Universidad Nacional del Litoral

Facultad de Ingeniería y Ciencias Hídricas

Proyecto Final de Carrera - Ingeniería en Informática

Desarrollo de un método para delimitar zonas de manejo dentro de un lote productivo agrícola a través del procesamiento de datos georreferenciados

Autores: Galarza, Romina
Mastaglia, Nicolás
Director: Martínez, César
Co-Director: Albornoz, Enrique Marcelo
Asesor Temático: Kemerer, Alejandra

14 de agosto de 2013

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
R. Galarza, N. Mastaglia, C. E. Martínez, E. M. Albornoz & A. Kemerer; "Desarrollo de un método para delimitar zonas de manejo dentro de un lote productivo agrícola a través del procesamiento de datos georreferenciados (Undergraduate project)"
Facultad de Ingeniería y Ciencias Hídricas - Universidad Nacional del Litoral, 2013.

Agradecimientos

"No solo no hubiéramos sido nada sin ustedes, sino con toda la gente que estuvo a nuestro alrededor desde el comienzo, algunos siguen hasta hoy... Gracias..."

Romina Galarza y Nicolás Mastaglia
Santa Fe, setiembre de 2013.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
R. Galarza, N. Mastaglia, C. E. Martínez, E. M. Albornoz & A. Kemerer; "Desarrollo de un método para delimitar zonas de manejo dentro de un lote productivo agrícola a través del procesamiento de datos georreferenciados (Undergraduate project)"
Facultad de Ingeniería y Ciencias Hídricas - Universidad Nacional del Litoral, 2013.

Resumen

La Agricultura de Precisión provee un conjunto de principios y herramientas que permiten manejar la variabilidad espacio-temporal en la producción agrícola. Dentro de sus incumbencias está la identificación de zonas de manejo en un lote productivo. Éstas son subregiones de un lote particular, que poseen características homogéneas y para las cuales una proporción de insumos resulta apropiada.

En el presente trabajo se desarrolla un método integrador que comprende toda la cadena de identificación/clasificación de zonas de manejo. En primer lugar permite fusionar, en una estructura común, datos de entrada georreferenciados provenientes de diferentes sensores y con diferentes resoluciones. Luego, un algoritmo de agrupamiento de lógica difusa permite identificar y agrupar diferentes regiones. Mediante procesamiento de imágenes se mejoran los resultados a fin de identificar los bordes de las zonas de manejo. Finalmente, el método provee un archivo *shape* que es utilizado en las máquinas agrícolas con aplicación de dosis variable.

Se realizaron diferentes pruebas con conjuntos de datos provistos por la Estación Experimental Agropecuaria Paraná del INTA (Entre Ríos), los cuales se analizaron utilizando distintos parámetros en cada etapa. Los resultados demostraron que el método logra automatizar el proceso, siendo una herramienta más rápida y sencilla que el procedimiento llevado a cabo en la actualidad. Se pretende que este trabajo sirva como base para el desarrollo de un software que brinde a los expertos una gran ventaja en la operatividad de las máquinas agrícolas.

El trabajo realizado en este proyecto final dio lugar a las siguientes publi-

caciones:

- Aceptado para presentación en el 5º Congreso de Agroinformática, 42 Jornadas Argentinas de Informática (JAIIO), 16 al 20 de setiembre de 2013, Córdoba.
- Enviado al XVII Encuentro de Jóvenes Investigadores de la UNL, 4 y 5 de setiembre de 2013, Santa Fe.

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1

Introducción

La agricultura de precisión (AP) es el conjunto de herramientas y principios que permiten manejar la variabilidad espacio-temporal en la producción agrícola, con el fin de maximizar el rendimiento y reducir el impacto ambiental, entre otros [1]. Una de las tareas que se llevan a cabo dentro de la AP es la identificación de las zonas de manejo que posee un lote productivo. Las zonas de manejo son subregiones dentro de un mismo lote que poseen características homogéneas, para las cuales, una proporción única de insumos resulta apropiada [2].

El cultivo sembrado en un campo registra un comportamiento dispar a lo largo de toda su extensión debido a múltiples factores que influyen de diferente manera en el crecimiento y posterior rendimiento del mismo. La condición química y física del suelo, la topografía y su efecto en la disponibilidad de agua para los cultivos, el exceso de lluvias, las sequías, granizo y las enfermedades son sólo algunos de los orígenes de la variabilidad.

Identificar las áreas que, dentro de un mismo lote, se comportan de manera homogénea posibilita un adecuado manejo y un uso eficiente de los insumos (fertilizantes, semillas, tierra, tiempo, etc.). De esta manera, el productor puede obtener mayores rendimientos y/o maximizar el beneficio económico al contar con datos que permitan realizar un manejo diferenciado, poniendo más énfasis en las zonas que poseen un mayor potencial productivo [3].

Podemos nombrar las cuatro etapas más relevantes que posibilitan obtener un buen resultado con la AP [4]:

- Medición de variables que caracterizan la variabilidad de la productividad: es el proceso que da inicio al ciclo de AP. Involucra el uso de tecnologías como sistemas de posicionamiento global (GPS), sistemas de información geográfica (SIG), instrumentos topográficos, sensores remotos, sensores directos y otros medios electrónicos para obtener datos del cultivo.

- Análisis de datos: los programas son los actores principales en la etapa de análisis, brindando a los encargados de tomar las decisiones, información técnica del campo analizado.

- Toma de decisiones: en base a la información obtenida de los análisis realizados, un especialista es el que finalmente decide la cantidad de zonas de manejo que se emplearán en el lote en función del interés económico y del impacto sobre el ambiente.

- Implementación de las decisiones: la aplicación diferencial de insumos requiere de maquinaria especializada tales como cosechadoras, sembradoras, pulverizadoras asistidas por GPS.

La AP, además de pretender realizar un apropiado tratamiento agronómico que se ajuste a las necesidades del cultivo, procura reducir el impacto ambiental que ocasiona la aplicación desmedida de insumos. Los ejemplos que se presentan en [5] demuestran que el manejo por ambientes que realiza la AP, no solo puede producir un aumento en la productividad del campo, sino que confirma la reducción de costos.

El proyecto se organiza como se detalla a continuación. En el resto del Capítulo 1 se desarrolla la motivación, estado del arte, objetivos y alcances del presente trabajo. Luego, en el Capítulo 2 se presentan los fundamentos teóricos de los diferentes temas que abarca el proyecto. En el Capítulo 3 se presenta el diseño del método propuesto. Seguidamente, el Capítulo 4 muestra los experimentos y resultados obtenidos con diferentes lotes y variables. Finalmente, en el Capítulo 5 se encuentran las conclusiones y trabajos futuros.

1.1 Motivación

Este proyecto nace a partir del problema planteado por profesionales de la Estación Experimental Agropecuaria (EEA) Paraná del INTA (Entre Ríos) en el proceso de identificación de zonas de manejo que llevan a cabo en la actualidad. La idea es automatizar alguno de los procesos llevados adelante en el tratamiento diferenciado aplicado a un lote productivo agrícola, para generar un proceso que integre toda la cadena de identificación/clasificación de zonas de manejo que llevan adelante hoy día.

Con esta herramienta se pretende que los especialistas de la EEA Parana del INTA puedan unificar o minimizar el uso de diferentes programas a lo largo de toda la etapa de clasificación. Además, se beneficiaría al sector agrícola y a los usuarios de la AP con una herramienta sencilla y eficiente que posibilite el uso de los datos georreferenciados de manera inteligente, sugiriendo la aplicación justa de insumos en el lugar exacto. El uso de esta herramienta permitirá maximizar la productividad y la rentabilidad de cada ambiente con sustentabilidad, mejorando la gestión del campo y el cuidado de los recursos naturales y el ambiente.

1.2 Estado del arte

El continuo avance y perfeccionamiento tecnológico ha posibilitado que cualquier sembradora o pulverizadora disponible en el mercado cuente con un sistema tecnológico de dosis variable que posibilita, mediante una computadora incorporada, indicar cuánto fertilizante usar y dónde se lo debe aplicar. Empresas como John Deere y Case/New Holland implementan en sus maquinarias los sistemas integrales AMS[1] y AFS [2] respectivamente, que utilizan métodos para la clasificación que permiten agrupar en un número preestablecido de ambientes diferentes. Para ésto se necesita tener un conocimiento previo de las características del lote, lo que puede incurrir en el error de sobre o subestimar la cantidad de zonas [6, 7].

[1] `http://www.deere.com.ar`

[2] `http://www.caseih.com/argentina/Products/AFS/ANALISIS/Pages/Intro.aspx`

En el mercado informático existe software comercial de manejo de información agrícola, como el caso de SSToolbox[3] y Farm Works[4]. Este último muestra detalladamente la representación gráfica de las distintas variables analizadas, pero carece de la posibilidad de realizar un análisis en profundidad sobre las posibles zonas de manejo que se encuentran dentro de un lote [8]. El SSToolbox posee diversas aplicaciones agronómicas, entre ellas se pueden citan la generación, el procesamiento y el manejo de información georreferenciada, el análisis de la misma y la generación de recomendaciones de manejo agronómico. Es un software complejo que delimita las zonas de manejo únicamente por posición de paisaje [9].

En la EEA Paraná del INTA, el proceso de delimitación de zonas de manejo se lleva a cabo actualmente en varias etapas. Algunas de las mismas constituyen un laborioso trabajo manual y se realiza con el apoyo de software diverso, entre los cuales podemos nombrar: MZA (Management Zone Analyst) [10], gvSIG, procesador de textos, procesador de cálculos. El software gratuito MZA sólo realiza la clasificación e identificación de las sub-regiones, lo cual implica el uso de otros software para el pre y post procesamiento de los resultados. Esto conlleva a una compleja y tediosa tarea para arribar al resultado final.

1.3 Objetivos del Proyecto Final

El objetivo general del proyecto es desarrollar un método para automatizar las distintas etapas llevadas a cabo por la EEA Paraná del INTA, que identifique los distintos ambientes presentes en un lote a través del ingreso de datos georreferenciados.

Los objetivos específicos son los siguientes:

- Seleccionar un algoritmo de conversión de coordenadas.
- Diseñar un algoritmo para fusionar el conjunto de datos de entrada.
- Diseñar un algoritmo para clasificar las zonas de manejo.

[3] http://www.sstsoftware.com/sstoolbox.htm
[4] http://www.farmworks.com

- Seleccionar índices que permitan determinar, de una manera objetiva, el número óptimo de zonas de manejo.
- Determinar las etapas necesarias para generar un archivo que se pueda utilizar en las máquinas agrícolas.
- Desarrollar una interfaz gráfica que permita acceder a las funcionalidades del método.
- Evaluar el desempeño del método propuesto sobre casos reales.

1.4 Alcances del Proyecto Final

Nuestro trabajo se centra en la etapa de Análisis de Datos, por lo que se limitará a la detección y delimitación de ambientes, lo que representa sólo uno de los pasos en la aplicación de sistemas de manejo de sitio-específico. Una vez delimitados correctamente los ambientes, éstos deben ser manejados de manera independiente, ajustando la utilización de insumos de acuerdo al potencial de cada uno, tarea que es llevada a cabo por personas capacitadas en materia agronómica [11].

En este trabajo se hará uso de la información suministrada por la EEA Paraná del INTA, quedando fuera de propósito las tareas de medición de variables en el lote productivo.

El proyecto no avanzará sobre la realización de un sistema comercial.

CAPÍTULO 2

Fundamentos teóricos

2.1 Geodesia cartográfica

La geodesia es la ciencia encargada de la medición y representación cartográfica de la superficie terrestre [12]. Se define como "Geoide" la superficie teórica de la Tierra que une todos los puntos que tienen igual gravedad (Figura 2.1). Esta superficie no es uniforme, sino que representa una serie de irregularidades causadas por las distintas composiciones minerales del interior de la Tierra y de sus distintas densidades [13].

En geodesia, es usual hacer referencia a la posición espacial de puntos sobre una superficie que aproxime la forma de la Tierra. Para esto se define como superficie geométrica de referencia la que se corresponde con un elipsoide de revolución. De esta forma se logra representar toda la superficie terrestre mediante un modelo matemático. La forma y dimensión del elipsoide de revolución terrestre queda determinada por dos parámetros (por ejemplo, el semieje mayor a y el semieje menor b), además es necesario definir su ubicación y orientación [12].

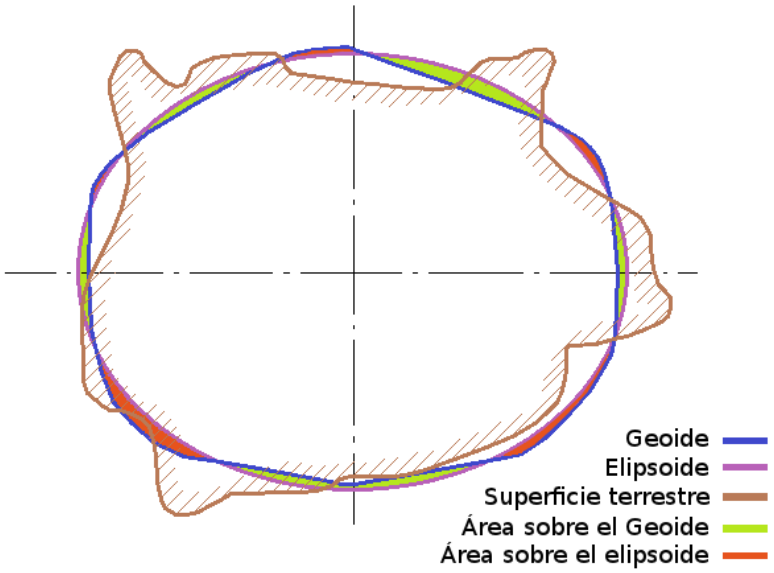


Figura 2.1: Geoide: representación de áreas de la Tierra según su posición. Figura adaptada de [13].

2.1.1 Sistemas de referencia

Cuando se trata de definir la forma, dimensión y ubicación de un objeto irregular, (como es el caso de la superficie de la Tierra) se reduce el problema a la determinación de la posición espacial de puntos. Justamente en topografía y geodesia, el punto es la entidad generadora de la superficie terrestre [12]. Existen diferentes formas de expresar la posición espacial de un punto. A continuación se desarrollan brevemente algunas de ellas.

Coordenadas tridimensionales cartesianas

Las coordenadas tridimensionales cartesianas es un sistema que está basado en tres ejes perpendiculares entre sí. Un punto P se ubica mediante los valores X, Y, Z que se miden generalmente en metros [14]. Desde el punto de vista geométrico, las coordenadas cartesianas, aunque adecuadas para el cálculo, no proporcionan una idea clara e inmediata de la posición de los puntos sobre la superficie terrestre [12].

Coordenadas tridimensionales geográficas

Otro sistema es el de coordenadas tridimensionales geográficas en el cual se necesita definir un centro de la tierra (O), radio mayor (a) y radio menor (b) del elipsoide y un coeficiente de aplastamiento (e) [14]. En la Figura 2.2 se puede ver cómo se sitúa un punto P mediante tres coordenadas:

λ: longitud, ángulo entre el plano del meridiano de origen y el meridiano sobre el cual se sitúa P.

ω: latitud, ángulo entre la perpendicular al elipsoide que pasa por P y el plano ecuatorial.

h: altura de P por encima del elipsoide, medida sobre la perpendicular entre P y el elipsoide.

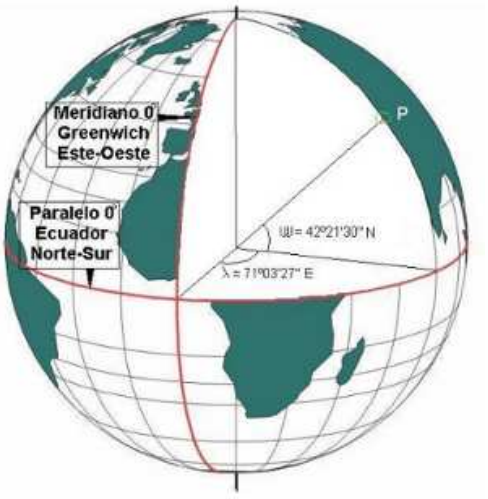


Figura 2.2: Coordenadas tridimensionales geográficas. Figura adaptada de [13].

El DATUM (Figura 2.3) se define como el punto tangente al elipsoide y al Geoide, donde ambos son coincidentes [13]. Cada DATUM está compuesto por:

- un elipsoide, definido por a, b y e.
- un punto llamado "fundamental" en el que el elipsoide y la Tierra son tangentes.

La República Argentina adoptó en mayo de 1997 el marco de referencia geodésico denominado POSGAR 94. Este marco usa el elipsoide denominado

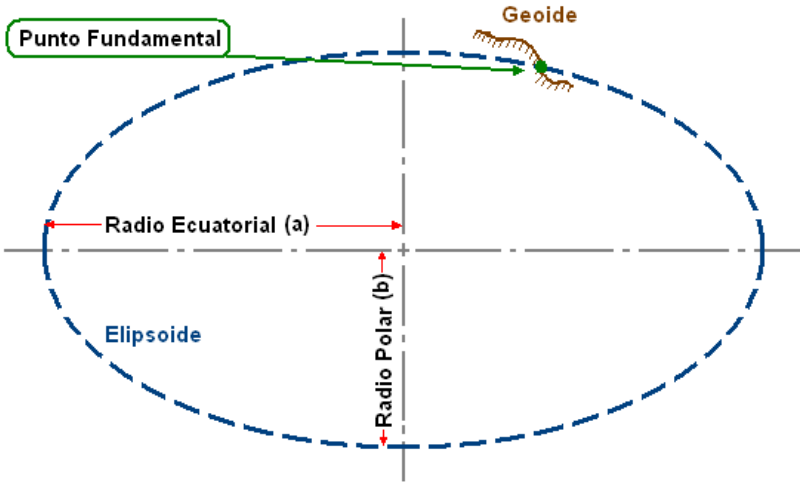


Figura 2.3: Representacion gráfica del DATUM. Figura adaptada de [13].

WGS84 (World Geodetic System 1984) definido por los siguientes parámetros:

- $a = 6378137$ m.
- $e = 1/298.257223563$.

Aunque la utilización de las coordenadas geográficas resulta de gran utilidad en geodesia, su uso resulta incómodo para emplearlo en determinadas aplicaciones. En efecto, teniendo en cuenta que estas coordenadas se expresan en unidades angulares (grados, minutos y segundos de latitud y longitud), cuando se pretende determinar distancias entre puntos o direcciones definidas por éstos, se presentan limitaciones [12].

Coordenadas planas

Para representar sobre un plano la superficie del elipsoide se utilizan las coordenadas planas. De esta manera se obtiene una proyección o representación plana de la Tierra. Cuando se realiza esta operación, se pierde parte de la información, es decir h. Pero ésto no representa una complicación ya que este dato se maneja por separado [14].

La proyección del globo terrestre supone un problema ya que no existe modo alguno de representar fielmente toda la superficie desarrollada sin deformarla. Esto obedece a que la superficie de una esfera no es desarrollable en

una representación plana. Existen diferentes proyecciones que tratan de minimizar, en la medida de lo posible, las deformaciones sufridas al representar la superficie terrestre de manera plana (proyecciones conforme, equivalentes, afilácticas) [13].

Un punto P ahora será situado con respecto a un punto O arbitrario y a dos distancias, una al Norte y otra al Este, medidas sobre dos ejes perpendiculares a partir de O. Generalmente, estas distancias están medidas en metros. Existen numerosos tipos de proyecciones diferentes como las proyecciones Mercator (utilizada por el SOHMA) y Gauss (utilizada por el Geográfico Militar) [14].

En general, en territorios extendidos en dirección Norte-Sur como el de la República Argentina, son adecuadas las proyecciones cilíndricas transversales. En ellas se utiliza como superficie intermedia un cilindro tangente a la superficie de referencia terrestre a lo largo de un meridiano, llamado meridiano de tangencia o meridiano central. En nuestro país se adoptó, en el año 1925, la proyección conforme Gauss-Krüger (cilíndrica transversal conforme) como sistema de representación plano.

Las deformaciones lineales crecen rápidamente con la distancia al meridiano central, por lo que representar los puntos de toda una superficie elipsóidica muy extendida en dirección Este-Oeste no es aconsejable, ya que las deformaciones (aunque calculables) serían demasiado grandes y esto distorsionaría apreciablemente las figuras. En este sentido, y con la finalidad de limitar las deformaciones, es recomendable la división de la superficie en zonas, que tendrán una representación plana independiente una de otra.

Es por esto que el territorio de la República Argentina se divide en 7 fajas meridianas de 3° de ancho cada una, con meridianos centrales en las longitudes: 72°, 69°, 66°, 63°, 60°, 57°, 54° al oeste de Greenwich.

Para distinguir a cada una de las fajas se emplean sólo números positivos y se asignan a los meridianos centrales las siguientes ordenadas:

Al meridiano 72° de la 1ra. faja, la ordenada 1.500.000 metros,

Al meridiano 69° de la 2ra. faja, la ordenada 2.500.000 metros,

Al meridiano 66° de la 3ra. faja, la ordenada 3.500.000 metros,

Al meridiano 63° de la 4ta. faja, la ordenada 4.500.000 metros,

Al meridiano 60° de la 5ta. faja, la ordenada 5.500.000 metros,

Al meridiano 57° de la 6ta. faja, la ordenada 6.500.000 metros,

Al meridiano 54° de la 7ma. faja, la ordenada 7.500.000 metros.

Debe tenerse en cuenta que en la representación plana de una región con

más de una faja meridiana se pierde continuidad espacial por lo que carece de sentido integrar puntos cuyas coordenadas planas fueron calculadas con distintos meridianos centrales. En estos casos, será necesario recalcular las coordenadas proyectivas de los puntos mencionados utilizando un único meridiano central [12].

2.1.2 GPS

La mayoría de los datos que se utilizan para el desarrollo de nuestro proyecto son obtenidos con la ayuda de sistemas GPS. El programa NAVSTAR, GPS (Navigation System Timing And Ranging, Global Positioning System) fue iniciado en diciembre de 1973. La responsabilidad del desarrollo y mantenimiento del sistema recae en el Departamento de Defensa de los Estados Unidos, ya que el sistema fue concebido para uso militar. Cuenta con una constelación de 24 satélites que se ubican en 6 órbitas planas prácticamente circulares.

GPS es un sistema que tiene como objetivo la determinación de las coordenadas espaciales de puntos respecto a un sistema de referencia mundial. Para la obtención de coordenadas el sistema se basa en la determinación simultánea de las distancias a cuatro satélites como mínimo. Estas distancias se obtienen a partir de las señales emitidas por los satélites, las que son recibidas por receptores especialmente diseñados. Los puntos pueden estar ubicados en cualquier lugar del planeta, pueden permanecer estáticos o en movimiento y las observaciones pueden realizarse en cualquier momento del día [12].

2.1.3 Imágenes geoTIFF

El formato TIFF es el formato de imágenes raster más popular y versátil actualmente. Se puede utilizar tanto para almacenar como para transferir imágenes satelitales, fotografías aéreas, modelos de elevación, mapas o los resultados de muchos tipos de análisis geográficos. En los últimos años muchos usuarios de este tipo de imágenes han instado a los proveedores de datos geográficos para proporcionar imágenes en formato TIFF. Es el único formato de dominio público, capaz de soportar la compresión y la inclusión de metadatos geográficos.

El formato GeoTIFF es totalmente abierto, de dominio público y no pro-

pietario. El propósito de GeoTIFF es implementar los metadatos geográficos formalmente, usando etiquetas y estructuras compatibles con TIFF [15]. La información adicional que se incluye puede ser el tipo de proyección, sistema de coordenadas, elipsoide, DATUM y todo lo necesario para que la imagen pueda ser automáticamente posicionada en un sistema de referencia espacial [16].

2.2 Fusión de datos

La utilización de un conjunto de datos multiparamétricos, los cuales son obtenidos a través de diversos sensores, obliga a fusionar la información de manera tal que se adecue a una estructura común. Hay que tener en cuenta que la dimensión de las variables de entrada puede ser diferente. Ésto significa que se debe realizar una transformación de los datos de entrada crudos en datos disponibles fácilmente, y relacionables con otras fuentes de datos georreferenciados. De esta manera se establece un marco de coordenadas global y común a los datos proporcionados por múltiples sensores, eventualmente heterogéneos.

Uno de los principales objetivos de la fusión de datos es combinar la información obtenida a través de las diferentes fuentes para tomar una mejor decisión, realizando para ello una reducción de la imprecisión y la incertidumbre mientras que se incrementa la robustez [17].

2.2.1 Interpolación

La representación de los valores que no han sido obtenidos experimentalmente, será lograda mediante la interpolación. Los métodos de interpolación a partir de puntos pueden dividirse en dos tipos fundamentales [18]:

1. Métodos globales, utilizan toda la muestra para estimar el valor en cada nuevo punto.

2. Métodos locales, utilizan sólo los puntos de muestreo más cercanos.

El conjunto de puntos utilizado en los métodos locales se llamará conjunto de interpolación; y constará de aquellos puntos cuya distancia al punto de

interpolación sea inferior a cierto umbral o bien, serán los n puntos más cercanos al punto de interpolación.

2.2.2 Métodos de interpolación

Krigeado

El krigeado es un método de interpolación exacto y local que pondera el peso de cada punto muestral X_i en un punto no muestral X_0 según una función estocástica de la distancia entre dichos puntos. Su fundamento conceptual deriva de la teoría de las variables regionalizadas. Se trata, en esencia, de un método geoestadístico consistente en la búsqueda de unos interpoladores óptimos que producen unos residuos insesgados y con mínima varianza. El krigeado presenta una ventaja sustancial con respecto a otros interpoladores, pues permite, una vez seleccionado el semivariograma que mejor explica la variabilidad de la variable a interpolar, seleccionar el tipo de malla de muestreo y el número de puntos muestrales mínimo para obtener un error predeterminado. Aunque el krigeado es un método de interpolación teóricamente muy recomendable para su inclusión en los Sistemas de Información Geográfica, algunos autores observan que, en la práctica, su eficacia es comparable a la de otros métodos más simples y de menor requerimiento computacional [19].

Medias móviles ponderadas por la distancia

El método de interpolación de medias móviles ponderadas por la distancia es ampliamente usado en la modelización de superficies. Se basa en la idea intuitiva de que las observaciones más cercanas deben tener más peso en la determinación del valor interpolado en un punto X_0. Se trata de un método exacto y local que estima el valor de la variable Z en un punto no muestral X_0. Probablemente el mayor problema que presenta este método es que los valores interpolados son medias ponderadas que siempre toman valores entre el máximo y el mínimo de los puntos muestrales, lo que reduce su eficacia para modelizar las cotas más altas o bajas de una superficie topográfica, en el caso de que las mismas no pertenezcan al conjunto de puntos muestrales [19].

Funciones de base radial

Las funciones de base radial (FBR) comprenden un amplio grupo de interpoladores exactos y locales que emplean una ecuación de base dependiente de la distancia entre el punto interpolado y los puntos muestrales vecinos. Entre las diversas FBRs que podemos encontrar, la función multicuadrática es la que mejores resultados obtiene en términos de evaluación estadística y visual de la superficie modelizada. Conviene observar que un factor de suavizado elevado producirá una superficie muy suavizada que probablemente se alejará sensiblemente de la geometría de la superficie real [19].

Triangulación lineal

La triangulación lineal es un método exacto de interpolación basado en la generación previa de una malla irregular de triángulos (TIN) cuyos vértices coinciden con los puntos muestrales. Dicha malla se obtiene mediante la conocida triangulación de Delaunay. La interpolación de puntos dentro de la topología obtenida se realiza suponiendo que dichos puntos pertenecen a la superficie plana de primer orden que se apoya en los vértices de cada triángulo [19].

Curvas adaptativas (Splines regulares)

La técnica de splines consiste en el ajuste local de ecuaciones polinómicas. La forma de la superficie final depende de un parámetro de tensión que hace que el comportamiento de la superficie interpolada tienda a asemejarse a una membrana más o menos tensa o aflojada que pasa por los puntos de observación. La ventaja fundamental del método de splines respecto a los basados en medias ponderadas es que, con estos últimos, los valores interpolados nunca pueden ser ni mayores ni menores que los valores de los puntos utilizados para interpolar. Por tanto resulta imposible interpolar correctamente máximos y mínimos [19].

2.3 Clustering

El análisis de agrupamiento o clustering se define como la tarea de aglomerar objetos en grupos (clusters) utilizando alguna medida de similitud entre

ellos. De este modo, dos objetos que pertenecen a un mismo grupo son más parecidos entre sí que a objetos de otros grupos.

Las técnicas de agrupamiento se pueden dividir en dos grandes categorías. Por un lado tenemos los algoritmos Jerárquicos que construyen una jerarquía de grupos iterativamente. Por otro lado están los algoritmos de Particionamiento en los que el número de grupos se determina de antemano y las observaciones se van asignando a los grupos en función de su cercanía [20].

Una distinción complementaria puede ser propuesta según la forma de clasificación: dura o difusa. Un algoritmo de agrupamiento duro asigna cada elemento a un solo grupo. Mientras que, un método de agrupamiento difuso asigna cada elemento a varios grupos, con diferentes grados de pertenencia. Un agrupamiento difuso se puede convertir en una agrupación dura mediante la asignación de cada elemento al grupo con el que tiene el mayor grado de pertenencia [21].

Existen técnicas de clustering supervisado que necesitan, además de los datos de entrada, información adicional como es el caso de la experiencia personal para determinar los agrupamientos presentes en las variables ingresadas [22]. En cambio, el clustering no supervisado no precisa conocimiento a priori para producir agrupaciones naturales en los datos del espacio de atributos [23, 24].

El clustering es una de las tareas más útiles para el análisis que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Uno de los pasos claves para el proceso de clustering es la elección de un algoritmo adecuado [25]. Un algoritmo de clustering se caracteriza principalmente por tener definido una medida de proximidad y un criterio de agrupamiento, y su eficacia para definir los grupos dependerá del conjunto de datos:

- *Medida de proximidad:* es una medida que cuantifica que tan "similares" son dos puntos del conjunto de datos.
- *El criterio de agrupamiento:* se puede expresar a través de una función de costo o algún otro tipo de regla.

Una gran cantidad de métodos de clustering se proponen en la literatura [2, 21, 26]. Los algoritmos de clustering se pueden clasificar según:

- El tipo de datos de entrada para el algoritmo.
- El criterio de agrupamiento para definir la similitud entre datos.
- La teoría y los conceptos fundamentales en los que las técnicas de agrupamiento se basan (por ejemplo, la teoría difusa, estadística, etc.).

2.3.1 Algoritmo K-Means

Dentro de los algoritmos de clustering particionales K-Means es uno de los algoritmos más comúnmente utilizado [27]. Se basa en la optimización de una función objetivo que se describe por la ecuación 2.1. La idea principal es definir k centroides, uno para cada grupo. Estos centroides deben colocarse de manera estratégica debido a que una ubicación diferente ocasiona un resultado diferente. Por lo tanto, la mejor opción es colocarlos lo más lejos posible uno de otro. El siguiente paso consiste en tomar cada dato que pertenece a un conjunto dado y asociarlo al centroide más cercano. Luego, con todos los puntos asociados a un centroide, éste es recalculado como el punto medio de este grupo de datos. Se repite este paso recalculando los centroides hasta que no se producen más reasignaciones. Por último, este algoritmo tiene como propósito minimizar una función objetivo, en este caso una función de error al cuadrado. La función objetivo es:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \tag{2.1}$$

donde $\left\| x_i^j - c_j \right\|^2$ es una medida de distancia elegida entre en punto x_i^j y el centro del cluster c_j, J es un indicador de la distancia total de los datos (n) a sus respectivos centroides.

2.3.2 Algoritmo Fuzzy C-Means

Fuzzy C-Means (FCM) es un método de agrupamiento del estado del arte en reconocimiento de patrones, permite que un dato pertenezca a dos o más grupos y se basa en la minimización de la siguiente función objetivo:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2 \tag{2.2}$$

donde $1 < m < \infty$, m es un índice de ponderación difuso que determina la falta de claridad de los grupos, x_i es el i-ésimo dato del conjunto medido, u_{ij}^m es el grado de pertenencia de x_i al grupo j, c_j es el centro del grupo d-dimensional, N la cantidad de elementos del conjunto de datos, C la cantidad de grupos y $|| * ||$ es cualquier norma que exprese la similitud entre los datos medidos y el centro.

FCM se lleva a cabo a través de un proceso iterativo de optimización de la función objetivo (2.2), con la actualización de la matriz de pertenencia u_{ij} y los centros del grupo c_j por:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{2.3}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{2.4}$$

El proceso se detiene cuando $max_{ij} \left\{ |u_{ij}^{k+1} - u_{ij}^k| \right\} < \varepsilon$, donde ε es un criterio de finalización entre 0 y 1, mientras que k es la iteración. Este procedimiento converge a un mínimo local [28, 29].

2.3.3 Algoritmo Jerárquico

Los algoritmos de clustering basados en agrupación jerárquica no proporcionan una única partición del conjunto de datos, sino que proporcionan una amplia jerarquía de grupos que se unen entre sí a distancias determinadas. Se basa en la idea central de vincular los elementos a los objetos que se encuentren más cercanos a ellos. Un grupo puede ser descrito en gran medida por la distancia máxima necesaria para conectar las partes del cluster. A diferentes distancias, diferentes grupos se forman, lo que se puede resumir mediante un dendograma (Figura 2.4). Un dendograma es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de clusters [30]. En este gráfico, el eje y marca la distancia a la que los grupos se fusionan, mientras que los objetos se colocan a lo largo del eje x de tal manera que los grupos no se mezclan.

2.3.4 Índices de validación de resultados

Uno de los temas más importantes en el análisis de cluster es la evaluación de los resultados del clustering para encontrar la partición que mejor

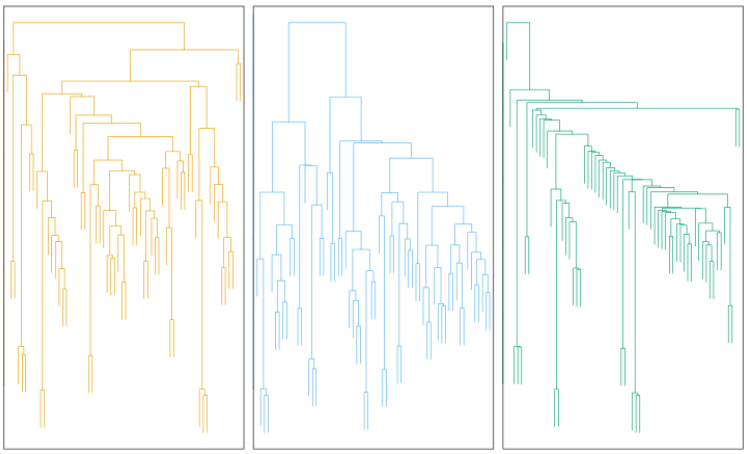


Figura 2.4: Ejemplos dendograma - Alg. jerárquico.

se ajuste a los datos subyacentes. La validación de los resultados tiene como objetivo verificar por medio de criterios y técnicas adecuadas la exactitud de los resultados del algoritmo de clustering. Dado que los algoritmos definen grupos que no se conocen a priori, independientemente del método seleccionado, la partición final de los datos requiere algún tipo de evaluación. Uno de los problemas que se enfrenta en el clustering es decidir el número óptimo de agrupaciones que se ajuste al conjunto de datos en cuestión. En esta sección, se discuten los métodos adecuados para la evaluación cuantitativa de los resultados del clustering, conocidos como métodos de validación de cluster. Sin embargo, se debe mencionar que dichos métodos proporcionan un índice de la calidad de la partición resultante, entendida en términos de centroides bien separados y conformación de grupos bien compactos. Por lo tanto sólo se pueden considerar como una herramienta más a disposición de los expertos con el fin de evaluar los resultados del agrupamiento.

A continuación se presentan los índices para evaluar al algoritmo de clustering FCM. Estos se pueden dividir en dos categorías, la primera utiliza sólo los valores de pertenencia de la matriz u_{ij} de la partición de datos difusa. Mientras que, el otro índice implica tanto la matriz u_{ij} como el conjunto de datos.

Índices de validez en los que sólo intervienen los valores de pertenencia

Para evaluar las características de la agrupación según el número de grupos, se calculan dos tipos de funciones de validez para los agrupamientos del algoritmo FCM.

El índice de performance difusa (FPI) [31] es la tasa de miembros compartidos entre clases y se define como:

$$FPI = 1 - \frac{nc}{nc-1}\left[1 - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{nc} u_{ij}^2\right]. \tag{2.5}$$

Donde nc es el número de cluster, N la cantidad de datos del conjunto y u_{ij} la matriz de pertenencia. Los valores de FPI pueden variar de 0 a 1. Un valor igual a 0 indica que las clases no comparten miembros, mientras que, un valor igual a 1 implica que existe un elevado número de miembros compartidos.

Bezdek [29] describió una segunda medida de validez del clustering conocida como la entropía de clasificación normalizada (NCE). Este índice es una estimación del grado de desorganización creada por un número de clases y se define como:

$$NCE = \frac{1}{1-(nc/N)}\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{nc} u_{ij}\cdot \log_a u_{ij}\right]. \tag{2.6}$$

Este índice también tiene valores entre 0 y 1. Si su valor es 0 existe un alto grado de organización, mientras que si es 1 existe una gran desorganización.

El número óptimo de zonas de manejo se obtiene cuando cada índice tiene el mínimo valor, lo que indica una baja cantidad de miembros compartidos y una mejor organización de clases.

Índice relacionado con los valores de pertenencia u_{ij} y el conjunto de datos

Dentro de los índices relacionados con los valores de pertenencia u_{ij} y el conjunto de datos, se encuentra el Xie-Beni (XB) [32]. Consideremos una partición difusa para el conjunto de datos $X = \{x_j; j = 1, ..., n\}$ donde $v_i\{i = 1, ..., nc\}$ representa los centros de cada cluster y u_{ij} la pertenencia de un elemento j perteneciente al cluster i. La desviación difusa d_{ij} del elemento x_j al cluster i, se define como la distancia entre x_j y el centro del cluster

ponderado por la pertenencia difusa del elemento j perteneciente al cluster i:

$$d_{ij} = u_{ij} \|x_j - c_i\| . \tag{2.7}$$

También, para un cluster i, la sumatoria de los cuadrados de la desviación difusa del elemento en X, denotado como σ_i, se denomina variación del cluster i, es decir:

$$\sigma_i = \sum_{j=1}^{n} d_{ij}^2. \tag{2.8}$$

Sea $\sum \sigma_i$ la *variación total* del conjunto de datos. La cantidad $\pi = \frac{\sigma}{n}$, se llama compactación del conjunto de datos, donde n es la cantidad de elementos del conjunto de datos. Asimismo, la separación de las particiones difusas se define como la distancia mínima entre los centros de grupo, es decir:

$$d_{min} = min \|c_i - c_j\| . \tag{2.9}$$

Entonces, el índice XB se define como:

$$XB = \frac{\pi}{(d_{min})^2}. \tag{2.10}$$

Es evidente que los valores pequeños de XB son esperados para grupos compactos y bien separados. Sin embargo, se observa que XB es monótono decreciente cuando el número de grupos nc se hace muy grande y cercano a n. Una forma de eliminar esta tendencia a la disminución del índice es determinar un punto de partida, $cmax$, del comportamiento monotónico y buscar el valor mínimo de XB en el intervalo [2, $cmax$]. Además, los valores del índice XB dependen del valor del índice difuso, entonces si $m \to \infty$ se tiene que $XB \to \infty$.

2.4 Procesamiento digital de imágenes

El procesamiento digital de imágenes es el conjunto de técnicas aplicadas a las imágenes digitales que tienen como objetivos mejorar la información pictórica para facilitar la interpretación humana o permitir la extracción de información de manera automática.

Provee un conjunto de métodos y algoritmos para manipular y transformar una imagen en una señal de utilidad. Los procesamientos en el dominio

espacial hacen referencia a operaciones que se aplican en forma directa sobre los píxeles. Se puede operar de manera individual o sobre la vecindad de cada píxel. Existen operaciones que se realizan sobre un dominio transformado, para lo cual es necesario aplicar las operaciones sobre la transformada de la imagen [33].

2.4.1 Filtrado de imágenes

Métodos de filtrado lineales

Los filtros espaciales lineales tienen la característica de ser, como su nombre lo indica, lineales e invariantes al desplazamiento. La salida del filtro es la convolución entre la imagen original y un kernel o máscara que produce un suavizado o acentuado de los detalles. Existe una correspondencia directa (uno-a-uno) entre el filtrado espacial lineal y el filtro en el dominio frecuencial [33].

Métodos de filtrado no lineales

Los filtros estadísticos de orden son filtros espaciales no lineales, cuya respuesta se basa en el ordenamiento (ranking) de los píxeles. Para cada píxel se toma una vencindad y a continuación, se sustituye su valor por el que resulte según el criterio de clasificación elegido [33]. Sea $g(x, y)$ la imagen a procesar y $\hat{f}(x, y)$ la imagen resultante, se aplican los siguientes procesos considerando los píxeles (s, t) de una vecindad S_{xy} centrada sobre cada píxel (x, y) original:

- Filtro de mediana: reemplaza el valor del píxel por la mediana estadística de los valores de intensidad en una vecindad del píxel (incluyendo el valor original del píxel en el cómputo), según

$$\hat{f}(x, y) = \underset{(s,t)\in S_{xy}}{mediana}\{g(s, t)\}. \tag{2.11}$$

- Filtro de máxima:

$$\hat{f}(x, y) = \underset{(s,t)\in S_{xy}}{max}\ \{g(s, t)\}. \tag{2.12}$$

  Este filtro es útil para encontrar los puntos más brillantes de una imagen. Además, se utiliza para reducir el ruido pimienta como resultado de la operación max.

- Filtro de mínima:

$$\hat{f}(x,y) = \min_{(s,t)\in S_{xy}} \{g(s,t)\}. \tag{2.13}$$

  Este filtro es útil para encontrar los puntos más oscuros en una imagen. Además, se utiliza para reducir el ruido sal como resultado de la operación *min*.

- Filtro de punto medio:

$$\hat{f}(x,y) = \frac{1}{2}\left[\max_{(s,t)\in S_{xy}} \{g(s,t)\} + \min_{(s,t)\in S_{xy}} \{g(s,t)\}\right]. \tag{2.14}$$

  El *filtro de punto medio* simplemente calcula el valor medio entre el máximo y el mínimo del área comprendida por el filtro.

- Filtro de alfa-media recortado:

$$\hat{f}(x,y) = \frac{1}{mn-d} \sum_{(s,t)\in S_{xy}} g_r(s,t). \tag{2.15}$$

  Este filtro calcula el promedio de los valores dentro de la máscara, pero con algunos valores iniciales y finales excluidos.

- Filtro de moda:

$$\hat{f}(x,y) = \operatorname*{moda}_{(s,t)\in S_{xy}} \{g(s,t)\}. \tag{2.16}$$

  La *moda* es el valor más frecuente de la zona.

2.4.2 Operaciones morfológicas

Las operaciones morfológicas simplifican imágenes y conservan las principales características de la forma de los objetos. La erosión y la dilatación son las operaciones morfológicas primarias.

Erosión

Sea un conjunto A y B de z^2, la erosión de A por B, denotada como $A \ominus B$ queda definida como [33]:

$$A \ominus B = \{z|(B)_z \subseteq A\}, \tag{2.17}$$

donde B comúnmente es conocido como máscara o kernel. De la ecuación (2.17) se puede interpretar que la erosión de A por B es el conjunto de todos los puntos z tales que B, trasladados por z, están contenidos en A.

Uno de los usos más simples de la erosión es la eliminación de detalles irrelevantes (en términos de tamaño) de una imagen binaria. En una imagen erosionada, el tamaño de los objetos se ve reducido y el ruido o detalles irrelevantes (aislados) es eliminado.

Dilatación

Sea A y B un conjunto en z^2, la dilatación entre A y B denotada como $A \oplus B$ queda definida como [33]:

$$A \oplus B = \{z|(\hat{B})_z \cap \ A \neq \varnothing\}. \tag{2.18}$$

Mediante la dilatación, los objetos crecen en su tamaño y algunos de los "espacios" dentro de ellos son rellenados.

2.4.3 Etiquetado por crecimiento de regiones

El crecimiento de regiones es un proceso que consiste en agrupar píxeles basándose en un criterio predefinido P. Los píxeles deben cumplir con un determinado criterio de conectividad o adyacencia, por ejemplo medida o criterio de similitud de los niveles de gris.

La técnica se inicia a partir de píxeles iniciales llamados "semillas" y se agrupan a éstas los píxeles vecinos que cumplen con una propiedad P seleccionada. El criterio es comprobado sobre una vecindad de 4 u 8 vecinos, de forma iterativa. El proceso termina cuando no existen más píxeles que cumplan con P o cuando se verifica algún criterio especificado (tamaño de la región, forma, etc.) [33].

2.4.4 Seguimiento continuo de contorno (borde)

Muchas aplicaciones requieren que los puntos presentes en el borde de una región sean ordenados en un sentido horario o antihorario. Los algoritmos para el reconocimiento del borde continuo devuelven una secuencia ordenada de puntos. Es necesario trabajar con imágenes binarias en donde los puntos

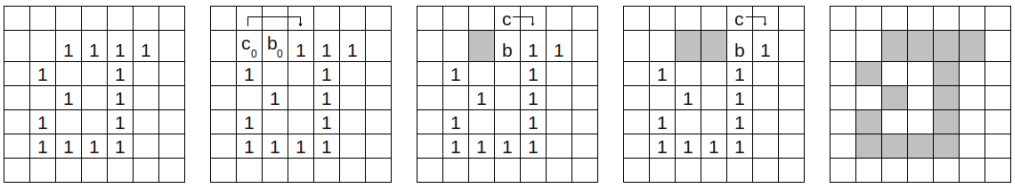


Figura 2.5: Ilustración del algoritmo de trazado de contorno: (a) sección de imagen original analizada; (b) paso 1; (c) paso 2; (d) paso 4; (e) fin del paso 5.

de los objetos y el fondo están identificados con 1 y 0 respectivamente. Dada una región binaria R o su contorno, un algoritmo para el seguimiento del borde de R consiste en los siguientes pasos [33]:

1. Tomar el punto inicial b_0, que es el punto superior izquierdo de la imagen con valor 1. Designamos c_0 al vecino oeste de b_0. Claramente, c_0 siempre es un punto del fondo. Se examinan los 8 vecinos de b_0 comenzando con c_0 y prosiguiendo en sentido horario. Se designa b_1 al primer vecino encontrado con valor 1, y c_1 es el punto (del fondo) inmediatamente precedente a b_1 en la secuencia. Se almacenan las ubicaciones de b_0 y b_1 para usarse en el paso 5.

2. Tomar $b = b_1$ y $c = c_1$.

3. Los 8 vecinos de b, comenzando por c y prosiguiendo en el sentido horario, se denominan n_1, n_2, ..., n_8. Encontrar el primer n_k con valor 1.

4. Tomar $b = n_k$ y $c = n_{k-1}$.

5. Repetir los pasos 3 y 4 hasta que $b = b_0$ y el siguiente punto de contorno sea b_1. La secuencia de puntos b encontrada cuando el algoritmo se detiene constituye el conjunto ordenado de puntos del contorno.

La Figura 2.5 ilustra los pasos del algoritmo descripto anteriormente. Algunas veces se hace referencia a éste algoritmo como *Algoritmo de Moore para el seguimiento de contorno* [33].

2.5 Archivos shape

El formato del archivo *shape* es un formato vectorial para almacenamiento de datos espaciales donde se guarda la localización de los elementos geográficos y los atributos asociados a ellos. Para visualizarlo, editarlo o convertirlo a otros formatos existen programas gratuitos y comerciales.

Fue creado por ESRI [1](Enviromental Systems Research Institute). Su implantación en la gama de productos de ESRI (ArcView, ArcInfo, actualmente ArcGIS) ha popularizado este formato hasta convertirlo en el más extendido dentro de los SIG vectoriales. Además se trata de un formato abierto con sus especificaciones disponibles en Internet, esto ha permitido que cada vez más compañías desarrollen aplicaciones compatibles con este formato convirtiéndolo en un estándar a la hora de representar información geográfica.

Un shapefile ESRI consta, como mínimo, de un archivo principal (.shp), un archivo de índice (.shx) y una tabla dBASE (.dbf). El archivo principal almacena las características geométricas de los elementos. Puede contener puntos, líneas o polígonos y cada vértice lleva implícitas sus coordenadas en un sistema de referencia concreto, que se establece en el fichero .prj. Este archivo posee longitud variable y está compuesto por registros, en el que cada uno describe una forma (*shape*) o un objeto geométrico con una lista de sus vértices. En el archivo de índice, cada registro contiene el desplazamiento del registro del archivo principal correspondiente desde el inicio del archivo principal. La tabla dBASE contiene los atributos de las características, con un registro por característica. La relación uno-a-uno entre la geometría y los atributos se basa en el número de registro. Los registros de los atributos en el archivo de dBASE deben estar en el mismo orden que los registros en el archivo principal [34].

2.5.1 Organización del archivo principal (.shp)

El archivo principal (.shp) contiene un encabezado de longitud fija seguido de registros de longitud variable. Cada registro de longitud variable se compone de una cabecera de longitud fija seguida por contenido de longitud variable.

[1] `http://www.esri.com`

El encabezado del archivo principal tiene una longitud de 100 bytes. La cabecera de cada registro almacena el número de registro y la longitud del contenido del registro. Este encabezado tiene una longitud fija de 8 bytes.

El contenido de los registros del shapefile consiste en un objeto (punto, línea o polígono) seguido de los datos geométricos del mismo. La longitud de este contenido depende del número de partes y los vértices del objeto [34].

2.5.2 Organización del archivo de índice (.shx)

Es un índice de las entidades geométricas que permite refinar las búsquedas dentro del archivo *shape* (.shp). Tiene una cabecera idéntica a la del archivo principal, tras la cual encontramos los registros que se encuentran en el mismo orden que en el archivo principal. Además contienen la posición del registro respectivo en el archivo principal (Offset) y su longitud.

El archivo de índice (.shx) contiene un encabezado de 100 bytes seguido de registros de 8 bytes. El registro i-ésimo en el archivo de índice almacena el desplazamiento y la longitud del contenido para el registro i-ésimo en el archivo principal. La longitud del contenido almacenado en el registro de índice es el mismo que el valor almacenado en la cabecera del archivo de registro principal [34].

2.5.3 Organización del archivo dBase (.dbf)

El archivo de dBASE (.dbf) contiene una tabla de datos en la que se registran los atributos de cada elemento. El formato dBase es un formato sencillo para almacenar datos estructurados.

En el caso de los shapefiles, las tablas dBase se emplean para asignar atributos numéricos, de texto o de fecha a los registros contenidos en el archivo principal. Cada registro debe estar asociado con una única entrada en la tabla, ambos archivos se vinculan mediante el número de registro en el archivo principal y el código en la tabla (OBJECTID) [34].

2.5.4 Archivos complementarios

Todos los archivos que componen un shapefile deben tener el mismo nombre, solo varía la extensión del archivo. Además de estos tres archivos reque-

ridos, opcionalmente se pueden utilizar otros para mejorar el funcionamiento en las operaciones de consulta a la base de datos, información sobre la proyección cartográfica, o almacenamiento de metadatos, estos archivos son [35]:

- Spatial Index (.sbn y .sbx) : se trata de un formato exclusivo de ESRI que almacena un índice espacial de los elementos.
- Metadatos (.xml) : en este archivo se almacenan los metadatos relativos al shapefile. Los metadatos guardan información sobre el contenido del archivo y su formato. Mediante el formato .xml se definen una serie de normas que permiten compatibilizar el intercambio de información entre distintos sistemas.
- Projection (.prj) : el archivo Projection guarda información para georreferenciar los datos geométricos que se posee en el *shape*. Si bien con el archivo *shape* (.shp) se define geométricamente una serie de elementos en un espacio bidimensional o tridimensional, si se quiere situar dicho elemento sobre el terreno se necesita referir los datos a un sistema de coordenadas. Estos datos necesarios por lo general están contenidos en este fichero.

CAPÍTULO 3

Desarrollo del método propuesto

En este capítulo se detallan las herramientas tecnológicas que intervienen a lo largo de la realización de este proyecto. Además se describe el diseño y la implementación del método para la sistematización del proceso, explicando el funcionamiento de las etapas constitutivas del método.

3.1 Herramientas utilizadas

Se ha decidido utilizar el lenguaje de programación C++ en base a la experiencia que hemos adquirido durante la carrera, por tratarse de uno de los lenguajes más utilizados para el desarrollo de este tipo de aplicaciones.

Para manipular las variables de entrada se utilizó la librería científica CGAL [36], que está desarrollada en C++. Además es una librería distribuída bajo la licencia de código abierto que facilita el acceso a los algoritmos geométricos de una forma robusta y eficiente. Esta librería ofrece estructuras de datos y algoritmos para llevar a cabo, por ejemplo, triangulaciones, diagramas de Voronoi, interpolaciones y análisis de formas; junto con una amplia variedad de algoritmos para procesamiento matemático complejo. Para el caso del manejo de las imágenes satelitales se utilizó la librería específica geotiff [37].

Para el procesamiento de imágenes se empleó la librería CImg [38], que es

una librería de código abierto y desarrollada en C++.

Para obtener un archivo *shape* una vez obtenidos los polígonos que delimitan las zonas de manejo, se utilizó la librería Shapefile C Library [39] con licencia LGPL. La misma ofrece la posibilidad de leer, escribir y actualizar el Shapefile de ESRI y el archivo de atributo asociado (.dbf).

Por otro lado, el desarrollo de la interfaz gráfica del prototipo se realizó a través del IDE Qt Creator en conjunto con la librería Qt [40]. Qt es una librería multiplataforma y un framework de interfaz de usuario para los desarrollos que utilizan C++ como lenguaje de programación.

3.2 Estructura del método

A continuación, se explica el funcionamiento del método de manera general tal como se observa en la Figura 3.1.

El método se divide en varias etapas consecutivas. Se utilizan conjuntos de datos multiparamétricos que son obtenidos a través de diversos sensores. El proceso comienza con la lectura de cada uno de los archivos de las variables utilizadas que contienen las coordenadas (longitud y latitud) y el valor muestreado en cada punto. Se transforman las coordenadas de todas las variables que no sean imágenes. Se crea un mallado con puntos distribuidos de forma equiespaciada a una distancia ingresada por el usuario y se calculan los valores correspondientes mediante algún método de interpolación de manera de fusionar la información. Luego se busca identificar y agrupar los elementos que poseen características similares a través de la utilización del algoritmo de agrupamiento de lógica difusa *Fuzzy C-Means* [29]. Se clasifican los puntos entre un valor mínimo y un máximo ingresado por el usuario y se validan los resultados mediantes indicadores que miden la calidad de la clasificación realizada. A continuación, comienza la etapa que filtra los resultados obtenidos anteriormente aplicando técnicas de procesamiento de imágenes. Se eliminan puntos espurios y zonas que no superan una determinada superficie. Por último se detectan los bordes de cada zona para obtener el archivo *shape* que es el resultado de todo el proceso.

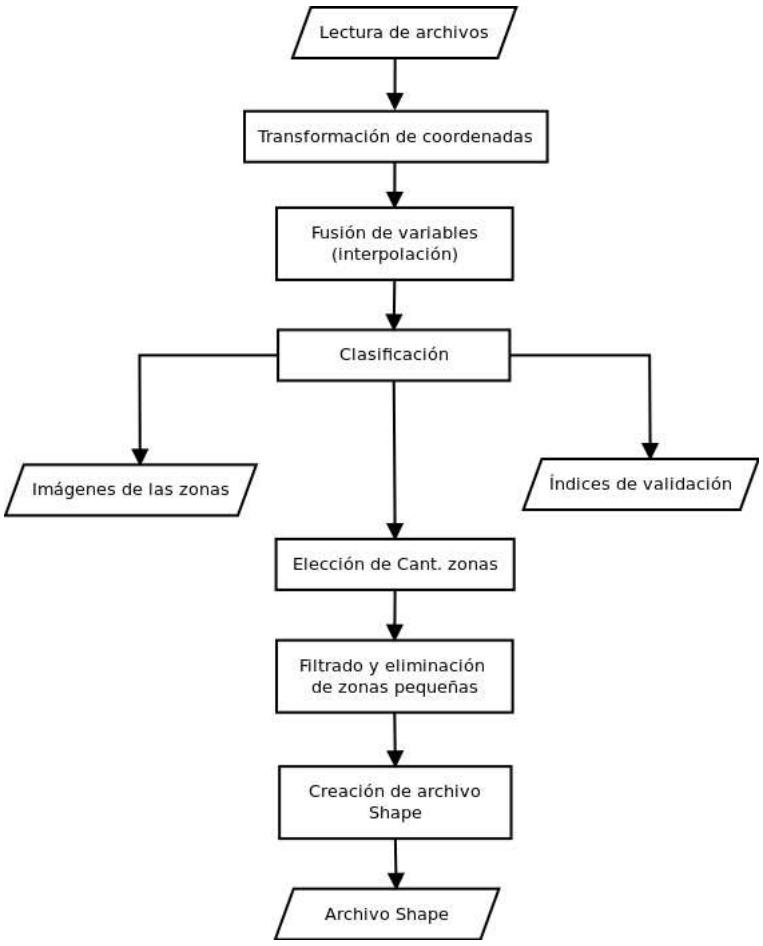


Figura 3.1: Diagrama general del algoritmo de detección de zonas de manejo.

3.2.1 Descripción de los datos de entrada

Se definen los datos de entrada con los cuales opera el método. Se determinan qué formatos de archivos son permitidos para su correcto funcionamiento.

Recolección y definición de los posibles datos de entrada

Los datos de entrada del método se pueden obtener utilizando diversas tecnologías tales como la teledetección (por ejemplo imágenes tomadas por satélites) o medición de muestras en el lote (por ejemplo rendimiento, altimetría, conductividad eléctrica, etc). Estas mediciones conducen a conjuntos de datos heterogéneos con diferentes representaciones digitales y resolucio-

nes, los cuales deben ser combinados para crear un único modelo espacial de la grilla de puntos en estudio.

Existen numerosos dispositivos que toman muestras de un lote para su posterior estudio. Estos equipos son capaces de generar mapas georreferenciados con la distribución espacial de alguna propiedad física del suelo. Por ejemplo, para medir la conductividad eléctrica del suelo existen instrumentos que, a partir de un circuito electrónico común, permiten medir la actividad eléctrica. Las máquinas agropecuarias son las encargadas de tomar las muestras del rendimiento en los lotes mediante los sistemas de monitoreo de cosecha. Éstos incorporan diversos programas propietarios que originan archivos de formatos particulares de cada marca. Algunos de los formatos de las marcas más reconocidas son: Ag Leader (.yld), AGCO (.rpt), CASE (.vy1, .yld, .vyg), CLASS (.dat), John Deere (.gsd, .gsy), New Holland (.yld, .log). Además de la diversidad de formatos se debe sumar el inconveniente de que algunos de ellos también suelen cambiar con los años y esto podría generar problemas de compatibilidad. Estos programas propietarios permiten exportar la información en diversos formatos para que pueda ser utilizada por otras personas y sistemas. Entre los formatos estándares más usados encontramos *txt*, *dat* y *csv*.

Por lo tanto se definen dos tipos de datos de entrada, por un lado las variables que fueron obtenidas mediante la medición de puntos, generalmente distribuidos irregularmente en el espacio y por otro las imágenes geotiff que se basan en una representación raster del terreno con una distribución regular de los puntos (píxeles).

Formato del archivo de entrada

Se utilizan como datos de entrada los archivos *txt*, *dat* y *csv* que en su interior contengan el valor de la coordenada (longitud y latitud con DATUM WGS84) y el valor de la variable en cada punto separados por coma.

Cada uno de estos archivos contiene las muestras de una variable en particular y se pueden ingresar tantos archivos como variables se tenga. Los datos dentro de cada archivo se ordenan de la siguiente manera: *Longitud, Latitud, Valor*.

En cuanto al procesamiento de imágenes se trabaja únicamente con imágenes georeferenciadas, debido a la necesidad de contar con información de la localización espacial de los píxeles. Este tipo de imágenes se representa mediante el formato *tiff* (GeoTiff, es decir tiff con coordenadas).

3.2.2 Transformación de coordenadas

Como se detalló anteriormente los datos de entrada son un conjunto de datos multiparamétricos que se encuentran georreferenciados. Se cuenta con dos grupos diferentes de datos de entrada, por un lado están las imágenes georreferenciadas con formato GeoTiff representadas mediante coordenadas planas y por otro los datos extraídos de los GPS, que se encuentran representados en coordenadas geográficas.

Aunque la utilización de las coordenadas geográficas resulta de gran utilidad en geodesia, su uso resulta incómodo para emplearlo en determinadas aplicaciones. Teniendo en cuenta que estas coordenadas se expresan en unidades angulares (grados, minutos y segundos de latitud y longitud), cuando se pretende determinar distancias entre puntos o direcciones definidas por éstos, se presentan limitaciones [12]. En efecto, las coordenadas geográficas necesitan una transformación a coordenadas planas para que todos los datos de entrada queden expresados en un mismo sistema geográfico. Esto permite operar rápida y fácilmente sobre la distancia de los puntos.

A partir de lo planteado se desarrolla una rutina para pasar de coordenadas geográficas a coordenadas planas con proyección Gauss-Kruger [41]. Se elige trabajar con esta proyección ya que en la Argentina se la adoptó mediante la Disposición Permanente Nro. 197 (24 abril 1925) del Instituto Geográfico Militar [14]. La rutina toma como entrada los archivos descritos en la sección anterior, transforma las coordenadas (longitud-latitud) y obtiene como salida un archivo que posee las coordenadas planas (norte-este en metros) junto al valor del punto muestreado.

3.2.3 Fusión de los datos de entrada

Los modelos digitales que representan las variables medidas en los campos pueden expresarse mediante diferentes estructuras vectoriales basadas en curvas de nivel o en una red irregular de triángulos (TIN). Debido a que la estructura de datos TIN es cada vez más utilizada para representar estos modelos, y dadas las características del conjunto de datos con el que se va a trabajar, se opta por la utilización del algoritmo de interpolación basado en TIN [42].

Esta interpolación tiene como objetivo unificar el mallado de todas las variables que intervienen en el proceso. Para ello, se interpolan sus valores de

manera de obtener una sola grilla espacial común y que cada punto contenga el valor de todas las variables.

Se crea una grilla espacial con una distribución equiespaciada de sus puntos. La densidad de los mismos queda determinada con el ingreso de un parámetro por parte del usuario. Dado el conjunto de datos se buscan los puntos (x,y) que representan los límites superior izquierdo e inferior derecho. En caso de existir más de un archivo se busca el máximo de los puntos mínimos y el mínimo de los máximos para determinar los límites de la grilla rectangular de puntos. Luego se procede a interpolar los nuevos puntos de cada variable. Sólo se calculan aquellos puntos del mallado regular que caen dentro de la envolvente convexa que forma cada variable.

Para interpolar se puede elegir entre los métodos de interpolación que proporciona la librería CGAL: interpolación Lineal, Cuadrática, Farin, Sibson y Sibson con raíz cuadrada. La salida de esta etapa consiste en un archivo que contiene todos los puntos del mallado regular. Cada uno consta de su respectiva ubicación espacial (x,y) y los valores que posee cada variable separados por coma, obtenidos en la interpolación.

3.2.4 Clasificación de las variables

Una vez obtenido el archivo interpolado se comienza con el proceso de clasificación e identificación de las zonas de manejo presentes en el lote productivo analizado. El algoritmo seleccionado para llevar adelante la tarea de clasificación es el Fuzzy C-Means de acuerdo a las pruebas preliminares que se exponen en el Apéndice A.

Se ingresa el archivo que se obtuvo en la etapa anterior y se configuran los siguientes parámetros para el algoritmo FCM: exponente difuso, criterio de convergencia, máximo número de iteraciones, cantidad mínima y máxima de grupos. Además, se calculan tres índices que permiten al usuario realizar una evaluación acerca de la calidad de cada agrupamiento.

Los índices aportan objetivamente una idea más clara acerca de cuál podría ser la clasificación óptima, aunque la selección final de la cantidad de grupos debe seguir una relación de compromiso entre lo sugerido por los índices y lo realmente practicable por la maquinaria de dosificación variable, la que que no tiene posibilidad de alcanzar cambios instantáneos.

Una clasificación de calidad busca que los índices XB, FPI y NCE sean mínimos. Para ello, se calcula la distancia euclídea ($\sqrt{XB^2 + FPI^2 + NCE^2}$) y se selecciona la menor. Todos los resultados obtenidos en esta etapa (resul-

tados de agrupaciones e índices de validación) son guardados en un archivo.

Para continuar con la etapa siguiente es necesario que el usuario indique la cantidad de zonas elegida para diferenciar su lote.

3.2.5 Procesamiento de imágenes

En esta etapa se realiza el procesamiento de los resultados del clustering para obtener zonas bien definidas y con una superficie mayor que un determinado valor. Se trabaja con técnicas de procesamiento digital de imágenes que contienen diferentes métodos y algoritmos que manipulan y transforman una imagen en una señal con información de utilidad.

Filtrado de imágenes

Se utilizan los filtros estadísticos de orden de mediana y moda, que son filtros espaciales no lineales cuya respuesta se basa en el ordenamiento (ranking) de los píxeles contenidos en una porción de la imagen. La elección se debe a la estructura que poseen las imágenes, donde cada píxel contiene el valor del grupo al que pertenece, los filtros lineales generarían valores fuera del rango elegido en el clasificador. Se recorre cada píxel de la imagen tomando una vecindad con máscaras de 3x3, 5x5 ó 7x7 píxeles y, a continuación, se sustituye el valor del píxel central por el valor que resulte del criterio de clasificación.

Erosión y dilatación

Existen casos en los que las zonas encontradas en los agrupamientos cumplen con la condición de área mínima impuesta por el usuario aunque, su morfología dificultaría la correcta aplicación de la dosis variable de la máquina agrícola. Para eliminar éstos casos, se deja a disposición del usuario la aplicación de las técnicas de erosión y dilatación, de manera conjunta, tantas veces como se indique.

Eliminación de superficies pequeñas

La capacidad de cambiar la dosis que aplica la máquina agrícola no es instantánea. Por ello, no se justifica tener grupos con superficies pequeñas que no lleguen a cubrir una determinada superficie.

Para identificar aquellos sectores pequeños se utiliza el etiquetado por crecimiento de regiones, basándose en un criterio de adyacencia o conectividad determinado por el nivel de gris. El criterio es comprobado sobre una vecindad de 8 vecinos y de acuerdo a la cantidad de píxeles que conforma cada región etiquetada, se elimina aquella que no alcanze el área mínima ingresada por el usuario.

3.2.6 Determinación de los polígonos de las zonas de manejo.

Para crear el archivo *shape* se necesita que los puntos que conforman el borde de cada zona estén ordenados. Para encontrar el trazado del contorno de cada zona y así determinar el polígono se desarrolla una rutina basada en el algoritmo de Moore [33].

3.2.7 Creación del archivo Shape

Culminando el proceso se convierten las coordenadas espaciales de los polígonos al sistema de referencia original (WGS84), para finalmente obtener el archivo *shape* (*shp*) que será utilizado en la maquinaria agrícola.

Se utiliza la librería Shapefile que ofrece la posibilidad de leer, escribir y actualizar el Shapefile de ESRI y el archivo de atributo asociado (*dbf*).

El proceso consiste en crear un objeto del tipo *Shape Polygon* al cual pasarle como parámetro las coordenadas (x,y) de los puntos que forman los polígonos que delimitan las zonas. Luego se crea el archivo *dbf* mediante una función de la librería a la cual se le pasa por parámetro los atributos que tiene cada polígono. Estos datos representan la dosificación variable que se aplicará a cada zona del lote.

3.2.8 Desarrollo de una interfaz gráfica

Se desarrolló una interfaz gráfica básica que muestra el funcionamiento del método de clasificación. Ésta permite al usuario ingresar los datos y visualizar tanto los resultados parciales como los finales. Algunas de las pantallas que serán visualizadas por el usuario son:

- Pantalla 1: es la primer pantalla con la que interactúa el usuario (Figura

3.2(a)). La misma permite ingresar los archivos de entrada, que serán utilizados para la posterior clasificación en zonas de manejo.

- Pantalla 2: esta pantalla (Figura 3.2(b)) indica al usuario cuáles son los paramétros que deberá completar para llevar a cabo el método de interpolación. Además opcionalmente podrá ingresar un archivo interpolado con anterioridad.

- Pantalla 3: a través del ingreso de los paramétros de esta pantalla (Figura 3.3(a)), se realiza el proceso de clasificación en la cantidad de zonas que determine el usuario.

- Pantalla 4: en esta pantalla se visualizan de forma gráfica los resultados de las distintas clasificaciones (Figura 3.3(b)). El usuario podrá elegir en cuantas zonas desea finalmente dividir su lote.

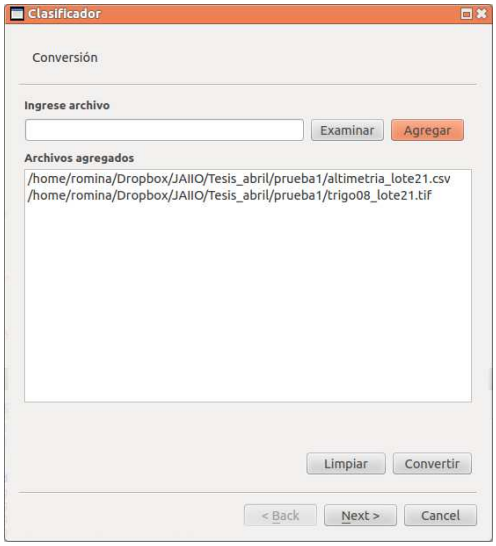
- Pantalla 5: aquí se presentan los resultados de los índices de validación de la clasificación (Figura 3.4(a)).

- Pantalla 6: una vez seleccionada la cantidad de zonas óptima se da comienzo a la siguiente etapa de filtrado (Figura 3.4(b)). Se debe ingresar el tipo de filtro, la máscara a utilizar y el tamaño de la superficie mínima.

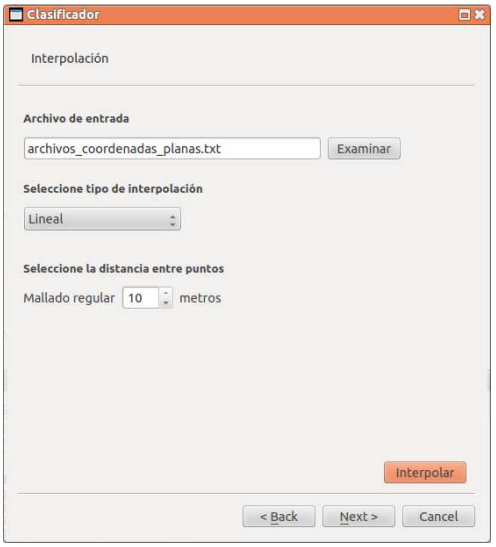
- Pantalla 7: se puede calcular el promedio de las variales en cada zona (Figura 3.5(a)), para tener una ayuda al momento de aplicar la cantidad de insumos en cada zona.

- Pantalla 8: se debe ingresar el nombre, tipo y cantidad de caracteres permitidos para el atributo que será aplicado en la zona (Figura 3.5(b)).

- Pantalla 9: la última pantalla brinda la posibilidad de ingresar los valores que se aplicarán a cada zona, según los atributos mencionados anteriormente. Finalmente al presionar el botón *guardar* se crea el archivo *shape*, el cual será utilizado por la maquinaria agrícola de dosificación variable.

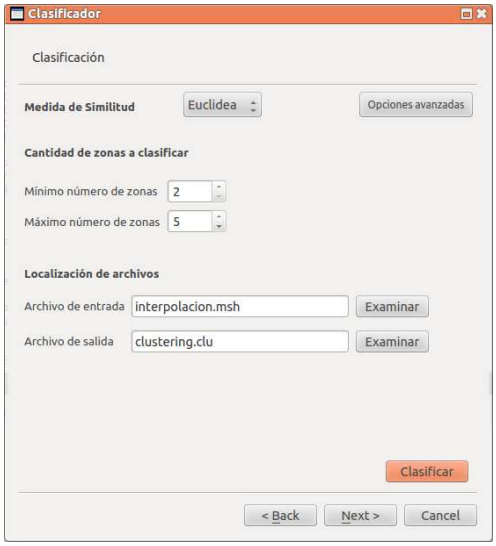


(a) Pantalla 1: ingreso de datos.

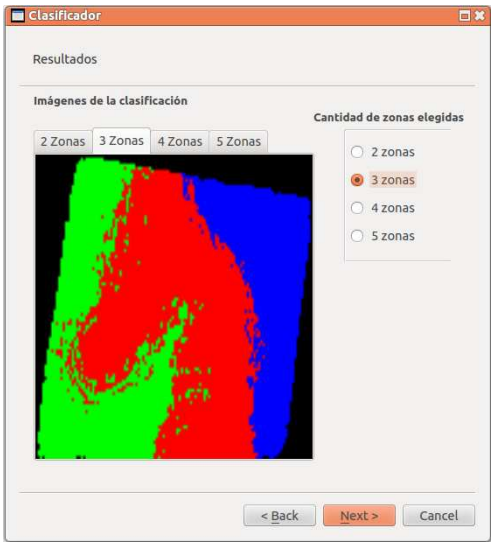


(b) Pantalla 2: interpolación.

Figura 3.2: Ingreso de datos e interpolación.

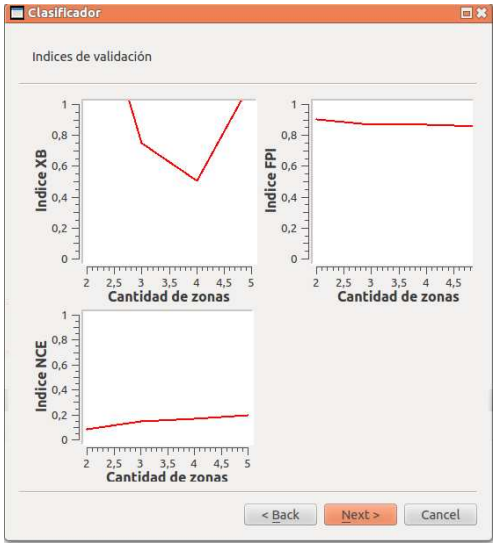


(a) Pantalla 3: clasificación.

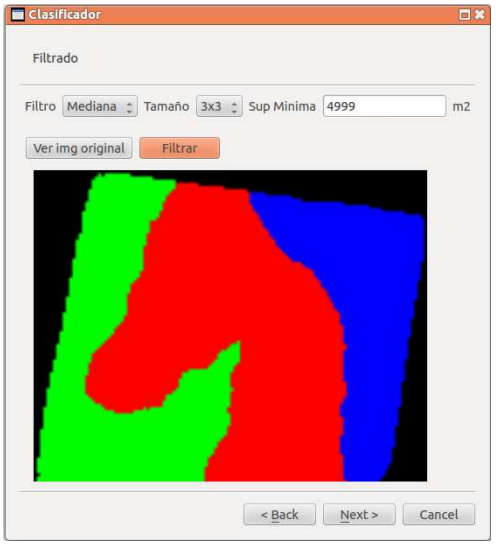


(b) Pantalla 4: resultados de clasificación.

Figura 3.3: Clasificación y resultados de clasificación.



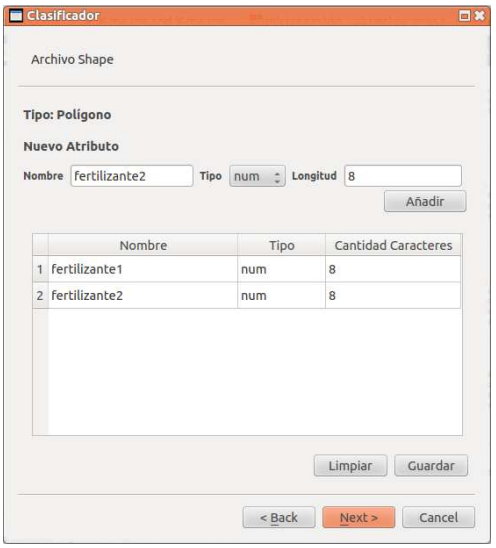
(a) Pantalla 5: índices de validez.



(b) Pantalla 6: filtrado.

Figura 3.4: Índices de validez y filtrado.

(a) Pantalla 7: promedio de las variables.



(b) Pantalla 8: configuración archivo *shape*.

Figura 3.5: Promedio de las variables y configuración *shape*.

CAPÍTULO 4

Experimentos y resultados

En este capítulo se presentan las pruebas realizadas con diversos conjuntos de datos, sobre los cuales se practicaron diferentes estudios y se analizaron los resultados obtenidos. Todos los conjuntos de pruebas fueron suministrados por la EEA Paraná del INTA.

4.1 Conjuntos de datos utilizados

Para realizar las pruebas se utilizaron datos provenientes de 3 lotes diferentes.

Conjunto 1

Este conjunto de datos pertence a un lote que posee una superficie de 110 hectáreas y en la campaña analizada estaba sembrado con trigo. Se cuenta con una imagen satelital de 74 x 69 píxeles que contiene el índice de vegetación diferencial normalizado (NDVI), un mapa de rendimiento con 124286 muestras y la altimetría con 4687 valores, estos últimos, distribuidos irregularmente sobre el lote.

Conjunto 2

Este conjunto de datos está conformado por una sola variable, la misma proviene de un lote perteneciente al INTA y contiene el rendimiento del maíz de la campaña 2009. De esta variable se tomaron 14495 muestras.

Conjunto 3

Este conjunto contiene la conductividad eléctrica con 7048 muestras y el rendimiento con 142948 muestras de un lote con 121 hectáreas aproximadamente.

4.2 Experimentos realizados

Se realizan 4 tipos de experimentos, por un lado se comprueba la eficacia de las interpolaciones, por otro se analiza el comportamiento del método de agrupamiento. Luego se analizan los resultados arrojados en el procesamiento de las imágenes que posibilitan obtener regiones apropiadas para crear las zonas de manejo del archivo *shape*. Y por último, se comprueba la correcta detección del bordes de cada zona de manejo.

4.2.1 Eficacia de la interpolación

Para realizar el anális de eficacia se toma un mismo archivo y se lo divide en dos, se toma 1 muestra cada 2 puntos de manera alternada. De esta manera se compara el valor original y el valor interpolado. Para evaluar el error de interpolación, se emplea el error cuadrático medio y el desvío estándar de cada método proporcionado por la librería CGAL. A los fines prácticos y únicamente para la realización de estas pruebas, el mallado se realiza con los puntos originales esparcidos irregularmente.

Prueba 1

Esta prueba se realiza a partir del archivo que contiene la *altimetría* en el conjunto de datos 1. Se lo divide en dos archivos de 2343 y 2344 puntos respectivamente y el mallado se conforma con todos los puntos de los dos

archivos creados recientemente. En total se interpolan 4655 puntos y los resultados obtenidos se observan en la Tabla 4.1.

Tabla 4.1: Comparación de diferentes tipos de interpolación - altimetría.

Interpolación	ECM	Desvío Estándar Promedio	Velocidad (s.)
Lineal	0.10	5.96	0.41
Sibson s/SQRT	0.32	5.82	0.44
Farin	0.25	5.80	0.45
Cuadrática	0.22	5.81	0.43
Sibson (Tradicional)	0.27	5.81	0.42

Prueba 2

Para realizar la segunda prueba acerca de la eficacia de los métodos de interpolación se utiliza el archivo que contiene 7048 valores de *conductividad eléctrica* del lote del conjunto 3. Se lo divide en dos archivos de 3524 puntos cada uno y el mallado se conforma con todos los puntos de los dos archivos creados recientemente. Los resultados obtenidos se observan en la Tabla 4.2.

Tabla 4.2: Comparación de diferentes tipos de interpolación - cond. eléctrica.

Interpolación	ECM	Desvío Estándar Promedio	Velocidad (s.)
Lineal	1.31	8.36	0.60
Sibson s/SQRT	1.26	8.36	0.62
Farin	1.25	8.35	0.66
Cuadrática	1.23	8.35	0.62
Sibson (Tradicional)	1.24	8.35	0.63

Prueba 3

En esta prueba se utiliza el archivo que contiene 124286 valores de *rendimiento* del lote del conjunto 1. Se lo divide en dos archivos de 62143 puntos cada uno y el mallado se conforma con todos los puntos de los dos archivos creados recientemente. Los resultados obtenidos se observan en la Tabla 4.3.

Prueba 4

En la última prueba sobre la eficacia de la interpolación se utiliza el mismo archivo de la Prueba 3. Se toma 1 muestra cada 4 puntos de manera alternada para lograr una mayor separación entre las muestras. De esta manera

Tabla 4.3: Comparación de diferentes tipos de interpolación - rendimiento.

Interpolación	ECM	Desvío Estándar Promedio	Velocidad (s.)
Lineal	0.55	3	20.30
Sibson s/SQRT	0.69	3.05	21.12
Farin	0.68	3.04	21.71
Cuadrática	0.66	3.03	20.88
Sibson (Tradicional)	0.68	3.04	21.02

obtenemos dos archivos con 15535 y 15536 muestras respectivamente con una mayor separación entre los mismos. Los resultados obtenidos se observan en la Tabla 4.4.

Tabla 4.4: Comparación de diferentes tipos de interpolación - rendimiento.

Interpolación	ECM	Desvío Estándar Promedio	Velocidad (s.)
Lineal	1.75	2.8	3.64
Sibson s/SQRT	1.86	2.71	3.74
Farin	1.80	2.69	3.95
Cuadrática	1.79	2.68	3.78
Sibson (Tradicional)	1.83	2.70	3.78

Análisis de la eficacia de la interpolación

El ECM es una medida estadística utilizada para evaluar qué tan bueno es el método de interpolación elegido. Surge a partir de las diferencias entre los valores reales y los estimados de la variable analizada. La prueba 1 tiene una marcada diferencia a favor de la interpolación lineal, esto se debe a la característica física de la variable interpolada. La altimetría de un campo tiene un comportamiento casi lineal por lo cual es esperable que esta interpolación logre los mejores resultados. Es común encontrar que determinados fenómenos medidos en un lote y utilizados en nuestro algoritmo tengan un comportamiento más aleatorio o menos predecible. Si bien la interpolación lineal logra el menor ECM en tres de las cuatro pruebas, sólo en la primera logra diferenciarse con claridad. Para el resto de las pruebas (Tablas 4.2, 4.3, 4.4), los valores de ECM en general no difieren mucho entre los distintos métodos de interpolación.

El desvío estándar es una medida que respresenta la desviación de distribución de los datos respecto a su media aritmética. Su comportamiento es

similar a lo largo de todas las interpolaciones de las diferentes pruebas y por lo tanto no nos aporta información útil al análisis.

En las pruebas 3 y 4 se interpoló la misma variable pero con diferentes densidades espaciales. Se puede notar un desmejoramiento en el resultado de los ECMs de la prueba 4 ya que hay más separación entre las muestras. El ECM resultó ser menor en la prueba 3 que tiene más puntos y por ende más información.

Si se analizan los tiempos que conlleva la ejecución de los métodos de interpolación se observa que existe una relación lineal entre la cantidad de puntos y la velocidad de interpolación que poseen todos los métodos analizados en las pruebas 1, 2, 3 y 4 (Figura 4.1). También se observa que la interpolación lineal es la más rápida, esto se debe, en parte, al cálculo del gradiente que se realiza en las demás interpolaciones.

Es por todo esto que se puede afirmar que todos los métodos analizados arrojan resultados aceptables y ninguno sobresale del resto. Los valores del ECM no presentan grandes diferencias entre los distintos tipos de interpolaciones. Podemos destacar la velocidad mínima que logra la interpolación lineal a lo largo de las 4 pruebas. Sin embargo, este parámetro no es de mayor importancia ya que nuestro algoritmo no requiere ejecutarse en tiempo real. Por lo tanto, la precisión de la interpolación dependerá del método, del origen de las variables y de la densidad espacial con la que se midan los puntos. La función que se utilice para realizar la interpolación de los valores queda a criterio del usuario.

Figura 4.1: Relación entre el tiempo de interpolación y la cantidad de puntos.

4.2.2 Pruebas de la clasificación

Estas pruebas consisten en analizar el comportamiento del método de agrupamiento utilizado. Para proceder con las mismas, se toman los conjuntos de datos fusionados con un determinado método de interpolación, se ingresan distintos parámetros de entrada y se analizan los resultados de acuerdo a los valores de los índices y de las gráficas.

Como ya se ha mencionado, para que la clasificación sea de calidad se busca que los índices XB, FPI y NCE sean mínimos. La selección de la mejor representación se realiza mediante la elección de la menor distancia euclídea ($\sqrt{XB^2 + FPI^2 + NCE^2}$). Para eliminar la preeminencia en los valores de unos índices sobre otros, se normaliza cada uno utilizando el valor máximo obtenido para ese índice en la clasificación actual, así quedan expresados con valores entre 0 y 1. Los resultados originales de estas pruebas se encuentran en el Apéndice B.

Para todas las pruebas de clasificación se consideran como parámetros de entrada los siguientes valores: criterio de convergencia = 0.0001, máximo número de iteraciones = 300, cantidad mínima de grupos = 2 y cantidad máxima de grupos = 5. El valor del exponente difuso se fija entre 1.2 y 1.5 ya que es apropiado cuando los datos utilizados contienen valores del suelo [31].

Prueba 5

Para realizar esta prueba se utiliza el conjunto de datos 1. Se crea el mallado regular con una separación de 10 metros entre punto y punto mediante el método de interpolación cuadrática, quedando un total de 10103 puntos.

En las Tablas 4.5 y 4.6 se observan los resultados de las clasificaciones realizadas utilizando los exponentes difusos 1.3 y 1.5 respectivamente. En negritas se señalan los menores valores de cada índice y de la distancia euclídea. La Figura 4.2 muestra la representación gráfica de la clasificación con el exponente difuso 1.3 para diferentes agrupamientos. Los resultados del exponente difuso 1.5 no se visualizan ya que son muy similares.

Prueba 6

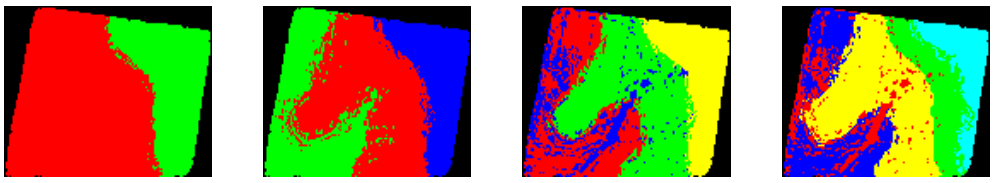
En esta prueba se utiliza el conjunto de datos 2 con la interpolación cuadrática y con un mallado de 5 metros. Esta variación respecto de la prue-

Tabla 4.5: Resultados prueba 5 con exponente 1.3.

Índices	2 clusters	3 clusters	4 clusters	5 clusters
XB	0.67	0.24	**0.16**	1.00
FPI	1.00	0.97	0.93	**0.91**
NCE	**0.44**	0.65	0.87	1.00
Dist. Euclídea	1.28	**1.19**	1.28	1.68
Tiempo (s)	1.93	3.85	9.38	14.01
Iteraciones	26	32	55	61

Tabla 4.6: Resultados prueba 5 con exponente 1.5.

Índices	2 clusters	3 clusters	4 clusters	5 clusters
XB	1.00	0.38	**0.26**	0.58
FPI	1.00	0.96	0.96	**0.94**
NCE	**0.42**	0.72	0.86	1.00
Dist. Euclídea	1.48	**1.26**	1.31	1.49
Tiempo (s)	1.88	4.9	9.19	28.2
Iteraciones	24	38	49	111



(a) Clasificación en 2 clusters. (b) Clasificación en 3 clusters. (c) Clasificación en 4 clusters. (d) Clasificación en 5 clusters.

Figura 4.2: Clasificación del lote 1 con exponente difuso igual a 1.3.

ba anterior se debe a que se cuenta con un menor número de puntos (4215 puntos).

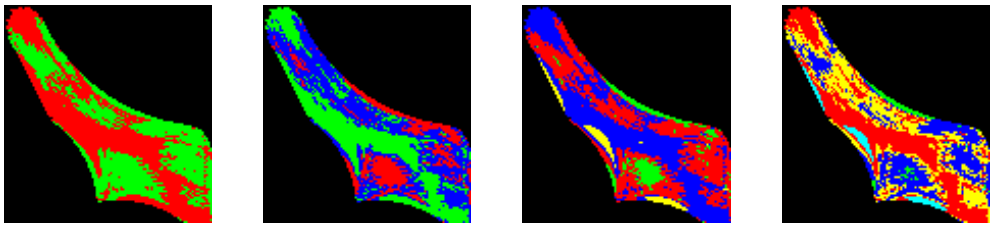
En las Tablas 4.7 y 4.8 se observan los resultados de las clasificaciones realizadas mediante los exponentes difusos 1.3 y 1.5 respectivamente. En negritas se señalan los menores valores de cada índice y de la distancia euclídea. La Figura 4.3 muestra la representación gráfica de la clasificación con el exponente difuso 1.5 para diferentes agrupamientos. Los resultados del exponente difuso 1.3 no se visualizan ya que son muy similares.

Tabla 4.7: Resultados prueba 6 con exponente 1.3.

Índices	2 clusters	3 clusters	4 clusters	5 clusters
XB	0.62	1.00	0.54	**0.28**
FPI	0.98	**0.95**	1.00	1.00
NCE	**0.60**	1.00	0.69	0.73
Dist. Euclídea	1.30	1.70	1.33	**1.27**
Tiempo (s)	0.36	8.14	2.78	6.03
Iteraciones	16	221	48	74

Tabla 4.8: Resultados prueba 6 con exponente 1.5.

Índices	2 clusters	3 clusters	4 clusters	5 clusters
XB	0.68	1.00	0.56	**0.31**
FPI	0.96	**0.95**	1.00	1.00
NCE	**0.67**	1.00	0.84	0.89
Dist. Euclídea	**1.35**	1.70	1.42	1.38
Tiempo (s)	0.34	3.4	3.98	5.27
Iteraciones	16	101	79	75



(a) Clasificación en 2 clusters. (b) Clasificación en 3 clusters. (c) Clasificación en 4 clusters. (d) Clasificación en 5 clusters.

Figura 4.3: Clasificación del lote 2 con exponente difuso igual a 1.5.

Prueba 7

Se utiliza el conjunto de datos 3 con la interpolación cuadrática y con un mallado de 10 metros, quedando un total de 12109 puntos.

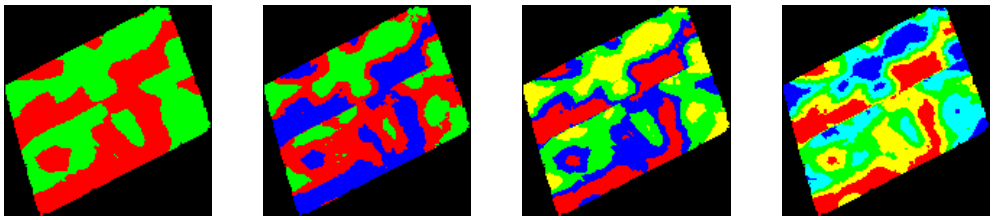
En las Tablas 4.9 y 4.10 se observan los resultados de las clasificaciones realizadas mediante los exponentes difusos 1.3 y 1.5 respectivamente. En negritas se señalan los menores valores de cada índice y de la distancia euclídea. La Figura 4.4 muestra la representación gráfica de la clasificación con el exponente difuso 1.5 para diferentes agrupamientos. Los resultados del exponente difuso 1.3 no se visualizan ya que son muy similares.

Tabla 4.9: Resultados prueba 7 con exponente 1.3.

Índices	2 clusters	3 clusters	4 clusters	5 clusters
XB	0.82	**0.38**	0.58	1.00
FPI	1.00	**0.98**	0.99	0.99
NCE	**0.57**	0.93	0.93	1.00
Dist. Euclídea	1.41	**1.40**	1.48	1.73
Tiempo (s)	0.86	6.56	7.98	18.54
Iteraciones	13	61	48	80

Tabla 4.10: Resultados prueba 7 con exponente 1.5.

Índices	2 clusters	3 clusters	4 clusters	5 clusters
XB	0.83	**0.38**	0.59	1.00
FPI	1.00	**0.97**	0.99	0.99
NCE	**0.57**	0.89	0.92	1.00
Dist. Euclídea	1.42	**1.37**	1.47	1.73
Tiempo (s)	0.87	5.58	7.02	18.8
Iteraciones	14	57	48	94



(a) Clasificación en 2 clusters. (b) Clasificación en 3 clusters. (c) Clasificación en 4 clusters. (d) Clasificación en 5 clusters.

Figura 4.4: Clasificación del lote 3 con exponente difuso igual a 1.5.

Análisis de la clasificación

Los valores mínimos de cada índice para cada prueba nos dan una idea de cual puede ser la cantidad óptima de zonas de manejo. Con dicha clasificación se tienen la menor cantidad de miembros compartidos entre clases y una mayor organización de las mismas.

Se debe destacar que los resultados obtenidos por los dos exponentes difusos en cada prueba, arrojan los mismos valores mínimos en cada índice. Exceptuando la prueba 6 que para el exponente difuso = 1.3 sugiere la clasificación en 5 zonas de manejo y con el exponente difuso = 1.5 en 2 zonas. En estos casos, se complementa la información de los índices con las represen-

taciones gráficas de las clasificaciones. De esta forma se visualizan las distribuciones espaciales de las zonas de manejo y se determina la más adecuada para la aplicación que realiza la maquinaria.

El tiempo de ejecución de la clasificación depende, además del número de iteraciones, de la cantidad de puntos y de la cantidad de variables que contenga el conjunto de datos.

Las diversas herramientas presentadas permiten al usuario decidir acerca de la cantidad óptima de zonas de manejo.

4.2.3 Pruebas de filtrado

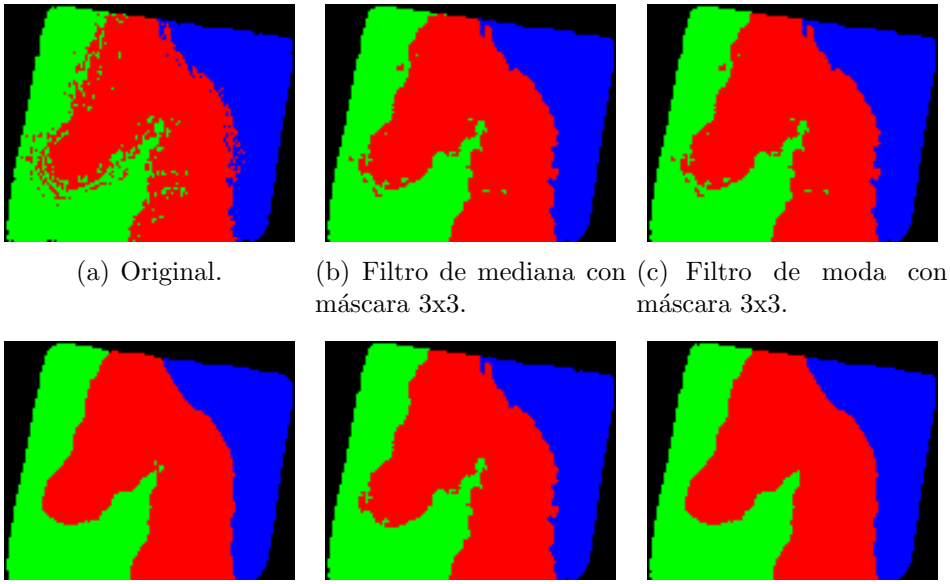
Debido a que la clasificación que produce el FCM se realiza sobre los valores de las variables sin tener en cuenta su ubicación espacial, se pueden obtener zonas definidas deficientemente, con sectores mal delimitados y no compactos. En estas pruebas se analizan las respuestas de los filtros de mediana y moda con máscaras de 3x3 y 7x7 para suplir estas imperfecciones. Además, como existen zonas con una superficie pequeña que el filtrado de imágenes no es capaz de eliminar, se considera apropiado utilizar un parámetro de superficie mínima que dependerá del tamaño del lote.

Prueba 8

Para la realización de esta prueba se eligió trabajar con la clasificación en 3 zonas realizada en la prueba 5, utilizando el exponente difuso 1.3 (Figura 4.5(a)). Se analiza el desempeño de los filtros de moda (Figura 4.5(b)) y mediana (Figura 4.5(c)) con máscaras de 3x3 y filtro de moda con máscara de 7x7 (Figura 4.5(d)). Por último, se trabaja con una superficie mínima de 5000 m^2 para la eliminación de zonas de pequeñas, en relación al área que posee el lote. El resultado final se presenta en las Figuras 4.5(e) y 4.5(f).

Prueba 9

En la Figura 4.6(a) se observa la imagen perteneciente a la clasificación en 2 zonas de manejo del conjunto de datos 2 utilizando el exponente 1.5. Se busca suplir las imperfecciones que resultaron del proceso de clustering. Se analiza el desempeño de los filtros de moda (Figura 4.6(b)) y mediana (Figura 4.6(c)) con máscaras de 3x3 y filtro de moda con máscara de 7x7 (Figura 4.6(d)). Por último, se trabaja con una superficie mínima de 5000



(a) Original. (b) Filtro de mediana con máscara 3x3. (c) Filtro de moda con máscara 3x3.

(d) Filtro de moda con máscara 7x7. (e) Eliminación sup. mínima 5000 m^2 - moda 3x3. (f) Eliminación sup. mínima 5000 m^2 - moda 7x7.

Figura 4.5: Filtrado y eliminación de superficies pequeñas - prueba 8.

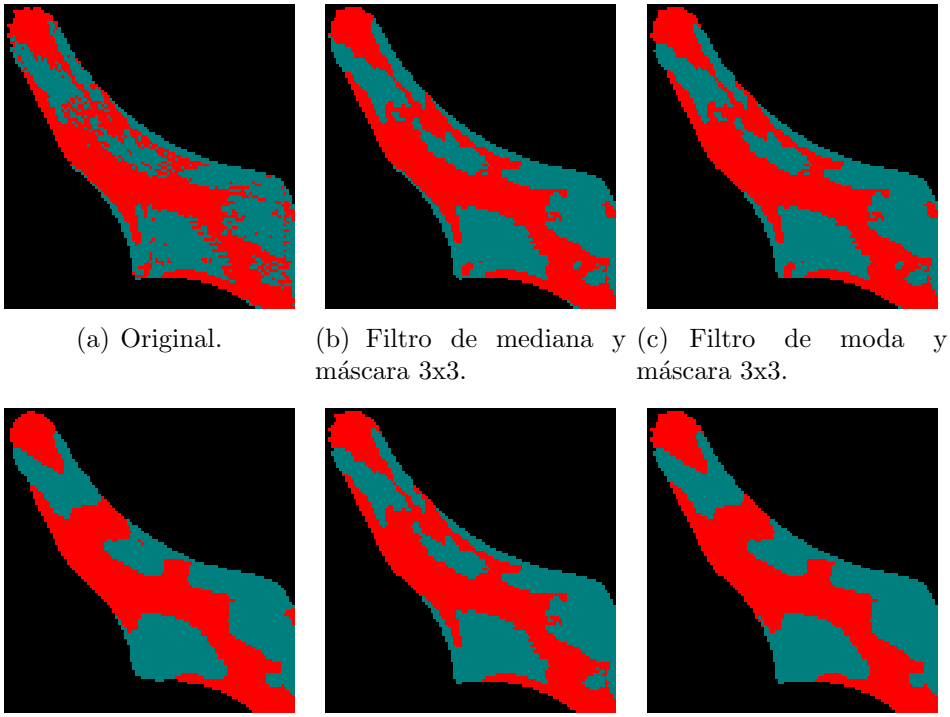
m^2 para la eliminación de zonas de pequeñas, en relación al área que posee el lote. El resultado final se presenta en las Figuras 4.6(e) y 4.6(f).

Prueba 10

La prueba se realiza a partir del resultado obtenido en la clasificación en 3 zonas de manejo de la prueba 7 utilizando el exponente 1.5 (Figura 4.7(a)). Se analizan los filtros de moda (Figura 4.7(b)) y mediana (Figura 4.7(c)) con máscaras de 3x3 y filtro de moda con máscara de 7x7 (Figura 4.7(d)). Adicionalmente se eliminan las zonas pequeñas que no superen una hectárea (10000 m^2). Los resultados finales de esta prueba se presentan en las Figuras 4.7(e) y 4.7(f).

Análisis del filtrado

El tamaño de las máscaras influye en la obtención de zonas más redondeadas o angulosas. Las máscaras de 7x7 son las que brindan zonas mejor definidas. El filtro de moda y mediana no tienen mayores diferencias, es más, para el caso de separaciones en dos grupos, ambos filtros arrojan los mismos resultados. A partir del etiquetado por crecimiento de regiones se garantiza



(a) Original. (b) Filtro de mediana y máscara 3x3. (c) Filtro de moda y máscara 3x3.

(d) Filtro de moda y máscara 7x7. (e) Eliminación superficie mínima 5000 m^2 - moda 3x3. (f) Eliminación superficie mínima 5000 m^2 - moda 7x7.

Figura 4.6: Filtrado y eliminación de superficies pequeñas - prueba 9.

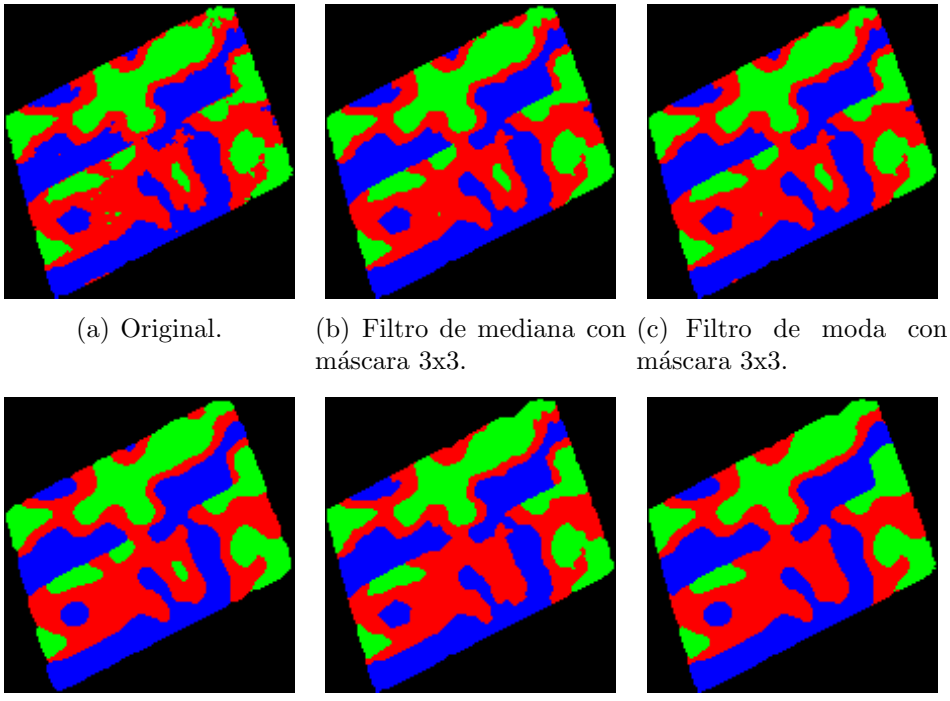
que las zonas resultantes superan una determinada superficie mínima. Los resultados muestran que se pueden suplir las imperfecciones producidas por el FCM y así obtener zonas bien definidas.

4.2.4 Pruebas de erosión-dilatación

El motivo de esta prueba es analizar cómo influye la aplicación de la erosión y dilatación junto a las técnicas de filtrado y eliminación de zonas pequeñas.

Prueba 11

Se toman los mejores resultados que cada lote obtuvo en las pruebas de filtrado (prueba 8, 9 y 10) y se realizan dos erosiones y dos dilataciones.



(a) Original. (b) Filtro de mediana con máscara 3x3. (c) Filtro de moda con máscara 3x3.

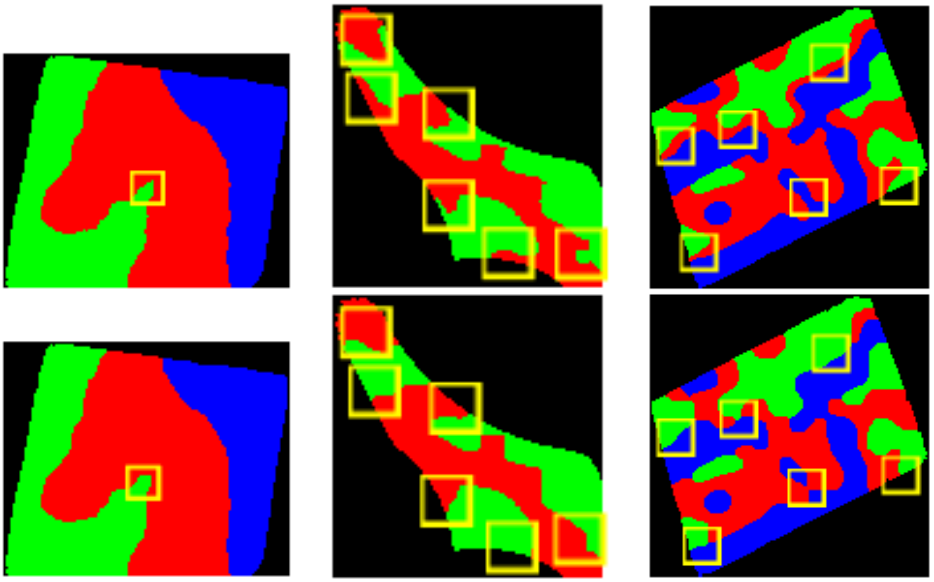
(d) Filtro de moda con máscara 7x7. (e) Eliminación sup. mínima 10000 m^2 - moda 3x3 (f) Eliminación sup. mínima 10000 m^2 - moda 7x7

Figura 4.7: Filtrado y eliminación de superficies pequeñas - prueba 10.

Los resultados obtenidos se observan en la Figura 4.8. En la parte superior se presentan las imágenes utilizadas y debajo se muestran los resultados de aplicar estas operaciones.

Análisis de la erosión-dilatación

Como se puede observar en las comparaciones realizadas en la Figura 4.8 los mayores cambios se observan en las Figuras 4.8(b) y 4.8(c) (los cambios son resaltados en amarillo). Los casos en los que se aplica erosión y dilatación logran reducir las zonas que poseen una sección angosta. Evidentemente la aplicación de la erosión y la dilatación mejora los resultados en campos que poseen zonas angostas y no genera cambios sustanciales en aquellos lotes con zonas amplias. Por tal motivo, y a pesar de que el uso de estas operaciones es optativo, se sugiere su aplicación para obtener mejores resultados.



(a) Lote 1. (b) Lote 2. (c) Lote 3.

Figura 4.8: Comparación aplicación erosión-dilatación - prueba 11.

4.2.5 Detección de bordes

En estas pruebas se identifica el borde de cada zona. Partiendo de la imagen filtrada de las pruebas anteriores, se procede a ejecutar la rutina que implementa el algoritmo de Moore. En las Figuras 4.9, 4.10 y 4.11 se indica el sentido de los polígonos encontrados con un aumento en la intensidad de gris a medida que se recorre su trayectoria.

Prueba 12

Se trabaja con el resultado obtenido en la prueba 8 con el filtro de moda con máscara 7x7. La zona de la Figura 4.9(a) está conformada por 327 puntos. La zona de la Figura 4.9(b) contiene 339 puntos en el borde y la zona de la Figura 4.9(c) 241 puntos.

Prueba 13

Aquí se utiliza el resultado obtenido en la prueba 9 con el filtro de moda con máscara 7x7.

(a) Borde zona 1. (b) Borde zona 2. (c) Borde zona 3.

Figura 4.9: Detección de bordes - prueba 12.

Las zonas del grupo 1 (Figura 4.10(a)) tienen las siguientes características:

- La zona *a* contiene 65 puntos.
- La zona *b* contiene 339 puntos.

Las zonas del grupo 2 (Figura 4.10(b)) tienen las siguientes características:

- La zona *a* contiene 100 puntos.
- La zona *b* contiene 218 puntos.
- La zona *c* contiene 107 puntos.

(a) Borde zona 1. (b) Borde zona 2.

Figura 4.10: Detección de bordes - prueba 13.

Prueba 14

En esta prueba se trabaja con el resultado obtenido en la prueba 10 con el filtro de moda con máscara 7x7

Las zonas del grupo 1 (Figura 4.11(a)) tienen las siguientes características:

- La zona a contiene 671 puntos.
- La zona b contiene 48 puntos.
- La zona c contiene 202 puntos.

Las zonas del grupo 2 (Figura 4.11(b)) tienen las siguientes características:

- La zona a contiene 224 puntos.
- La zona b contiene 97 puntos.
- La zona c contiene 67 puntos.
- La zona d contiene 107 puntos.
- La zona e contiene 59 puntos.
- La zona f contiene 41 puntos.

Las zonas del grupo 3 (Figura 4.11(c)) tienen las siguientes características:

- La zona a contiene 462 puntos.
- La zona b contiene 47 puntos.
- La zona c contiene 125 puntos.
- La zona d contiene 36 puntos.

(a) Borde zona 1. (b) Borde zona 2. (c) Borde zona 3.

Figura 4.11: Detección de bordes - prueba 14.

Análisis de la detección de bordes

Se puede observar que las zonas de cada lote fueron perfectamente identificadas por el algoritmo de Moore, permitiendo conseguir los polígonos que delimitan cada una de ellas. El algoritmo respondió adecuadamente tanto para zonas con bordes sencillos (prueba 12), como para bordes con demasiados curvas (prueba 14). A partir de estos resultados es posible crear un archivo *shape* a través de la rutina de creación del mismo. Se pasan por parámetros las coordenadas de los puntos que conforman cada polígono de las distintas zonas y los atributos que tiene cada zona.

CAPÍTULO 5

Conclusiones

En este último capítulo, se describen las conclusiones obtenidas tras el desarrollo del método y las pruebas efectuadas con diferentes conjuntos de datos. Además, se presentan diversos trabajos futuros que permitirán mejorar la operabilidad del mismo.

5.1 Conclusiones finales

En este trabajo se ha diseñado y desarrollado un método para delimitar zonas de manejo dentro de un lote productivo agrícola a través del procesamiento de datos georreferenciados.

El método ha resuelto satisfactoriamente el problema de identificar zonas de manejo en un lote productivo, integrando diferentes etapas, automatizándolas y logrando un procedimiento más rápido y sencillo.

En la etapa de transformación y fusión se logró obtener una malla regular utilizando distintos tipos de datos georreferenciados, evitando el tedioso trabajo manual por parte del usuario.

Se ha implementado un procedimiento que mediante *Fuzzy C-Means* y el uso de los índices de validación de la calidad de los grupos permiten la evaluación simultánea de una serie de propuestas de agrupamientos. Así, el

usuario puede comparar y elegir la cantidad de zonas de manejo que mejor se adapte a sus necesidades, utilizando información subjetiva (visualización de gráficas sobre el terreno) y objetiva (índices de calidad).

Mediante el uso de técnicas de procesamiento de imágenes se logró eliminar información espuria y superficies inadecuadas, con el fin de garantizar la generación de zonas de forma compacta, que representan una gran ventaja para la operatividad de las máquinas agrícolas.

La parte final del método posibilita la generación del archivo con formato *shape* que puede ser utilizado en cualquier máquina que realice dosificación variable. Además, el usuario cuenta con la posibilidad de personalizar los atributos necesarios para realizar un tratamiento diferenciado de las zonas de manejo.

5.2 Trabajos futuros

A fin de mejorar el desempeño general del método, en esta sección se listan algunas modificaciones que podrían ser incorporadas en las diferentes etapas:

- Realizar un módulo de adaptación de formatos de los datos de entrada para todas aquellas variables que poseen formatos particulares (no estándares).
- Desarrollar un módulo para la edición gráfica de las variables de entrada que permita seleccionar sólo determinadas áreas.
- Incorporar nuevos tipos de proyecciones de coordenadas que amplíen los territorios donde se pueda aplicar el método.
- Desarrollar un software comercial a partir del prototipo diseñado, incorporando requerimientos aportados por los usuario de la AP.

Bibliografía

[1] F. J. Pierce and P. Nowak. Aspects of precision agriculture. *Advances in Agronomy*, 67:1–85, 1999.

[2] T. A. Doerge. Management zone concepts. site-specific management guideline. *Information Agriculture Conference*, 1999.

[3] R. Bongiovanni, E. Mantovani, S. Best, and A. Roel. *Agricultura de precisión: integrando conocimientos para una agricultura moderna y sustentable*. PROCISUR/IICA, 2006.

[4] Colaboradores de Wikipedia. Agricultura de precisión. `http://es.wikipedia.org/wiki/Agricultura_de_precision`, 2009. [Accedido 20-12-2012].

[5] E. Chartuni y F. A. de Carvalho y D. Marçal y E. Ruz. Agricultura de precisión: Nuevas herramientas para mejorar la gestión tecnología en la empresa agropecuaria. *Comunica*, 1:24–65, enero-abril 2007.

[6] M. Flowers, R. Weisz, and J. G. White. Yield-based management zones and grid sampling strategies: Describing soil test and nutrient variability. *Agronomy Journal*, 97:968–982, 2005.

[7] J. C. Taylor, G. A. Wood, R. Earl, and R. J. Godwin. Soil factors and their influence on within-yield crop variability. part ii: Spatial analysis and determination of management zones biosystems engineering. *Agronomy Journal*, 84:441–453, 2003.

[8] M. E. Bengolea. Demostración de utilización de Farm Works Farm Site. http://www.agriculturadeprecision.org/articulos/software/Farm-Works-Prec.asp, 1999. [Accedido 20-12-2012].

[9] T. Gotthold. Demostración de utilización de SSToolbox y sus posibles aplicaciones. AgriMax S.A. http://www.agriculturadeprecision.org/articulos/software/SSToolbox.asp, 2005. [Accedido 02-08-2012].

[10] J. J. Fridgen, K. A. Sudduth N. R. Kitchen, S. T. Drummond, W. J. Wiebold, and C. W. Fraisse. Software Management Zone Analyst (MZA): Software for Subfield Management Zone Delineation. *Agronomy Journal*, 96:101–107, 2004.

[11] Instituto Nacional de Tecnología Agropecuaria. 7mo. curso de agricultura de precisión y 2do. expo de máquinas precisas. In *Agricultura de Precisi ón*, Manfredi, Córdoba, 2007.

[12] E. Huerta, A. Mangiaterra, and G. Noguera. *GPS Posicionamiento Satelital*. UNR Editora – Universidad Nacional de Rosario, 2005.

[13] I. Fernandez Coppel. *Localizaciones Geográficas – Las Coordenadas Geográficas y la Proyección UTM*. Universidad de Valladolid, 2001.

[14] A. Francois. *Sistemas de coordenadas y transformaciones. Bases para el trabajo en SIG*. Universidad de la República – Uruguay, 2000.

[15] M. Ruth. GeoTIFF FAQ. http://www.remotesensing.org/geotiff/faq.html, 2011. [Accedido 05-10-2012].

[16] N. Ritter. GeoTIFF Format Specification. http://www.remotesensing.org/geotiff/spec/geotiff1.html#1.1. [Accedido 05-10-2012].

[17] I. Bloch. Information combination operators for data fusion: A comparative review with classification. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 26:52–67, 1996.

[18] J. Fallas. *Modelo de elevación digital para las hojas cartográficas tilarán y juntas*. Instituto Geográfico Nacional – Costa Rica, 2003.

[19] F. J. Aguilar Torres, M. A. Aguilar Torres, F. Aguera Vega, F. Carvajal Ramírez, and P. L. Sánchez Salmerón. Efectos de la morfología del terreno, densidad muestral y métodos de interpolación en la calidad de

los modelos digitales de elevaciones. *Anales de la ingeniería gráfica*, 17:25–35, 2005.

[20] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, Inc., NJ – USA, May 2008.

[21] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31:265–323, September 1999.

[22] P. W. Mausel, W. J. Kamber, and J. K. Lee. Optimum band selection for supervised classification of multispectral data. *Photogrammetric Engineering and Remote Sensing*, 56:55–60, 1990.

[23] B. J. Irvin, S. J. Ventura, and B. K. Slater. Fuzzy and isodata classification of landform elements from digital terrain data in pleasant valley, wisconsin. *Geoderma*, 77:137–154, 1997.

[24] J. R. Jensen. *Introductory digital image processing: A remote sensing perspective*. Prentice Hall, NJ, 1996.

[25] M. U. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.

[26] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

[27] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *In Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability*, I: Statistics:281–297, 1967.

[28] J. C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4:95–104, 1974.

[29] J. C. Bezdeck, R. Ehrlich, and W. Full. FCM: Fuzzy C-Means Algorithm. *Computers and Geoscience*, 10:191–203, 1984.

[30] A. Gulli. Antonio Gulli's coding playground. `http://codingplayground.blogspot.com.ar/2009/04/fuzzy-clustering.html`, 2009. [Accedido 02-08-2012].

[31] I.O.A Odeh, A.B McBratney, and D.J Chittleborough. Soil pattern recognition with fuzzy-c-means: Application to classification and soilâlandform interrelationships. *Soil Science Society of America Journal*, pages 505–516, 1992.

[32] X. L. Xie and G. Beni. A Validity Measure for Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:841–847, 1991.

[33] R. C. Gonzalez and R.E. Woods. *Digital Image Processing*. Pearson Prentice Hall, Upper Saddle River – NJ, 2008.

[34] ESRI Shapefile Technical Description. *Digital Image Processing*. Environmental Systems Research Institute, Inc., 1998.

[35] Colaboradores de Wikipedia. Shapefile. `http://en.wikipedia.org/wiki/Shapefile`, 2006. [Accedido 16-01-2012].

[36] Computational Geometry Algorithms Library. `http://www.cgal.org`, 1998. [Accedido 05-10-2012].

[37] N. Ritter. listgeo - Dump GeoTIFF Metadata. `http://www.remotesensing.org/geotiff/listgeo.html`. [Accedido 05-10-2012].

[38] CImg Library. `http://cimg.sourceforge.net`. [Accedido 10-10-2012].

[39] Shapefile C Library. `http://shapelib.maptools.org`. [Accedido 08-11-2012].

[40] Librería QT. `http://qt-project.org/`. [Accedido 20-10-2012].

[41] R. Miretti, E. Cerati, and L. Coronel. *Cartografía Matemática*. Universidad Nacional del Litoral, Ediciones UNL, 2012.

[42] A. M. Felicísimo. *Modelos digitales del terreno. Introducción y aplicaciones en las ciencias ambientales*. Pentalfa, Oviedo, 1994.

APÉNDICE A

Evaluación de los algoritmos de clasificación

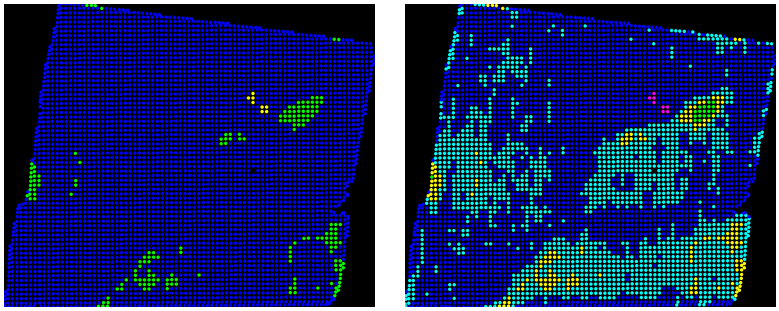
A.1 Selección del algoritmo de Clustering

Se realizaron pruebas comparativas para analizar el comportamiento y rendimiento de los siguientes algoritmos:

- Algoritmo Jerárquico.
- Algoritmo K-means.
- Algoritmo C-means difuso.

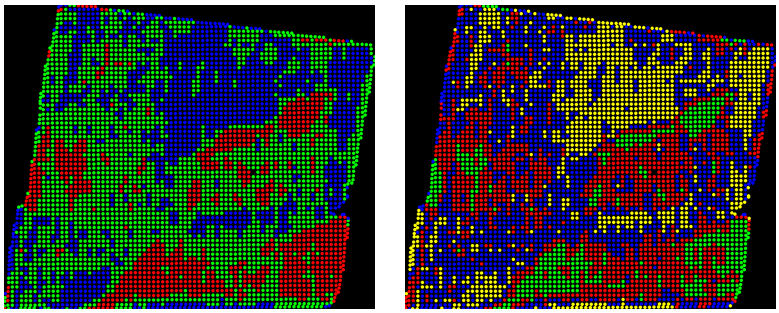
A continuación se presentan las representaciones gráficas de las zonas encontradas por los distintos algoritmos a partir de la clasificación en 3 y 5 clusters. Para esto se utilizó un conjunto de datos de 4680 puntos que contienen tres variables: rendimiento, altimetría y una imagen satelital. En las figuras A.1(a) y A.1(b) se pueden observar las clasificaciones producidas por el algoritmo Jerárquico para 3 y 5 zonas de manejo respectivamente. En tanto que las figuras A.2(a) y A.2(b) corresponden a los resultados de la clasificación del algoritmo K-means en las mismas zonas. Por último, en las figuras A.3(a) y A.3(b) se muestran las zonas encontradas por el algoritmo FCM.

Los resultados obtenidos con el algoritmo Jerárquico para la clasificación en 3 zonas no son los esperados, ya que define una gran cantidad de puntos para una sola zona, la segunda zona se encuentra esparcida por diversos



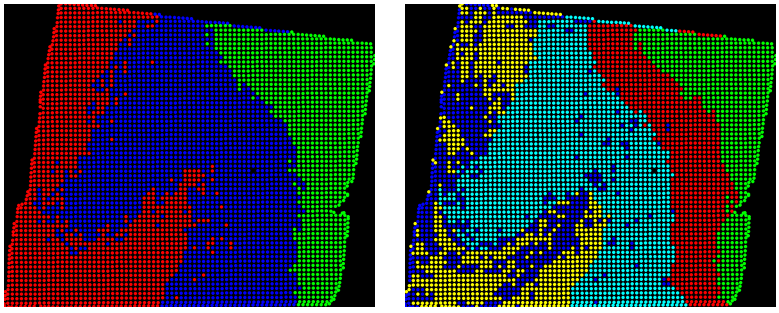
(a) Clasificación en 3 zonas. (b) Clasificación en 5 zonas.

Figura A.1: Clasificación con algoritmo Jerárquico.



(a) Clasificación en 3 zonas. (b) Clasificación en 5 zonas.

Figura A.2: Clasificación con algoritmo K-means.



(a) Clasificación en 3 zonas. (b) Clasificación en 5 zonas.

Figura A.3: Clasificación con algoritmo FCM.

lugares del campo y la tercera contiene escasa cantidad de puntos. Esto no justifica llevar adelante un proceso diferenciado de las zonas. En la figura A.1(b) se pueden observar dos zonas que poseen la mayor cantidad de puntos (zona azul y zona celeste) y el resto de las zonas (zona amarilla, zona

verde y zona rosada) son muy pequeñas y se encuentran contenidas dentro de las zonas más grandes, pudiendo ser absorbidas por éstas. La clasificación realizada por el algoritmo K-Means resultó ser una agrupación muy dispersa, con zonas que no están bien definidas y que contienen puntos dispersos por distintas partes del campo. Esto ocurre para las dos clasificaciones ya sea en 3 o 5 cluster. Por último en las figuras A.3(a) y A.3(b) se muestran los resultados de la clasificación con el algoritmo FCM, aquí se observan zonas bien definidas y con contornos delimitados. Si bien existen puntos espúreos, cada zona esta concentrada en un sector del campo.

Es por ésto, y por la eficacia demostrada en programas como MZA que el algoritmo seleccionado para llevar adelante la tarea de clasificación es el FCM.

APÉNDICE B

Resultados originales de los índices de validación

B.1 Pruebas de la clasificación

Para todas las pruebas de clasificación se consideran como parámetros de entrada los siguientes valores: criterio de convergencia = 0.001, máximo número de iteraciones = 300, cantidad mínima de grupos = 2 y cantidad máxima de grupos = 5.

Prueba 5

Para realizar esta prueba se utiliza el conjunto de datos 1. Se crea un mallado regular con una separación de 10 metros entre punto y punto mediante el método de interpolación cuadrática, quedando un total de 10103 puntos.

Tabla B.1: Resultados prueba 5 con exponente 1.3.

Índices	2 cluster	3 cluster	4 cluster	5 cluster
XB	1.72	0.62	**0.41**	2.56
FPI	0.75	0.73	0.70	**0.69**
NCE	**0.21**	0.32	0.42	0.48
Dist. Euclídea	1.89	1.01	**0.91**	2.69

En las tablas B.1 y B.2 se observa que tanto para la clasificación con el exponente 1.3, como para el 1.5, la clasificación en 4 grupos es la que obtiene

Tabla B.2: Resultados prueba 5 con exponente 1.5.

Índices	2 cluster	3 cluster	4 cluster	5 cluster
XB	1.91	0.72	**0.49**	1.10
FPI	0.88	0.84	0.84	**0.83**
NCE	**0.11**	0.18	0.21	0.25
Dist. Euclídea	2.10	1.12	**0.99**	1.40

los mejores resultados. Seguida por la clasificicación en 3 grupos.

Prueba 6

Se utilizó el conjunto de datos 2 con la interpolación cuadrática y con un mallado de 5 metros. Esta variación respecto de la prueba anterior se debe a que se cuenta con un menor número de puntos (4215 puntos). En las tablas B.3 y B.4 se observan los resultados de las clasificaciones realizadas utilizando los exponentes difusos 1.3 y 1.5 respectivamente.

Tabla B.3: Resultados prueba 6 con exponente 1.3.

Índices	2 cluster	3 cluster	4 cluster	5 cluster
XB	1.34	2.16	1.16	**0.60**
FPI	0.91	**0.89**	0.93	0.93
NCE	**0.08**	0.13	0.09	0.09
Dist. Euclídea	1.62	2.34	1.49	**1.11**

Tabla B.4: Resultados prueba 6 con exponente 1.5.

Índices	2 cluster	3 cluster	4 cluster	5 cluster
XB	1.27	1.88	1.06	**0.58**
FPI	0.84	**0.83**	0.87	0.88
NCE	**0.13**	0.20	0.17	0.18
Dist. Euclídea	1.53	2.06	1.38	**1.06**

Se puede observar en las tablas B.3 y B.4 que la agrupación en 5 clusters obtiene la menor distancia euclídea, seguida de la clasificación en 4 grupos.

Prueba 7

Se utilizó el conjunto de datos 3 con la interpolación cuadrática y con un mallado de 10 metros, quedando un total de 12109 puntos. En las tabla B.5

y B.6 se observan los resultados de las clasificaciones realizadas utilizando los exponentes difusos 1.3 y 1.5 respectivamente.

Tabla B.5: Resultados prueba 7 con exponente 1.3.

Índices	2 cluster	3 cluster	4 cluster	5 cluster
XB	0.95	**0.44**	0.68	1.17
FPI	0.93	**0.91**	0.92	0.92
NCE	**0.06**	0.10	0.10	0.11
Dist. Euclídea	1.33	**1.02**	1.15	1.49

Tabla B.6: Resultados prueba 7 con exponente 1.5.

Índices	2 cluster	3 cluster	4 cluster	5 cluster
XB	0.90	**0.41**	0.63	1.08
FPI	0.87	**0.85**	0.86	0.86
NCE	**0.11**	0.17	0.18	0.20
Dist. Euclídea	1.25	**0.96**	1.08	1.40

Para esta prueba los índices reflejan que la clasificación en 3 clusters para los dos exponentes difusos, es que la mejor resultados arroja, debido a la menor distancia euclídea lograda.

Para todas las pruebas realizadas los valores que obtiene el índice *XB* supera ampliamente el rango de valores que los otros índices pueden obtener, originando q la distancia euclídea esté totalmente ligada al valor del índice *XB*. Por este motivo se necesita normalizar los índices para eliminar la preeminencia que existe de unos sobre otros.