista probabilístico se asume que los objetos son generados de acuerdo a algunas distribuciones probabilísticas. Objetos en distintos grupos son generados por distintas distribuciones de probabilidad o son derivados de distintos tipos de funciones de densidad, o de las mismas familias pero con distintos parámetros.

- **Algoritmos de agrupamiento basados en densidades:** estos algoritmos aplican criterios locales de grupo. Los grupos son regiones en el espacio de datos de gran densidad de objetos, que están separados por regiones de menor densidad (ruido). Estas regiones pueden tener cualquier forma y pueden estar distribuidas de cualquier manera.

Además, los algoritmos de agrupamiento pueden ser duros o difusos (fuzzy). Un algoritmo de agrupamiento duro asigna cada objeto a un único grupo durante su operación y a la salida. Por el contrario, un algoritmo de agrupamiento difuso puede asignar cada objeto o dato a más de un grupo con un cierto grado de pertenencia. Un algoritmo de agrupamiento difuso se puede convertir en uno duro si se asigna cada objeto al cluster con mayor grado de pertenencia.

### 1.3.3. K-Means

Es un algoritmo duro, el más conocido entre los algoritmos de agrupamiento particionales, que usa como función criterio una función de error cuadrático [24]. La ecuación 1 define el criterio de error cuadrático dado un conjunto de entrada

con $d$ características, $\mathrm{x_j} \in \mathbb{R}^{\mathrm{d}}, \mathrm{j} = 1, \ldots, \mathrm{N}$, de N individuos, que se agrupa en un conjunto de *K* grupos, $\mathrm{C} = \{\mathrm{C_1}, \ldots, \mathrm{C_K}\}$

$$J(\Gamma, M) = \sum_{i=1}^{k} \sum_{j=1}^{N} \gamma_{ij} \left\| x_j - m_i \right\|^2 \quad (1)$$

Donde, $\Gamma = [\gamma_{ij}]$ es una matriz de particiones, siendo $\gamma_{ij} = \begin{Bmatrix} 1, & si\ \mathrm{x_j} \in cluster\ i \\ 0, & otro \end{Bmatrix}$, con $\sum_{i=1}^{K} \gamma_{ij} = 1$ para todo $j$, y donde $M = [m_1, \ldots, m_K]$ es la matriz de medias, centroides o prototipos de los *K* grupos, siendo $m_i = \frac{1}{N_i} \sum_{j=1}^{N} \gamma_{ij} \mathrm{x_j}$ una matriz de la media del grupo $i$ dados $N_i$ objetos en ese grupo.

Los grupos resultantes de esta función de error cuadrática son frecuentemente denominados particiones.

Existen diversas maneras de asignar cada objeto del conjunto de datos al grupo $\mathrm{C_w}$ más cercano. Una de las técnicas más usadas es que un objeto $x_j$ pertenezca al grupo $C_w$ si $\left\| x_j - m_w \right\| < \left\| x_j - m_i \right\|$ para $j = 1, \ldots, N, i \neq w\ e\ i, w = 1, \ldots, K$.

### 1.3.4. EM-Expectation Maximization: Esperanza-Maximización

El algoritmo EM pertenece al tipo de clustering probabilístico, puesto que se trata de tener una *FDP* (Función de Densidad de Probabilidad) desconocida a la que pertenecen el conjunto completo de datos o en este caso, de pixeles, que posee la base de datos. Según [25], esta *FDP* se puede aproximar mediante una combinación lineal de *NC* componentes definidas a falta de una serie de parámetros, $\{\theta\} = U\{\theta_j \forall j = 1 \ldots NC\}$ que son los que hay que averiguar. La FDP está dada por

$$P(x) = \sum_{j=1}^{NC} \pi_j p(x; \theta_j)\ con\ \sum_{j=1}^{NC} \pi_j = 1 \quad (2)$$

Donde, $\pi_j$ se entiende como las probabilidades a priori de cada cluster cuya suma debe ser igual a 1, que también forman parte de la solución buscada, $P(x)$ denota la *FDP* arbitraria y $p(x; \theta_j)$ la función de densidad del componente $j$.

EM hace una suposición acerca de los datos de entrada $x$, que después de iterar cada cluster corresponde con las respectivas muestras de datos que pertenecen a cada una de las densidades que se mezclan. Se pueden estimar *FDP* de forma arbitraria utilizando *FDP* normales n-dimensionales, T-Student, Bernoulli, Poisson, y log-normales. En esta investigación se modelan los datos de reflectancias mediante distribuciones normales multivariantes, ya que se realiza una mezcla gaussiana, en donde se efectúa el proceso que se evidencia en la Figura 7.

El ajuste de los parámetros del modelo requiere alguna medida de su bondad, es decir qué tan bien encajan los datos sobre la distribución que los representa. Por tanto, se busca maximizar este valor, que se calcula a partir de la siguiente expresión:

$$L(\theta, \pi) = \log \prod_{n=1}^{NI} P(x_n) \quad (3)$$

Donde $NI$ es el número de pixeles, que suponemos independientes entre sí. El algoritmo EM, procede en dos pasos que se repiten de forma iterativa:

- **Esperanza - Expectation**: utiliza los valores de los parámetros, iniciales o proporcionados por el paso Maximization de la iteración anterior, obteniendo diferentes formas de la *FDP* buscada.

- **Maximización - Maximization**: obtiene nuevos valores de los parámetros a partir de los datos proporcionados por el paso anterior.

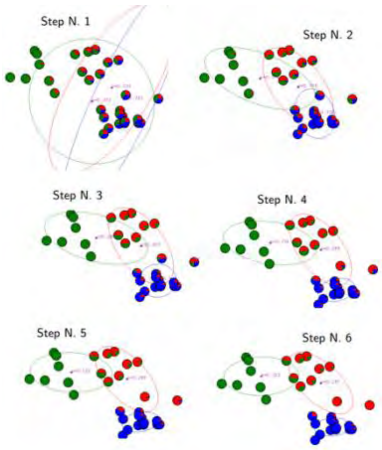



**Figura 7.** Proceso de mezcla gaussiana, gaussian mixture model. **Fuente:** [23].

Después de una serie de iteraciones, EM tiende a un máximo local de la función $L$. Finalmente, se obtiene un conjunto de clusters que agrupan el conjunto de pixeles original. Cada uno de estos clusters está definido por los parámetros de una distribución normal. Para mayor información del funcionamiento de este algoritmo con sus diferentes variaciones se recomienda consultar [26].

### 1.3.5. Isodata

Isodata es el acrónimo de Iterative Self-Organizing Data Analysis Techniques. Se trata de un algoritmo iterativo cuyo objetivo final es la minimización del criterio SSW (Sum of Squared Within o suma del cuadrado interior que hace referencia al

índice de cohesión entre clusters). Al igual que en los algoritmos mencionados anteriormente, los patrones se procesan repetitivamente y en cada iteración se asignan al grupo más idóneo, recalculándose los centros de los agrupamientos después de cada asignación. La diferencia más relevante de este es que incorpora la definición de umbrales con los que establece una serie de heurísticas con tres objetivos:

- **Eliminar:** es un sub-proceso en el que se eliminan agrupamientos poco numerosos teniendo en cuenta un criterio establecido.

- **Mezclar:** en esta etapa se mezclan agrupamientos cercanos de acuerdo a la distancia entre grupos.

- **Dividir:** hace referencia a la separación de agrupamientos dispersos.

Para controlar estas operaciones se requieren los siguientes parámetros:

$ON$: umbral del número de elementos para la eliminación de un agrupamiento.
$OC$: umbral de distancia entre los centros para la unión de agrupamientos.
$OS$: umbral de desviación estándar típica para la división de un agrupamiento.

Otros parámetros necesarios son:

$K$: número máximo de agrupamientos.
$L$: máximo número de agrupamientos que pueden mezclarse en una sola iteración.
$I$: máximo número de iteraciones permitidas.

El algoritmo utilizado en este trabajo, además de los parámetros descritos anteriormente, cuenta con un parámetro adicional, *min,* correspondiente a la mínima distancia entre un punto y cada uno de los centros. Éste parámetro interfiere en la eliminación de individuos que hacen parte de un cierto agrupamiento, es decir que si no se desea eliminar ningún individuo, *min* debe tener un valor elevado.

Las líneas generales de este algoritmo son las siguientes:

- En cada iteración se procesan todos los patrones, asignándose al agrupamiento más cercano y recalculándose los centros.

- Se eliminan los agrupamientos que tengan menos de $ON$ miembros. La eliminación no conlleva la mezcla con el agrupamiento más cercano.

- Se aplican una serie de heurísticas para aumentar (por división) o reducir (por mezcla) el número de agrupamientos. Los criterios que se siguen son:

- Se divide un agrupamiento por la variable máxima varianza si ésta supera el umbral $OS$.
- Se unen agrupamientos si las distancias entre los centros son inferiores al umbral $OC$. Se impone un máximo de $L$ mezclas por iteración.

De esta manera, el algoritmo tiene como entradas el conjunto de datos y los parámetros descritos anteriormente. Para mayor información de la ejecución y desarrollo del algoritmo consultar [27].

### 1.3.6. FCM: Fuzzy C-Means

FCM se considera un algoritmo probabilístico difuso, en el que se asume que cada objeto $x_j$ tiene algún grado de pertenencia $u_i(x_i)$ al cluster $C_i$, con $0 \leq u_i(x_i) \leq 1$. Se busca minimizar la función de coste global que se muestra en la ecuación 4.

$$J(\mathrm{U}, M) = \sum_{i=1}^{C} \sum_{j=1}^{N} {u_{ij}}^{b} \left\| x_j - m_i \right\|^2 \tag{4}$$

Donde, $U = \left[u_{ij}\right]_{C \times N}$ es la matriz de particiones difusa, $u_{ij} \in [0, 1]$ el grado de pertenencia del objeto $x_j$ al grupo $C_i$, $M = [m_1, \ldots, m_C]$ la matriz de centroides y $b > 1$ un parámetro de elección libre escogido para ajustar la "mezcla" de los distintos grupos.

Las probabilidades de pertenencia de los objetos a los grupos son normalizadas según la ecuación 5. El mínimo de la función de coste se obtiene cuando $\frac{\partial J}{\partial m_i} = 0$ y cuando $\frac{\partial J}{\partial \hat{P}_j} = 0$ llevándonos a las condiciones expresadas en la ecuación 6, teniendo en cuenta que $P\left(C_i \middle| x_j\right)$ es la probabilidad de pertenencia de un objeto $x_j$ a un grupo $C_i$.

$$\sum_{i=1}^{C} \hat{P}\left(C_i \middle| x_j\right) = 1, \mathrm{j} = 1, \ldots, \mathrm{N} \tag{5}$$

$$m_j = \frac{\sum_{i=1}^{N} \left[P\left(C_i \middle| x_j\right)\right]^b x_j}{\sum_{i=1}^{N} \left[P\left(C_i \middle| x_j\right)\right]^b} \qquad P\left(C_i \middle| x_j\right) = \frac{\left(1 / \left\| x_j - m_i \right\|^2\right)^{1/(b-1)}}{\sum_{r=1}^{C} \left(1 / \left\| x_j - m_r \right\|^2\right)^{1/(b-1)}} \tag{6}$$

En general, la función de coste es minimizada cuando los centros de los grupos $m_j$ están cerca de aquellos puntos que tienen una elevada probabilidad de pertenecer al grupo $C_j$. Las medias de los grupos y las probabilidades de los puntos se estiman iterativamente según el algoritmo como se expone en [24].

### 1.3.7. GK: Gustafson Kessel

El algoritmo GK hace parte del clustering duro, es una versión extendida del FCM contextualizado anteriormente, con la diferencia de que éste algoritmo busca clusters hiperhelipsoidales. Es bastante adecuado para el propósito de la identificación, ya que tiene las siguientes propiedades:

- La dimensión de los clusters viene limitada por la medida de la distancia entre puntos y por la definición del prototipo de los clusters como un pixel ideal.

- En comparación con otros algoritmos, GK es relativamente insensible a la inicialización de la matriz de partición.

- Como el algoritmo está basado en una norma adaptativa, no es sensible al escalado de los datos, con lo que se hace innecesaria la normalización previa.

Sin embargo, también tiene sus desventajas:

- La carga computacional es bastante elevada, sobre todo en el caso de grandes cantidades de datos.

- El algoritmo GK puede detectar clusters de diferentes formas, no solo subespacios lineales que son los que en principio nos interesan. Cuando el número de datos disponibles es pequeño, o cuando los datos son linealmente dependientes, pueden aparecer problemas numéricos ya que la matriz de covarianzas se hace casi singular.

- El algoritmo GK no puede ser aplicado a problemas puramente lineales y en el caso ideal de no existir ruido. Si no hay información al respecto, los volúmenes de los clusters se inicializan a valores todos iguales. De esta forma, no se pueden detectar clusters con grandes diferencias en tamaño.

### 1.3.8. Dbscan

Pertenece al grupo de clustering que se basa en la densidad. Dentro de su metodología está el hacer crecer un cluster siempre y cuando la densidad en el entorno del objeto no exceda un umbral.

Tiene como fundamento buscar puntos centrales o *core* en los grupos. Para el agrupamiento realizado por este algoritmo existen 2 parámetros fundamentales que son *MinPts* y $\epsilon$, que equivalen al número de puntos mínimo y radio, respectivamente. Por lo anterior se puede decir que Dbscan hace distinción de tres tipos de puntos:

- **Puntos centrales o core:** son los puntos que están dentro de un grupo determinado limitado por el radio épsilon, $\epsilon$, y además el número de sus integrantes es mayor a *MinPts*.

- **Puntos de borde:** estos, como su nombre lo indica, son los puntos que no son centrales y pertenecen a la región de vecindad de uno o más puntos centrales.

- **Puntos de ruido:** son aquellos que no pertenecen a ninguno de los anteriores grupos y que están fuera de las condiciones impuestas por $\epsilon$ y *MinPts*.

Según [24], para que el algoritmo funcione de manera correcta, es necesario que para cada punto $p$ de un grupo $C$, exista otro punto del mismo grupo tal que $p$ esté dentro de la región de vecindad de $q$, y que esta región de vecindad contenga un número mínimo de puntos. En la Figura 8 se ilustra el procedimiento efectuado. Se evidencia cómo un punto $p$ es densamente alcanzable desde otro punto $q$, mientras que $q$ no es densamente alcanzable desde $p$, puesto que la región de vecindad de $p$ no contiene el número minimo de puntos. En el lado izquierdo de la misma figura se muestra como $p$ y $q$ están densamente conectados desde el punto $o$.

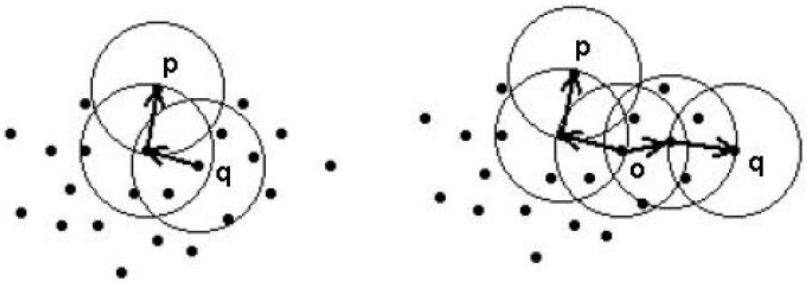



**Figura 8.** Proceso efectuado en Dbscan, a la izquierda se muestra el punto densamente alcanzable, a la derecha los puntos densamente conectados. **Fuente:** [24].

Cuando el algoritmo encuentra un punto que con radio $\epsilon$, contiene la cantidad de *MinPts* o más, entonces se consideran todos los puntos dentro de ese radio como parte de un mismo cluster. En seguida, se expande este grupo mediante la

comprobación de todos los nuevos puntos, verificando que ellos también tienen más puntos *MinPts* a una distancia épsilon, creciendo el cluster en caso afirmativo.

Entre sus desventajas están el asumir que los clusters detectados poseen densidades similares y que puede presentar problemas a la hora de separar grupos. La información detallada de este algoritmo se puede consultar en [28].

### 1.3.9. Mean - shift

También se lo conoce como algoritmo de búsqueda de modas. Es un método iterativo no paramétrico; es decir, que su distribución no puede ser definida a priori, pues son los datos observados los que la determinan. Su finalidad es encontrar las modas de unas distribuciones, pero sin necesitar saber cuántas modas se tienen, por ello no necesita definir el número de clusters a priori. Considera que el espacio de datos es una función de densidad de probabilidad muestreada. Para cada punto del conjunto de datos, encuentra la moda más cercana, para lo que define una región alrededor de ese punto, a partir del único parámetro de entrada que se define como el criterio de mezcla y encuentra su media, cambiando la situación de la media actual a la nueva (shift). Repite el proceso hasta que converja. Sin embargo, cuanto el conjunto de datos es grande (más de 1.000 puntos o pixeles) la complejidad computacional se vuelve demasiado costosa [29]. Los dominios de aplicación incluyen análisis de cluster en visión por computadora y procesamiento de imágenes. Para información acerca de su implementación se recomienda consultar [30], [31].

### 1.3.10. Region growing

Este es un algoritmo de segmentación que se basa en la técnica de crecimiento de región o region growing, en la que se almacena una buena división y se producen etiquetas semánticas de manera interactiva. En [32] se plantea que este método agrupa pixeles a partir de un criterio de semejanza. El proceso empieza con unos valores de intensidad de pixel dentro de la imagen que son definidos inicialmente y que se les conoce como semillas. La técnica busca agrupar pixeles vecinos a las semillas que cumplan con el criterio de similitud determinado, proceso de agrupación que se realiza hasta que no se encuentre más valores de intensidad de pixel que cumplan con el criterio de similitud.

## 1.4. PCA: PRINCIPAL COMPONENT ANALYSIS: ANÁLISIS DE COMPONENTES PRINCIPALES

PCA es un proceso estadístico ampliamente utilizado para reducción de dimensión de forma no supervisada en imágenes [33]. También se define como una técnica muy difundida en el tratamiento de grandes masas de datos, que permite reducir su dimensionalidad con el fin de evitar redundancias y destacar relaciones. Según [34], en la mayoría de los casos, tomando sólo los primeros componentes se puede explicar la mayor parte de la variación total contenida en los datos originales. PCA transforma el conjunto de $p$ variables originales en otro conjunto de $q$ variables sin correlación entre ellas ($q \leq p$) llamadas componentes principales. Las $q$ variables son medidas sobre cada uno de los $n$ individuos, obteniendo una matriz de datos de orden $n \times q$.

En PCA existe la opción de usar la matriz de correlaciones o la matriz de covarianzas. En la primera opción se le está dando la misma importancia a todas y cada una de las variables, que puede ser conveniente cuando el investigador considera que todas las variables son igualmente relevantes. La segunda opción se puede utilizar cuando todas las variables tengan las mismas unidades de medida y cuando el investigador juzga conveniente destacar cada una de las variables en función de su grado de variabilidad.

Las $q$ componentes principales son obtenidas como combinaciones lineales de las variables originales. Los componentes se ordenan en función del porcentaje de varianza, en este sentido, el primer componente será el más importante por ser el que explica mayor porcentaje de la varianza de los datos. Queda a criterio del investigador decidir cuántos componentes se eligen en el estudio, pero una técnica destacada para esto es el análisis del gráfico de autovalores.

### 1.4.1. Gráfico de los autovalores

Consiste en representar el porcentaje de variación entre los datos obtenidos con PCA y los datos originales, contra el número de componentes principales. En el eje de las ordenadas se registra el porcentaje de variación, mientras que en el eje de las abscisas se coloca el número de componentes según su orden de importancia. Por lo general, los puntos del gráfico presentan una figura similar al perfil de una bota. Al analizar este gráfico se busca el punto de quiebre donde el cambio de la pendiente se hace mayor. La abscisa correspondiente a este punto indica el número de componentes a retener.

### 1.4.2. Método basado en correlaciones

Es una forma básica de aplicar PCA que parte de la matriz de correlaciones. Se considera el valor $F_j$ de cada una de las $m$ características que posee un pixel. Para cada uno de los $n$ pixeles se toma el valor de estas variables y se escribe el conjunto de datos en forma de matriz, como se muestra en la ecuación 7, donde cada conjunto descrito por la ecuación 8 puede considerarse una muestra aleatoria para la variable $F_j$.

$$\binom{F_j^{\beta}}{} \begin{matrix} \beta = 1, \dots, n \\ j = 1, \dots, m \end{matrix} \tag{7}$$

$$M_j = \left\{ F_j^{\beta} \,/\, \beta = 1, \dots, n \right\} \tag{8}$$

A partir de los $m \times n$ datos puede construirse la matriz de correlación muestral, definida por la ecuación 9.

$$R = [r_{ij}] \in M_{m\times m} \; donde \; r_{ij} = \frac{cov(F_i, F_j)}{\sqrt{var(F_i) var(F_j)}} \tag{9}$$

Puesto que la matriz de correlaciones es simétrica, resulta diagonalizable y sus valores propios $\lambda_i$, verifican la ecuación 10.

$$\sum_{i=1}^{m} \lambda_i = 1 \tag{10}$$

Debido a la propiedad anterior, estos $m$ valores propios reciben el nombre de pesos de cada uno de los $m$ componentes principales.

## 1.5. VALIDACIÓN DE CLUSTERING

En el caso de estudio, realizar algoritmos de clustering utilizando diferentes métodos requiere de la inspección visual de los resultados que depende del actor que la realiza y no es absolutamente confiable a la hora de elegir la mejor clasificación. Por esto, se hace necesario utilizar algunas métricas para la validación que cuantifiquen que tan bueno es el método implementado.

La principal finalidad del clustering es agrupar objetos similares dentro del mismo cluster y a los objetos diferentes ubicarlos en otros grupos. Por esta razón, se realiza validación interna que se basa en dos criterios: cohesión y separación, considerados en esta investigación como los principales índices de desempeño. Cabe resaltar que no existe un patrón definido a partir del cual los valores de cohesión y separación se evalúen como buenos o malos, por lo que la validación

por medio de estas técnicas busca siempre el mínimo valor para la cohesión y el máximo valor para la separación dependiendo de los resultados comparados.

### 1.5.1. Cohesión

Índice donde se analiza que cada miembro del grupo debe ser lo más cercano posible a los otros miembros del mismo cluster. Para calcularlo se utiliza la denominada Sum of Squared Within (SSW) o suma del cuadrado interior. Su fórmula está dada por la ecuación 11.

$$\mathrm{SSW} = \sum_{i=1}^{k} \sum_{x \epsilon C_i} dist^2(m_i, x) \quad (11)$$

Siendo $k$ el número de clusters, $x$ un punto del cluster $C_i$ y $m_i$ el centroide del cluster $C_i$.

### 1.5.2. Separación

Su concepto se basa en que cada cluster debe estar lo más distante posible del otro. Entre mayor sea la distancia se interpreta como un mejor método. Existen varios enfoques para medir esta distancia entre los cluster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides. En vista de que es una medida de separación, se utiliza Sum of Squared Between (SSB) para evaluar la distancia inter-cluster. Su fórmula se muestra en la ecuación 12.

$$\mathrm{SSB} = \sum_{j=1}^{k} n_j dist^2(c_j - \overline{x}) \quad (12)$$

Siendo $k$ el número de clusters, $n_j$ el número de elementos en el cluster $j$, $c_j$ el centroide del cluster $j$ y $\overline{x}$ es la media del data set.

En trabajos previos se encuentra que ningún índice de validación es fiable por sí solo y además se evidencia que el óptimo puede descubrirse sólo con la comparación entre diferentes resultados de diferentes métodos. Para esto se propone un indicador de comparación visual que detalla la correspondencia de la clase “agua”, claramente identificable en mapas genéricos (Street Map de QGIS y el mapa de Corponariño).

### 1.5.3. Visualización

Para evidenciar gráficamente los resultados de los diferentes algoritmos, es necesario trabajar con una herramienta práctica que posea una georreferenciación, es decir, que permita manipulación de información geográfica, y que contenga un sistema de manejo de base de datos. En vista de que la base de datos proporcionada por el equipo de trabajo de Alternar posee por cada pixel una latitud y longitud con un vector de valores de reflectancia para todas las bandas, se utiliza QGIS, que es un sistema de información geográfica (SIG) libre y de código abierto. En esta aplicación se puede crear, editar, visualizar y publicar información geoespacial en Windows, Mac, Linux y BSD, y soporta numerosos formatos y funcionalidades de datos vector, ráster y bases de datos.

QGIS posee una interfaz amigable donde se puede encontrar marcadores espaciales, identificar y seleccionar los grupos que se quieren visualizar, se pueden etiquetar atributos o grupos, así como verlos y editarlos, como también guardar proyectos. Además permite la creación, edición y exportación de datos espaciales, entre muchas otras funcionalidades.

Una de las principales características de los SIG, es la posibilidad de ubicar unívocamente cualquier punto dentro de la superficie terrestre. Como se mencionó anteriormente, esto se hace con la ayuda del par de coordenadas $X$ e $Y$ que pueden ser coordenadas geográficas, cartesianas o planas. Por esta razón es necesario tener claro la extensión de la zona de estudio y el sistema de referencia de coordenadas que se trabaja. Para este caso se elige a Nariño con el sistema de referencia de coordenadas con proyección EPSG (European Petroleum Survey Group) 3857.

## 2. METODOLOGÍA Y RESULTADOS

Teniendo en cuenta la gran cantidad de datos necesarios para representar el departamento de Nariño, se opta por considerar una zona de estudio representativa (en aspectos de biodiversidad y dimensión territorial), en la que se apliquen los algoritmos de clustering y los mejores resultados se puedan extrapolar a todo Nariño. De esta manera, la metodología propuesta para el presente trabajo de grado se resume en la Figura 9, teniendo un proceso que se desarrolla en cuatro etapas con el fin de obtener el mapa final de clasificación de coberturas del departamento de Nariño. Posteriormente se detalla cada uno de los procesos.

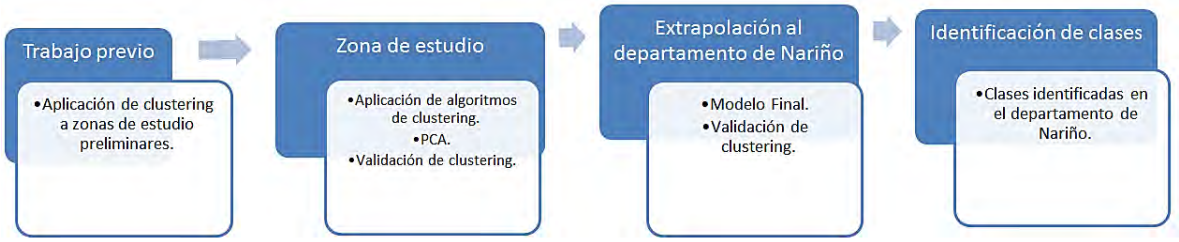



**Figura 9.** Etapas de la metodología desarrollada en el presente trabajo de investigación.

La aplicación de algoritmos en esta investigación tiene como resultado final un archivo de extensión *.csv*, que contiene los datos de la ubicación geográfica de cada pixel y la clase a la que pertenece.

### 2.1. VALIDACIÓN DE CLUSTERING

Para seleccionar los mejores resultados de la aplicación de algoritmos de clustering, se define dar un mismo peso cuantitativo a los índices de desempeño. De esta manera, cada prueba realizada presenta un porcentaje de rendimiento descrito por la ecuación 13.

$$\%Rendimiento\ prueba\ X\ = \left(\frac{MC\times 0{,}5}{CX} + \frac{SX\times 0{,}5}{MS}\right)\times 100 \qquad (13)$$

Donde se calcula el porcentaje de rendimiento de una prueba determinada, $MC$ y $MS$ representan los mejores valores de cohesión y separación de todas las pruebas que se comparan entre sí, y $CX$ y $SX$ son los valores de cohesión y separación de la prueba en cuestión. Por tanto aquella prueba que posee como índices de desempeño la mejor cohesión y la mejor separación, presentará un porcentaje de rendimiento del 100%.

Luego, se selecciona la prueba con el más alto porcentaje de rendimiento y se visualiza su resultado en QGIS. Como un indicador adicional (indicador de comparación visual con mapas de referencia) se tiene en cuenta la comparación de los ríos (grupo "*agua*") del resultado con los del mapa genérico de Street Map (herramienta de QGIS) y/o el mapa propuesto por Corponariño, dado que los ríos son elementos de referencia fácilmente comparables. Además, se tiene en cuenta qué tan compactos y definidos son los clusters encontrados, para determinar si la prueba es el resultado con mejor fitness. De lo contrario se continúa con el siguiente porcentaje más alto y se efectúa el mismo procedimiento.

En el caso de la zona de estudio, para cada algoritmo aplicado se eligen dos mejores resultados: uno en el que los datos de entrada corresponden a la base de datos original y el segundo al resultado obtenido por PCA. Para cada algoritmo de clustering se analiza el uso de componentes principales y se determina la viabilidad de extrapolación a todo el departamento.

## 2.2. TRABAJO PREVIO

En esta etapa del proyecto se desarrollan dos fases principales que se describen a continuación.

### 2.2.1. Pre-procesamiento

El proceso de obtención de datos se realiza tomando imágenes satelitales libres provistas por el sensor Landsat 8. Para cubrir el departamento de Nariño en su totalidad, con una extensión aproximada de 33.268 $km^2$ [35], se hizo necesario descargar imágenes de cinco escenas diferentes como se muestra en la Figura 10 con los respectivos identificadores (Path ID y Row ID) de las imágenes satelitales utilizadas. Se tomaron varias imágenes satelitales de cada escena, de diferentes días de los años 2015 y 2016, con el fin de obtener las imágenes más limpias posibles. Luego se realiza un trabajo básico de procesamiento sobre las imágenes adquiridas que incluye un filtrado espacial y temporal para la eliminación de nubosidad y pixeles no válidos [36]. En primer lugar, dada la extensión del área de estudio, las escenas descargadas tenían diferentes sistemas de coordenadas (EPSG: 32618 y EPSG: 32617). Por motivos de visualización se decide unificar el sistema de coordenadas usando EPSG: 3857. Muchas de las escenas cubren una gran área del Océano Pacífico así como de otros departamentos de la región, por tanto, se recortó las imágenes para obtener solo los datos referentes al departamento de Nariño.

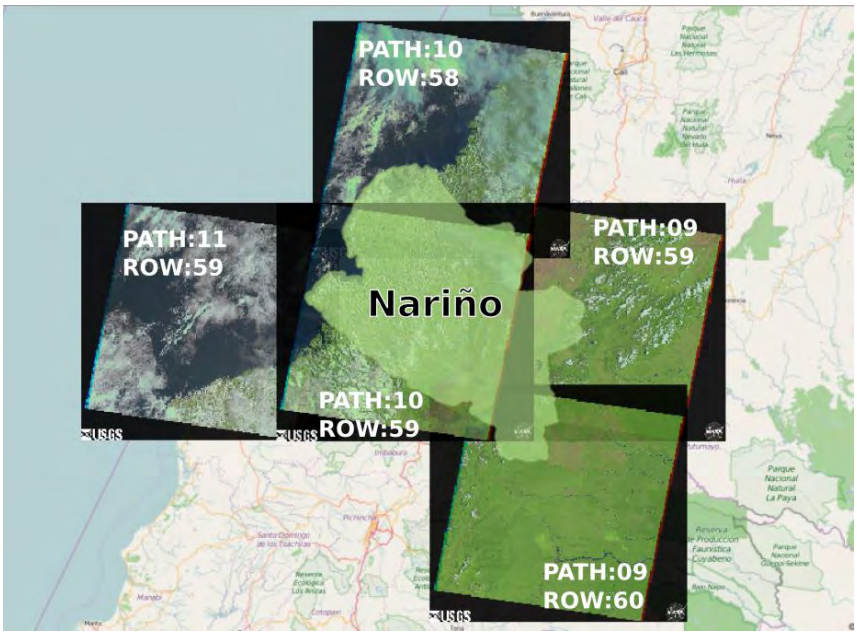



**Figura 10**. Escenas seleccionadas para descargar imágenes satelitales que representan el departamento de Nariño. **Fuente:** [37].

El resultado final del recorte de las imágenes se ilustra en la Figura 11. Luego del recorte, se tiene una matriz de características con localización y reflectancia para las 7 bandas espectrales por cada pixel que constituye la entrada al procesamiento y ejecución de los algoritmos de clustering.

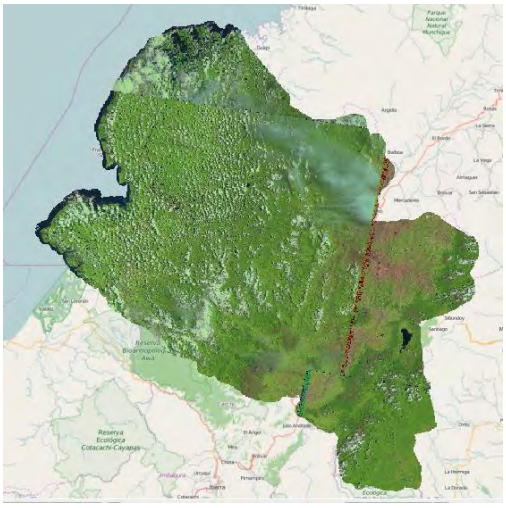


**Figura 11.** Efecto del recorte de las imágenes satelitales para obtener la información que representa únicamente el departamento de Nariño. **Fuente:** [37].

Posteriormente, con el objetivo de organizar los datos adquiridos se hace uso de una base de datos. El procesamiento de las imágenes y alimentación de la base

de datos se realiza a través de scripts y archivos procesados por lotes, mediante el uso de Python. Entre los procesos se transforma los valores originales (o digital numbers) a su correspondiente valor de reflectancia solar, usando la información presente en el archivo MTL de la imagen satelital, la banda Quality (BQA) para filtrar nubosidad y se ejecuta el algoritmo ACCA [38]. Finalmente, para almacenar los valores de reflectancia de todas las imágenes se usa el gestor de bases de datos Postgres SQL, que facilita las consultas para obtener un promedio de todas las imágenes correspondientes a una misma escena. Con estos resultados se conforma la base de datos final de todo el departamento, correspondiente a la matriz de características con la que trabaja la presente investigación. A partir del proceso anterior, se toman las primeras 7 bandas de la imagen satelital, puesto que las bandas 8 en adelante, según [39], se usan para otros estudios como mapeo térmico y humedad estimada del suelo. Sin embargo, las bandas superiores se usan en esta etapa de pre-procesamiento para hacer correcciones y filtrado de las imágenes y los cálculos necesarios para determinar los valores de reflectancia. Para mayor información, se sugiere revisar [35], [37].

### 2.2.2. Aplicación de clustering a zonas de estudio preliminares

Inicialmente se trabajó con varias bases de datos preliminares correspondientes a zonas de los municipios de San Juan de Pasto, Sandoná y Tumaco, que se seleccionan teniendo en cuenta la calidad de los datos obtenidos y la representación de los municipios de todo el departamento. La región de San Juan de Pasto se elige porque define claramente grupos como *“agua”* y *“zona urbana”.* Por su parte, Sandoná es una de las zonas que posee los datos más limpios debido a que presenta menor nubosidad en las imágenes satelitales. Finalmente, Tumaco reúne las dos características mencionadas, además de poseer una hidrología representativa.

Para estas zonas de estudio preliminares, cada pixel tiene datos de reflactancia para cada una de las 7 bandas espectrales, y otras características como temperatura, altura, NDVI y msi, que se consideran atributos tentativos para identificar tipos de vegetación mediante clustering. La intención de este proceso fue probar algunos algoritmos de clustering consultados hasta el momento, como K-means, Dbscan, Fuzzy C-Means, Mean-Shift, con el fin de conocer su comportamiento, determinar los efectos en variación del número de clases, analizar la distancia utilizada para minimizar la separación entre los individuos de una misma clase y evidenciar los resultados de combinar algunas de las características. Finalmente, al tener varias zonas de diferentes municipios, se estudian los resultados de las pruebas y por medio de la validación de clustering

se escoge una zona del departamento como la zona prototipo de estudio de la investigación.

Para mostrar las pruebas y los resultados más relevantes de esta sección, primero se muestran en tablas y mapas las tres bases de datos principales sobre las que se aplicó clustering como trabajo previo, como se especifican en la Tabla II. Las zonas de aplicación se muestran en la Figura 12.

TABLA II
BASES DE DATOS DE LAS ZONAS PRELIMINARES

| Municipio | No. de pixeles | No. de características | Características |
|---|---|---|---|
| Pasto | 2’019.046 | 13 | “latitude”, “longitude”, “b1”, “b2”, “b3”, “b4”, “b5”, “b6”, “b7”, “temperatura”, “altura”, “ndvi”, “msi” |
| Sandoná | 118.462 | 10 | “latitude”, “longitude”, “b1”, “b2”, “b3”, “b4”, “b5”, “b6”, “b7”, “ndvi” |
| Tumaco | 2’500.000 | 9 | “latitude”, “longitude”, “b1”, “b2”, “b3”, “b4”, “b5”, “b6”, “b7” |

La característica *“temperatura”* hace referencia a la temperatura en grados Kelvin del pixel*, “altura”* es la altura identificada en el punto exacto del pixel en m.s.n.m., *“ndvi”* es el índice de vegetación de diferencia normalizada usado para estimar la cantidad, calidad y desarrollo de la vegetación con base a la medición por medio de sensores remotos, instalados desde una plataforma espacial, y *“msi”* es el índice de estrés hídrico que también funciona como estimador de cambios de diferentes tipos de cobertura vegetal.

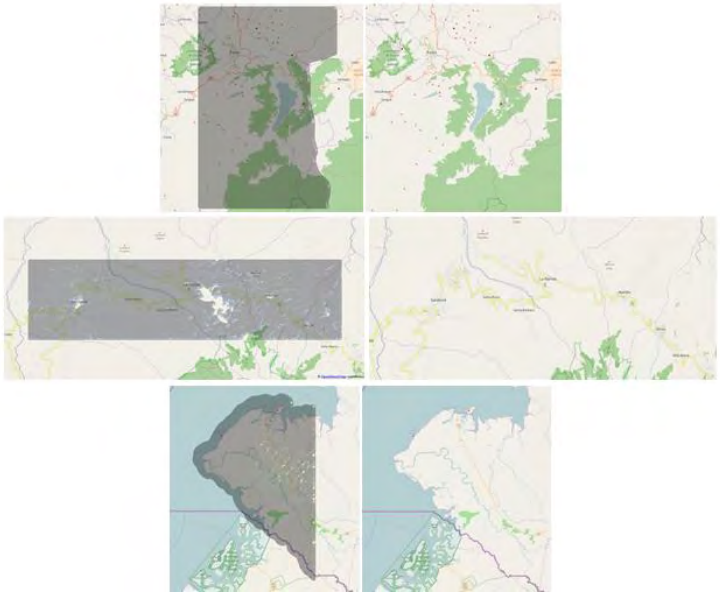


**Figura 12.** Zonas de aplicación preliminares. De arriba hacia abajo, San Juan de Pasto, Sandoná y Tumaco. A la izquierda el área que cubren las bases de datos expuestas en la Tabla II y a la derecha las correspondientes zonas de prueba en QGIS.

Inicialmente se consideró que el municipio de San Juan de Pasto era el lugar adecuado para seleccionar como zona de estudio, por lo que se inició con la prueba de clustering sobre esta zona. En la Tabla III y en las Figuras 13, 14 y 15, se presentan las pruebas más importantes. Los resultados permitieron el dominio de algunos algoritmos y obtener conclusiones de ellos.

En las Tablas III, IV y V, la columna “No. de clusters” hace referencia a la cantidad de grupos que se define a priori a la ejecución de los algoritmos.

TABLA III
PRUEBAS DE CLUSTERING PRELIMINARES EN ZONA DE ESTUDIO: SAN JUAN DE PASTO

| Prueba No. | Algoritmo | Datos de entrada | No. de clusters | Especificaciones del algoritmo | Cohesión | Separación |
|---|---|---|---|---|---|---|
| 1 | K-Means | Todas las características | 5 | Distancia: cityblock, MaxIter = 100 | 31.240’310.000 | $2{,}7773060e^{11}$ |
| 2 | K-Means | Todas las características | 5 | Distancia: correlation, MaxIter = 100 | 45.405’750.000 | $2{,}63565e^{11}$ |
| 3 | K-Means | Todas las características | 5 | Distancia: sqeuclidean, MaxIter = 100 | 25.751’160.000 | $2{,}8322e^{11}$ |
| 4 | K-Means | Todas las características | 25 | Distancia: sqeuclidean, MaxIter = 100 | 1.423’944.000 | $3{,}07547e^{11}$ |
| 5 | K-Means | Todas las características normalizadas | 25 | Distancia: sqeuclidean, MaxIter = 100 | 551.134,7 | 74.554,8 |
| 6 | K-Means | 7 Bandas espectrales | 5 | Distancia: sqeuclidean, MaxIter = 100 | 46.409,79 | 603.817,60 |
| 7 | K-Means | Todas las características | 10 | Distancia: sqeuclidean, MaxIter = 100 | 7.639’985.000 | 30.133’090.000 |
| 8 | K-Means | Todas las características, altura normalizada | 10 | Distancia: sqeuclidean, MaxIter = 100 | 2’686.244 | 152’866.800 |
| 9 | K-Means | 7 Bandas espectrales, ndvi, msi | 10 | Distancia: sqeuclidean, MaxIter = 100 | 51.629,04 | 745.385,80 |
| 10 | K-Means | 7 Bandas espectrales | 10 | Distancia: sqeuclidean, MaxIter = 100 | 24.513,60 | 625.713,80 |
| 11 | GK | Todas las características | 10 | K = 10, pi = 5, MaxIter = 100, Tol = $1e^{-7}$ | $1{,}821272e^{13}$ | $1{,}790376e^{13}$ |
| 12 | GK | Todas las características normalizadas | 10 | K = 10, pi = 5, MaxIter = 100, Tol = $1e^{-7}$ | $8{,}655427e^{5}$ | $2{,}153153e^{5}$ |
| 13 | EM | Todas las características | 5 | Tol = $1e^{-12}$, iter = 5 | $3{,}05875e^{11}$ | 479’859.700 |
| 14 | EM | 7 Bandas espectrales | 5 | Tol = $1e^{-12}$, iter = 5 | 632.377,60 | 598,1885 |
| 15 | FCM | Todas las características | 5 | Exp = 2, MaxIter = 100, Criterio = $1e^{-5}$ | $3{,}01937e^{11}$ | 7.033’650.000 |
| 16 | FCM | Todas las características normalizadas | 5 | Exp = 2, MaxIter = 100, Criterio = $1e^{-5}$ | 127.639,60 | 477.208 |
| 17 | FCM | 7 Bandas espectrales | 5 | Exp = 2, MaxIter = 100, Criterio = $1e^{-5}$ | 45.552,06 | 609.900,1 |

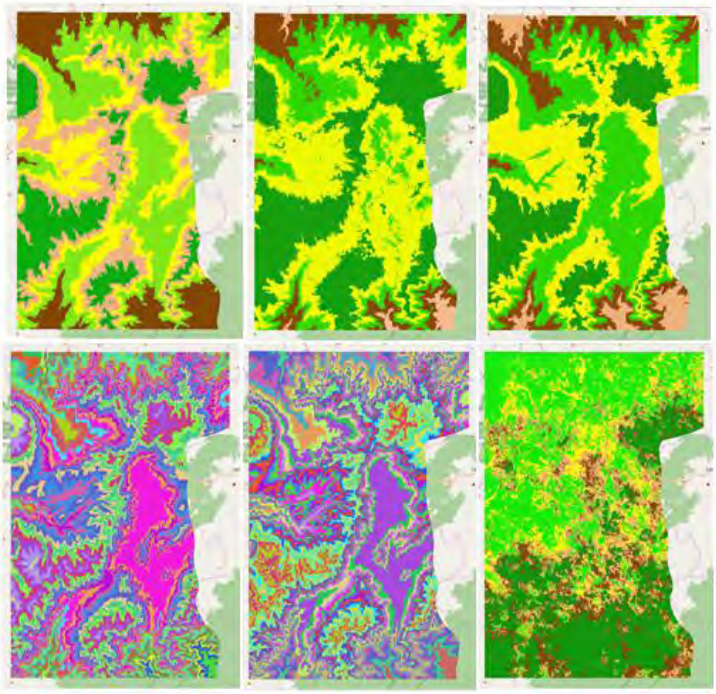


**Figura 13.** Resultado gráfico de la aplicación de k-means sobre el municipio de San Juan de Pasto, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas 1 a 6 expuestas en la Tabla III.

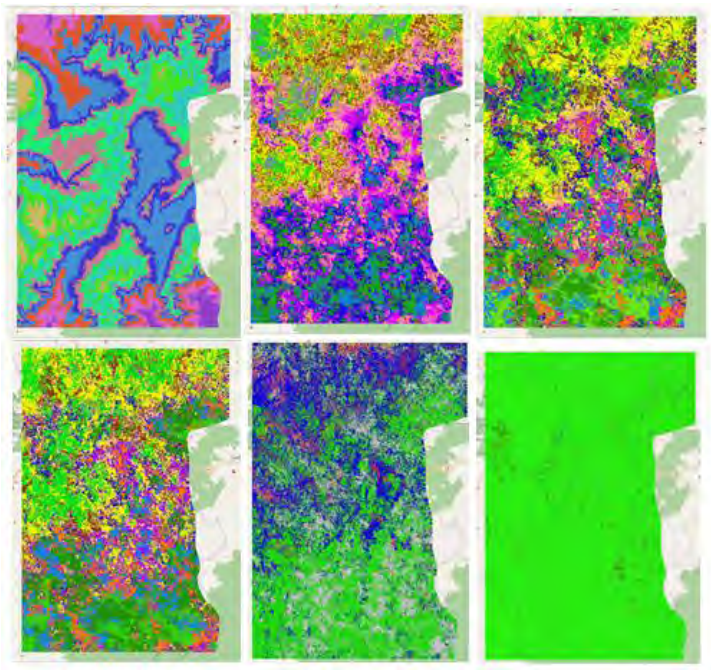


**Figura 14.** Resultado gráfico de la aplicación de k-means y GK sobre el municipio de San Juan de Pasto, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, de las pruebas 7 a 12 expuestas en la Tabla III.

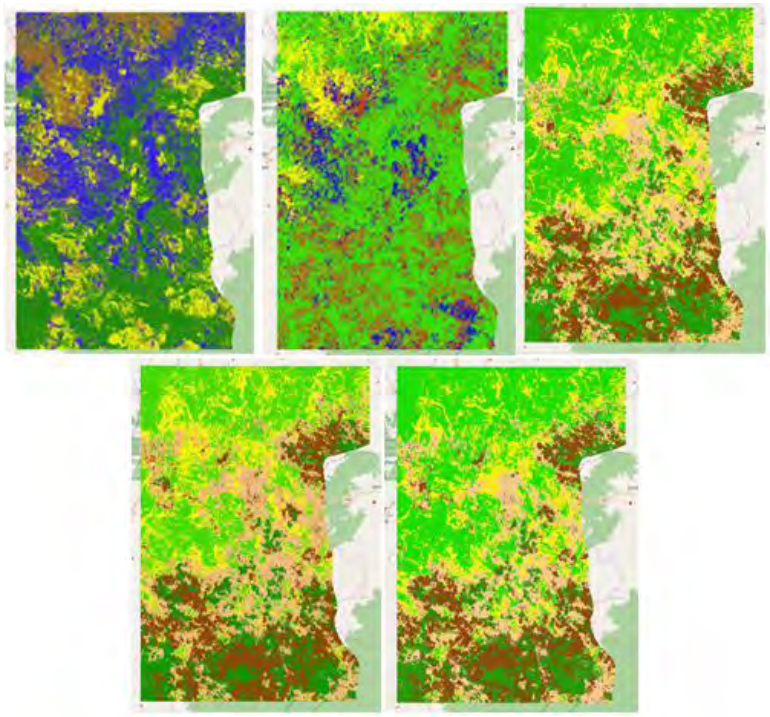


**Figura 15.** Resultado gráfico de la aplicación de EM y FCM sobre el municipio de San Juan de Pasto, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, de las pruebas 13 a 17 expuestas en la Tabla III.

Enseguida para la zona preliminar de Sandoná, las pruebas realizadas y sus resultados se muestran en la Tabla IV y en la Figura 16.

TABLA IV
PRUEBAS DE CLUSTERING PRELIMINARES EN ZONA DE ESTUDIO: SANDONÁ

| Prueba No. | Algoritmo | Datos de entrada | No. de clusters | Especificaciones del algoritmo | Cohesión | Separación |
|---|---|---|---|---|---|---|
| 1 | K-Means | 7 Bandas espectrales | 2 | Distancia: sqeuclidean, MaxIter = 100 | 591,5574 | 525,9952 |
| 2 | K-Means | 7 Bandas espectrales | 5 | Distancia: sqeuclidean, MaxIter = 100 | 340,7434 | 776,8093 |
| 3 | K-Means | Todas las características | 2 | Distancia: sqeuclidean, MaxIter = 100 | 1.202,7350 | 760,9295 |
| 4 | K-Means | Todas las características | 5 | Distancia: sqeuclidean, MaxIter = 100 | 570,1891 | 1.393,4760 |
| 5 | GK | Todas las características | 6 | K = 10, pi = 5, MaxIter = 100, Tol = $1e^{-7}$ | 1.963,6650 | $8{,}11619e^{-20}$ |
| 6 | GK | 7 Bandas espectrales | 6 | K = 10, pi = 5, MaxIter = 100, Tol = $1e^{-7}$ | 1.117,553 | $9{,}13125e^{-20}$ |
| 7 | Mean -- Shift | 7 Bandas espectrales | X | Criterio de mezcla = 0,1 | X | X |
| 8 | EM | Todas las características | 5 | Tol = $1e^{-12}$, iter = 5 | 1.675,2700 | 86,23849 |
| 9 | EM | 7 Bandas espectrales | 5 | Tol = $1e^{-12}$, iter = 5 | 1.079,8800 | 14,99243 |

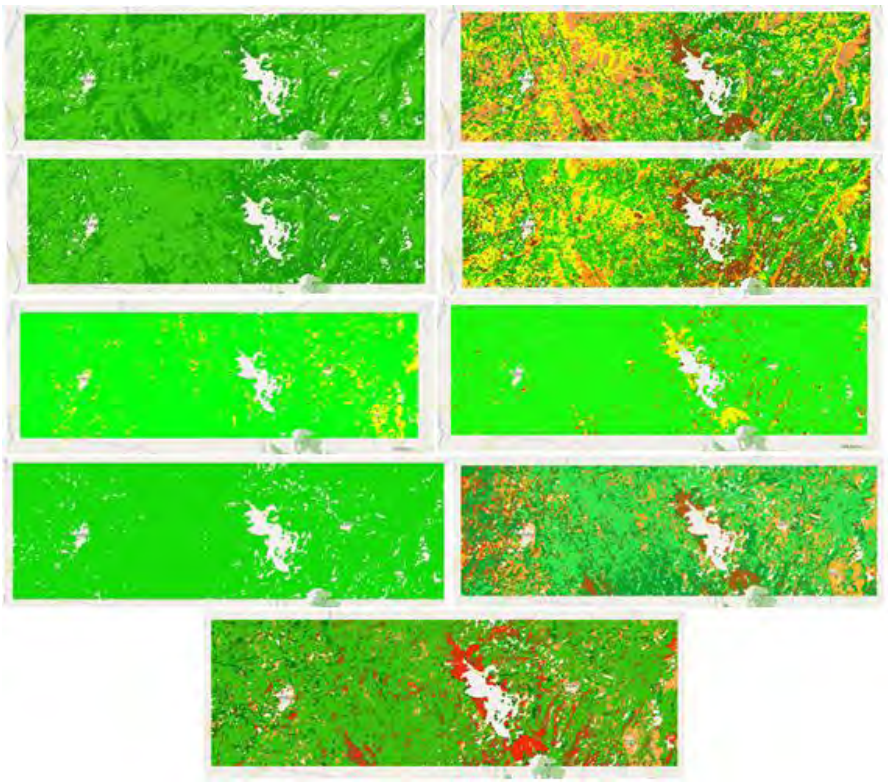


**Figura 16.** Resultado gráfico de la aplicación de k-means, GK, Mean Shift y EM sobre el municipio de Sandoná, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas expuestas en la Tabla IV.

Finalmente para la zona preliminar del municipio de Tumaco, se efectuaron las pruebas que se describen en la Tabla V y sus respectivas visualizaciones en las Figuras 17 y 18.

TABLA V
PRUEBAS DE CLUSTERING PRELIMINARES EN ZONA DE ESTUDIO: TUMACO

| Prueba No. | Algoritmo | Datos de entrada | No. de clusters | Especificaciones del algoritmo | Cohesión | Separación |
|---|---|---|---|---|---|---|
| 1 | K-Means | 7 Bandas espectrales | 5 | Distancia: sqeuclidean, MaxIter = 100 | 16.311,370 | 70.856,460 |
| 2 | K-Means | 7 Bandas espectrales normalizadas | 5 | Distancia: sqeuclidean, MaxIter = 100 | 40.253,830 | 209.237,000 |
| 3 | K-Means | Bandas espectrales 3, 4 y 5 | 5 | Distancia: sqeuclidean, MaxIter = 100 | 9.151,542 | 5.3621,130 |
| 4 | K-Means | 7 Bandas espectrales | 10 | Distancia: sqeuclidean, MaxIter = 100 | 7.867,040 | 79.300,790 |
| 5 | K-Means | Bandas espectrales 3, 4 y 5 | 10 | Distancia: sqeuclidean, MaxIter = 100 | 4.709,389 | 58.063,280 |
| 6 | GK | 7 Bandas espectrales | 5 | K = 10, pi = 5, MaxIter = 100, Tol = $1e^{-7}$ | 87.167,1 | $1{,}04191e^{-8}$ |
| 7 | Dbscan | 7 Bandas espectrales | 5 | MinPts = 100.000, eps = 0,001 | 8.713,541 | 32,41689 |
| 8 | EM | 7 Bandas espectrales | 5 | Tol = $1e^{-12}$, iter = 5 | 81.802,370 | 3.832,237 |
| 9 | EM | 7 Bandas espectrales | 10 | Tol = $1e^{-6}$, iter = 10 | 105.286,900 | 124.037,900 |

| CONTINUACIÓN TABLA V | | | | | | |
|---|---|---|---|---|---|---|
| 10 | EM | 7 Bandas espectrales normalizadas | 5 | Tol = $1e^{-12}$, iter = 5 | 234.183,800 | 6.938,127 |
| 11 | FCM | 7 Bandas espectrales | 5 | Exp = 2, MaxIter = 100, Criterio = $1e^{-5}$ | 16.614,460 | 71.534,340 |
| 12 | FCM | 7 Bandas espectrales | 7 | Exp = 2, MaxIter = 100, Criterio = $1e^{-5}$ | 11.734,550 | 7.565,200 |
| 13 | FCM | 7 Bandas espectrales normalizadas | 5 | Exp = 2, MaxIter = 100, Criterio = $1e^{-5}$ | 41.528,760 | 215.353,800 |

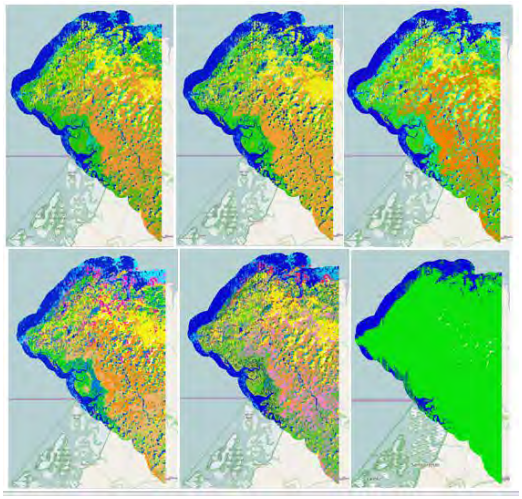


**Figura 17.** Resultado gráfico de la aplicación de k-means y GK sobre el municipio de Tumaco, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas 1 a 6 expuestas en la Tabla V.

Para la interpretación de resultados en esta investigación es necesario tener en cuenta que los índices de desempeño son solo uno de los factores de evaluación de las pruebas, y que estos no tienen un patrón de comparación, si no que se evalúa su rendimiento de acuerdo al valor más pequeño de cohesión y grande de separación presentado en las pruebas realizadas.

El análisis de los resultados anteriores es el siguiente:

- Se determinó que el incremento en el número de grupos ocasionaba que los individuos se dispersen demasiado haciendo que los clusters no sean compactos, aspecto reflejado en los índices de desempeño y la comparación visual con los mapas de referencia.

- Se evidenció que para el caso de k-means resulta eficiente el uso de la distancia euclidiana, puesto que no solo mejora el porcentaje de rendimiento sino que también considera con igual importancia todas las características.

- En cuanto a la combinación de diferentes características, se determinó que resulta conveniente trabajar únicamente con las bandas espectrales, puesto

que los otros índices contienen información redundante al ser obtenidos de cálculos efectuados a partir de las bandas. Si se hace una comparación entre seleccionar solo las bandas espectrales y con todas las características y sus combinaciones, las dos últimas distorsionaron los grupos compactos, visualizándose una diseminación de las diferentes clases. Por ejemplo, se encontró que la característica de altura, al tener un valor grande, predomina sobre el resto de atributos a pesar de haberse aplicado la normalización. Como resultado la altura predomina arrojando un mapa de clasificación en donde se visualizan curvas de nivel, es decir, una clasificación más por altura que por vegetación, como se observa en las Figuras 13 y 14.

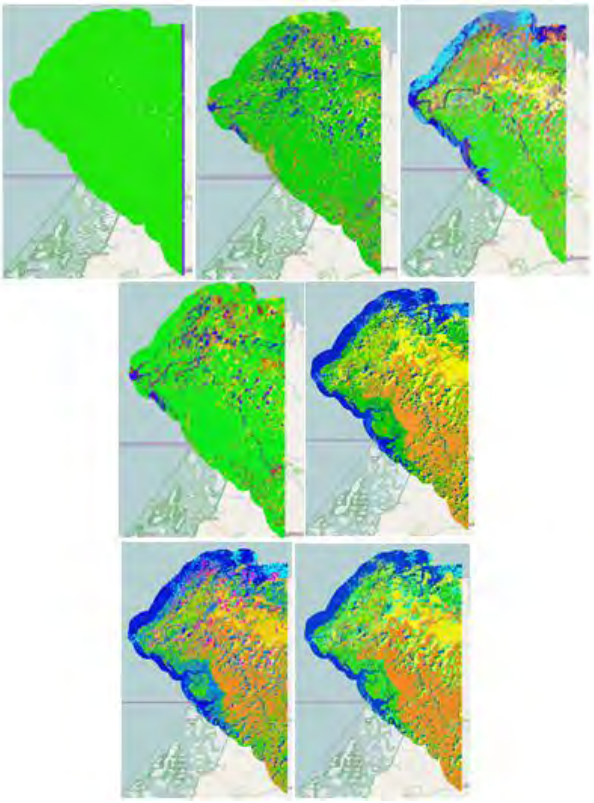


**Figura 18.** Resultado gráfico de la aplicación de Dbscan, EM y FCM sobre el municipio de Tumaco, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas 7 a 13 expuestas en la Tabla V.

- Adicionalmente se probaron los algoritmos Dbscan, mean-shift y región growing. La ejecución de Dbscan no fue conveniente debido a problemas con la memoria requerida para procesamiento de los datos y el tiempo de ejecución del algoritmo. Puesto que Dbscan visita cada punto de la base de datos, la complejidad temporal radica en que se ejecuta exactamente una consulta por cada punto, y en este caso, debido a la cantidad de datos que

se manejan, la matriz de distancias es de tamaño considerable y se construye para evitar recalcularlas en intermedios de la ejecución. Para el caso del algoritmo Mean-Shift, se tuvieron problemas frente al gasto computacional y la memoria requerida para la ejecución del algoritmo. Sin embargo, se aplicó al municipio de Sandoná y como se aprecia en la Figura 16, no produce los mejores resultados puesto que aunque se variaban los parámetros, el algoritmo asignó la misma clase para todos los miembros del "data set". El algoritmo Region-Growing se trató de implementar, pero las codificaciones encontradas no son compatibles con las bases de datos trabajadas, especialmente porque es un algoritmo de segmentación de imágenes por regiones y utiliza como entrada una imagen. Por lo tanto se registra su consulta, pero su implementación de acuerdo a los requerimientos de esta investigación no es satisfactoria. Por las razones anteriores, estos tres algoritmos no se usan más adelante.

- Como resultado del pre-procesamiento, de las conclusiones encontradas por el proyecto Alternar y a partir del trabajo anterior, se define que las imágenes satelitales para Pasto no son las mejores. En su mayoría, no están limpias por la elevada nubosidad. En Sandoná los resultados encontrados no se pueden interpretar fácilmente, debido a que la zona se trabajó con una base de datos filtrada, y además no se puede apreciar el comportamiento de los algoritmos al identificar el cluster que define el grupo *"agua".* Por su parte, las imágenes satelitales de Tumaco presentan una caracterización más despejada de la zona y la región contiene gran cantidad de ríos, característica que permite detallar mejor el rendimiento de los algoritmos y sus resultados son más coherentes con la realidad.

A partir de los resultados, la calidad de los datos (nubosidad), la extensión territorial, los fines de esta investigación que se basan en aportar información para el análisis de oportunidades energéticas de zonas no interconectadas de la región, se define el municipio de Tumaco como el área de estudio, ya que cumple con las características mencionadas y además posee gran cantidad de biomasa que podría tener correlación con gran parte de la vegetación del departamento de Nariño, siendo viable extrapolar sus resultados. En la Figura 19 se observa la división política de Nariño, destacando el municipio de Tumaco y una representación del espacio de trabajo en QGIS para la visualización de resultados de cada una de las pruebas. De esta manera, en la zona de estudio se usa una matriz de características de 3'971.007 pixeles (un 11,6% del total de datos del departamento) por 9 atributos, correspondientes a las coordenadas, latitud y longitud, y las 7 primeras bandas espectrales del satélite Landsat 8. Además se establece que los algoritmos a trabajar en esta zona son: k-means, fuzzy c-means, EM y Gustafson Kessel.

**Figura 19.** Municipio de Tumaco (Nariño) - Colombia. A la izquierda el municipio de Tumaco y a la derecha la zona de prueba en QGIS.

## 2.3. ZONA DE ESTUDIO

A partir del trabajo anterior y gracias a la colaboración del proyecto Alternar se determina el municipio de Tumaco como la zona de estudio a trabajar, en la que se realizan las pruebas necesarias para establecer el mejor modelo a ser extrapolado a todo Nariño.

La metodología utilizada para esta etapa se muestra en la Figura 20, teniendo en cuenta que se seleccionan únicamente los datos de cada municipio de la base de datos final de todo el departamento.

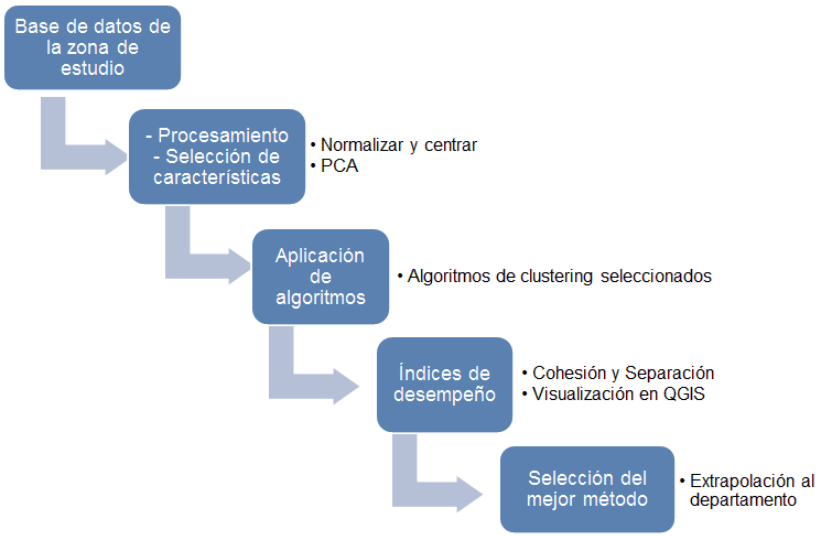



**Figura 20:** Metodología utilizada para la aplicación de los algoritmos en la zona de estudio, con el fin de seleccionar los mejores resultados.

### 2.3.1. Procesamiento

Para cada uno de los algoritmos de clustering implementados se realiza una etapa de procesamiento constituida por dos elementos principales: la normalización y el centrado de los datos. Este procesamiento del data set se efectúa ya que los datos podrían tener valores de escalas diferentes entre sí.

- **Normalización de los datos:** se ajustan los valores medidos en diferentes escalas, respecto a una escala común. En este caso se normaliza todas las características en el rango de 0 a 1 usando la expresión descrita en la ecuación 14.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{14}$$

  Donde $X'$ representa la característica normalizada, $X$ el valor de la característica original, $X_{min}$ y $X_{max}$ son los valores mínimo y máximo respectivamente, de todo el conjunto de valores conformado por la característica en cuestión para cada uno de los individuos.

- **Centrado de los datos:** se ajustan los valores medidos teniendo en cuenta el promedio en cada una de las características como centro de las mismas, a partir de la siguiente expresión:

$$Xc = X - \bar{X} \tag{15}$$

  Donde, $Xc$ representa la característica centrada, $X$ el valor de la característica original y $\bar{X}$ es el valor promedio de la característica en cuestión.

Para la ejecución de los diferentes algoritmos de clustering, el procesamiento se varía. En primer lugar, se prueba la base de datos original. Seguidamente, se prueba por separado cada una de las técnicas de normalizado y centrado, para posteriormente aplicar su combinación. Esto se realiza con el fin de comparar y observar la influencia del procesamiento en los resultados obtenidos.

### 2.3.2. PCA (Principal Component Analysis)

A partir de los datos originales y las técnicas de procesamiento de datos, se ejecuta el procedimiento de PCA. Para ello se usa la función *pca* de Matlab modificando sus argumentos, como el algoritmo utilizado para la prueba, centrado y el número de componentes principales a usar.

Para observar el comportamiento de PCA se opta por calcular el error entre los datos reconstruidos a partir de su salida y los datos originales, usando la ecuación 16 con la función *norm* de Matlab. Se pretende que el error sea lo más pequeño posible con el fin no solo de reducir la dimensión de los datos sino también de sacar el mayor provecho de los datos originales y extraer la información más relevante.

$$ERROR = \frac{(Norma\ matricial\ de\ Frobenius)^2}{n} \times 100\% \quad (16)$$

Una vez se elige el mejor PCA, se calcula el error definiendo el número de componentes principales $k$, desde 1 hasta $m-1$, con el fin de definir a través del gráfico de los autovalores, cuántas componentes es conveniente utilizar en el estudio.

En cuanto a los resultados, en primer lugar se aplica PCA y se deja por defecto el número de componentes principales $k = 7$. Los resultados para determinar la mejor configuración de parámetros y por ende, elegir el mejor PCA, se resumen en la Tabla VI.

TABLA VI
PRUEBAS DE PCA CON 7 COMPONENTES PRINCIPALES

| Prueba No. | Procesamiento | | 'Algorithm' | 'Centered' | Error (%) |
|---|---|---|---|---|---|
| | Normalizar | Centrar | | | |
| 1 | Si | No | 'svd' | 'true' | $3{,}7612e^{-26}$ |
| 2 | Si | No | 'eig' | 'true' | $2{,}9254e^{-31}$ |
| 3 | Si | No | 'als' | 'true' | $4{,}3454e^{-18}$ |
| 4 | Si | No | 'svd' | 'false' | 26,5130 |
| 5 | Si | No | 'eig' | 'false' | 26,5130 |
| 6 | Si | Si | 'svd' | 'false' | $3{,}7612e^{-26}$ |
| 7 | Si | Si | 'eig' | 'false' | $2{,}9546e^{-31}$ |
| 8 | No | No | 'eig' | 'false' | 24,7198 |

A partir de las pruebas realizadas se determinó que la función de Matlab centra los datos de igual forma que si se centraran con la técnica de procesamiento. Además a partir de las pruebas 1 a 3 de la Tabla VI, se determina que el algoritmo *'als'* no es el mejor en este caso de estudio. Por su parte, el hecho de no centrar los datos genera una diferencia elevada entre los datos originales y el resultado de PCA. Por otro lado, el error obtenido en las pruebas 2 y 7 que conducen a la misma configuración, es el adecuado puesto que la diferencia entre los datos reconstruidos a partir de PCA y los originales es mínima. La prueba 8 se hace, conociendo que *'eig'* es el mejor algoritmo para aplicar PCA y observando el efecto de no emplear un procesamiento en los datos. Se selecciona la prueba 2 como el mejor resultado que presenta el menor error e incluye tanto la

normalización como el centrado de los datos. Por tanto, para la fase de aplicación de PCA en los diferentes algoritmos de clustering, se toma este resultado como base de datos de entrada.

A partir de la prueba 2, se efectúan los cálculos de error, variando el número de componentes principales como se muestra en la Tabla VII. Se puede observar que a medida que el número de componentes principales disminuye, el error aumenta ya que aunque las componentes principales se organizan en orden de importancia, al elegir menos componentes principales, se está perdiendo información de los datos originales. La Figura 21 muestra el gráfico de autovalores, a partir del que se considera que el número de componentes principales a usar para representar adecuadamente los datos está entre 3 y 5 ya que es allí donde el cambio de la pendiente es mayor y el error es menor.

TABLA VII
PRUEBAS DE PCA VARIANDO EL NÚMERO DE COMPONENTES PRINCIPALES

| Prueba No. | Procesamiento | | 'Algorithm' | 'Centered' | No. de componentes principales | Error (%) |
|---|---|---|---|---|---|---|
| | Normalizar | Centrar | | | | |
| 1 | Si | No | 'eig' | 'true' | 6 | $6{,}5621e^{-5}$ |
| 2 | Si | No | 'eig' | 'true' | 5 | $3{,}8793e^{-4}$ |
| 3 | Si | No | 'eig' | 'true' | 4 | 0,0023 |
| 4 | Si | No | 'eig' | 'true' | 3 | 0,0098 |
| 5 | Si | No | 'eig' | 'true' | 2 | 0,1934 |
| 6 | Si | No | 'eig' | 'true' | 1 | 0,8111 |

### 2.3.3. Aplicación de algoritmos

La etapa de aplicación de algoritmos incluye la ejecución de diferentes pruebas de clustering sobre la base de datos de la zona de estudio con diferentes parámetros, e inclusive modificando los datos de entrada mediante su procesamiento, con el fin de seleccionar el mejor resultado de cada algoritmo aplicado.

El proceso de variación de parámetros se hace mediante ensayo y error, donde inicialmente cada algoritmo se prueba sobre la base de datos original, con sus valores por defecto, y posteriormente, se modifica un primer parámetro, con el objetivo de fijar un valor apropiado. En seguida, se efectúa el mismo proceso para los demás parámetros, obteniendo finalmente una configuración que ajuste el resultado a una clasificación adecuada de tipos de cobertura presentes en el municipio.

Como segunda parte de la aplicación de algoritmos, se ejecutan las pruebas enumeradas en la Tabla VIII sobre el resultado obtenido con PCA. En primer lugar se aplica el algoritmo de clustering con un número fijo de 7 componentes

principales (pruebas 1 a 4 de la Tabla VIII) debido al uso de las 7 bandas espectrales y se validan los resultados con los criterios establecidos. Posteriormente, partiendo del resultado anterior se ejecuta la última prueba en la que el número de componentes principales empleado varía de 3 a 5 como se determinó en el gráfico de los autovalores, esto se hace con el fin de analizar la factibilidad de reducir la dimensión de los datos y facilitar la extrapolación a todo el departamento de Nariño.

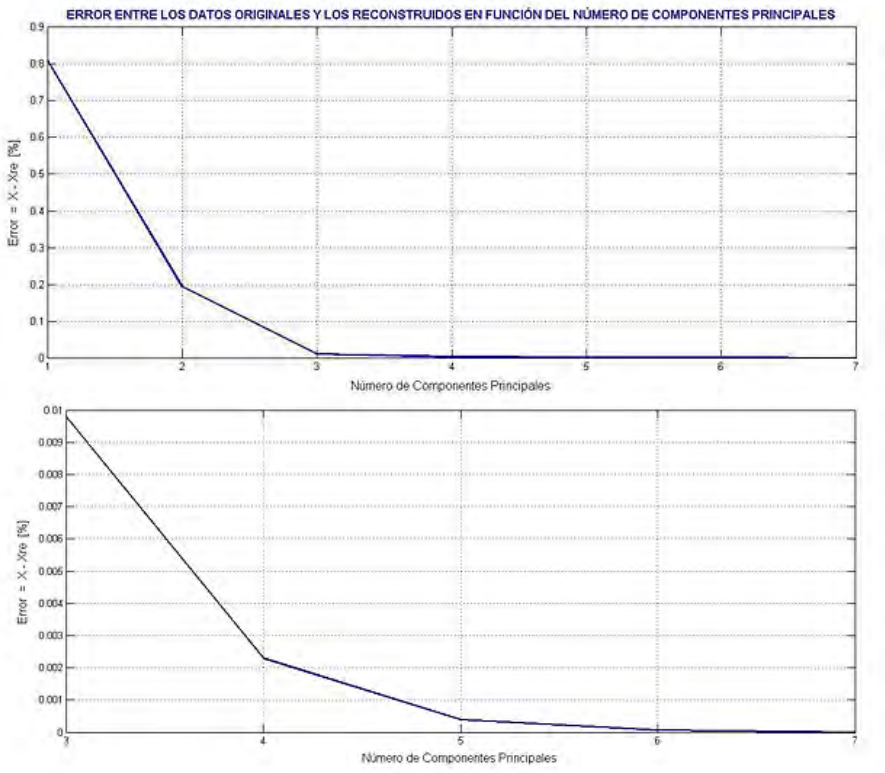



**Figura 21.** Gráfico de los autovalores para determinar el número de componentes principales para aplicar PCA. En la parte superior el gráfico completo y en la parte inferior, un zoom a partir de 3 componentes principales.

TABLA VIII
PRUEBAS A REALIZAR CON CADA ALGORITMO A PARTIR DEL RESULTADO DE PCA

| Prueba No. | Procesamiento | | Número de componentes principales |
|---|---|---|---|
| | Normalizar | Centrar | |
| 1 | No | No | 7 |
| 2 | Si | No | 7 |
| 3 | No | Si | 7 |
| 4 | Si | Si | 7 |
| 5 | | | 3, 4 o 5 |

#### 2.3.3.1. K-Means

Para ejecutar el algoritmo K-Means se usa la función de Matlab *kmeans* puesto que es una función completa y fácil de utilizar.

El algoritmo requiere definir el número de grupos a priori, por tanto se decide ejecutarlo variando el número de clases entre 2 y 13. Este número se escoge dado que según la base de datos facilitada por Corponariño, los municipios no exceden las 13 clases. Se usa una configuración sencilla de la función *kmeans*, donde se emplea la distancia euclidiana y un número máximo de 1000 iteraciones, con los demás parámetros por defecto. Para cada prueba se calculan los índices de desempeño y se hace una gráfica de dichos valores versus el número de clases, determinando un buen desempeño con un número de clases definido.

El resultado de este procedimiento se muestra en la Figura 22. Se observa que a medida que el número de clases aumenta, los índices de desempeño se comportan mejor. Sin embargo, teniendo en cuenta que según el estudio realizado por Corponariño en Tumaco hay alrededor de 6 clases predominantes y la gráfica muestra que el número de clases apropiado se puede considerar entre 4 y 7 clases, se determina que 5 es un número apropiado para realizar las pruebas en la zona.

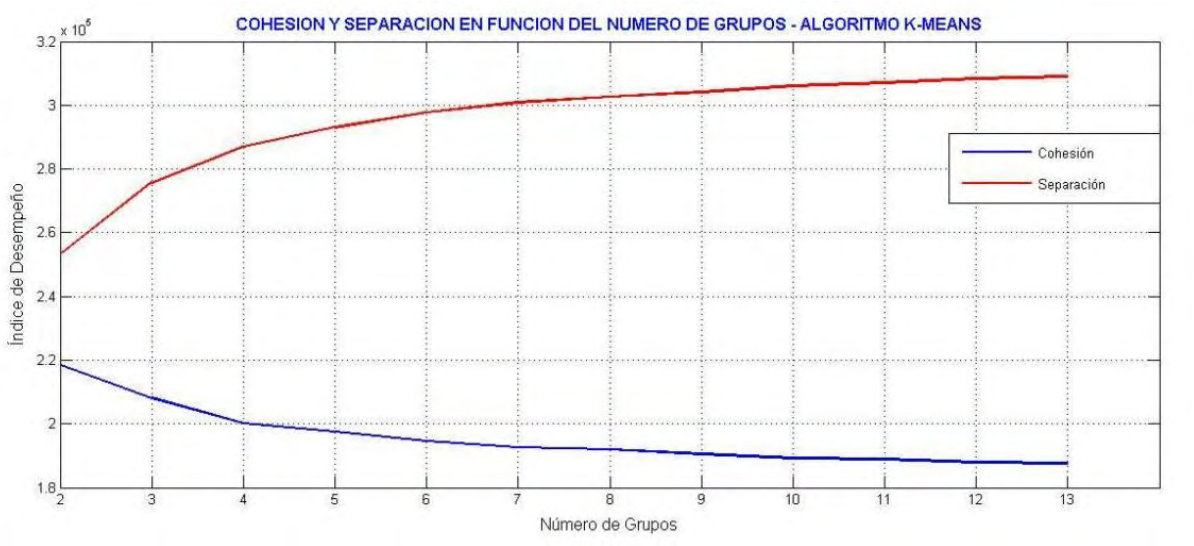



**Figura 22.** Gráfica de los índices de desempeño frente al número de grupos empleando el algoritmo k-means con distancia euclidiana y 1000 iteraciones.

Las pruebas más importantes de la aplicación de k-means sobre la base de datos original se presentan en la Tabla IX.

TABLA IX
PRUEBAS MÁS IMPORTANTES DE K-MEANS EN LOS DATOS ORIGINALES DE TUMACO

| Prueba No. | Procesamiento | | Número de grupos | Número de iteraciones | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|---|
| | Normalizar | Centrar | | | | | |
| 1 | Si | No | 5 | 100 | 32.236,60 | 128.693,70 | 81,90398 |
| 2 | Si | No | 5 | 500 | 32.236,60 | 128.693,70 | 81,90398 |
| 3 | Si | No | 5 | 1000 | 196.713,40 | 130.631,10 | 55,34982 |
| 4 | No | No | 5 | 100 | 21.047,62 | 90.288,92 | 84,55874 |
| 5 | No | No | 5 | 500 | 21.047,62 | 90.288,92 | 84,55874 |
| 6 | No | No | 5 | 1000 | 21.047,62 | 90.288,92 | 84,55874 |
| 7 | No | No | 5 | 2000 | 21.047,62 | 90.288,92 | 84,55874 |
| 8 | Si | No | 5 | 2000 | 32.236,60 | 128.693,70 | 81,90398 |
| 9 | No | Si | 5 | 500 | 21.047,62 | 90.288,92 | 84,55874 |
| 10 | Si | Si | 5 | 500 | 32.236,60 | 128.693,70 | 81,90398 |
| 11 | Si | Si | 5 | 500 | 32.236,60 | 128.693,70 | 81,90398 |

Los resultados de todas las pruebas son parecidos, se evidencia la demarcación del río Mira en Tumaco, y los clusters restantes se dividen de una manera acertada, como se muestra en la Figura 23. Se puede observar que las pruebas de la 4 a la 7 y la número 9 presentan el mayor porcentaje de rendimiento, sin embargo, se elige la prueba 5 puesto que la 6 y la 7 se descartan por implicar mayor tiempo de ejecución debido al número de iteraciones. De igual manera, la prueba 9 requiere que los datos se centren, y aun así el resultado es similar al obtenido en la prueba 5. La prueba 4 no es conveniente puesto que el número de iteraciones es bajo (100 iteraciones) y el algoritmo no converge.

Las pruebas 2, 5, 9, 10 y 11 (Figura 23), muestran el efecto del procesamiento en los datos de entrada. Se observa que el porcentaje de rendimiento es igual para las pruebas 2, 10 y 11 y que es menor al de las pruebas 5 y 9, que poseen el mismo valor. Esto quiere decir que para el algoritmo k-means resulta mejor no aplicar un procesamiento completo o únicamente centrar los datos. Las pruebas 10 y 11 se hacen con el fin de evaluar el efecto del orden en que se efectúa el procesamiento. En la prueba número 10, primero se normalizan los datos y luego se centran, mientras que en la prueba 11 es al contrario. Los resultados indican que el desempeño es igual, independientemente del proceso que se realice primero, si normalizar o centrar los datos, debido a que son operadores lineales.

Finalmente, se determina que la prueba 5 es el mejor resultado del algoritmo k-means, tanto por los índices de desempeño como por el indicador de comparación visual con los mapas de referencia, donde los ríos se demarcan apropiadamente.

Los resultados de la prueba de k-means con PCA se presentan en la Tabla X y la Figura 24. Gracias a la visual por medio de QGIS, se encontró que en este caso

influye el número de iteraciones, por lo que se comparan los resultados obtenidos con 500 y 1000 iteraciones.

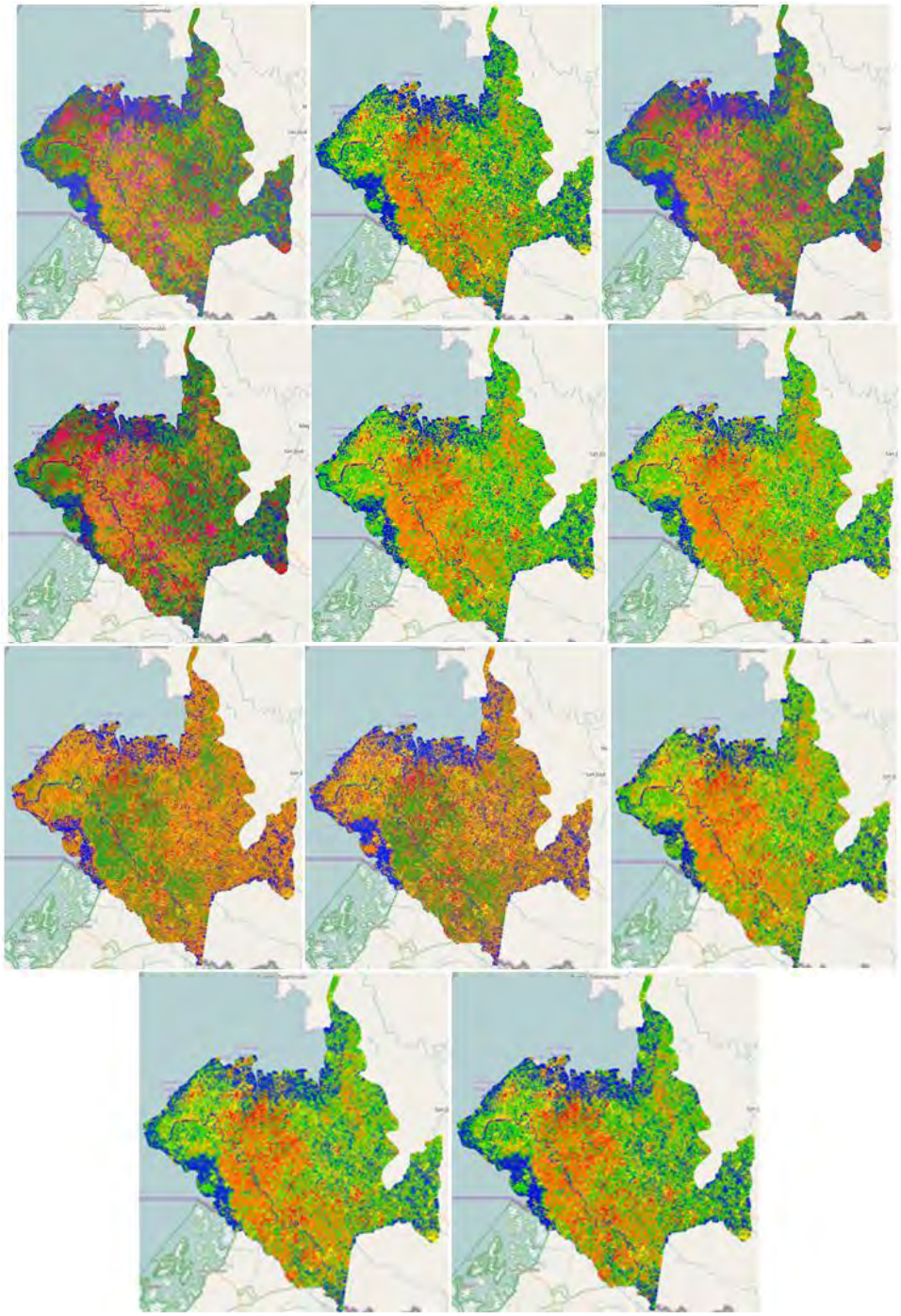


**Figura 23.** Resultado gráfico de la aplicación de k-means, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas expuestas en la Tabla IX.

TABLA X

PRUEBAS DE K-MEANS EN LOS DATOS RESULTADO DE PCA

| Prueba No. | Número de componentes principales | Procesamiento | | Número de iteraciones | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|---|
| | | Normalizar | Centrar | | | | |
| 1 | 7 | No | No | 1000 | 32.236,60 | 128.693,70 | 100,00000 |
| 2 | 7 | No | No | 500 | 32.236,60 | 128.693,70 | 100,00000 |
| 3 | 7 | Si | No | 1000 | 32.379,68 | 37.029,59 | 64,16577 |
| 4 | 7 | Si | No | 500 | 32.306,73 | 37.102,55 | 64,30652 |
| 5 | 7 | Si | Si | 1000 | 32.306,73 | 37.102,55 | 64,30652 |
| 6 | 7 | Si | Si | 500 | 32.343,35 | 37.065,92 | 64,23580 |
| 7 | 7 | No | Si | 1000 | 32.236,60 | 128.693,70 | 100,00000 |
| 8 | 7 | No | Si | 500 | 32.236,60 | 128.693,70 | 100,00000 |
| 9 | 3 | Si | Si | 1000 | 18.728,64 | 36.940,04 | 99,90067 |
| 10 | 4 | Si | Si | 1000 | 22.348,09 | 37.013,57 | 91,90210 |
| 11 | 5 | Si | Si | 1000 | 26.256,90 | 36.965,41 | 85,59917 |

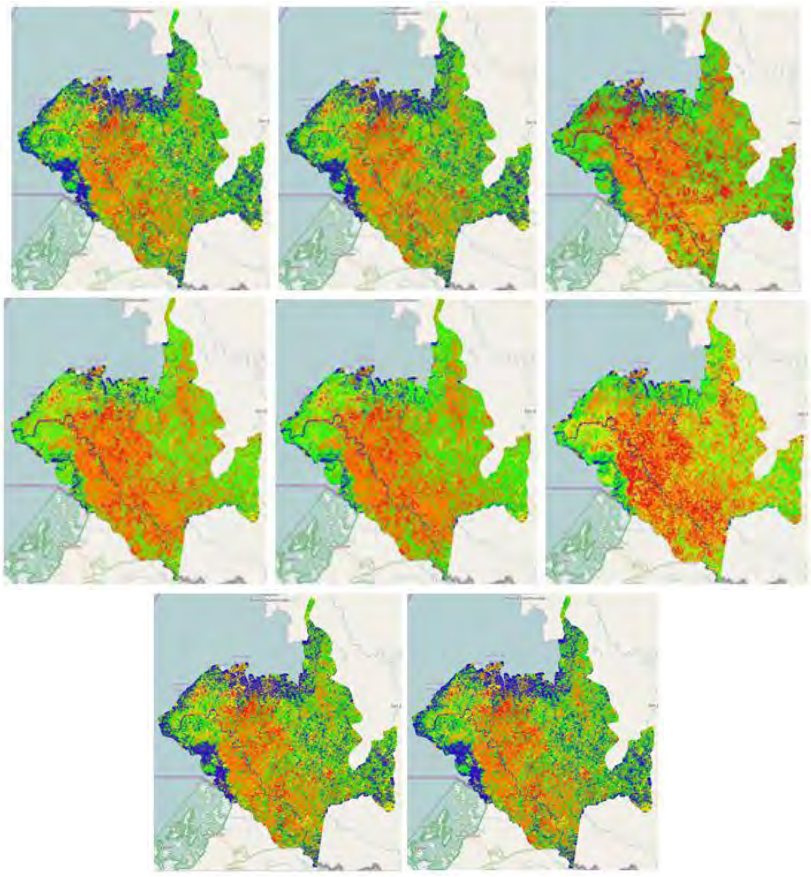


**Figura 24.** Resultado gráfico de la aplicación de k-means sobre el resultado de PCA con 7 componentes principales, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas 1 a 8 de la Tabla X.

Las pruebas 1, 2, 7 y 8 presentan un porcentaje de rendimiento de 100%, pero el resultado final no es el mejor teniendo en cuenta que la agrupación de los ríos no está bien definida. Se considera que esto se debe a que los datos no están normalizados ya que se determinó que en PCA influye la etapa de procesamiento. Por lo tanto, se decide elegir entre las cuatro pruebas restantes, donde las pruebas 4 y 5 presentan el mismo porcentaje de rendimiento y una visual muy similar. Se determina que el resultado más adecuado se da en la prueba 5 debido a que los datos se normalizan y se centran. En seguida, con la prueba 5 se varía el número de componentes principales entre 3 y 5, cuyos resultados son las pruebas 9, 10 y 11 de la Tabla X. Estos resultados se comparan entre sí y se deduce que se deben tomar 3 componentes principales para obtener tanto un buen resultado como una reducción en la dimensión de los datos. Además se observa que el resultado en la visualización es muy similar e inclusive mejor al más adecuado de k-means aplicado a la base de datos original y al de 7 componentes principales, ver Figura 25.

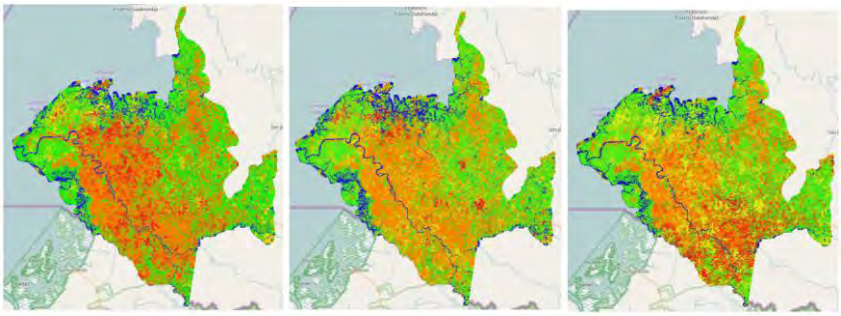


**Figura 25.** Resultado gráfico de la aplicación de k-means sobre el resultado de PCA, tomando 3, 4 y 5 componentes principales, representado en QGIS, en orden, de izquierda a derecha, las pruebas 9 a 11 de la Tabla X.

Se plantea además el desarrollo de una prueba que muestre el efecto causado por la inicialización de centroides facilitada por la función de Matlab (métodos *'cluster'*, *'uniform'* y *'numeric matrix'*) con el método por defecto ('sample') de la función *kmeans*. En este caso, para el método *'numeric matrix'* es necesario definir una matriz que incluya las características de los centroides iniciales. Para ello se toma como referencia la base de datos del mapa de clasificación de Corponariño con 13 clases, lo que facilita la reducción a cinco clases con las que se efectúan las pruebas. Las 13 clases expuestas en la Figura 3 se reducen a 5 tipos como se muestra en Tabla XI. En PostgreSQL se ejecutan las consultas para conformar la matriz de dimensiones $5 \times 7$, ya que 7 es el número de características que se usan y cuyo valor es el promedio por banda espectral de cada una de las 5 clases definidas en la columna "Cobertura reducción" de la Tabla XI. Las pruebas realizadas para observar el efecto de la inicialización de centroides se describen en la Tabla XII.

TABLA XI

REDUCCIÓN DE CLASES Corpocorin PARA INICIALIZAR CENTROIDES

| Corpocorin Clase No. | Corpocorin | Reducción Clase No. | Cobertura reducción |
|---|---|---|---|
| 1 | Aguas continentales | 5 | Agua |
| 2 | Aguas marítimas | 5 | Agua |
| 3 | Áreas abiertas, sin o con poca vegetación | 4 | Áreas abiertas y/o suelo desnudo |
| 4 | Áreas agrícolas heterogéneas (misceláneos) | 3 | Áreas agrícolas y cultivos |
| 5 | Áreas con vegetación herbácea y/o arbustiva | 2 | Vegetación herbácea y/o arbustiva |
| 6 | Áreas húmedas continentales | 5 | Agua |
| 7 | Bosques | 2 | Vegetación herbácea y/o arbustiva |
| 8 | Cultivos anuales o transitorios | 3 | Áreas agrícolas y cultivos |
| 9 | Cultivos permanentes | 3 | Áreas agrícolas y cultivos |
| 10 | Pastos | 2 | Vegetación herbácea y/o arbustiva |
| 11 | Vegetación de páramo | 2 | Vegetación herbácea y/o arbustiva |
| 12 | Zonas de extracción minera y escombros | 4 | Áreas abiertas y/o suelo desnudo |
| 13 | Zonas urbanizadas | 1 | Zonas urbanizadas |

TABLA XII

PRUEBAS DE K-MEANS EN LOS DATOS ORIGINALES DE TUMACO – MÉTODOS DE INICIALIZACIÓN DE CENTROIDES

| Prueba No. | Procesamiento Normalizar | Método de inicialización de centroides | Cohesión | Separación | % de Rendimiento | % de Rendimiento 1 |
|---|---|---|---|---|---|---|
| 1 | No | 'sample' | 21047,62 | 90288,92 | | 85.07900 |
| 2 | No | 'cluster' | 21047,62 | 90288,92 | 85,07900 | 85,07900 |
| 3 | Si | 'cluster' | 32236,60 | 128693,70 | 82,64553 | 82,64553 |
| 4 | No | 'uniform' | 21781,59 | 89554,96 | 83,10900 | 83,10900 |
| 5 | Si | 'uniform' | 33686,46 | 127243,80 | 80,67716 | 80,67716 |
| 6 | No | 'numeric matrix' | 21047,62 | 90288,92 | 85,07900 | 85,07900 |
| 7 | Si | 'numeric matrix' | 32236,60 | 128693,70 | 82,64553 | 82,64553 |

En la tabla anterior, la prueba número 1 hace referencia al mejor resultado de k-means con la base de datos original (prueba 5 de la Tabla IX). La columna "% de Rendimiento" compara las pruebas 2 a 7 mientras que la columna "% de Rendimiento 1" contrasta todas las pruebas de la tabla. Se observa que las pruebas 2 y 6 presentan el más alto porcentaje de rendimiento, correspondiente a una visualización adecuada como se muestra en la Figura 26. Además, el porcentaje de rendimiento es igual en comparación al de la prueba 1. Por otro lado el método *'numeric matrix'* da un indicio de que las clases que se obtienen son apropiadas puesto que la inicialización de centroides utilizada se obtiene del mapa de *cobertura* de Corponariño. Adicionalmente, se considera que aunque el método *'uniform'* presenta un buen resultado, no es mejor que los otros debido no solo al

porcentaje de rendimiento, sino que también la visualización presenta una perturbación de las diferentes clases, especialmente del grupo *"agua".*

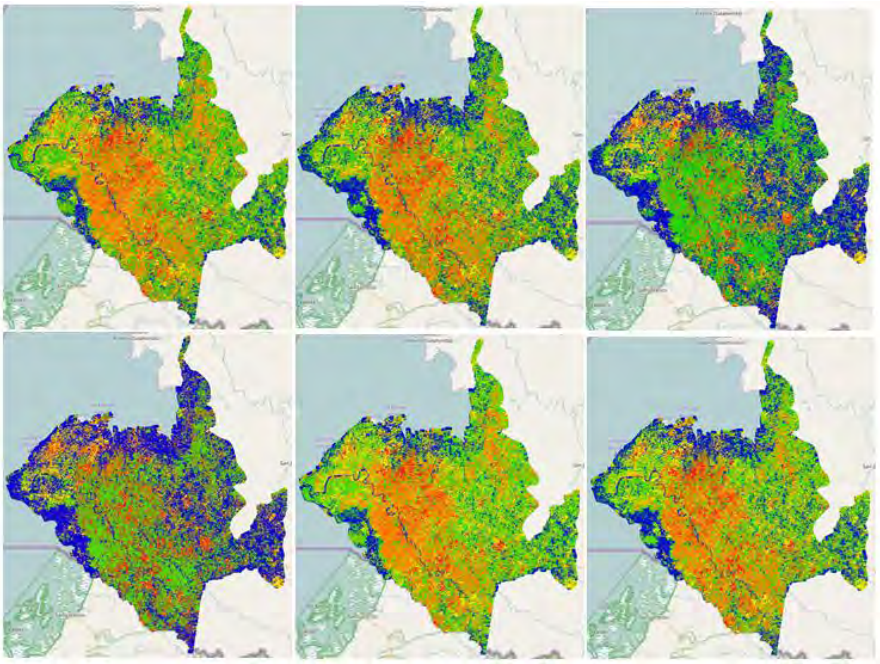


**Figura 26.** Resultados gráficos de la aplicación de k-means variando el método de inicialización de centroides, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 2 a la 7 expuestas en la Tabla XII.

De esta manera, las pruebas en adelante para k-means se continúan haciendo sin necesidad de un método definido de inicialización de centroides, sino utilizando la técnica por defecto "*sample*". Además, partiendo de los resultados obtenidos con este tipo de clustering, se considera viable su extrapolación al departamento de Nariño.

#### 2.3.3.2. Fuzzy C-Means

Para ejecutar este algoritmo se usa la función de Matlab *fcm* que proporciona una matriz *U*, que indica el grado de pertenencia de cada pixel a cada cluster. El grupo final se identifica con el mayor grado de pertenencia.

Las pruebas más importantes de la aplicación de este algoritmo se resumen en la Tabla XIII, donde el número de clusters definido a priori es igual a 5. En las Figuras 27 y 28, se observan los resultados obtenidos para las pruebas 1 a 14. La prueba número 15 no se grafica porque el resultado visual es el mismo que se muestra para la prueba 14.

TABLA XIII

PRUEBAS MÁS IMPORTANTES DE FCM EN LOS DATOS ORIGINALES DE TUMACO

| Prueba No. | Procesamiento | | Número de iteraciones | Exponente de la matriz U | Criterio | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|---|---|
| | Normalizar | Centrar | | | | | | |
| 1 | Si | No | 100 | 2 | 0,001 | 32.966,01 | 121.303,20 | 55,79472 |
| 2 | No | No | 100 | 2 | 0,001 | 21.609,12 | 85.389,69 | 65,50503 |
| 3 | Si | No | 200 | 2 | 0,001 | 32.965,89 | 121.284,00 | 55,79105 |
| 4 | No | No | 200 | 2 | 0,001 | 21.609,21 | 85.388,90 | 65,50467 |
| 5 | No | No | 100 | 1,1 | 0,001 | 21.047,81 | 90.380,84 | 67,78601 |
| 6 | No | No | 100 | 1,1 | 100 | 32.243,51 | 127.929,80 | 57,81409 |
| 7 | No | No | 100 | 10 | 0,001 | 108.293,60 | 42,71733 | 9,72634 |
| 8 | No | No | 100 | 2 | 0,00001 | 21.609,33 | 85.399,79 | 65,50654 |
| 9 | Si | No | 100 | 2 | 0,00001 | 32.965,93 | 121.299,00 | 55,79397 |
| 10 | No | No | 500 | 1,7 | 0,0001 | 208.367,30 | 252.341,80 | 54,70891 |
| 11 | Si | No | 100 | 1,1 | 0,001 | 32.237,09 | 128.831,10 | 57,99796 |
| 12 | No | Si | 100 | 1,1 | 0,001 | 21.047,81 | 90.380,50 | 67,78595 |
| 13 | Si | Si | 100 | 1,1 | 0,001 | 33.687,14 | 127.271,40 | 56,28582 |
| 14 | No | No | 500 | 1,01 | 0,0001 | x | x | x |
| 15 | Si | No | 100 | 1,01 | 0,00001 | x | x | x |

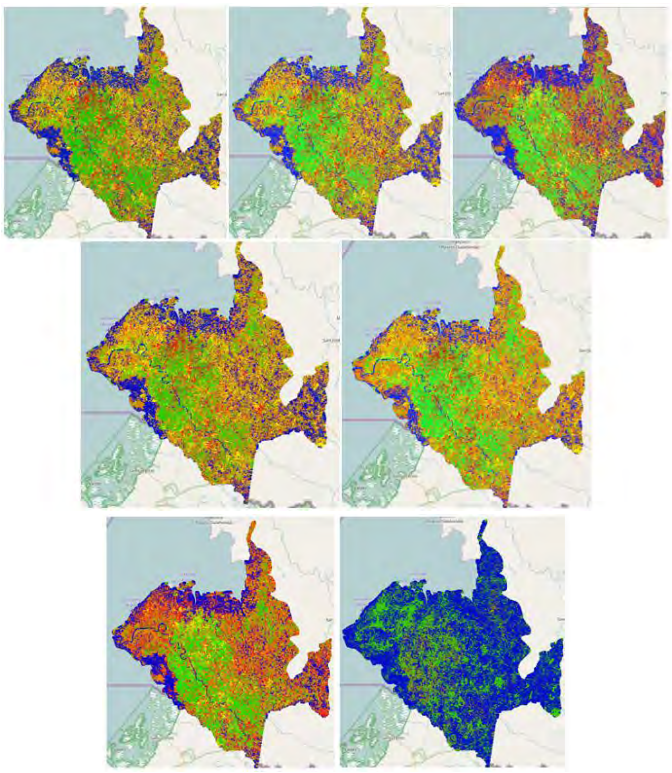


**Figura 27.** Resultados gráficos de la aplicación de fuzzy c-means, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 1 a la 7 expuestas en la Tabla XIII.

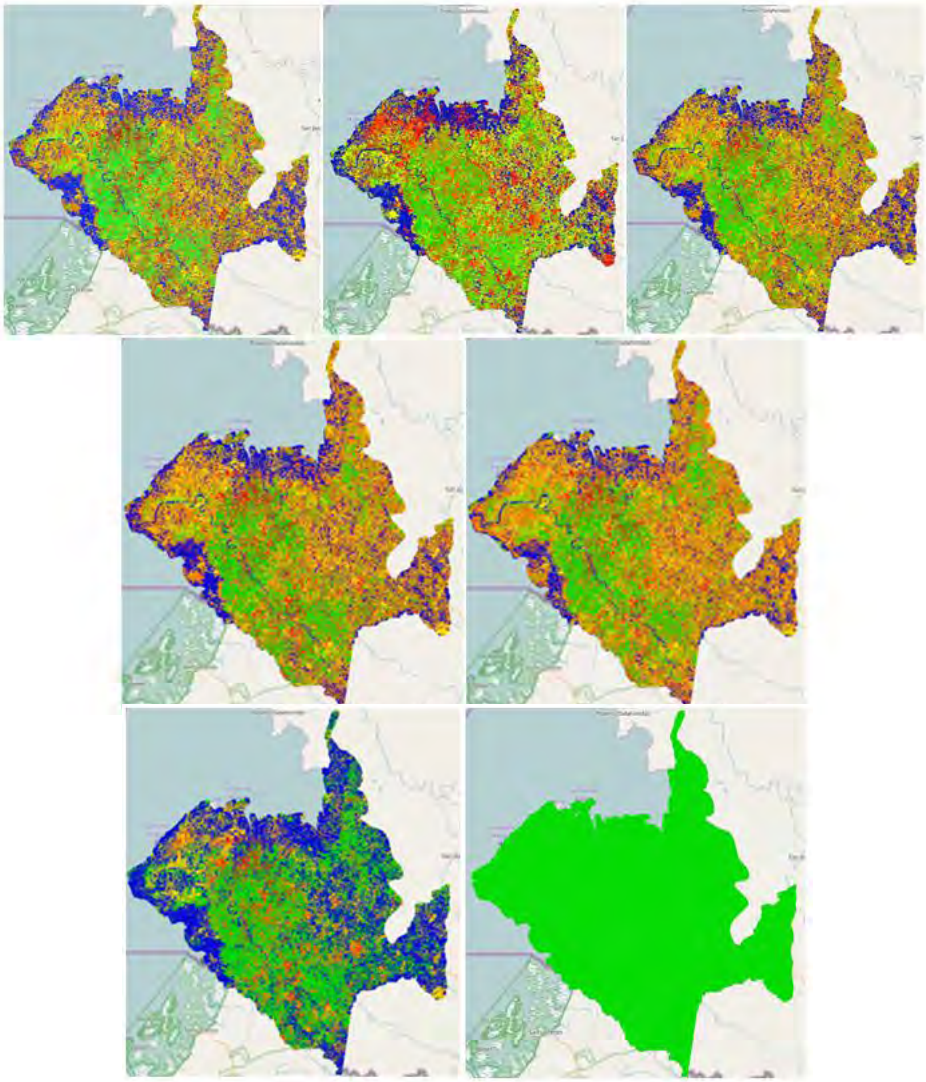


**Figura 28.** Resultados gráficos de la aplicación de fuzzy c-means, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 8 a la 14 expuestas en la Tabla XIII.

El valor de separación de la prueba 10 es el valor máximo obtenido en el transcurso de las pruebas y el valor de cohesión mínimo es el correspondiente a la prueba número 5, que es la seleccionada como el mejor resultado para el algoritmo fuzzy c-means.

Entre las pruebas realizadas se encontró que cuando el exponente de la matriz $U$ tiende a 1, la mejor partición es cada vez más cercana a una partición exclusiva, es decir se tiende a formar un único grupo o un grupo predominante. Como se muestra en las pruebas 14 y 15, todos los puntos hacen parte de un único grupo,

mientras que si el exponente tiende a infinito la mejor partición se aproxima a una matriz $U$ con todos sus valores iguales a $1/\mathrm{Nc}$, donde Nc es el número de grupos.

Por otro lado, el efecto del procesamiento en los datos de entrada se puede analizar a partir de las pruebas 5, 11, 12 y 13. Se demuestra tanto en la visualización como en los índices de desempeño, que la clasificación es mejor empleando los datos sin normalizar. Además, al igual que en k-means, se presenta un resultado apropiado cuando los datos solamente se centran (prueba 12, con un porcentaje de rendimiento muy cercano al de la prueba 5). Por su parte, la prueba número 7 es un ejemplo de lo importante que es el exponente de la matriz de pertenencias, puesto que un valor muy elevado del mismo ocasiona una partición sin coherencia como se observa en la Figura 27.

Se concluye que a excepción de las pruebas 7, 13, 14 y 15, los resultados de FCM son muy parecidos entre sí. Además, teniendo en cuenta el indicador de comparación visual con los mapas de referencia, la prueba número 5 presenta un resultado similar al de k-means que utiliza los mismos datos de entrada. Por tanto, los dos algoritmos presentan un buen desempeño, aunque k-means posee porcentajes más elevados que fuzzy c-means.

Para el caso de PCA, los resultados se muestran en la Tabla XIV y sus visualizaciones en las Figuras 29 y 30.

TABLA XIV
PRUEBAS MÁS IMPORTANTES DE FCM EN LOS DATOS RESULTADO DE PCA

| Prueba No. | Número de componentes principales | Procesamiento | | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|
| | | Normalizar | Centrar | | | |
| 1 | 7 | No | No | 32.237,07 | 128.827,30 | 100,00000 |
| 2 | 7 | Si | No | 32.311,72 | 36.953,80 | 64,22686 |
| 3 | 7 | No | Si | 32.309,74 | 36.976,80 | 64,23885 |
| 4 | 7 | Si | No | 32.381,49 | 36.999,54 | 64,13713 |
| 5 | 3 | No | Si | 18.698,04 | 36.966,06 | 99,85063 |
| 6 | 4 | No | Si | 22.496,92 | 36.910,23 | 91,33223 |
| 7 | 5 | No | Si | 26.077,04 | 37.076,82 | 85,85154 |

En la tabla anterior se observa que la prueba número 1 tiene un 100% en su desempeño, sin embargo en la visualización presenta una identificación errónea del cluster *"agua".* La prueba con el siguiente mejor porcentaje de rendimiento es la número 3, cuya visualización en QGIS es mejor que la prueba 1, por resaltar los ríos del municipio. En cuanto al número de componentes principales se decide tomar como mejor resultado la prueba 5 tanto por los índices de desempeño como

por la visualización. Se observa que el resultado es muy parecido al de k-means, por lo que se decide realizar su extrapolación a todo el departamento de Nariño.

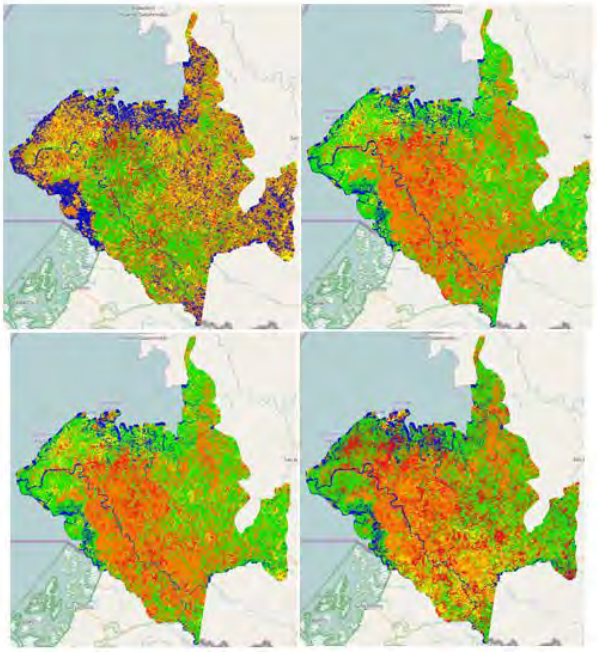


**Figura 29.** Resultado gráfico de la aplicación de fuzzy c-means sobre el resultado de PCA, tomando 7 componentes principales, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas 1 a 4 expuestas en la Tabla XIV.

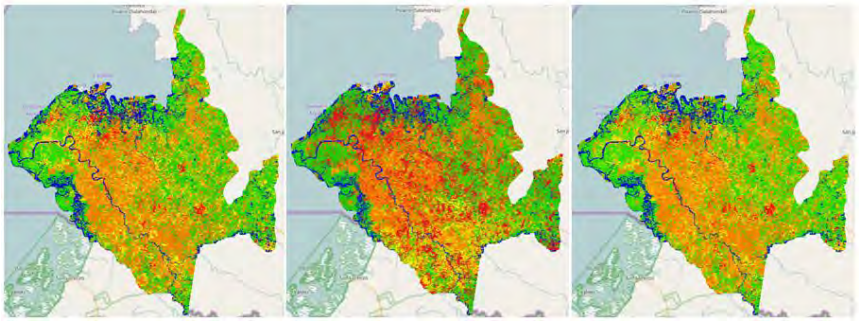


**Figura 30.** Resultado gráfico de la aplicación de fuzzy c-means sobre el resultado de PCA, tomando 3, 4 y 5 componentes principales respectivamente, representado en QGIS las pruebas 5 a 7 expuestas en la Tabla XIV.

Los algoritmos que resultaron de la revisión bibliográfica y del trabajo previo como EM, Dbscan y GK se ejecutan a partir de modificaciones realizadas a códigos abiertos expuestos en [30], con el propósito de adaptarlos a los requerimientos de datos de entrada, salida y finalidad del presente estudio. A continuación se presentan los resultados obtenidos.

### 2.3.3.3. EM: Expectation - Maximization

En la Tabla XV se evidencian las pruebas relevantes del algoritmo EM y en las Figuras 31 y 32, se aprecian sus visualizaciones.

TABLA XV
PRUEBAS DESTACADAS DE EM EN LOS DATOS ORIGINALES DE TUMACO

| Prueba No. | Procesamiento | | Máximo número de iteraciones | Grupos | Mínimo valor de perfeccionamiento Tol | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|---|---|
| | Normalizar | Centrar | | | | | | |
| 1 | No | No | 30 | 5 | 0,000001 | 38.453,56 | 68.097,220 | 100,0 |
| 2 | No | No | 10 | 13 | 0,000001 | 43.820,05 | 61.013,760 | 88,7 |
| 3 | No | No | 20 | 5 | 0,000001 | 58.897,54 | 40.607,220 | 62,5 |
| 4 | No | No | 10 | 5 | 0,0000001 | 45.367,17 | 54.215,550 | 82,2 |
| 5 | No | No | 5 | 5 | 0,0000001 | 108.147,3 | 161,442 | 17,9 |
| 6 | No | No | 3 | 5 | 0,000001 | 39.320,15 | 1.361,250 | 49,9 |
| 7 | No | No | 10 | 5 | 0,000001 | 41.975,05 | 8.128,650 | 51,8 |
| 8 | No | No | 40 | 5 | 0,000001 | 50.781,14 | 57.702,660 | 80,2 |
| 9 | No | Si | 40 | 5 | 0,000001 | 43.500,01 | 63.265,550 | 90,7 |
| 10 | Si | Si | 10 | 5 | 0,000001 | 56.986,55 | 4.567,560 | 37,1 |
| 11 | Si | No | 10 | 5 | 0,000001 | 49.259,66 | 2.355,090 | 40,8 |
| 12 | Si | Si | 5 | 5 | 0,000001 | 67.964,54 | 4.567,890 | 31,6 |

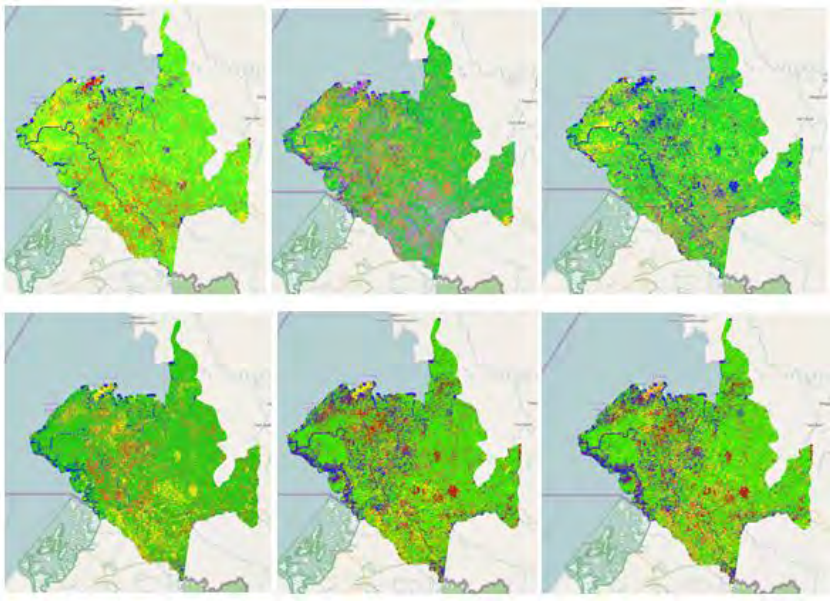


**Figura 31.** Resultado gráfico de la aplicación de EM sobre la base de datos original, representado en QGIS, en orden, de arriba abajo y de izquierda a derecha, las pruebas 1 a 6 expuestas en la Tabla XV.

Los resultados de este algoritmo producen 5 clusters cuyo desempeño mejora si el número de iteraciones está alrededor de 30. En diferentes pruebas se varía el mejor valor de perfeccionamiento, destacando los resultados en los que se fijó un valor de 0,000001. Este parámetro, propio de este algoritmo, indica la mínima mejora de la función objetivo entre dos iteraciones consecutivas.

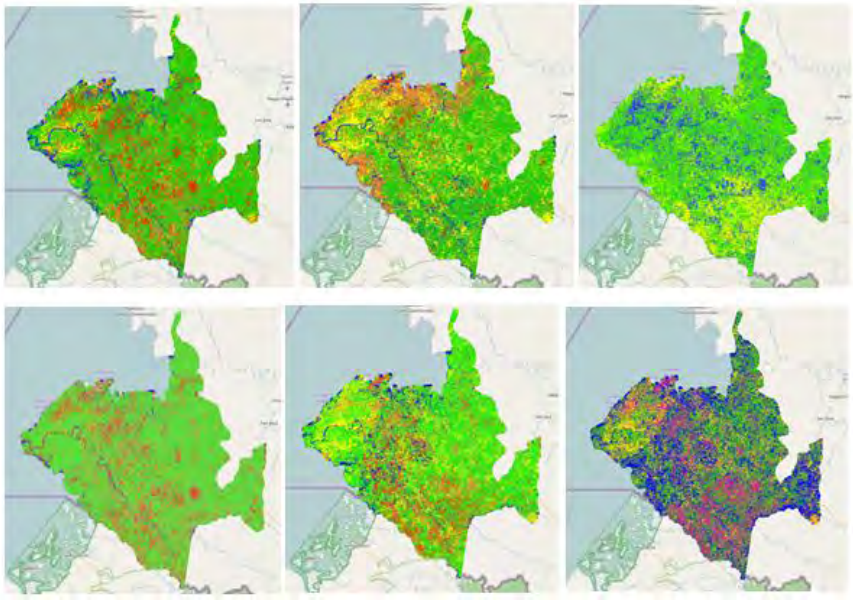


**Figura 32.** Resultado gráfico de la aplicación de EM sobre la base de datos original, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas 7 a 12 expuestas en la Tabla XV.

En cuanto al efecto del procesamiento de los datos, se evidencia que no es relevante, como se nota en las pruebas 9, 10, 11 y 12 de la Tabla XV. Es más, en la Figura 32 se aprecia que el centrado de los datos influye negativamente mientas que el normalizado no logra superar los resultados en desempeño y de agrupación de agua de la prueba número 1. Finalmente, se evidencia que efectuar un procesamiento completo, con un número de iteraciones pequeño (prueba 12), arroja un resultado inadecuado tanto en visual como en índices de desempeño.

La prueba 1, presenta los mejores valores de cohesión y separación obtenidos en el transcurso de las pruebas y el indicador de comparación con los mapas de referencia permite detallar de mejor forma los ríos de la zona. De esta manera, se escoge esta prueba como el resultado más eficiente del algoritmo EM y se decide que es viable para ser extrapolado a todo el departamento de Nariño.

Para el caso de PCA, se realizaron las pruebas respectivas aplicando el algoritmo EM con su mejor configuración, para un total de 5 grupos. Los resultados se muestran en la Tabla XVI y sus visuales se presentan en las Figuras 33 y 34.

Se determina que al usar las 7 componentes principales, la prueba 2 de la Tabla XVI posee la mejor visual e índices de desempeño. Por tanto, se realizan las pruebas tomando solo 3, 4 y 5 componentes principales, resultados evidenciados en la Figura 34.

Finalmente, se determina que la prueba número 6 de la Tabla XVI, posee la mejor visual y porcentaje de rendimiento. Además realiza una adecuada aproximación a la prueba 2 de la misma tabla, por tanto se elige para ser extrapolada a todo el departamento de Nariño.

TABLA XVI
PRUEBAS MÁS IMPORTANTES DE EM EN LOS DATOS COMO RESULTADO DE PCA

| Prueba No. | Número de componentes principales | Procesamiento | | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|
| | | Normalizar | Centrar | | | |
| 1 | 7 | No | No | 71.975,05 | 84.129,65 | 82,37010 |
| 2 | 7 | No | Si | 63.918,68 | 88.196,35 | 89,04613 |
| 3 | 7 | Si | No | 49.915,54 | 15.830,87 | 58,97479 |
| 4 | 7 | Si | Si | 52.934,74 | 14.585,04 | 55,41669 |
| 5 | 3 | No | Si | 67.012,77 | 80.578,30 | 84,73286 |
| 6 | 4 | No | Si | 56.250,02 | 94.214,47 | 100,00000 |
| 7 | 5 | No | Si | 72.524,72 | 81.617,46 | 82,09461 |

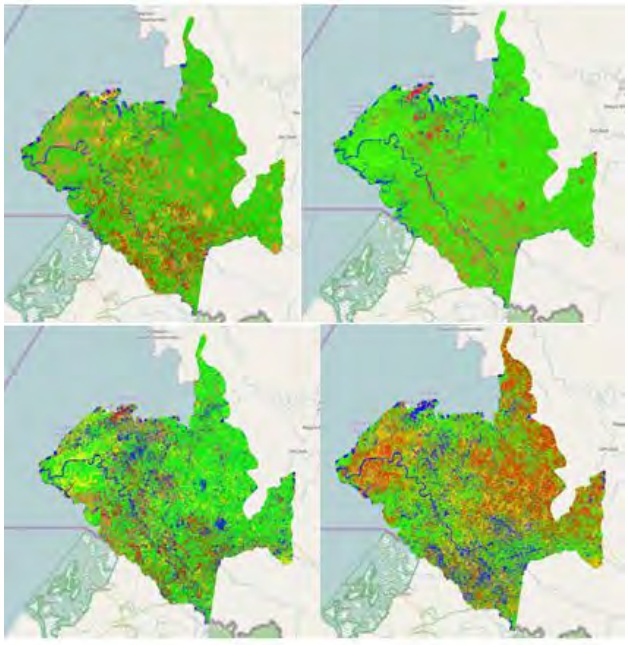


**Figura 33.** Resultado gráfico de la aplicación de EM sobre el resultado de PCA, tomando 7 componentes principales, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas 1 a 4 expuestas en la Tabla XVI.

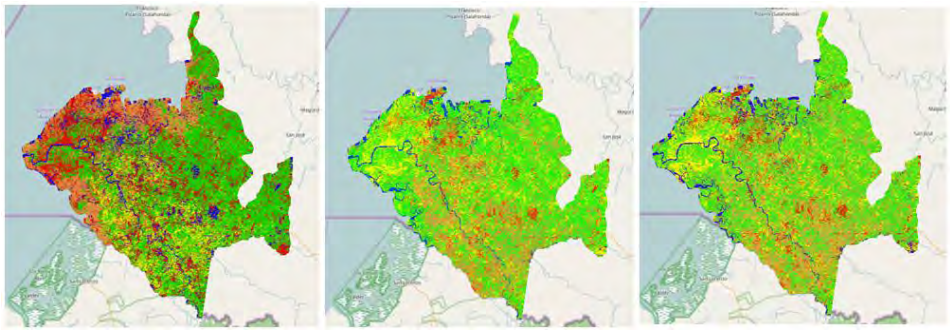


**Figura 34.** Resultado gráfico de la aplicación de EM sobre el resultado de PCA, tomando 3, 4 y 5 componentes principales respectivamente (pruebas 5 a 7 de la Tabla XVI).

#### 2.3.3.4. Isodata

Las pruebas que se realizaron para el algoritmo Isodata se presentan en la Tabla XVII. La columna "A" indica la cantidad de grupos que se obtienen en la clasificación final, diferente de "K" que define el número máximo de grupos que puede arrojar el algoritmo.

TABLA XVII
PRUEBAS DE ISODATA EN LOS DATOS ORIGINALES DE TUMACO

| Prueba No. | Procesamiento | | ON | OC | OS | K | L | I | A | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Normalizar | Centrar | | | | | | | | | | |
| 1 | Si | No | 10 | 0,05 | 0,07 | 5 | 2 | 1 | 2 | 241.893,10 | 220.438,20 | 44,00990 |
| 2 | Si | No | 10 | 0,05 | 0,07 | 5 | 2 | 20 | 3 | 208.116,30 | 275.752,10 | 54,69515 |
| 3 | Si | No | 10 | 0,05 | 0,07 | 5 | 2 | 50 | 3 | 208.107,10 | 275.752,20 | 54,69538 |
| 4 | Si | No | 10 | 0,05 | 0,07 | 5 | 5 | 20 | 4 | 38.337,57 | 122.592,70 | 47,71660 |
| 5 | Si | No | 10 | 0,05 | 0,07 | 5 | 20 | 20 | 4 | 38.337,57 | 122.592,70 | 47,71660 |
| 6 | Si | No | 10 | 0,05 | 0,07 | 10 | 2 | 20 | 4 | 38.337,57 | 122.592,70 | 47,71660 |
| 7 | Si | No | 10 | 0,05 | 0,07 | 20 | 2 | 20 | 4 | 38.337,57 | 122.592,70 | 47,71660 |
| 8 | Si | No | 10 | 0,00001 | 0,07 | 5 | 2 | 20 | 3 | 49.558,98 | 111.371,30 | 39,91082 |
| 9 | Si | No | 10 | 0,0005 | 0,07 | 5 | 2 | 20 | 3 | 49.558,98 | 111.371,30 | 39,91082 |
| 10 | Si | No | 10 | 2 | 0,07 | 5 | 2 | 20 | 3 | 65.028,49 | 95.901,77 | 32,41547 |
| 11 | Si | No | 10 | 0,05 | 0,0000001 | 5 | 2 | 20 | 3 | 49.558,98 | 111.371,30 | 39,91082 |
| 12 | Si | No | 10 | 0,05 | 0,0001 | 5 | 2 | 20 | 3 | 49.558,98 | 111.371,30 | 39,91082 |
| 13 | Si | No | 10 | 0,05 | 0,09 | 5 | 2 | 20 | 3 | 72.193,14 | 88.737,12 | 29,62511 |
| 14 | Si | No | 10 | 0,05 | 2 | 5 | 2 | 20 | 3 | 160.930,30 | 0,00 | 6,07183 |
| 15 | Si | No | 10 | 0,01 | 0,01 | 20 | 5 | 20 | 11 | 30.404,73 | 130.525,50 | 55,80497 |
| 16 | Si | No | 10 | 0,01 | 0,05 | 20 | 2 | 20 | 11 | 30.404,73 | 130.525,50 | 55,80497 |
| 17 | Si | No | 100 | 0,0001 | 0,005 | 13 | 1 | 30 | 7 | 30.405,18 | 130.525,10 | 55,80442 |
| 18 | No | No | 100 | 0,05 | 0,07 | 20 | 2 | 20 | 4 | 38.337,57 | 122.592,70 | 47,71660 |

| CONTINUACIÓN TABLA XVII | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | No | No | 100.000 | 0,0005 | 0,005 | 9 | 2 | 20 | 5 | 32.242,48 | 128.687,80 | 53,63997 |
| 20 | No | No | 100.000 | 0,0005 | 0,005 | 13 | 2 | 20 | 7 | 27.151,46 | 133.778,80 | 60,24559 |
| 21 | No | No | 100 | 0,0001 | 0,005 | 13 | 1 | 30 | 7 | 19.542,82 | 91.793,73 | 66,64424 |
| 22 | No | No | 10.000 | 0,0005 | 0,05 | 13 | 2 | 30 | 7 | 25.604,09 | 135.326,20 | 62,70112 |
| 23 | Si | No | 100.000 | 0,0005 | 0,005 | 9 | 2 | 20 | 5 | 32.242,48 | 128.687,80 | 53,63997 |
| 24 | No | Si | 100.000 | 0,0005 | 0,005 | 9 | 2 | 20 | 5 | 21.053,75 | 90.282,80 | 62,78201 |
| 25 | Si | Si | 100.000 | 0,0005 | 0,005 | 9 | 2 | 20 | 5 | 32.307,77 | 128.622,50 | 53,56688 |
| 26 | Si | No | 100.000 | 0,0005 | 0,005 | 13 | 2 | 20 | 7 | 27.151,46 | 133.778,80 | 60,24559 |
| 27 | No | Si | 100.000 | 0,0005 | 0,005 | 13 | 2 | 20 | 7 | 20.172,27 | 95.164,28 | 65,69521 |
| 28 | Si | Si | 100.000 | 0,0005 | 0,005 | 13 | 2 | 20 | 7 | 26.308,20 | 134.622,10 | 61,55205 |

La prueba número 1 se realizó con los parámetros por defecto definidos para el algoritmo, corriendo únicamente una iteración, con el fin de observar el comportamiento de Isodata. Como se aprecia en la Figura 35, el resultado no es adecuado pero se pueden mejorar sus parámetros para obtener una clasificación apropiada. En las pruebas 2 y 3 se evidencia el impacto del número de iteraciones. La diferencia entre 1 y 20 iteraciones es de alrededor del 10%, mientras que la diferencia entre 20 y 50 iteraciones es del 0,0002%. Además, como se aprecia en la Figura 35, la visualización no presenta cambios bruscos y por lo tanto se decide continuar trabajando con 20 iteraciones. Las pruebas 4 y 5 muestran que el valor para el máximo número de agrupamientos que pueden mezclarse en una sola iteración no debe ser mayor que 2, puesto que el porcentaje de rendimiento se reduce. En cuanto a la visualización, el efecto no es considerable. El cambio en el valor de $L$, adiciona un agrupamiento a la clasificación final, sin embargo, se aprecia que al parecer uno de los grupos de la prueba 2 se subdivide para formar el cuarto cluster.

En las pruebas 6 y 7 se incrementa el número de clusters, pero el resultado final se compone únicamente de 4 grupos, indicio de que en Tumaco no es oportuno encontrar un mayor número de clases. Además teniendo en cuenta que se trabaja con una base de 5 grupos, se decide utilizar este número en la mayoría de las pruebas. Las pruebas 8, 9 y 10 muestran el efecto del parámetro $OC$, donde las dos primeras presentan un resultado similar a la prueba 2. Un valor muy pequeño para $OC$ reduce el porcentaje de rendimiento pero un valor muy elevado como $OC = 2$ perjudica totalmente la clasificación (Figura 36) puesto que la distancia entre centros debe ser adecuada para considerar pixeles dentro de un mismo cluster. Las pruebas entre la 11 y la 14 muestran el efecto del parámetro $OS$, que es determinante en la clasificación. Entre más se incrementa $OS$, el porcentaje de rendimiento disminuye y el efecto se evidencia en el mapa de clasificación. Un $OS = 2$ aunque produce 3 clases, posee un solo grupo predominante (Figura 37). El resultado de las pruebas indica que el valor de $OS$ debe ser también pequeño

debido a que si la desviación estándar entre los individuos de un grupo supera este umbral, probablemente los pixeles hacen parte de clusters diferentes. La conclusión anterior se evidencia en la visual de las Figuras 36 y 37.

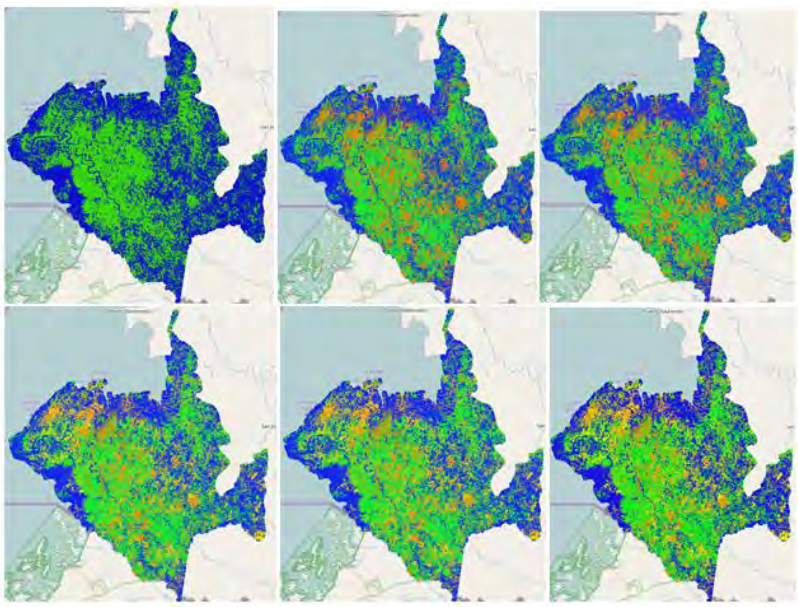


**Figura 35.** Resultados gráficos de la aplicación de isodata, representados en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 1 a la 6 expuestas en la Tabla XVII.

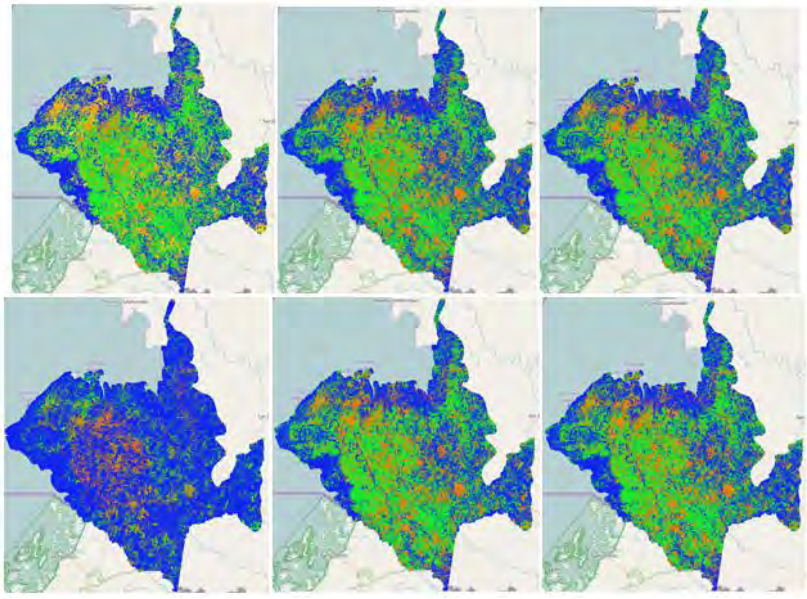


**Figura 36.** Resultados gráficos de la aplicación de isodata, representados en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 7 a la 12 expuestas en la Tabla XVII.

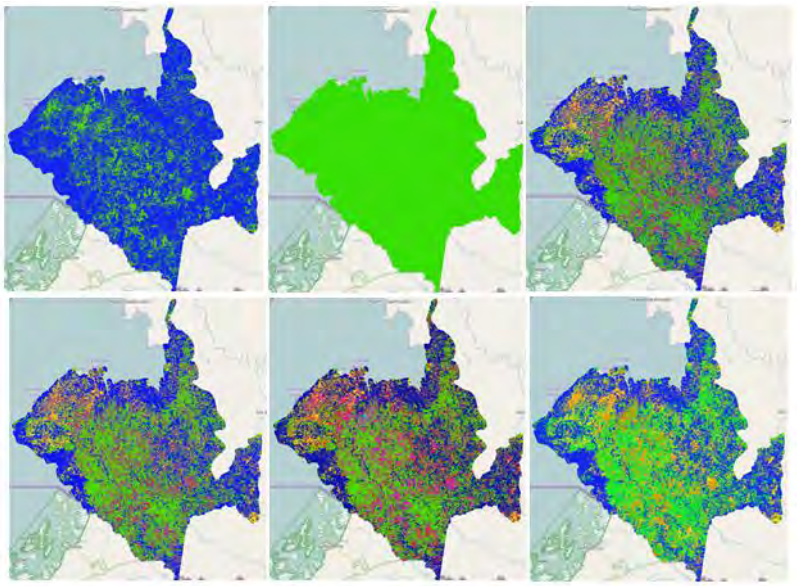


**Figura 37.** Resultados gráficos de la aplicación de isodata, representados en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 13 a la 18 expuestas en la Tabla XVII.

Teniendo en cuenta los resultados anteriores, a partir de la prueba 15 hasta la prueba número 22, se obtienen una serie de efectos como producto de diferentes configuraciones de los parámetros de Isodata (Figuras 37 y 38). Se observa el efecto de la variable $ON$ y el número máximo de agrupamientos, con las asignaciones adecuadas para los demás parámetros que producen altos porcentajes de rendimiento entre las pruebas realizadas, y que presentan una visual acorde a los resultados que se han venido obteniendo. Se considera que $ON$ está relacionado con el número total de elementos del conjunto de datos de entrada, debido a que si el valor de $ON$ es muy pequeño en comparación a la cantidad de puntos de entrada, la variable no tendrá un efecto relevante en la ejecución del algoritmo. Por el contrario, si el valor de $ON$ es muy grande en comparación al número de pixeles del conjunto de datos de entrada, se comete un error al eliminar un grupo compuesto por gran parte del total de datos.

En conclusión, a partir de las pruebas 1 a la 22 de la Tabla XVII, y de los resultados detallados anteriormente, se determina que la configuración más adecuada de los parámetros de Isodata se presenta en las pruebas 19 y 20. En este caso, la prueba número 21 presenta el mayor porcentaje de rendimiento, pero el indicador de comparación con el mapa de referencia no es el más adecuado (Figura 38). De igual forma, la siguiente prueba con un porcentaje de funcionamiento del 62,7%, es la número 22, cuyo resultado visual es superado por la prueba 20 (Figura 38) que arroja un total de 7 grupos. Esta información es relevante teniendo en cuenta que en Isodata no se define el número de clases a

priori, lo que puede dar un indicio del número de clases óptimo presentes en la zona. Sin embargo, como en los algoritmos anteriores se escogieron 5 clases, se establece la prueba 19 como otra opción de configuración del algoritmo. Hay pruebas que según la Tabla XVII, presentan un mayor porcentaje de rendimiento, pero el indicador de comparación visual con Street Map y el mapa de Corponariño, muestra que la clasificación obtenida en la prueba 19 es mejor, arrojando un resultado similar al alcanzado con los anteriores algoritmos.

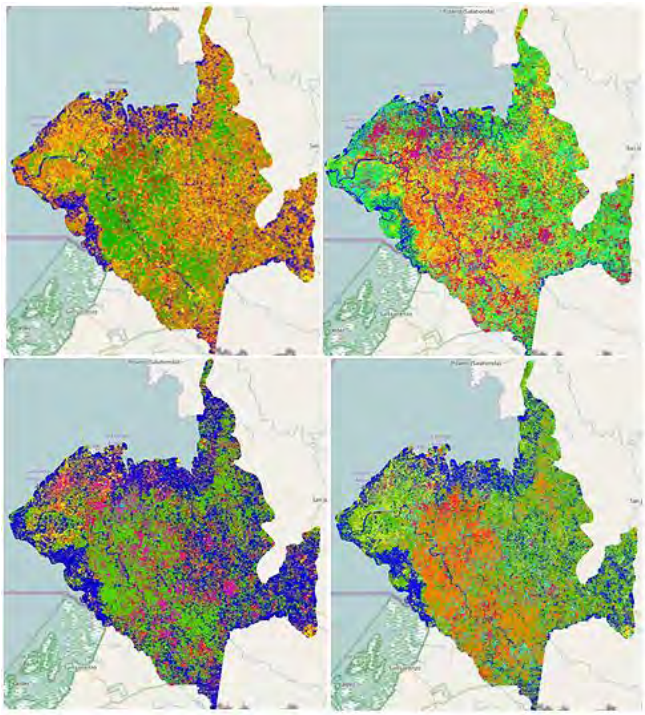


**Figura 38.** Resultados gráficos de la aplicación de isodata, representados en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 19 a la 22 expuestas en la Tabla XVII.

Las pruebas de la 23 en adelante, presentan el efecto del procesamiento de los datos. Para ambos casos, con 5 y 7 clusters se tiene el mayor porcentaje de rendimiento cuando los datos no se normalizan pero si se centran. Sin embargo, la comparación con los mapas de referencia y con los mejores resultados obtenidos, permite apreciar que no se presentan mayores cambios, aunque se resalta que el cluster *"agua"* se define mejor en las pruebas 19 y 20 (Figuras 38 y 39).

Gracias a los anteriores resultados, se logra adaptar el algoritmo Isodata al estudio. Luego, la aplicación mediante el uso de PCA, es más fácil. Los resultados obtenidos para esta fase se presentan en la Tabla XVIII. La columna "Número de agrupamientos" indica la cantidad de grupos que se obtienen en el resultado y la columna "Prueba" obedece a una de las dos opciones correspondientes a la

configuración de parámetros expuesta en las pruebas número 19 y 20 de la Tabla XVII, respectivamente.

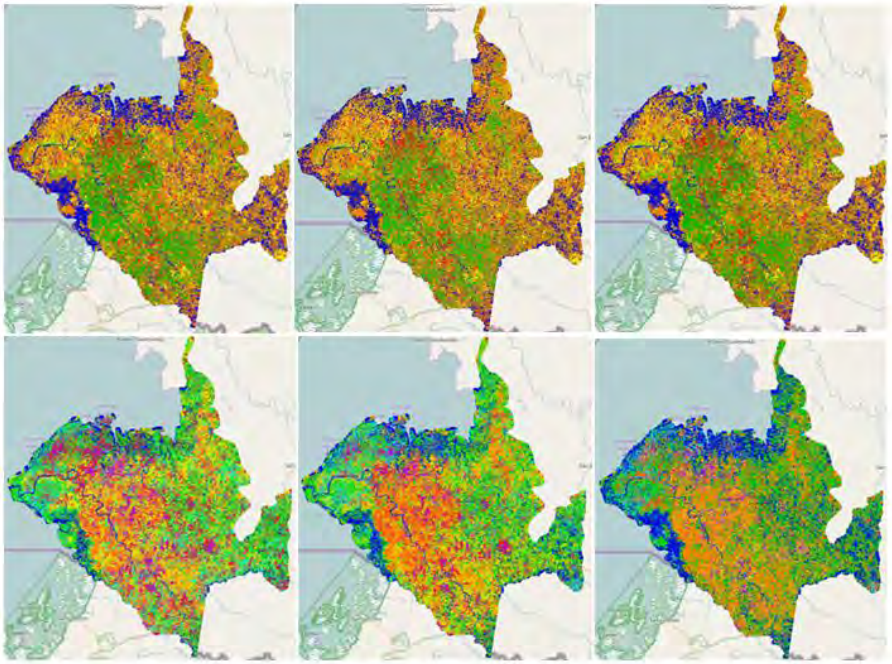


**Figura 39.** Resultados gráficos de la aplicación de isodata, representados en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 23 a la 28 expuestas en la Tabla XVII.

TABLA XVIII
PRUEBAS DE ISODATA EN LOS DATOS COMO RESULTADO DE PCA

| Prueba No. | Número de componentes principales | Procesamiento | | Número de agrupamientos | Prueba | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|---|---|
| | | Normalizar | Centrar | | | | | |
| 1 | 7 | No | No | 5 | 1 | 32.242,42 | 128.687,80 | 100,00000 |
| 2 | 7 | Si | No | 5 | 1 | 32.469,98 | 36.939,29 | 64,00187 |
| 3 | 7 | Si | Si | 5 | 1 | 32.597,06 | 37.029,59 | 63,75894 |
| 4 | 7 | No | Si | 5 | 1 | 32.255,33 | 128.674,90 | 99,97498 |
| 5 | 7 | No | No | 7 | 2 | 27.706,45 | 133.223,80 | 93,22734 |
| 6 | 7 | Si | No | 7 | 2 | 28.426,95 | 40.982,32 | 58,27570 |
| 7 | 7 | Si | Si | 7 | 2 | 28.164,16 | 41.245,11 | 58,77559 |
| 8 | 7 | No | Si | 7 | 2 | 24.587,36 | 136.342,90 | 100,00000 |
| 9 | 3 | Si | No | 5 | 1 | 18.911,47 | 36.757,21 | 99,64516 |
| 10 | 4 | Si | No | 5 | 1 | 22.516,24 | 36.845,42 | 91,75948 |
| 11 | 5 | Si | No | 5 | 1 | 26.202,39 | 37.019,93 | 86,08730 |
| 12 | 3 | No | Si | 7 | 2 | 24.207,41 | 136.335,20 | 99,99721 |
| 13 | 4 | No | Si | 7 | 2 | 24.498,35 | 136.340,40 | 99,40532 |
| 14 | 5 | No | Si | 7 | 2 | 24.572,07 | 136.342,80 | 99,25798 |

Las pruebas 1 a 4, 5 a 8, 9 a 11 y 12 a 14 se comparan entre sí con el fin de obtener el resultado más adecuado para las dos diferentes configuraciones de Isodata. En la primera comparación, el porcentaje de rendimiento indica que la prueba 1 es el mejor resultado, sin embargo en la Figura 40 se observa que la demarcación de los ríos no es la más propicia e inclusive similar al siguiente más alto porcentaje de rendimiento, correspondiente a la prueba 4. Es por ello que se selecciona la prueba 2, que presenta la visualización más adecuada al observar grupos compactos en el mapa, y buena diferenciación del cluster *"agua"*.

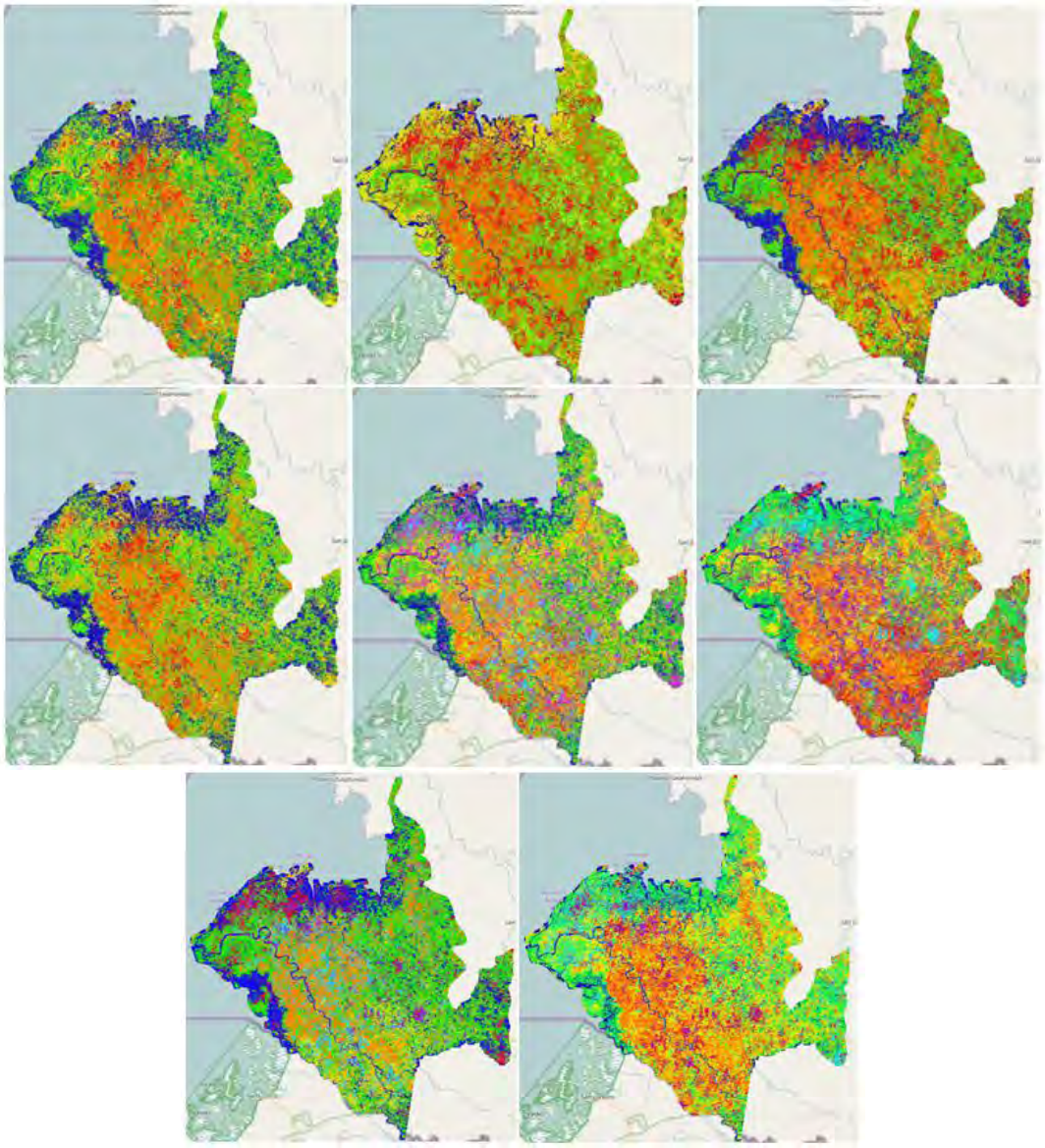


**Figura 40.** Resultado gráfico de la aplicación de Isodata sobre el resultado de PCA, tomando 7 componentes principales, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas 1 a 8 expuestas en la Tabla XVIII.

La segunda comparación resulta más fácil puesto que la decisión del resultado más adecuado, no sólo se respalda con los índices de desempeño, sino también con la visual. La prueba 8 identifica 7 grupos en el municipio de Tumaco, pero se observan solo algunas clases predominantes. De aquí se asumen únicamente 5 clusters. Por esto, la visual es acertada y el resultado presenta un desempeño del 100%, ver Figura 40.

En cuanto a la configuración de parámetros especificados con la opción 1, se determina que el efecto de PCA es relevante puesto que en comparación con la prueba 19 expuesta en la Tabla XVII, el resultado es más detallado y con un mapa de clasificación adecuado. Por otro lado, la prueba definida con la opción 2 no presenta cambios apreciables en la clasificación y el resultado es muy parecido al obtenido en la prueba 20 expuesta en la Tabla XVII.

En cuanto al número de componentes principales, se obtienen los resultados de la Figura 41. La Tabla XVIII indica que tanto para la opción 1 como para la 2, el porcentaje de rendimiento es más alto cuando se toman 3 componentes principales. Sin embargo, en el caso de la opción 1, la visualización es más detallada y define más claramente el cluster *"agua",* en el mapa de 4 componentes principales.

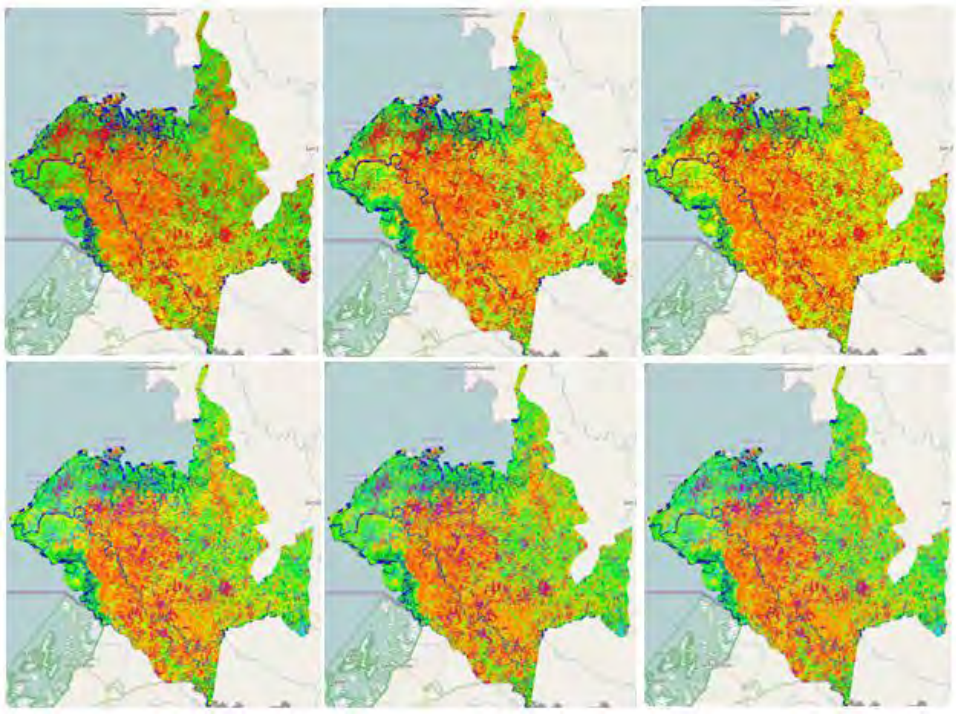


**Figura 41.** Resultado gráfico de la aplicación de isodata sobre el resultado de PCA, tomando 3, 4 y 5 componentes principales, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas 9 a 14 expuestas en la Tabla XVIII.

Teniendo en cuenta que se asumen únicamente 5 clusters, se resuelve extrapolar el modelo del algoritmo Isodata a todo el departamento de Nariño, a partir de la opción 1 de configuración de parámetros considerada en esta sección.

#### 2.3.3.5. GK: Gustafson Kessel

En la Tabla XIX se evidencian las pruebas destacadas para el algoritmo GK y en las Figuras 42 y 43 se observan sus resultados. La prueba número 13 no se grafica porque el resultado visual es el mismo que se muestra para la prueba 12.

TABLA XIX
PRUEBAS DESTACADAS DE GK EN LOS DATOS ORIGINALES DE TUMACO

| Prueba No. | Procesamiento | | Número de iteraciones | Exponente de la matriz U | Mínimo valor de perfeccionamiento | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|---|---|
| | Normalizar | Centrar | | | | | | |
| 1 | No | No | 10 | 2 | 0,000001 | 111335,6 | $4.607\times4e^{-4}$ | 100,00000 |
| 2 | No | No | 50 | 2 | 0,000001 | 111336,4 | $1.781e^{-5}$ | 51,93199 |
| 3 | No | No | 200 | 2 | 0,000001 | 111336,5 | $6.603e^{-8}$ | 50,00676 |
| 4 | No | No | 1000 | 2 | 0,000001 | 111336,5 | $4.747e^{-16}$ | 49,99960 |
| 5 | No | No | 1000 | 2 | 0,01 | 111336,5 | $5.985e^{-10}$ | 49,99966 |
| 6 | No | No | 1000 | 2 | 0,0001 | 111336,5 | $8.325e^{-14}$ | 49,99960 |
| 7 | No | No | 1000 | 2 | 0,0000001 | 111336,5 | $7.189e^{-16}$ | 49,99960 |
| 8 | No | No | 1000 | 2 | 0,00000001 | 111336,5 | $2.856e^{-16}$ | 49,99960 |
| 9 | No | No | 1000 | 5 | 0,0000001 | 111336,5 | $1.037e^{-15}$ | 49,99960 |
| 10 | No | No | 1000 | 10 | 0,0000001 | 111336,5 | $5.351e^{-16}$ | 49,99960 |
| 11 | Si | No | 1000 | 5 | 0,0000001 | 160930,3 | $1.063e^{-16}$ | 34,59125 |
| 12 | No | Si | 1000 | 5 | 0,0000001 | X | X | X |
| 13 | Si | Si | 1000 | 5 | 0,0000001 | X | X | X |

Los mejores índices de desempeño obtenidos en el transcurso de las pruebas los presenta la prueba número 1. En contraste, la prueba 9 es la configuración seleccionada como el mejor resultado para el algoritmo GK dada su visualización en QGIS, que permite distinguir mejor la clase correspondiente al grupo *"agua".* Las pruebas con mayor porcentaje de rendimiento como la número 1, 2, 3 o 5 presentan una visual uniforme donde no se aprecia una buena clasificación.

Los resultados de este algoritmo mejoran si el exponente de la matriz $U$ tiende a 5. Si el exponente es menor a 2, el algoritmo presenta un error debido a que no existe convergencia, y por el contrario, si tiende a 10, la clasificación empeora. El valor de perfeccionamiento debe ser ajustado de acuerdo a la escala presente en la matriz de datos de entrada. En este caso, de acuerdo a las pruebas el mejor

valor fue de 0,0000001, puesto que representa un mínimo cambio en el valor de la función objetivo entre una iteración y la siguiente identificando una convergencia del algoritmo. Por otra parte el resultado visual mejora cuando se incrementa el número de iteraciones.

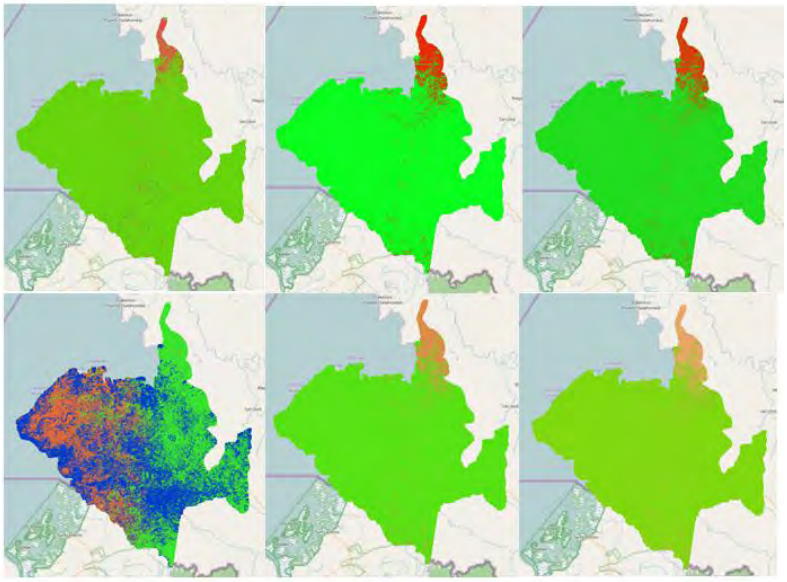


**Figura 42.** Resultados gráficos de la aplicación de GK, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 1 a la 6 expuestas en la Tabla XIX.

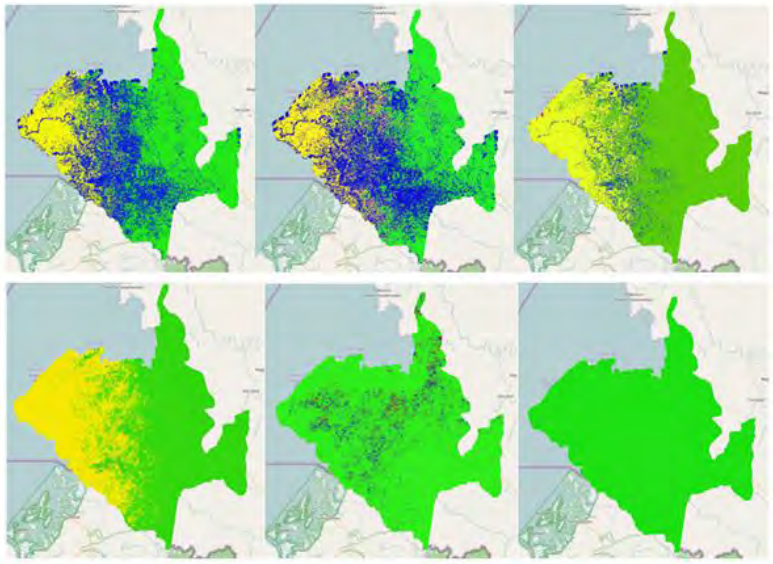


**Figura 43.** Resultados gráficos de la aplicación de GK, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, pruebas de la 7 a la 12 expuestas en la Tabla XIX.

En las pruebas 11, 12 y 13 de la Tabla XIX se nota el efecto negativo del procesamiento de los datos. El efecto de la normalización no es bueno (Figura 43) en la medida en que se pierde la distinción del grupo *"agua",* los clusters no son compactos y no presentan concordancia alguna con el mapa de clasificación de Corponariño. En cuanto al centrado de los datos, se determina que tampoco es conveniente, puesto que el algoritmo identifica todo el data set como un solo grupo. Lo mismo pasa al efectuar un procesamiento completo (pruebas 12 y 13).

Para el caso de PCA los resultados se muestran en la Tabla XX. En este algoritmo los resultados de la aplicación de PCA no son los mejores. En las pruebas 1, 3 y 4, el resultado es erróneo debido a que GK determinó que solo existe un grupo en Tumaco (Figura 44 donde se presenta solo la prueba 1 ya que las otras dos poseen la misma visual). En la prueba número 2, gracias a la normalización de los datos de entrada, se evidencia que sí existen las 5 clases planteadas, sin embargo, éstas no son compactas y no se realiza una buena distinción entre agua y vegetación. Es por ello que no se realizan las pruebas con la variación del número de componentes principales, ya que si los resultados no son buenos con toda la información, la reducción de dimensión no aportaría un resultado relevante.

TABLA XX
PRUEBAS MÁS IMPORTANTES DE GK EN LOS DATOS COMO RESULTADO DE PCA

| Prueba No. | Número de componentes principales | Procesamiento | | Cohesión | Separación | % de Rendimiento |
|---|---|---|---|---|---|---|
| | | Normalizar | Centrar | | | |
| 1 | 7 | No | No | X | X | X |
| 2 | 7 | Si | No | 3'679.686 | 3'518.755 | 100 |
| 3 | 7 | No | Si | X | X | X |
| 4 | 7 | Si | No | X | X | X |

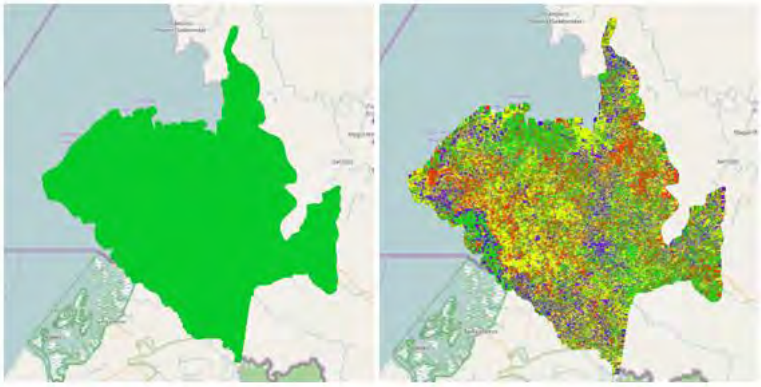


**Figura 44.** Resultado gráfico de la aplicación de GK sobre el resultado de PCA, tomando 7 componentes principales, representado en QGIS. A la izquierda se muestra el resultado de la prueba 1 y a la derecha la prueba número 2, expuestas en la Tabla XX.

Se concluye que el algoritmo no es eficiente a la hora de clasificar este tipo de datos, tanto con la base de datos original de Tumaco, como con el resultado de PCA, debido a su forma y escala. Paralelamente, realizando la comparación visual de su desempeño respecto a los demás algoritmos probados, se determina que no es un método apto para ser extrapolado a todo el departamento de Nariño.

## 2.4. EXTRAPOLACIÓN AL DEPARTAMENTO DE NARIÑO

Para llevar a cabo la extrapolación de los algoritmos de clustering al departamento de Nariño, únicamente se parte de los resultados obtenidos en Tumaco. Se usa la base de datos de todo Nariño, proporcionada por el proyecto Alternar, y se ejecutan los diferentes algoritmos de clustering con la mejor configuración de parámetros obtenida en la sección anterior.

Se hace la extrapolación de los modelos de los algoritmos k-means, EM e Isodata. No se obtiene resultado de la extrapolación de FCM, puesto que se corrió durante una semana, y finalmente se consideró la limitación de tiempo de esta investigación y el elevado tiempo de ejecución para detenerla. Se determina que aunque los resultados obtenidos anteriormente de FCM son buenos, al ejecutarlo con una mayor cantidad de datos no se asegura que converja, y si lo hace tarda un tiempo considerable. En la Tabla XXI se presentan las especificaciones de los algoritmos empleados.

TABLA XXI
CARACTERÍSTIAS DE LOS ALGORITMOS EXTRAPOLADOS A TODO EL DEPARTAMENTO DE NARIÑO

| **BASE DE DATOS ORIGINAL** | | | | | |
|---|---|---|---|---|---|
| **Extrapolación No.** | **Algoritmo de clustering** | **Procesamiento** | | | **Características** |
| | | **Normalizar** | **Centrar** | | |
| 1 | K-Means | No | No | | $k = 13, 'Distance' =$ sqeuclidean, $MaxIter = 500$, $Start' =$ sample |
| 2 | EM | No | No | | $k = 13, \quad iter = 30, \quad Tol = 0{,}000001$ |
| 3 | Isodata | No | No | | $ON = 10e^5, OC = 5e^{-4}, OS = 5e^{-3}, K = 25, L = 2, I = 20, A = 11$ |
| **RESULTADO DE PCA** | | | | | |
| **Extrapolación No.** | **Algoritmo de clustering** | **Procesamiento** | | **Número de componentes principales** | **Características** |
| | | **Normalizar** | **Centrar** | | |
| 4 | K-Means | Si | Si | 3 | $k = 13, 'Distance' =$ sqeuclidean, $MaxIter = 1000$, $Start' =$ sample |
| 5 | EM | No | Si | 4 | $k = 13, \quad iter = 30,$ $Tol = 0{,}000001$ |
| 6 | Isodata | Si | No | 4 | $ON = 100.000, OC = 0{,}0005, OS = 0{,}005,$ $K = 25, L = 2, I = 20, A = 11$ |

Ahora, la matriz de entrada a la ejecución del clustering cuando se usa la base de datos original contiene 34’202.925 puntos o pixeles por 9 atributos, correspondientes a las 7 bandas espectrales y las coordenadas de latitud y longitud. Cuando se usa el resultado de PCA, la matriz contiene igual cantidad de pixeles por $m$ características, siendo $m$ el número de componentes principales a usar.

Para realizar las extrapolaciones se mantiene la mejor configuración de parámetros para cada uno de algoritmos, excepto el número de clases. A diferencia de la región de prueba, y con base en el mapa teórico (clasificación *corpocorin* del mapa de Corponariño), el departamento tendría 13 clases definidas que representan el territorio.

En la Tabla XXII se expone el porcentaje de rendimiento de cada extrapolación. La columna “% de Rendimiento” compara por separado los resultados obtenidos cuando se aplica el clustering a la base de datos original y los obtenidos con PCA. Finalmente se comparan los dos mejores resultados como se indica en la columna “% de Rendimiento 1”. En las Figuras 45 y 46 se presenta la visualización de los resultados obtenidos con la extrapolación de los algoritmos k-means y EM, respectivamente.

TABLA XXII
DESEMPEÑO DE LOS ALGORITMOS DE CLUSTERING EN LA EXTRAPOLACIÓN A TODO EL DEPARTAMENTO DE NARIÑO

| Extrapolación No. | Algoritmo | Datos de entrada | Cohesión | Separación | % de Rendimiento | % de Rendimiento 1 |
|---|---|---|---|---|---|---|
| 1 | K-Means | Base de datos original | 1’799.905,00 | 3’553.792,0 | 54,11525 | 51,68104 |
| 2 | | Resultado de PCA | 60.514,15 | 513.384,5 | 63,61276 | 57,22305 |
| 3 | EM | Base de datos original | 964.679,20 | 2’809.107,0 | 42,20094 | |
| 4 | | Resultado de PCA | 323.566,20 | 1’885.674,0 | 59,35112 | |
| 5 | Isodata | Base de datos original | 148.141,30 | 2’170.612,0 | 80,53938 | |
| 7 | | Resultado de PCA | 87.021,15 | 486.877,5 | 47,67970 | |

Como se observa, a pesar de la gran cantidad de datos y aunque la estructura de ejecución de los algoritmos k-means y EM está basada en un concepto diferente, la extrapolación de estos métodos presenta resultados satisfactorios y similares tanto en su visualización como en índices de desempeño. Se observa que el efecto de PCA en k-means es significativo en cuanto a la mejora de los índices de cohesión y separación, que se ve reflejado en el porcentaje de rendimiento final. Esto se observa en los mapas, donde se realiza una mayor demarcación entre los

diferentes clusters (especialmente en la región sur), a pesar de no percibirse cambios sustanciales con una comparación visual.

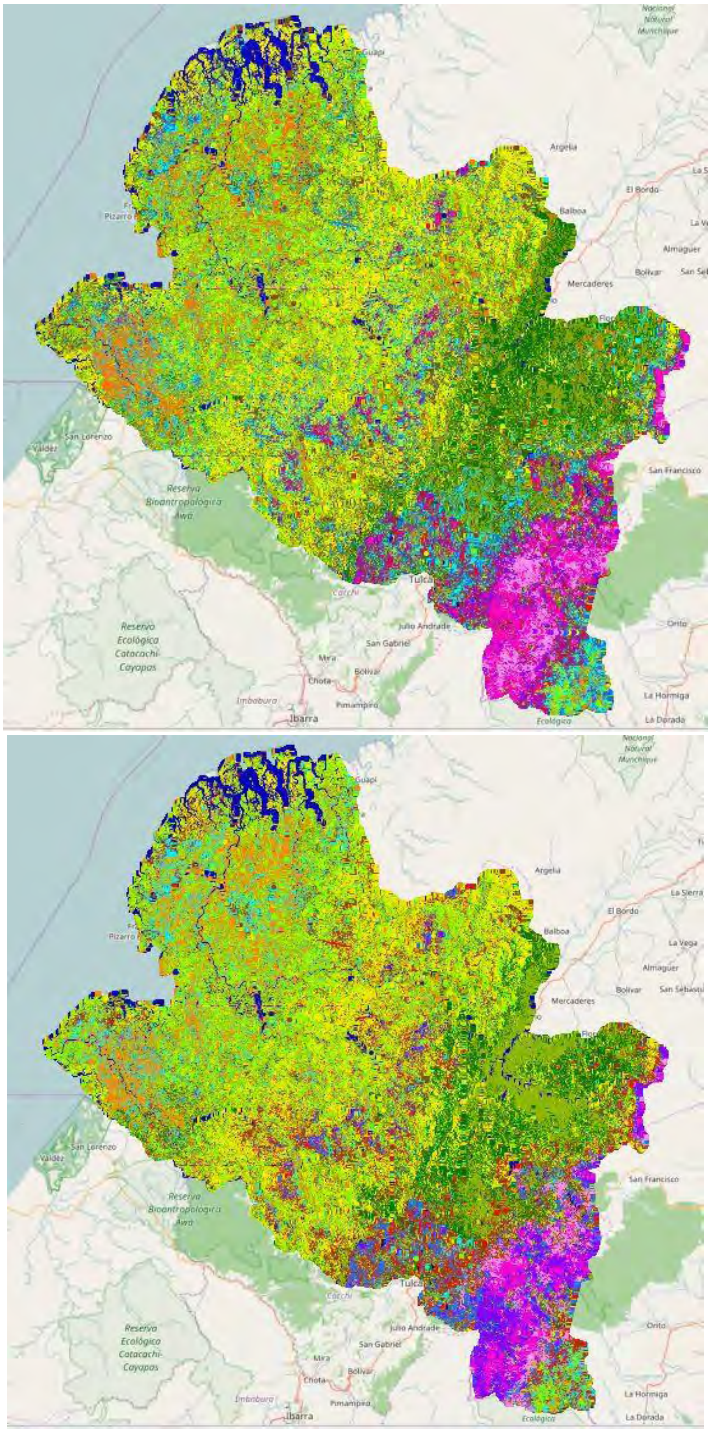



**Figura 45.** Mapa del departamento de Nariño, resultado de la aplicación de K-Means, en la parte superior a partir de la base de datos original y en la parte inferior a partir de los datos obtenidos como resultado de PCA.

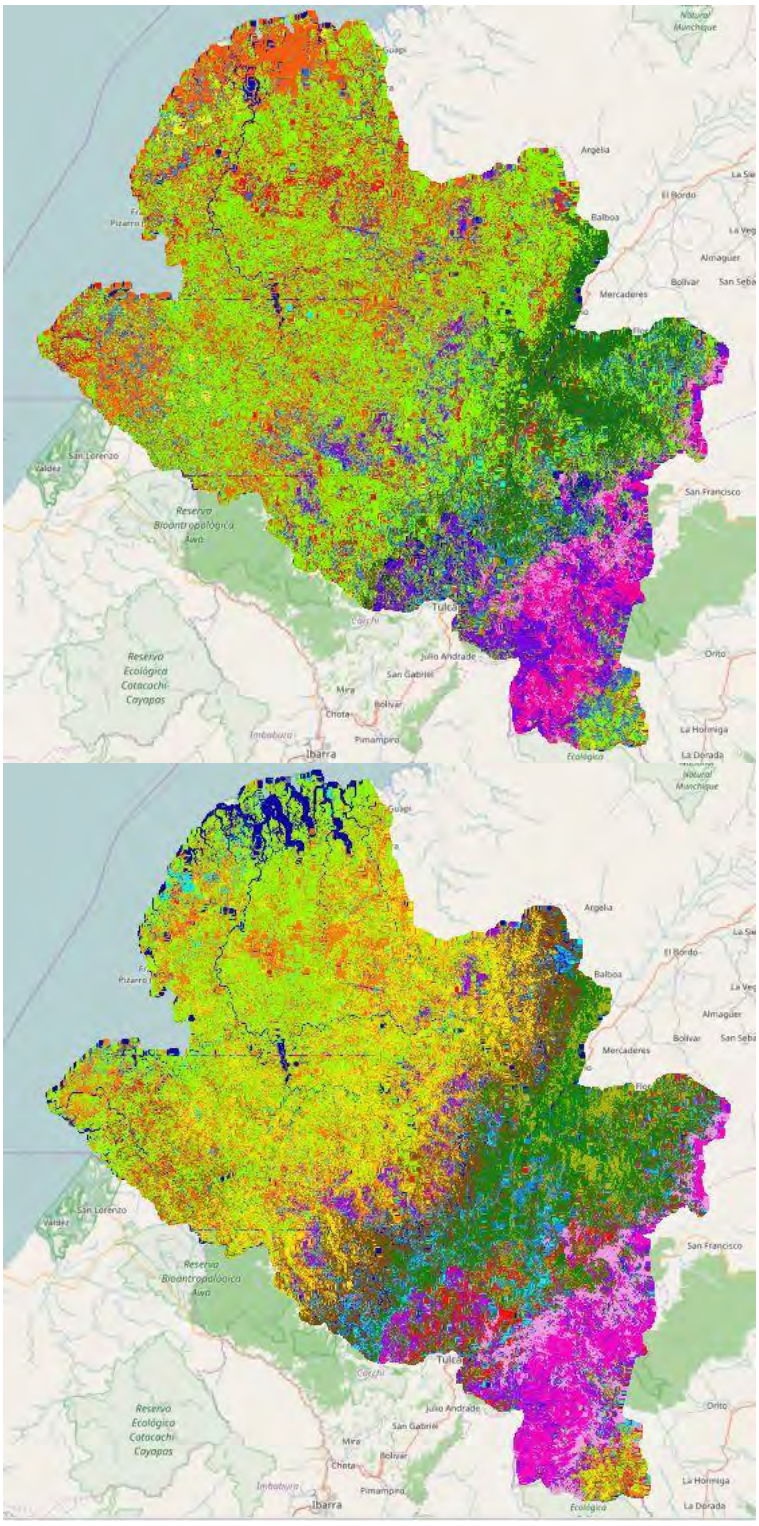



**Figura 46.** Mapa del departamento de Nariño, resultado de la aplicación de EM, en la parte superior a partir de la base de datos original y en la parte inferior a partir de los datos obtenidos como resultado de PCA.

Por su parte, en el algoritmo EM la influencia de la aplicación de PCA es más notoria, especialmente en la demarcación de los clusters. El método EM aplicado sin PCA pierde la correcta definición del grupo “agua” (Figura 46), mientras que con PCA mejora la definición de los ríos. Este indicador de comparación muestra un funcionamiento adecuado debido a la capacidad del algoritmo para conformar grupos con diferentes formas y tamaños, a la vez que controla la varianza y covarianza de los datos. Sin embargo, la métrica de rendimiento de EM es significativamente menor a la obtenida con k-means.

La extrapolación del algoritmo Isodata, con la base de datos original, se hace a partir de la configuración de las dos mejores pruebas obtenidas en Tumaco, puesto que poseen los mismos parámetros, a diferencia del número máximo de clusters. En este caso se asigna un $K = 25$ para que el algoritmo esté en libertad de proponer una clasificación, que probablemente pueda establecer la cantidad de clusters que hay en el departamento de Nariño. Esta extrapolación es la número 3 de la Tabla XXI, y su visualización se presenta en la Figura 47, donde se observa que el resultado disminuye notablemente su desempeño en la identificación del grupo “agua” en comparación a los otros dos algoritmos.

Por su parte, la aplicación del algoritmo a la base de datos resultado de PCA, asumiendo únicamente 5 clusters, parte de la configuración establecida por la prueba 10 de la Tabla XVIII, su visualización se muestra en la Figura 47, y se puede apreciar que esta extrapolación (número 6 de la Tabla XXI), mejora drásticamente el resultado, no solo identifica de una manera más coherente el cluster *“agua”,* sino que también agrupa los demás clusters de una forma semejante a los resultados obtenidos con k-means y EM. Sin embargo, se observa una mezcla entre coberturas dado que este método establece que algunos de los pixeles de cierto cluster no están lo suficientemente distantes de otros grupos para considerarse coberturas independientes al no superar los umbrales definidos y pueden ser clasificados erróneamente.

A diferencia de los otros dos algoritmos, la aplicación de Isodata con PCA produce disminución en el rendimiento ya que las características de la base de datos original tienen mayor información que puede ser usada por el algoritmo para identificar grupos más definidos y compactos.

Es así como al comparar los resultados de rendimiento (Tabla XXII), el mejor desempeño se obtiene en las extrapolaciones de Isodata con los datos originales y k-means con PCA. Sin embargo, el indicador de comparación con los mapas de referencia muestra que la extrapolación de Isodata no tiene una definición adecuada de los clusters, por lo explicado anteriormente.

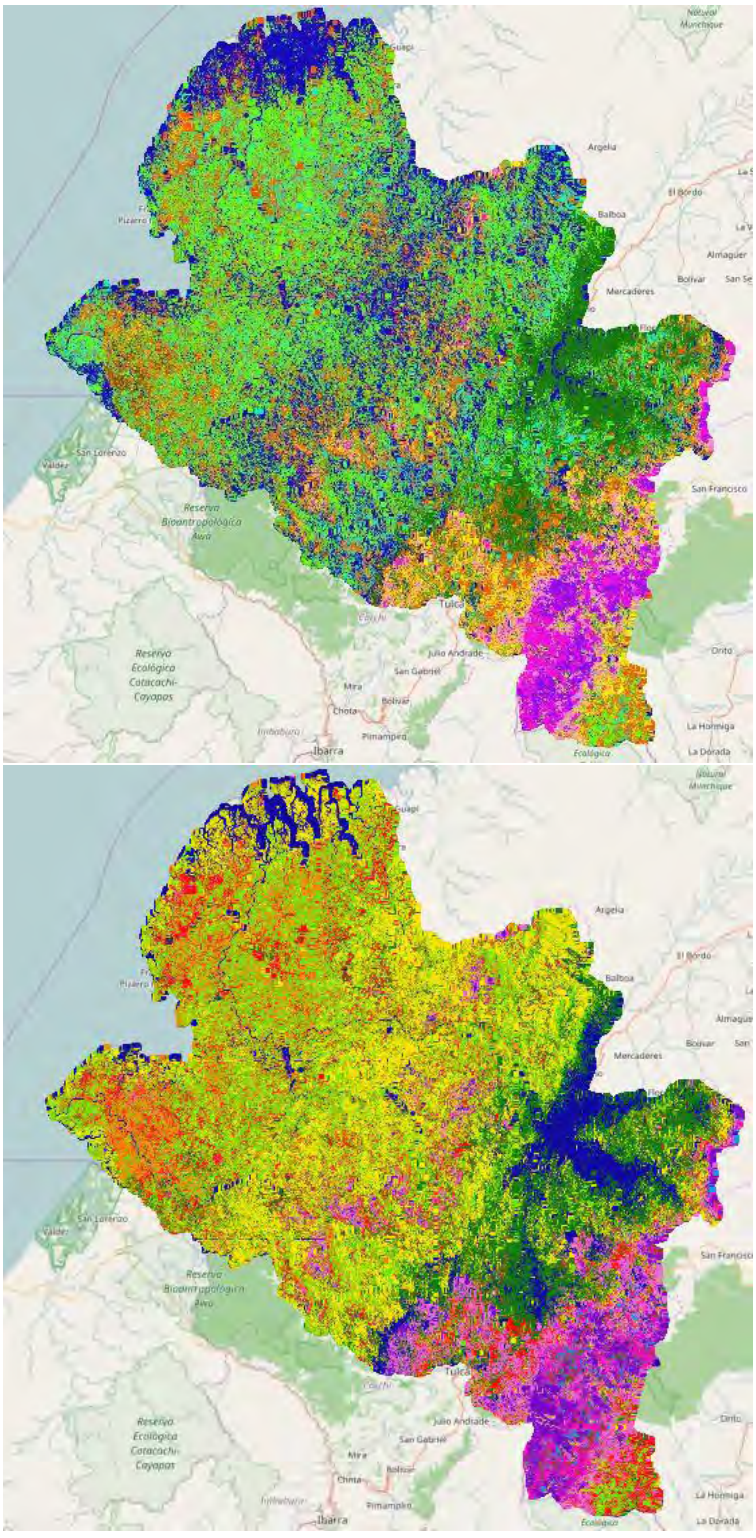



**Figura 47.** Mapa del departamento de Nariño, resultado de la aplicación de Isodata, en la parte superior a partir de la base de datos original y en la parte inferior a partir de los datos obtenidos como resultado de PCA.

De esta manera, se concluye que el algoritmo k-means presenta los mejores resultados independiente de los datos de entrada propuestos. Además, es un algoritmo que tiene un comportamiento consistente a lo largo de las pruebas dado que la distancia euclidiana considera todas las características para medir la separación inter-cluster e intra-cluster, y se puede adaptar fácilmente a diferentes condiciones de los datos, ya sea en su forma original o con procesamiento.

Por otra parte, para comparar la carga computacional de los algoritmos se tiene en cuenta que la aplicación del clustering se realiza en un computador portátil con un procesador Intel Core i5 de 2.2GHz, 8GB de RAM y 1TB de disco duro. El tiempo de ejecución de los métodos es similar ya que para la extrapolación de la base de datos original, k-means y EM tardaron 51 y 46 horas, respectivamente, y la mitad del tiempo para los datos con PCA. Para el caso de Isodata, el rendimiento computacional fue superior, dado que los procesos tardaron 23 y 10 horas para la base original y los datos con PCA, respectivamente.

Finalmente, siguiendo la metodología planteada con base en el porcentaje de rendimiento, se elige k-means con PCA (extrapolación número 2 de la Tabla XXII, Figura 45) como el mejor resultado comparativo y con el que se genera el mapa final de clasificación de tipos de cobertura del departamento de Nariño.

## 2.5. IDENTIFICACIÓN DE CLASES

En esta etapa se identifica el tipo de cobertura de terreno que representa cada una de las 13 clases definidas en el proceso anterior. Para esto se hace uso del servidor y la herramienta PostgreSQL, con el fin de ejecutar consultas que permiten establecer las siguientes fases:

- **Extracción de base de datos teórica:** la base para proponer el resultado final es el mapa de clasificación de Corponariño. Se selecciona la clasificación por *cobertura* puesto que es más detallada que *corpocorin*. Teniendo en cuenta que el modelo propuesto plantea únicamente 13 grupos, se decide extraer una nueva base de datos en la que se almacene una reducción de clases similares de los datos de *cobertura.* Esto se realiza ya que en los 63 grupos del mapa existen muchas subdivisiones de una misma clase que probablemente poseen una firma espectral semejante. Por ejemplo, el caso de *“bosque guandal”, “bosque guandal intervenido”,* y otros, que pueden resumirse en una clase un poco más general llamada *“Bosque”*.

- **Caracterización de clases:** Se efectúa el cruce entre las dos bases de datos, la tabla de reducción de grupos a partir de *cobertura* del mapa de Corponariño y la tabla de la clasificación final obtenida mediante clustering, con el propósito de evaluar la correspondencia de las dos clasificaciones

para determinar a qué clase de cobertura pertenecen cada uno de los grupos encontrados. Así, se toman los 13 clusters uno a uno y se contabiliza la cantidad de pixeles que pertenecen a cada tipo de *cobertura*. Finalmente se obtiene el porcentaje de los diferentes tipos de *cobertura* con respecto a una misma clase del modelo final, y aquella cobertura que cuenta con el porcentaje más alto se toma como la cobertura que representa a ese cluster.

- **Consolidar el mapa final:** para el caso en que se presenta una reducción en la cantidad de grupos del modelo final, se plantea seleccionar el siguiente porcentaje más alto, repitiendo el proceso si es necesario, y comparar el resultado con el mapa de clasificación de coberturas de Corponariño para definir el tipo de cobertura correspondiente a la clase en cuestión.

De esta manera, el procedimiento para obtener el mapa final de clasificación con su respectiva caracterización de tipos de biomasa del departamento de Nariño se muestra a continuación.

En el servidor del grupo de investigación GIIEE se encuentra la tabla denominada *bdcoberturas*, que contiene 34’202.925 filas correspondientes a diferentes puntos geo-referenciados en el mapa de Corponariño. De las 5 columnas, las dos primeras corresponden a la latitud y la longitud que identifican el punto, la tercera columna es la ID del municipio, la cuarta columna identifica la clase que se le asigna al punto dependiendo de la reducción de clases que se hace a partir de *cobertura* del mapa de Corponariño, y por último se muestra la clase según *corpocorin*. La Figura 48 muestra 20 filas que pertenecen a la tabla *bdcoberturas*.

| | latitude integer | longitude integer | id_municipio integer | cobertura character varying | corpocorin character varying |
|---|---|---|---|---|---|
| 1 | -8641380 | 198960 | 48 | 5 | Cultivos permanentes |
| 2 | -8641380 | 198990 | 48 | 5 | Cultivos permanentes |
| 3 | -8641380 | 199020 | 48 | 5 | Cultivos permanentes |
| 4 | -8641380 | 199440 | 48 | 5 | Cultivos permanentes |
| 5 | -8641380 | 199470 | 48 | 5 | Cultivos permanentes |
| 6 | -8641350 | 198930 | 48 | 5 | Cultivos permanentes |
| 7 | -8641350 | 198960 | 48 | 5 | Cultivos permanentes |
| 8 | -8641350 | 198990 | 48 | 5 | Cultivos permanentes |
| 9 | -8641350 | 199020 | 48 | 5 | Cultivos permanentes |
| 10 | -8641350 | 199050 | 48 | 5 | Cultivos permanentes |
| 11 | -8641350 | 199080 | 48 | 5 | Cultivos permanentes |
| 12 | -8641350 | 199140 | 48 | 5 | Cultivos permanentes |
| 13 | -8641350 | 199200 | 48 | 5 | Cultivos permanentes |
| 14 | -8641350 | 199230 | 48 | 5 | Cultivos permanentes |
| 15 | -8641350 | 199260 | 48 | 5 | Cultivos permanentes |
| 16 | -8641350 | 199290 | 48 | 5 | Cultivos permanentes |
| 17 | -8641350 | 199440 | 48 | 5 | Cultivos permanentes |
| 18 | -8641350 | 199470 | 48 | 5 | Cultivos permanentes |
| 19 | -8641350 | 200100 | 48 | 5 | Cultivos permanentes |
| 20 | -8641350 | 200130 | 48 | 5 | Cultivos permanentes |

**Figura 48.** Tabla compuesta de 20 muestras pertenecientes a la tabla *bdcoberturas*.

Mediante la ejecución de consultas en PostgreSQL, se conforma la base de datos final en la que se basa la identificación de clases a partir del mapa teórico de Corponariño. En la Tabla XXIII se muestran las correspondencias utilizadas para formar la base de datos.

TABLA XXIII
REDUCCIÓN DE CLASES *Cobertura* PARA CONFORMAR LA TABLA *bdcoberturas*

| **Cobertura - Clase No.** | **Cobertura** | **Reducción – Clase No.** | **Cobertura reducción** |
|---|---|---|---|
| 1 | Afloramientos rocosos | 1 | Zonas de extracción minera y escombros |
| 2 | Arbustos y matorrales | 2 | Áreas con vegetación herbácea y/o arbustiva |
| 3 | Bosque bajo | 3 | Bosques |
| 4 | Bosque Bajo | 3 | Bosques |
| 5 | Bosque de colina | 3 | Bosques |
| 6 | Bosque de guandal | 3 | Bosques |
| 7 | Bosque de guandal intervenido | 3 | Bosques |
| 8 | Bosque de mangle | 3 | Bosques |
| 9 | Bosque de palma naidi | 3 | Bosques |
| 10 | Bosque de piedemonte amazónico | 4 | Bosque de piedemonte amazónico |
| 11 | Bosque guandal | 3 | Bosques |
| 12 | Bosque guandal con predominio de palma naidi | 3 | Bosques |
| 13 | Bosque guandal intervenido | 3 | Bosques |
| 14 | Bosque húmedo piedemonte pacifico | 3 | Bosques |
| 15 | Bosque plantado | 3 | Bosques |
| 16 | Bosque primario | 3 | Bosques |
| 17 | Bosque primario cuangarial intervenido | 3 | Bosques |
| 18 | Bosque primario de colinas bajas | 3 | Bosques |
| 19 | Bosque primario intervenido | 3 | Bosques |
| 20 | Bosque primario sajal | 3 | Bosques |
| 21 | Bosque primario sajal intervenido | 3 | Bosques |
| 22 | Bosque ripario | 3 | Bosques |
| 23 | Bosque secundario | 3 | Bosques |
| 24 | Bosque secundario alto andino | 3 | Bosques |
| 25 | Bosque secundario intervenido | 3 | Bosques |
| 26 | Café | 5 | Cultivos mixtos |
| 27 | Caña panelera | 5 | Cultivos mixtos |
| 28 | Cañales | 5 | Cultivos mixtos |
| 29 | Centros poblados | 6 | Zonas urbanizadas |
| 30 | Cultivo mixto con predominio de Café | 5 | Cultivos mixtos |
| 31 | Cultivos anuales o transitorios | 7 | Cultivos anuales o transitorios |
| 32 | Cultivos de clima cálido | 5 | Cultivos mixtos |
| 33 | Cultivos mixtos | 5 | Cultivos mixtos |

| CONTINUACIÓN TABLA XXIII | | | |
|---|---|---|---|
| 34 | Cultivos mixtos con predominio Café | 5 | Cultivos mixtos |
| 35 | Cultivos mixtos con predominio Cana | 5 | Cultivos mixtos |
| 36 | Cultivos mixtos con predominio de Palma | 8 | Cultivos de palma |
| 37 | Cultivos permanentes | 9 | Cultivos permanentes |
| 38 | Estanques piscícolas | 10 | Agua |
| 39 | Esteros | 11 | Áreas húmedas continentales |
| 40 | Estuarios | 10 | Agua |
| 41 | Lagunas, lagos y ciénagas | 10 | Agua |
| 42 | Mares y océanos | 10 | Agua |
| 43 | Mosaico de cultivos, pastos y espacios naturales | 12 | Áreas agrícolas heterogéneas |
| 44 | Mosaico de pastos y cultivos | 12 | Áreas agrícolas heterogéneas |
| 45 | Palma africana | 8 | Cultivos de palma |
| 46 | Pastos arbolados | 13 | Pastos |
| 47 | Pastos clima cálido | 13 | Pastos |
| 48 | Pastos enmalezados o enrastrojados | 13 | Pastos |
| 49 | Pastos limpios | 13 | Pastos |
| 50 | Pastos naturales | 13 | Pastos |
| 51 | pendiente | 12 | Áreas agrícolas heterogéneas |
| 52 | pendiente (cultivos) | 12 | Áreas agrícolas heterogéneas |
| 53 | Playas y arenales | 14 | Áreas abiertas, sin o con poca vegetación |
| 54 | Rastrojo alto | 2 | Áreas con vegetación herbácea y/o arbustiva |
| 55 | Rastrojo bajo | 2 | Áreas con vegetación herbácea y/o arbustiva |
| 56 | Ríos | 10 | Agua |
| 57 | Tierras desnudas o degradadas | 14 | Áreas abiertas, sin o con poca vegetación |
| 58 | Vegetación achaparrada | 2 | Áreas con vegetación herbácea y/o arbustiva |
| 59 | Vegetación de paramo | 15 | Vegetación de páramo |
| 60 | Zonas de extracción minera | 1 | Zonas de extracción minera y escombros |
| 61 | Zonas pantanosas | 11 | Áreas húmedas continentales |
| 62 | Zonas quemadas | 14 | Áreas abiertas, sin o con poca vegetación |
| 63 | Zonas quemadas (Pastos o Rastrojos) | 14 | Áreas abiertas, sin o con poca vegetación |

Las 63 clases de *cobertura* del mapa de Corponariño se agrupan en las 15 clases especificadas en la columna “Cobertura reducción” como se muestra en la tabla anterior. De esta manera, la tabla *bdcoberturas* contiene la base de datos reducida a 15 clases, cuyo resultado se grafica en QGIS para representar el mapa teórico del departamento de Nariño (Figura 49). En la Tabla XXIV se muestra la cantidad de puntos que pertenecen a cada una de las 15 clases reducidas.

Por otro lado, la tabla *kmfinal_dpto* contiene 34’202.925 filas correspondientes a diferentes pixeles geo-referenciados en el modelo final obtenido en la presente investigación. De las 3 columnas, las dos primeras corresponden a la latitud y la longitud que identifican el pixel, y la tercera columna es la clase a la que pertenece

según la clasificación obtenida mediante clustering, es decir, un número entre 1 y 13 identificado por *type*. La Figura 50 muestra 20 filas que pertenecen a la tabla *kmfinal_dpto*.

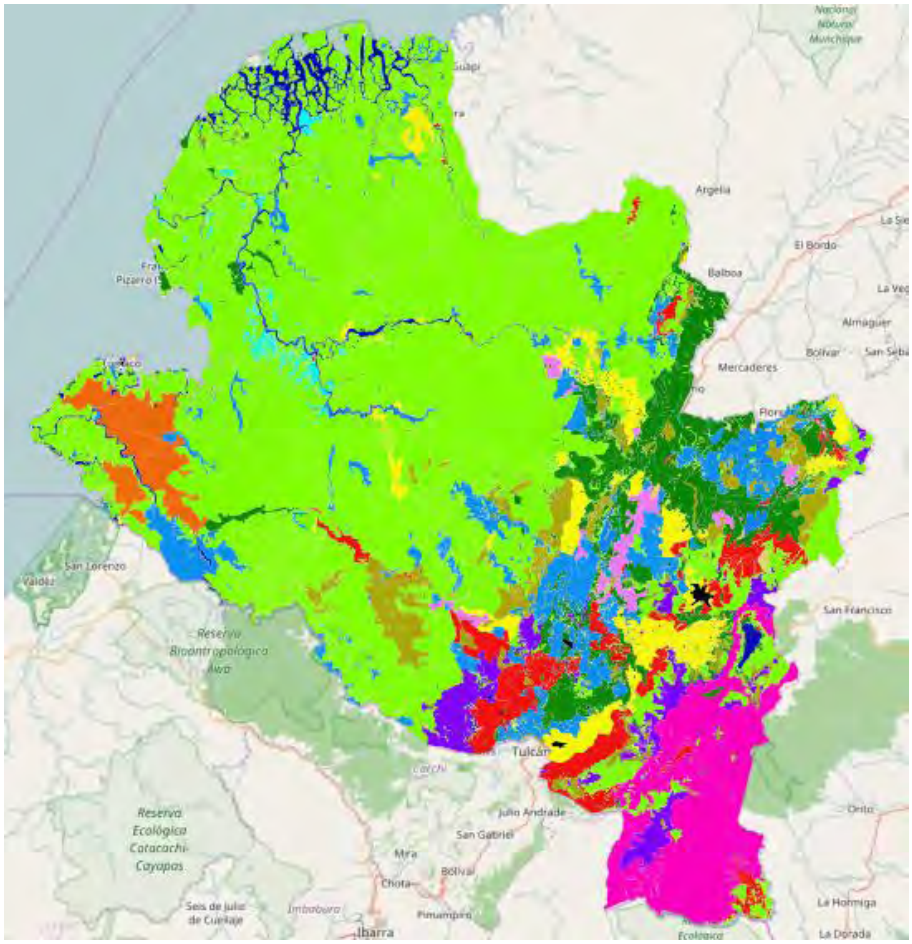



**Figura 49.** Mapa teórico obtenido con 15 clases definidas en la Tabla XXIII a partir de la clasificación por *cobertura* del mapa de Corponariño.

TABLA XXIV
CANTIDAD DE PUNTOS PERTENECIENTES A CADA CLASE REDUCIDA

| **Reducción – Clase No.** | **Color mapa Figura 49** | **Cobertura reducción** | **Total puntos** |
|---|---|---|---|
| 1 | Café | Zonas de extracción minera y escombros | 10.267 |
| 2 | Verde pasto | Áreas con vegetación herbácea y/o arbustiva | 1’496.541 |
| 3 | Verde fluorescente | Bosques | 19’441.232 |
| 4 | Fucsia | Bosque de piedemonte amazónico | 2’049.904 |
| 5 | Azul claro | Cultivos mixtos | 2’601.168 |
| 6 | Negro | Zonas urbanizadas | 57.428 |
| 7 | Rojo | Cultivos anuales o transitorios | 1’382.562 |

| CONTINUACIÓN TABLA XXIV | | | |
|---|---|---|---|
| 8 | Naranja | Cultivos de palma | 758.743 |
| 9 | Rosado | Cultivos permanentes | 328.883 |
| 10 | Azul oscuro | Agua | 607.234 |
| 11 | Azul fluorescente | Áreas húmedas continentales | 246.097 |
| 12 | Amarillo | Áreas agrícolas heterogéneas | 1’295.404 |
| 13 | Verde oscuro | Pastos | 2’855.125 |
| 14 | Piel | Áreas abiertas, sin o con poca vegetación | 46.565 |
| 15 | Morado | Vegetación de páramo | 875.372 |
| | | Total puntos | 34’052.525 |
| | Blanco | Vacíos | 150.400 |
| | | Total puntos | 34’202.925 |

| | latitude integer | longitude integer | type integer |
|---|---|---|---|
| 1 | -8659830 | 112230 | 2 |
| 2 | -8659830 | 112260 | 2 |
| 3 | -8659830 | 112290 | 2 |
| 4 | -8659830 | 112320 | 2 |
| 5 | -8659830 | 112350 | 2 |
| 6 | -8659830 | 112380 | 6 |
| 7 | -8659830 | 112410 | 9 |
| 8 | -8659830 | 112440 | 2 |
| 9 | -8659830 | 112470 | 9 |
| 10 | -8659830 | 112500 | 8 |
| 11 | -8659830 | 112530 | 12 |
| 12 | -8659830 | 112560 | 12 |
| 13 | -8659830 | 112590 | 9 |
| 14 | -8659830 | 112620 | 9 |
| 15 | -8659830 | 112650 | 2 |
| 16 | -8659830 | 112680 | 2 |
| 17 | -8659830 | 112710 | 2 |
| 18 | -8659830 | 112740 | 6 |
| 19 | -8659830 | 112770 | 6 |
| 20 | -8659830 | 112800 | 6 |

**Figura 50.** Tabla compuesta de 20 muestras pertenecientes a la tabla *kmfinal_dpto*.

De igual forma que para la tabla *bdcoberturas*, se contabilizan los pixeles pertenecientes a cada uno de los clusters (Tabla XXV).

El cruce efectuado entre las dos tablas, *bdcoberturas* y *kmfinal_dpto*, resulta en la correspondencia de puntos, con el fin de evaluar a qué tipo de cobertura pertenecen las clases definidas mediante clustering. En la Figura 51 se muestra un ejemplo del resultado que arroja el cruce entre las dos tablas. Se seleccionan únicamente los atributos de interés, columnas “*latitude*, *longitude*, *cobertura* y *type*” (Figura 51). De esta manera, se contabiliza la cantidad de puntos por

cobertura pertenecientes a cada una de las clases obtenidas en el mapa final, resultado que se muestra en la Tabla XXVI.

TABLA XXV
CANTIDAD DE PUNTOS PERTENECIENTES A CADA CLASE OBTENIDA MEDIANTE CLUSTERING

| Cluster No. | Color en el mapa de la Figura 45 | Total puntos |
|---|---|---|
| 1 | Verde fluorescente | 7'366.299 |
| 2 | Azul claro | 974.361 |
| 3 | Verde pasto | 2'332.017 |
| 4 | Rosado | 628.017 |
| 5 | Naranja | 4'057.274 |
| 6 | Café | 2'902.454 |
| 7 | Morado | 892.270 |
| 8 | Azul oscuro | 1'003.868 |
| 9 | Rojo | 1'854.276 |
| 10 | Fucsia | 596.225 |
| 11 | Azul fluorescente | 2'451.715 |
| 12 | Amarillo | 6'151.122 |
| 13 | Verde oscuro | 2'993.027 |
| | Total puntos | 34'202.925 |

| | latitude integer | longitude integer | id_municipio integer | cobertura character varying | corpocorin character varying | type integer |
|---|---|---|---|---|---|---|
| 1 | -8702670 | 283020 | 50 | 10 | Aguas continentales | 12 |
| 2 | -8702670 | 283050 | 50 | 10 | Aguas continentales | 12 |
| 3 | -8702670 | 283080 | 50 | 10 | Aguas continentales | 12 |
| 4 | -8702670 | 283110 | 50 | 10 | Aguas continentales | 12 |
| 5 | -8702670 | 283140 | 50 | 10 | Aguas continentales | 12 |
| 6 | -8702670 | 283170 | 50 | 10 | Aguas continentales | 1 |
| 7 | -8702670 | 283200 | 50 | 10 | Aguas continentales | 1 |
| 8 | -8702670 | 283980 | 50 | 10 | Aguas continentales | 8 |
| 9 | -8702670 | 284010 | 50 | 10 | Aguas continentales | 8 |
| 10 | -8702670 | 284040 | 50 | 10 | Aguas continentales | 8 |
| 11 | -8702670 | 284070 | 50 | 10 | Aguas continentales | 8 |
| 12 | -8702670 | 284100 | 50 | 10 | Aguas continentales | 12 |
| 13 | -8591580 | 150870 | 40 | 13 | Areas con vegetacion herbacea y/o arbustiva | 13 |
| 14 | -8591580 | 150900 | 40 | 13 | Areas con vegetacion herbacea y/o arbustiva | 13 |
| 15 | -8591580 | 150930 | 40 | 13 | Areas con vegetacion herbacea y/o arbustiva | 3 |
| 16 | -8591580 | 150960 | 40 | 13 | Areas con vegetacion herbacea y/o arbustiva | 3 |
| 17 | -8591580 | 150990 | 40 | 13 | Areas con vegetacion herbacea y/o arbustiva | 3 |
| 18 | -8591580 | 151020 | 40 | 13 | Areas con vegetacion herbacea y/o arbustiva | 13 |
| 19 | -8591580 | 151050 | 40 | 13 | Areas con vegetacion herbacea y/o arbustiva | 13 |
| 20 | -8591580 | 151080 | 40 | 13 | Areas con vegetacion herbacea y/o arbustiva | 3 |

| | latitude integer | longitude integer | cobertura character varying | type integer |
|---|---|---|---|---|
| 1 | -8702670 | 283020 | 10 | 12 |
| 2 | -8702670 | 283050 | 10 | 12 |
| 3 | -8702670 | 283080 | 10 | 12 |
| 4 | -8702670 | 283110 | 10 | 12 |
| 5 | -8702670 | 283140 | 10 | 12 |
| 6 | -8702670 | 283170 | 10 | 1 |
| 7 | -8702670 | 283200 | 10 | 1 |
| 8 | -8702670 | 283980 | 10 | 8 |
| 9 | -8702670 | 284010 | 10 | 8 |
| 10 | -8702670 | 284040 | 10 | 8 |
| 11 | -8702670 | 284070 | 10 | 8 |
| 12 | -8702670 | 284100 | 10 | 12 |
| 13 | -8591580 | 150870 | 13 | 13 |
| 14 | -8591580 | 150900 | 13 | 13 |
| 15 | -8591580 | 150930 | 13 | 3 |
| 16 | -8591580 | 150960 | 13 | 3 |
| 17 | -8591580 | 150990 | 13 | 3 |
| 18 | -8591580 | 151020 | 13 | 13 |
| 19 | -8591580 | 151050 | 13 | 13 |
| 20 | -8591580 | 151080 | 13 | 3 |

**Figura 51.** Muestra en 20 líneas del cruce efectuado entre las tablas *bdcoberturas* y *kmfinal_dpto* mediante consulta en SQL, a la izquierda todos los atributos, a la derecha las características de interés.

De acuerdo con esto, se calcula el porcentaje de aporte de cada una de las coberturas en cada uno de los tipos identificados por el algoritmo de clustering, como se presenta en la Tabla XXVII.

TABLA XXVI

CANTIDAD DE PUNTOS PERTENECIENTES A CADA CLASE DISCRIMINADOS POR COBERTURA

| Type Cobertura | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 1.823 | 101 | 1.460 | 63 | 798 | 850 | 4 | 2.172 | 261 | 0 | 406 | 1.335 | 994 |
| **2** | 251.474 | 47.004 | 142.347 | 35.252 | 125.371 | 118.770 | 26.463 | 20.842 | 77.717 | 23.221 | 64.267 | 313.930 | 249.883 |
| **3** | 5’723.903 | 303.658 | 136.141 | 136.982 | 2’686.955 | 2’136.014 | 151.602 | 521.849 | 816.962 | 44.109 | 1’607.551 | 4’576.947 | 598.559 |
| **4** | 120.573 | 198.402 | 14.769 | 305.600 | 56.419 | 56.668 | 522.943 | 17.062 | 178.715 | 397.917 | 98.495 | 55.904 | 26.437 |
| **5** | 427.127 | 27.457 | 285.580 | 2.168 | 332.905 | 129.406 | 2.748 | 41.172 | 123.703 | 154 | 172.164 | 336.411 | 720.173 |
| **6** | 1.428 | 6.700 | 21.776 | 303 | 1.638 | 1.790 | 677 | 4.453 | 6.904 | 151 | 1.317 | 2.835 | 7.456 |
| **7** | 97.862 | 145.059 | 172.156 | 19.867 | 93.516 | 83.303 | 17.324 | 11.619 | 243.374 | 4.469 | 113.328 | 123.429 | 257.256 |
| **8** | 108.222 | 7.519 | 2.218 | 2.088 | 348.074 | 26.346 | 186 | 8.301 | 35.696 | 22 | 170.737 | 22.168 | 27.166 |
| **9** | 36.449 | 7 | 74.423 | 9.108 | 17.939 | 11.245 | 5.475 | 2.775 | 1.212 | 5.018 | 4.448 | 63.313 | 117.072 |
| **10** | 100.650 | 8.418 | 3.724 | 360 | 65.164 | 39.358 | 154 | 232.621 | 13.168 | 72 | 32.353 | 86.243 | 5.934 |
| **11** | 81.903 | 1.185 | 937 | 8.311 | 47.298 | 17.054 | 3.882 | 24.559 | 3.973 | 729 | 24.108 | 38.053 | 6.441 |
| **12** | 186.190 | 42.051 | 193.938 | 28.029 | 125.773 | 88.840 | 15.557 | 17.615 | 122.689 | 6.289 | 61.610 | 147.314 | 296.462 |
| **13** | 197.667 | 63.093 | 1’129.798 | 159 | 142.040 | 114.301 | 328 | 49.892 | 119.695 | 26 | 88.051 | 270.045 | 630.668 |
| **14** | 3.232 | 134 | 17.983 | 71.097 | 2.263 | 1.594 | 132.315 | 2.435 | 1.133 | 107.921 | 762 | 3.525 | 12.991 |
| **15** | 9.352 | 113.745 | 121.626 | 8.630 | 1.937 | 64.051 | 12.612 | 35.119 | 97.679 | 6.127 | 2.421 | 88.991 | 29.118 |
| **Vacío** | 18.444 | 9.828 | 13.141 | 63 | 9.184 | 12.864 | 4 | 11.382 | 11.395 | 23.221 | 9.697 | 20.679 | 6.417 |

TABLA XXVII

PORCENTAJE DE PERTENENCIA DE CADA COBERTURA EN CADA CLASE (Clustering)

| Cobertura | Type | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 1 | 0,02 | 0,01 | 0,06 | 0,01 | 0,02 | 0,03 | 0,00 | 0,22 | 0,01 | 0,00 | 0,02 | 0,02 | 0,03 |
| 2 | 3,41 | 4,82 | 6,10 | 5,61 | 3,09 | 4,09 | 2,97 | 2,08 | 4,19 | 3,89 | 2,62 | 5,10 | 8,35 |
| 3 | 77,70 | 31,16 | 5,84 | 21,81 | 66,23 | 73,59 | 16,99 | 51,98 | 44,06 | 7,40 | 65,57 | 74,41 | 20,00 |
| 4 | 1,64 | 20,36 | 0,63 | 48,66 | 1,39 | 1,95 | 58,61 | 1,70 | 9,64 | 66,74 | 4,02 | 0,91 | 0,88 |
| 5 | 5,80 | 2,82 | 12,25 | 0,35 | 8,21 | 4,46 | 0,31 | 4,10 | 6,67 | 0,03 | 7,02 | 5,47 | 24,06 |
| 6 | 0,02 | 0,69 | 0,93 | 0,05 | 0,04 | 0,06 | 0,08 | 0,44 | 0,37 | 0,03 | 0,05 | 0,05 | 0,25 |
| 7 | 1,33 | 14,89 | 7,38 | 3,16 | 2,30 | 2,87 | 1,94 | 1,16 | 13,13 | 0,75 | 4,62 | 2,01 | 8,60 |
| 8 | 1,47 | 0,77 | 0,10 | 0,33 | 8,58 | 0,91 | 0,02 | 0,83 | 1,93 | 0,00 | 6,96 | 0,36 | 0,91 |
| 9 | 0,49 | 0,00 | 3,19 | 1,45 | 0,44 | 0,39 | 0,61 | 0,28 | 0,07 | 0,84 | 0,18 | 1,03 | 3,91 |
| 10 | 1,37 | 0,86 | 0,16 | 0,06 | 1,61 | 1,36 | 0,02 | 23,17 | 0,71 | 0,01 | 1,32 | 1,40 | 0,20 |
| 11 | 1,11 | 0,12 | 0,04 | 1,32 | 1,17 | 0,59 | 0,44 | 2,45 | 0,21 | 0,12 | 0,98 | 0,62 | 0,22 |
| 12 | 2,53 | 4,32 | 8,32 | 4,46 | 3,10 | 3,06 | 1,74 | 1,75 | 6,62 | 1,05 | 2,51 | 2,39 | 9,91 |
| 13 | 2,68 | 6,48 | 48,45 | 0,03 | 3,50 | 3,94 | 0,04 | 4,97 | 6,46 | 0,00 | 3,59 | 4,39 | 21,07 |
| 14 | 0,04 | 0,01 | 0,77 | 11,32 | 0,06 | 0,05 | 14,83 | 0,24 | 0,06 | 18,10 | 0,03 | 0,06 | 0,43 |
| 15 | 0,13 | 11,67 | 5,22 | 1,37 | 0,05 | 2,21 | 1,41 | 3,50 | 5,27 | 1,03 | 0,10 | 1,45 | 0,97 |
| Vacío | 0,25 | 1,01 | 0,56 | 0,01 | 0,23 | 0,44 | 0,00 | 1,13 | 0,61 | 3,89 | 0,40 | 0,34 | 0,21 |

De esta manera se determina a qué cobertura corresponde cada clase identificada mediante clustering, tomando el mayor porcentaje de aporte. Los resultados son resaltados en la Tabla XXVII con color gris. Debido a que en algunos casos más de un cluster, tiene un alto porcentaje de concordancia con el mismo grupo del mapa teórico, las clases se reducen. Sin embargo, se considera que ese resultado no es propicio de acuerdo con las clases obtenidas en las pruebas realizadas, que conducen a establecer el clustering como un buen método para la identificación de tipos de biomasa en el departamento. Por lo anterior, se decide equiparar los tipos 2, 5, 8, 9 y 11 con las coberturas marcadas con color naranja en la tabla. De esta manera, se evita la reducción de clases, se logra concordancia en los resultados y se obtienen 7 grupos en total, definidos por las coberturas 3, 4, 5, 7, 8, 10 y 13. Esto como resultado del análisis de porcentajes y el indicador de comparación visual con los mapas de referencia.

En seguida, mediante el uso de Matlab, se conforma la base de datos del mapa final, una matriz de 34’202.925 de filas por 3 columnas, donde la tercera columna define la clase final a la que pertenece el pixel, teniendo en cuenta la Tabla XXVIII. El detalle del mapa final caracterizado se muestra en la Figura 52 y los datos se almacenan en el servidor, en una tabla con el nombre de *km_mod_finalk8*.

TABLA XXVIII

CLASE ASIGNADA A CADA CLUSTER DESPUÉS DE LA IDENTIFICACIÓN DE CLASES

| Clase Final No. | Conformada por cluster No. | Color en el mapa de la Figura 52 | Tipo de cobertura | Total puntos por clase |
|---|---|---|---|---|
| 1 | 1, 6 y 12 | Verde fluorescente | Bosque | 16’419.875 |
| 2 | 4, 7 y 10 | Fucsia | Bosque de piedemonte amazónico | 2’116.512 |
| 3 | 11 y 13 | Azul claro | Cultivos mixtos | 5’444.742 |
| 4 | 5 | Naranja | Cultivos de palma | 4’057.274 |
| 5 | 2 y 9 | Morado | Cultivos anuales o transitorios | 2’828.637 |
| 6 | 3 | Verde oscuro | Pasto | 2’332.017 |
| 7 | 8 | Azul oscuro | Agua | 1’003.868 |
| | | | Total puntos | 34’202.925 |

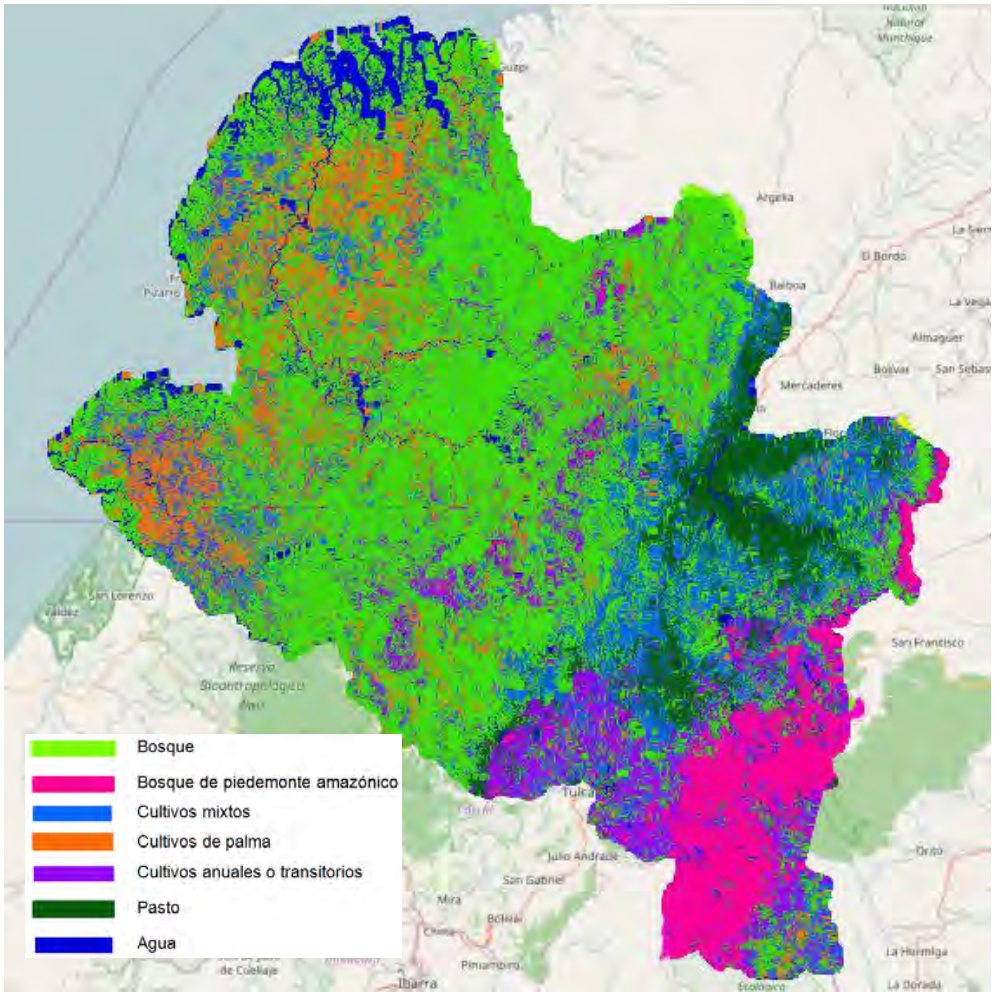



**Figura 52.** Mapa final del departamento de Nariño, que contiene 7 grupos obtenidos después de la identificación de clases.

A continuación se describen los tipos de cobertura encontrados:

- **Bosque:** este grupo es un ecosistema donde la vegetación predominante la constituyen árboles y arbustos. Comprende tipos de bosque como: bosque de colina, bosque de guandal, bosque de mangle, bosque húmero piedemonte pacífico, bosque ripario, bosque primario de sajal, bosque bajo, entre otros. Este grupo contribuye con el 48% del territorio departamental.

- **Bosque de piedemonte amazónico:** comprende un tipo específico de bosque que pertenece a la zona de piedemonte Andino – Amazónico, que es uno de los sitios con la mayor diversidad en los Andes de Colombia. Esta clase cubre el 6,19% del territorio del departamento.

- **Cultivos mixtos:** consiste en la plantación de diferentes variedades de plantas. Se puede encontrar tipos de cultivos como: cultivos mixtos con predominio de café, con predominio de caña, cultivos de clima cálido, entre otros. Este tipo de cobertura se ubica en el 15,92% de todo Nariño.

- **Cultivos de palma:** es un grupo específico de cultivos destinados a la siembra de palma, que ocupan el 11,86% del territorio departamental.

- **Cultivos anuales o transitorios:** son los cultivos con un ciclo de crecimiento menor a un año y que tienen la característica de destruir la planta al ser cosechada. Entre ellos se encuentran el cultivo de maíz amarillo, el maíz blanco, la papa, la arveja, cebolla rama y cebolla bulbo. Este grupo contribuye con el 8,27% del territorio departamental.

- **Pasto:** es un grupo en el que cabe cualquier planta que sirve para el sustento de los animales, especialmente la hierba que el ganado come en el mismo terreno donde se cría. Se pueden encontrar diferentes tipos de pasto como: pastos arbolados, pastos de clima cálido, pastos enmalezados o enrastrojados, pastos limpios, pastos naturales y rastrojos. Esta clase cubre el 6,82% del departamento de Nariño.

- **Agua:** este grupo se destinó a zonas como son los ríos, lagos, lagunas, ciénagas, estanques piscícolas y estuarios, que se expanden en el departamento con un porcentaje del 2,94%.

Finalmente, para optimizar el resultado respecto al mapa teórico se decide añadir una clase más, destinada a las zonas urbanizadas, que hace referencia a los centros poblados del departamento de Nariño. Mediante consultas de PostgreSQL, se identifica las coordenadas de los puntos que pertenecen a este tipo de cobertura a partir de la tabla *bdcoberturas* y el resultado se almacena en una nueva tabla con el nombre de *zonas_urbanizadas*. Posteriormente se altera la

tabla *km_mod_finalk8* modificando la clase de un total de 57.428 registros, mediante la asignación de una nueva clase. Por lo anterior, la clase número 7 de la Tabla XXVIII pasa a ser la clase 8 para dar paso al tipo de cobertura, zonas urbanizadas. Adicionalmente se efectúa el mismo procedimiento para insertar en el mapa final aquellos puntos que representan la clase de cobertura *Lagunas, lagos y ciénagas* de la Tabla XXIII, en el tipo de cobertura Agua. La diferencia con el proceso anterior radica en que las coordenadas de los puntos no se extraen de la tabla *bdcoberturas* sino de una tabla diferente que lleva el nombre de *gfc30m*, que contiene la clasificación original para cada punto asignada por el mapa de Corponariño. El resultado de la optimización del mapa final se observa en la Figura 53, donde se asigna el color negro a la cobertura Zonas Urbanizadas. Cabe aclarar que los pixeles modificados resaltan en la figura en mención debido al sistema de geo-referenciación que posee el mapa de Corponariño. Un detalle del municipio de Pasto se muestra en la Figura 54 para apreciar la resolución del mapa cuando se realizan acercamientos a zonas fácilmente identificables.

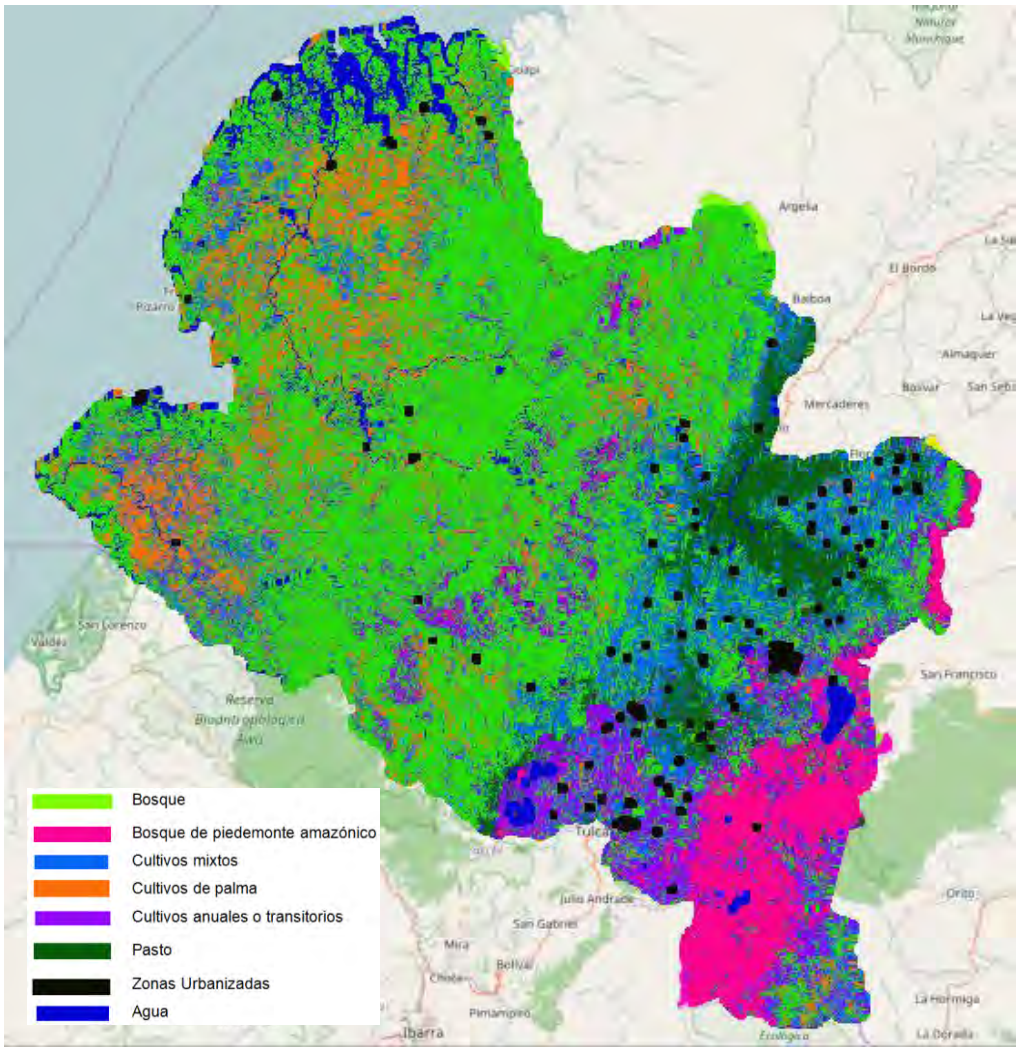



**Figura 53.** Mapa final del departamento de Nariño, que contiene 8 grupos obtenidos después de la optimización del resultado.

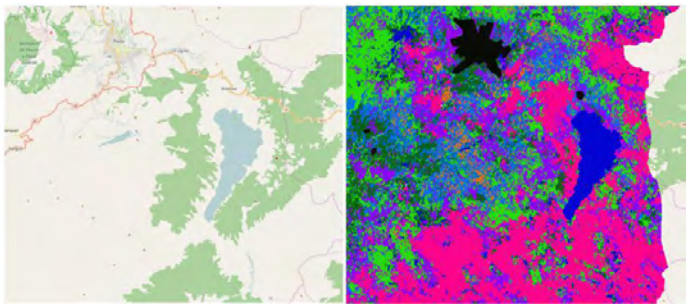


**Figura 54.** Ampliación de una zona del municipio de Pasto, del Mapa final del departamento de Nariño, que contiene pixeles modificados.

# 3. CONCLUSIONES

En este tipo de trabajos se emplean con frecuencia los valores de reflectancia de las bandas espectrales. De acuerdo a los resultados obtenidos se determina que es suficiente trabajar con estos datos, puesto que emplear características que se calculan a partir de las bandas espectrales, aumenta la redundancia en el conjunto de datos de entrada, además características como la altura y la temperatura no son las mejores para clasificar tipos de cobertura vegetal en una determinada zona.

Entre los resultados se encontró que el procesamiento de los datos no se debe aplicar cuando se emplean las características originales de la zona de estudio puesto que la normalización y centrado pueden afectar la interpretación de un pixel con respecto a los demás, dificultando la discriminación entre diferentes grupos, debido a la reducción de los intervalos de los valores de reflectancias. Sin embargo, en algunas pruebas efectuadas sobre el resultado de PCA resultó mejor efectuar el procesamiento en los datos, debido a que se está trabajando en un nuevo espacio de estados con otras características.

Los resultados demuestran que el algoritmo *K-Means* tiene una robustez comparable al algoritmo *EM*, aunque el primero se destaca aún mejor por sus índices de desempeño. Por su parte, aunque *Isodata* pudo trabajar con toda la base de datos del departamento de Nariño distinguiendo diferentes clases, hace una discriminación que no concuerda en su totalidad con la realidad, donde un tipo de vegetación se confunde con la cobertura Agua.

El resultado de los algoritmos de clustering depende en gran medida de factores como la evolución del estado fenológico de la vegetación en estudio, el clima presente cuando se capturaron las diferentes imágenes satelitales, entre otros aspectos; que afectan en gran medida la respuesta espectral, pudiendo ocasionar errores en valores de reflectancia.

Una de las complicaciones más relevantes de la presente investigación es la definición adecuada del número de clases, en primer lugar porque es un dato importante en algoritmos que requieren su definición a priori y en segundo lugar porque se posee escasa información acerca de la cantidad de grupos de cobertura, biomasa o vegetación que posee el departamento de Nariño. Esto se corrobora dado que en la inspección bibliográfica realizada, se encontró que en la región existe poca información y estudios relacionados con estas temáticas.

# 4. RECOMENDACIONES Y TRABAJO FUTURO

Para hacer una clasificación más profunda y exacta de los tipos de biomasa del departamento de Nariño, se recomienda utilizar imágenes satelitales de mayor resolución, que puedan aportar información más detallada y en lo posible aumentar la calidad de los datos de entrada a la ejecución del clustering.

Con el fin de explorar otros métodos de extrapolación de resultados, se recomienda que aspectos como determinar el número de clusters y seleccionar una zona de estudio, se realicen teniendo en cuenta toda la base de datos con la que se trabaja, en este caso todo el departamento de Nariño.

Se considera que es importante destacar como trabajo futuro, el análisis e implementación de otras técnicas para la identificación de clases, como por ejemplo análisis de respuestas espectrales. Con esto, se podría reconocer a qué coberturas vegetales pertenecen las 13 clases obtenidas en el mapa del departamento de Nariño como resultado del clustering, y de esta manera mejorar la clasificación de tipos de biomasa encontrada en este estudio.

Finalmente, se menciona que cabe la posibilidad de obtener una mejor clasificación si se realiza trabajo de campo y se extrae información acertada de los tipos de vegetación presentes en la región. Si no se encuentra esta información, como ocurrió en este trabajo, se sugiere abordar el problema de factores como nubosidad y estado fenológico desde las imágenes de laboratorio, donde probablemente se pueden controlar estas variables entre el sensor y el elemento sensado.

## BIBLIOGRAFÍA


[1] C. LOPEZ LOPEZ y M. V. SANCHEZ QUITIAN, «Diagnóstico de las centrales termoeléctricas en Colombia y evaluación de alternativas tecnológicas para el cumplimiento de la norma de emisión de fuentes fijas,» Universidad de la Salle, Bogotá, Colombia, 2007.

[2] H. ALTOMONTE, «Las energías renovables no convencionales en la matriz de generación eléctrica: tres estudios de caso,» *Comisión Económica para América Latina y el Caribe (CEPAL),* pp. 9-41, Febrero, 2017.

[3] A. M. CÁRCAMO y J. G. REJAS, «Análisis multitemporal mediante teledetección espacial y SIG del cambio de cobertura del suelo en el municipio de Danlí, El Paraíso, en los años 1987-2011,» *Revista Ciencias Espaciales,* vol. 8, nº 2, pp. 259-271, 2015.

[4] D. F. PEREZ, «Identificación de ecosistemas en la provincia de Napo - Ecuador mediante análisis digital de imágenes satelitales,» Universidad San Francisco de Quito, Colegio de Postgrados, Quito, Ecuador, Octubre, 2012.

[5] J. L. RODRÍGUEZ SOTELO, D. H. PELUFFO y D. CUESTA, «Unsupervised feature relevance analysis applied to improve ECG heartbeat clustering,» *Computer methods and programs in biomedicine,* vol. 108, nº 1, pp. 250-261, 2012.

[6] I. L. CASTILLEJO GONZÁLES, «Evaluación de métodos basados en píxeles y objetos para la clasificación de usos de suelo con imágenes de satélite quickbird, para el seguimiento de medidas agroambientales y la optimización del uso de herbicidas con agricultura de precisión,» Universidad de Córdoba, Córdoba, 2011.

[7] T. J. GARCÍA MORA y J. F. MAS, «Evaluación de imágenes del sensor MODIS para la cartografía de la cobertura del suelo en una región altamente diversa de México,» *Boletín de la Sociedad Geológica Mexicana,* vol. 63, nº 1, pp. 83-94, 2011.

[8] M. J. ÁLVAREZ, P. TRISTÁN, J. M. MASSA y R. WAINSCHENKER, «Clasificación automática de cubiertas terrestres en imágenes satelitales,» Universidad Nacional del Centro de la Provincia de Bs.As, Tandil, Argentina, 2011.

[9] J. A. ÁLVAREZ, P. E. VILLAGRA, E. M. CESCA, F. ROJAS y S. DELGADO,

«Estructura, distribución y estado de conservación de los bosques de Prosopis flexuosa del Bolsón de Fiambalá (Catamarca),» *Boletín de la Sociedad Argentina de Botánica,* vol. 50, nº 2, pp. 193-208, Junio, 2015.

[10] S. K. PAL y P. MITRA, «Multispectral image segmentation using Rough-Set-Initialized EM algorithm,» *IEEE Transactions on Geoscience and Remote Sensing,* 2002.

[11] A. SRIPAKAGOM y C. SIKAM, «Design and performance of a moderate temperature difference Stirling engne,» *Renew Energy,* 2011.

[12] L. A. TOSCANO MORALES y A. BARRIGA, «Análisis de los parámetros y selección de hornos para la combustión de biomasa,» *Tecnológica ESPOL,* pp. 1-10.

[13] Plan de desarrollo del departamento de Nariño;, «"La fuerza continua"-NARIÑO, territorio para querer, primera parte, diagnóstico del departamento de Nariño, dimension ambiental,» Nariño, Colombia, 2004-2007.

[14] Gobernación de Nariño;, «Plan participativo de desarrollo departamental de Nariño,» Nariño, Colombia, Mayo, 2016.

[15] A. E. ORDOÑEZ y J. P. SERNA, «Análisis superficial y multiespectral de imágenes Landsat 7 ETM+ y Landsat 8 OLI TIRS en el proyecto carbonífero de la luna entre los años 2001 y 2015,» Facultad de Ciencias e Ingeniería, Universidad de Manizales, Manizales, Colombia, 2015.

[16] S. ACOSTA, A. APONTE y D. TORRES, «Percepción remota o teledetección,» Universidad Santo Tomás, 22 Agosto 2016. [En línea]. Available: https://prezi.com/hj7d1y8t9jk8/percepcion-remota-o/. [Último acceso: 2016].

[17] S. F. A. C. W. USGS, «What are the band designations for the Landsat satellites,» [En línea]. Available: https://landsat.usgs.gov/what-are-band-designations-landsat-satellites. [Último acceso: Enero 2017].

[18] A. ARIZA, «Descripción y corrección de productos Landsat 8 LDCM,» Centro de Investigación y Desarrollo - CIAF, Bogotá, Colombia, 2013.

[19] C. GONZAGA AGUILAR, «Aplicación de índices de vegetación derivados de imágenes satelitales Landsat 7 ETM+ y ASTER para la caracterización de la cobertura vegetal en la zona centro de la provincia de Loja, Ecuador,» 8 Abril 2014. [En línea]. Available: http://sedici.unlp.edu.ar/handle/10915/34487.

[Último acceso: 2017].

[20] E. R. CERVANTES GÓMEZ, «Clasificación de imágenes satelitales mediante el uso de memorias asociativas,» Departamento de Telecomunicaciones, Instituto Politécnico Nacional, México, 2014.

[21] S. f. a. c. w. USGS, «EarthExplorer,» [En línea]. Available: https://earthexplorer.usgs.gov/. [Último acceso: 2016].

[22] J. A. WATANABE CABRERA, «Manejo de ENVI V4.5,» Monografías, [En línea]. Available: http://www.monografias.com/trabajos82/manejo-envi/manejo-envi2.shtml. [Último acceso: 2016].

[23] M. CÁRDENAS MONTES, «Clustering: clasificación no supervisada. Gráficas estadística y mineria de datos con Python,» 22-26 Abril 2013. [En línea]. Available: http://wwwae.ciemat.es/~cardenas/curso_MD/clustering.pdf. [Último acceso: 2017].

[24] M. CAMPOS G, «Aplicación de técnicas de clustering para la mejora del aprendizaje,» Universidad Carlos III, Leganés - Madrid, España, 2009.

[25] M. GARRE, J. CUADRADO y A. SICILIA, «Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software,» Departamento de las Ciencias de la Computación, Universidad de Alcalá, Barcelona, España.

[26] J. A. CAMARENA IBARROLA, «El algoritmo E-M,» 2011. [En línea]. Available: http://dep.fie.umich.mx/~camarena/Expectation-Maximization.pdf. [Último acceso: 2017].

[27] F. J. CORTIJO BON, «Técnicas no supervisadas: métodos de agrupamiento,» Noviembre, 2001.

[28] F. BERZAL, «Clustering basado en densidad,» Departamento de Ciencias de la Comunicación e I.A., Universidad de Granada.

[29] M. I. NAKAMA, «Un estudio basado en la técnica de Mean - Shift para agrupamiento y seguimiento en video,» Departamento de Computación, FCEyN, Universidad de Buenos Aires, Buenos Aires, Argentina, 13 Agosto, 2011.

[30] MATLAB, «MathWorks,» [En línea]. Available: https://es.mathworks.com/. [Último acceso: 2016 - 2017].

[31] UNAMDSP, «Implementación algoritmo Mean-Shift en Matlab,» Procesamiento de señales, visión por computadora, 25 Junio 2009. [En línea]. Available: https://fierdetregauche.wordpress.com/2009/06/25/implementacion-algoritmo-mean-shift-en-matlab/. [Último acceso: 2017].

[32] J. AVENDAÑO PÉREZ, J. A. PARRA PLAZAS y J. F. BAYONA, «Segmentación y clasificación de imágenes SAR en zonas de inundación en Colombia, una herramienta computacional para prenvención de desastres,» *Revista Facultades de Ingeniería, Universidad Antonio Nariño,* pp. 24-38, 2013.

[33] T. CELIK, «Unsupervised change detection in satellite images using principal component analysis and k-means clustering,» *IEEE Geoscience and Remote Sensing Letters,* vol. 6, nº 4, pp. 772-776, Octubre, 2009.

[34] P. GONZÁLES MARTÍN, A. DÍAZ DE PASCUAL, E. TORRES LEZAMA y E. GARNICA OLMOS, «Una aplicación del análisis de componentes principales en el área educativa,» *Economía,* nº 9, pp. 56-72.

[35] U. d. N. GIIEE, «Análisis de regresión para el cálculo de biomasa en el departamento de Nariño (Colombia) utilizando imágenes satelitales Landsat,» Universidad de Nariño, Pasto, Colombia, 29 de Junio, 2016.

[36] O. CABRERA, B. CHAMPUTIZ, A. CALDERÓN y A. PANTOJA, «Landsat and MODIS Satellite Image Processing for Solar Irradiance Estimation in the Department of Narino-Colombia,» de *XXI Symposium on Signal Processing, Images and Artificial Vision*, Bucaramanga, Colombia, 2016.

[37] O. E. CABRERA ROSERO, «Análisis de oportunidades energéticas con fuentes alternativas en el departamento de Nariño - Alternar. Estrategia biomasa,» Universidad de Nariño, Pasto, Colombia, 2015.

[38] C. CONTELL, J. C. VAYÁ, S. LANJERI y F. CAMACHO, «Optimización del algoritmo ACCA para la detección en imágenes Landsat de nubes, sombras y agua,» *Teledetección: agua y desarrollo sostenible. XIII congreso de la asociación española de teledetección,* pp. 453-456, 23-26 de Septiembre, 2009.

[39] S. F. A. C. W. USGS, «What are the best spectral bands to use for my study?,» [En línea]. Available: https://landsat.usgs.gov/what-are-best-spectral-bands-use-my-study. [Último acceso: Enero 2017].

[40] R. C. ROMERO ZALIZ, «Reconocimiento de perfiles de regulación genética

mediante algoritmos evolutivos multiobjetivo,» Universidad de Granada, Granada, 2005.

[41] A. A. NAVARRO ESPINOSA, «Planificación de redes de distribución: aproximación vía clustering, diagramas de voronoi y búsqueda tabú,» Pontificia Universidad Católica de Chile, Santiago de Chile, Chile, 2007.

[42] G. LORCA, J. ARZOLA y O. PEREIRA, «Segmentación de imágenes médicas digitales mediante técnicas de clustering,» *Aporte Santiaguino ,* pp. 108-116, 2010.

[43] W. HASPERUÉ, «Extracción de conocimiento en grandes bases de datos utilizando estrategias adaptativas,» Universidad Nacional de la Plata, La Plata, Buenos Aires, Argentina, Marzo, 2012.

## ANEXO 1

Este anexo contiene el artículo seleccionado para sustentación en modalidad ponente en el "Congreso Internacional Multimedia 2017" desarrollado en Cajicá – Cundinamarca, durante los días 28 y 29 de septiembre de 2017. Adicionalmente se presenta el certificado de ponencia.

**ARTÍCULO:** Análisis de imágenes satelitales para clasificación de biomasa en el departamento de Nariño.

*Congreso Internacional Multimedia 2017*

# *Análisis de imágenes satelitales para clasificación de biomasa en el departamento de Nariño*

Multimedia International Conference 2017 Analysis of satellite images for biomass classification in the department of Nariño

Alison Bastidas, Andrea Bravo, Andrés Pantoja

**Resumen**. *En este trabajo se evalúan distintas técnicas de clustering para la determinación de tipos de cobertura vegetal en el suroccidente colombiano a partir del procesamiento de la reflectancia de diferentes bandas de imágenes satelitales libres. Para esto, se comparan los algoritmos K-Means, EM e Isodata mediante índices de desempeño en una zona característica y se expanden los mejores resultados a un mapa de clasificación a todo el territorio teniendo en cuenta la información de un mapa teórico de la región.*

*Palabras claves: Clasificación de Biomasa, Clustering, Imágenes Satelitales*.

**Abstract.** *In this work, we evaluate distinct clustering methods to determine the types of natural cover in Southwest Colombia by processing the reflectance of different bands of free satellite imagery database. For this purpose, the K-Means, EM and Isodata algorithms are compared with performance indexes in a characteristic zone and then, the best results are extended to the whole territory taking into account a theoretical map of the region***.**

*Key words: Biomass Classification, Clustering, Satellite Images.*

## *1. Introducción.*

La caracterización de los recursos renovables para la generación de energía eléctrica constituye el principal aporte para la proposición de alternativas en la creciente necesidad de diversificación de la matriz energética de los países [1]. En cuanto a la biomasa vegetal, para estimar su potencial es necesario determinar los diferentes tipos de cobertura, aunque usualmente la determinación de estas clases implica el procesamiento de imágenes sobre áreas pequeñas, se presenta en [2].

En contraste, en este trabajo se analiza una zona de gran extensión territorial que implica una cantidad elevada de información proveniente de varias bandas de imágenes satelitales, que requiere de técnicas computacionales eficientes para su procesamiento.

Es importante destacar que la respuesta espectral útil de las fotografías para la determinación de los tipos de biomasa (i.e., valores de reflectancia), puede ser afectada por la evolución del estado fenológico de la vegetación, el clima presente cuando se capturan las fotos y la evolución temporal de los terrenos. Por lo tanto, la caracterización de las coberturas es un problema de investigación abierto y altamente dependiente de los lugares de aplicación [3].

Diferentes métodos de agrupación de datos en clases congruentes (clustering) se usan en el procesamiento de características de imágenes.

En particular, los métodos con análisis no supervisado son relevantes en estas aplicaciones puesto que no requieren un etiquetado previo en el conjunto de individuos y el proceso de clasificación es más flexible [4]. Entre los algoritmos de clustering más utilizados en estudios similares se encuentran K-Means e Isodata [5], [6], en donde se clasifica el uso del suelo en diferentes regiones utilizando las propiedades de imágenes QuickBird y MODIS, respectivamente. Por su parte, los autores en [7] proponen la clasificación automática de cubiertas terrestres usando Region Growing y K-Means, mostrando la adaptabilidad de estos métodos a los requerimientos establecidos. Usando también este último método para obtener los mejores resultados, en [8] se identifican unidades boscosas utilizando imágenes Landsat.

Teniendo en cuenta que el resultado de los diferentes algoritmos de clustering depende en gran medida de la calidad de los datos de entrada, en este artículo se propone una metodología para analizar imágenes en diferentes bandas de Landsat 8 con los algoritmos K-Means, EM (expectation–maximization) e Isodata. Los métodos se aplican inicialmente a una base de datos pre-procesada de reflectancias en un área representativa del departamento, determinando el método con mejor desempeño de acuerdo a los índices de cohesión y separación de las clases identificadas. Con estos resultados se aplican los algoritmos con los mejores parámetros a la base de datos de todo el departamento con el fin de establecer un mapa de alta resolución de clasificación de tipos de cobertura natural. El resultado final se contrasta con un mapa oficial de coberturas vegetales para determinar los tipos finales de biomasa, que pueden interpretarse para el estudio de su aprovechamiento energético en la región.

## *2. Metodología de procesamiento.*

### 2.1. *Pre-procesamiento de datos*

El conjunto de datos proviene de las 7 primeras bandas espectrales de imágenes libres Landsat 8, que tienen una resolución de 30 metros por pixel en tomas realizadas cada 16 días. Para cubrir toda la extensión del departamento de Nariño son necesarias 5 escenas que requieren de un recorte al área precisa del mapa y la aplicación de un filtrado basado en continuidad espacial y temporal para la eliminación de nubosidad y pixeles no válidos [9]. El filtro usa el algoritmo denominado ACCA (automated cloud cover assessment) [10], para cuya aplicación se obtuvo una serie de imágenes en el periodo 2015-2016. Los datos resultantes se organizan en una base de datos PostgreSQL que facilita la consulta de las coordenadas y reflectancias de las 7 bandas para cada uno de los 34’202.925 pixeles necesarios para cubrir el departamento de Nariño de 33.268 $km^2$ [11].

Para la aplicación de los algoritmos de clustering, se realiza normalización y centrado de las características (reflectancias) de cada pixel. Las operaciones están descritas por

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \text{y} \quad X_c = X - \bar{X},$$

donde $X'$ y $X_c$ representan las características normalizada y centrada, respectivamente; $X$ el valor de la característica original; $X_{min}$ y $X_{max}$ son los valores mínimo y máximo del conjunto de valores de la característica para todos los pixeles, y $\bar{X}$ es el valor promedio del mismo conjunto.

Finalmente, se aplica el análisis de componentes principales (PCA), proceso estadístico ampliamente utilizado para reducción de dimensión de forma no supervisada en imágenes [12]. En este caso, a partir del gráfico de autovalores obtenido del PCA, se determina que el número de componentes principales para representar adecuadamente los datos está entre 3 y 5. Así, el uso de las nuevas características permite disminuir la elevada carga computacional que implica la aplicación de clustering en la base de datos original.

### *2.2. Clustering y Validación.*

Teniendo en cuenta los estudios presentados en [5], [6], [7], [8] y [12], se seleccionan los algoritmos de K-Means e Isodata que están basados en minimizar distancias entre los individuos a un número determinado de centroides que se mueven alrededor del espacio de trabajo. A diferencia de K-Means, Isodada puede cambiar la cantidad de clusters mediante procedimientos de fusión o división para minimizar las distancias de objetos lejanos o de posibles nuevas clases. Por su parte, el algoritmo de EM supone un modelo probabilístico que define la clasificación de los individuos que se va mejorando en dos pasos denominados *Expectation* de cada punto y *Maximization* de la veracidad del nuevo modelo probabilístico definido por la medida de expectativa [13].

Dada la gran cantidad de datos del mapa completo, para la sintonización de los algoritmos de clustering, se realizaron pruebas en diferentes zonas del departamento. Con base en los resultados, la calidad de los datos (nubosidad) y la extensión territorial, se define el municipio de Tumaco como una región representativa de Nariño, que contiene 3'971.007 pixeles (un 11,6% del total de datos del departamento). En esta región de prueba se aplican los tres métodos de clustering, teniendo en cuenta la base de datos original (normalizada y centrada) y luego con la aplicación del PCA. Estas dos pruebas se realizan con el fin de comparar los resultados y validar el uso de las componentes principales como herramienta de reducción de dimensión y de discriminante adicional de clases.

El proceso de PCA, así como los métodos de clustering se adaptan de las funciones de Matlab realizadas en códigos abiertos propuestos en [14], analizando las características de los datos de entrada, salida y las clases esperadas.

Para comparar la eficiencia de los algoritmos implementados, se calculan los índices de desempeño de cohesión y separación, con igual importancia. Estas métricas de validación interna cuantifican la dispersión de los pixeles a nivel inter-cluster e intra-cluster. La cohesión analiza que cada miembro del grupo debe ser lo más cercano posible a los otros miembros del mismo clúster, mientras que la separación se basa en que cada cluster debe estar lo más distante posible del resto de los grupos [13].

Para medir el rendimiento ($\%Ren$) de cada algoritmo, se propone el cálculo de un porcentaje definido por

$$\%Ren = \left(\frac{0.5C_{best}}{C_X} + \frac{0.5S_X}{S_{best}}\right)100\%,$$

donde, $C_{best}$ y $S_{best}$ representan los mejores valores de cohesión y separación de todas las pruebas que se comparan entre sí, y $C_X$ y $S_X$ son los valores de cohesión y separación de la prueba en cuestión. El proceso de sintonización de los parámetros se hace mediante ensayo y error, iniciando con valores por defecto, y modificando un parámetro a la vez hasta fijar un valor que maximice el rendimiento, para continuar con los parámetros restantes.

Una vez seleccionada la prueba con el mejor desempeño, se construye un mapa con las clases definidas y ayuda del visor de sistemas de información geográfica QGIS. Además, se define un indicador de comparación que detalla la correspondencia de la clase "agua", claramente identificable en mapas genéricos. Para valorar este indicador en la región de prueba se determina la demarcación de los ríos de los mapas resultantes y se compara con el mapa de referencia de la herramienta *Street Map* de QGIS.

Finalmente, con el resultado de mejor desempeño se realiza la expansión a todo departamento de Nariño con base en el número de clases definidas en el mapa oficial de 2015 sobre el uso del suelo en la región. Este mapa teórico adaptado para 15 clases de cobertura vegetal se presenta en la Figura 1, cuya base de datos para comparar se procesa también en PostgreSQL y Matlab.

Para la identificación de las clases obtenidas, se contabiliza la cantidad de pixeles que pertenecen a cada tipo de cobertura. Así, se obtiene el porcentaje de los diferentes tipos de biomasa y se compara cluster por cluster hasta obtener el porcentaje más cercano de una clase en el mapa teórico, que correspondería a la etiqueta de cada grupo encontrado (e.g., bosques, pastos, cultivos y agua, entre otros).

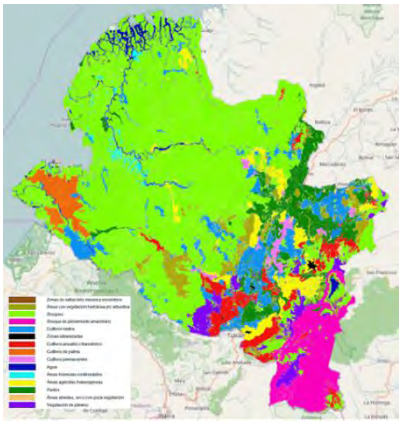


*Figura 1: Mapa teórico con 15 clases obtenido a partir del mapa oficial de coberturas vegetales en 2015 de la Corporación Autónoma Regional – Corponariño.*

## 3. *Resultados y análisis.*

Para las tablas que se presentan en esta sección, la columna de pre-procesamiento (Proces) indica si se utilizó la normalización (N) o el centrado (Ce). El número de clusters final se presenta en la columna K, el número de iteraciones realizadas está en Iter y los valores de cohesión y separación se muestran en C y S. Para PCA, el número de componentes principales utilizadas se simboliza por Co. Otras

columnas, hacen referencia a parámetros específicos de los algoritmos.

### *3.1. K-Means.*

Las pruebas de la aplicación de K-Means se muestran en la Tabla 1. Se nota que el proceso de normalización no es adecuado en este algoritmo a pesar de incrementar el número de iteraciones (prueba 1) y que el centrado no ayuda a mejorar el rendimiento (prueba 4).

Los resultados del análisis del grupo “agua” en todas las pruebas (Figura 2) son similares, mostrando la demarcación acertada de los ríos en Tumaco. Con base en el mejor porcentaje de rendimiento y el indicador de comparación con el mapa de referencia, se elige la prueba 2 como la mejor, mientras la prueba 3 se descarta ya que requiere que los datos se centren, lo que implica un gasto computacional adicional. De esta manera, se determina que para esta base de datos el algoritmo K-Means no requiere normalización ni centrado.

| Prueba No. | Proces | | K | Iter | C | S | % Ren |
|---|---|---|---|---|---|---|---|
| | N | Ce | | | | | |
| 1 | Si | No | 5 | 1000 | 196.713,4 | 130.631,1 | 55,4 |
| 2 | No | No | 5 | 500 | 21.047,6 | 90.288,9 | 84,6 |
| 3 | No | Si | 5 | 500 | 21.047,6 | 90.288,9 | 84,6 |
| 4 | Si | Si | 5 | 500 | 32.236,6 | 128.693,7 | 81,9 |

*Tabla 1: Pruebas más importantes de K-Means aplicadas a la base de datos original del municipio de Tumaco.*

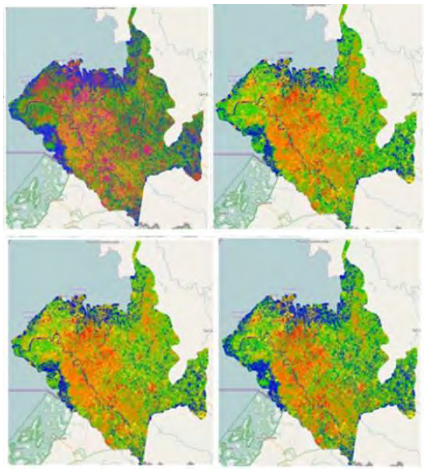


*Figura 2: Resultado gráfico de la aplicación de K-Means, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas expuestas en la Tabla 1.*

Los resultados para la base de datos con PCA se presentan en la Tabla 2 y la Figura 3. Es de resaltar que aunque la prueba 1 tiene un porcentaje de rendimiento de 100, la agrupación de los ríos no está bien definida. Al variar el número de componentes principales, se llega a la conclusión que el mejor resultado se tiene en la prueba 4, corroborado con la mejora en la demarcación de los ríos en el mapa de referencia.

Cabe resaltar que entre las pruebas realizadas para este algoritmo, se encontró que tiene un mejor desempeño utilizando como medida de distancia la euclidiana.

| Prueba No. | Co | Proces | | Iter | C | S | % Ren |
|---|---|---|---|---|---|---|---|
| | | N | Ce | | | | |
| 1 | 7 | No | No | 1000 | 32.236,6 | 128.693,7 | 100,0 |
| 2 | 7 | Si | No | 1000 | 32.379,7 | 37.029,6 | 64,2 |
| 3 | 7 | Si | Si | 1000 | 32.306,7 | 37.102,6 | 64,3 |
| 4 | 3 | Si | Si | 1000 | 18.728,6 | 36.940,0 | 99,9 |

*Tabla 2: Pruebas más importantes de K-Means aplicadas al data set como resultado de aplicar PCA a la base de datos de Tumaco*.

### *3.2. EM.*

En la Tabla 3 se presentan las pruebas realizadas con el algoritmo EM y los mapas resultantes en la Figura 4. Los resultados de este algoritmo producen 5 clusters cuyo desempeño mejora si el número de iteraciones está alrededor

de 30, mientras que el mejor valor de perfeccionamiento (Tol) es del orden de $10^{-6}$. Este parámetro, propio de este algoritmo, indica la mínima mejora de la función objetivo entre dos iteraciones consecutivas.

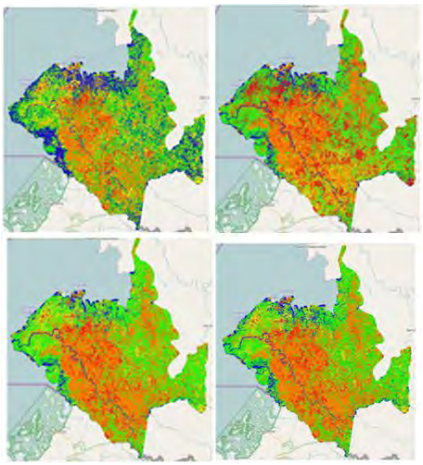


*Figura 3: Resultado gráfico de la aplicación de K-Means sobre el resultado de PCA, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas expuestas en la Tabla 2.*

Nuevamente se observa que no es necesario realizar un pre-procesamiento de los datos como se nota en las pruebas 2, 3 y 4. En la Figura 4, se aprecia que el centrado de los datos influye negativamente, mientas que el normalizado no logra superar los resultados en desempeño y de agrupación de agua de la prueba número 1. Finalmente, se evidencia que efectuar un procesamiento completo con un número de iteraciones pequeño (prueba 4), produce un resultado inadecuado. En conclusión, la prueba 1 presenta los mejores valores de cohesión y separación y el indicador de comparación con el mapa de referencia permite detallar de mejor forma los ríos de la zona.

Para el caso de PCA, se realizaron las pruebas respectivas con la configuración óptima de EM para 5 clases. Los resultados se muestran en la Tabla 4 y Figura 5. El mejor resultado se presenta en la prueba número 4, con 4 características principales.

| Prueba No. | Proces | | Iter | K | Tol | C | S | % Ren |
|---|---|---|---|---|---|---|---|---|
| | N | Ce | | | | | | |
| 1 | No | No | 30 | 5 | $1e^{-6}$ | 38.454 | 68.097 | 100 |
| 2 | No | Si | 40 | 5 | $1e^{-6}$ | 43.500 | 63.266 | 91 |
| 3 | Si | No | 10 | 5 | $1e^{-6}$ | 49.260 | 2.355 | 41 |
| 4 | Si | Si | 5 | 5 | $1e^{-6}$ | 67.964 | 4.568 | 32 |

*Tabla 3: Pruebas más importantes de EM aplicadas a la base de datos original del municipio de Tumaco.*

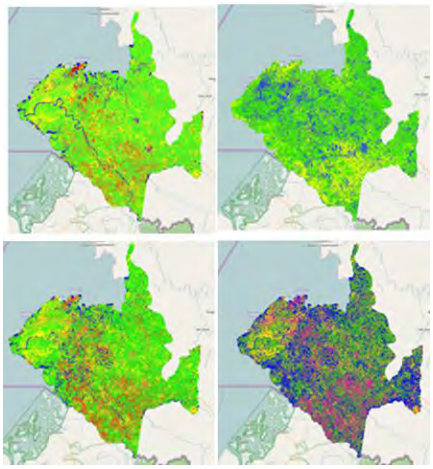


*Figura 4: Resultado gráfico de la aplicación de EM, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas expuestas en la tabla 3.*

### *3.3. Isodata.*

Las pruebas de la aplicación de Isodata se presentan en la Tabla 5. La columna A indica la cantidad de grupos que se obtienen en la clasificación final, diferente de K que define el número máximo de grupos que puede arrojar el algoritmo.

| Prueba No. | Co | Proces | | C | S | % Ren |
|---|---|---|---|---|---|---|
| | | N | Ce | | | |
| 1 | 7 | No | Si | 63.918,68 | 88.196,35 | 89 |
| 2 | 7 | Si | No | 49.915,54 | 15.830,87 | 59 |
| 3 | 7 | Si | Si | 52.934,74 | 14.585,04 | 55 |
| 4 | 4 | No | Si | 56.250,02 | 94.214,47 | 100 |

*Tabla 4: Pruebas más importantes de EM aplicadas al data set como resultado de aplicar PCA a la base de datos de Tumaco.*

Este método tiene varios umbrales para el agrupamiento como ON que limita el número de elementos para la eliminación de un grupo, OC y OS que refieren a la distancia entre centros y la desviación estándar típica para la unión o

división de clusters, y L hace referencia al máximo número de agrupamientos que pueden mezclarse en una sola iteración. En este caso, teniendo en cuenta que se trabaja con una base de 5 grupos, se decide utilizar este número en la mayoría de las pruebas. Un valor muy pequeño para OC reduce el porcentaje de rendimiento, pero un valor muy elevado como OC=2 perjudica totalmente la clasificación, puesto que la distancia entre centros debe ser adecuada para considerar pixeles dentro de un mismo cluster. El resultado de las pruebas indica que el valor de OS debe ser también pequeño, debido a que si la desviación estándar entre los individuos de un grupo supera este umbral, probablemente los pixeles hacen parte de clusters diferentes.

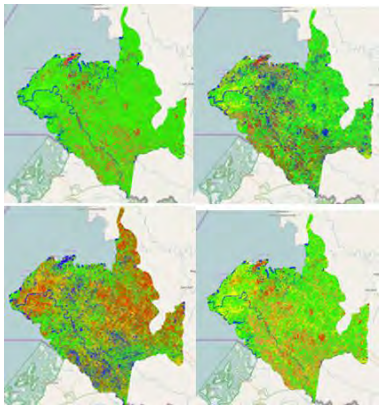


*Figura 5: Resultado gráfico de la aplicación de EM sobre el resultado de PCA, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas expuestas en la Tabla 4.*

Con estos resultados, la configuración más adecuada de los parámetros de Isodata se presenta en las pruebas 2 y 3. En este caso, la prueba número 4 presenta el mayor porcentaje de rendimiento, pero el indicador de comparación con el mapa de referencia no es el más adecuado, como se muestra en la Figura 6. La prueba 3 tiene el mejor índice de desempeño, pero arroja un total de 7 grupos. Esta información es relevante teniendo en cuenta que en Isodata no se define el número de clases a priori, lo que puede dar un indicio del número de clases óptimo presente en la zona. Sin embargo, como en los algoritmos anteriores se escogieron 5 clases, se establece la prueba 2 como la configuración adecuada.

| Prueb No. | Proces | | ON | OC | OS | K | L | I | A | % Ren |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | Ce | | | | | | | | |
| 1 | Si | No | 10 | 0,05 | 0,07 | 5 | 2 | 20 | 3 | 55 |
| 2 | No | No | $10e^5$ | 0,0005 | 0,005 | 9 | 2 | 20 | 5 | 54 |
| 3 | No | No | $10e^5$ | 0,0005 | 0,005 | 13 | 2 | 20 | 7 | 60 |
| 4 | No | No | 100 | 0,0001 | 0,005 | 13 | 1 | 30 | 7 | 67 |

*Tabla 5: Pruebas más importantes de Isodata aplicadas a la base de datos original del municipio de Tumaco.*

Los resultados de la aplicación de Isodata, mediante el uso de PCA se presentan en la Tabla 6. El porcentaje de rendimiento indica que la prueba número 1 es el mejor resultado. Sin embargo, en la Figura 7 se observa que la demarcación de los ríos no es la más propicia en comparación con el mapa teórico. Entonces se selecciona la prueba 2 que presenta la mejor visualización del cluster “agua”.

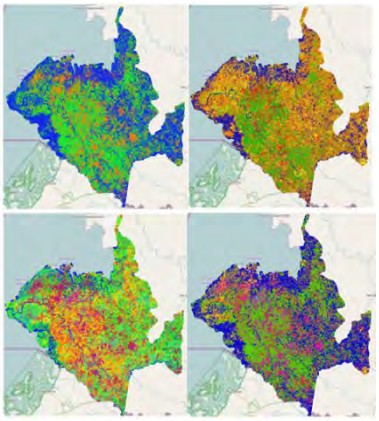


*Figura 6: Resultado gráfico de la aplicación de Isodata, representado en QGIS, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas expuestas en la Tabla 5.*

En cuanto al número de componentes principales, el porcentaje de rendimiento es más alto cuando se toman solo 3 características. A pesar de esto, la visualización más detallada se obtiene con 4 características, correspondiente a la prueba número 4 de la Tabla 6.

### 3.4. *Expansión del modelo al departamento de Nariño.*

Con base en las mejores configuraciones obtenidas en la región de estudio para cada algoritmo, se aplican los métodos a todo el departamento con la base de datos original (34’202.925 pixeles por 9 atributos, incluyendo

longitud, latitud y las 7 reflectancias), y para los mismos pixeles con las características obtenidas para el mejor caso de aplicación de PCA. Se mantiene la configuración de parámetros óptima para cada uno de algoritmos, excepto el número de clases. A diferencia de la región de prueba, y con base en el mapa teórico, el departamento tendría 13 clases definidas que representan la mayor parte del territorio, debido a que de las 15 clases originales del mapa de la Figura 1, las zonas urbanizadas y áreas abiertas, únicamente representan el 0,17% del total de los datos.

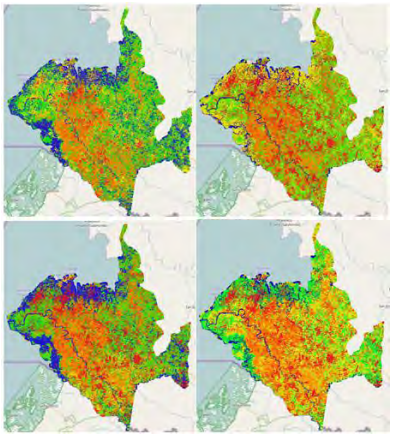


*Figura 7: Resultado gráfico de la aplicación de Isodata sobre el resultado de PCA, en orden, de arriba hacia abajo y de izquierda a derecha, las pruebas expuestas en la Tabla 6.*

Los resultados de las expansiones de K-Means, EM e Isodata se muestran en la Figura 8 y la Tabla 7 expone el porcentaje de rendimiento de cada expansión.

A pesar de la gran cantidad de datos y aunque la estructura de ejecución de los algoritmos K-Means y EM está basada en un concepto diferente, la expansión de estos métodos presenta resultados similares tanto en su visualización como en índices de desempeño. En general, K-Means tiene un comportamiento consistente a lo largo de las pruebas dado que la distancia euclidiana considera todas las características para medir la separación inter-cluster e intra-cluster.

La normalización de los datos en este tipo de técnicas, puede afectar la interpretación de un pixel con respecto a los demás, dificultando la discriminación entre diferentes grupos, debido a la reducción de los intervalos de los valores de reflectancias.

El efecto de PCA en K-Means es significativo en cuanto a la mejora de los índices de cohesión y separación, que se ve reflejado en el porcentaje de rendimiento final. Esto se observa en los mapas, donde se realiza una mayor demarcación entre los diferentes clusters (especialmente en la región sur), a pesar de no percibirse cambios sustanciales con una comparación visual. Por su parte, en el algoritmo EM la influencia de la aplicación de PCA es más notoria, especialmente en la demarcación de los clusters. El método EM aplicado sin PCA pierde la correcta definición del grupo “agua” (Figura 8), mientras que con PCA mejora la definición de los ríos. Este indicador de comparación muestra un funcionamiento adecuado debido a la capacidad del algoritmo para conformar grupos con diferentes formas y tamaños, a la vez que controla la varianza y covarianza de los datos. Sin embargo, la métrica de rendimiento de EM (%Ren) es significativamente menor al obtenido con K-Means.

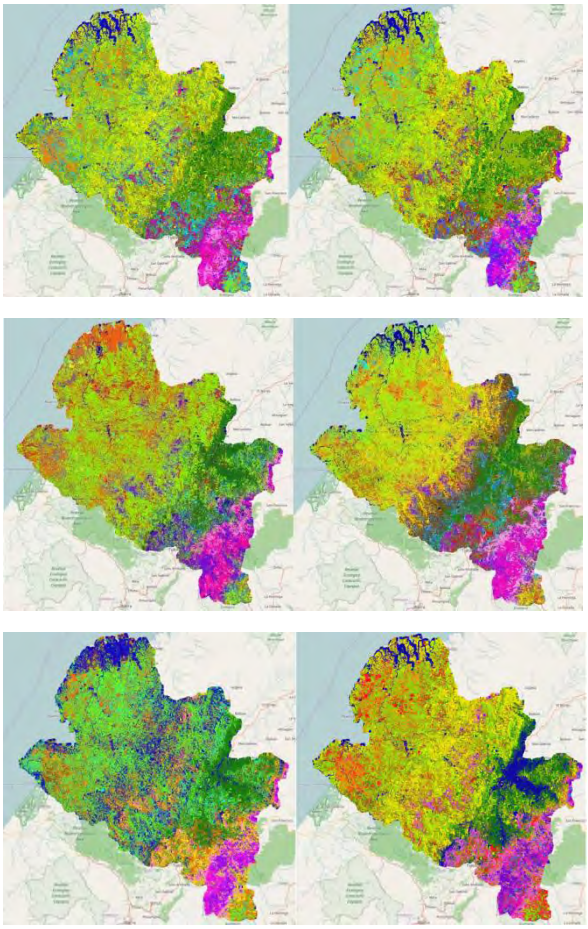


*Figura 8: Mapa del departamento de Nariño, resultado de la aplicación de K-Means (fila superior), EM (fila central) e Isodata (fila inferior), utilizando la base de datos original (columna*

*izquierda) y con los datos obtenidos con PCA (columna derecha).*

La expansión de Isodata identifica la mayoría de los clusters de una forma semejante a los resultados obtenidos con K-Means y EM. Es de destacar que sin aplicar PCA, el resultado disminuye notablemente su desempeño en la identificación del grupo "agua". Al emplear PCA, aunque mejora la demarcación de los ríos, disminuye considerablemente el porcentaje de rendimiento. En particular, se observa una mezcla entre coberturas dado que este método establece que algunos de los pixeles de cierto cluster no están lo suficientemente distantes de otros grupos para considerarse coberturas independientes al no superar los umbrales definidos y pueden ser clasificados erróneamente.

A diferencia de los otros dos algoritmos, la aplicación de Isodata con PCA produce disminución en el rendimiento ya que las características de la base de datos original tienen mayor información que puede ser usada por el algoritmo para definir grupos más definidos y compactos.

| Exp No. | Algoritmo | Datos de entrada | C | S | % Ren |
|---|---|---|---|---|---|
| 1 | K-Means | Base original | 1'799.905 | 3'553.792 | 54 |
| 2 | | PCA | 60.514 | 513.385 | 64 |
| 3 | EM | Base original | 964.679 | 2'809.107 | 42 |
| 4 | | PCA | 323.566 | 1'885.674 | 59 |
| 5 | Isodata | Base original | 148.141 | 2'170.612 | 81 |
| 6 | | PCA | 87.021 | 486.877 | 48 |

*Tabla 7: Desempeño de los algoritmos de clustering en las expansiones a todo el departamento de Nariño.*

Al comparar los resultados de rendimiento, el mejor desempeño se obtiene en las expansiones de Isodata con los datos originales y K-Means con PCA. Sin embargo, el indicador de comparación con el mapa de referencia muestra que la expansión de Isodata no tiene una definición adecuada de los clusters, por lo explicado anteriormente. De esta manera, se concluye que el algoritmo K-Means presenta los mejores resultados independiente de los datos de entrada propuestos. Además, se muestra que es un algoritmo que se puede adaptar fácilmente a diferentes condiciones de los datos, ya sea en su forma original o con pre-procesamiento.

Por otra parte, para comparar la carga computacional de los algoritmos se tiene en cuenta que la aplicación del clustering se realiza en un computador portátil con un procesador Intel Core i5 de 2.2GHz, 8GB de RAM y 1TB de disco duro. El tiempo de ejecución de los métodos es similar ya que para la expansión de la base de datos original, K-Means y EM tardaron 51 y 46 horas, respectivamente, y la mitad del tiempo para los datos con PCA. Para el caso de Isodata, el rendimiento computacional fue superior, dado que los procesos tardaron 23 y 10 horas para la base original y los datos con PCA, respectivamente.

De esta manera, se elige el mapa generado por K-Means con PCA como la expansión con los mejores resultados comparativos y con la que se genera el mapa final de clasificación de tipos de cobertura del departamento de Nariño.

| Prueba No. | Co | Proces | | C | S | % Ren |
|---|---|---|---|---|---|---|
| | | N | Ce | | | |
| 1 | 7 | No | No | 32.242 | 128.688 | 100 |
| 2 | 7 | Si | No | 32.470 | 36.939 | 64 |
| 3 | 7 | Si | Si | 32.597 | 37.030 | 64 |
| 4 | 4 | Si | No | 22.516 | 36.845 | 92 |

*Tabla 6: Pruebas más importantes de Isodata realizadas al data set como resultado de aplicar PCA a la base de datos de Tumaco.*

### *3.5. Identificación de clases en el mapa final.*

En esta etapa se identifica el tipo de cobertura de terreno que representa cada una de las 13 clases definidas en el proceso anterior. Para esto, mediante el uso de un servidor y la ejecución de consultas a través de PostgreSQL, se contabiliza automáticamente la cantidad de puntos pertenecientes a cada una de los clusters del mapa final, y se compara con la misma medida en el mapa teórico de la Figura 1 para hacer el apareamiento. Debido a que en algunos casos más de un cluster, tenía un alto porcentaje de concordancia con el mismo grupo del mapa teórico, las clases se redujeron finalmente a 7 categorías, correspondientes a bosque general, bosque de pie de monte amazónico, cultivos

mixtos, cultivos de palma, cultivos anuales o transitorios, pasto y agua. El detalle del mapa final caracterizado se muestra en la Figura 9 con resultados satisfactorios, ya que por ejemplo, la consulta de la base de datos final asigna un 48% de la superficie del departamento a la cobertura de bosque, correspondiendo con los datos oficiales.

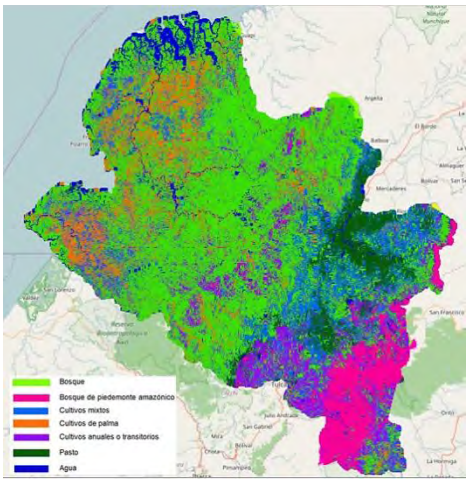



*Figura 9: Mapa final del departamento de Nariño, que contiene 7 tipos de cobertura, obtenidos después de la identificación de clases.*

## *4. Conclusiones.*

En este trabajo se presenta una metodología para caracterizar los tipos de biomasa en una región amplia de Colombia por medio de distintos tipos de algoritmos de clustering. La comparación realizada muestra que el algoritmo K-Means tiene una robustez comparable al algoritmo EM, aunque el primero se destaca en los índices de desempeño de la mayoría de las pruebas realizadas. Por su parte, los resultados del método de Isodata no permiten distinguir visualmente las clases en comparación con el mapa de referencia, aunque su eficiencia computacional es superior a los demás algoritmos probados.

Se muestra además la importancia de aplicar el proceso de PCA para reducir eficientemente la cantidad de atributos, extraer la información más relevante de los datos de entrada y disminuir el costo computacional de los algoritmos.

Para hacer una clasificación más profunda de los tipos de biomasa o coberturas en una zona con presencia de alta nubosidad (tal como la región central del departamento), se recomienda como trabajo futuro utilizar imágenes satelitales de mayor resolución e incluir otro tipo de características de la zona que puedan ser útiles a la hora de agrupar los pixeles con información precaria en las fotografías.

## *Referencias*


1. ALTOMONTE, H., *Las energías renovables no convencionales en la matriz de generación eléctrica*. Documentos de Proyectos, Comisión Económica para América Latina y el Caribe (CEPAL), 2017, ch. Las energías renovables: panorama mundial, latinoamericano y síntesis de tres estudios de caso, pp. 9-41.

2. CÁRCAMO, A., AND. REJAS, J., Análisis multitemporal mediante teledetección espacial y SIG del cambio de cobertura del suelo en el municipio de Danlí, El Paraíso, en los años 1987-2011. *Revista Ciencias Espaciales,* 8, 2 (2015), 259-271.

3. PÉREZ, D., *Identificación de ecosistemas en la provincia de Napo – Ecuador mediante análisis digital de imágenes satelitales*. Universidad San Francisco de Quito, Ecuador, (Octubre de 2012).

4. RODRÍGUEZ, J., PELUFFO, D., AND CUESTA, D. Unsupervised feature relevance analysis applied to improve ECG heartbeat clustering. *Journal of Computer methods and programs in biomedicine,* 108, 1 (October 2012), 250-261.

5. CASTILLEJO-GONZÁLEZ, I., ET AL., Object and pixel-based classification for mapping crops and their agro-environmental associated measures in QuickBird images. *Computers and Electronics in Agriculture,* 68, 2 (2009), 207-215.

6. GARCÍA-MORA, T., AND. MAS, J., Evaluación de imágenes del sensor MODIS para la cartografía de la cobertura del suelo en una región altamente diversa de México. *Boletín de la Sociedad Geológica Mexicana*, 63, 1 (2011), 83-94.

7. ÁLVAREZ, M. J., TRISTÁN, P., MASSA J. M., AND WAINSCHENKER, R. Clasificación automática de cubiertas terrestres en imágenes satelitales. *En XVII Congreso Argentino de Ciencias de la Computación*, *Tandil, Argentina* (2011).

8. ALVAREZ, J., VILLAGRA, P., CESCA, E., ROJAS, F., AND DELGADO, S. Estructura, distribución y estado de conservación de los bosques de Prosopis flexuosa del Bolsón de Fiambalá

(Catamarca). *Boletín de la Sociedad Argentina de Botánica,* 50, 2 (Junio 2015), 193-208.

9. CABRERA, O., CHAMPUTIZ, B., CALDERÓN, A., PANTOJA, A., Landsat and MODIS Satellite Image Processing for Solar Irradiance Estimation in the Department of Narino-Colombia. *En XXI Symposium on Signal Processing, Images and Artificial Vision, Bucaramanga, Colombia* (2016).

10. IRISH, R., BARKER, J., GOWARD, S., AND ARVIDSON, T., Characterization of the landsat-7 ETM+ automated cloud cover assessment (ACCA) algorithm. *Photogrammetric Engineering & Remote Sensing*, 72, 10 (2006), 1179–1188.

11. UNIVERSIDAD DE NARIÑO, Análisis de oportunidades energéticas con fuentes alternativas en el departamento de Nariño - ALTERNAR. http://alternar.udenar.edu.co, 2016.

12. CELIK, T., Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *IEEE Geoscience and Remote Sensing Letters,* 6, 4 (October 2009), 772-776.

13. CHARU, C. A., AND CHANDAN, K. R. *Data Clustering, Algorithms and Applications*. Chapman & Hall, 2014.

14. MATHWORKS INC. Matlab Documentation. https://www.mathworks.com/help/index.html, 2017.

**CERTIFICADO DE PONENCIA:** “Análisis de imágenes satelitales para clasificación de biomasa en el departamento de Nariño” – Congreso Internacional Multimedia 2017.

UNIVERSIDAD MILITAR NUEVA GRANADA

LA UNIVERSIDAD MILITAR NUEVA GRANADA

Certificación de presentación del ponencia titulada:

*ANÁLISIS DE IMÁGENES SATELITALES PARA CLASIFICACIÓN DE BIOMASA EN EL DEPARTAMENTO DE NARIÑO*

Alison Giovanna Bastidas, Andrea Lorena Bravo y Andrés Darío Pantoja

***CONGRESO INTERNACIONAL MULTIMEDIA***

28 y 29 de septiembre de 2017

Evento organizado por el programa Ingeniería en Multimedia

Ingeniera Carol E. Arevalo Daza M.SC
Decana Facultad de Ingeniería - Campus
Universidad Militar Nueva Granada

2017 Congreso Internacional Multimedia