



UNIVERSIDAD NACIONAL DE COLOMBIA

Diseño de una metodología de clasificación automática de unidades geomorfológicas en la geografía colombiana utilizando técnicas de reconocimiento de patrones

Diego Alberto Patiño Cortés

Universidad Nacional de Colombia - Sede Medellín
Facultad de Minas, Escuela de Sistemas
Medellín, Colombia
2012

Diseño de una metodología de clasificación automática de unidades geomorfológicas en la geografía colombiana utilizando técnicas de reconocimiento de patrones

Diego Alberto Patiño Cortés

Tesis presentada como requisito parcial para optar al título de:
Magister en Ingeniería de Sistemas

Director:

Ph.D, John Willian Branch Bedoya

Codirectora:

Ph.D, Verónica Botero Fernández

Línea de Investigación:

Inteligencia Artificial

Grupo de Investigación:

GIDIA - Grupo de Investigación en Inteligencia Artificial

Universidad Nacional de Colombia - Sede Medellín

Facultad de Minas, Escuela de Sistemas

Medellín, Colombia

2012

Dedicatoria:

Dedico este trabajo a mi familia, mi verdadero, sincero y más grande apoyo en los momentos más importantes de mi vida; los buenos y los malos. A mi hermano Andrés, a quien quiero, respeto y admiro profundamente; a mi madre Nora, que sabe quererme como nadie más en el mundo (y tolerarme además); a mi padre, Carlos, que supo inculcarme los principios, la tenacidad y el coraje para hacer de mis aspiraciones una realidad.

Todo mi esfuerzo tiene sentido gracias a ustedes.

Agradecimientos

En primer lugar agradezco el profesor Domingo Mery por sus valiosas ideas para el desarrollo de esta investigación; además agradezco su amabilidad, su buen trato, sus útiles y muy valiosas enseñanzas y su constante guía en el tiempo de mi estancia como pasante en el grupo de investigación GRIMA en la Universidad Católica de Chile.

Agradezco al director de esta tesis, John Willian Branch Bedoya, por sus consejos y por apoyarme, seguir y valorar mi trabajo a través de todo este proceso investigativo. Asimismo, agradezco a la escuela de sistemas de la Facultad de Minas y a todos los profesores, empleados y compañeros que de forma directa o indirecta contribuyeron con mi trabajo.

Al Posgrado en Aprovechamiento de Recursos Hidráulicoa. Al profesor Jaime Ignacio Vélez por permitirme formar parte del equipo de trabajo de HidroSIG y de muchos otros proyectos durante tanto tiempo; esta experiencia fue fundamental para poder desarrollar esta investigación, y lo será en etapas posteriores de mi vida; a la profesora Verónica Botero Fernández, por ser una gran persona, una excelente docente y por propiciar tantos momentos divertidos que tan necesarios son en todo grupo de trabajo. A todos mis compañeros allí, algunos de los cuales se han convertido en grandes amigos: Carlos Restrepo, Mario Jiménez, Cristian Ortiz, Leidy Yepes, Juan Camilo Castro, Nicolás Velásquez, Cahola Ramírez.

A la Universidad Nacional de Colombia y en especial a la Facultad de Minas por confiar en mi trabajo y ofrecerme un soporte institucional y económico para llevar a cabo mis estudios de pregrado y posgrado; así como por permitirme tener la muy importante experiencia de ser docente de tan prestigiosa institución.

Finalmente, agradezco el apoyo de toda mi familia, quienes siempre han sabido confiar en mí. A los amigos y compañeros que han estado ahí acompañándome durante los mejores años de mi vida, mi vida universitaria: David Saldaña, Maria Isabel Marín, Juan Carlos Rodríguez, Mariana Vásquez, Daniela Zapata.

Resumen

Este trabajo presenta los resultados de la aplicación de las texturas de Haralick al proceso de clasificación automática de unidades geomorfológicas en la geografía colombiana. En proyectos de índole ambiental y de ordenamiento territorial es importante conocer las clases de relieve existentes en una zona geográfica. El objetivo de este trabajo es incrementar el desempeño de este tipo de clasificaciones añadiendo información de texturas al conjunto de características comúnmente utilizado en este tipo de problemas. La información de texturas fue extraída utilizando el método de Haralick con ventanas móviles parametrizadas sobre mapas en formato Raster de la morfología de terreno. Se realizaron varias pruebas con diferentes clasificadores y validación cruzada con 30000 datos de diferentes muestras sobre la información disponible. Se demostró que las características basadas en texturas son útiles para el problema planteado ya a que su desempeño es superior (alrededor del 97 %) al alcanzado si se utiliza solo información morfológica del terreno (alrededor del 61 %).

Palabras clave: Clasificación automática, Unidad geomorfológica, Reconocimiento de patrones, Texturas de Haralick, Validación cruzada, Análisis de componentes principales, extracción y selección de características.

Abstract

This paper presents the results of applying Haralick's textures in the automatic landform classification process of Colombia's geography. For environmental and land use projects, is important to know the landform available in a geographic area. The objective of this work is to improve the performance of such classifications, by adding additional information to the commonly used data features in such problems. The texture information was extracted using the method of Haralick, parameterized with moving windows of geo-referenced maps in raster format. Those maps contains information of the morphology of terrain such as elevation, slopes, etc. Several tests were performed using different classifiers and cross-validation over several dataset with 30000 samples extracted from the study case data. It was shown that texture-based features are useful for the problem, because its performance is higher (about 97 %) compared to that achieved when using only morphological information of the terrain (about 61 %).

Keywords: Automatic classification, Landform, Pattern recognition, Haralick's textures, Cross-validation, Principal Component Analysis, Feature selection and extraction.

Contenido

Agradecimientos	VII
Resumen	IX
1. Introducción	2
1.1. Motivación	2
1.2. Definición del problema	4
1.3. Objetivos	6
1.3.1. Objetivo general	6
1.3.2. Objetivos específicos	6
1.4. Alcance o contribuciones	6
1.5. Organización del documento	7
2. Marco Teórico	8
2.1. Clasificación y reconocimiento de patrones	8
2.1.1. Clasificación supervisada (Inductiva)	8
2.1.2. Clasificación no supervisada (Deductiva)	10
2.2. Geomorfología básica	11
2.3. Estadística multivariada	12
2.4. Texturas de Haralick	13
3. Trabajos Relacionados	15
3.1. Antecedentes	15
3.2. Casos de estudio relacionados	16
3.3. Aspectos principales de la revisión de literatura	17
4. Definición del Caso de Estudio	18
4.1. Principales geo-formas de la geografía Colombiana	18
4.2. Clases objetivo para el problema de clasificación	19
4.3. Caso de estudio	22
4.3.1. Conjunto de datos para la clasificación	23
4.3.2. Subdivisión del conjunto de datos	24

5. Metodología de Clasificación	27
5.1. Criterio de segmentación	29
5.1.1. División del área de estudio	29
5.1.2. Definición del criterio de segmentación	30
5.1.3. Super clase 1	32
5.1.4. Super clase 2	32
5.2. Extracción de características basadas información de texturas	33
5.2.1. Descriptores de textura de Haralick	33
5.2.2. Extracción de texturas a partir de descriptores geomorfológicos	34
6. Experimentos y Resultados	37
6.1. Algoritmos, parámetros y selección de características	37
6.1.1. Selección de características	37
6.1.2. Elección de clasificadores y pruebas de validación	40
6.2. Clasificación de unidades geo-morfológicas	43
6.2.1. Separación de las 2 super clases	44
6.2.2. Clasificación de la super clase 1	45
6.2.3. Clasificación de la super clase 2	47
7. Conclusiones y Trabajo Futuro	51
7.1. Conclusiones	51
7.2. Trabajo futuro	52
Bibliografía	54

Lista de Figuras

2-1.	Estructura de una red neuronal artificial. Tomada de [17].	9
2-2.	Dendograma para la visualización la agrupación de datos. Tomada de [32].	10
2-3.	Representación de clusters. Tomada de [40].	11
2-4.	8 de las direcciones posibles para la relación espacial entre píxeles.	14
2-5.	Ejemplo de cálculo de la matriz de co-ocurrencia.	14
4-1.	Representación del relieve Colombiano con un SIG.	19
4-2.	Imágenes aereas de algunas de las principales unidades geo-morfológicas de la geografía Colombiana	21
4-3.	Cuenca del río Sinú.	23
4-4.	Representación de los 7 regiones de los subconjuntos de datos sobre el mapa de elevaciones del caso de estudio.	25
5-1.	Histograma del descriptor: CTI, para las 8 clases.	27
5-2.	Histograma del descriptor: Perfil de curvatura, para las 8 clases.	28
5-3.	Histograma del descriptor: Curvatura plana, para las 8 clases.	28
5-4.	Histograma del descriptor: Pendientes en grados, para las 8 clases.	29
5-5.	Descriptores morfológicos una vez agrupadas las clases en las dos super clases definidas.	31
5-6.	Ejemplo de algunos de los descriptores de textura extraídos.	36
6-1.	Gráficas de diferentes combinaciones de los 3 primeros descriptores de las diferentes selecciones para la super clase 1	41
6-2.	Gráficas de diferentes combinaciones de los 3 primeros descriptores de las diferentes selecciones para la super clase 2	42
6-3.	Histogramas de frecuencia de las pendientes en grados para las clases 5, 6, 7 y 8.	48

Lista de Tablas

4-1. Geo-formas seleccionadas para la realización de este estudio. [3]	20
4-2. Criterios para delimitar unidades geomorfológicas. [3]	22
4-3. Mapas del modelo de elevación de la cuenca del Río Sinú	24
4-4. Descripción de los subconjuntos de datos para la clasificación	26
5-1. Lista de los descriptores propuestos por Haralick.	34
6-1. Descriptores seleccionados para la super clase 1.	39
6-2. Descriptores seleccionados para la super clase 2.	39
6-3. Desempeño de la clasificación usando sólo descriptores geo-morfológicos . . .	44
6-4. Desempeños de la clasificación de las 2 super clases.	45
6-5. Desempeño de la clasificación de las geo-formas en la super clase 1	46
6-6. Proporciones y origen de los individuos de la muestra de la super clase 2. . .	49
6-7. Desempeño de la clasificación de las geo-formas de la super clase 2	49

1 Introducción

A menudo ocurre en proyectos ingenieriles que los costos de conseguir la información necesaria para el proyecto son elevados en términos de tiempo, esfuerzo y dinero. Además, la cantidad de información necesaria para llevar a cabo sus objetivos no siempre está disponible, ya sea porque es difícil de recolectar, almacenar, manipular o interpretar.

Esta situación y el acelerado crecimiento de mejores y más adecuadas tecnologías para diferentes tareas impulsa el uso de enfoques de resolución de problemas de todo tipo desde un punto de vista computacional; por lo tanto resulta útil buscar mecanismos automáticos que se apoyen en modelos computacionales y que faciliten la consecución o generación de la información necesaria para desarrollar dichos proyectos.

1.1. Motivación

Las unidades geo-morfológicas (geo-formas o tipos de relieve) son elementos topográficos que componen el paisaje y tienen características naturales físicas de la superficie de la Tierra. Como ejemplo se tienen los valles, montañas, planicies, colinas, glaciares, volcanes, etc. La clasificación de estos tipos de relieve es realizada frecuentemente con el objetivo de saber qué tipos de relieves existen en una determinada región geográfica. Este proceso es llevado a cabo por expertos geo-morfólogos que, a través de su experiencia, son capaces de reconocer cuándo un terreno es de un tipo de unidad geo-morfológica determinada.

La clasificación automática de unidades geomorfológicas surge de la necesidad de conocer la forma de un territorio evitando recorrerlo por completo; ya sea por ser de gran tamaño, que requiera un esfuerzo considerable o que sea de difícil o imposible acceso [9, 10].

Este tipo de clasificaciones constituyen un insumo importante usado por investigadores e instituciones en proyectos de toma de decisiones en áreas ambientales, socio-económicos o de ordenamiento territorial. Además, sirven como información base para diversos modelos hidrológicos y geológicos [3, 13], para estudios de erosión, estudios de predicción de riesgos y en otros campos de aplicación [8].

La clasificación de geo-formas se realiza manualmente con ayuda de expertos geomorfólogos que aplican su conocimiento para determinar cuándo una zona pertenece a cierto tipo de relieve. Esta es la manera en la que la mayoría de clasificaciones de tipo de relieve son realizadas actualmente. Los mapas manuales son entonces digitalizados para su posterior usos en sistemas de información geográfica (SIG) [53, 11].

De esta experiencia han surgido una serie de reglas empíricas que se utilizan como criterios computables para lograr clasificar los diferentes tipos de relieve. Comúnmente esto se realiza sobre datos derivados de modelos digitales de elevación que se almacenan en forma de mapas georeferenciados de variables geomorfológicas en formato **Raster**; e.g MDE (Modelo digital de elevaciones), pendientes, curvaturas, mapas de relieve relativo, etc.

Sin embargo, estas reglas no entregan resultados totalmente satisfactorios puesto que no logran capturar todo el conocimiento necesario para hacer una separación efectiva y precisa de las geo-formas de una región, y por tanto, la clasificación no se adapta totalmente a la realidad física del terreno.

El reconocimiento de patrones (**RP**) es el proceso a través del cual se logra construir un modelo que aproxime la forma en que los individuos de un conjunto de datos son asignados a clases (patrones) previamente definidas o desconocidas. Este proceso se realiza con el objetivo de llevar a cabo una clasificación de dicha información. La clasificación es la etapa final del proceso de reconocimiento y consiste en asignar una etiqueta a un vector n-dimensional donde cada dimensión representa una característica (descriptor o feature en inglés) de la información que se está analizando. Cada etiqueta representa una clase o agrupación de características comunes que pueden o no tener sentido en el mundo real en el contexto de los datos clasificados.

El reconocimiento de patrones es un área de investigación activa desde hace alrededor de 60 años, con temas de investigación como la reducción de dimensión de un espacio de datos, aprendizaje de máquina, selección y extracción de descriptores más adecuados para el proceso de clasificación, etc. Aún hoy esta área plantea desafíos importantes para los investigadores [32], y cuenta también con importantes campos de aplicación como la biometría, la seguridad informática, inspección visual industrial, visión por computador, etc [18].

Automatizar el proceso de clasificación de información geográfica a través de técnicas tomadas del área de reconocimiento de patrones conlleva a la reducción de esfuerzo y trabajo de campo requerido cuando se realiza la clasificación con una metodología manual; además, reduce los errores sujetos a los diferentes puntos de vista de quienes realizan el proceso de clasificación manualmente [48]. Sin embargo, y pese a que existen diferentes aproximaciones al tema; aún no se logra una automatización completa del proceso de clasificación debido a problemas tales como:

1. No existe un conjunto de unidades geomorfológicas estándar; de manera que para un estudio realizado en una determinada región geográfica, es factible la aparición de otros tipos de relieve no concebidos en el estudio, y que están presentes en otra región; reduciendo por tanto la generalidad de la solución del enfoque utilizado en el estudio hipotético.
2. No hay un acuerdo sobre el conjunto de características de las geo-formas que son más adecuados para la clasificación. En general cada autor elige y argumenta de manera diferente el conjunto de descriptores que usa.

3. Aunque hay una tendencia hacia algunas de las técnicas de reconocimiento de patrones en la literatura, no hay una argumentación clara que indique cuál es la mejor técnica de acuerdo a cada situación particular, ni una comparación concluyente entre diferentes enfoques con los que se aborda el problema.
4. Pese a que con los resultados obtenidos en diferentes aproximaciones al tema se han logrado índices de desempeño elevado; aún hoy las clasificaciones en dichos estudios se mantienen considerablemente alejadas del óptimo 100 %. Esto genera errores que son propagados de los datos clasificados hacia otro tipo de información cuando esta la clasificación es usada.

De lo anterior se deduce que todavía existe una gran cantidad de frentes de investigación en el problema de clasificar unidades geo-morfológicas; en estos frentes se pueden desarrollar varias investigaciones con el objetivo de mejorar todas o algunas de las etapas del proceso de clasificación, y de esta manera obtener resultados más precisos y adaptados a la realidad geo-morfológica de la zona que se desea clasificar.

1.2. Definición del problema

La clasificación automática de tipos de relieve ha sido un tema de interés en la comunidad científica desde hace alrededor de 15 años, como lo demuestran los estudios de Irvin et al. [31], MacMillan et al. [41] y Burrough et al. [10]. Estas clasificaciones se realizan a partir de información que describe las características del terreno, como su forma, geometría, propiedades del suelo, apariencia, etc.

En la comunidad científica existen numerosos trabajos que hacen referencia a este tipo de estudios utilizando diversos enfoques de reconocimiento de patrones para diferentes clases objetivo y en varias regiones geográficas. Una gran cantidad de estos trabajos sobresalen por el uso de técnicas de clasificación no supervisadas (clustering) para resolver el problema; empleando algoritmos como K-Means [25], ISODATA [6], fuzzy-K-Means [7] y redes neuronales auto-organizativas SOM [1, 5, 10, 4, 31]. Es importante resaltar que no existe una definición estándar de tipos de relieve, por lo tanto, cada clasificación lleva consigo un componente de subjetividad inherente a la geografía del caso de estudio para el que fue desarrollado.

En otros estudios sobre el tema abunda principalmente el uso de lógica difusa en conjunción con reglas heurísticas muy ligadas a conocimiento experto. Entre ellos se encuentra el trabajo de Arell et al. [4] y el trabajo de Dragut et al. [16]. Este último realiza una clasificación basada en características geométricas del terreno como la concavidad y la orientación. De esa manera definen nueve clases surgidas de las posibles combinaciones de los valores de curvatura plana y perfil de curvatura.

En la literatura los autores que abordan el problema suelen presentar siempre un conjunto de descriptores típico para este problema. Entre ellos se encuentran descriptores con información derivada de modelos digitales del terreno, tales como: Mapas de elevaciones, mapas de

pendientes, mapas de curvaturas, etc. Además de esta información también son abundantes las investigaciones en donde se estudia la generación de nuevos descriptores que ofrezcan mejores resultados al momento de realizar la clasificación. Para dar algunos ejemplos se pueden relacionar los trabajos de Acciani et al. [1] y Ardiansyah et al. [48] y Renno et al. [50].

En este trabajo se desarrolla una metodología de clasificación de unidades geo-morfológicas sobre una región geográfica que se encuentra dentro del territorio Colombiano. Se busca discriminar ocho tipos diferentes de relieve pertenecientes a la geo-morfología de la cuenca del Río Sinú en el Noroccidente de Colombia. Las geo-formas elegidas son Vertientes, cañones, pie de monte abierto, terrazas y colinas, deltas y estuarios, ciénagas con control fluvial, ciénagas con control del mar y litorales. Las 8 geo-formas representan ocho clases objetivo en el problema de clasificación que se planteó. Los descriptores que fueron utilizados para la clasificación corresponden a características de la forma del terreno derivadas del modelo de elevación digital de los datos.

Pese a la cantidad de aproximaciones al tema en los enfoques utilizados en trabajos anteriores, estos no resuelven totalmente el problema de las clasificación de tipos de relieve. Esto ocurre porque los descriptores comúnmente utilizados no ofrecen una separabilidad satisfactoria de las clases; conllevando a desempeños de clasificación de alrededor del 85% de los datos bien clasificados [47, 46, 42, 19, 48]. Además algunos autores recurren al uso de técnicas de clasificación no supervisadas[32], lo que les obliga a presentar una evaluación cualitativa o visual de los resultados de clasificación obtenidos, sujeta al punto de vista de quien visualiza estos resultados [16], [27], [31], [52].

Para obtener un desempeño significativo en la separación automática de las clases propuestas se considera necesario explorar alguna otra técnica o técnicas que permitan extraer nuevas características del conjunto de datos más adecuados para la clasificación. Los nuevos descriptores deben ofrecer un considerable grado de separabilidad de las clases objetivo mejorando los resultados obtenidos en comparación con los descriptores usados en trabajos previos. Para tal propósito se ha elegido una metodología de extracción de características basadas en la textura del terreno. Una textura de terreno puede ser definida como los patrones de comportamiento que se perciben de la forma de una región geográfica; de esta manera una unidad geo-morfológica será clasificada teniendo en cuenta información obtenida a partir de un entorno centrado en el punto que se desea clasificar.

Uno de los problemas más simples de clasificación es aquel en el que solo se tienen dos clases objetivo (clasificación binaria); una vez que el número de clases aumenta la dificultad de separar las clases conjuntamente crece. En este trabajo se propone una metodología de clasificación de tipo jerárquico, de manera que las ocho unidades geo-morfológicas son separadas en varios grupos que llamaremos **super clases**. De esta manera hemos dividido el problema de clasificación en varios sub problemas, cada uno de ellos con un número menor de clases objetivo. A través de esta división podemos diseñar varios experimentos de clasificación, utilizando un conjunto diferente de descriptores para cada super clase, reduciendo la complejidad del problema global.

Los datos que utilizados para los experimentos en esta investigación quedan definidos como la unión de un sub-conjunto de los descriptores más utilizados en la literatura, con los nuevos descriptores basados en texturas. Luego de tener los datos consolidados se realizaron varias selecciones sobre este, para determinar cuales de ellos ofrecían la mejor separación de las clases objetivo en cada una de las sub-clases para cada super-clase. Para tal propósito se utilizó el algoritmo de selección de características SFS [33] con diferentes conjuntos de parámetros. Posteriormente se utilizaron varios clasificadores supervisados conocidos con el fin de encontrar aquel que obtuviera el mejor desempeño.

1.3. Objetivos

Para el desarrollo de esta investigación se plantearon los siguientes objetivos.

1.3.1. Objetivo general

Definir un método semi-automático fundamentado en técnicas de reconocimiento de patrones que sea capaz de aproximar una clasificación de las principales geo-formas presentes en la geografía Colombiana.

1.3.2. Objetivos específicos

1. Identificar cuales son las geo-formas más relevantes que se encuentran en la geografía Colombiana para diferentes escalas; además definir cuales son sus principales características.
2. Caracterizar las diferentes técnicas de reconocimiento de patrones según su respuesta al tratar de identificar las diferentes geoformas de interés.
3. Definir uno a varios criterios de segmentación para identificar sub-zonas en la zona definida por el conjunto de datos de entrada.
4. Esquematizar un algoritmo que se base en el criterio de segmentación para aplicar la técnica de clasificación más adecuada a cada región.
5. Comparar los resultados del método propuesto con clasificaciones geomorfológicas realizados por expertos para evaluar sus bondades.

1.4. Alcance o contribuciones

Nuestra principal contribución con el desarrollo de este trabajo es la consolidación de un conjunto de descriptores apto para el problema de la clasificación de ciertos tipos de relieve; de

manera que el desempeño obtenido con los descriptores seleccionados sea considerablemente mayor que los reportados en la literatura en aproximaciones de solución al mismo problema. El enfoque utilizado usa información de la textura del terreno similar a como lo haría un experto geo-morfológico a través de la observación y análisis de una región geográfica. Se resalta también la puesta en práctica de una clasificación jerárquica en la cual el problema planteado es atacado a través de la división en varios sub-problemas de complejidad menor y características suficientemente diferentes. De tal forma fue posible mejorar el rendimiento en los resultados obtenidos, e incluso disminuir los tiempos de ejecución de la fase de entrenamiento y evaluación de las diferentes clasificaciones realizadas en los experimentos. Este trabajo también presenta una comparación de desempeño de la clasificación propuesta a través de la utilización de diferentes tipos de clasificadores supervisados, para diferentes conjuntos de características. Esta comparación incluye aquella clasificación que se realiza sólo con los descriptores clásicos presentados en trabajos anteriores.

1.5. Organización del documento

Este documento está organizado de la siguiente manera: En el capítulo 2 se presenta el marco teórico para la investigación presentada. Este contiene un síntesis de los conceptos sobre reconocimiento de patrones, geomorfología básica, estadística multivariada y análisis de texturas en imágenes; que fueron utilizados para el desarrollo de esta investigación. En el capítulo 3 se da a conocer la revisión de literatura consultada sobre el problema abordado; se resaltan los trabajos más importantes y se dan a muestran algunos casos de aplicación que ilustren el interés científico en el área. En el capítulo 4 se establece y se argumenta cuales serán los tipos de relieve que se quiere clasificar. Los tipos de relieve serán las clases objetivo del problema de clasificación. En la sección 5 se propone y explica detalladamente la metodología que se siguió para realizar la clasificación de los tipos de relieve que son objeto de estudios; como parte de la metodología se define un criterio de división en super clases del conjunto de clases objetivo en el problema de clasificación planteado. La sección 6 explica los experimentos que fueron llevados a cabo en este trabajo, y se dan detalles de todas las etapas del proceso de clasificación y de los datos utilizados para dicho propósito; se presentan también los resultados de dichos experimentos con su respectiva discusión. Finalmente en la sección 7 se dejan algunas conclusiones y se plantean varios frentes de trabajo futuro.

2 Marco Teórico

2.1. Clasificación y reconocimiento de patrones

Dentro de la literatura se mencionan 4 enfoques principales con los cuales se puede abordar el problema del reconocimiento de patrones: Redes neuronales, ajuste de plantillas, reconocimiento de patrones sintáctico y reconocimiento estadístico de patrones [34]. A su vez se denomina a un clasificador como supervisado o de aprendizaje inducido cuando el entrenamiento del mismo se realiza minimizando el error entre la salida real del clasificador y la salida esperada conocida a priori; en contraposición, un clasificador no supervisado es aquel que debe estimar no sólo los parámetros para clasificar sino también el conocimiento de que clases existen en el conjunto de datos de entrada [34]. La elección de un enfoque depende del tipo de problema que se vaya a abordar puesto que cada uno posee ventajas y debilidades. El reconocimiento de patrones puede subdividirse en diferentes etapas:

Pre-procesamiento de los datos de entrada: Consiste en seleccionar un subconjunto de los descriptores (también llamados features en inglés) de los datos, o la generación de nuevos a partir de los existentes. Otra técnica habitual es la reducción de la dimensionalidad a través de herramientas como el análisis de componentes principales.

Selección y aprendizaje del algoritmo: El algoritmo elegido es determinado por el tipo de información que se desea clasificar. Para cada algoritmo existen una serie de mecanismos para realizar el aprendizaje según sus características.

Análisis del error: En el caso de la clasificación supervisada se realiza con un subconjunto de los datos clasificados a priori; y en el caso no supervisado se realiza con la ayuda de un experto que le da sentido físico a la clasificación realizada y metodologías como las matrices de confusión entre otras.

A continuación se presenta una breve descripción de las técnicas supervisadas y no supervisadas:

2.1.1. Clasificación supervisada (Inductiva)

Existen muchas técnicas de clasificación supervisada, incluyendo técnicas híbridas; sin embargo, las más utilizadas en la literatura se exponen en [39] y se resumen a continuación:

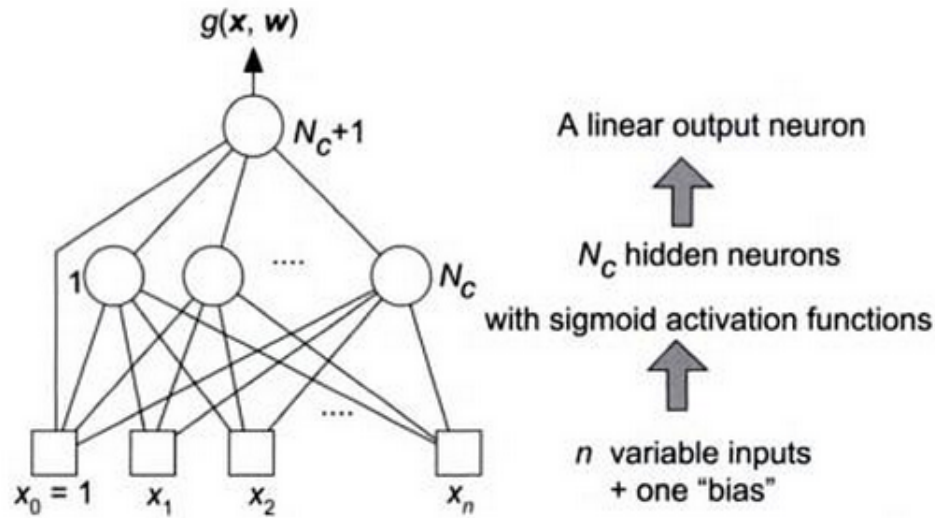


Figura 2-1: Estructura de una red neuronal artificial. Tomada de [17].

La primera técnica agrupa los llamados métodos lógicos como los *árboles de decisión* donde cada nodo del árbol representa una característica del conjunto de datos, y sus hojas son las clases existentes en él. La clasificación se lleva a cabo recorriendo las ramas del árbol hasta llegar a las hojas, que presentan las clases. Son susceptibles de *overfitting* [26] que puede evitarse si no se permite que el algoritmo de entrenamiento complete la estructura del árbol, truncándose. Las diferentes rutas entre la raíz de los árboles y sus hojas pueden verse como las reglas de decisión que conforman el clasificador. Se han desarrollado métodos que pueden extraer estas reglas desde el conjunto de datos, un ejemplo de ellos es el algoritmo RIPPER. Otra metodología de clasificación supervisada son las *redes neuronales artificiales*, aunque estas sirven para múltiples propósitos, se utilizan también como método supervisado de clasificación. Las arquitecturas comunes para esta tarea son los perceptrones simples para clasificación binaria y lineal; y los perceptrones multicapa para problemas de múltiples clases y no linealidad (Ver figura 2-1). Un problema común con los perceptrones multicapa es que su entrenamiento es lento; por lo que existen una serie de criterios de parada que truncan el entrenamiento de la red para asumirla correctamente entrenada. Son también susceptibles al *overfitting*. Un problema de las redes neuronales es su falta de habilidad para razonar sobre su salida (network output) en una manera en que dicho razonamiento pueda ser efectivamente comunicado a la red.

Otra metodología de clasificación son los métodos estadísticos que se diferencian del resto por tener un modelo probabilístico definido; y que asignan una clase a un vector de manera que su probabilidad de pertenencia a ella sea mayor que la probabilidad de pertenencia a las otras clases. Entre ellos se encuentra el *Clasificador simple de Bayes* y las *Redes Bayesianas*. Estas últimas poseen un gran inconveniente al momento de determinar su estructura puesto que el número de posibles arquitecturas crece exponencialmente con n (número de nodos de la red). Varios trabajos han sido realizados para eludir este problema; principalmente,

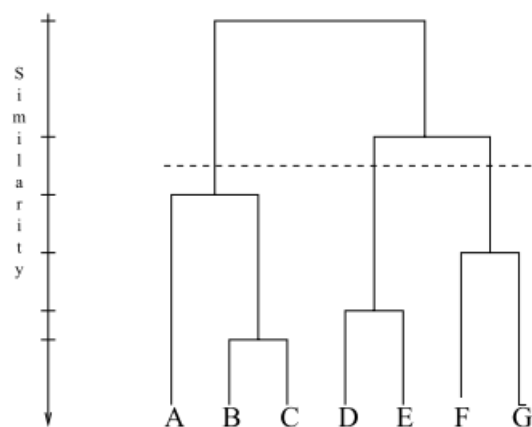


Figura 2-2: Dendrograma para la visualización la agrupación de datos. Tomada de [32].

creando arquitecturas basadas en el conocimiento de la independencia entre las dimensiones del espacio de datos de entrada.

Los clasificadores basados en instancias son tipos de clasificadores estadísticos. Los algoritmos más representativos de este tipo de clasificadores son el *Vecino cercano (NN)*, por sus siglas en inglés, y el *K-Vecinos-Cercanos (KNN)*. Ambos se basan en que dados ciertos vectores de entrada clasificados, se etiquetan el resto de vectores a través de un proceso de selección con un criterio de distancia. Si el vector está cerca de otro vector clasificado la etiqueta se propaga. El principal problema de KNN es que resulta complicado determinar el valor óptimo de k con el que se obtienen mejores resultados en la clasificación.

2.1.2. Clasificación no supervisada (Deductiva)

Los métodos de clasificación no supervisada, también llamados *análisis exploratorio de datos* o *clustering*, fueron revisados principalmente de [34] y [32] donde se describen sus características:

En Jain et al. [32] se define una categorización principal de técnicas de clustering dividiéndolas en dos categorías: Jerárquica y particional. Las técnicas jerárquicas consisten en la construcción de una estructura de niveles de similitud con cada uno de los vectores del conjunto de datos de entrada. Cada elemento se agrupa con otro que comparta sus características para formar un nuevo elemento que a su vez puede agruparse con otro, formando así la jerarquía. Para visualizar la forma en que la información se organiza se usan los dendogramas. En ellos se puede seleccionar un nivel de similitud y truncar la estructura jerárquica tomando como clusters individuales los valores hacia abajo del dendrograma. La figura 2-2 muestra un dendrograma típico para un conjunto de 7 datos (A, B, C, D, E, F, G).

Aunque existen varios algoritmos de clasificación jerárquica, la mayoría derivan del *Agglomerative Single-link algorithm* y el *Agglomerative Complete-link algorithm* descritos en [32]. Las técnicas particionales actúan creando k particiones independientes del espacio de los

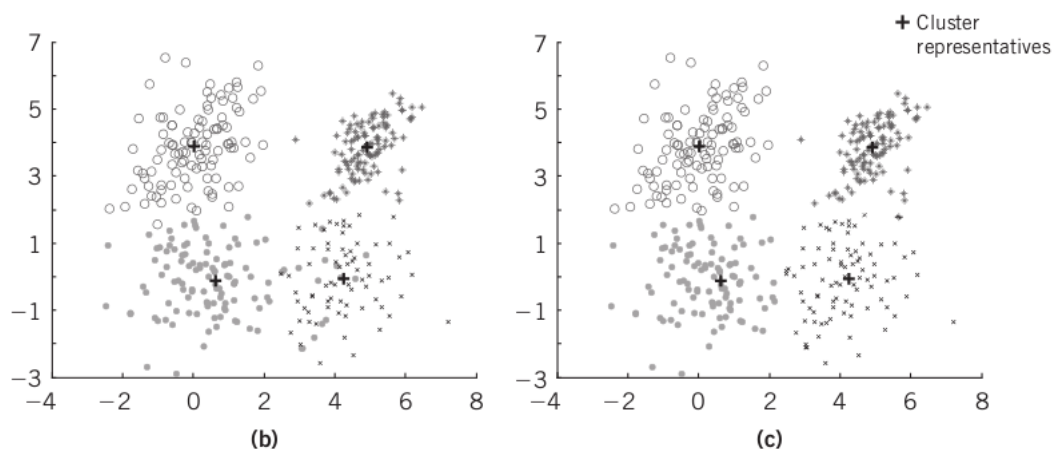


Figura 2-3: Representación de clusters. Tomada de [40].

datos en vez de una estructura como en los casos anteriores; estas particiones son llamadas *clusters* (Ver figura 2-3). Esta propiedad los hace útiles cuando se trabaja con grandes cantidades de datos; pero también introducen un problema importante que es la determinación del número apropiado de clusters (k) que será utilizado.

El enfoque particional más utilizado es aquel en el cual se seleccionan los clusters de manera que se minimice el error cuadrático medio entre los elementos que lo componen. A este tipo de algoritmos pertenece el *k-means* y su sucesor el algoritmo *ISODATA*; que es una versión mejorada de *k-means*. *ISODATA* fue diseñado para forzar al algoritmo hacia la creación de un número de agrupaciones predefinido [32] a través de la división o agrupación de otros clusters cuando superan ciertos umbrales que son entregados como parámetros al algoritmo. Existen algunos otros enfoques que combinan técnicas supervisadas y no supervisadas (semi-supervisadas); e incluso que interactúan con otras disciplinas para formar metodologías evolutivas y genéticas [28], [54].

2.2. Geomorfología básica

La forma de la superficie de la tierra es el resultado de un conjunto de procesos o fenómenos que ocurren y han ocurrido siempre sobre el planeta; estas fuerzas pueden dividirse en procesos internos y externos, y ambas son las responsables de moldear el relieve del planeta. Los procesos internos son aquellos cuya energía proviene de las capas internas de la tierra; y los procesos externos son aquellos que se dan sobre la superficie terrestre, y cuya fuente de energía proviene principalmente del clima y de la interacción gravitacional.

Los procesos internos son los encargados de la formación de nuevo relieve debido a que produce fuerzas que cambian la forma superficial de la tierra. Aunque algunos de ellos son lentos en tiempo perceptible por los humanos, como la formación de cordilleras (algunos milímetros al año), otros son abruptos y producen cambios significativos en poco tiempo. e.g

las erupciones volcánicas y los terremotos de gran magnitud. Los principales mecanismos que llevan a cabo este tipo de procesos son la tectónica, el diapirismo, la isostacia y el vulcanismo. Contrario a los procesos internos, los externos son los encargados de suavizar el relieve de la tierra. Los mecanismos a través de los cuales se llevan a cabo son el agua, el hielo, el viento, etc. Dentro de los procesos externos más importantes se encuentran la meteorización (física y química), la erosión, el transporte y la sedimentación[30, 14].

Meteorización: Es el proceso a través del cual se descomponen las rocas de la corteza de la tierra. Existen dos tipos de meteorización: Química y física. En la química la roca cambia su composición cuando sus componentes reaccionan con otros elementos y compuestos. En la física su estructura química se mantiene intacta y sólo se divide en fragmentos (detritos) de la misma composición.

Erosión: Este fenómeno es a través del cual los detritos son separados de la roca meteorizada cambiando la forma del terreno al ir disminuyendo paulatinamente el volumen de la roca inicial. En ocasiones la erosión puede modificar las condiciones de estabilidad de las roca y el terreno en general, produciendo desastres.

Transporte: En este proceso los detritos y componentes extraídos de la roca erodada son trasladados a otros lugares a través de medios como el agua (en forma de escorrentía o caída de lluvia), el aire (corrientes de viento) o el hielo (movimiento glaciar).

Sedimentación: Es el proceso a través del cual los componentes transportados se acumulan para formar diferentes tipos de geoformas. e.g abanicos fluviales, rocas sedimentarias, y los lechos de los ríos.

En la génesis de la geomorfología Colombiana intervienen en gran medida procesos fluviales debido a las características geográficas y climáticas del país. Por lo tanto, entre las principales geoformas que pueden ser encontradas en el territorio nacional están los diferentes tipos de valles de acuerdo las características de estos, llanuras de inundación, terrazas aluviales, abanicos aluviales, zonas de depositación fluvial y ciénagas. Cabe resaltar que en general estas geoformas están asociadas a cierto riesgo de desastre por inundaciones en algunas épocas del ciclo climático de la región.

2.3. Estadística multivariada

Teniendo en cuenta que en la gran mayoría de los casos los datos de entrada utilizados en reconocimiento de patrones conforman conjuntos multivariados; es necesario para el desarrollo de la investigación hacerse con herramientas propias de esta área.

Dentro de los conceptos útiles de estadística, aplicados al análisis de los datos multivariados, encontramos por ejemplo la matriz de varianza/covarianza (Σ) que juega un papel decisivo

al momento de analizar entre los descriptores del conjunto de datos de entrada aspectos como la variabilidad, la correlación, la distancia entre individuos, etc. Aún más, los vectores y valores propios (λ_i, e_i) de Σ son herramientas útiles para diversos análisis pertinentes en el campo del RP, como la distancia estadística de Mahalanobis [35] y el análisis de componentes principales [36].

Las técnicas de Clustering [32] son ampliamente utilizadas y discutidas en el campo de reconocimiento de patrones; incluso, en el problema específico de la clasificación de geofor-mas juega un papel fundamental al ser la técnica no supervisada que más se utiliza, como se verá más adelante. Estas técnicas son consideradas como métodos de clasificación estadísticos, y por eso, al momento de ser utilizadas se debe tener en cuenta aspectos como las distribuciones de probabilidad de las poblaciones de las cuales provienen los datos y la variabilidad de los diferentes descriptores utilizados al momento de comparar individuos, etc. En cualquier análisis estadístico es importante tener en cuenta cual es la distribución de probabilidad de la población de la cual provienen los datos. La distribución de probabilidad *Normal* es tal vez la más utilizada o la más intuitiva; sin embargo, en ocasiones se suele dar por sentado torpemente el supuesto de que los datos provienen de una distribución *Normal Multivariada*. Aunque esto en ocasiones se cumple, y aunque para muestras grandes puede asumirse normalidad [35] es importante conocer y saber aplicar pruebas de normalidad o realizar transformaciones para aproximar la normalidad cuando es necesario, con el objetivo de facilitar posteriores análisis sobre los datos.

2.4. Texturas de Haralick

En la década de los 70's Haralick[24], desarrolló un modelo para analizar texturas en una imagen por medio del uso de la matriz de co-ocurrencia (COM) de los valores de intensidad de la imagen.

La matriz de co-ocurrencia describe la frecuencia de un nivel de gris que aparece en una relación espacial específica con otro valor de gris, dentro del área de una imagen determinada. La matriz de co-ocurrencia es un resumen de la forma en que los valores de los píxeles ocurren al lado de otro valor en una misma imagen. La relación espacial se define como la dirección en la que la comparación de los píxeles es realizada; de manera que se cuentan con 8 posibles direcciones en el que un pixel puede ser comparado con sus vecinos. Esta situación se ilustra en la figura 2-4.

Aunque es común que la relación espacial entre píxeles se haga entre los 8 vecinos mas cercanos la matriz de co-ocurrencia no tiene que estar limitada a esta situación, pudiendo relacionarse un pixel con vecinos muchos más alejados.

La matriz de concurrencia que se denotará como P_{kl} , donde el elemento $P_{kl}(i, j)$ otorga el valor de frecuencia (divido por NT) de ocurrencia de los valores de color i y j en dos píxeles ubicados en una posición relativa dada por el vector (k, l) . La variable NT significa el número de píxeles que fueron necesarios para calcular P_{kl} , con esto se normaliza la matriz

		(-1,-1)	(0,-1)	(+1,-1)
		(-1,0)	(0,0)	(+1,0)
		(-1,+1)	(0,+1)	(+1,+1)


Figura 2-4: 8 de las direcciones posibles para la relación espacial entre píxeles.

de concurrencia ya que la suma de todos sus elementos es uno.

Si la variable de color tiene una resolución de 256, por ejemplo de 0 a 255, el tamaño de la matriz de concurrencia P_{kl} será 256x256. Ya que esto implica un costo computacional muy alto, es común que se utilicen matrices más pequeñas empleando sólo los bits más significativos de la variable de color [12]. A manera de ejemplo, se puede tener una matriz de co-ocurrencia de 8x8 agrupando el valor de la variable de color x en $[0, \dots, 31]$, $[32, \dots, 63]$, ... $[224, \dots, 255]$.

A modo de ejemplo en la figura 2-5 se muestra una una imagen hipotética con 4 niveles de gris (entre 0 y 3) y su respectiva matriz de confusión en la dirección (1,0).

Valores de intensidad de la imagen (0-3)				
0	1	0	3	0
1	3	2	1	1
0	1	1	3	2
2	2	1	3	2
1	1	0	3	1



Matriz de co-ocurrencia				
0	2	0	2	2
2	3	0	3	3
0	2	1	0	0
1	1	3	3	0

Figura 2-5: Ejemplo de cálculo de la matriz de co-ocurrencia.

Una vez que se tiene la matriz de co-ocurrencia calculada Haralick propone 14 métricas extraídas a través de cálculos sobre los valores de la COM. Estas métricas resumen de manera cuantificable parte de la información de texturas presentes en la imagen. En el capítulo 5 se describirán estos descriptores.

3 Trabajos Relacionados

3.1. Antecedentes

La clasificación automática de geoformas ha sido un tema de interés en la comunidad científica desde hace alrededor de 20 años, como lo demuestran los estudios de Irvin et al. [31], MacMillan et al. [41] y Burrough et al. [10].

En el caso de Irvin et al. su trabajo ha servido como antecedente para muchos otros trabajos de publicación posterior donde se discute y cita; convirtiéndose en un referente importante en el tema. El propósito de Irvin et al. fue realizar una comparación de dos metodologías de clasificación: El algoritmo ISODATA y reglas difusas; esta última es denominada por los autores como *clasificación continua* debido a la naturaleza de los conjuntos difusos.

Los autores afirman en el estudio que para realizar el análisis de medias y de varianzas de los clusters, los datos de entrada deben ser aproximadamente Gaussianos en cada descriptor; de esta manera, cuando los datos no cumplían este supuesto fueron divididos en subconjuntos para aplicar el algoritmo individualmente para cada subconjunto.

Los resultados para ambas metodologías arrojaron resultados satisfactorios al ser comparados con clasificaciones manuales; sin embargo, en el caso de la clasificación con ISODATA las clases debieron ser examinadas manualmente para asignarles un sentido físico; en ocasiones varios clusters fueron mezclados para representar una misma geoforma. Por el contrario, en la clasificación continua no hubo necesidad de hacer una análisis extra de los datos.

Un poco más tarde, MacMillan et al. [41] presentan una metodología de clasificación que utiliza un conjunto de reglas heurísticas y el uso de lógica difusa. En su estudio se presenta un conjunto de los descriptores derivados de DEM's más utilizados para clasificación de geoformas. Entre estos descriptores se encuentran: El gradiente de la pendiente, el perfil de curvatura y curvatura plana, el mapa de aspecto, el mapa de iluminación solar, el índice de humedad (Wetness Index) o índice topográfico compuesto (Compound Topographic Index), posición de la pendiente y longitud de la pendiente.

Algunos otros descriptores mas que los autores consideraron útiles para el estudio fueron incluidos en del conjunto de datos de entrada para la clasificación. Estos nuevos descriptores son el porcentaje Z relativo a la parte superior e inferior de cada cuenca, porcentaje Z relativo a los picos y sumideros, porcentaje Z relativo a la corriente más cercana y a la divisoria más cercana, altura absoluta (Z) sobre la celda de sumidero local y sumidero máximo absoluto con respecto al pico local.

Otro estudio de gran importancia es el de Burrough et al. [10]; tal estudio pretende en parte

confirmar los resultados obtenidos en [31], y en parte establecer una nueva metodología difusa que permita asignar a un individuo un grado de pertenencia a cada clase y no sólo una pertenencia única a una de ellas. Para lograr lo anterior se usa en el estudio el algoritmo *fuzzy c-means* para llevar a cabo la clasificación; y se utilizan varios valores de k en cierto rango para encontrar diferentes *clusters* significativos.

Para elegir el valor de k que mejores resultados entregara en la clasificación se usaron artefactos como el coeficiente de partición F y la entropía de la clasificación H . El trabajo plantea además que algunos de los descriptores que son derivados de los DEM afectan la clasificación debido a la forma en la que son calculados ya que no reflejan las condiciones reales del terreno debido al algoritmo que se usa para generarlos. Un ejemplo de esto son los modelos de extracción de redes de drenaje que asumen que una celda puede ser o no una celda de red drenaje, descartando situaciones en las que la corriente es más ancha o más angosta que la resolución del mapa.

En este estudio los autores se vieron enfrentados al tamaño de los datos, utilizando entonces muestras pequeñas para estimar el número de clusters para los algoritmos fuzzy c-means; sin embargo, teniendo en cuenta la fecha de publicación del estudio, es plausible considerar que tamaños de datos incluso superiores a los presentados en el caso de estudio del artículo pueden ser fácilmente computable por ordenadores con tecnologías actuales.

3.2. Casos de estudio relacionados

Existen otros estudios donde abunda principalmente el uso de lógica difusa en conjunto con reglas heurísticas muy ligadas a conocimiento experto; así como también ligados a técnicas de reconocimiento de patrones estadístico (*Fuzzy C-Means*). Entre esos estudios se encuentra el de Arell et al. [4] y el trabajo de Dragut et al. [16].

En [4] se utilizó la técnica *fuzzy C-Means* varias veces sobre un mismo conjunto de descriptores (pendiente, curvatura plana y perfil de curvatura) pero con diferentes resoluciones. El estudio arrojó buenos resultados, confirmando así los trabajos de [31] y [10]. En este estudio se considera que la *elevación* y el *mapa de aspecto* sólo tienen una contribución limitada para caracterizar unidades morfológicas y por eso fueron omitidas del conjunto de descriptores utilizados.

Una conclusión relevante del trabajo de Arrell et al. es la existencia de algunas geo-formas, como picos y zonas de baja pendiente, que son persistentes en la clasificación sin importar la resolución de los datos.

En otro estudio [16] realiza una clasificación basada en objetos con características geométricas del terreno como la concavidad y la orientación. De esa manera definen 9 clases surgidas de las posibles combinaciones de los valores de curvatura plana y perfil de curvatura. Estas clases definen los objetos que pueden existir en el terreno, y que posteriormente serán clasificados utilizando una combinación de reglas difusas con funciones de pertenencia flexibles, y algunas otras reglas no difusas.

3.3. Aspectos principales de la revisión de literatura

Aun en la actualidad se sigue trabajando en el área debido a que la información sobre geformas es necesaria por ejemplo para la evaluación del paisaje, estudios de erosión, estudios de predicción de riesgos y de varios otros campos de planeamiento territorial. Entre algunos de los trabajos recientes en el área se encuentra el de Deng et al. [15] y el de Klingseisen et al. [37].

Según la literatura existen dos temas importantes que se resaltan en cada trabajo sobre clasificación de formas de relieve: (1) La selección y generación de los datos de entrada que se utilizarán para clasificar; y (2) la elección del algoritmo adecuado para la clasificación.

En cuanto a la selección y generación de los datos de entrada, los autores presentan siempre un conjunto de datos típico entre los que se encuentra información derivada de modelos digitales de elevación como mapas de pendientes, mapas de curvaturas, etc. Además de estas capas de información, también son abundantes las investigaciones en donde se estudia la generación de nuevas capas (como mapas raster) que ofrezcan mejores resultados al momento de realizar la clasificación. Para dar algunos ejemplos están los trabajos de Acciani et al. [1] y Ardiansyah et al. [48], y especialmente los trabajos de Nobre et al. [45] y Renno et al. [50], donde se introduce y utiliza un nuevo descriptor de relieve relativo llamado HAND (Height Above Nearest Drainage).

En cuanto a la selección del algoritmo de clasificación, no se evidencia una predilección específica por las técnicas supervisadas o no supervisadas. Los autores las usan de acuerdo a que tan bien se adapten a las zonas y la información seleccionada como se puede observar en [10] y [48].

No obstante, los documentos seleccionados dejan en evidencia una tendencia desde hace por lo menos 20 años en la utilización de técnicas que incluyen componentes difusos en combinación con las técnicas actuales [52], [31], [10].

4 Definición del Caso de Estudio

En esta sección se establece detalladamente el caso de estudio sobre el cual se le aplicó el proceso de clasificación de unidades geo-morfológicas desarrollado en esta investigación. Se delimita la zona geográfica del territorio Colombiano sobre la cual se realizó la extracción de los datos morfológicos del terreno, y se describen las características físicas de algunas de las unidades geo-morfológicas más importantes presentes en esta zona. Estas unidades conforman el conjunto de clases objetivo del problema de clasificación planteado; de esta manera los experimentos de clasificación tendrán el objetivo de separar la información morfológica de los datos de la zona de estudio en los diferentes tipos de geo-formas definidas a continuación.

4.1. Principales geo-formas de la geografía Colombiana

En esta sección se presentan algunos aspectos más relevantes de los diferentes tipos de relieves existentes en la geografía Colombiana. En este punto se debe aclarar que la estructura física del territorio Colombiano obedece a una serie de eventos tecto-dinámicos generadores de relieve y una morfología definida básicamente por procesos bio-climáticos locales y globales que se desarrollan lentamente en tiempo geológico. Actualmente los cambios en el paisaje también pueden atribuirse a la intervención antrópica.

El territorio colombiano puede dividirse en dos grandes tipos de relieve: Montañas y llanuras bajas; siendo éstas últimas las que ocupan la mayor cantidad de área. Dentro de su topografía se pueden distinguir tres sectores principales. El primero sector corresponde al sistema montañoso andino, conformado por las tres cordilleras y los diferentes valles interandinos, abarcando gran parte del occidente del país.

Las 3 cordilleras se ubican en el occidente del país alineada con las costas pacífica y atlántica; se originan en el sur cerca del nudo de los pastos y se extienden hasta la Guajira en su brazo más largo (la oriental). Este sistema de cordillera está partido por los valles de dos de los ríos más importantes de Colombia que fluyen de sur a norte y se unen para desembocar en el océano atlántico. Estos ríos son: El río Magdalena que separa a la cordillera oriental de la central, y el río Cauca que separa la cordillera central de la occidental. El río Atrato, también de gran importancia en Colombia, fluye de norte a sur al occidente de la cordillera occidental. La cordillera central es la más alta de las 3 cordilleras, y una de sus principales características es la presencia de varios volcanes y nevados a lo largo de su extensión, así como también la existencia de muchos ecosistemas de tipo páramo.

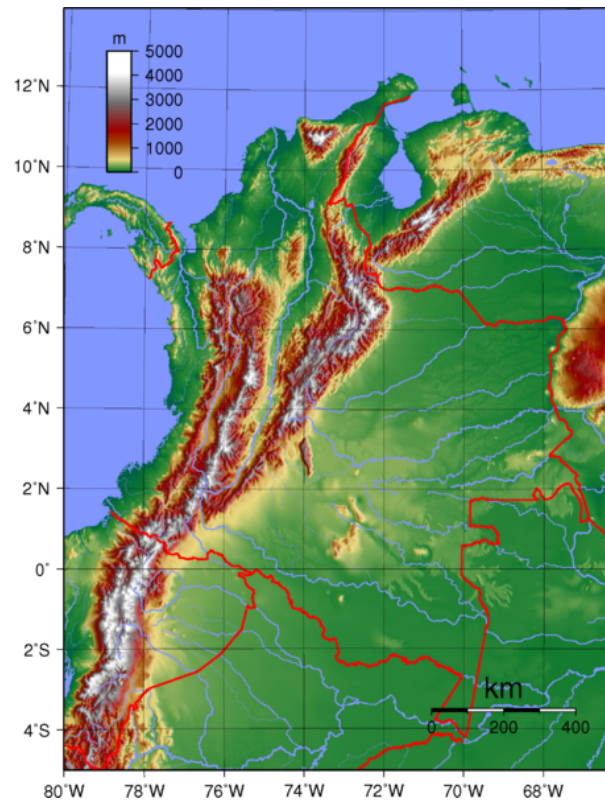


Figura 4-1: Representación del relieve Colombiano con un SIG.

El segundo sector lo conforman las extensas llanuras bajas, ubicadas en el oriente en las regiones de Orinoquía y Amazonía, así como también, las llanuras de las costas Pacífica y Caribe (al Occidente y Norte del país respectivamente). El último sector corresponde al sistema periférico, el cual cubre los sistemas montañosos aislados, como el de la sierra Nevada de Santa Marta y la Macarena [20]. En la figura 4-1 se muestra una representación con la ayuda de un SIG del relieve Colombiano donde se pueden apreciar las zonas descritas anteriormente.

4.2. Clases objetivo para el problema de clasificación

Para llevar a cabo los objetivos de esta investigación se han seleccionado ocho tipos de relieve que describen adecuadamente la geo-morfología de la zona de estudio. Esta categorización fue realizada manualmente asistida por expertos geomorfológicos de la Universidad Nacional de Colombia - Sede Medellín. Estos expertos poseen un alto grado de conocimiento sobre los procesos de generación y moldeamiento del relieve en todo el territorio Colombiano. Las ocho unidades geomorfológicas seleccionadas se presentan en la tabla 4-1, cada una de ellas con una pequeña descripción de sus características principales.

El problema de clasificación que se plantea en este trabajo tendrá entonces como objetivo

Tabla 4-1: Geo-formas seleccionadas para la realización de este estudio. [3]

Unidad	Nombre de la geo-forma	Descripción
1	Vertientes.	Agrupar gran parte de las unidades geomorfológicas de baja, media y alta montaña como: Superficies tabulares, glaciares, altiplanos, etc.
2	Cañones	Depresiones de profundidad, con respecto a las divisorias, superiores a 100 m a lo largo de pequeñas distancias horizontales.
3	Pie de monte abierto.	Pendientes entre 1° y 6° . Cambios fuertes de pendientes; es decir, pasar de pendientes mayores de 20° a pendientes menores de 6° .
4	Terrazas y colinas.	Zonas generadas por procesos fluviales que ya no son alcanzados por el río.
5	Zona litoral.	Lugares cercanos a la costa con valores de altura menores a los 10 msnm en la costa Pacífica y menores de 5 msnm en la costa Atlántica.
6	Ciénagas y llanuras aluviales con control fluvial.	Pendientes medias inferiores a 5°. Distancias mayores a 30 km a partir de la costa.
7	Ciénagas y llanuras con control del mar.	Pendientes medias inferiores a 1°. Alturas menores a los 50 msnm. Distancias horizontales no mayores a los 30 km desde la costa.
8	Deltas y estuarios.	Lugares de confluencia fluvial y marina. Alturas menores a los 20msnm y pendientes medias inferiores a 1°.

mapear la información de las características del terreno en cada punto de la zona de estudio hacia alguna de las unidades geo-morfológicas descritas con el menor error posible con respecto a la realidad geo-morfológica del terreno. Para ilustrar mejor cuales son los tipos de relieve elegidos para la clasificación la figura 4-2 muestra fotografías aéreas de algunas de ellas en donde se puede apreciar sus características.

Es importante tener en cuenta que la delimitación de las unidades geo-morfológicas propuestas en el presente trabajo fue asistida por personas calificadas que tienen amplio conocimiento de la geografía Colombiana y de los procesos que han generado los tipos de relieve presentes. De la experiencias de los expertos geo-morfólogos han surgido una serie de reglas empíricas que son utilizadas en la actualidad para realizar cierto tipo de clasificación automática, estas reglas no son absolutas y pueden variar con facilidad a través de los diferentes puntos de vista de quien realiza la clasificación. Estas reglas son aplicadas sobre variables que describen la forma que tiene el terreno en un área específica. Algunos de esos criterios para las geo-formas seleccionadas en este trabajo se resumen en la tabla 4-2.



(a) Delta del río Atrato



(b) Ciénaga en el río Magdalena



(c) Laderas y cañones en alta montaña



(d) Colinas en el departamento de Risaralda

Figura 4-2: Imágenes aéreas de algunas de las principales unidades geo-morfológicas de la geografía Colombiana

Tabla 4-2: Criterios para delimitar unidades geomorfológicas. [3]

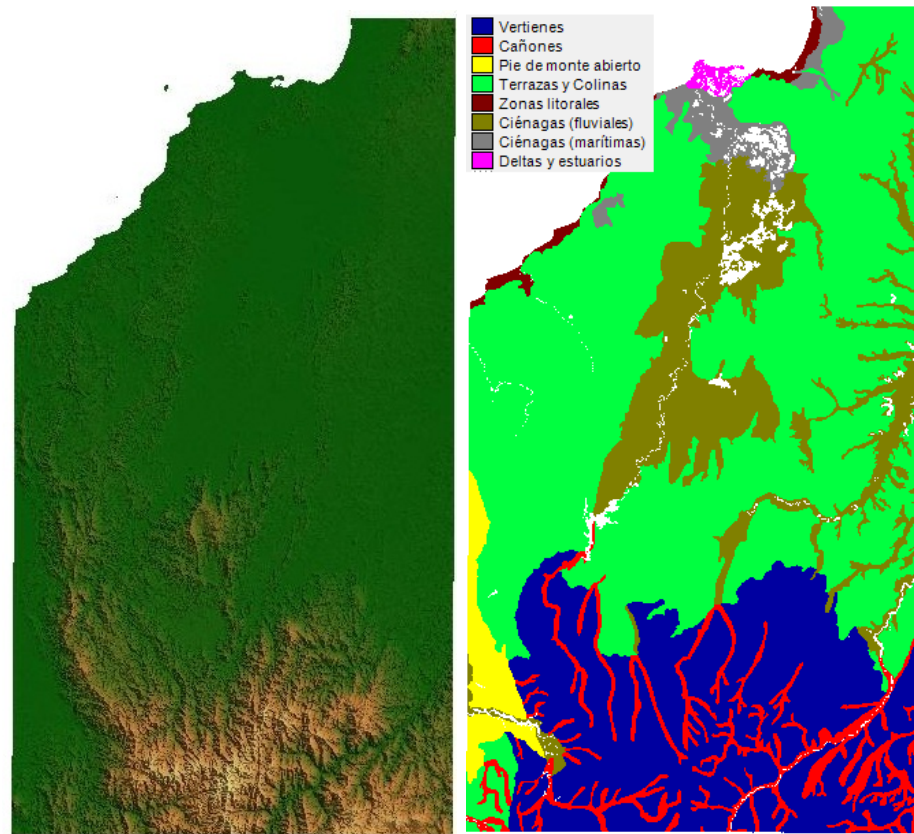
Unidad	Criterios		
	Altura (msnm)	Pendiente	Otros
Zona Litoral Caribe	< 5		Longitudes <1 km de la costa
Zona Litoral Pacífico	< 10		Longitudes <1 km de la costa
Deltas y Estuarios	< 20	< 1°	
Ciénagas interiores y llanuras aluviales con control del mar	< 50	< 1°	Longitudes horizontales < 30 km de la costa
Ciénagas interiores y llanuras aluviales con control Fluvial		< 5°	Longitudes horizontales > 30 km de la costa
Piedemonte Abierto		1° a 6°	Cambios fuertes de pendientes pasar 20° o más a 6° o menos
Piedemonte Confinado		1° a 6°	Cambios fuertes de pendientes pasar 20° o más a 6° o menos
Cañones	Cambios relativos: >100 m		
Montaña		> 10°	

4.3. Caso de estudio

La información fuente para los experimentos de este trabajo corresponde a una sección del modelo de elevación digital de la cuenca hidrográfica del río Sinú. Esta cuenca está ubicada en Colombia entre los departamentos de Antioquia y Córdoba y cuenta con un área de alrededor de 13700 km^2 . El río Sinú discurre de sur a norte, comenzado su recorrido en el “Nudo del Paramillo” (Antioquia) a 3000 msnm en la una zona alta montaña de la Cordillera de lo Andes occidental [51].

Pese a que existen diferentes tipo de de relieve en la geografía de la cuenca, la zona de estudios fue elegida de manera que las ocho geoformas predominantes fueran aquellas definidas en la tabla 4-1 y explicadas previamente. En la figura 4-3 muestra el mapa de elevaciones de la cuenca del río Sinú y sus alrededores (izquierda) con su respectiva clasificación (derecha) en términos de las ocho unidades elegidas. El mapa con con las clasificaciones es también llamado “mapa geomorfológico”.

La clasificación del caso de estudio fue obtenida a través de información de la geo-morfología de la zona almacenada en formato vectorial (ESRI Shapefile) *cita requerida* y extraída del proyecto [3]. Esta información vectorial fue transformada a formato Raster (ESRI Grid) *Cita requerida* para tener los datos organizados en forma de matriz bidimensional de forma que en cada celda de la matriz se almacena una etiqueta d que toma los valores $d = 1, 2, 3, \dots, 8$, los valores de d corresponden a cada una de las clases objetivo.



(a) Elevaciones de la cuenca del río Sinú (b) Mapa geomorfológico de la cuenca del río Sinú

Figura 4-3: Cuenca del río Sinú.

4.3.1. Conjunto de datos para la clasificación

El conjunto de datos para la clasificación de las unidades geo-morfológicas está constituido por un modelo digital de terreno de la zona de estudio con siete mapas Raster geográficamente referenciados que corresponden a las diferentes características morfológicas del terreno. Cada mapa tiene un tamaño de 3207 filas por 1751 columnas que abarcan un poco más del área de toda la cuenca del Sinú. Cada celda de estos mapas representa un área de 92 m^2 . Los siete mapas con una pequeña descripción de cada uno se resumen en la tabla 4-3.

Todos los mapas fueron generados con la ayuda del software ESRI ARCGIS y almacenados en formato ASCII Grid. Este formato es ampliamente utilizado y compatible con otros sistemas de información geográfica (SIG) y con otros software capaces de manejar información geoespacial.

Tabla 4-3: Mapas del modelo de elevación de la cuenca del Río Sinú

	Nombre	Unidades	Descripción
1	Elevaciones	msnm	Alturas en metros sobre el nivel del mar.
2	Perfil de curvatura	NA	Medida de la convexidad, concavidad o planaridad de una región en la dirección de máxima pendiente. Se corresponde con la segunda derivada.
3	Curvatura plana	NA	Medida de la convexidad, concavidad o planaridad de una región en la dirección perpendicular a la de máxima pendiente. Se corresponde con la segunda derivada.
4	Pendientes en grados	Grados °	Inclinación máxima del terreno en un punto.
5	Tangente de las pendientes	NA	Valor de la tangente para el ángulo de inclinación máxima del terreno.
6	Índice topográfico compuesto (CTI)	NA	Métrica que relaciona el tamaño del área de drenaje una cuenca en un punto del mapa con la pendiente en ese mismo punto.
7	Height Above Nearest Neighbor (HAND)	msnm	Altura del terreno en metros calculada con respecto al punto de drenaje más cercano [45, 50].

4.3.2. Subdivisión del conjunto de datos

Debido al tamaño de la zona de estudio y a la resolución de los mapas Raster que contiene la información geo-morfológica se decidió dividir el caso de estudio en varias secciones que fueran representativas de las ocho geo-formas objetivo presentes en el área. Se eligieron siete regiones diferentes dentro del área de estudio; cada una de ellas representa un rectángulo de 451×451 celdas de cada mapa Raster, lo que equivale a un área de $1721,5869 \text{ km}^2$ con un total de $n = 203401$. Cada uno de los descriptores geo-morfológicos así como el mapa con las etiquetas de las ocho clases objetivo fue dividido de tal forma que se consolidaran siete nuevos conjuntos de datos, $m = 7$, cada uno de ellos con los ocho descriptores de la tabla 4-3 de la región que representan, así como también con el conjunto de etiquetas de las clases presentes allí.

En el caso de estudio las diferentes geo-formas se distribuyen de manera que una solo geo-forma agrupa un área bastante grande en comparación con la resolución de los mapas. En esta distribución algunas geo-formas se encuentran cerca de otras; sin embargo, ocurre que algunas de ellas se encuentran bastante alejadas unas de otras. También sucede que el área que corresponde a algunas de ellas es bastante mayor que el área correspondiente a otras.

Lo anterior introduce un problema conocido en reconocimiento de patrones como inbalance de clases [21]. Otro problema derivado de esta situación es la imposibilidad de definir una región dentro del caso de estudio que contenga al mismo tiempo todas las clases objetivo.

La subdivisión realizada en los siete subconjuntos de datos soluciona los dos problemas mencionados. Por un lado la suma de los datos de cada uno de los subconjuntos de datos es menor que la cantidad de datos del caso de estudio en total. Además, las regiones que

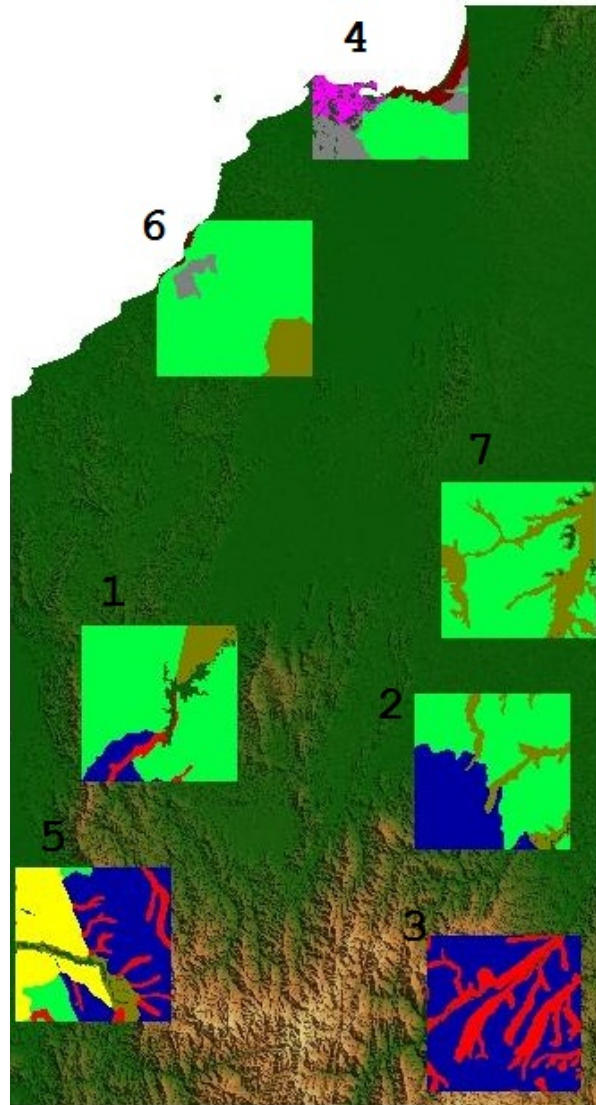


Figura 4-4: Representación de los 7 regiones de los subconjuntos de datos sobre el mapa de elevaciones del caso de estudio.

representan cada subconjunto fueron elegidas estratégicamente para que en conjunto las regiones resumieran las ocho geo-formas que se desean clasificar. Finalmente, el problema de inbalance de clases se mitiga debido a que cada región contiene en proporciones un poco más balanceadas las cantidades de celdas que pertenecen a una geo-forma u otra.

Para crear los subconjuntos de datos se diseñaron varias rutinas en el software de cálculo MatLab©. Estas rutinas extraen los subconjuntos a partir de un punto que se considera con el centro de la región (celda central), y un parámetro r que corresponde a un radio al rededor del centro que definirá el tamaño de la región. Las rutinas desarrolladas también guardan la información acerca de donde está ubicado el subconjunto de datos con respecto al caso de estudio global, de esta manera es posible relacionar los cada subconjunto con la

Tabla 4-4: Descripción de los subconjuntos de datos para la clasificación

Región	Nombres de la geo-formas	Número de geo-formas
1	Vertientes, cañones, terrazas y colinas, ciéngas y llanuras aluviales con control fluvial.	4
2	Vertientes, terrazas y colinas, ciénagas y llanuras aluviales con control fluvial.	3
3	Vertientes y cañones.	2
4	Terrazas y colinas, zona litoral, ciénagas y llanuras aluviales con control del mar, deltas y estuarios.	4
5	Cañones, vertientes, pie de monte abierto, terrazas y colinas, ciénagas y llanuras aluviales con control fluvial.	5
6	Terrazas y colinas, zona litoral, ciénagas y llanuras aluviales con control fluvial, ciénagas y llanuras aluviales con control del mar.	4
7	Terrazas y colinas, ciénagas y llanuras aluviales con control fluvial.	2

totalidad del caso de estudio.

La gráfica 4-4 muestra cuales son y donde están ubicadas las siete regiones correspondientes a la división del caso de estudio. Cada región en la imagen muestra las etiquetas de las clases que ella pertenecen; los colores coinciden con los de la gráfica 4-3. Las unidades geomorfológicas presentes en cada una de las regiones se muestran en la tabla 4-4.

Las diferentes geo-formas elegidas como clases para el problema de clasificación de este estudio resumen adecuadamente las características geo-morfológicas de la zona de estudio según criterios de expertos en el tema[3]. La división realizada a la información del caso de estudios permitió un mejor tratamiento de la información en términos de costo computacional, debido al menor tamaño de cada subconjunto de datos; y permitió además eludir algunos problemas que se presentan en reconocimiento de patrones como el inbalance de clases.

5 Metodología de Clasificación

En este capítulo se describe la metodología utilizada para llevar a cabo la clasificación de las ocho unidades geo-morfológicas anteriormente descrita. Se propone y detalla el proceso de extracción de nuevas características a partir de información de la textura del terreno, usando los descriptores de textura de Haralick. Se explicará como fueron formulados los experimentos de clasificación y con que datos fueron ejecutados.

La motivo fundamental para utilizar una técnica de extracción de características de textura surge principalmente de la necesidad de consolidar un conjunto de descriptores más amplio que el que se tenía originalmente, y estaba constituido sólo por características morfológicas. Estas características morfológicas demostraron según los análisis realizados una precaria separabilidad de las ocho unidades geo-morfológicas objetivo. Se analizaron los histogramas y algunas de las gráficas de dispersión para varias combinaciones de ellos. Algunas de estas gráficas se muestran en las figuras 5-1 a 5-4. Se resalta el caso particular de los histogramas de los descriptores perfil de curvatura y curvatura plana, en donde las clases tienen la peor separación con distribuciones muy similares y medias prácticamente idénticas.

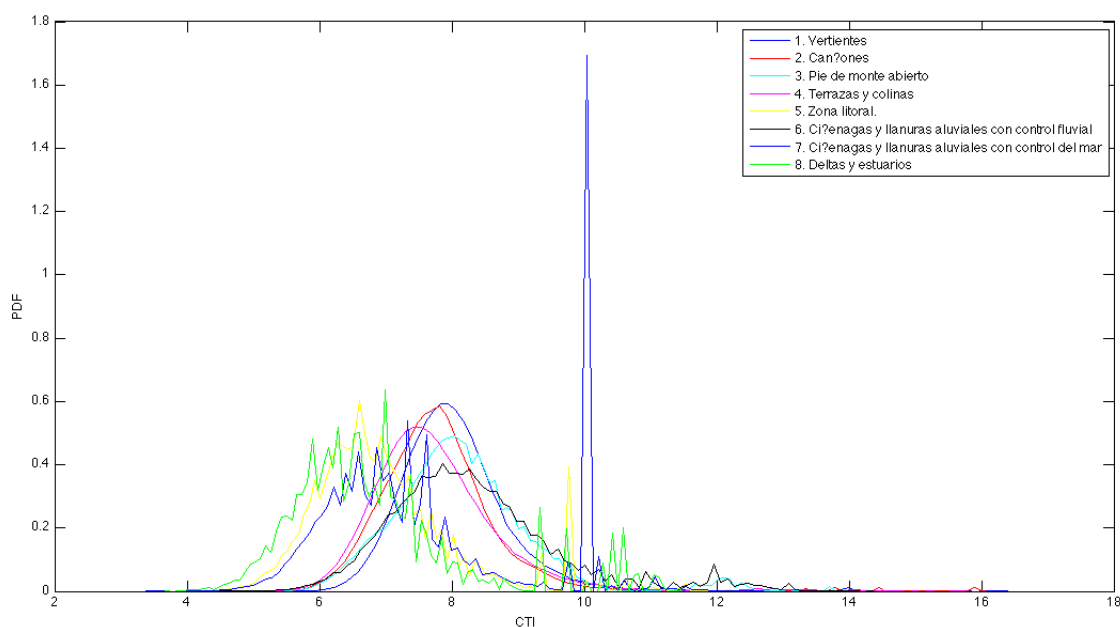


Figura 5-1: Histograma del descriptor: CTI, para las 8 clases.

Es bien sabido en la literatura que un problema de clasificación con pocas clases es menos complejo que un problema en el que el número de clases objetivo es muy elevado; lo anterior

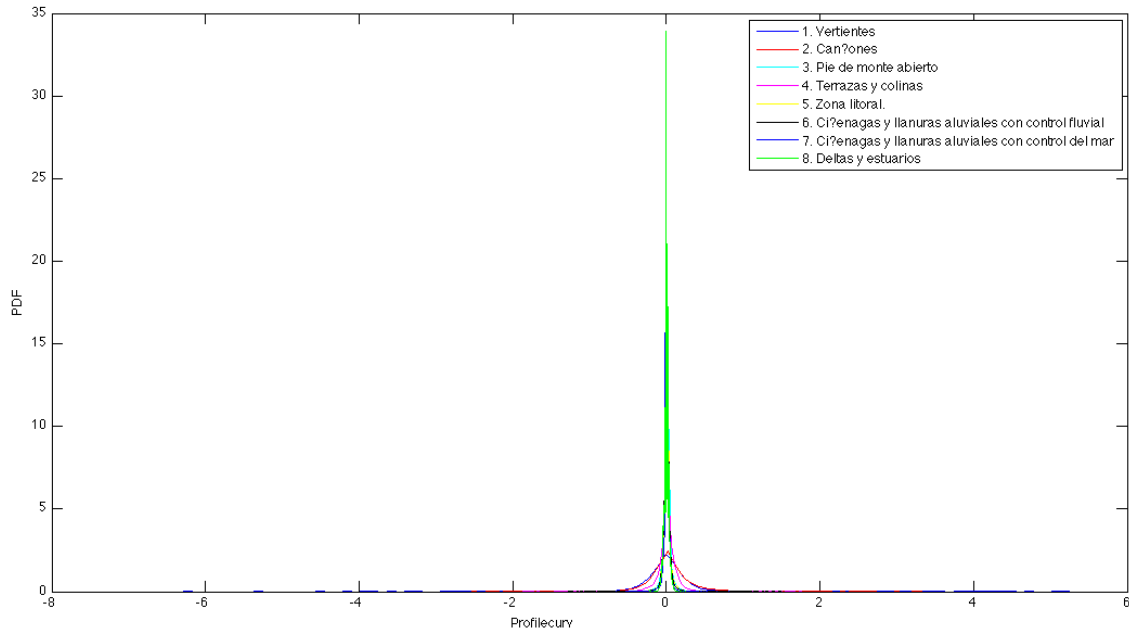


Figura 5-2: Histograma del descriptor: Perfil de curvatura, para las 8 clases.

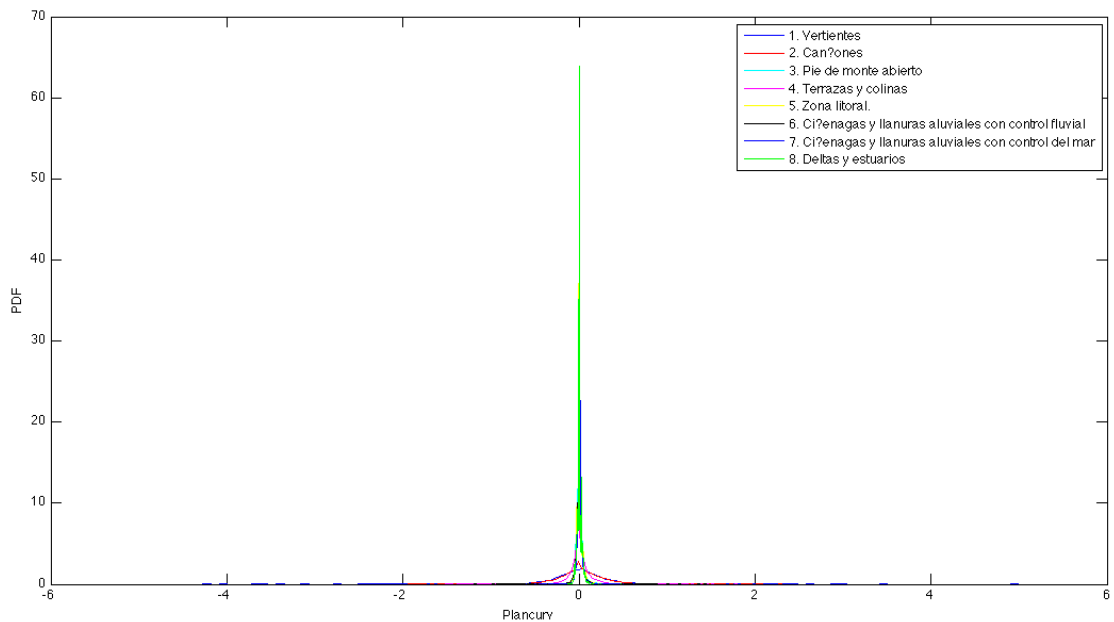


Figura 5-3: Histograma del descriptor: Curvatura plana, para las 8 clases.

debido a que para la mayoría de los clasificadores la clasificación binaria es un caso especial y trivial de la clasificación multi clase [34, 18].

Por lo anterior en esta investigación se propone llevar a cabo una división inicial del conjunto de clases objetivo, pasando de un problema de 8 clases a dos problemas de 2 clases y 6 clases respectivamente. Esta agrupación se realiza de acuerdo a criterios empíricos que sugie-

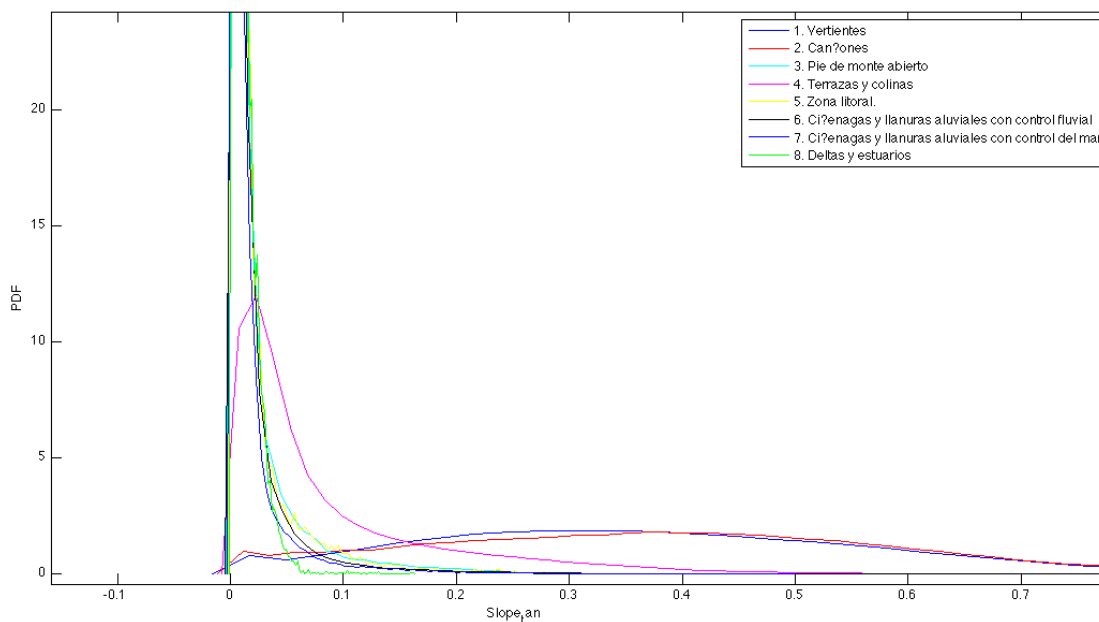


Figura 5-4: Histograma del descriptor: Pendientes en grados, para las 8 clases.

ren que pueden obtenerse mejores resultados agrupando geo-formas que comparten ciertas características. En las siguientes secciones se explicara como se llevó a cabo tal división.

5.1. Criterio de segmentación

En esta sección se establece un criterio lo más simple posible para dividir las 8 clases del caso de estudio en dos super clases que agrupen las diferentes geo-formas existentes. La definición de este criterio fue empírica, formulado a partir de la observación de los mapas que representan los descriptores, así como también el análisis de los histogramas y gráficos de dispersión de los descriptores para las clases objetivo.

5.1.1. División del área de estudio

El primer aspecto determinante para este proceso fue que las geo-formas contenidas en una super clase guardaran una relación espacial entre si; es decir, que deben estar geográficamente cerca unas de otras. Esto conduce a que cada super clase sea una gran región continua de un subconjunto de geoformas.

La cuenca del río Sinú, que constituye el caso de estudio, discurre de sur a norte, pasando de una zona montañosa alta en la cordillera de los Andes, a una región baja conformada por un valle amplio para luego desembocar en el mar. Teniendo en cuenta lo anterior y las características de las geo-formas en las tablas 4-1 y 4-2; se concluyó que existe una división natural entre geo-formas que aparecen en zonas montañosas y otras que aparecen en zonas

de planicies y valles amplios.

Con estos dos conceptos en mente se decidió establecer dos super clases para dividir el terreno correspondiente al área de estudio. La super clase 1 intuitivamente pertenece a zonas altas como cordilleras y serranías, y la super clase 2 a zonas bajas y planas como los valles de grandes ríos.

5.1.2. Definición del criterio de segmentación

Dos super clases fueron definidas, la super clase 1 contiene las unidades geo-morfológicas:

- Vertientes.
- Cañones.

La super clase 2 contiene el resto de geo-formas que son:

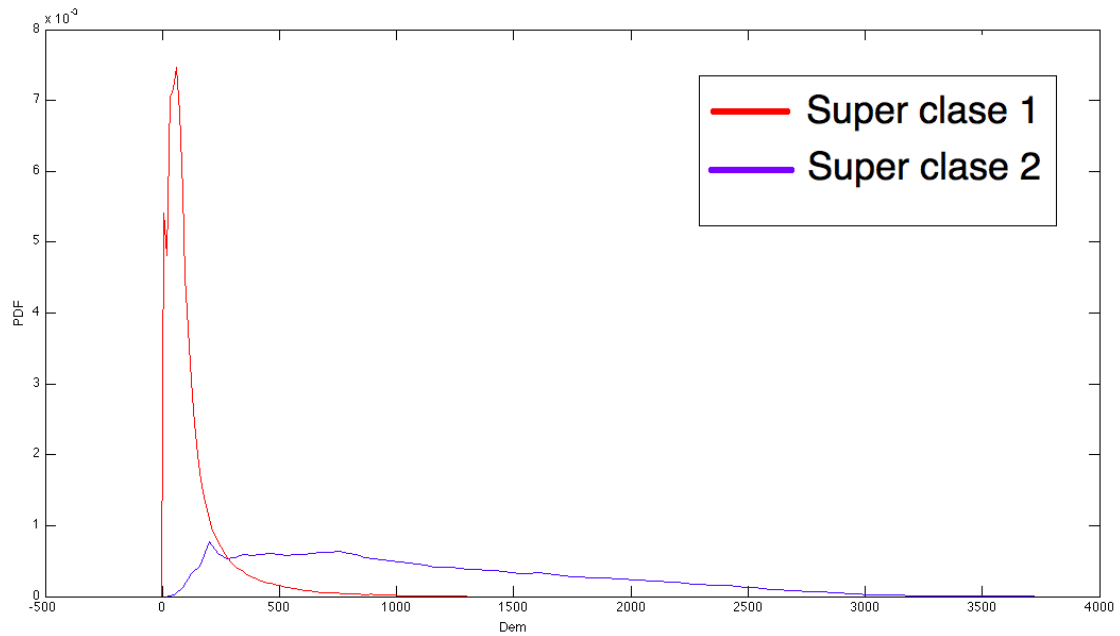
- Pie de monte abierto.
- Terrazas y colinas.
- Zona litoral.
- Ciénagas y llanuras aluviales con control fluvial.
- Ciénagas y llanuras con control del mar.
- Deltas y estuarios.

Se analizaron todas las combinaciones posibles de los descriptores morfológicos para determinar cual o cuales de ellos mostraban una buena separabilidad una vez agrupadas las geo-formas en las super clases. Para ello se utilizó el algoritmo de búsqueda exhaustiva con criterios de separabilidad, como Fisher y KNN con $k = 5$. Este algoritmo se encuentra en el paquete de selección de características implementado en el toolbox Balú[43] implementado en MatLab©.

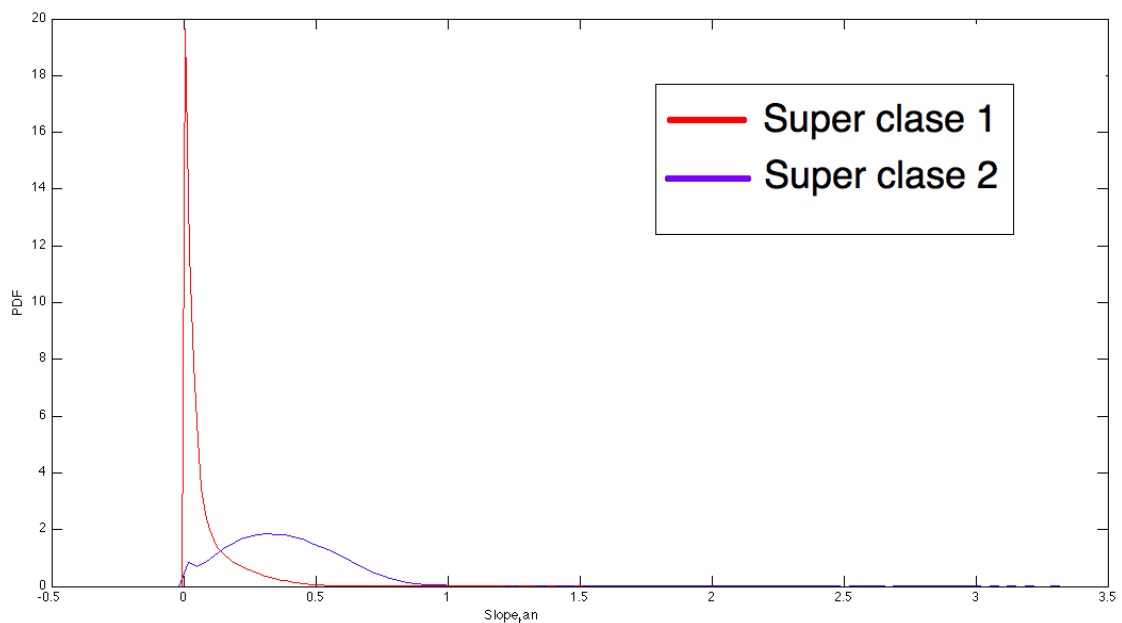
Los resultados obtenidos son subconjuntos de los descriptores: Elevaciones, pendientes en grados, tangente de la pendiente y HAND. La figura 5-5 muestra los histogramas de las elevaciones y las pendientes en grados, donde puede apreciarse que hay una mayor separabilidad de las clases agrupadas, en comparación con los histogramas de la figuras 5-1, 5-2, 5-3 y 5-4.

Con cada uno de los subconjuntos que arrojó la búsqueda exhaustiva se formuló un experimento de clasificación para indicar que individuos pertenecen a una super clase u otra.

Para fortalecer el criterio de segmentación se utilizó el mecanismo de Análisis de componentes principales para extraer nuevos descriptores útiles a partir de los subconjuntos encontrados por el algoritmo de búsqueda exhaustiva. Para elegir las componentes principales



(a) Histogramas de las elevaciones para las 8 clases agrupadas en las super clases.



(b) Histogramas de las pendientes para las 8 clases agrupadas en las super clases.

Figura 5-5: Descriptores morfológicos una vez agrupadas las clases en las dos super clases definidas.

que se incluirán en el conjunto de datos, se estableció un valor de variabilidad acumulada de $\lambda_{acumulada} = 99,9\%$.

El desempeño de un proceso de clasificación automática se mide contando el número de individuos de la muestra clasificada que fueron bien etiquetados, es decir, aquellos que fueron clasificados correctamente en su respectiva clase; este valor se divide por el número de muestras totales, para obtener un nuevo valor en el intervalo $p = [0, 1]$ que indica que tan buena fue la clasificación. Un valor de $p = 1$ sugiere una clasificación perfecta.

Es importantes resaltar que una vez realizada la división en super clases con el criterio establecido, el desempeño final de la clasificación de las 8 geo-formas dependerá no solo de la clasificación individual de cada una de las unidades geo-morfológicas dentro de cada super clase; sino también del desempeño obtenido en la separación en super clases. El error en esta etapa se propaga a través de las siguientes como sucede siempre en problemas de clasificación jerárquica.

5.1.3. Super clase 1

La super clase 1 comprende las geoformas 1 y 2 que corresponden a las vertientes y cañones. Estas geo-formas se agrupan comúnmente en zonas montañosas como las cordilleras. En el caso de estudio sucede que la zona sur de la cuenca del río Sinú y sus alrededores se ubica en el extremo norte de la cordillera occidental por lo que las unidades geo-morfológicas que corresponden a las clases 1 y 2 se agrupan homogéneamente en esta región. Ver figura 4-3b. Como se explicó en el capítulo 4 los datos del caso de estudio de esta investigación se dividieron en subconjuntos para facilitar su análisis; sin embargo, a raíz del tamaño de estas divisiones no es posible tener un subconjunto que posea las 8 clases al tiempo. Por lo tanto, para realizar la tarea de clasificar las geo-formas de la super clase 1 se deben utilizar los subconjuntos de datos que contienen las clases vertientes y cañones, que son el 1, 2, 3 y 5. Ver figura 4-4. El subconjunto de datos 3 es el más adecuado para la clasificación pues sólo contienen instancias de las dos clases requeridas.

Un aspecto particular de la super clase 1 es que representa un problema de clasificación binaria menos complejo que el problema inicial de 8 clases, e incluso menos complejo que el de la super clase 2.

5.1.4. Super clase 2

Las unidades geo-morfológicas de la super clase 2 incluyen los dos tipos de ciénagas, las zonas litorales, los deltas y estuarios, y los pie de monte. Se ubica en el caso de estudio básicamente en zonas bajas y planas cerca de la desembocadura del río Sinú. La clase pie de monte abierto; sin embargo, se halla cerca de la cordillera, justo donde esta termina y comienza la planicie. Las ciénagas son zonas aledañas al río que normalmente están inundadas por su cercanía al cauce además de otros fenómenos geo-morfológicos. Las terrazas y colinas

rodean comúnmente a las ciénagas y son zonas que habitualmente no se inundan.

Los datos para realizar la clasificación de la super clase 2 se extrajeron de los subconjuntos de datos 4, 5 y 7. Ver figura 4-4. A diferencia de la super clase 1, en esta clasificación no fue posible utilizar un solo subconjunto de datos que tuviera todas las clases requeridas; por lo que se realizó un muestreo en la información contenida en 4, 5 y 7 para extraer los datos que se usaron para los experimentos. En el capítulo 6 se ofrecen mas detalles de esta situación.

Una vez realizada la separación de las 8 clases iniciales en sub problemas de menos clases se logra reducir la complejidad de cada uno de las clasificaciones individuales. No obstante, aún hay un problema restante, que es la necesidad de encontrar nuevos descriptores a parte de los morfológicos iniciales que permitan separar las clases adecuadamente.

5.2. Extracción de características basadas información de texturas

La hipótesis fundamental de este trabajo es que se puede mejorar el desempeño de la clasificación de tipos de relieve a través del cálculo de características basadas en la textura del terreno; estas texturas son extraídas a partir de los descriptores morfológicos del terreno descritos en la tabla 4-3.

Existen en la literatura varias metodologías de extracción de texturas ampliamente utilizadas en áreas como visión por computador, procesamiento digital de imágenes, reconocimiento de voz, etc. Entre las más destacadas se encuentran los métodos de extracción de texturas de Gabor[49], Local Binary Pattern[2] y las texturas de Haralick[24, 49].

5.2.1. Descriptores de textura de Haralick

La matriz de concurrencia descrita en el capítulo 2 contiene en su estructura información sobre los patrones de textura; esta información se guarda en ella contabilizando cuantas veces se repite una combinación de valores dentro de una imagen en una dirección dada de 8 posibles. Haralick presenta 14 métricas que son obtenidas realizando diferentes operaciones sobre la COM. La tabla 5-1 muestra cada uno de los descriptores propuestos por Haralick.

Pese a que estas métricas de textura fueron desarrolladas para analizar texturas en imágenes, su conceptualización es trasladable al caso de los mapas raster georeferenciados puesto que al tener estos últimos la forma de una matriz bidimensional coincide con la definición discreta de una imagen, acomodándose a los cálculos y mecanismos de la metodología. Previamente es necesario discretizar los datos de cada mapa en valores enteros para que puedan ser tratados como si fueran niveles de gris en una imagen.

Tabla 5-1: Lista de los descriptores propuestos por Haralick.

Id	Nombre
1	Angular Second Moment
2	Contrast
3	Correlacion
4	Sum of squares
5	Inverse Difference Moment
6	Sum Average
7	Sum Entropy
8	Sum Variance
9	Entropy
10	Difference Variance
11	Difference Entropy
12	Information Measures of Correlation 1
13	Information Measures of Correlation 2
14	Maximal Correlation Coefficient

5.2.2. Extracción de texturas a partir de descriptores geomorfológicos

Para cada una de las 14 métricas de Haralick de la tabla 5-1 se generó un nuevo mapa (descriptor) del mismo tamaño que el mapa original, donde el valor de cada celda es la métrica calculada en una ventana cuadrada con $lado = 2r + 1$, centrada en esa celda. r es un parámetro del algoritmo de extracción de texturas que representa un determinado número de celdas.

Cada uno de los siete descriptores morfológicos contienen valores reales, con excepción de las elevaciones y el HAND; por lo tanto, para poder aplicar la metodología de texturas tal y como fue definida para el caso de imágenes, es necesario discretizar sus valores y escalarlos. Este escalamiento tuvo además el propósito de reducir dimensiones de las matrices de co-ocurrencia evitando al máximo la pérdida de información. Por lo anterior, los valores de todos los descriptores morfológicos fueron escalados y discretizados en dos diferentes rangos:

$$d_1 = [0, 7] , d_2 = [0, 15] \quad (5-1)$$

Esta situación da como resultado matrices de co-ocurrencia de 8x8 y 16x16 respectivamente, por cada una de las 14 métricas de textura, por cada uno de los descriptores morfológicos.

Análogamente fueron elegidos varios valores para el tamaño de la ventana sobre la que se extrajo la información de texturas. El rango de valores de r fue:

$$r = [10, 15, 25, 30, 50] \quad (5-2)$$

Dado que cada celda en los descriptores representa un área de 92 m^2 , se entiende entonces que las ventanas representan áreas de alrededor de $3,5 \text{ km}^2$, $7,6 \text{ km}^2$, 21 km^2 , 30 km^2 y $82,6 \text{ km}^2$ respectivamente.

La ejecución del algoritmo de extracción de texturas bajo cada combinación de los parámetros descritos anteriormente entregó un total de 980 descriptores adicionales. Con el fin de facilitar el análisis de los resultados de la clasificación con los nuevos descriptores en etapas posteriores de este trabajo, se estableció en la ecuación 5-3 una nomenclatura para denominar a cada uno de ellos.

$$H^{(\#)}_{\mathbf{f}\#, \mathbf{r}\#, \mathbf{d}\#} \quad (5-3)$$

En dicha nomenclatura el superíndice indica una de las 14 métricas de textura de la tabla 5-1, f se refiere al descriptor morfológico de la tabla 4-3 al que se le extrajo la información de textura, r representa el tamaño de la ventana utilizada y d indica el tamaño de la matriz de co-ocurrencia. Esta nomenclatura será utilizada en el resto del documento.

Para ejemplificar los resultados obtenidos la figura 5-6 muestra algunas de las características de Haralick para una ventana con $r = 10$, $d = 8$, que corresponden al mapa de elevaciones. Se puede observar en la figura que algunos de ellos resaltan adecuadamente las clases objetivo. En la figura también se muestran las etiquetas correspondientes a esta región, así como una imagen generada en un SIG del descriptor perfil de curvatura.

Finalmente, el algoritmo básico de Haralick fue optimizado para este problema gracias a una propiedad de la matriz de co-ocurrencia que permite que su cálculo para una ventana centrada en una celda $C_{i,j}$ se realice a partir de pequeñas modificaciones a la matriz de co-ocurrencia para una ventana de igual tamaño y forma centrada en la celda $C_{i,j-1}$ como se muestra a continuación:

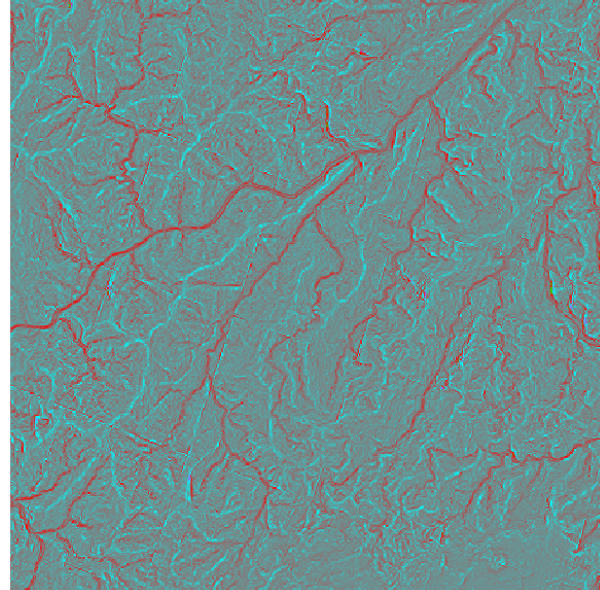
$$C_{i,j} = C_{i,j-1} - C_{izquierda}_{i,j-1} + C_{derecha}_{i,j} \quad (5-4)$$

Donde $C_{i,j}$ es la matriz de co-ocurrencia de una ventana centrada en la celda (i, j) del mapa, $C_{izquierda}_{i,j-1}$ es la matriz de co-ocurrencia de la primera columna de la ventana centrada en $(i, j - 1)$, y $C_{derecha}_{i,j}$ es la matriz de co-ocurrencia de la última columna de la ventana centrada en la celda (i, j) . De esta manera se eliminan operaciones redundantes, y se agiliza el proceso de cálculo.

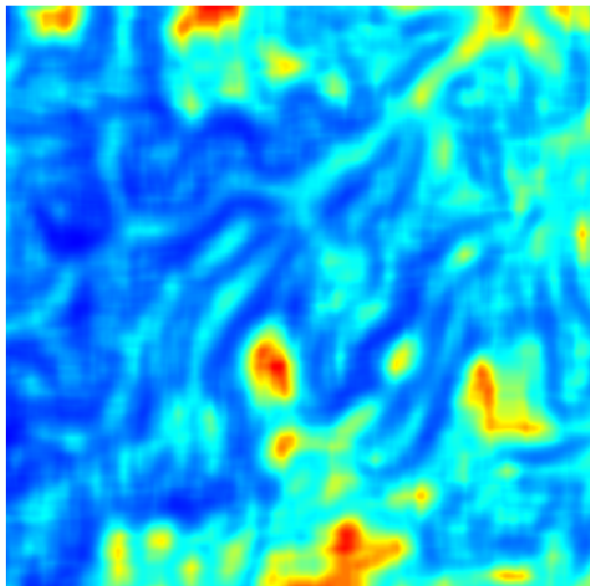
Es importante anotar que el proceso de selección fue exhaustivo y costoso computacionalmente debido principalmente al tamaño de los conjunto de datos y a la gran cantidad de combinaciones de parámetros.



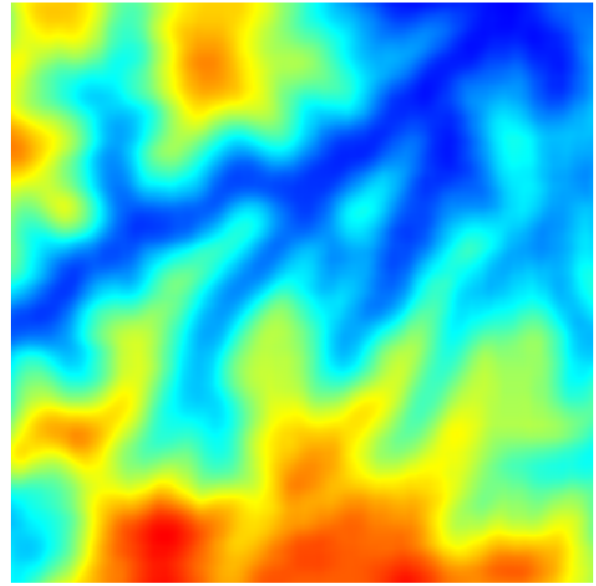
(a) Conjunto de etiquetas para el modelo



(b) Descriptor: Perfil de curvatura



(c) Descriptor $H_{fl,r10,d08}^{(1)}$



(d) Descriptor $H_{fl,r10,d08}^{(6)}$

Figura 5-6: Ejemplo de algunos de los descriptores de textura extraídos.

6 Experimentos y Resultados

En este capítulo se presentan los detalles del diseño e implementación de los experimentos realizados utilizando la metodología de la sección 5. También se resumen los principales resultados mostrando las tablas con los desempeños de las diferentes clasificaciones llevadas a cabo; y se comparan los resultados de la clasificación usando descriptores de textura con la clasificación usando solo los descriptores morfológicos reportados en la literatura.

6.1. Algoritmos, parámetros y selección de características

Una vez que se han extraído los descriptores de las diferentes geo-formas a partir de la información de la textura del terreno, el paso siguiente en el proceso de clasificación es seleccionar cuales de esos descriptores son los más adecuados en términos de que tan bien separan las clases objetivo. Cuando los mejores descriptores para este propósito han sido elegidos, se procede con la ejecución de diferentes pruebas utilizando varios clasificadores, comparando sus desempeños para evaluar los resultados.

6.1.1. Selección de características

El conjunto de datos para la clasificación de las unidades geo-morfológicas contenidas en cada una de las super clases queda definido como la unión de los descriptores morfológicos con los descriptores basados en texturas de Haralick. Para separar las geo-formas sólo se usaron los descriptores morfológicos debido a que se demostró en el capítulo 5 que estos junto con sus componentes principales separan bien las dos super clases definidas.

Una vez consolidados los datos, el proceso de clasificación continua seleccionando varios subconjuntos de todos los descriptores disponibles que ofrezcan una mejor separación de las clases objetivos. Para ello se usó el algoritmo de selección de características conocido como SFS [40, 33], Sequential Forward Search.

El objetivo del algoritmo SFS es encontrar los k primeros descriptores que maximizan la separabilidad de las clases de acuerdo a un criterio establecido (criterios wrapper[38] y filter[23]). SFS entrega los descriptores seleccionados en orden de mayor a menor grado de separabilidad; teniendo en cuenta que la separabilidad conjunta de los l primeros es menor que la separabilidad de los $l + 1$ descriptores. Es decir que, el primer descriptor entregado por el algoritmo es aquel que mejor resultado ofrece por si solo al separar las clases; el segundo

descriptor entregado es aquel que en conjunto con el anterior (el primero) ofrece una mejor separación que el primero solo; y de esta manera hasta alcanzar el número de features requeridos, o no encontrar alguno que mejores los resultados.

Para este trabajo se usó la implementación del algoritmo SFS que se encuentra en el toolbox Balú [43], y se estableció un valor de $k = 40$ como máximo número de características para cuatro criterios de separabilidad [39, 34]. Estos criterios se eligieron por ser rápidos y comúnmente utilizados en otros estudios sobre clasificación [49, 22, 29, 44]. Los mecanismos de selección fueron entonces:

Selección 1 (S1): Selección de tipo filter utilizando el criterio de separabilidad de Fisher.

Selección 2 (S2): Selección de tipo wrapping utilizando el clasificador LDA (Linear Discriminant Analysis).

Selección 3 (S3): Selección de tipo wrapping utilizando el clasificador KNN (K-Nearest-Neighbor) con un valor de $k = 5$.

Selección 4 (S4): Selección de tipo wrapping utilizando el clasificador KNN (K-Nearest-Neighbor) con un valor de $k = 7$.

La selección se realizó de manera diferente en el caso de la división del caso de estudio en la super clase 1 y la super clase 2. Para esta separación solo se utilizaron los ocho descriptores morfológicos, que son una cantidad muy menor de características comparada con los descriptores de textura. Por lo anterior, el mecanismo de selección en este caso fue búsqueda exhaustiva [33], que arrojó como resultado la dos mejores selecciones. Adicionalmente se incluyó dos selecciones más que corresponden a los resultados de la extracción de las primeras componentes principales de los descriptores entregados por la búsqueda exhaustiva. Para calcular las componentes principales se utilizó un valor de variabilidad acumulada de $\lambda_{acumulada} = 99,9\%$ de la variabilidad total de los datos. Los resultados obtenidos en esta selección son:

Selección 1: (1 4 5). Elevaciones, Pendientes en grados y tangente de la pendiente.

Selección 2: (1 5 7). Elevaciones, tangente de la pendiente y Height Above Nearest Neighbor (HAND).

PCA(1,4,5): Componentes principales de la selección 1. Se obtuvo una sola componente que resumía el 99,9% de la variabilidad de la selección.

PCA(1,5,7): Componentes principales de la selección 2. Se obtuvieron dos componente que resumía el 99,9% de la variabilidad de la selección.

Tabla 6-1: Descriptores seleccionados para la super clase 1.

Selección	Criterio	Descriptores seleccionados (Solo los primeros 5)	Total
S1	Fisher	$H_{f2,r25,d16}^{(6)}$, $H_{f7,r15,d16}^{(2)}$, $H_{f7,r50,d16}^{(12)}$, $H_{f3,r10,d16}^{(6)}$, $H_{f2,r25,d08}^{(12)}$	40
S2	LDA	$H_{f2,r25,d16}^{(6)}$, $H_{f3,r30,d08}^{(1)}$, $H_{f7,r25,d08}^{(14)}$, $H_{f2,r25,d08}^{(5)}$, $H_{f2,r30,d16}^{(7)}$	27
S3	Knn, $k = 5$	$H_{f2,r25,d16}^{(6)}$, $H_{f7,r50,d08}^{(2)}$, $H_{f4,r25,d08}^{(7)}$, $H_{f6,r25,d08}^{(7)}$, $H_{f7,r25,d08}^{(8)}$	21
S4	Knn, $k = 7$	$H_{f2,r25,d16}^{(6)}$, $H_{f7,r15,d16}^{(2)}$, $H_{f6,r25,d08}^{(7)}$, $H_{f4,r25,d08}^{(7)}$, $H_{f7,r25,d08}^{(6)}$	16

Tabla 6-2: Descriptores seleccionados para la super clase 2.

Selección	Criterio	Descriptores seleccionados (Solo los primeros 5)	Total
S1	Fisher	$H_{f3,r10,d08}^{(6)}$, $H_{f2,r10,d08}^{(6)}$, $H_{f2,r25,d16}^{(6)}$, $H_{f6,r25,d08}^{(7)}$, $H_{f2,r50,d16}^{(1)}$	100
S2	LDA	$H_{f6,r25,d08}^{(3)}$, $H_{f2,r50,d16}^{(5)}$, $H_{f6,r25,d08}^{(13)}$, $H_{f6,r10,d16}^{(8)}$, $H_{f5,r25,d08}^{(6)}$	25
S3	Knn, $k = 5$	$H_{f3,r15,d08}^{(7)}$, $H_{f6,r25,d08}^{(7)}$, $H_{f5,r30,d16}^{(13)}$, $H_{f1,r25,d08}^{(8)}$, $H_{f1,r25,d16}^{(2)}$	12
S4	Knn, $k = 7$	$H_{f3,r25,d16}^{(6)}$, $H_{f6,r50,d16}^{(8)}$, $H_{f5,r25,d08}^{(8)}$, $H_{f4,r25,d16}^{(5)}$, $H_{f6,r25,d08}^{(7)}$	13

En la tabla **6-1** se muestran los diferentes subconjuntos de descriptores entregados por el algoritmo SFS para la super clase 1; y en la tabla **6-2** se muestran los resultados de la selección de características para la super clase 2. Por simplicidad y legibilidad sólo se muestran los 5 primeros descriptores seleccionados; sin embargo se reporta el total entregado por el algoritmo.

Tres aspectos del resultado del proceso de selección de características son resaltados en este punto:

1. Para cada una de las selecciones para la super clase 1, el primer descriptor es siempre el mismo; es decir, el descriptor que ofrece una mejor separabilidad de las clases fue persistente en cada una de las cuatro selecciones realizadas.
2. Para ambas super clases el número total descriptores seleccionados en S3 y S4 es significativamente menor que los obtenidos en S1; donde de hecho, el algoritmo SFS no converge, entregando el máximo número de descriptores pedido. S2 en ambos casos entregó un número menor de descriptores que S1, pero no tan pequeño como S3 y S4.
3. Después de realizar la selección de características la dimensionalidad de los problemas de clasificación bajó considerablemente, de 987 descriptores a 30 en promedio; esto representa una reducción de aproximadamente el 3% de su tamaño inicial; esta situación facilitó el ejercicio de clasificación y pruebas.

El objetivo de la extracción de características de texturas es mejorar la separabilidad de las clases a través de nueva información que tenga en cuenta más elementos del entorno de cada una de las celdas de los mapas que representan el área de estudio. En este orden de ideas, y después de haber realizado la selección, se muestran algunos de los histogramas y gráficas 2-D Y 3-D de los 3 primeros descriptores de las selecciones de características de las super clases. En la figura **6-1** se puede apreciar que los 3 primeros descriptores no ofrecen una separación adecuada de las clases vertientes y cañones; sin embargo las diferentes selecciones de la super clase 1 arrojaron cada una más de 10 descriptores, por lo tanto es posible que una separación más adecuada se pueda hallar en un espacio dimensional mayor.

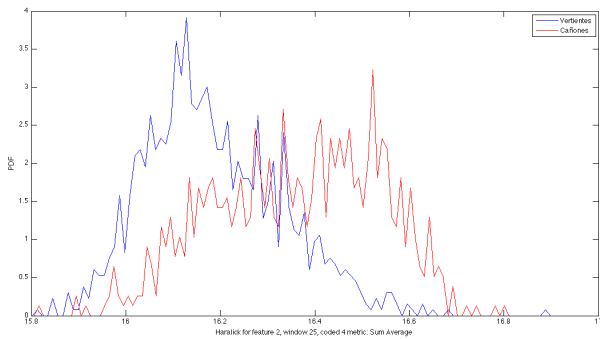
En el caso de la super clase 2 la figura **6-2** muestra que los 3 primeros descriptores de las selecciones ofrecen una buena separación de las 6 clases contenidas en ellas. Las geo-formas pie de monte y ciénagas (fluviales) quedan bastante separadas entre si por estos descriptores; así mismo, una combinación de las clases ciénaga (marítima), zona litoral y deltas y estuarios queda también bien separa del resto. Se espera que la separación total se alcance con el resto de descriptores obtenidos en las selecciones de características para la super clase 2.

En las selecciones para la super clase 2 con el criterio de Fisher los descriptores seleccionados en la posición 7 y 8 son las Elevaciones y el HAND respectivamente, en las selecciones de la super clase 1 el HAND aparece como el descriptor 18 en la selección con Fisher. Estos resultado son interesantes pues son los únicos casos en los que los descriptores morfológicos fueron seleccionados en comparación con el resto de las selecciones, tanto para la super clase 1 como para la super clase 2.

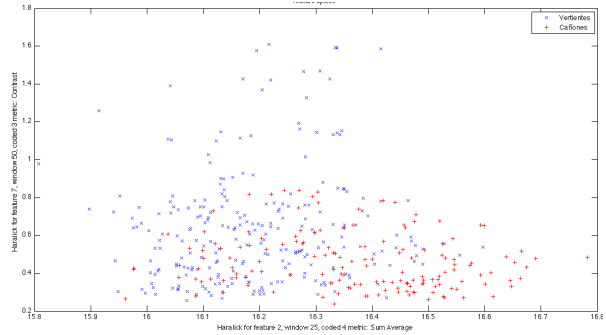
6.1.2. Elección de clasificadores y pruebas de validación

Para la realización experimentos de este trabajo se utilizaron nueve clasificadores de tipo supervisado, de manera que constituyeran un conjunto suficientemente representativo de los diferentes enfoques de reconocimiento de patrones presentes en la literatura. Los clasificadores utilizados para los experimentos con sus respectivos parámetros son:

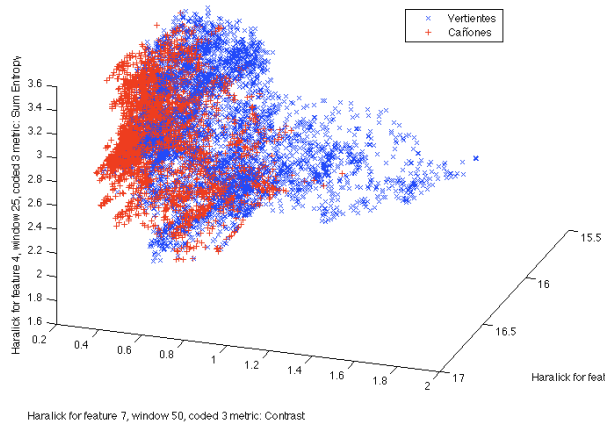
1. K-Nearest-Neighbor con un valor de $k = 5$.
2. K-Nearest-Neighbor con un valor de $k = 7$.
3. K-Nearest-Neighbor con un valor de $k = 9$.
4. LDA (Linear Discriminant Analysis).
5. QDA (Quadratic Discriminant Analysis).
6. Red Neuronal (Perceptrón multicapa).
7. SVM (Support Vector Machine).



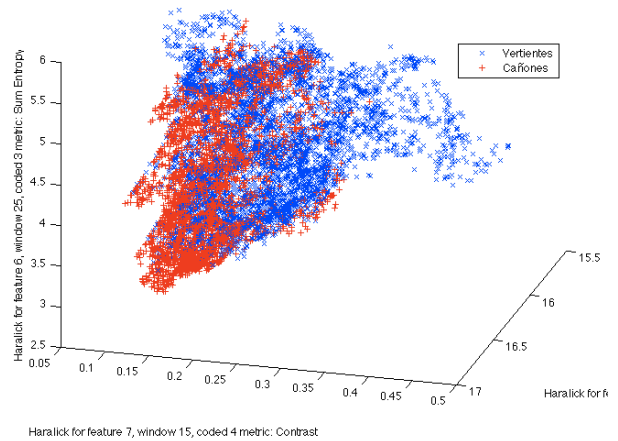
(a) Histogramas de $H_{f2,r25,d16}^{(6)}$ para las 8 clases.



(b) Gráfico de dispersión de $H_{f2,r25,d16}^{(6)}$ vs $H_{f7,r50,d08}^{(2)}$ para las ocho clases.

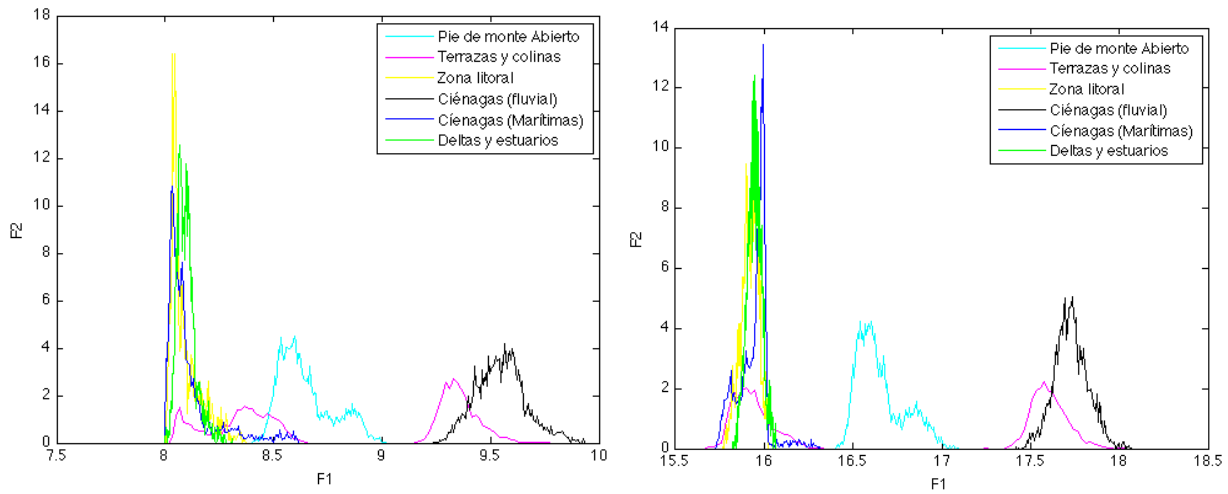


(c) Gráfico de dispersión de $H_{f2,r25,d16}^{(6)}$ vs $H_{f7,r50,d08}^{(2)}$ vs $H_{f4,r25,d08}^{(7)}$ para las ocho clases.

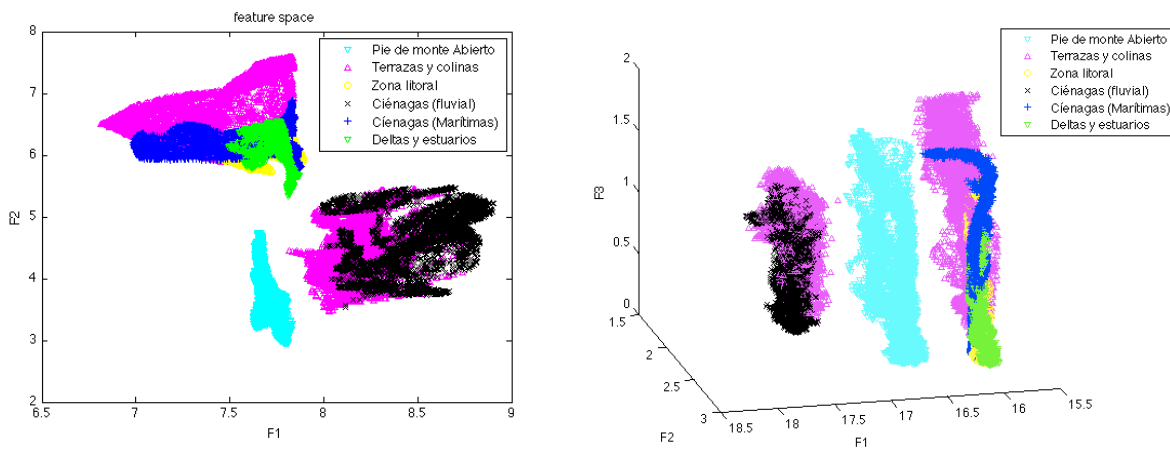


(d) Gráfico de dispersión de $H_{f2,r25,d16}^{(6)}$ vs $H_{f7,r15,d16}^{(2)}$ vs $H_{f6,r25,d08}^{(7)}$ para las ocho clases.

Figura 6-1: Gráficas de diferentes combinaciones de los 3 primeros descriptores de las diferentes selecciones para la super clase 1



(a) Histogramas de $H_{F3,r10,d08}^{(6)}$ para las 8 clases. (b) Histogramas de $H_{F3,r25,d16}^{(6)}$ para las 8 clases.



(c) Gráfico de dispersión de $H_{F3,r15,d08}^{(7)}$ vs $H_{F6,r25,d08}^{(7)}$ para las ocho clases. (d) Gráfico de dispersión de $H_{F3,r25,d16}^{(6)}$ vs $H_{F6,r50,d16}^{(8)}$ para las 8 clases.

Figura 6-2: Gráficas de diferentes combinaciones de los 3 primeros descriptores de las diferentes selecciones para la super clase 2

8. Clasificador de Mahalanobis.
9. Clasificador de distancia mínima.

La implementación de cada uno de ellos que se usó en el desarrollo de investigación hace parte del toolbox Balú[43] desarrollado en la Universidad Católica de Chile por el profesor Domingo Mery y el grupo de investigación en inteligencia de Máquina (GRIMA)¹. Este toolbox contiene las implementaciones de todos los algoritmos de selección de características y los métodos de clasificación usados en esta investigación, y está disponible en la web para su uso en ambientes académicos.

La calidad de los resultados de clasificación fue evaluada mediante la metodología de validación cruzada con un total de 10 divisiones de los datos. La validación cruzada reportó el promedio del desempeño de cada uno de los nueve clasificadores y construyó intervalos con el 95 % de confianza para ayudar a analizar y comprender mejor la estabilidad de los resultados obtenidos.

Los clasificadores elegidos fueron utilizados con cada una de las selecciones obtenidas de manera que se pudiera comparar no solamente cual es el mejor clasificador para las ocho unidades geo-morfológicas; sino también, que se se pudiera determinar cual o cuales de las selecciones ofrecen los mejores resultados.

6.2. Clasificación de unidades geo-morfológicas

En esta sección se presentan y analizan los resultados de las diferentes clasificaciones desarrolladas para alcanzar los objetivos de esta investigación. Antes de presentar los desempeños obtenidos con la metodología de descriptores basados en texturas de Haralick es necesario mostrar cual es el desempeño de la clasificación de las ocho unidades geo-morfológicas usando sólo los features morfológicos extraídos de la revisión de literatura. La tabla **6-3** resume los resultados de clasificar las geo-formas con 3 subconjuntos de los siete descriptores morfológicos de la tabla **4-3**. De esta manera se logra tener un punto de comparación para evaluar las bondades de la metodología propuesta en este trabajo.

En la tabla **6-3** se puede observar que ninguno de los desempeños es bastante alto, alcanzando un máximo de 61,83 % de las muestras bien clasificadas con el algoritmo KNN ($k = 9$). El intervalo de confianza de las mejor clasificación es

$$60,61 \% \leq p \leq 63,06 \%$$

con un 95 % de confianza. Aproximadamente el 40 % de los datos fueron mal clasificados; esto, en términos del área del caso de estudio constituye una región de aproximadamente $19000km^2$ de área en los que las unidades geo-morfológicas no fueron bien aproximadas.

¹Página web del grupo GRIMA. <http://grima.ing.puc.cl>

Tabla 6-3: Desempeño de la clasificación usando sólo descriptores geo-morfológicos

CLASIFICADOR	Subconjuntos de descriptores morfológicos		
	[1 6 5 4 2 3]	[5 1 6 4]	[5 1 6 2 3]
knn, $k = 5$	0.6025	0.6067	0.6090
knn, $k = 7$	0.6178	0.6162	0.6092
knn, $k = 9$	0.6180	0.6198	0.6183
LDA	0.4566	0.4591	0.4550
QDA	0.4185	0.4249	0.4177
Red Neuronal	0.5043	0.5039	0.5007
SVM, kernel = RBF	0.5551	0.5597	0.5568
Mahalanobis	0.4774	0.4718	0.4752
Mínima distancia	0.3514	0.3512	0.3510

6.2.1. Separación de las 2 super clases

Utilizando el criterio definido en la sección 5 se separaron los datos de todo el caso de estudio en dos super clases; de esta manera el problema de clasificar las ocho unidades geo-morfológicas se transforma en dos problemas donde cada uno de ellos posee una complejidad menor. Esta clasificación se llevó a cabo utilizando las cuatro selecciones detalladas en la sección 6.1.1, donde las dos primeras corresponden a dos subconjuntos de descriptores de los siete descriptores morfológicos; las dos restantes son el resultado de la extracción de las componentes principales de las primeras dos selecciones. Los nueve clasificadores descritos anteriormente fueron utilizados con cada una de las selecciones, obteniendo así los desempeños que se muestran en la tabla 6-4.

La división de las super clases se considera satisfactoria debido al un aumento en la tasa de individuos bien clasificados en comparación con el desempeño obtenido al tratar de separar las ocho clases de manera directa (6-3). Este aumento una vez realizada la agrupación en super clases es de aproximadamente 31 % con el clasificador SVM y la selección 2. Esta selección corresponde a los descriptores 1, 5, 7 que son las elevaciones, la tangente de la pendiente y el HAND respectivamente. Ver detalles en la tabla 6-4. El intervalo de confianza del 95 % del desempeño de este clasificador es:

$$91,90 \% \leq p \leq 93,09 \%$$

Los resultados obtenidos son consistentes con respecto a los histogramas de las dos super clases mostrados en la sección 5.1. Adicionalmente, los resultados son congruentes con la reducción de complejidad de la clasificación, puesto que se pasó de un problema con ocho clases a un problema binario una vez que las clases fueron agrupadas.

Se debe ser cuidadoso al interpretar los resultados de la tabla 6-4 debido a que este no es el desempeño final del problema planteado. El desempeño final dependerá de las clasifica-

Tabla 6-4: Desempeños de la clasificación de las 2 super clases.

CLASIFICADOR	S1	S2	S3	S4
knn, $k = 5$	0.9136	0.9184	0.9112	0.9183
knn, $k = 7$	0.9164	0.9230	0.9098	0.9193
knn, $k = 9$	0.9179	0.9226	0.9132	0.9195
LDA	0.9095	0.9093	0.8842	0.8856
QDA	0.9148	0.9203	0.9172	0.9188
Red Neuronal	0.9214	0.9245	0.9174	0.9193
SVM, kernel = RBF	0.9225	0.9249	0.9204	0.9230
Mahalanobis	0.8682	0.8687	0.9112	0.8882
Mínima distancia	0.9035	0.9037	0.9037	0.9033

ciones de las unidades geo-morfológicas pertenecientes a cada super clase, ponderadas por el obtenido en los experimentos de segmentación en las dos super clases. De este modo

$$p_{total} = p_g \cdot \left(\frac{p_{sc1} + p_{sc2}}{2} \right) \quad (6-1)$$

Donde p_{total} es la proporción de datos clasificados correctamente en toda el área de estudio, p_g es la proporción de datos bien clasificados para el problema de la división de los datos del caso de estudio en las super clases 1 y 2, p_{sc1} es la tasa de datos correctamente clasificados pertenecientes sólo a las super clase 1 (limitada a las geo-formas que en ella se contienen), y p_{sc2} es la tasa de datos correctamente clasificados pertenecientes sólo a la super clase 2.

6.2.2. Clasificación de la super clase 1

La super clase 1 contiene las unidades geo-morfológicas: Vertientes y cañones. La clasificación inherente a esta super clase es notablemente menos compleja que la clasificación de las ocho geo-formas simultáneamente, debido a que es bien sabido por la literatura sobre reconocimiento de patrones [18] que un problema de clasificación binaria (dos clases) es uno de los tipos de problemas de clasificación más simple.

Para separar las vertientes de los cañones se utilizaron los datos del subconjunto de datos 3 (ver tabla 4-4 y figura 4-4); este contiene solo instancias de estas dos unidades geo-morfológicas en una proporción aproximada de 70 %-30 % respectivamente. Por lo anterior el subconjunto de datos 3 se adapta muy bien al tipo de clasificación que se plantea. Como se mencionó en el capítulo 4 cada subconjunto de datos cuenta con un total de individuos (n_{total}) de:

$$n_{total} = 451 * 451 = 203401$$

Dado que en los experimentos se utilizaron nueve clasificadores para cuatro selecciones de características, y que la metodología de evaluación fue validación cruzada con diez divisiones; se hizo necesario realizar un muestreo a al subconjunto de datos para reducir el costo computacional de los experimentos. Se eligieron entonces $n_{muestra} = 15000$ individuos para las pruebas de clasificación, de manera que la muestra fuera suficientemente representativa, y al mismo tiempo el costo computacional de realizar las validaciones cruzadas de todos los experimentos fuera suficientemente bajo.

La clasificación de la super clase 1 se realizó con cada uno de los nueve clasificadores para cada una de las cuatro selecciones explicadas en la tabla 6-1; adicionalmente los clasificadores se evaluaron también utilizando sólo los siete descriptores morfológicos; de esta manera es posible comparar la metodología de clasificación con información de la textura del terreno contra la metodología utilizada en trabajos relacionados. Los resultados obtenidos en esta subclasificación se presentan en la tabla 6-5 resaltando el mejor desempeño obtenido con cada una de las diferentes selecciones de características.

Tabla 6-5: Desempeño de la clasificación de las geo-formas en la super clase 1

CLASIFICADOR	SÓLO DESCRIPTORES MORFOLÓGICOS	S1	S2	S3	S4
knn, $k = 5$	0.7210	0.8985	0.6936	0.9682	0.9674
knn, $k = 7$	0.7314	0.8921	0.7028	0.9657	0.9651
knn, $k = 9$	0.7345	0.8873	0.6992	0.9637	0.9638
LDA	0.7198	0.8672	0.8168	0.7875	0.7770
QDA	0.6989	0.8677	0.8223	0.8077	0.7801
Red Neuronal	0.7195	0.8683	0.8190	0.7883	0.7791
SVM, kernel = RBF	0.7283	0.9558	0.9607	0.9729	0.9737
Mahalanobis	0.6846	0.8695	0.8396	0.8240	0.7969
Mínima distancia	0.6705	0.6243	0.5018	0.6127	0.5730

En general el desempeño de la clasificación con información de texturas para la super clase 1 arrojó resultados satisfactorios con índices de clasificación por encima del 80 %, y llegando a un máximo de 97,37 % de desempeño con un intervalo de confianza de

$$97,07\% \leq p \leq 97,66\%$$

Este resultado fue alcanzado con la selección S4 usando el clasificador SVM para el cual se usó como kernel una función de base radial. Existen sólo algunas pocas excepciones donde el desempeño estuvo por debajo del 70 %, particularmente en el caso del clasificador de mínima distancia.

La idea detrás de la utilización de las técnicas selección de características con criterios wrapper es que el mismo clasificador que se utilizó para realizar la selección sea utiliza-

do para clasificar, pudiendo obtener así mejores resultados. Los resultados de la tabla **6-1** son consecuentes con la afirmación anterior ya que para los casos de los clasificadores KNN con $k = 5, 7, 9$ se obtuvieron desempeños altos para las selecciones S4 y S3; estas selecciones corresponden a las realizadas con el algoritmo SFS y como criterio KNN.

Sin embargo es de resaltar que el algoritmo de clasificación SVM ganó en todos los casos obteniendo los mejores desempeños en todos los casos de selección, excepto en aquella en que sólo se utilizaron descriptores morfológicos. SVM no fue utilizado como criterio para la selección de características debido que es un algoritmo de clasificación de complejidad alta, y está asociado a un gran costo computacional. Su uso podría mejorar los resultados obtenidos.

Finalmente, para dar una idea de la precisión del desempeño de los clasificadores de tipo KNN se presentan los intervalos de confianza del 95 % de todos ellos, que fueron obtenidos con la selección S3:

$$KNN, k = 5 : 92,17 \% \leq p \leq 93,98 \%$$

$$KNN, k = 7 : 92,14 \% \leq p \leq 93,96 \%$$

$$KNN, k = 9 : 91,75 \% \leq p \leq 93,62 \%$$

6.2.3. Clasificación de la super clase 2

La super clase 2 contiene las 6 unidades geo-morfológicas restantes una vez que son extraídas las 2 que están presentes en la super clase 1, estas unidades son:

1. Terrazas y colinas.
2. Pie de monte abierto.
3. Zona litoral.
4. Ciénagas y llanuras aluviales con control fluvial.
5. Ciénagas y llanuras aluviales con control del mar.
6. Deltas y estuarios.

Pese a que la super clase 2 contiene aún 6 clases, la complejidad de su clasificación es menor que el caso en el que se tenían las 8 clases simultáneamente. Sin embargo, los análisis realizados los descriptores morfológicos de la super clase 2 revelan que los dos diferentes tipos de ciénagas y los deltas y estuarios poseen distribuciones muy similares con medias muy cercanas en sus histogramas (Ver figura **6-3**). Adicionalmente, las ciénagas con control fluvial pueden ser diferenciadas de las ciénagas con control del mar a través de criterios de distancia horizontal medida desde la línea de costa del mar. Esta misma situación ocurre con las clases Zona litoral y deltas y estuarios (Ver criterios en la tabla **4-2**).

Lo anterior sugiere entonces que una nueva clasificación puede realizarse agrupando las geoformas 5, 6, 7, 8; teniendo entonces un problema de 3 clases en la super clase 2. La etapa restante sería entonces la separación de las las clases 5, 6, 7 y 8 con el criterio de distancia horizontal al mar. Esta situación representa una clasificación con un solo descriptor, que es considerada un caso más sencillo de resolver[18], pero que no se realizará en este trabajo.

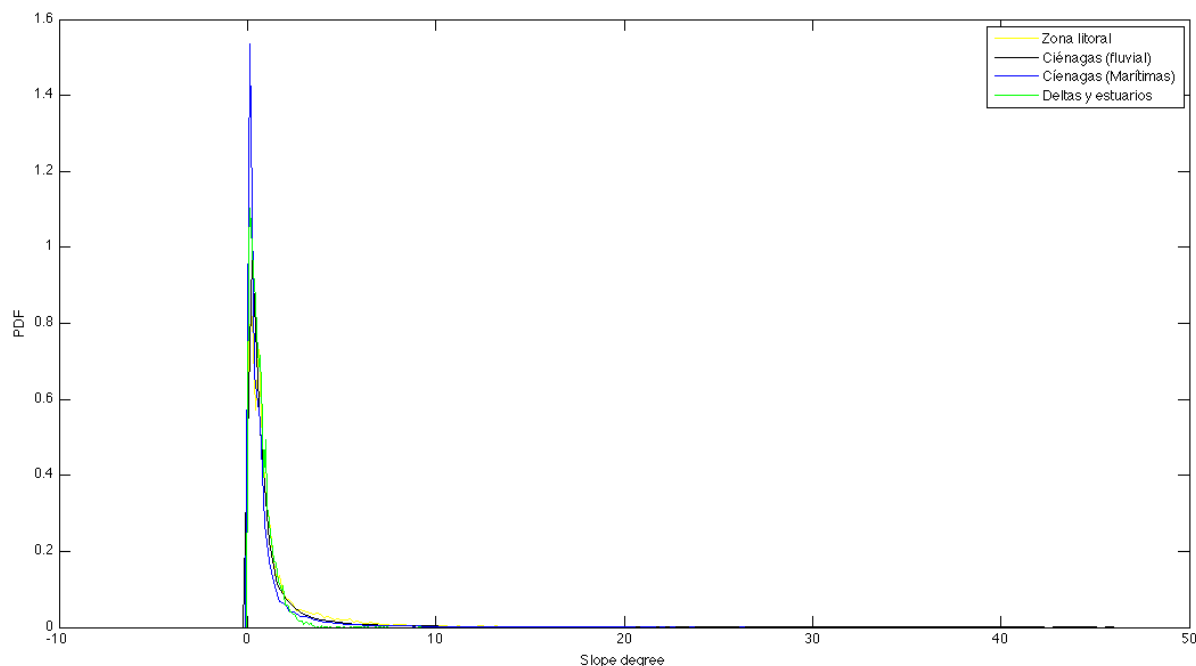


Figura 6-3: Histogramas de frecuencia de las pendientes en grados para las clases 5, 6, 7 y 8.

Los datos utilizados para separar las geo-formas de la super clase 2 se extrajeron de varios de los subconjuntos de datos: El subconjunto 4, el 5 y el 7 (Ver figura 4-4 y tabla 4-4). La razón de esto es que no existe uno sólo de los conjuntos que posea todas las geo-formas contenidas en la super clase. Análogamente al caso de la super clase 1, se tomó en este caso una muestra con un total de $n_{total} = 15000$ individuos para realizar la clasificación. Las proporciones de elementos de cada clase en la muestra y los subconjuntos de datos de los que se extrajeron se muestran en la tabla 6-6.

Los experimentos para la super clase 2 se diseñaron de manera muy similar a como se hizo para la super clase 1. Se utilizaron nueve clasificadores y las 4 selecciones de las tabla 6-2. Adicionalmente se realizó una clasificación utilizando sólo los descriptores morfológicos para comparar la metodología propuesta. Una diferencia en este caso debe ser resaltada: El clasificador SVM tiene una implementación especial, ya que este está originalmente definido sólo para problemas de clasificación binaria [18, 39]; por lo tanto, la forma de clasificar las 6 clases en este problema es utilizando una versión modificada del algoritmo SVM, implementada en el toolbox Balú, que realiza clasificaciones binarias con subconjuntos de las clases objetivo,

Tabla 6-6: Proporciones y origen de los individuos de la muestra de la super clase 2.

Conjunto de datos	Geoforma	Cantidad de individuos	Proporción
Conjunto 4	4	2000	14.28 %
Conjunto 4	5	1000	7.14 %
Conjunto 4	7	2000	14.28 %
Conjunto 4	8	1000	7.14 %
Conjunto 5	3	4000	28.57 %
Conjunto 7	4	2000	14.28 %
Conjunto 7	6	2000	14.28 %
Total	-	14000	100 %

y genera un árbol de clasificaciones hasta que los nodos del árbol sean grupos de una sola clase.

Tabla 6-7: Desempeño de la clasificación de las geo-formas de la super clase 2

CLASIFICADOR	SÓLO DESCRIP- TORES MORFOLÓGICOS	S1	S2	S3	S4
knn, $k = 5$	0.5993	0.9189	0.8909	0.9770	0.9784
knn, $k = 7$	0.6079	0.9131	0.8869	0.9741	0.9766
knn, $k = 9$	0.6043	0.9096	0.8821	0.9726	0.9726
LDA	0.4543	0.9456	0.9343	0.8452	0.8636
QDA	0.3987	0.2857	0.9154	0.9190	0.9290
Red Neuronal	0.4890	0.9653	0.9429	0.9139	0.9273
SVMPLUS, kernel = RBF	0.5417	0.9587	0.9711	0.9748	0.9729
Mahalanobis	0.4700	0.9511	0.9203	0.9235	0.9198
Mínima distancia	0.3496	0.6580	0.5169	0.5907	0.6008

La clasificación de la super clase 2 con la metodología propuesta en esta investigación mostró un alto desempeño con índices de clasificación por encima del 85 %; obteniéndose el mejor resultado con la selección S4 y algoritmo KNN con $k = 5$. La tasa de individuos bien clasificados llegó al 97,84 % con un intervalo de confianza de

$$97,57\% \leq p \leq 98,12\%$$

La selección S4 fue obtenida con el algoritmo SFS con criterio de separabilidad: KNN con $k = 7$, lo que es consecuente con lo que se espera de las técnicas de selección de características wrapper.

En comparación con la clasificación de la super clase 1, en este caso el algoritmo SVM no ganó en todas las pruebas; no obstante obtuvo un desempeño notablemente mayor que el resto de clasificadores de la selección S2. Con el clasificador de mínima distancia se obtuvieron los resultados más bajos de la misma manera en que ocurrió en la super clase 1. Un caso especial ocurrió con el algoritmo QDA; su desempeño con la selección S1 fue el más bajo de todos (28 %); esto posiblemente se deba a problemas de precisión numérica del algoritmo en el cálculo de las matrices de covarianza y de covarianza inversa, como lo advirtieron los mensajes arrojados por la plataforma Matlab© en la cual está implementado.

En cada una de las clasificaciones de este capítulo que hacen parte de la metodología propuesta se puede apreciar que los resultados siempre fueron iguales o superiores al 95 %; lo que es una muy aceptable tasa de individuos bien clasificados en cualquier tipo de problema. El desempeño total de clasificar las 8 clases debe ser estimado de forma especial debido a la forma jerárquica en que el proceso fue conducido. El desempeño total es entonces calculado siguiendo la fórmula de la ecuación 6-1 en donde $p_g = 0,9249$ y representa el desempeño de la división del caso de estudio en las super clases; $p_{sc1} = 0,9737$ y $p_{sc2} = 0,9784$ y corresponden a los desempeños de la clasificación de las unidades geo-morfológicas de la super clase 1 y 2 respectivamente. El desempeño total es entonces:

$$p_{total} = p_g \cdot \left(\frac{p_{sc1} + p_{sc2}}{2} \right) = 0,9249 \cdot \left(\frac{0,9737 + 0,9784}{2} \right) = (0,9249) \cdot (0,9761) = 0,9027 \quad (6-2)$$

El desempeño total de la clasificación ascendió a un 90,27 % que representa una región del área de estudio bien clasificada de aproximadamente $43885 km^2$, y un error de clasificación del 9,72 % para un área aproximada de $4725 km^2$.

7 Conclusiones y Trabajo Futuro

En este capítulo se resume a manera de conclusiones los aspectos más importantes encontrados en el desarrollo de esta investigación; adicionalmente se sugieren varios temas que pueden ser focos potenciales de trabajo para futuras investigaciones.

7.1. Conclusiones

En clasificación y reconocimiento de patrones uno de los aspectos más importantes es la información disponible para discriminar las clases objetivo. Cuando no existe información suficiente, o cuando la información disponible no es adecuada para separar los datos en una clase u otra, es necesario llevar a cabo un proceso de extracción de características nuevas adecuadas para el problema planteado.

En esta tesis se contó inicialmente con siete descriptores morfológicos que son de uso común en la literatura en muchos trabajos. Se probó que estos descriptores no son suficientemente buenos para discriminar las ocho unidades geo-morfológicas que fueron definidas como clases objetivos. Por lo tanto, el esfuerzo en esta investigación se enfocó básicamente en encontrar nuevos descriptores de las clases a través de la utilización de métricas de textura del terreno de la región a clasificar, esperando que la introducción de esta nueva información pudiera mejorar los resultados de la clasificación.

La razón por la que se eligió una metodología de extracción de características basada en texturas es que esta metodología es similar a la forma en que un geomorfolólogo analiza una región geográfica a través de la observación de esta. En una clasificación geomorfológica manual un experto observa una zona en busca de patrones bien establecidos en su conocimiento, y determina a que tipo de unidad geo-morfológica pertenece esa zona. Con el análisis de texturas de terreno se busca, análogamente, encontrar patrones de comportamiento cuantificables que den cuenta de la forma del terreno en una región geográfica.

La metodología de extracciones de características a partir de texturas de Haralick entrega una gran cantidad de descriptores nuevos, donde algunos de ellos aportan una considerable mejora en el desempeño de la clasificación cuando son incluidos. Según revelan los desempeños en las clasificaciones realizadas, las clasificaciones con información de texturas es superior a la realizada solo con descriptores morfológicos en la mayoría de los casos con desempeños de alrededor de 25 % más precisión al separar las clases.

A partir de los análisis realizados a los descriptores morfológicos iniciales se puede deducir que ninguno de ellos ofrece una separabilidad aceptable para separar los tipos de relieve; un

ejemplo de ello son los descriptores de curvatura que se muestran en las figuras 5-3 y 5-2. Sin embargo, los resultados de la extracción y selección de características para las dos super clases revelan que los descriptores de textura que mejor separan los datos en las diferentes selecciones realizadas son extraídos a partir de la información de la curvatura. Tal es el caso de: $H_{f2,r25,d16}^{(6)}$ en la selección S1 para la super clase 1 y de $H_{f3,r25,d16}^{(6)}$ en la selección S4 para la super clase 2.

Se puede decir entonces que la contribución principal de este trabajo es el establecimiento de un conjunto de descriptores apto para el problema de la clasificación de tipos de relieve. Este conjunto ofrece desempeños de clasificación mayores que los reportados en la literatura en aproximaciones de solución al mismo problema.

En cada uno de los experimentos de clasificación con texturas se resalta el buen desempeño del clasificador SVM; manteniéndose siempre entre aquellos que entregaban los mejores resultados. En contraste se tiene el clasificador de distancia mínima cuyo desempeño fue en general el más bajo en todos los experimentos. Por otro lado, ninguno de ellos arrojó resultados por encima del 80 % con la selección de características que no incluía información de texturas. De lo anterior se infiere entonces la utilidad y efectividad de la metodología propuesta frente al enfoque típico para el problema abordado.

El desempeño final de la clasificación de las ocho clases presentado en la ecuación 6-2 sugiere un alto grado de individuos bien clasificados, superando bastante la clasificación sin usar información de textura. No obstante, en la ecuación se puede apreciar, que el desempeño de la separación de las super clases es significativamente menor que el desempeño de las clasificaciones dentro de cada super clase; afectándolo y reduciéndolo considerablemente.

7.2. Trabajo futuro

Para la realización de los experimentos de este trabajo se utilizó un número razonable de clasificadores de manera que fuera un conjunto suficientemente representativo de los diferentes enfoques de reconocimiento de patrones. Sin embargo, los parámetros de cada uno de ellos se eligieron teniendo en cuanto su efectividad en en otros problema de clasificación. Un foco importante de trabajo futuro es encontrar un conjunto de parámetros para cada clasificador que sea más adecuado específicamente para el problema presentado; e incluso, utilizar clasificadores híbridos que puedan ayudar a mejorar el desempeño.

Los algoritmos utilizados para la selección de características fueron limitados, y se eligieron como los más usados y recomendados en la literatura. Sin embargo, existen una gran cantidad de algoritmos de selección de características y de criterios de separabilidad que aplicados al problema planteado podrían mejorar los resultados de la clasificación o mejorar la eficiencia computacional de la misma. Algunos clasificadores no fueron utilizados como criterios de tipo wrapper debido al costo computacional que implicaba realizar un gran cantidad de clasificaciones en el proceso de selección; en condiciones de hardware más adecuadas se pueden realizar experimentos más precisos en la etapa de selección de características.

El criterio de división que se utilizó para separar la super clase 1 de la super clase 2 no usó información de textura y solo se usaron los descriptores morfológicos. Pese a que los resultados no fueron bajos al aplicar el criterio, la inclusión de este tipo de información pudiera elevar los resultados considerablemente.

Una vez que ha sido comprobada la eficacia de extraer información de texturas para clasificar tipos de relieve, una dirección de trabajo futuro es la exploración de otras técnicas, como LBP o texturas de Gabor, para extraer dicha información y aportar más descriptores útiles para la clasificación.

Tal vez uno de los aspectos más importantes que se plantean en este y en cualquier otro problema de reconocimiento de patrones es la capacidad de generalización de la solución de clasificación presentada. Por lo tanto, las bondades de esta investigación deben ser puestas a prueba utilizando este mismo mecanismo con nueva información en otras zonas del territorio Colombiano con características morfológicas similares.

Bibliografía

- [1] ACCIANI, Giuseppe ; CHIARANTONI, Ernesto ; FORNARELLI, Girolamo ; VERGURA, Silvano: 2003 Special Issue A feature extraction unsupervised neural network for an environmental data set. En: *Neural Networks* 16 (2003), p. 427–436
- [2] AHONEN, Timo ; HADID, Abdenour ; PIETIKÄINEN, Matti: Face description with local binary patterns: application to face recognition. En: *IEEE transactions on pattern analysis and machine intelligence* 28 (2006), Dezember, Nr. 12, p. 2037–41. – ISSN 0162–8828
- [3] DE GEOCIENCIAS Y MEDIO AMBIENTE, Escuela: IMPLEMENTACIÓN DE LA METODOLOGÍA DE ZONIFICACIÓN DE AMENAZAS POR INUNDACIONES PARA TRES GRANDES CUENCAS DEL PAÍS (Colombia) / Universidad Nacional de Colombia - Sede Medellín. 2011. – Informe de Investigación
- [4] ARRELL, K.E. ; FISHER, P.F. ; TATE, N.J. ; BASTIN, L.: A fuzzy c-means classification of elevation derivatives to extract the morphometric classification of landforms in Snowdonia, Wales. En: *Computers & Geosciences* 33 (2007), Oktober, Nr. 10, p. 1366–1381. – ISSN 00983004
- [5] BACAO, F ; LOBO, V ; PAINHO, M: The self-organizing map, the Geo-SOM, and relevant variants for geosciences. En: *Computers & Geosciences* 31 (2005), März, Nr. 2, p. 155–163. – ISSN 00983004
- [6] BALL, Geoffrey H.: ISODATA: A novel method for data analysis and pattern classification / Stanford Research Institute. 1965. – Informe de Investigación
- [7] BEZDEK, James C.: FCM : THE FUZZY c-MEANS CLUSTERING ALGORITHM 1 ; yk E Y ~ l. 10 (1984), Nr. 2, p. 191–203
- [8] BI, Jacek S.: Landform Characterization with Geographic Information Systems. En: *Most* 63 (1997), Nr. 2, p. 183–191
- [9] BUE, B.D. ; STEPINSKI, T.F.: Automated classification of landforms on Mars. En: *Computers & Geosciences* 32 (2006), Juni, Nr. 5, p. 604–614. – ISSN 00983004
- [10] BURROUGH, P: High-resolution landform classification using fuzzy k-means. En: *Fuzzy Sets and Systems* 113 (2000), Juli, Nr. 1, p. 37–52. – ISSN 01650114

-
- [11] BURROUGH, Peter A. ; MCDONNELL, Rachael A.: Principles of Geographical Information Systems. (1998)
- [12] CASTLEMAN, K.R.: *Digital image processing*. Prentice Hall, 1996 (Prentice-Hall signal processing series). – ISBN 9780132114677
- [13] CASTRO, Juan C.: *Metodología para la obtención de modelos digitales de terreno hidrológica y geomorfológicamente coherentes*, Escuela de geociencias y medio ambiente, Universidad Nacional de Colombia, Tesis de Grado, 2011
- [14] CHORLEY, R.J. ; SCHUMM, S.A. ; SUGDEN, D.E.: *Geomorphology*. Methuen, 1985. – ISBN 9780416325904
- [15] DENG, Y. ; WILSON, J. P. ; BAUER, B. O.: DEM resolution dependencies of terrain attributes across a landscape. En: *International Journal of Geographical Information Science* 21 (2007), Nr. 2, p. 187–213. – ISSN 1365–8816
- [16] DRAGUT, L ; BLASCHKE, T: Automated classification of landform elements using object-based image analysis. En: *Geomorphology* 81 (2006), November, Nr. 3-4, p. 330–344. – ISSN 0169555X
- [17] DREYFUS, G.: *Neural networks: methodology and applications*. Springer, 2005. – ISBN 9783540229803
- [18] DUDA, R.O. ; HART, P.E. ; STORK, D.G.: *Pattern classification*. Vol. 2. Citeseer, 2001
- [19] EHSANI, Amir H. ; QUIEL, Friedrich: A semi-automatic method for analysis of landscape elements using Shuttle Radar Topography Mission and Landsat ETM+ data. En: *Computers & Geosciences* 35 (2009), Februar, Nr. 2, p. 373–389. – ISSN 00983004
- [20] FLÓREZ, A.: *Colombia: Evolución de Sus Relieves y Modelados*. Universidad Nacional de Colombia, Red de Estudio de Espacio y Territorio, RET, 2003 (Espacio y territorio). – ISBN 9789587013122
- [21] GARCÍA, V. ; SÁNCHEZ, J.S. ; MOLLINEDA, R.a.: On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. En: *Knowledge-Based Systems* 25 (2012), Februar, Nr. 1, p. 13–21. – ISSN 09507051
- [22] GUO, Gongde ; NEAGU, Daniel ; CRONIN, Mark T D.: Using kNN Model for Automatic Feature Selection. (2005), Nr. 3
- [23] HALL, Mark A. ; SMITH, Lloyd A.: Practical Feature Subset Selection for Machine Learning. (1997)
- [24] HARALICK, Robert M.: Statistical and Structural Approaches to Texture. En: *Proceedings of The IEEE* 67 (1979), Nr. 5

-
- [25] HARTIGAN, J. A. ; WONG, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm. En: *Journal of the Royal Statistical Society* (1979)
- [26] HAWKINS, Douglas M.: The problem of overfitting. En: *Journal of chemical information and computer sciences* 44 (2004), Nr. 1, p. 1–12. – ISSN 0095–2338
- [27] HOSOKAWA, Masafumi ; HOSHI, Takashi: Landform Classification Method Using Self-Organizing Map and its application to Earthquake Damage Evaluation. 00 (2001), Nr. C, p. 1684–1686. ISBN 0780370317
- [28] HRUSCHKA, Eduardo R. ; CAMPELLO, Ricardo J G B. ; FREITAS, Alex A. ; PONCE, C ; CARVALHO, Leon F D.: A Survey of Evolutionary Algorithms for Clustering. 39 (2009), Nr. 2, p. 133–155
- [29] HUANG, T.: Face localization via hierarchical CONDENSATION with fisher boosting feature selection. En: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* 2 (2004), p. 719–724. ISBN 0–7695–2158–4
- [30] HUGGETT, R.J.: *Fundamentals of Geomorphology*. Taylor & Francis, 2011. – ISBN 9780415567749
- [31] IRVIN, B: Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. En: *Geoderma* 77 (1997), Juni, Nr. 2-4, p. 137–154. – ISSN 00167061
- [32] JAIN, a. K. ; MURTY, M. N. ; FLYNN, P. J.: Data clustering: a review. En: *ACM Computing Surveys* 31 (1999), September, Nr. 3, p. 264–323. – ISSN 03600300
- [33] JAIN, Anil ; ZONGKER, Douglas: Feature Selection: Evaluation, Application, and Small Sample Performance. 19 (1997), Nr. 2, p. 153–158
- [34] A.K. JAIN ; DUIN, P.W.: Statistical pattern recognition: a review. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2000), Nr. 1, p. 4–37. – ISSN 01628828
- [35] JOHNSON, R.A. ; WICHERN, D.W.: *Applied multivariate statistical analysis*. Vol. 5. Prentice Hall Upper Saddle River, NJ, 2002
- [36] JOLLIFFE, I T.: *Principal Component Analysis, Second Edition*
- [37] KLINGSEISEN, B ; METTERNICHT, G ; PAULUS, G: Geomorphometric landscape analysis using a semi-automated GIS-approach. En: *Environmental Modelling & Software* 23 (2008), Januar, Nr. 1, p. 109–121. – ISSN 13648152

- [38] KOHAVI, Ron ; JOHN, H: Artificial Intelligence Wrappers for feature subset selection. 97 (2011), Nr. 97, p. 273–324
- [39] KOTSIANTIS, S B.: Supervised Machine Learning : A Review of Classification Techniques. En: *Informatica* 31 (2007), p. 249–268
- [40] KOUTROUMBAS, K. ; THEODORIDIS, S. ; ELSEVIER (Ed.): *An Introduction to Pattern Recognition: A Matlab Approach*. 2009. – 219 p.
- [41] MACMILLAN, R: A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. En: *Fuzzy Sets and Systems* 113 (2000), Juli, Nr. 1, p. 81–109. – ISSN 01650114
- [42] MAULIK, Ujjwal ; SAHA, Indrajit: Modified differential evolution based fuzzy clustering for pixel classification in remote sensing imagery. En: *Pattern Recognition* 42 (2009), September, Nr. 9, p. 2135–2149. – ISSN 00313203
- [43] MERY, Domingo. *BALU: A toolbox Matlab for computer vision, pattern recognition and image processing* (<http://dmery.ing.puc.cl/index.php/balu>). 2011
- [44] NAVOT, Amir ; SHPIGELMAN, Lavi ; TISHBY, Naftali ; VAADIA, Eilon: Nearest Neighbor Based Feature Selection for Regression and its Application to Neural Activity, MIT Press, 2006
- [45] A.D. NOBRE ; CUARTAS, L.a. ; HODNETT, M. ; RENNÓ, C.D. ; RODRIGUES, G. ; SILVEIRA, a. ; WATERLOO, M. ; SALESKA, S.: Height above the Nearest Drainage, a hydrologically relevant new terrain model. En: *Journal of Hydrology* 404 (2011), April, Nr. 1-2, p. 13–29. – ISSN 00221694
- [46] PAL, M. ; MATHER, P. M.: Support vector machines for classification in remote sensing. En: *International Journal of Remote Sensing* 26 (2005), März, Nr. 5, p. 1007–1011. – ISSN 0143–1161
- [47] PAL, Mahesh ; MATHER, Paul M.: An assessment of the effectiveness of decision tree methods for land cover classification. En: *Remote Sensing of Environment* 86 (2003), August, Nr. 4, p. 554–565. – ISSN 00344257
- [48] PRIMA, O ; ECHIGO, a ; YOKOYAMA, R ; YOSHIDA, T: Supervised landform classification of Northeast Honshu from DEM-derived thematic maps. En: *Geomorphology* 78 (2006), August, Nr. 3-4, p. 373–386. – ISSN 0169555X
- [49] RANDEN, Trygve: Filtering for Texture Classification : A Comparative Study. 21 (1999), Nr. 4, p. 291–310

-
- [50] RENNO, C ; NOBRE, a ; CUARTAS, L ; SOARES, J ; HODNETT, M ; TOMASELLA, J ; WATERLOO, M: HAND, a new terrain descriptor using SRTM-DEM: Mapping terra-firme rainforest environments in Amazonia. En: *Remote Sensing of Environment* 112 (2008), September, Nr. 9, p. 3469–3481. – ISSN 00344257
- [51] RUIZ-OCHOA, Mauricio ; BERNAL, Gladys ; POLANÍA, Jaime: Influence of Sinú River and the Caribbean Sea over the Cispatá Lagoon System. En: *Boletín de Investigaciones Marinas y Costeras - INVEMAR* (2008)
- [52] SCHMIDT, J: Fuzzy land element classification from DTMs based on geometry and terrain position. En: *Geoderma* 121 (2004), August, Nr. 3-4, p. 243–256. – ISSN 00167061
- [53] TAGIL, Sermin ; JENNESS, Jeff: GIS-Based Automated Landform Classification and Topographic, Landcover and Geologic Attributes of Landforms Around the Yazoren Polje, Turkey. En: *Journal of Applied Sciences* (2008)
- [54] XU, Rui ; WUNSCH, Donald: Survey of clustering algorithms. En: *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* 16 (2005), Mai, Nr. 3, p. 645–78. – ISSN 1045–9227