Universidad de la República – Facultad de Ingeniería

Proyecto de Grado

"Desarrollo de un módulo de generación de zonas de manejo para un Sistema de Información Geográfico (SIG) de Gestión Agrícola"

INFORME FINAL

Marzo 2008

Estudiantes:
Anastasia Rava – 3.305.333-3
Claudia Santa Ana – 4.646.121-8

Tutores:
Mercedes Berterretche
Omar Viera

RESUMEN

Es sabido que en los últimos años la Informática se ha aplicado en muchas y diversas áreas con el propósito de brindar beneficios y mejoras a la situación actual. La agricultura tampoco escapa al avance tecnológico e informático.

La respuesta de la investigación en las áreas relacionadas a la agricultura para enfrentar nuevos desafíos, ha sido la generación de tecnología que permita desarrollar técnicas que consideren y manejen diferenciadamente la variabilidad del área productiva.

Históricamente, las herramientas de producción agrícola han sido desarrolladas para cultivar grandes espacios de una manera uniforme, no diferenciada. Pero desde hace un tiempo los productores han reconocido que una chacra no es una unidad homogénea y que varía por su naturaleza y por la actividad del hombre.

La Agricultura de Precisión se refiere al ajuste en el manejo de los cultivos a las diferentes condiciones del suelo dentro de una chacra, es decir, suministrar a cada parte del suelo los insumos necesarios para optimizar la producción. Podría ejemplificarse como una especie de "atención personalizada" de las diferentes partes de nuestra superficie productiva de acuerdo a sus necesidades. Es un objetivo que no es nuevo, pero que puede llevarse adelante de una manera más eficiente hoy en día gracias a las tecnologías disponibles.

El proyecto presentado forma parte de un proyecto de mayor tamaño desarrollado por el ICA[1], con el propósito de incursionar en la aplicación de la Agricultura de Precisión en nuestro país. En particular, corresponde a la generación de zonas de manejo homogéneas para un área de cultivo a partir de medidas de factores productivos. Las zonas se obtienen mediante la utilización de técnicas de Data Mining las cuales permiten obtener información a partir de los datos. La información obtenida resulta relevante a la hora de tomar decisiones sobre un área de cultivo, permitiendo cosechar beneficios respecto a la sustentabilidad ambiental y económica del proceso de producción.

Si bien la dificultad radica en la obtención de datos y los costos de las herramientas involucradas, los resultados obtenidos resultan de cierta forma alentadores, ya que representan un paso hacia adelante para continuar con este desafío.

Palabras claves: Agricultura de Precisión, Data Mining, zonas de manejo, área de cultivo.

[1] Ingenieros Consultores Asociados.

Tabla de contenido

Índice de Figuras

CAPÍTULO 1: Introducción

En este capítulo se define y describe el Proyecto de Grado en términos generales, así como también los principales conceptos que se relacionan y dan origen al mismo. También se exponen los objetivos y motivación para llevar adelante la propuesta. Se presentan resumidamente las conclusiones obtenidas y por último, se describe la organización general de este documento.

Descripción General del Proyecto

El proyecto propuesto consiste en la implementación de un módulo de generación de zonas de manejo de un área de cultivo a partir de datos obtenidos en una zafra[2]. Dicho módulo, será utilizado por un Sistema de Información Geográfico (SIG) (1) de Gestión Agrícola que permitirá visualizar las zonas generadas.

Los datos que utiliza la aplicación corresponden a medidas de los factores productivos relevantes para el desarrollo del cultivo, como pueden ser: suelo, rendimiento, conductividad eléctrica, etc.

Como resultado de su ejecución, la aplicación genera archivos de texto con la especificación y descripción de las zonas de manejo para el área que se está considerando y en base a los datos de la zafra dados. También se generan índices de performance que permiten determinar si la cantidad de zonas generadas y las características de las mismas son las óptimas.

Para generar las diferentes zonas de manejo se utilizó un algoritmo de Data Mining (2), los cuales permiten detectar patrones o modelos de comportamiento ocultos.

Objetivos y Motivación

Como estudiantes de la carrera de Ingeniería en Computación, la principal motivación de este proyecto, es la aplicación de la Informática en un área tan "diferente" como lo es la agricultura. Puede resultar difícil pensar que estas áreas pueden tener algo que ver o que, mejor aún, se pueden complementar para obtener todo tipo de beneficios.

En el proceso tradicional de la agricultura, todas las áreas productivas se tratan de forma uniforme; primero se establecen objetivos de producción y luego se aplica el mismo tratamiento en toda el área. En la Agricultura de Precisión (AP) (3) cada campo es "gerenciado" independientemente. Primero se obtienen y almacenan las características específicas de cada zona. Luego se manipulan los datos utilizando por ejemplo, Sistemas de Información Geográfica. Y por último se aplica un tratamiento específico en cada zona dependiendo de sus características.

La AP consiste en la gestión agronómica diferenciada del terreno en función de la variabilidad espacial[3] presente. Si el campo fuera uniforme no haría falta este tipo de

[2] Época, tiempo o período de gran intensidad laboral en que transcurre el desarrollo del cultivo.

[3] Expresa las diferencias de producción en un mismo campo, en una misma campaña y cosecha.

agricultura, pero como la mayor parte de los predios son heterogéneos, se considera la variabilidad espacial como criterio.

La aplicación del concepto de AP está siendo posible gracias a la evolución de diferentes tecnologías. Esto permite la aplicación de insumos agrícolas como fertilizantes, semillas, plaguicidas, etc., en forma variable dentro de un área de cultivo, de acuerdo a los requerimientos y/o potencial productivo de varios sectores homogéneos, pre-definidos dentro del mismo. Es en la obtención de dichos sectores homogéneos donde la Informática ocupa un rol fundamental.

El análisis, procesamiento e interpretación de la información recolectada incluye varias actividades. Una de ellas es el análisis de factores como el tipo de suelo, fertilidad, acidez, variedad de semilla, entre otros, que influyen en la productividad. Todos estos factores varían en el espacio y en el tiempo, por lo tanto, las decisiones de manejo deben ser relativas a un lugar y momento específico, y no rígidamente programadas, como ocurre en la actualidad. (4)

El objetivo principal del proyecto, consiste en la generación de zonas de manejo o sectores homogéneos de un área de cultivo como resultado del análisis de los factores productivos. Dichas zonas son un apoyo a la toma de decisiones a la hora de aplicar insumos en el área de cultivo.

Objetivo general: que el módulo generado pueda ser utilizado por un Sistema de Información Geográfica de Gestión Agrícola.

Objetivos específicos:

- ✓ Realizar y documentar una investigación sobre Agricultura de Precisión como base para comprender el contexto del proyecto.
- ✓ Realizar y documentar una investigación sobre Data Mining como técnica que nos proveerá una estrategia para implementar la solución al problema planteado.
- ✓ Que la implementación del módulo cumpla con ciertas características o requerimientos definidos por el cliente, los cuales se describen más detalladamente en el Capítulo 4 de este documento.

Metodología

Para poder cumplir los objetivos del proyecto, se recopilaron y analizaron diferentes fuentes de información que permitieran obtener el conocimiento necesario para comprender el contexto del proyecto y la aplicabilidad del mismo, así como los nuevos conceptos y terminología involucrada.

También se investigaron herramientas existentes que implementaran una funcionalidad igual o semejante a la que se pretende para poder tener como referencia.

Con el conocimiento obtenido y ya con un panorama más amplio del problema a resolver, se realizó un análisis y se planteó una propuesta de solución al problema. Luego, se llevó la misma a un nivel más detallado de diseño, donde se buscó que se

cumpliera con los criterios que debe cumplir una buena solución de desarrollo de software.

Finalmente se realizó la implementación y correspondiente verificación y validación, evaluando así los resultados obtenidos.

Resultados Obtenidos

Se realizaron las investigaciones planteadas, las cuales fueron parte fundamental del proyecto para la inserción en el tema a tratar y el aprendizaje de los conceptos y objetivos relacionados a la agricultura.

Se logró la realización del módulo de generación de zonas de manejo con las características especificadas. No obstante, consideramos que la validación del mismo no fue suficiente. Esto se debió a los pocos datos de prueba con los que se contó, lo cual sabemos que se debe a la dificultad y el costo de la obtención.

Organización del Documento

El presente documento continúa de la siguiente forma: en el Capítulo 2 se realiza una descripción del proyecto, planteando de forma más detallada el problema a resolver.

En el Capítulo 3, se exponen los resultados de las investigaciones realizadas para el proyecto. Se analiza el contexto del problema a resolver, introduciendo el concepto de Agricultura de Precisión y se presenta el concepto de Data Mining como técnica a utilizar para la resolución del problema.

En el Capítulo 4 se plantean los requerimientos y restricciones que debe cumplir el sistema a construir.

En el Capítulo 5 se expone la solución propuesta y sus principales características.

En el Capítulo 6 se detalla el diseño y la implementación realizada. Se especifican las funcionalidades, restricciones y decisiones tomadas en esta etapa del proceso. Finalmente se propone un plan de pruebas preliminar para la verificación de la solución.

En el Capítulo 7 se muestran las pruebas realizadas, comparación y análisis de las mismas y se verifica si se lograron los objetivos y en qué forma.

Finalmente, en el Capítulo 8, se presenta la conclusión final del trabajo, evaluando los resultados obtenidos. Se describen las principales dificultades encontradas a lo largo del proyecto y se mencionan posibles extensiones. Se realiza una autocrítica mencionando qué se hizo y qué no. Por último se plantea una opinión personal de las integrantes en cuanto al proyecto realizado.

Al final del documento se incluyen las referencias bibliográficas, un glosario y anexos.

CAPÍTULO 2: Descripción del problema

Como se mencionó anteriormente, el proyecto propuesto consiste en la implementación de un Sistema, en particular un módulo reutilizable, de generación de zonas de manejo para un área de cultivo a partir de datos obtenidos en una zafra, teniendo en cuenta su variabilidad espacial y las relaciones entre las diferentes variables. Dicho módulo, forma parte de un proyecto de mayor tamaño desarrollado por el ICA (5), y será utilizado por un Sistema de Información Geográfica de Gestión Agrícola que permitirá visualizar las zonas generadas.

El proyecto original de mayor tamaño mencionado anteriormente, se realiza en el marco del Programa de Desarrollo Tecnológico y tiene por objetivo principal el desarrollo de un conjunto de herramientas y metodologías que faciliten la aplicación de la Agricultura de Precisión, considerando las condiciones de producción de Uruguay. En el proyecto participan La Hectárea, una consultora de servicios agronómicos de Dolores, ICA, una empresa de consultoría en informática, cuatro productores agropecuarios en cuyos establecimientos se realizan los trabajos de campo y se cuenta con la colaboración del INIA (Instituto Nacional de Investigación Agropecuaria) (6).

Se considera una chacra y un conjunto de variables medidas en la misma durante una zafra. Dichas variables corresponden a medidas de los factores productivos relevantes para el desarrollo del cultivo y medidas de resultado como pueden ser: propiedades del suelo como conductividad eléctrica, propiedades de los cultivos como reflectancia, o el rendimiento final de la cosecha. La salida es una división espacial de la chacra en zonas de manejo homogéneas para esa zafra en particular.

En primer lugar, se debe realizar un análisis de las variables. Esto implica analizar las correlaciones espaciales de las diferentes capas de datos, de forma de encontrar las variables más correlacionadas y por lo tanto más influyentes.

En cuanto a la generación de zonas, lo que se busca es dividir el conjunto de datos de entrada en diferentes zonas de manera de agrupar los datos con características similares para que puedan ser tratados colectivamente como un grupo. Esto implica obtener zonas de forma que los elementos posean características similares dentro de una zona y diferentes entre zonas.

Se debe poder especificar el número de zonas a generar, es decir, la cantidad de agrupaciones o subdivisiones a formar. También se deben generar índices de performance que permitan determinar si la cantidad de zonas generadas y las características de las mismas son las óptimas.

Existe un requerimiento muy importante a tener en cuenta, y es el tamaño de las zonas generadas. El mismo debe cumplir un mínimo para que dicha zona pueda ser tratada. No es viable el tratamiento de una zona pequeña en lo que se refiere a la relación costo-beneficio. Igualmente, el tamaño mínimo no debe exceder el tamaño de la chacra que se está considerando.

La entrada y salida de la aplicación es por medio de archivos de textos. En el caso de la entrada, el archivo contiene los datos de las variables. Como resultado de su ejecución, genera archivos de texto con la especificación y descripción de las zonas de manejo para el área que se está considerando y en base a los datos de la zafra dados.

La especificación se refiere a decir a qué zona pertenece cada punto o coordenada. La descripción de las zonas, brinda para cada variable y cada zona, información correspondiente a las estadísticas básicas de las variables (máximo, mínimo, media, desviación estándar). Esto servirá a los agrónomos para comprender las causas de la variabilidad del rendimiento.

Para entender en mayor profundidad el problema de la generación de zonas de manejo y el proceso de producción agrícola en general, se realizó y documentó una investigación sobre Agricultura de Precisión. Para encontrar un algoritmo que permita manipular la información para obtener zonas con las características planteadas, se realizó y documentó una investigación sobre Data Mining en general y sobre varios algoritmos en esa área. En el siguiente capítulo se presenta un resumen de ambas investigaciones. Los documentos completos son presentados como anexos de este documento.

Herramientas disponibles

Si bien se encontró que existen muchas y variadas herramientas de software para la Agricultura de Precisión, para el caso particular de generación de zonas de manejo se investigó la siguiente herramienta.

MZA (Management Zone Analyst) (7)

MZA es un programa de software libre que fue desarrollado utilizando el algoritmo de clustering no supervisado fuzzy c-means, el cual asigna la información de campo a clases similares o potenciales zonas de manejo. Una ventaja de MZA sobre muchos otros programas de software es que provee una salida concurrente para un rango de números de clases de forma que el usuario puede evaluar cuantas zonas de manejo deberían ser usadas.

Management Zone Analyst fue desarrollado utilizando Microsoft Visual Basic 6.0[4] (8) y funciona en cualquier computadora con Microsoft Windows 95 o superior. Los conceptos y teorías detrás de MZA son exhibidos, ya que conforman los pasos secuenciales del programa.

Management Zone Analyst calcula estadísticas descriptivas, ejecuta el algoritmo de clasificación borroso no supervisado para un rango de números de clases, y provee al usuario dos índices de performance [Fuzziness Performance Index (FPI) y Normalized Classification Entropy (NCE)] para ayudarlo a decidir cuantas clases o agrupaciones son las más apropiadas para crear las zonas de manejo.

[4] Lenguaje de programación.

CAPÍTULO 3: Estado del Arte

En éste capítulo se busca brindar un panorama general de los resultados obtenidos en las investigaciones realizadas sobre Agricultura de Precisión (AP) y Data Mining (DM), que permitirán entender tanto el problema a resolver y como la solución planteada.

Introducción

Dado que uno de los beneficios de aplicar AP es uno de los fines de nuestro sistema, optimizar el uso de insumos agrícolas en función de la variabilidad espacial y temporal de la producción (3); se realizó en primera instancia una investigación de AP buscando obtener una visión global sobre el tema.

Luego se realizó un estado del arte de DM, cuyo objetivo fue plasmar el resultado de la investigación correspondiente, que permitió conocer, de alguna forma, la variedad de técnicas que existen para obtener conocimiento a partir de los datos. Y encontrar la más apropiada para aplicar en el contexto del Proyecto de Grado al cual está relacionado este trabajo. La especificación de la técnica seleccionada y su aplicación particular al problema en cuestión se describen en el Capítulo 5.

Agricultura de Precisión

La Agricultura de Precisión, también conocida como Manejo de Sitio Específico (SSCP: Site Specific Crop Management) (9), es un proceso en el que se combinan la tecnología, la Informática y la agricultura para producir cultivos de forma eficiente. Esta combinación permite dar a cada zona del campo el tratamiento agronómico más apropiado, tanto desde el punto de vista económico-productivo como del ambiental. (10)

Por medio de su correcta aplicación, es posible incrementar la producción, reducir los costos de aplicación de insumos, tener en cuenta las necesidades del cultivo, reducir los impactos ambientales, mejorar la planificación del tiempo a nivel de cultivo, etc. según sea el o los objetivo planteados.

Dicha aplicación está siendo posible gracias a la evolución de tecnologías como los Sistemas de Información Geográfica (SIG), Sistemas de Posicionamiento Global (GPS) (11), computadores, maquinarias, percepción remota[5] (12), monitores de rendimiento, tecnologías de dosis variable y aplicación variable de insumos (VRA) para automatizar el manejo de sitios específicos. (3) Esto permite la aplicación de insumos agrícolas en forma variable dentro de un área de cultivo, de acuerdo a los requerimientos y/o potencial productivo de varios sectores homogéneos, pre-definidos dentro del mismo.

Los SIG permiten generar mapas con diferentes parámetros del suelo permitiendo una visualización que es fundamental para tomar decisiones del cultivo. También

[5] Adquisición de información sobre las propiedades de un objeto empleando instrumentos que no están en contacto directo con el objeto estudiado.

permiten determinar superficies trabajadas, presentar visualmente información para facilitar su análisis, realizar mapas de lectura, etc.

La percepción remota permite medir la calidad de la tierra, determinar la aptitud del suelo, separar especies de vegetales, etc.

Los monitores de rendimiento son la base para la delimitación de sectores del suelo con productividad similar. Permiten recolectar datos sobre la cosecha y brindar información sobre el rendimiento.

Las tecnologías de dosis variable permiten aplicar el insumo apropiado o la cantidad correcta del mismo, a un sitio dependiendo de su rendimiento.

Proceso

La AP consiste en la gestión agronómica diferenciada del terreno en función de la variabilidad espacial [6] presente. En la misma se pueden distinguir tres etapas importantes que forman un ciclo: obtención de datos, análisis de datos y aplicación de insumos. En cada una de ellas es generada una "salida" que se utiliza como "entrada" en la etapa siguiente. Dicho ciclo está representado en la Figura 1.

Figura 1: Ciclo de Agricultura de Precisión

La primera etapa, obtención de datos, es iniciada en la cosecha. En ella son recolectados los datos relevantes por medio de herramientas específicas. El uso de maquinaria adecuada genera información de productividad metro a metro del área cultivada. Después es necesario el levantamiento de información directamente en el campo, asociada, por ejemplo, al análisis de fertilidad de la tierra, al impacto de procesos erosivos y a la interferencia de otros factores que puedan causar variaciones o alteraciones en la productividad.

La segunda etapa, análisis de datos, corresponde al análisis, procesamiento e interpretación de la información recolectada. Existen diferentes técnicas y métodos para analizar en detalle la información obtenida y delimitación de zonas de

[6] Expresa las diferencias de producción en un mismo campo, en una misma campaña y cosecha.

comportamiento productivo diferente dentro de un área de cultivo. Dentro de ellas se encuentran la geoestadística y el análisis de clústeres.

La tercera y última etapa, aplicación de insumos, comprende la aplicación diferencial de insumos e incluye la aplicación variable de fertilizantes, pesticidas y semillas.

Finalmente, el proceso se completa con la nueva cosecha (volviendo así al comienzo del ciclo).

La Agricultura de Precisión en Uruguay

La AP está utilizándose en muchos países, entre los cuales se encuentran Estados Unidos, Europa, Argentina, Reino Unido, Dinamarca, Alemania, Suecia, Francia, Holanda, Bélgica, Brasil, Chile, Australia, Sudáfrica y China.

Uruguay tampoco ha escapado a esta tecnología, aunque se encuentra en un nivel de escasa adopción, se está instalando cada vez más en los últimos años. Esto se debe, entre otros, al gran aumento en el área agrícola que se ha producido en este tiempo.

Los productores de trigo, soja y arroz, sobretodo del litoral sur, están utilizando actualmente AP.

Una de las tantas razones es el aumento progresivo del área agrícola y de productores argentinos, dado que en Argentina la AP quedó inmersa y es aplicada por gran parte de los productores, sobre todo aquellos que vienen a buscar tierras a nuestro país.

Data Mining

Data Mining o Minería de Datos, surge como la necesidad de obtener conocimiento, factor fundamental para mejorar la productividad. En muchos casos, se almacenan grandes cantidades de datos, que contienen información valiosa pero que, a simple vista, no es evidente. Se necesitan medios automáticos o semiautomáticos que permitan procesar grandes volúmenes de datos buscando descubrir información valiosa en los mismos.

Definición

Existen distintas definiciones de Data Mining, que pueden resumirse en la siguiente: “DM es el análisis semiautomático de relaciones existentes en el contenido de una base de datos de gran tamaño, buscando encontrar información oculta, patrones, modelos, estructuras e información útil en general, para la toma de decisiones”.

Su utilización se centra en aquellas organizaciones o empresas que recolectan gran cantidad de información sobre sus productos, insumos o servicios.

DM tiene distintas y variadas aplicaciones, algunas relacionadas al objetivo de nuestro Proyecto de Grado. Por ejemplo, con la información existente sobre la variable

y datos históricos relacionados a la misma se elaboran modelos que permitan estimar como será la evolución del comportamiento de la variable en el futuro.

Etapas del proceso de Data Mining

DM no se refiere solo la aplicación de un algoritmo, sino que es un proceso que incluye la selección y el pre-procesamiento de los datos, la obtención de un modelo de conocimiento, la interpretación y evaluación de los resultados, y la aplicación del conocimiento obtenido. (13) (14) (15) (16) En la Figura 2 se muestran las etapas de dicho proceso.

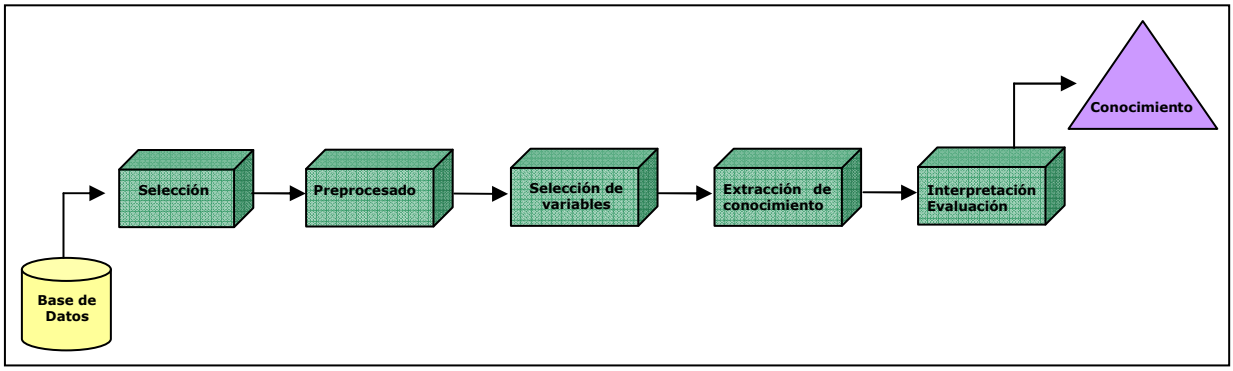


Figura 2: Etapas del proceso de Data Mining.

El primer paso consiste en seleccionar las fuentes de datos que son relevantes para el área donde se quiere aplicar DM.

Luego, en el pre-procesamiento, se filtran los datos de forma de eliminar valores incorrectos, no válidos o desconocidos. Se realiza la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos, según las necesidades y el algoritmo a utilizar.

En la selección de variables, se reduce el tamaño de los datos eligiendo las variables más influyentes en el problema.

La etapa de extracción de conocimiento se refiere a la aplicación de alguna/s técnica/s de DM mediante la cual se obtiene un modelo de conocimiento.

Luego se debe proceder a la validación de lo obtenido en el punto anterior, comprobando que las conclusiones que muestran sean válidas, coherentes y suficientemente satisfactorias. Para la interpretación de los resultados obtenidos, generalmente resulta de ayuda utilizar una técnica de visualización que permite ver los resultados de una manera más descriptiva.

Por último, es importante incorporar el conocimiento descubierto al sistema, normalmente para mejorarlo, lo cual puede incluir resolver conflictos potenciales con el conocimiento existente.

Técnicas

Los modelos y algoritmos aplicados en DM se apoyan o basan principalmente en los campos de estadística e Informática, donde las técnicas son conocidas y estudiadas hace mucho tiempo. Se decidió adoptar la clasificación propuesta en la bibliografía (17) y (18), donde se expone que las técnicas más frecuentes de DM pueden ser catalogadas en dos tipos dependiendo de su finalidad:

- *Descriptivas*: el objetivo de estos procedimientos es la búsqueda de la caracterización o discriminación de un conjunto de datos. Las técnicas más conocidas son: Agrupamiento o Clustering, Reglas de Asociación, Análisis de Patrones Secuenciales, Análisis de Componentes Principales y Detección de Desviación.
- *Predictivas*: el propósito de estos métodos es aprender una hipótesis la cual pueda clasificar a nuevos individuos. Están orientados a estimar valores de salida. Los algoritmos principales son: Regresión y Clasificación (Árboles de Decisión, Clasificación Bayesiana, Redes Neuronales, Algoritmos Genéticos y Lógica Difusa).

De todas las técnicas presentadas se decidió profundizar en aquellas que parecían adaptables al problema de la generación de zonas de manejo homogéneas.

Reglas de Asociación

Las reglas de asociación detectan eventos asociados que se ocultan en las bases de datos. Son reglas que relacionan un conjunto de pares atributo-valor con otros pares atributo-valor. Una regla de asociación se define de la siguiente forma:

$$A \rightarrow B \mid n\%$$

donde → significa implica, n% es el factor de confianza de la regla y A y B son conjuntos de ítems. A es llamado antecedente y B es llamado consecuencia. Y se lee como: “El n% de las veces que ocurre A, ocurre B”.

Análisis de Patrones Secuenciales

Se define el análisis o detección de patrones secuenciales como el intento de modelar la evolución temporal de alguna variable, con fines descriptivos o predictivos. (2) Equivale a detectar la asociación entre eventos con determinadas relaciones temporales y se basa en el concepto de secuencia de conjunto de elemento.

Se trata de establecer asociaciones del estilo:

“si ocurre X en T, ocurrirá Y en T + P”. (19)

Podemos decir que para esta técnica, el objetivo de analizar una base de datos es encontrar secuencias temporales de eventos tales que su significancia estadística sea mayor que un umbral especificado por el usuario.

Para ello, existen varios algoritmos propuestos, como GSP, PrefixSpan, SPADE, CloSpan, FreSpan. (20) (21)

Análisis de Componentes Principales (ACP)

Permite analizar la estructura de los datos y proporcionar herramientas de visualización. (22) Es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible, donde los nuevos componentes principales o factores son independientes entre sí y serán una combinación lineal de las variables originales. (23)

Un aspecto clave en ACP es la interpretación de los factores, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los factores con las variables iniciales. Tiene sentido siempre que existan altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante y, por tanto, pocos factores explicarán gran parte de la variabilidad total.

Detección de desviación

Se centra en la detección de desviaciones, es decir, datos raros que producen ruido. Estos métodos consideran que las desviaciones pueden esconder conocimientos verdaderamente inesperados e interesantes. (24) El problema está en determinar cuándo una desviación es significativa para ser de interés. (15)

Típicamente, los métodos para la detección de desviaciones emplean información adicional a los datos, por ejemplo: condiciones preestablecidas o restricciones de integridad. Y en algunas ocasiones estos métodos aprovechan la propia redundancia de los datos.

Regresión

Esta técnica persigue la obtención de un modelo que permita predecir el valor numérico de alguna variable continua. Es posible predecir el valor de dicha variable a partir de la evolución sobre otra variable continua, generalmente el tiempo, o sobre un conjunto de variables. (2) (19)

Dentro de regresión consideramos la regresión lineal, que es una técnica estadística comúnmente empleada para ajustar un conjunto de observaciones o puntos de n dimensiones con la variable objetivo y.

Clasificación

Se trata de obtener un modelo que permita asignar un caso de clase desconocida a una clase concreta (seleccionada de un conjunto predefinido de clases). (2)

Los algoritmos de clasificación agrupan individuos o variables en clases que muestran un comportamiento homogéneo y, por lo tanto, permiten descubrir patrones de comportamiento. (22)

Agrupamiento o Clustering

Las técnicas de agrupamiento son utilizadas tanto para identificar grupos con determinadas características como para determinar a qué grupo pertenece un elemento nuevo. (25) (26) Este procesamiento de los datos, ayuda a los usuarios a entender el agrupamiento natural o estructura en un conjunto de datos.

Pertenece a las técnicas de clasificación no supervisadas, esto significa, que al momento de hacer la clasificación no conocemos las propias clases y probablemente no sepamos la cantidad de clases. Para medir la similitud de los elementos y poder agrupar los mismos se utilizan distintas formas de distancia: Euclideana, de Manhatan, etc. Mientras más cerca estén los objetos según la distancia utilizada estos son más similares.

Los algoritmos de agrupamiento se pueden clasificar en los siguientes tipos (27): algoritmos basados en particiones, basados en rejillas, basados en modelos, basados en densidades y algoritmos jerárquicos.

Algoritmos basados en particiones

Construyen varias particiones y las evalúan utilizando algún criterio. Cada partición tiene al menos un elemento y cada elemento pertenece a una sola partición. Crean una partición inicial e iteran hasta un criterio de paro.

Los algoritmos de clustering basados en particiones organizan los objetos dentro de k clústeres de tal forma que sea minimizada la desviación total de cada objeto desde el centro de su clúster o desde una distribución de clústeres. La desviación de un punto puede ser evaluada en forma diferente según el algoritmo, y es llamada generalmente función de similitud.

K-medias

Se eligen una serie de valores del espacio y a partir de ellos se comienzan a generar clases. Cada vez que se presenta una nueva instancia se calcula su distancia a todas las medias y se le asigna la clase cuya media sea la más cercana. (28)

K-medianas

En este tipo de algoritmos, cada clúster está representado por uno de los objetos en el clúster. No utiliza un punto del espacio como centroide, sino una instancia real perteneciente a los datos (medoide o mediana).

Busca objetos “representativos”, llamados medianas[7], en los clústeres (usa como centros las medianas y no las medias). Para ello solo se necesita la definición de distancia entre dos objetos.

Existen otros algoritmos basados en particiones que no serán explicados en este documento.

[7] Instancia del clúster más centrada.

Algoritmos jerárquicos

Los algoritmos jerárquicos crean una descomposición jerárquica de los datos agrupándolos en un árbol de clústeres llamado Dendograma, que divide recursivamente el conjunto de datos en conjuntos cada vez más pequeños. El árbol puede ser creado de dos formas: de abajo hacia arriba (bottom-up) o de arriba hacia abajo (top-down). (29)

Algoritmos basados en densidades

Los algoritmos basados en densidades agrupan objetos mientras su densidad (número de objetos) en la "vecindad" este dentro de un cierto umbral (parámetro definido por el usuario). Consideran los clústeres como regiones densas de objetos en el espacio, que están separados por regiones de baja densidad (elementos aislados que representan ruido). Es decir, para cada punto en un clúster, debe haber otro punto en el clúster cuya distancia a éste es menor que el umbral. Y debe haber una cantidad suficiente de puntos como para formar un clúster.

Este tipo de métodos es muy útil para filtrar ruido y encontrar clústeres de diversas formas. La mayoría de los métodos de particionamiento, realizan el proceso de clustering en base a la distancia entre dos objetos, esto hace que solo puedan encontrar clústeres esféricos y se les dificulte hallar clústeres de formas diversas.

Algoritmos basados en rejillas

Dividen el espacio en un número finito de celdas que forman una estructura de rejilla en la que se llevan a cabo las operaciones del clustering. (30)

Algunos algoritmos que pertenecen a esta clasificación son: STING (STatistical INformation Grid approach), WaveCluster y CLIQUE.

Algoritmos basados en modelos

Encuentran un modelo matemático para cada clúster que mejor ajuste los datos de ese grupo. Intentan optimizar el ajuste entre el modelo y los datos. Habitualmente, asumen que el espacio de instancias está gobernado por una mezcla de distribuciones de probabilidades. (29) Existen 2 aproximaciones principales: probabilísticos y redes neuronales.

Selección de un algoritmo

Para elegir un algoritmo de clustering adecuado para una determinada aplicación deben considerarse varios factores. Entre ellos el objetivo de la aplicación, la relación esperada entre calidad y velocidad y las características de los datos.

Objetivo de la aplicación: el objetivo de la aplicación a menudo afecta la elección de un algoritmo de clustering.

Elección entre calidad y velocidad: existe siempre un problema al elegir entre velocidad de procesamiento y calidad de los clústeres obtenidos. Un algoritmo adecuado debe cumplir con estas dos características, pero a veces la cantidad de instancias a procesar juega un papel importante en el tiempo de ejecución del algoritmo de clustering. Un algoritmo que produce clústeres de calidad, por lo general es incapaz de manejar grandes cantidades de información. A su vez, cuando se trabaja con grandes volúmenes de datos, se realiza una especie de compresión sobre la información inicial, perdiendo así calidad.

Características de los datos: las características de los datos a los cuales se quiere aplicar clustering, también son un factor importante para la elección del método adecuado. Los tipos de datos de los atributos, dimensionalidad y cantidad de ruido son algunas de las características a considerar.

Para obtener mayor información sobre los temas Data Mining, Agricultura de Precisión o los términos utilizados en este capítulo, consulte los anexos correspondientes a los estados del arte (anexos 1 y 2) y el glosario.

CAPÍTULO 4: Requisitos

Paralelamente a las investigaciones de DM y AP se realizaron reuniones con el cliente que permitieron ir entendiendo las necesidades y características requeridas de la solución a construir.

En pocas palabras, el cliente requiere de un componente, en particular una DLL[8], que permitiera generar automáticamente un conjunto de zonas de manejo homogéneas a partir de datos de una zafra, teniendo en cuenta su variabilidad espacial y un conjunto de variables medidas durante una zafra. Dicho componente debe ser integrado con el sistema que está actualmente construyendo el cliente.

En el siguiente capítulo se expone un resumen de las características que fueron solicitadas y relevadas con el cliente, en relación al componte a desarrollar.

Requisitos

El sistema a desarrollar consiste en una biblioteca que contiene las operaciones necesarias para generar las zonas de manejo mencionadas. Dicha biblioteca se debe comunicar con los usuarios solo por medio de archivos de texto.

Requisitos no funcionales

Si bien no se especificó un tiempo de respuesta para la generación de zonas de manejo, el mismo debe ser razonable y no quitar usabilidad al sistema.

Requisitos funcionales

Archivo de Entrada

El sistema recibe los datos a procesar por medio de un archivo de texto llamado <archivo_entrada>.ZF**,** el mismo contiene información de la chacra a para la cual se van a generar las zonas. El formato de dicho archivo está definido en el *Anexo 3: Formato de Archivos*.

Área a homogenizar

El archivo de entrada que será utilizado para una zafra debe brindar información de la misma zona geográfica y solo de ella. Debe brindar la información de los valores de los factores productivos medidos dentro de la chacra, que son obtenidos cada X metros (siendo X un dato de entrada al sistema).

[8] Biblioteca de enlace dinámico (Dynamik Link Library), que contiene funciones que pueden ser utilizadas desde los programas, y que son cargadas sólo en el momento en que se necesitan.

Factores productivos

La aplicación debe utilizar distintos factores productivos que representan medidas cuantitativas realizadas a la chacra.

No es necesario para calcular las zonas en determinada zafra que existan medidas de cada uno de los factores productivos posibles, el sistema debe poder obtener las zonas de manejo para un subconjunto de factores productivos de todos los posibles a utilizar. Dicho subconjunto es especificado en el archivo de entrada.

Variabilidad Espacial, Correlación

Se debe realizar un análisis estadístico de los factores productivos. Se deben obtener valores máximo, mínimo, medio y desviación estándar para cada variable, y coeficientes de correlación y varianza-covarianza. Dicho resultado se almacena en un archivo de texto. El formato de dicho archivo está definido en el *Anexo 3: Formato de Archivos*.

Este análisis permitirá seleccionar la medida de similitud a usarse para agrupar los puntos geográficos.

Generación de zonas

El sistema debe, por medio de procedimientos sistemáticos, utilizando algún algoritmo de Data Mining, dividir la chacra en un conjunto de zonas homogéneas, de forma que el rendimiento total obtenido sea el mayor rendimiento que se puede obtener para esa chacra en esa zafra. La información de dichas zonas se almacena en el archivo <archivo_salida>.ZF. El formato de dicho archivo está definido en el *Anexo 3: Formato de Archivos*.

Las zonas generadas en la división representarán áreas geográficas dentro de la chacra, las cuales presentan alguna similitud.

Descripción de zonas

Con el conjunto de factores productivos a analizar en una zafra y las zonas generadas, el sistema debe analizar para cada zona la variabilidad espacial de dichos factores. Para esto deberá obtener los valores máximos, mínimos, medio y desviación estándar de las variables. Se debe brindar en un archivo de salida <archivo_salida>.ZAF dicha información. El formato de dicho archivo está definido en el *Anexo 3: Formato de Archivos*.

También, deberán ser calculados los índices de performance y almacenados en el mismo archivo de forma que el usuario pueda validar el resultado de la operación de zonificación.

Herramientas

Se debe utilizar para desarrollar Visual Studio 2005 (31) e implementar en el lenguaje C#.

Restricciones

Las restricciones para la generación de las zonas son las siguientes:

- El número de zonas a generar debe ser mayor que 1.
- El tamaño mínimo de zona debe ser menor que el tamaño de la chacra a procesar.
- El número de zonas no debe ser mayor a un número especificado.
- El área mínima de puntos adyacentes, para cada subgrupo dentro de una zona, debe ser mayor a un tamaño mínimo dado en metros.

CAPÍTULO 5: Análisis de la Solución

En este capítulo se describe el análisis que se realizó para llegar a la solución implementada. En particular, se explica el algoritmo elegido para la implementación, el cual es parte central de la misma, y cómo se seleccionó.

Data Mining Aplicado a la Agricultura de Precisión

El objetivo de la investigación de Data Mining fue conocer, de alguna forma, la variedad de técnicas que existen para obtener conocimiento a partir de los datos. La técnica seleccionada dentro de las técnicas de Data Mining para la aplicación de este Proyecto de Grado, corresponde a clustering. La elección de la misma, se basó principalmente en su adaptación al problema que nos concierne, que es el de generar zonas de manejo. Esta decisión fue acompañada por diversos artículos y documentos que sugerían la técnica de clustering como la más apropiada en este contexto.

En *The Australian Centre for Precision Agriculture* (32) se expresa que "un método muy utilizado para distinguir patrones espaciales y temporales en datos asociados con SSCM (Site Specific Crop Managment) o manejo de sitio específico, está dado por el análisis de clústeres o clustering".

En 1998, Lark definió clasificación como el proceso de reorganizar un conjunto de clases entre un número de individuos. Clustering, es el agrupamiento de objetos similares en distintas clases llamadas clústeres. Tou y Gonzalez (1974) y Hartigan (1975) investigaron los algoritmos que podían ser utilizados para agrupar datos no clasificados.

El análisis de clúster ha sido utilizado para extraer información sobre el terreno, desde imágenes remotas digitales. En aplicaciones de manejo de sitio específico o Agricultura de Precisión, las técnicas de clustering pueden ser utilizadas para identificar regiones de un campo que son similares basados en atributos de la tierra, fertilidad o propiedades físicas del suelo. (33)

Luego, dentro de esta técnica, se debió seleccionar qué tipo de algoritmo sería utilizado, dado que dentro de clustering, existen decenas de algoritmos. En este caso, se optó primeramente por utilizar una técnica de clustering no supervisado, dado que a priori no se conocen las clases; la generación de las mismas debe formar parte del algoritmo. Finalmente, se decidió que el algoritmo a utilizar seria Fuzzy-c-Means. La decisión de utilizar este algoritmo se basó en su adaptación al problema en cuestión y en la investigación de trabajos previos en el área de Agricultura de Precisión y SSCM, donde se expresa que se han obtenido muy buenos resultados con la aplicación del mismo.

Clustering

El análisis de clústeres o clustering es un conjunto de métodos estadísticos que abarca una cantidad de algoritmos diferentes con el fin de agrupar datos de tipos similares en grupos respectivos (clústeres). Su objetivo es crear agrupaciones tal que el

grado de asociación sea alto entre elementos del mismo clúster y bajo entre elementos de diferentes clústeres. (32)

Algoritmos de Clustering Supervisado

Las técnicas de clasificación supervisadas son aquellas en las cuales al momento de realizar la clasificación, están definidas las clases y también el número de clases.

En clasificación supervisada, se combinan el trabajo de campo, mapas, análisis de imágenes espaciales y/o experiencia personal para caracterizar sitios específicos que representan ejemplos homogéneos de los datos no clasificados. Estas áreas son llamadas conjuntos de entrenamiento, dado que la información del suelo o la vegetación es utilizada para entrenar el algoritmo de clasificación para mapear los datos restantes. Luego de calcular los parámetros (ej. media, desviación estándar, matrices de covarianza, matrices de correlación) para cada conjunto de entrenamiento, cada dato tanto dentro del conjunto de entrenamiento como fuera de éste, es evaluado y asignado a la clase que tiene más probabilidad de pertenecer. (33)

Algoritmos de Clustering No Supervisado

Los algoritmos de clustering no supervisado son aquellos en los que al momento de realizar la clasificación no se conocen las propias clases y probablemente tampoco se conozca la cantidad de clases.

A diferencia de las técnicas supervisadas, los algoritmos de clustering no supervisados no requieren que el usuario especifique las medias de las clases y las matrices de covarianza a ser utilizadas en la clasificación. Las técnicas de clasificación no supervisadas producen agrupaciones naturales de los datos. Casi siempre, la clasificación no supervisada se utiliza para ganar entendimiento en la estructura inherente de los datos. (33)

El algoritmo ISODATA (Iterative Self-Organizing Data Analysis Technique) es uno de los algoritmos de clustering no supervisado más utilizados. Este algoritmo calcula la media de las clases y luego agrupa iterativamente el resto de los datos minimizando la distancia Euclideana[9] de cada punto a la media de la clase. Cada iteración resulta en la re calculación de las medias de las clases y la reclasificación de los datos con respecto a la nueva media. Este proceso continúa hasta que se alcanza un número máximo de iteraciones o el número de datos de cada clase cambia en un valor menor a un parámetro especificado. Para caracterizar efectivamente las clases que se obtienen como resultado del algoritmo por vectores de media y una matriz de covarianza, ISODATA requiere que cada variable utilizada presente, a grandes rasgos, una distribución Gaussiana. Además, se obtienen mejores resultados si todos los datos presentan varianzas similares. Estos dos requerimientos pueden resultar en una preparación más compleja del conjunto de datos previa a la clasificación. (33)

[9] Índice cuantitativo que mide la separación existente entre dos unidades de observación según los valores que ellas posean en un conjunto de variables.

Pese a las buenas características de este algoritmo, a priori no se cumplen ninguno de los requerimientos mencionados, por lo cual no fue seleccionado.

A diferencia del algoritmo de clasificación no supervisado ISODATA, el algoritmo k-Means o k-Medias (también conocido como c-means) no requiere que las variables utilizadas en la clasificación contengan varianzas similares o sigan una distribución Gaussiana. El algoritmo k-Means, se basa en minimizar una función objetivo o un índice de performance definido como la suma de los cuadrados de las distancias desde todos los puntos en un clúster al centro de éste. Al igual que ISODATA, el algoritmo k-Means utiliza un proceso iterativo para re calcular el centro del clúster y asignar los datos a éste. El algoritmo finaliza cuando se cumple el criterio especificado de convergencia (por ejemplo, no cambia el centro del clúster). (33)

Clasificación Fuzzy o Clasificación borrosa

El objetivo del clustering es, esencialmente, el de particionar un conjunto de datos de entrada en un número de clústeres homogéneos, con respecto a una medida de similitud. Debido a la naturaleza borrosa de varios problemas prácticos, se han desarrollado algunos métodos de clustering borroso siguiendo la teoría general de conjuntos borrosos (34) delineada por Zadeh. (35)

Zadeh (1965) introdujo la teoría de conjuntos fuzzy o conjuntos borrosos como una generalización de la teoría de conjuntos. A diferencia de la teoría de conjuntos convencional, que permite a un elemento pertenecer a un único conjunto, la teoría de conjuntos fuzzy permite a cada elemento pertenecer parcialmente a un conjunto. (33)

Utilizando esta teoría, Ruspini (1969) introduce el concepto de fuzzy clustering o agrupamiento borroso. La principal diferencia entre el clustering tradicional y el clustering borroso puede ser expresada de la siguiente forma: mientras en el clustering tradicional un elemento pertenece a un único clúster, en el clustering borroso los elementos pueden pertenecer a varios clústeres con diferentes grados de pertenencia.

La aplicación de este tipo de algoritmos ha permitido a los investigadores mejorar las justificaciones sobre la variabilidad continua en fenómenos naturales.

Uno de los algoritmos de agrupamiento borroso más utilizado es Fuzzy c-Means (FCM), el cual fue propuesto inicialmente por Duna y generalizado por Bezdek y otros autores. Es una generalización del algoritmo c-means que aplica la teoría de conjuntos borrosos. Este utiliza un exponente ponderado para controlar el grado de pertenencia que ocurre entre las clases.

Usualmente, las funciones de pertenencia están definidas en base a una función de distancia, de tal forma que el grado de pertenencia expresa la proximidad de un elemento con respecto al centro del clúster. Eligiendo una función de distancia apropiada, se pueden identificar clústeres con diferentes formas.

Según las investigaciones, la clasificación fuzzy ha sido utilizada para clasificar datos del suelo, rendimiento e imágenes remotas. (33)

Medida de similitud

Antes de que una agrupación de datos pueda ser formada, es necesario definir una medida de similitud que establecerá una regla para asignar elementos[10] al dominio de un clúster particular.

La medida de similitud más usada frecuentemente es la distancia desde un punto hasta el centro del clúster. Así, mientras la distancia entre el punto y el centro del clúster decrece, más alta es la similitud entre los dos.

La **distancia Euclideana** (36) es una de las medidas de similitud más utilizada. Ésta se define como:

$$d(v,w) = \sqrt{(v_1 - w_1)^2 + (v_2 - w_2)^2 + \ldots + (v_n - w_n)^2}$$

donde v y w son dos matrices fila (o columna) de dimensión (1 x n) y v_i (w_i) un número real para todo i perteneciente al intervalo [1, n].

Asigna los mismos pesos o ponderaciones a todas las variables y no es sensible a variables correlacionadas[11]. Geométricamente, la distancia Euclideana usualmente genera clústeres con forma esférica, lo que en realidad, ocurre raramente en el ambiente de suelo. (33)

El **método Diagonal** (37) para medir la similitud, fue descrito por McBratney, Moore Odeh. Al igual que la distancia Euclideana, el método Diagonal no es sensible a variables correlacionadas. Sin embargo, compensa las distorsiones en las asumidas formas esféricas de los clústeres ponderando la varianza de las variables. (33)

Otra alternativa es la **distancia de Mahalanobis** (38), la cual tiene en cuenta varianzas desiguales así como también las correlaciones entre las variables. Logra esto incluyendo la matriz de varianza-covarianza como una parte del cálculo de la distancia. (33)

La distancia de Mahalanobis entre dos variables aleatorias con la misma distribución de probabilidad $\vec{x}$ y $\vec{y}$ con matriz de covarianza ∑ se define como:

$$d_m(\vec{x},\vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}$$

La norma o distancia utilizada depende de las diferentes clases de características de los datos: (39)

Distancia Euclideana: se usa cuando las características son estadísticamente independientes, y varían en la misma medida para clústeres con una forma hiperesférica[12].

[10] En el contexto de este documento, consideramos un elemento al dato correspondiente al valor de una variable (ej., rendimiento) en una coordenada.

[11] Dos variables están correlacionadas si los valores de una variable tienden a ser más altos o más bajos para valores más altos o más bajos de la otra variable.

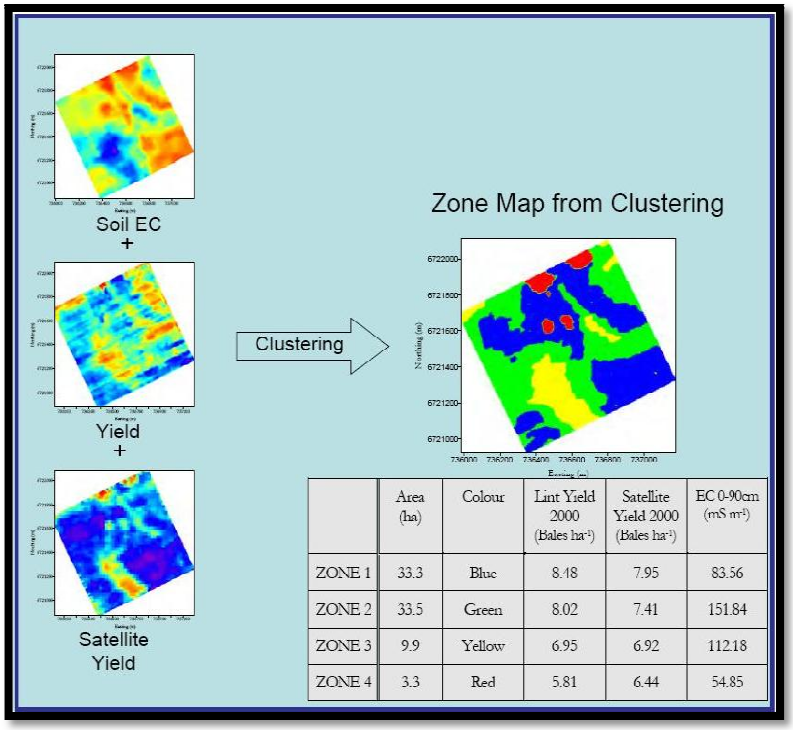
[12] Una hiperesfera es un análogo en mayor número de dimensiones de una esfera.

Distancia Diagonal: se usa cuando las características son estadísticamente independientes, y varían en medidas desiguales para clústeres con una forma hiperelipsoidal.

Distancia de Mahalanobis: se usa cuando las características son estadísticamente dependientes, y varían en medidas desiguales para clústeres de una forma hiperelipsoidal.

Diseño de la Solución

En la Figura 3, se puede ver el esquema de la generación de zonas de manejo de una manera general. Se tiene como entrada los datos de las variables (o factores productivos) y luego de la aplicación del algoritmo se obtiene como salida la división del campo en zonas con sus respectivas características.



	Area (ha)	Colour	Lint Yield 2000 (Bales ha⁻¹)	Satellite Yield 2000 (Bales ha⁻¹)	EC 0-90cm (mS m⁻¹)
ZONE 1	33.3	Blue	8.18	7.95	83.56
ZONE 2	33.5	Green	8.02	7.41	151.84
ZONE 3	9.9	Yellow	6.95	6.92	112.18
ZONE 4	3.3	Red	5.81	6.44	54.85

Figura 3: Esquema de generación de zonas de manejo. Fuente: (40)

Sin embargo, este proceso requiere de algunas etapas o pasos que se presentan en el esquema de la Figura 4.

1) Cálculo de estadísticas descriptivas

El primer paso es el cálculo de estadísticas, las cuales consisten en información necesaria para establecer los parámetros del algoritmo de clasificación no supervisada.

2) Aplicación del algoritmo

Se aplican las funciones y estrategias del algoritmo seleccionado y se realizan test de performance. A pesar de que existen muchos y variados procedimientos para generar zonas de manejo, la literatura generalmente apoya el uso de clasificación no supervisada para este propósito.

3) Validación del método

Por último, se discuten los métodos utilizados para evaluar y validar el algoritmo de clustering no supervisado.

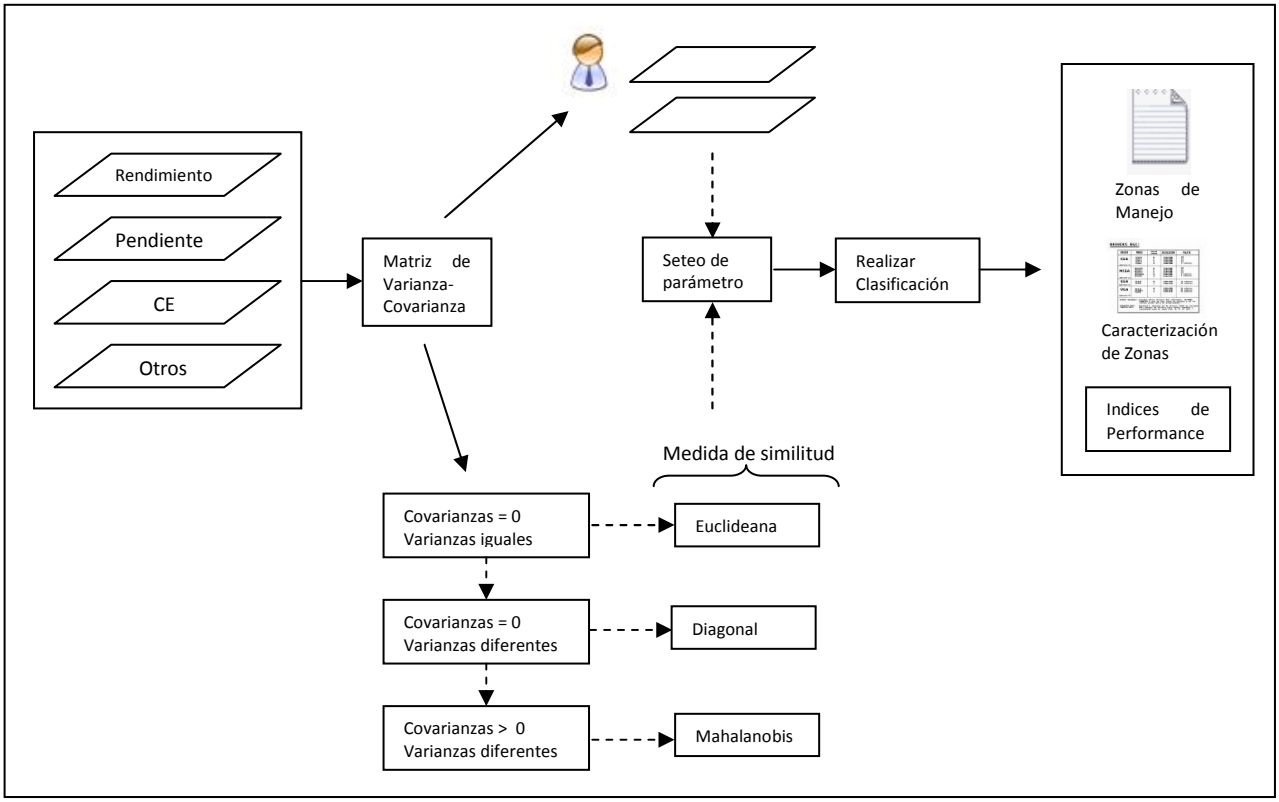


Figura 4: Proceso a seguir para la generación de zonas de manejo.

Cálculo de estadísticas descriptivas

Antes de aplicar el algoritmo seleccionado, se realizará el cálculo de estadísticas descriptivas de los atributos dados.

Estas incluyen:

- valores máximos y mínimos de cada variable
- media
- desviación estándar

- matriz de varianza-covarianza
- matriz de correlación

La medida más común que indica el centro es la **media** (μ), y se define como la suma de todos los valores de una variable dada, dividida la cantidad de valores.

$$\mu_k = \frac{\sum_{i=1}^{n} y_{ik}}{n}$$ Ecuación 1

donde y_{ik} representa el i-ésimo valor de la variable k y n es la cantidad de valores.

La **desviación estándar** se utiliza para medir la cantidad de variabilidad en una variable. La desviación estándar de la k-ésima variable (σ_k) se define como:

$$\sigma_k = \sqrt{\frac{\sum_{i=1}^{n} (y_{ik} - \mu_k)^2}{n-1}}$$ Ecuación 2

Donde el término $\sum_{i=1}^{n} (y_{ik} - \mu_k)^2$ es la suma de los cuadrados.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & & \ddots & \\ \sigma_{p1} & \sigma_{p2} & & \sigma_{pp} \end{bmatrix}$$ Ecuación 3

La **matriz de varianza-covarianza** mostrada como ejemplo en la Ecuación 3, brinda una medida de la variación conjunta entre dos variables. La matriz de covarianza es una matriz cuadrada que es siempre simétrica. Los elementos diagonales de la matriz (σ_{kk}) representan la varianza de y_k, mientras que el resto de los elementos (σ_{jk}) son la covarianza de y_j e y_k. La covarianza se calcula como

$$\sigma_{jk} = \frac{SP_{jk}}{n-1}$$ Ecuación 4

donde SP_{jk} es la suma de los productos calculados utilizando la fórmula

$$SP_{jk} = \sum_{i=1}^{n} (y_{ij} \times y_{ik}) - \frac{\sum_{i=1}^{n} y_{ij} \sum_{i=1}^{n} y_{ik}}{n}$$ Ecuación 5

Si $\sigma_{jk} > 0$, hay dependencia directa (positiva), es decir, a grandes valores de j corresponden grandes valores de k.

Si $\sigma_{jk} = 0$, no existe una relación lineal entre las dos variables.

Si $\sigma_{jk} < 0$, hay dependencia inversa o negativa, es decir, a grandes valores de j corresponden pequeños valores de k.

Para estimar la cantidad de interrelación que existe entre las variables en un método no influenciado por las unidades de medida, se utiliza el **coeficiente de correlación** r. Como se define en (33), el coeficiente de correlación entre dos variables, r_{jk}, es:

$$r_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$$ Ecuación 6

El valor del coeficiente de correlación varía en el intervalo [-1, +1]:

Si r = 0, no existe ninguna correlación, el coeficiente indica una independencia total entre las dos variables, es decir, que la variación de una de ellas no influye en absoluto en el valor que pueda tomar la otra.

Si r = 1, existe una correlación positiva perfecta. El coeficiente indica una dependencia total entre las dos variables denominada relación directa: cuando una de ellas aumenta, la otra también lo hace en idéntica proporción.

Si 0 < r < 1, existe una correlación positiva.

Si r = -1, existe una correlación negativa perfecta. El coeficiente indica una dependencia total entre las dos variables llamada relación inversa: cuando una de ellas aumenta, la otra disminuye en idéntica proporción.

Si -1 < r < 0, existe una correlación negativa.

Estos resultados se muestran gráficamente en la Figura 5.

Figura 5: Tipos de correlación entre variables. Fuente: (41)

Una consideración que debe tenerse en cuenta, es que el coeficiente de correlación que se está considerando es lineal. Por lo tanto representa el grado de asociación lineal entre dos variables. Aunque el grado de correlación sea cercano a cero, eso no significa que no haya relación entre las dos variables. Puede que dicha relación no sea lineal. (41)

Relación entre matriz de varianza-covarianza y matriz de correlación:

Si las n variables tienen medidas incompatibles (kg, m, s, ...) las varianzas no son comparables. Entonces se recurre a la matriz de correlación. La correlación es la covarianza medida para valores estandarizados. Por eso la correlación de una variable consigo misma da uno; es la varianza de cualquier variable estandarizada. (41)

Algoritmo Fuzzy c-Means

El algoritmo de clustering Fuzzy c-Means fue seleccionado con el propósito de particionar el conjunto de datos en grupos de clústeres.

Este algoritmo requiere pocos parámetros de entrada por parte del usuario dado que es un algoritmo de clasificación no supervisada. Como ya lo mencionamos, Rupsini (1969) introdujo la teoría de fuzzy clustering o agrupamiento borroso permitiendo a un elemento compartir su pertenencia entre clases. El espacio de partición c-fuzzy para una matriz de datos Y se define como el siguiente conjunto:

$$M_{fc} = \left\{ U \in V_{cn} \middle| u_{ik} \in [0,1] \forall i,k; \sum_{i=1}^{c} u_{ik} = 1 \forall k; 0 < \sum_{k=1}^{n} u_{ik} < n \forall i \right\} \qquad \text{Ecuación 7}$$

donde

U = una partición fuzzy de c clases fuzzy y un total de n elementos.

V_{cn} = el conjunto de matrices c x n.

c = la cantidad de clústeres ($2 \leq c < n$).

u_{ik} = el valor de pertenencia del elemento y_k en la i-ésima clase.

El algoritmo Fuzzy c-Means se basa en la función objetivo J_m, la cual corresponde a la suma de los cuadrados y es una medida ponderada de la distancia cuadrada entre los elementos y el centro del clúster.

$$J_m(U,v) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m (d_{ik})^2 \qquad \text{Ecuación 8}$$

donde

U = c-partición fuzzy de Y ($U \in M_{fc}$).

V = vector con los centroides de los clústeres ($v_1, v_2, \ldots, v_c$) de tamaño c x p, donde p representa la cantidad de variables utilizadas en la clasificación.

m = exponente fuzzy ($1 \leq m < \infty$).

$(d_{ik})^2$ = distancia entre y_k y v_i, calculada como:

$$(d_{ik})^2 = \|y_k - v_i\|^2 = (y_k - v_i)' A (y_k - v_i) \qquad \text{Ecuación 9}$$

v_i = centroide del clúster i.

A = matriz ponderada de tamaño p x p (induce la norma)

El exponente fuzzy (m) controla la medida de pertenencia entre las clases. A medida que el valor de m crece hacia infinito, la medida de pertenencia aumenta y las clases resultantes resultan menos distintas. Los clústeres "duros" (por ejemplo, aquellos que no comparten la pertenencia) ocurren cuando el valor de m se aproxima a uno.

La matriz ponderada (A) define la norma inducida del producto. Una norma representa la distancia entre dos puntos en un vector lineal del espacio. En este caso, la matriz ponderada puede tener tres diferentes formas dependiendo de la covarianza de los datos. Se proveerá la información necesaria para seleccionar la norma apropiada en el proceso de cálculo de estadísticas descriptivas previo a la aplicación del algoritmo.

La **norma Euclideana** es utilizada para variables estadísticas independientes que presentan el mismo valor para la varianza. Bajo la norma Euclideana, la matriz ponderada (denotada A_E) toma la forma de una matriz identidad de tamaño p x p. La norma Euclideana se define como:

$$(d_{ik})^2 = (y_k - v_i)' A_E (y_k - v_i) \qquad \text{Ecuación 10}$$

Para variables estadísticamente independientes que tienen diferentes varianzas, se utiliza la **norma Diagonal**. Esta es inducida por la matriz diagonal de tamaño p x p:

$$A_D = \begin{bmatrix} (1/\sigma_1)^2 & 0 & \cdots & 0 \\ 0 & (1/\sigma_2)^2 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & (1/\sigma_p)^2 \end{bmatrix} \qquad \text{Ecuación 11}$$

donde σ_p es la desviación estándar de la variable p utilizada en la clasificación. Por lo tanto, la norma diagonal es

$$d^2{}_{ik} = (y_k - v_i)' A_D (y_k - v_i) \qquad \text{Ecuación 12}$$

Una tercera alternativa es la **distancia de Mahalanobis**, la cual se define como:

$$d^2{}_{ik} = (y_k - v_i)' \sum^{-1} (y_k - v_i) \qquad \text{Ecuación 13}$$

donde $\sum^{-1}$ es la inversa de la matriz p x p de varianza-covarianza de Y. La distancia de mahalanobis se considera para las variables estadísticamente independientes con varianzas diferentes.

La minimización de J_m (Ecuación 7) está acompañada por un cálculo iterativo de los valores de pertenencia al clúster (Ecuación 14) y el centroide del clúster (Ecuación 15):

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right]^{-1}$$ Ecuación 14

Y

$$v_i = \sum_{k=1}^{n} (u_{ik})^m y_k \Big/ \sum_{k=1}^{n} (u_{ik})^m , 1 \le i \le c.$$ Ecuación 15

A pesar de que el algoritmo Fuzzy c-Means se basa en la minimización de la función objetivo J_m (Ecuación 8), el mínimo de J_m no es aceptable para la validación estadística de los clústeres sugeridos por el algoritmo. Para determinar el número apropiado de clústeres, se calculan dos tipos de funciones de validación de clústeres para cada partición fuzzy de Y producida por el algoritmo de clustering Fuzzy c-Means.

El **Índice de Performance Fuzzy** (FPI, Fuzzy Performance Index) es una medida del grado de separación (por ejemplo, borrosidad) entre las c particiones fuzzy de Y, y se define como:

$$FPI = 1 - \frac{c}{(c-1)} [1 - F(U;c)]$$ Ecuación 16

donde F es la función

$$F(U;c) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^2 \Big/ n$$ Ecuación 17

Los valores de FPI varían de cero a uno. Valores cercanos a cero indican distintas clases con una baja pertenencia, mientras que valores cercanos a uno indican clases similares con alto grado de pertenencia.

Bezdek (1981) describe una segunda medida de validación de clústeres conocida como **Entropía de Clasificación Normalizada** (NCE, Normalizad Classification Entropy). Esta medida representa el grado de desorganización de una partición fuzzy de Y. La entropía de clasificación (H) está definida por la función:

$$H(U;c) = -\sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log_a (u_{ik}) \Big/ n$$ Ecuación 18

donde la base a del logaritmo es cualquier entero positivo. Los valores de H varían entre 0 y $\log_a(c)$. Adicionalmente, Bezdek (1981) reportó que los valores finales del rango de H no representan apropiadamente el grado de desorganización presente (por ejemplo, cuando c = 1, H = 0 o c = n, H = 0). Para remediar esto, fue sugerido que H se normalizara:

$$NCE = H(U;c) / [1 - (c/n)]$$ Ecuación 19

Los valores de NCE serán similares a los valores de H cuando c es relativamente pequeño comparado con n (por ejemplo, (c/n) tiende a 0). Sin embargo, en situaciones

donde (c/n) es grande (por ejemplo, cercano a 1), NCE producirá resultados bastante diferentes.

A continuación se presenta el seudo-código del algoritmo Fuzzy c-Means:

1. Elegir el número de clústeres c, con $2 \leq c < n$.
2. Elegir el exponente fuzzy m, con $1 \leq m < \infty$.
3. Elegir una norma apropiada para la medida de distancia d_{ik}^2.
4. Elegir un valor para el criterio de parada ε (ε = 0.0001 asegura una buena convergencia).
5. Elegir un valor para la máxima cantidad de iteraciones l_{max}.
6. Inicializar la matriz de pertenencia $U^{(0)}$ ϵ M_{fc} con valores aleatorios de pertenencia entre 0 y 1.
7. Para cada iteración l = 1, 2, 3, …, calcular los c centroides de los clústeres $\{v_i^l\}$ utilizando la Ecuación 15y $U^{(l-1)}$.
8. Re calcular $U^{(l)}$ utilizando la Ecuación 14 y $\{v_i^l\}$.
9. Parar cuando $l = l_{max}$ o cuando $||U^{(l)} - U^{(l-1)}|| \leq \varepsilon$, sino volver al paso 7.
10. Calcular las funciones de validación de clúster (FPI y NCE) para la partición c fuzzy de Y.

Post Procesamiento

Para cumplir con el requerimiento especificado de que las sub-zonas generadas deben cumplir que su tamaño sea mayor que un número dado, se decidió aplicar un post-procesamiento de las sub-zonas obtenidas luego de aplicar el algoritmo, ya que el mismo no cuenta con el manejo de este requerimiento.

El post-procesamiento de las zonas obtenidas es el siguiente:

1. Obtener sub-zonas cuyo tamaño es menor que el tamaño mínimo.
2. Mientras existan sub-zonas cuyo tamaño es menor que el tamaño mínimo
    a. Obtener sub-zona de menor tamaño.
    b. Asignar sub-zona a la zona con mayor perímetro circundante
    c. Volver a 2.

Una zona, corresponde a un número de zona y ciertas características que se presentan en todos los datos que tienen asignada dicha zona.

Una sub-zona, es un área geográfica que tiene un valor de zona asignada.

Gráficamente, en la Figura 6 se muestra una imagen en donde se generaron 3 zonas de manejo que se distribuyen en 5 sub-zonas geográficas.

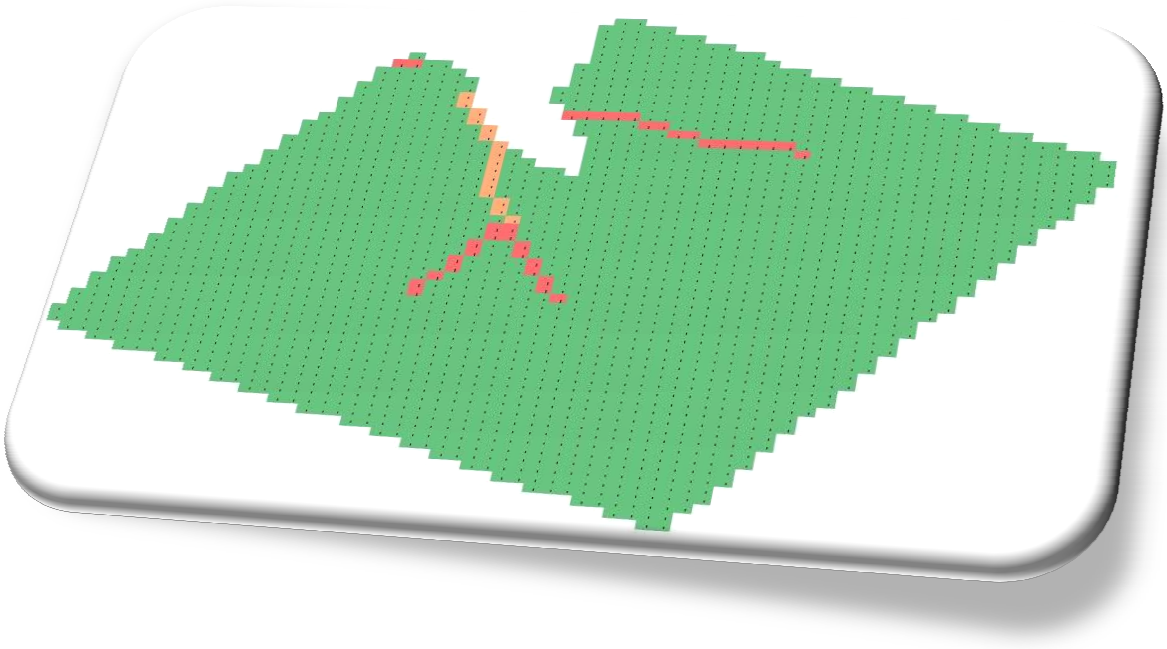


Figura 6: Generación de zonas de manejo.

Es importante mencionar, que como resultado de este post-procesamiento la cantidad de zonas generadas puede resultar menor al número pasado como parámetro. Es por esto, que se decidió, en conjunto con el cliente, generar 2 archivos de salida .ZF: el que contiene las zonas generadas sin aplicar post-procesamiento y el que contiene las zonas generadas luego de aplicar el post-procesamiento. El nombre de este último archivo es <archivo_salida>_post.ZF.

Observar y comparar ambos resultados puede resultar de mucha ayuda al productor a la hora de decidir en cuantas zonas dividir el terreno.

CAPÍTULO 6: Diseño e Implementación

En este capítulo se describirá la solución propuesta mostrando sus características generales, arquitectura, componentes principales y su funcionamiento. También se describen las principales decisiones de implementación tomadas.

Introducción

El sistema a desarrollar consiste de una biblioteca o DLL conteniendo distintas clases interconectadas que brindarán una solución integral al mencionado problema de generación de zonas de manejo.

Una biblioteca es un conjunto de procedimientos y funciones (subprogramas) agrupadas en un archivo con el fin de que puedan ser aprovechadas por otros programas.

Dicha biblioteca interactuará con los usuarios o agentes externos únicamente a través de archivos de texto y no contendrá interfaz de usuario para recibir información de los mismos. Esto se debe a que el propósito es que la biblioteca forme parte de una aplicación de mayor tamaño la cual invoca sus servicios mediante llamadas a funciones con parámetros.

Ambiente de Desarrollo

La implementación se realizó en el lenguaje C# utilizando el entorno de desarrollo MS Visual Studio 2005 (31) sobre el sistema operativo Windows XP. El mismo fue de gran utilidad y de trascendente importancia para el éxito del proyecto, ya que permitió automatizar los ambientes de prueba y depurar con facilidad la aplicación desarrollada.

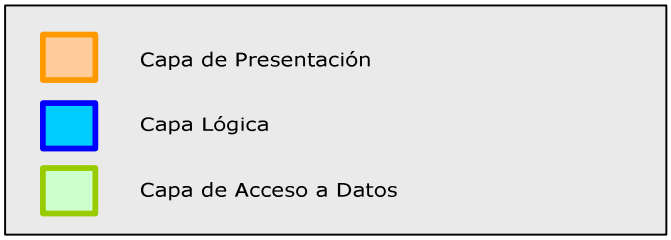
Para facilitar el manejo de archivos por parte de los estudiantes, se instaló y configuró la herramienta de control de versiones[13] SVN (Subversion) (42) y se utilizó el cliente Windows TortoiseSVN (43).

Arquitectura Propuesta

Para la implementación de la biblioteca, se decidió utilizar una arquitectura en capas. La misma consiste en 3 capas, donde se delimitan perfectamente la interface, los procesos y el acceso al repositorio de datos.

En la Figura 7 se muestra un diagrama de la arquitectura propuesta.

[13] Los sistemas de control de versiones vigilan las diferentes versiones de un archivo fuente.



Criterios utilizados en los diagramas de arquitectura.

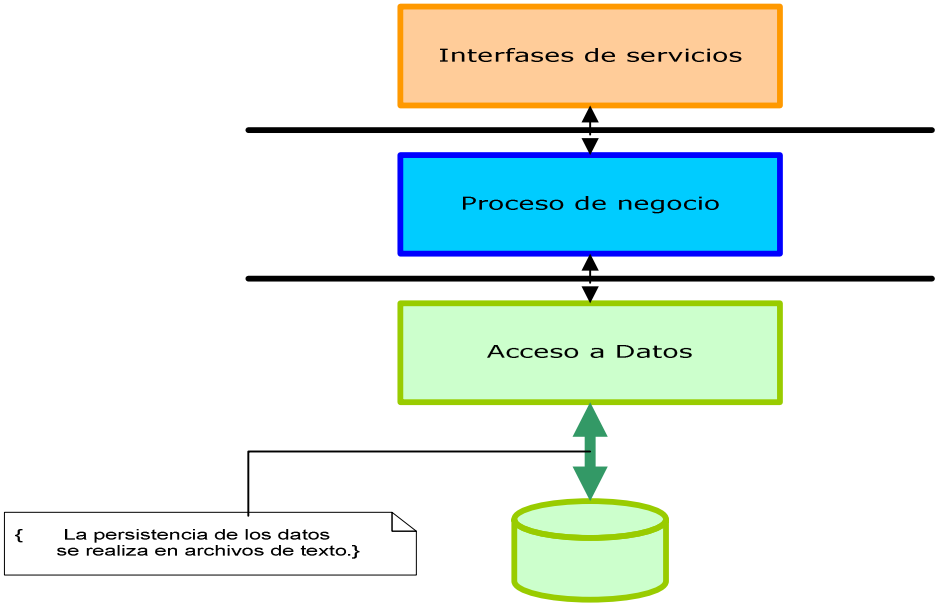


Figura 7: Arquitectura de la solución.

Capa de Presentación

Esta capa es la que utilizarán quienes quieran invocar las operaciones disponibles de la librería. Contiene las clases que se encargan de capturar las solicitudes. Para resolver dichas solicitudes se comunican con la capa lógica y los servicios que ella expone. A la capa de presentación de la solución se accederá por medio de operaciones definidas, incluyendo antes la librería en el proyecto donde quiera ser utilizada.

Capa Lógica

Esta capa tiene como principal cometido la implementación de la lógica del sistema. Allí se resolverán los servicios que serán consumidos por la capa de presentación.

Capa de Acceso a Datos

Esta capa contiene la lógica necesaria para el acceso a los datos. Expondrá los servicios que serán consumidos por la capa de lógica. Está capa obtendrá y persistirá la información en archivos de texto especificados.

Casos de Uso

El sistema o biblioteca a construir brinda dos funcionalidades, las cuales se describen a continuación.

Estas funcionalidades pueden verse también como los casos de uso[14] (44) relevantes del sistema.

Cálculo de estadísticas

Consiste en el cálculo de estadísticas que brindarán al usuario la información necesaria para seleccionar las variables y (opcionalmente) la distancia a utilizar en la generación de zonas de manejo.

Operación	*CalcularEstadísticas*
Entrada	- ruta absoluta al archivo de entrada (.ZF) - ruta absoluta al archivo de salida
Salida	Archivo de texto con los siguientes datos: - matriz de correlación - matriz de varianza-covarianza - máximo (para cada variable) - mínimo (para cada variable) - desviación estándar (para cada variable) - media (para cada variable)

Generación de zonas de manejo

Esta operación consiste en la generación de las zonas de manejo.

Operación	*GenerarZonas*
Entrada	- ruta absoluta al archivo de entrada - ruta absoluta al archivo de salida (solo el nombre, sin extensión)[15]

[14] En ingeniería del software, un caso de uso es una técnica para la captura de requisitos potenciales de un nuevo sistema o una actualización software. Cada caso de uso proporciona uno o más escenarios que indican cómo debería interactuar el sistema con el usuario o con otro sistema para conseguir un objetivo específico.

[15] Los dos archivos de salida generados tienen este nombre y diferente extensión.

	- cantidad de zonas a generar (> 1) - tamaño mínimo de zona (en m^2) - distancia a utilizar (opcional)
Salida	- archivo de texto con la especificación de las zonas (<archivo_salida>.ZF) - archivo de texto con la especificación de las zonas luego del post-procesamiento (<archivo_salida>_POST.ZF) - archivo de texto con la descripción de las zonas generadas (<archivo_salida>.ZAF) - archivo de texto con la descripción de las zonas generadas luego del post-procesamiento (<archivo_salida>_POST.ZAF)

En caso de que la distancia no sea pasada como parámetro o el valor pasado no sea válido (debe ser 'E', 'D' o 'M'), ésta será calculada en base al criterio definido en el diseño de la solución.

Diseño de módulos

En esta sección se presentan los módulos o componentes del sistema. En la Figura 8 se muestra la interacción entre los mismos. A continuación se describe cada uno de ellos y se especifican sus operaciones principales.

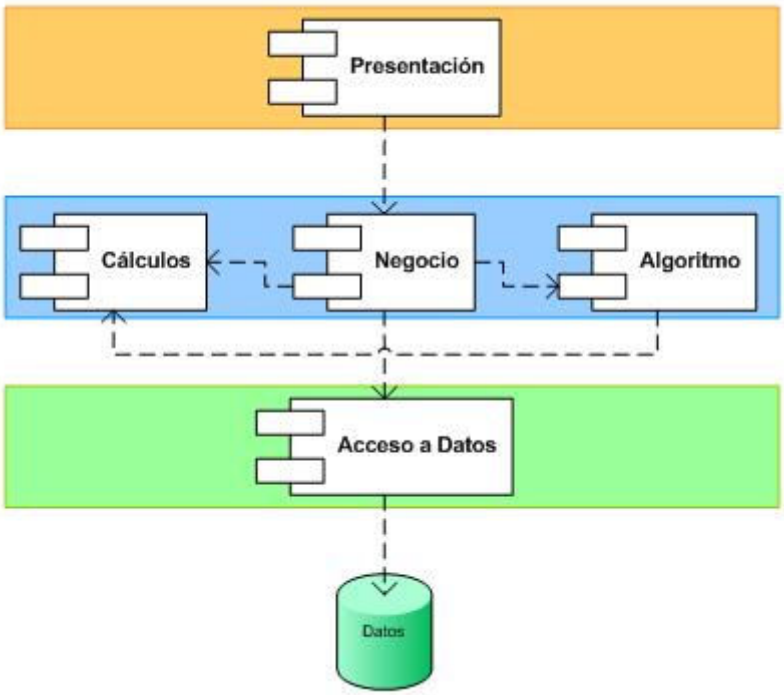


Figura 8: Interacción entre módulos.

Módulo Presentación

Este módulo contiene únicamente las funciones que va a brindar el sistema. Es decir, se definen las funciones de la biblioteca que pueden ser invocadas por un agente externo.

CalcularEstadísticas

Se corresponde al caso de uso Cálculo de Estadísticas.

GenerarZonas

Se corresponde al caso de uso Generar Zonas de Manejo.

Módulo Negocio

Este módulo contiene la implementación de las principales funciones del sistema. Su objetivo es independizarlo de la capa presentación y que interactúe directamente con módulos en el mismo nivel y/o en el nivel inferior.

CalcularEstadisticas(In: archivo_entrada, archivo_salida)

Realiza las invocaciones necesarias para obtener los datos a mostrar como resultado de la invocación de esta operación y genera el archivo de salida correspondiente.

GenerarZonas(In: archivo_entrada, archivo_salida, cant_zonas, tam_min_zona, distancia)

Este procedimiento compone la estructura global de la generación de zonas. Se encarga de generar las zonas invocando las distintas operaciones que se encuentran en los módulos que utiliza. También realiza el post-procesamiento de las zonas generadas. Por último, se comunica con el módulo de Acceso a Datos para generar los archivos de salida correspondientes.

PostProcesamiento(In: zonas, cant_zonas, tam_min_zona)

Este procedimiento realiza el post-procesamiento de las zonas obtenidas de forma de reasignar los elementos de una sub-zona que no cumple con el tamaño mínimo a la zona con el mayor borde circundante. Para ello, primero obtiene la sub-zona de menor tamaño que no cumple la restricción, la reasigna y vuelve a obtener las sub-zonas de tamaño menor que el especificado. Itera hasta que se cumpla que no existan sub-zonas cuyo tamaño sea menor que tam_min_zona. El resultado es la actualización de la matriz de zonas, la cual especifica a qué zona pertenece cada punto o coordenada.

Módulo Cálculos

En este módulo se implementan todas las funciones correspondientes a cálculos. A continuación se especifican las más importantes.

CalcularMatrizVarianzaCovarianza(In: v_1, …, v_p; Out: matriz)

Dados los datos correspondientes a las variables, esta operación devuelve la matriz de varianza-covarianza.

CalcularMatrizCorrelación(In: v_1, …, v_p; Out: matriz)

Dados los datos correspondientes a las variables, esta operación devuelve la matriz de correlación de dichas variables.

ObtenerSubZonasMenorTamMin(In: zonas, tam_min_zona; Out: list<sub-zona>)

Este procedimiento recibe la matriz de zonas y genera una lista de sub-zonas cuyo tamaño es menor al tamaño mínimo de zona especificado.

ObtenerNorma(In: matriz_var_cov; Out: norma)

Dada la matriz de varianza covarianza, devuelve el tipo de distancia a utilizar (Euclideana, Diagonal o Mahalanobis).

El cálculo se realiza de la siguiente forma:

- covarianzas = 0 y varianzas iguales -> Euclideana
- covarianzas = 0 y varianzas diferentes -> Diagonal
- covarianzas > 0 y varianzas diferentes -> Mahalanobis

Como decisión de implementación se considera que covarianzas = 0 cuando el 80% de los valores están entre -0,1 y 0,1. Para el caso de las varianzas, se considera que son iguales, si el 80% de los valores no difieren en un número mayor a 0,001.

TieneDatos(In: índice, v_1, …, v_p; Out: bool)

Dado un índice correspondiente a una coordenada, devuelve true si todas las variables tienen asignado un valor distinto de NaN (valor nulo) para dicha coordenada.

Módulo Algoritmo

El objetivo de este módulo es la implementación del algoritmo Fuzzy c-Means para la generación de zonas de manejo.

FuzzyCMeans(In: cant_zonas, distancia)

Esta operación realiza la implementación del algoritmo Fuzzy c-Means. Recibe como parámetro de entrada la cantidad de zonas a generar y la distancia a utilizar. En caso de que la distancia sea un valor distinto de ‘E’, ‘D’ o ‘M’, está será calculada.

CalcularIndices(In: U, c;ant_zonas Out: fpi, nce)

Esta operación realiza el cálculo de los índices de performance. Recibe como parámetros de entrada la última matriz de pertenencia generada por el algoritmo, y la cantidad de zonas a generar.

Módulo Acceso a Datos

En este módulo se implementarán todas las funciones correspondientes al acceso a datos, ya sea para la lectura de los datos de entrada, como para la generación de archivos de salida.

La estructura de todos los archivos involucrados está definida en el *Anexo 3: Formato de Archivos*.

A continuación se describen las operaciones principales de este módulo.

GenerarArchivoCalculoEstadisticas

Dados los datos correspondientes a las características de las variables: media, máximo, mínimo, desviación estándar, coeficiente de variación, matriz de varianza-covarianza y matriz de correlación, genera el archivo de texto correspondiente al Cálculo de Estadísticas almacenando dicha información.

GenerarArchivoZonas

Dadas las coordenadas con su correspondiente zona, genera el archivo de texto que contiene el resultado de la operación *GenerarZonas*.

GenerarArchivoDescripcionZonas

Dados los datos de las zonas de manejo generadas por el algoritmo, y las características (media, máximo, mínimo y desviación estándar) de las variables utilizadas, genera el archivo de texto correspondiente a la descripción de las zonas de manejo almacenando dicha información.

CargarDatos(In: archivo_datos_entrada)

Dado el nombre del archivo de texto que contiene los datos de entrada, se encarga de almacenar los mismos en las estructuras internas del sistema.

Características de la Solución

Valor Nulo

La constante definida para cuando no existen datos para un punto o variable es “NaN”. En la aplicación esta constante se mapea con el valor NaN que existe dentro de la clase Double de C# (`double.NaN`).

Configuración

La aplicación utiliza ciertos parámetros, los cuales pueden ser modificados fácilmente cambiando los valores de los mismos en el archivo *app.config*[16].

Dado que estos parámetros no se cambian frecuentemente, se decidió ponerlos en un archivo de configuración y no que sean pasados como parámetro a las funciones.

[16] Application Configuration File.

Parámetro	Descripción	Valor asignado
MAX_ITER	Máxima cantidad de iteraciones del algoritmo	300
EXP_FUZZY	Exponente fuzzy del algoritmo	1,3
STOP	Criterio de parada del algoritmo	0,0001

Manejo de errores

La aplicación realiza manejo de errores logueando los mismos en un archivo llamado *zonas.log*. La ubicación de este archivo puede modificarse cambiando la ruta asignada al parámetro **archivo_log** en el archivo *app.config.*

Decisiones tomadas

Durante la etapa de implementación surgieron algunos cambios a los requerimientos definidos en etapas anteriores del proyecto.

Generación de 2 archivos de salida

En la especificación de los requerimientos, se definió un único archivo de salida .ZF con la especificación de las zonas. Durante la etapa de implementación, primero se implementó la generación de zonas sin aplicar el post-procesamiento. Luego, se implementó la parte del post-procesamiento y fue ahí donde se decidió mantener ambas salidas (antes y después) para poder comparar los resultados y verificar que la reasignación de zonas se estaba realizando correctamente.

De la misma forma, la aplicación genera dos archivos de salida .ZAF con la caracterización de zonas antes y después del post-procesamiento.

Al validarlo con el cliente, se decidió que esto podía resultar muy útil y se realizaron los cambios necesarios a la aplicación para generar los 2 archivos de salida.

Datos

Durante la verificación de le aplicación, ocurrió que para los datos dados, existían puntos que no tenían valores asignados para todas las variables. Por ejemplo, la coordenada (x,y) tenía un valor asignado para las variables 1, 2 y 3, pero para la variable 4 tenía el valor NaN.

La asignación de valores a dichos puntos se hace mediante interpolación. No se mide en cada punto el valor de la variable, sino en algunos y luego se interpolan los valores para los puntos intermedios. Sin embargo, esto no está del todo resuelto desde el lado del cliente.

Surgió entonces la duda de que debía hacerse con dichos puntos, es decir, si debían asignarse a una zona o no. Se puede deducir que los cálculos realizados sobre estos datos no iban a ser “coherentes” con otros que si tenían valores asignados para todas las variables. En conjunto con el cliente, se decidió que la aplicación no los considerara. Es decir, solo se le asigna una zona a las coordenadas que tienen valores distintos de NaN para todas las variables que se están considerando.

Está resolución puede “achicar” la zona dada, ya que en estos casos hay puntos que se pasan como datos de entrada los cuales no son asignados a ninguna zona.

Distancia utilizada por el algoritmo

Como parte de la implementación del algoritmo, en cada iteración se obtiene una matriz de distancias cuyo cálculo depende del tipo de distancia utilizada. En el caso de Mahalanobis, para obtener la matriz de distancias se debe utilizar la matriz inversa de la matriz de varianza-covarianza. Dado que puede ocurrir que no exista la inversa, en este caso la aplicación devuelve una excepción indicando en el archivo de log la causa correspondiente.

Plan preliminar de pruebas

En esta sección se establece el plan de testing para el sistema de generación de zonas homogéneas. Describe los procedimientos y lineamientos que se considerarán para efectuar el testing del sistema.

Las actividades de testing tienen como objetivo principal lograr que el producto a entregar al cliente se ajuste a las normativas del mismo.

Verificación

En Ingeniería de Software, la verificación es la confirmación de que se han cumplido los requisitos especificados, o dicho de otra forma, la comprobación de que se está construyendo el producto correctamente.

Alcance

A continuación se detallan las características especificadas que serán testeadas.

- Cálculo de estadísticas
- Generación de zonas homogéneas
- Post-procesamiento de las zonas generadas
- Índices de performance
- Caracterización de zonas

Estrategia

A continuación se definen los diferentes enfoques de testing que se tendrán en cuenta para la realización del mismo. Estos se utilizarán para testear las características definidas en el alcance.

Testing de integración

Se realizará testing de integración, para verificar la correcta integración de la aplicación con el sistema que se encuentra desarrollando el cliente y el resto de los proyectos de grado involucrados.

Testing del sistema

Es aquel llevado a cabo sobre el sistema ya completo para probar que cumple con las funciones de acuerdo a los requerimientos.

Testing de desempeño

Se realizarán pruebas de tiempo de respuesta de la aplicación para verificar que el desempeño del mismo no le quite usabilidad.

Características a no ser testeadas

No será materia del alcance del testing probar la escalabilidad y usabilidad del sistema (requisitos no funcionales.)

Validación

En Ingeniería de Software, la validación se refiere a la adecuación al uso del resultado del diseño y desarrollo, o dicho de otra forma, la comprobación de que se está construyendo el producto correcto.

En una primera instancia se entregará al cliente el sistema para que realice testing y valide así la aplicación desarrollada.

Recursos requeridos

Se lista a continuación las características de software y hardware donde se testeará el sistema, el cliente debería disponer de requerimientos similares o superiores para validar correctamente la usabilidad y el comportamiento del mismo.

Software

- Windows XP Service Pack 2
- Microsoft .NET Framework 2.0

Hardware

- Doble Procesador 1.86 GHz, 1 GB de RAM.

CAPÍTULO 7: Verificación

En este capítulo se exponen las pruebas realizadas y los resultados obtenidos para la verificación de la aplicación. También se presenta un análisis de los datos utilizados para las pruebas.

Datos de prueba

Para probar la aplicación, el cliente nos brindó archivos con datos correspondientes a 2 potreros, a los cuales de aquí en adelante se hará referencia como "Potrero 1" y "Potrero 6". Ambos contienen datos de 30 variables. El Potrero 1 tiene 411.800m^2 y el Potrero 6 tiene 396.000m^2.

En el *Anexo 5: Datos de prueba* se presentan gráficamente las variables utilizadas. La visualización gráfica se logró utilizando la herramienta Excel 2007. Para cada una de ellas se definió una escala con el color asignado según el valor de la variable. La escala definida se muestra en la Figura 9, donde el color verde representa valores altos mientras que el rojo representa valores bajos de la variable.

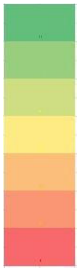


Figura 9: Escala de colores definida para las variables.

Pruebas realizadas

Pruebas de performance

Se realizaron pruebas de performance del sistema, para verificar que el tiempo de ejecución de la aplicación fuera razonable de forma de no afectar su usabilidad. Las mismas se realizaron variando la cantidad de zonas a generar y la distancia utilizada, comprobando que el tiempo de ejecución no superara los 10 minutos en ninguno de los casos de prueba (con los requerimientos de hardware especificados - *Ver: Recursos requeridos*).

Los parámetros utilizados para realizar las pruebas fueron:

Máxima cantidad de iteraciones: 300

Exponente Fuzzy: 1,3

Tamaño mínimo de zona: 10.000 m^2 = 1 Ha

Criterio de parada del algoritmo: 0,0001

El siguiente esquema muestra el tiempo de ejecución para cada uno de los potreros y cada una de las distancias posibles. La cantidad de zonas a generar se varió entre 2 y 6. No consideramos necesario realizar pruebas para una mayor cantidad, dado que 6 ya es un número bastante grande, para zonas homogéneas distintas dentro de una chacra uruguaya (por razones de costos de producción y tamaño de los campos):

Potrero 1

Distancia Euclideana

Cantidad de zonas	Nº iteraciones	Tiempo de ejecución
2	25	00’39
3	59	02’10
4	42	02’02
5	54	03’13
6	67	04’46

Distancia Diagonal

Cantidad de zonas	Nº iteraciones	Tiempo de ejecución
2	31	00’53
3	178	07’21
4	115	06’30
5	55	03’46
6	90	07’24

Distancia de Mahalanobis

Cantidad de zonas	Nº iteraciones	Tiempo de ejecución
2	30	00’59
3	138	06’09
4	129	07’38
5	108	08’06
6	63	05’41

Potrero 6

Distancia Euclideana

Cantidad de zonas	Nº iteraciones	Tiempo de ejecución
2	22	00’31
3	16	00’33
4	16	00’44
5	47	02’36
6	31	02’03

Distancia Diagonal

Cantidad de zonas	Nº iteraciones	Tiempo de ejecución
2	32	00’53
3	28	01’09
4	75	04’05
5	39	02’38
6	74	06’03

Distancia de Mahalanobis

Cantidad de zonas	Nº iteraciones	Tiempo de ejecución
2		
3		
4		
5		
6		

Figura 10: Resultados de las pruebas de performance.

Se puede ver que el tiempo de ejecución no supera en ninguno de los casos los 10 minutos, verificando así que la performance del sistema es la esperada para los casos testeados.

No se realizaron para el Potrero 6, las pruebas con la distancia de Mahalanobis, dado que dicha distancia necesita utilizar la inversa de la matriz de varianza-covarianza, y para los datos de las variables dadas para dicho potrero no existe la inversa de la matriz.

Pruebas de sistema

Se realizaron pruebas del sistema, variando la cantidad de zonas a generar y la distancia utilizada, para verificar que las zonas generadas y el cálculo de estadísticas fueran correctos.

A continuación se presenta el análisis que se realizó utilizando los datos del Potrero 1. El análisis correspondiente al Potrero 6 se encuentra en el *Anexo 6: Análisis Potrero 6*.

Potrero 1

En esta sección se presenta un análisis de las pruebas realizadas para el Potrero 1, el cual pretende demostrar que los resultados obtenidos son coherentes con los datos de entrada.

Las variables utilizadas son:

Nº	Variable	Descripción
1	a_curvature(Banda 1)	Curvatura
2	a_dem_corr(Banda 1)	Elevación
3	a_orientac(Banda 1)	Orientación
4	a_pendiente(Banda 1)	Pendiente
5	a_plan(Banda 1)	Plano de curvatura
6	a_profile(Banda 1)	Perfil de curvatura
7	a_sca(Banda 1)	Área específica de cuenca
8	a_spi(Banda 1)	Índice topográfico SPI
9	a_stci(Banda 1)	Índice topográfico STCI
10	a_wi(Banda 1)	Índice topográfico WI
11	a_basin(Banda 1)	Cuenca
12	trigo05rend(Banda 1)	Rendimiento normalizado 2006 trigo
13	trigo05hum(Banda 1)	Humedad normalizada 2006 trigo
14	maizrend(Banda 1)	Rendimiento normalizado 2007 maíz
15	maizhum(Banda 1)	Humedad normalizada 2007 maíz
16	ec30(Banda 1)	CE a 30 cm
17	ec30a90(Banda 1)	CE de 30 a 90 cm
18	ec90(Banda 1)	CE a 90 cm
19	ec9030(Banda 1)	CE relación 90-30
20	20060127_reflect.img - Layer_2(Banda 1)	Reflectancia banda 2
21	20060127_reflect.img - Layer_3(Banda 1)	Reflectancia banda 3
22	20060127_reflect.img - Layer_4(Banda 1)	Reflectancia banda 4
23	20060127_reflect_tc_1a3.img - 24Layer_1(Banda 1)	Tasseled cap banda 1
24	20060127_reflect_tc_1a3.img - Layer_2(Banda 1)	Tasseled cap banda 2

25	20060127_reflect_tc_1a3.img - Layer_3(Banda 1)	Tasseled cap banda 3
26	20060127_ndvi_rec.img(Banda 1)	Índice NDVI
27	20060127_gndvi_rec.img(Banda 1)	Índice GNDVI
28	raster_cone1(Banda 1)	Índice CONEAT
29	drenaje_Clip.img(Banda 1)	Drenaje
30	pbray(Banda 1)	Muestreo de PBray (set) 2007

Figura 11: Variables Potrero 1

Primero se realizó el cálculo de estadísticas.

La matriz de correlación permite estimar la cantidad de correlación que existe entre las variables en un método no influenciado por las unidades de medida. El coeficiente de correlación varía en el intervalo [-1, +1]:

Si r = 0, no existe ninguna correlación.

Si r = 1, existe una correlación positiva perfecta.

Si 0 < r < 1, existe una correlación positiva.

Si r = -1, existe una correlación negativa perfecta.

Si -1 < r < 0, existe una correlación negativa.

Mientras más cercano a uno sea el valor del coeficiente de correlación, más fuerte será la asociación lineal entre las dos variables. Mientras más cercano a cero sea el coeficiente de correlación indicará que más débil es la asociación entre ambas.

En la Figura 12 se muestra la matriz de correlación coloreada según el siguiente criterio:

	1 < r < 0,7	Existe correlación positiva
	-1 < r < -0,7	Existe correlación negativa

VAR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	1	0,23	-0,06	-0,24	0,78	-0,94	-0,29	-0,37	-0,45	-0,32	-0,04	-0,18	-0,09	-0,04	0,06	-0,03	0,05	0,03	0,05	0,22	0,18	-0,16	0,09	-0,19	-0,15	-0,17	-0,20	-0,14	-0,17	-0,01
2	0,23	1	0,11	-0,01	0,16	-0,23	-0,17	-0,09	-0,11	-0,29	0,02	-0,10	-0,42	-0,33	0,34	0,27	0,34	0,36	-0,03	0,45	0,42	-0,15	0,36	-0,30	-0,27	-0,35	-0,36	-0,59	-0,38	0,44
3	-0,06	0,11	1	0,20	-0,04	0,06	0,02	0,04	0,06	-0,04	-0,02	0,06	0,01	0,10	0,08	-0,21	-0,02	-0,08	0,23	-0,03	-0,05	0,01	-0,05	0,02	0,02	0,04	0,03	-0,16	-0,02	0,02
4	-0,24	-0,01	0,20	1	-0,21	0,22	0,16	0,33	0,42	-0,27	0,10	-0,16	-0,03	-0,09	0,00	0,21	0,05	0,11	-0,10	0,14	0,11	0,09	0,20	0,03	-0,04	-0,06	-0,07	0,14	0,01	0,02
5	0,78	0,16	-0,04	-0,21	1	-0,53	-0,17	-0,21	-0,28	-0,33	-0,02	-0,12	-0,05	-0,01	0,06	-0,03	0,01	-0,01	0,02	0,14	0,09	-0,07	0,06	-0,09	-0,07	-0,07	-0,11	-0,08	-0,11	0,00
6	-0,94	-0,23	0,06	0,22	-0,53	1	0,31	0,39	0,47	0,26	0,04	0,19	0,09	0,04	-0,06	0,03	-0,07	-0,04	-0,05	-0,23	-0,19	0,18	-0,09	0,21	0,16	0,19	0,21	0,15	0,17	0,01
7	-0,29	-0,17	0,02	0,16	-0,17	0,31	1	0,71	0,68	0,41	-0,04	0,13	0,03	0,02	-0,05	0,03	-0,05	-0,04	-0,07	-0,18	-0,21	0,23	-0,10	0,26	0,25	0,24	0,23	0,07	0,06	-0,04
8	-0,37	-0,09	0,04	0,33	-0,21	0,39	0,71	1	0,98	0,16	-0,02	0,10	0,02	0,04	-0,03	0,01	-0,02	-0,01	-0,03	-0,08	-0,11	0,15	-0,03	0,15	0,14	0,13	0,12	0,06	0,02	0,00
9	-0,45	-0,11	0,06	0,42	-0,28	0,47	0,68	0,98	1	0,18	0,01	0,12	0,03	0,04	-0,03	0,04	-0,02	0,00	-0,04	-0,09	-0,11	0,16	-0,03	0,17	0,15	0,14	0,13	0,08	0,06	0,00
10	-0,32	-0,29	-0,04	-0,27	-0,33	0,26	0,41	0,16	0,18	1	0,01	0,22	0,15	0,06	-0,06	-0,10	-0,08	-0,10	0,01	-0,24	-0,21	0,08	-0,20	0,15	0,17	0,18	0,20	0,07	0,22	-0,07
11	-0,04	0,02	-0,02	0,10	-0,02	0,04	-0,04	-0,02	0,01	0,01	1	0,09	0,32	-0,08	0,37	0,10	-0,42	-0,31	-0,38	0,24	0,26	-0,03	0,30	-0,12	-0,21	-0,22	-0,21	0,43	0,50	0,49
12	-0,18	-0,10	0,06	-0,16	-0,12	0,19	0,13	0,10	0,12	0,22	0,09	1	0,19	0,29	0,09	-0,25	-0,24	-0,27	0,02	-0,30	-0,31	0,29	-0,23	0,34	0,41	0,35	0,35	0,03	0,30	0,21
13	-0,09	-0,42	0,01	-0,03	-0,05	0,09	0,03	0,02	0,03	0,15	0,32	0,19	1	0,08	0,09	-0,23	-0,32	-0,34	0,00	-0,11	-0,10	0,02	-0,06	0,06	0,01	0,08	0,08	0,34	0,52	0,03
14	-0,04	-0,33	0,10	-0,09	-0,01	0,04	0,02	0,04	0,04	0,06	-0,08	0,29	0,08	1	-0,26	-0,45	-0,52	-0,56	0,02	-0,26	-0,31	0,13	-0,24	0,22	0,21	0,28	0,27	0,17	0,10	-0,34
15	0,06	0,34	0,08	0,00	0,06	-0,06	-0,05	-0,03	-0,03	-0,06	0,37	0,09	0,09	-0,26	1	0,12	-0,04	-0,01	-0,14	0,26	0,23	-0,08	0,24	-0,16	-0,20	-0,21	-0,23	0,04	0,14	0,62
16	-0,03	0,27	-0,21	0,21	-0,03	0,03	0,03	0,01	0,04	-0,10	0,10	-0,25	-0,23	-0,45	0,12	1	0,36	0,61	-0,64	0,19	0,23	0,14	0,33	0,02	-0,04	-0,13	-0,09	0,05	-0,12	0,30
17	0,05	0,34	-0,02	0,05	0,01	-0,07	-0,05	-0,02	-0,02	-0,08	-0,42	-0,24	-0,32	-0,52	-0,04	0,36	1	0,94	0,40	0,08	0,18	-0,17	0,03	-0,19	-0,11	-0,20	-0,14	-0,49	-0,36	-0,01
18	0,03	0,36	-0,08	0,11	-0,01	-0,04	-0,04	-0,01	0,00	-0,10	-0,31	-0,27	-0,34	-0,56	-0,01	0,61	0,94	1	0,14	0,12	0,21	-0,10	0,12	-0,15	-0,10	-0,20	-0,14	-0,38	-0,32	0,09
19	0,05	-0,03	0,23	-0,10	0,02	-0,05	-0,07	-0,03	-0,04	0,01	-0,38	0,02	0,00	0,02	-0,14	-0,64	0,40	0,14	1	-0,13	-0,09	-0,23	-0,28	-0,14	-0,03	-0,02	-0,01	-0,39	-0,14	-0,31
20	0,22	0,45	-0,03	0,14	0,14	-0,23	-0,18	-0,08	-0,09	-0,24	0,24	-0,30	-0,11	-0,26	0,26	0,19	0,08	0,12	-0,13	1	0,88	-0,43	0,69	-0,67	-0,62	-0,82	-0,90	-0,06	-0,14	0,36
21	0,18	0,42	-0,05	0,11	0,09	-0,19	-0,21	-0,11	-0,11	-0,21	0,26	-0,31	-0,10	-0,31	0,23	0,23	0,18	0,21	-0,09	0,88	1	-0,50	0,69	-0,76	-0,65	-0,94	-0,87	-0,09	-0,10	0,32
22	-0,16	-0,15	0,01	0,09	-0,07	0,18	0,23	0,15	0,16	0,08	-0,03	0,29	0,02	0,13	-0,08	0,14	-0,17	-0,10	-0,23	-0,43	-0,50	1	0,13	0,94	0,68	0,74	0,75	0,23	0,17	0,07
23	0,09	0,36	-0,05	0,20	0,06	-0,09	-0,10	-0,03	-0,03	-0,20	0,30	-0,23	-0,06	-0,24	0,24	0,33	0,03	0,12	-0,28	0,69	0,69	0,13	1	-0,16	-0,53	-0,48	-0,46	0,15	0,08	0,39
24	-0,19	-0,30	0,02	0,03	-0,09	0,21	0,26	0,15	0,17	0,15	-0,12	0,34	0,06	0,22	-0,16	0,02	-0,19	-0,15	-0,14	-0,67	-0,76	0,94	-0,16	1	0,75	0,91	0,90	0,23	0,18	-0,08
25	-0,15	-0,27	0,02	-0,04	-0,07	0,16	0,25	0,14	0,15	0,17	-0,21	0,41	0,01	0,21	-0,20	-0,04	-0,11	-0,10	-0,03	-0,62	-0,65	0,68	-0,53	0,75	1	0,75	0,76	0,04	0,00	-0,11
26	-0,17	-0,35	0,04	-0,06	-0,07	0,19	0,24	0,13	0,14	0,18	-0,22	0,35	0,08	0,28	-0,21	-0,13	-0,20	-0,20	-0,02	-0,82	-0,94	0,74	-0,48	0,91	0,75	1	0,94	0,14	0,11	-0,21
27	-0,20	-0,36	0,03	-0,07	-0,11	0,21	0,23	0,12	0,13	0,20	-0,21	0,35	0,08	0,27	-0,23	-0,09	-0,14	-0,14	-0,01	-0,90	-0,87	0,75	-0,46	0,90	0,76	0,94	1	0,13	0,13	-0,23
28	-0,14	-0,59	-0,16	0,14	-0,08	0,15	0,07	0,06	0,08	0,07	0,43	0,03	0,34	0,17	0,04	0,05	-0,49	-0,38	-0,39	-0,06	-0,09	0,23	0,15	0,23	0,04	0,14	0,13	1	0,40	0,22
29	-0,17	-0,38	-0,02	0,01	-0,11	0,17	0,06	0,02	0,06	0,22	0,50	0,30	0,52	0,10	0,14	-0,12	-0,36	-0,32	-0,14	-0,14	-0,10	0,17	0,08	0,18	0,00	0,11	0,13	0,40	1	0,14
30	-0,01	0,44	0,02	0,02	0,00	0,01	-0,04	0,00	0,00	-0,07	0,49	0,21	0,03	-0,34	0,62	0,30	-0,01	0,09	-0,31	0,36	0,32	0,07	0,39	-0,08	-0,11	-0,21	-0,23	0,22	0,14	1

Figura 12: Matriz de correlación para el Potrero 1.

Por lo tanto, podemos ver que las variables que tienen una correlación positiva para el criterio tomado son:

1	Curvatura	Y	5	Plano de curvatura
7	Área específica de cuenca	Y	8	Índice topográfico SPI
8	Índice topográfico SPI	Y	9	Índice topográfico STCI
17	CE de 30 a 90 cm	Y	18	CE a 90 cm
20	Reflectancia banda 2	Y	21	Reflectancia banda 3
22	Reflectancia banda 4	Y	24	Tasseled cap banda 2
22	Reflectancia banda 4	Y	26	Índice NDVI
22	Reflectancia banda 4	Y	27	Índice GNDVI
24	Tasseled cap banda 2	Y	25	Tasseled cap banda 3
24	Tasseled cap banda 2	Y	26	Índice NDVI
24	Tasseled cap banda 2	Y	27	Índice GNDVI
25	Tasseled cap banda 3	Y	26	Índice NDVI
25	Tasseled cap banda 3	Y	27	Índice GNDVI
26	Índice NDVI	Y	27	Índice GNDVI

Figura 13: Variables con correlación positiva para el Potrero 1.

Y las variables que tienen una correlación negativa son:

1	Curvatura	Y	6	Perfil de curvatura
20	Reflectancia banda 2	Y	26	Índice NDVI
20	Reflectancia banda 2	Y	27	Índice GNDVI
21	Reflectancia banda 3	Y	24	Tasseled cap banda 2
21	Reflectancia banda 3	Y	26	Índice NDVI
21	Reflectancia banda 3	Y	27	Índice GNDVI

Figura 14: Variables con correlación negativa para el Potrero 1.

Estos resultados son coherentes con las imágenes que se presentan de las variables (*ver: Anexo 5: Datos de prueba*). Es decir, para las variables que presentan una correlación positiva, puede observarse que los colores de una variable se corresponden con la otra, ambas tienen valores altos o bajos en los mismos puntos. Para las variables que tienen una correlación negativa ocurre lo inverso, cuando una variable tiene un color, la otra variable tiene el color opuesto para los mismos puntos.

Ahora analizaremos la matriz de varianza-covarianza para establecer la medida de similitud a utilizar.

La matriz de varianza-covarianza es una matriz simétrica que brinda una medida de la variación conjunta entre dos variables. Los elementos diagonales de la matriz representan la varianza de la variable, mientras que el resto de los elementos son la covarianza de las variables involucradas.

A continuación se muestran los resultados obtenidos:

1,08	1,05	-6,78	-0,79	0,34	-0,74	-445,72	-105,71	-5,42	-0,56	-0,12	-0,49	-0,35	-0,36	0,10	-0,43	0,62	0,33	0,64	0,00	0,00	0,00	0,67	-2,66	-2,13	-2,11	-1,77	-0,14	-0,17	0,00
1,05	19,21	58,00	-0,08	0,29	-0,76	-1.118,89	-109,72	-5,70	-2,17	0,26	-1,10	-6,87	-13,79	2,32	14,59	17,17	16,91	-1,96	0,02	0,03	-0,02	11,38	-17,57	-16,50	-18,12	-13,52	-2,57	-1,62	1,06
-6,78	58,00	13.370,85	70,76	-1,78	5,00	3.296,65	1.150,29	79,85	-8,04	-5,29	18,33	3,40	113,30	15,05	-291,08	-25,89	-102,00	351,28	-0,04	-0,08	0,02	-42,65	27,99	26,94	51,26	27,76	-18,20	-1,86	1,21
-0,79	-0,08	70,76	9,88	-0,28	0,52	712,55	282,69	15,27	-1,41	0,84	-1,30	-0,29	-2,60	0,00	8,23	1,83	3,78	-4,30	0,00	0,01	0,01	4,46	1,11	-1,56	-2,23	-1,79	0,44	0,02	0,03
0,34	0,29	-1,78	-0,28	0,17	-0,17	-101,55	-24,02	-1,34	-0,23	-0,02	-0,12	-0,08	-0,05	0,04	-0,16	0,04	-0,03	0,11	0,00	0,00	0,00	0,17	-0,48	-0,39	-0,36	-0,38	-0,03	-0,04	0,00
-0,74	-0,76	5,00	0,52	-0,17	0,58	344,17	81,70	4,08	0,34	0,10	0,37	0,26	0,31	-0,07	0,26	-0,58	-0,37	-0,53	0,00	0,00	0,00	-0,51	2,18	1,73	1,74	1,39	0,11	0,13	0,00
-445,72	-1.118,89	3.296,65	712,55	-101,55	344,17	2.141.570,84	282.735,97	11.470,26	1.019,52	-147,08	492,87	182,87	246,75	-109,68	508,51	-858,57	-583,46	-1.283,42	-2,39	-4,48	8,92	-1.099,52	5.039,48	5.034,29	4.186,00	2.856,04	106,24	87,14	-33,42
-105,71	-109,72	1.150,29	282,69	-24,02	81,70	282.735,97	74.699,19	3.081,10	71,31	-12,28	68,93	16,22	109,19	-11,23	39,78	-70,77	-37,38	-116,62	-0,19	-0,42	1,07	-60,15	559,13	546,57	420,03	276,83	15,99	5,08	-0,25
-5,42	-5,70	79,85	15,27	-1,34	4,08	11.470,26	3.081,10	132,64	3,52	0,24	3,37	1,37	4,81	-0,62	4,87	-2,16	0,11	-6,24	-0,01	-0,02	0,05	-2,44	25,70	24,15	18,78	12,83	0,91	0,62	0,01
-0,56	-2,17	-8,04	-1,41	-0,23	0,34	1.019,52	71,31	3,52	2,83	0,06	0,93	0,95	0,95	-0,14	-1,98	-1,54	-1,73	0,28	0,00	-0,01	0,00	-2,38	3,43	3,37	3,65	2,83	0,12	0,35	-0,07
-0,12	0,26	-5,29	0,84	-0,02	0,10	-147,08	-12,28	0,24	0,06	7,91	0,61	3,38	-2,25	1,65	3,53	-13,67	-9,38	-14,58	0,01	0,01	0,00	5,19	-4,69	-8,04	-7,47	-5,03	1,22	1,36	0,75
-0,49	-1,10	18,33	-1,30	-0,12	0,37	492,87	68,93	3,37	0,93	0,61	6,51	1,86	7,15	0,37	-7,80	-7,20	-7,57	0,75	-0,01	-0,01	0,02	-4,30	11,64	14,60	10,67	7,76	0,07	0,75	0,28
-0,35	-6,87	3,40	-0,29	-0,08	0,26	182,87	16,22	1,37	0,95	3,38	1,86	14,09	2,83	0,56	-10,71	-14,08	-13,68	-0,11	0,00	-0,01	0,00	-1,68	3,09	0,55	3,70	2,56	1,29	1,90	0,06
-0,36	-13,79	113,30	-2,60	-0,05	0,31	246,75	109,19	4,81	0,95	-2,25	7,15	2,83	91,07	-3,87	-52,00	-56,91	-57,82	2,67	-0,02	-0,04	0,03	-16,75	27,94	28,22	32,48	22,02	1,64	0,90	-1,74
0,10	2,32	15,05	0,00	0,04	-0,07	-109,68	-11,23	-0,62	-0,14	1,65	0,37	0,56	-3,87	2,45	2,19	-0,80	-0,10	-3,04	0,00	0,01	0,00	2,73	-3,41	-4,41	-3,93	-3,14	0,06	0,21	0,53
-0,43	14,59	-291,08	8,23	-0,16	0,26	508,51	39,78	4,87	-1,98	3,53	-7,80	-10,71	-52,00	2,19	149,53	51,24	80,93	-105,84	0,02	0,04	0,05	29,29	2,46	-7,34	-18,53	-9,89	0,63	-1,38	2,02
0,62	17,17	-25,89	1,83	0,04	-0,58	-858,57	-70,77	-2,16	-1,54	-13,67	-7,20	-14,08	-56,91	-0,80	51,24	133,90	118,18	62,86	0,01	0,03	-0,05	2,24	-30,03	-18,07	-27,15	-14,25	-5,68	-4,00	-0,06
0,33	16,91	-102,00	3,78	-0,03	-0,37	-583,46	-37,38	0,11	-1,73	-9,38	-7,57	-13,68	-57,82	-0,10	80,93	118,18	117,86	19,91	0,01	0,03	-0,03	9,28	-22,16	-15,36	-25,58	-13,30	-4,11	-3,35	0,51
0,64	-1,96	351,28	-4,30	0,11	-0,53	-1.283,42	-116,62	-6,24	0,28	-14,58	0,75	-0,11	2,67	-3,04	-105,84	62,86	19,91	182,71	-0,02	-0,02	-0,08	-27,51	-24,65	-4,95	-3,72	-1,03	-5,21	-1,80	-2,24
0,00	0,02	-0,04	0,00	0,00	0,00	-2,39	-0,19	-0,01	0,00	0,01	-0,01	0,00	-0,02	0,00	0,02	0,01	0,01	-0,02	0,00	0,00	0,00	0,05	-0,08	-0,08	-0,09	-0,07	0,00	0,00	0,00
0,00	0,03	-0,08	0,01	0,00	0,00	-4,48	-0,42	-0,02	-0,01	0,01	-0,01	-0,01	-0,04	0,01	0,04	0,03	0,03	-0,02	0,00	0,00	0,00	0,07	-0,15	-0,13	-0,16	-0,11	0,00	0,00	0,00
0,00	-0,02	0,02	0,01	0,00	0,00	8,92	1,07	0,05	0,00	0,00	0,02	0,00	0,03	0,00	0,05	-0,05	-0,03	-0,08	0,00	0,00	0,00	0,02	0,34	0,26	0,24	0,17	0,01	0,00	0,00
0,67	11,38	-42,65	4,46	0,17	-0,51	-1.099,52	-60,15	-2,44	-2,38	6,19	-4,30	-1,68	-16,75	2,73	29,29	2,24	9,28	-27,51	0,05	0,07	0,02	52,78	-15,73	-53,31	-42,07	-28,46	1,08	0,57	1,55
-2,66	-17,57	27,99	1,11	-0,48	2,18	5.033,48	559,13	25,70	3,43	-4,69	11,64	3,09	27,94	-3,41	2,46	-30,03	-22,16	-24,65	-0,08	-0,15	0,34	-15,73	181,05	140,40	147,19	104,06	3,09	2,29	-0,57
-2,13	-16,50	26,94	-1,56	-0,39	1,73	5.034,29	546,57	24,15	3,37	-8,04	14,60	0,55	28,22	-4,41	-7,34	-18,07	-15,36	-4,95	-0,08	-0,13	0,26	-53,31	140,40	195,28	125,67	91,20	0,49	0,00	-0,84
-2,11	-18,12	51,26	-2,23	-0,36	1,74	4.186,00	420,03	18,78	3,65	-7,47	10,67	3,70	32,48	-3,93	-18,53	-27,15	-25,58	-3,72	-0,09	-0,16	0,24	-42,07	147,19	125,67	143,24	97,25	1,66	1,30	-1,35
-1,77	-13,52	27,76	-1,79	-0,38	1,39	2.856,04	276,83	12,83	2,83	-5,03	7,76	2,56	22,02	-3,14	-9,89	-14,25	-13,30	-1,03	-0,07	-0,11	0,17	-28,46	104,06	91,20	97,25	74,24	1,14	1,11	-1,05
-0,14	-2,57	-18,20	0,44	-0,03	0,11	106,24	15,99	0,91	0,12	1,22	0,07	1,29	1,64	0,06	0,63	-5,68	-4,11	-5,21	0,00	0,00	0,01	1,08	3,09	0,49	1,66	1,14	0,99	0,39	0,12
-0,17	-1,62	-1,86	0,02	-0,04	0,13	87,14	5,08	0,62	0,35	1,36	0,75	1,90	0,90	0,21	-1,38	-4,00	-3,35	-1,80	0,00	0,00	0,00	0,57	2,29	0,00	1,30	1,11	0,39	0,94	0,07
0,00	1,06	1,21	0,03	0,00	0,00	-33,42	-0,25	0,01	-0,07	0,75	0,28	0,06	-1,74	0,53	2,02	-0,06	0,51	-2,24	0,00	0,00	0,00	1,55	-0,57	-0,84	-1,35	-1,05	0,12	0,07	0,29

Figura 15: Matriz de varianza-covarianza para el Potrero 1

Como puede verse, los valores de varianza son bastante diferentes. En cuanto a la covarianza, los elementos coloreados que no están en la diagonal, indican la existencia de valores positivos y negativos distintos de cero. Por lo tanto, es coherente, que la operación que calcula la norma a utilizar devuelva como resultado la distancia de

Mahalanobis para este caso. Esto concuerda con lo planteado en (33) para el tipo de variables utilizadas.

Luego, se generaron las zonas de manejo estableciendo el tamaño mínimo de zona en 10.000 m^2.

En las imágenes presentadas a continuación, pueden verse las zonas generadas antes y después de aplicar el post-procesamiento. Se puede observar claramente, que los puntos o sub-zonas que no cumplen con el tamaño mínimo son reasignadas a la zona adyacente con mayor borde circundante.

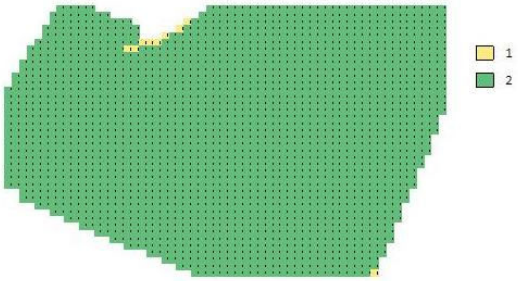
También se presentan los índices de performance obtenidos para cada caso. Es importante recordar, que los índices son calculados antes de realizar el post-procesamiento.

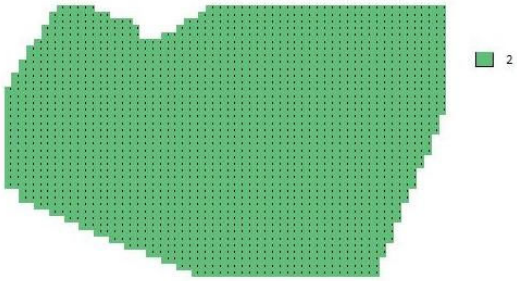
El índice FPI (Fuzzy Performance Index) es una medida del grado de separación entre las zonas. Valores cercanos a cero indican distintas zonas con una baja pertenencia, mientras que valores cercanos a uno indican clases similares con alto grado de pertenencia. El índice NCE (Normalizad Classification Entropy) representa el grado de desorganización de la partición obtenida. Sus valores varían entre 0 y $\log_a(c)$, donde el valor de a elegido para la implementación fue 2.

Generalmente, la mejor clasificación es aquella para la cual el grado de pertenencia y/o la desorganización es mínima con el menor número de zonas utilizadas.

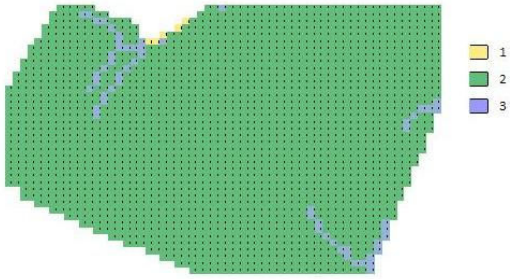
Distancia Euclideana

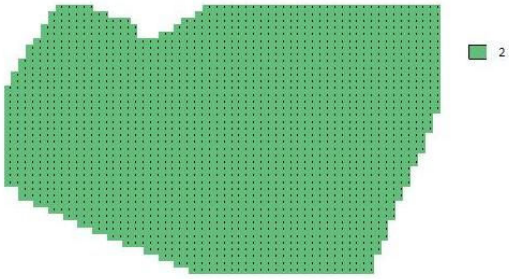
Cantidad de zonas = 2



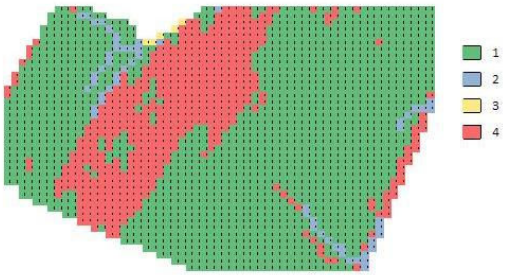


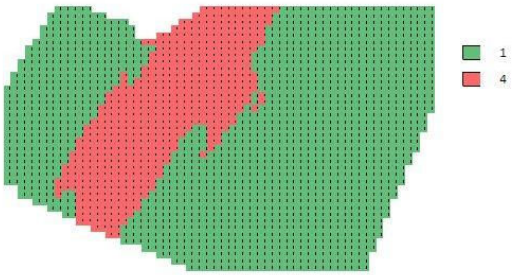
Cantidad de zonas = 3



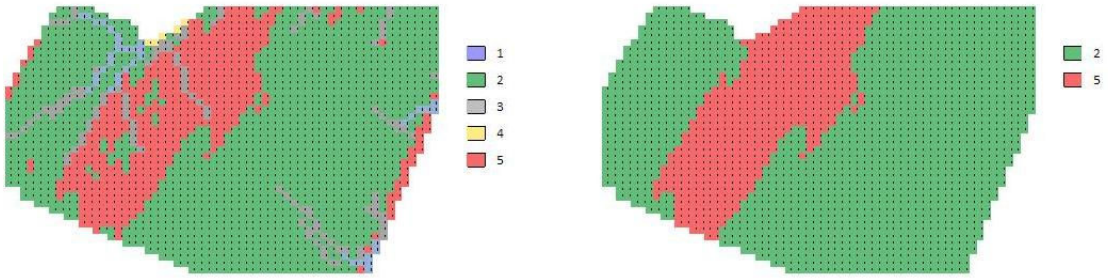


Cantidad de zonas = 4

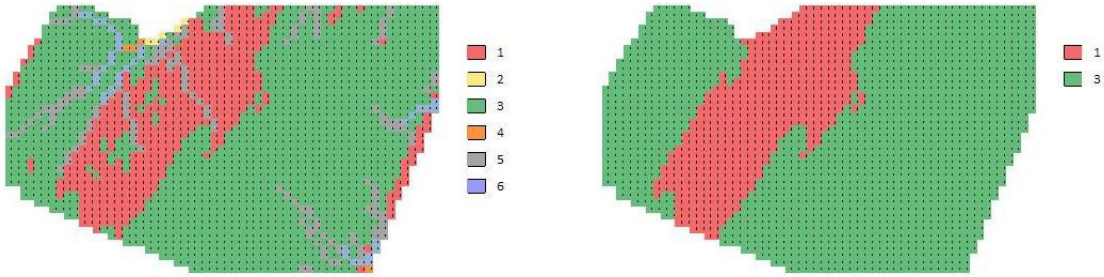




Cantidad de zonas = 5



Cantidad de zonas = 6



Índices de Performance

Los índices de performance obtenidos para estos casos se muestran en la Figura 16.

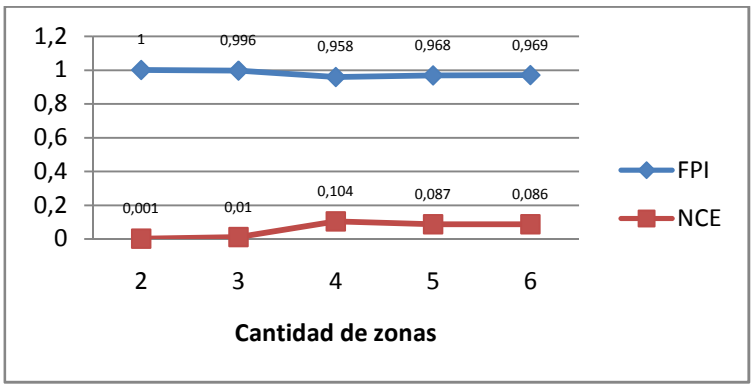
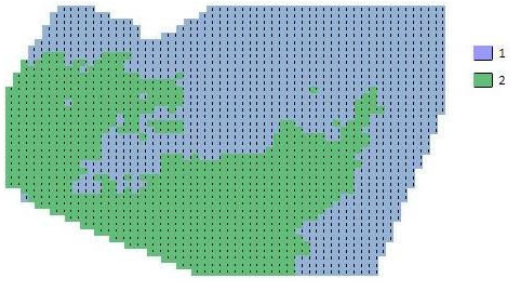


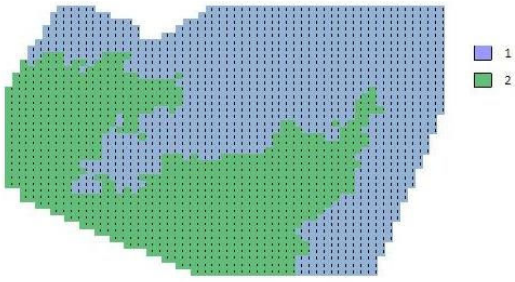
Figura 16: Índices de performance FPI y NCE calculados para el Potrero 1 utilizando la distancia Eculideana.

En este caso, no está muy claro cuál es la cantidad óptima de zonas. Esto confirma que la distancia Euclideana no es la mejor opción debido a las características presentadas por las variables.

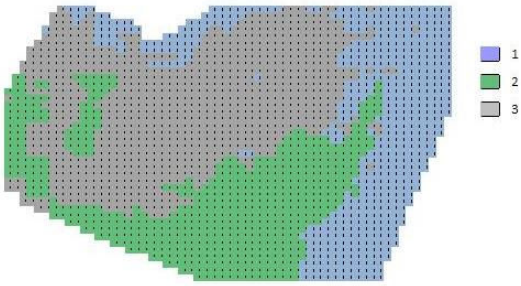
Distancia Diagonal

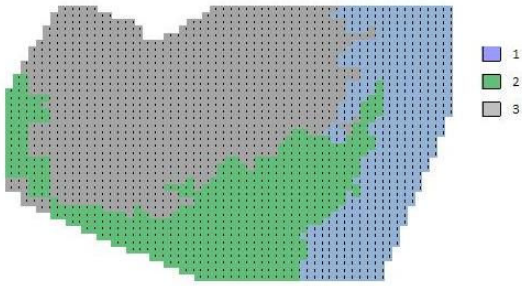
Cantidad de zonas = 2



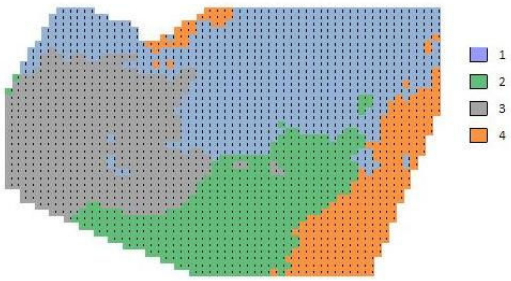


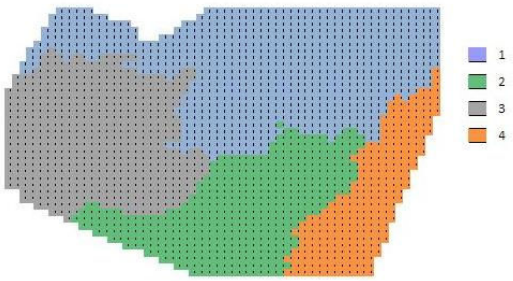
Cantidad de zonas = 3



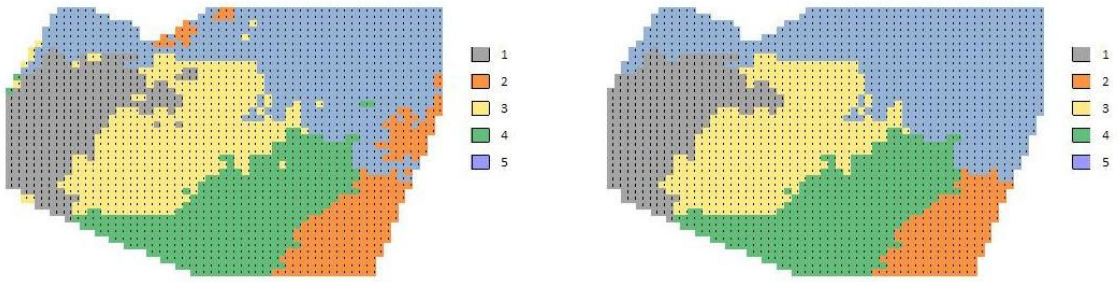


Cantidad de zonas = 4

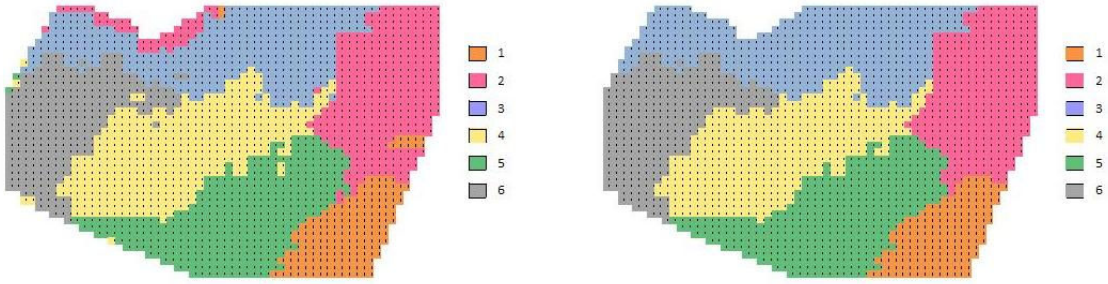




Cantidad de zonas = 5



Cantidad de zonas = 6



Índices de Performance

Los índices de performance obtenidos para estos casos se muestran en la Figura 17.

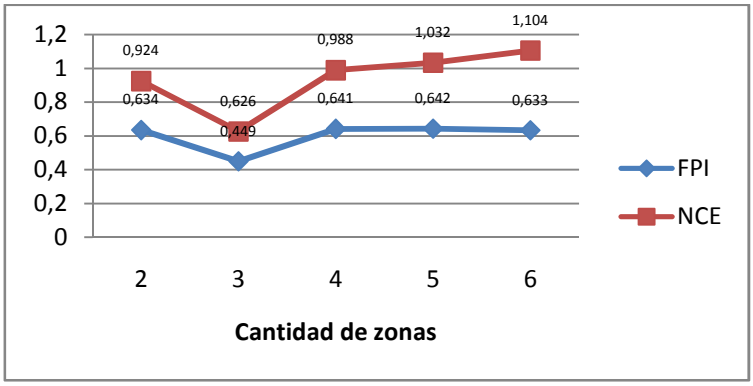
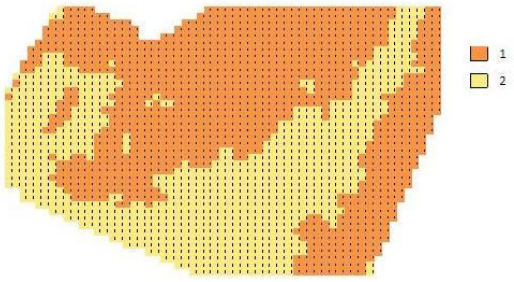


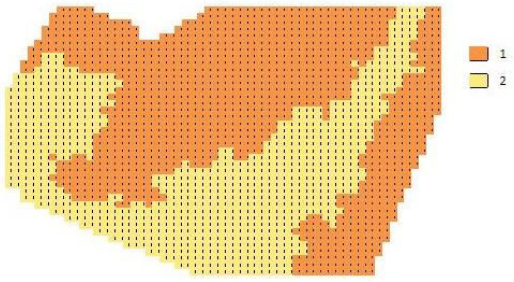
Figura 17: Índices de performance FPI y NCE calculados para el Potrero 1 utilizando la distancia Diagonal.

En este caso, la mejor clasificación corresponde a la generación de 3 zonas de manejo.

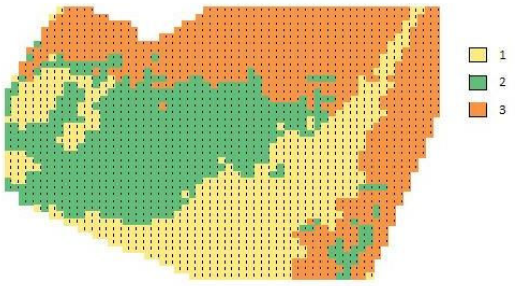
Distancia de Mahalanobis

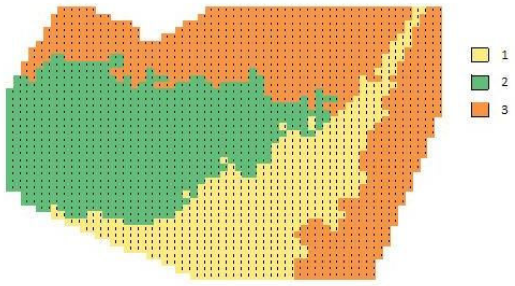
Cantidad de zonas = 2



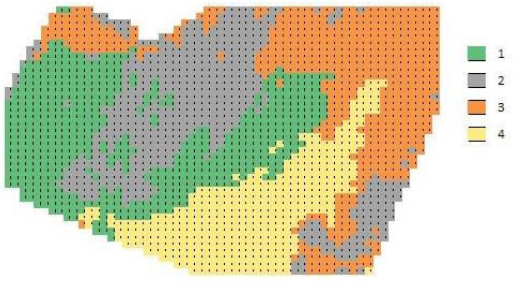


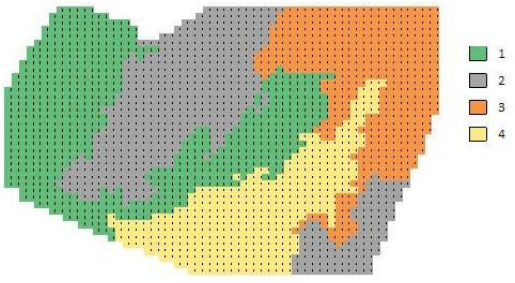
Cantidad de zonas = 3



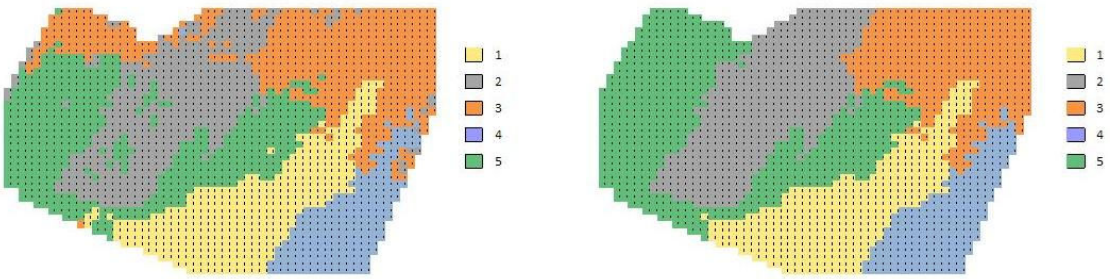


Cantidad de zonas = 4

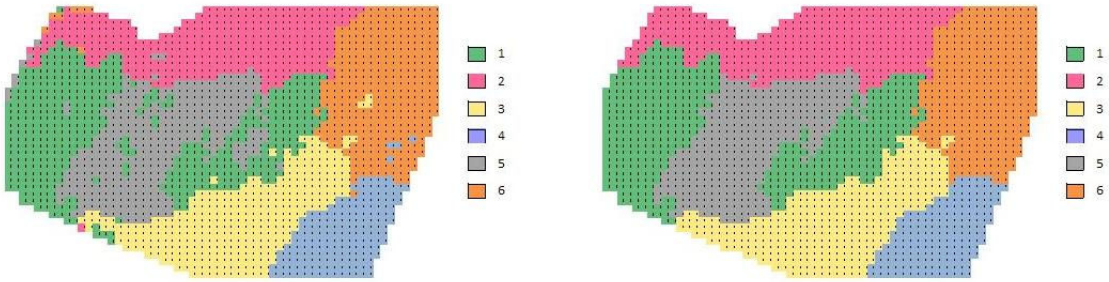




Cantidad de zonas = 5



Cantidad de zonas = 6



Índices de Performance

Los índices de performance obtenidos para estos casos se muestran en la Figura 18.

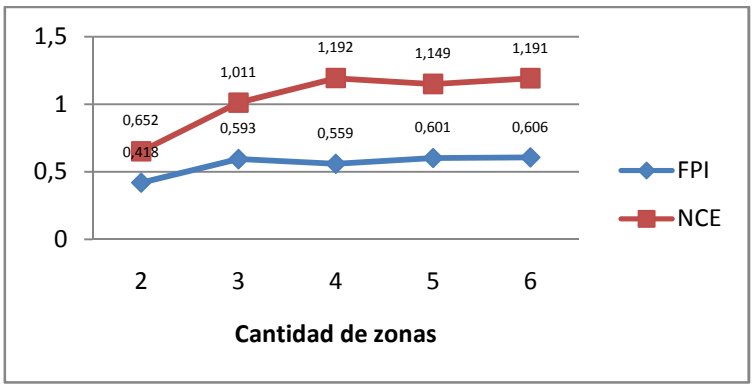


Figura 18: Índices de performance FPI y NCE calculados para el Potrero 1 utilizando la distancia de Mahalanobis.

En este caso, la mejor clasificación corresponde a la generación de 2 zonas de manejo.

CAPÍTULO 8: Conclusiones y trabajo futuro

Conclusiones

Los objetivos propuestos del proyecto fueron cumplidos en su totalidad. Se logró obtener una perspectiva del estado actual de la Agricultura de Precisión, comprendiendo sus conceptos y objetivos. Igualmente, se comprendieron y aplicaron las técnicas de Minería de Datos como forma de proveer una solución a la obtención de información a partir del procesamiento de los datos.

También se logró la implementación del módulo de generación de zonas de manejo con los requerimientos planteados, el cual va a poder ser integrado al Sistema de Información Geográfica de Gestión Agrícola.

El proyecto requirió de una gran parte de investigación. Primero dentro de lo que es la Agricultura de Precisión para tener una visión global del contexto en el que se aplicaba el proyecto. Esto nos permitió conocer y aprender de esta técnica, que puede decirse que va de la mano con la Informática. Además, se pudo comprender a fondo el tema de la generación de zonas, y qué es lo que se busca con ellas. Luego, para seguir profundizando, se investigó sobre las diferentes técnicas de Data Mining que nos pudieran brindar una solución al problema en cuestión. Paralelamente a las investigaciones, por medio de reuniones y mails con el cliente, se fueron definiendo las características y funcionalidades que debía proveer el sistema a construir.

Con el conocimiento obtenido, se realizó un análisis y se presentó el diseño de una solución que cumplía con los objetivos y requerimientos establecidos. Durante la etapa de implementación fueron surgiendo nuevos desafíos o dificultades que no se podían prever en etapas anteriores. Las mismas pudieron ser resultas sin afectar el objetivo del proyecto, e incluso permitiendo dar más flexibilidad a la aplicación final.

Si bien no se contó con una herramienta específica, el conseguir una herramienta que permitiera la visualización de las zonas y las variables, fue muy importante para poder verificar, que tanto las estadísticas como las zonas generadas eran coherentes con los datos de entrada.

El generar zonas de manejo homogéneo para un área de cultivo, constituye un pequeño pero importante paso dentro del proceso de Agricultura de Precisión. Las zonas determinan las diferentes áreas a ser tratadas, permitiendo un manejo diferenciado del área de cultivo, teniendo en cuenta la variabilidad espacial presente en la misma, y brindado así beneficios económicos y ambientales.

La integración de la aplicación con el Sistema de Información Geográfica de Gestión Agrícola, permitirá contar con una herramienta que puede usarse como soporte en la toma de decisiones a la hora de gestionar un campo. Por lo que podemos decir entonces, que logramos la obtención de conocimiento a través de la aplicación de la Informática a la agricultura.

Dificultades encontradas

Visualización de datos

No se contó con una herramienta que permitiera visualizar fácilmente las zonas generadas y las variables de forma descriptiva.

Por esa razón, se tuvieron que implementar procedimientos auxiliares que generaran matrices con los datos, tanto para las variables como para las zonas generadas. Una vez obtenidos los datos en forma de matriz, se utilizó la herramienta Excel 2007 para la visualización. Se aplicó formato condicional para colorear los datos según su valor, en el caso de las variables y según el número de zona en el caso de la zonificación.

La visualización de las zonas y las variables permitió verificar que las estadísticas y las zonas generadas fueran coherentes, así como también verificar las zonas obtenidas luego del post-procesamiento. Además permitió deducir la influencia de las variables en la generación de zonas.

Datos de prueba

Consideramos que la performance del sistema no se puedo verificar correctamente debido al tamaño de las chacras utilizadas para realizar las pruebas. Los datos entregados corresponden a potreros de 41 y 39 Ha aproximadamente. Estos valores no representan campos de gran tamaño en lo que se refiere a un área de cultivo.

Trabajo futuro

Respecto a la implementación se plantean como principales desafíos a futuro los siguientes:

- La optimización de la aplicación, está se podría realizar utilizando alguna herramienta adicional específica para cálculos matemáticos, en la que se realicen todas las cuentas necesarias para el funcionamiento del sistema.
- Se podría generar una interfaz de usuario que permita el mismo seleccionar las variables a utilizar y también ingresar los parámetros utilizados por el sistema. De la misma forma se podría crear otra interfaz que brinde al usuario los resultados de las estadísticas y caracterización de zonas de forma más amigable.
- Para el caso de las zonas generadas antes y después del post-procesamiento sería útil que el sistema se comunique con otro de visualización, que le permita al usuario obtener como respuesta no solo los archivos generados sino también una visualización gráfica de la información contenida en los mismos.
- Dar más flexibilidad a la hora de generar las zonas, permitiendo especificar más restricciones sobre las mismas o sobre los valores de las variables.

También sería interesante exponer la implementación a un conjunto de pruebas más exhaustivas, con mayor cantidad de datos.

Bibliografía

1. **Wikipedia.** Sistema de Información Geográfica. [En línea] [Citado el: 14 de marzo de 2008.] http://es.wikipedia.org/wiki/Sistema_de_Informaci%C3%B3n_Geogr%C3%A1fica.

2. Data Mining & Knowledge Discovery in Databases (KDD). [En línea] [Citado el: 06 de febrero de 2008.] http://elvex.ugr.es/etexts/spanish/kdd/KDD.html.

3. **Bongiovanni, R., y otros.** *Agricultura de Precisión: Integrado conocimientos para una una agricultura moderna y sustentable.* Montevideo : PROCISUR, 2006.

4. **Bragachini, M., von Martini, A. y Méndez, A.** Proyecto Nacional Agricultura de Precisión - INTA Manfredi. *Tecnología Disponible para Aplicaciones de Insumo Sitio Específico.* [En línea] [Citado el: 29 de marzo de 2008.] http://www.agriculturadeprecision.org/mansit/TecnologiaDisponibleParaAplicacion.htm.

5. **ICA.** Ingenieros Consultores Asociados. [En línea] [Citado el: 29 de marzo de 2008.] http://www.ica.com.uy/.

6. **INIA.** Instituto Nacional de Investigación Agropecuaria. [En línea] [Citado el: 29 de marzo de 2008.] http://www.inia.org.uy/.

7. Agronomy Journal. [En línea] [Citado el: 15 de marzo de 2008.] http://agron.scijournals.org/cgi/content/abstract/96/1/100.

8. **Wikipedia.** Microsoft Visual Basic. [En línea] [Citado el: 15 de marzo de 2008.] http://es.wikipedia.org/wiki/Visual_Basic.

9. **GRDC: Grains Research & Development Corporation; Australian Centre for Precision Agriculture.** *A Process for Implementing Site-Specific Crop Management.* 2006.

10. **Lowenberg-DeBoer, J. y Bongiovanni, R.** Agricultura de Precisión. *Agricultura de Precisión y Sustentabilidad - 2001.* [En línea] [Citado el: 15 de febrero de 2008.] http://www.agriculturadeprecision.org/analecon/AgriPrecySustentabilidad.htm.

11. **Wikipedia.** Sistema de Posicionamiento Global. [En línea] [Citado el: 15 de marzo de 2008.] http://es.wikipedia.org/wiki/GPS.

12. **Wikipedia**. Teledetección. [En línea] [Citado el: 14 de marzo de 2008.] http://es.wikipedia.org/wiki/Percepci%C3%B3n_remota.

13. **Vallejos, S.** *Minería de Datos.* Facultad de Ciencias Exactas, Naturales y Agrimensura, Universidad Nacional del Nordeste. Corrientes - Argentina : s.n., 2006. Trabajo de Adscripción. 38 p.

14. *Obtención y validación de modelos de estimación de software mediante técnicas de minería de datos.* **Moreno, M., y otros.** 1, 2003, Revista Colombiana de Computación, Vol. 3, págs. 55-56.

15. **Acosta, M.** *Minería de datos y descubrimiento de conocimiento.* Facultad de Economía, Universidad de La Habana. 2004. 9 p.

16. **Villena, J.** *Minería de datos.* 2006. 32 p.

17. **Nigro, H., González, S. y Xodo, D.** *Ontologías en el proceso de descubrimiento de conocimiento en bases de datos.* INTIA - Departamento de Computación y Ciencias, Facultad de Ciencias Exactas - UNICEN. Buenos Aires, Argentina, 2007. 9 p.

18. **Weiss, S.M. y Indurkhya, N.** Cap 1.2: Types of Data-Mining Problems. *Predictive Data Mining: a practical guide.* s.l. : Morgan Kaufmann Publishers, Inc. San Francisco, 1998.

19. **Martínez, A.** *Minería de datos.* 2006. pp. 1-18.

20. **Hernández, A.** *Evaluación y optimización de consultas inductivas: el caso de los patrones secuenciales.* 2006. 1 p.

21. **Ventura, S.** *Minería de datos en sistemas educativos.* Departamento de Informática y Análisis Numérico, Universidad de Córdoba. 2007. 48 p.

22. Servicio de minería de datos y de soporte a la toma de decisiones. [En línea] [Citado el: 26 de febrero de 2008.] http://einstein.uab.es/_c_serv_estadistica/cas/Assessoria/SMD_Presentacio.htm.

23. **Terrádez, M.** *Análisis de componentes principales.* Proyecto e-Math, UOC - La Universidad Oberta de Catalunya. 2002. 11 p.

24. **Montes y Gómez, M., Gelbukh, A. y López-López, A.** *Detección de los patrones raros en un conjunto de datos semiestructurados.* Centro de Investigación en Computación (CIC-IPN); Instituto Nacional de astrofísica, Optica y Electrónica (INAOE). México, 2003. págs. 1-3.

25. **Prieto, O. y Alonso, C.** *Errror de posición en clasificadores bayesianos para la clasificación de series temporales.* 2005. pp. 2-3.

26. Revisión de técnicas de agrupamiento de minería de datos especiales en un SIG. [En línea] [Citado el: 21 de febrero de 2008.] http://www.monografias.com/trabajos27/datamining/datamining.shtml.

27. **Zaiane, o.** Cap 8: Data Clustering . *Principles of Knowledge Discovery in Databases.* s.l. : 8 pp., 1999.

28. **González, C.** *Apendizaje inductivo no basado en el error. Métodos no supervisados: agrupamiento.* 2007. pág. 55.

29. **Alonso, C.** *Aprendizaje inductivo no basado en el error. Métodos no supervisados: agrupamiento.* 2007. 55 p.

30. **Velez-Langs, o y Staffetti, E.** *Computación neuronal y evolutiva.* 2006. 45 p..

31. **Microsoft Corporation.** Visual Studio. [En línea] 2008. [Citado el: 10 de febrero de 2008.] http://msdn2.microsoft.com/es-ar/vstudio/default.aspx .

32. **ACPA: The Australian Centre for Precision Agriculture.** [En línea] [Citado el: 22 de setiembre de 2007.] http://www.usyd.edu.au/su/agric/acpa.

33. *Management Zone Analyst (MZA): Software for sub-fiels management zone delineation.* **Fridgen, J., y otros.** 2004.

34. **Wikipedia.** Subconjunto difuso. [En línea] [Citado el: 16 de febrero de 2008.] http://es.wikipedia.org/wiki/Subconjunto_difuso.

35. *A Fuzzy Clustering Model of Data and Fuzzy c-Means.* **Nascimiento, S., Mirkin, B. y Moura-Pires, F.** 1999.

36. **Wikipedia.** Distancia euclideana. [En línea] [Citado el: 15 de febrero de 2008.] http://es.wikipedia.org/wiki/Distancia_euclideana.

37. **Patel, Amit.** Heuristics. [En línea] 2007. [Citado el: 14 de marzo de 2008.] http://theory.stanford.edu/~amitp/GameProgramming/Heuristics.html.

38. **Wikipedia.** Distancia de Mahalanobis. [En línea] [Citado el: 15 de febrero de 2008.] http://es.wikipedia.org/wiki/Distancia_de_Mahalanobis.

39. **Nettleton, David F.** Técnicas para el análisis de datos clínicos. *Búsqueda de libros de Google.* [En línea] [Citado el: 02 de marzo de 2008.] http://books.google.com.uy/books?id=QqfuCWT3h8cC&pg=PA148&lpg=PA148&dq=norma+diagonal&source=web&ots=ew-C9227Dx&sig=_-NDFK_IAVd9xrLotmnCUeF6_HA&hl=es#PPA149,M1.

40. *Precision on decisions for quality cotton - A guide to site specific cotton crop Management.* **Stewart, C., Boydell, B. y McBratney, A.** 2005.

41. **Iturrate, Eduardo.** Curso básico de teledetección con ENVI. *Base matemática y estadística.* [En línea] [Citado el: 06 de marzo de 2008.] http://www.innovanet.com.ar/gis/TELEDETE/TELEDETE/bmatyest.htm.

42. **CollabNet.** Open Source Software Engineering Tools. *Subversion.* [En línea] 2006. [Citado el: 10 de febrero de 2008.] http://subversion.tigris.org/.

43. **CollabNet**. Open Source Software Engineering Tools. *TortoiseSVN.* [En línea] 2006. [Citado el: 10 de febrero de 2008.] http://tortoisesvn.tigris.org/.

44. **Wikipedia.** Caso de Uso. [En línea] [Citado el: 10 de febrero de 2008.] http://es.wikipedia.org/wiki/Caso_de_uso.

Glosario

Agricultura: Es el arte de cultivar la tierra; son los diferentes trabajos de tratamiento de suelo y cultivo de vegetales, normalmente con fines alimenticios.

Agricultura de Precisión**:** Proceso que permite a través de herramientas y tecnologías obtener los datos necesarios para dar a cada zona del campo cultivado el tratamiento agronómico más apropiado, tanto desde el punto de vista económico-productivo como del ambiental. Permitiendo así, incrementar la producción, reducir los costos de insumos o en la producción, tener en cuenta las necesidades reales del cultivo, reducir los impactos ambientales, etc.

Algoritmo: Es una lista bien definida, ordenada y finita de operaciones que permiten llegar a la solución de un problema.

Algoritmo Fuzzy c-Means**:** Algoritmo de Data Mining perteneciente a las técnicas de clustering no supervisado. Los algoritmos no supervisados son aquellos en los que al momento de hacer la clasificación no se conocen las propias clases y probablemente tampoco la cantidad de clases.

Algoritmos genéticos: Técnica de clasificación capaz de llevar a cabo una búsqueda adaptada y sólida en un amplio espectro de topologías de espacios de búsqueda.

Análisis de Componentes Principales (ACP): Es una técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables).

Árbol de decisión: Estructura en forma de árbol que representa un conjunto de decisiones. Estas decisiones forman reglas que permiten clasificar un conjunto de datos.

Atributo: Es una medida posible sobre un elemento o tipo de elemento, por ejemplo, temperatura o cantidad de meses.

Base de datos: Conjunto organizado e integrado de datos almacenados en una computadora, con el fin de facilitar su uso para aplicaciones con múltiples finalidades.

Biblioteca: Colección o conjunto de subprogramas usados para desarrollar software. En general, las bibliotecas no son ejecutables, pero sí pueden ser usadas por ejecutables que las necesiten para poder funcionar correctamente.

Campo: Tierra laborable.

Centroide: Es el punto donde de se encuentra el "punto de equilibrio".

Chacra: Terreno preparado para la siembra y cultivo de diversos productos.

Clasificación bayesiana: Técnica no supervisada de clasificación de datos.

Clúster: Grupo de elementos interconectados, que están agrupados de alguna forma.

Clustering: Es la clasificación de instancias en distintos grupos, de modo que las instancias de cada grupo sean lo más similares posibles y que cada grupo sea lo más distinto posible a los demás.

Coeficiente de correlación: Establece una medida de asociación lineal entre variables.

Coeficiente de variación: Desviación estándar dividido la media.

Conductividad eléctrica: Capacidad de un cuerpo de permitir el paso de la corriente eléctrica a través de sí.

Conocimiento: Es un conjunto de datos sobre hechos, verdades o de información ganada, a través de la experiencia, del aprendizaje o de introspección. El conocimiento es una apreciación de la posesión de múltiples datos interrelacionados que por sí solos poseen menor valor cualitativo.

Constante: Dato que permanece invariable.

Cosechar: Recolección de frutos, semillas u hortalizas de los campos.

Covarianza: El análisis de la covarianza es una técnica estadística que, utilizando un modelo de regresión lineal múltiple, busca comparar los resultados obtenidos en diferentes grupos de una variable cuantitativa pero corrigiendo las posibles diferencias existentes entre los grupos en otras variables que pudieran afectar también al resultado.

Cultivar: Cuidar la tierra y las plantas para que fructifiquen.

Data Mining: El análisis semiautomático de relaciones existentes en el contenido de una base de datos de gran tamaño, buscando encontrar información oculta, patrones, modelos, estructuras e información útil en general, para la toma de decisiones. Las bases utilizadas están formadas por un conjunto de datos relevantes para la organización en cuestión, donde al ser de gran tamaño las relaciones en la misma no son triviales.

Dendograma: Es un tipo de representación gráfica o diagrama de datos en forma de árbol que organiza los datos en sub-categorías que se van dividiendo en otras hasta llegar al nivel de detalle deseado (asemejándose a las ramas de un árbol que se van dividiendo en otras sucesivamente).

Desviación estándar: Se utiliza para medir la cantidad de variabilidad.

Distancia de Manhattan: Es la suma de los valores absolutos de la diferencia entre observaciones para cada variable. Ej: La distancia de Manhattan entre el punto de coordenadas (x,y) y el punto de coordenadas (i,j), es: $|x-i|+|y-j|$.

Distancia Euclideana: Medida de similitud. La distancia Euclídeana entre dos puntos A y B es la longitud del segmento de recta con extremos en A y B.

Distancia de Mahalanobis: Medida de similitud que tiene en cuenta varianzas desiguales así como también las correlaciones entre las variables.

Distribución gaussiana: La distribución normal, también llamada distribución de Gauss o distribución gaussiana, es la distribución de probabilidad que con más frecuencia aparece en estadística y teoría de probabilidades. Esto se debe a dos razones fundamentalmente: su función de densidad es simétrica y con forma de campana, lo que favorece su aplicación como modelo a gran número de variables estadísticas, y sus propiedades matemáticas.

DLL: Dynamic Link Library ("Biblioteca de vínculos dinámicos") es un archivo que contiene funciones que se pueden llamar desde aplicaciones u otras DLLs. Las DLLs no pueden ejecutarse directamente, es necesario llamarlas desde un código externo.

Entropía: Es la medida de incertidumbre que existe en un sistema, o sea, la probabilidad de que ocurra cada uno de los posibles resultados.

Estadística: Ciencia que utiliza conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades.

Función: Es el término para describir una secuencia de órdenes que hacen una tarea específica de una aplicación más grande.

Geoestadística: Estudio de las variables numéricas distribuidas en el espacio.

Hardware: La parte "que se puede tocar" de una computadora: caja (y todo su contenido), teclado, pantalla, etc. Lo físico de una computadora.

ICA: Ingenieros Consultores Asociados.

Información: Es un conjunto organizado de datos, que constituyen un mensaje sobre un determinado ente o fenómeno. La información es todo aquello que permite adquirir cualquier tipo de conocimiento.

Informática: Disciplina que estudia el tratamiento automático de la información utilizando dispositivos electrónicos y sistemas computacionales.

Insumo: Elemento que afecta de alguna manera la producción agrícola, semillas, agroquímicos, etc.

Interpolación: Proceso de calcular valores numéricos desconocidos a partir de otros ya conocidos mediante la aplicación de algoritmos concretos.

ISODATA: Iterative Self-Organizing Data Analysis Technique Algorithm. Algoritmo de clustering no supervisado.

Lógica difusa: Técnica de clasificación que se basa en lo relativo de lo observado, permitiendo representar de forma matemática conceptos o conjunto imprecisos.

Manejo de Sitio Específico: Hacer lo correcto en el momento adecuado y en el sitio oportuno, aplicado a la producción de cultivos.

Matriz varianza-covarianza: Es una matriz cuadrada simétrica que brinda una medida de la variación conjunta entre dos variables. Los elementos diagonales de la matriz representan la varianza, mientras que el resto la covarianza.

Media: Suma de todos los valores de una variable divida la cantidad de valores.

Mediana: Instancia del clúster más centrada.

Método Diagonal: Se utiliza para medir similitud, fue descrito por McBratney, Moore Odeh. No es sensible a variables correlacionadas pero compensa las distorsiones en las asumidas formas esféricas de los clústeres ponderando la varianza de las variables.

Módulo: Un programa puede constar de diferentes módulos y cada cual actúa de independientemente del otro. Un módulo es un conjunto de rutinas bien definidas que se une a un componente u otro módulo por medio de sus interfaces.

Monitor de Rendimiento: Formato para almacenar y mostrar datos espaciales de SIG en los que los datos son almacenados como puntos, líneas, o áreas que crean los objetos en el terreno o el mapa. Mediante el uso de un sistema de coordenadas aproximadamente continuas.

Patrón: Cualquier relación existente entre los elementos de una base de datos.

Percepción remota: Acto de detección y/o identificación de un objeto, serie de objetos, o superficies sin tener el sensor en contacto directo con el objeto.

Predio: Propiedad inmueble que se compone de una porción delimitada de terreno.

Procedimiento: Serie de pasos o operaciones que cumplen un propósito y que son realizadas al ejecutar el procedimiento.

Redes neuronales: Es una técnica derivada de la investigación en inteligencia artificial que emplea la regresión generalizada y proporciona un método iterativo para llevarla a cabo.

Reflectancia: Medida de la capacidad de una superficie para reflejar energía electromagnética en una determinada longitud de onda. Es la razón existente entre el flujo reflejado y el incidente sobre dicha superficie.

Reglas de asociación: Técnica de Data Mining que detecta eventos asociados que se ocultan en las bases de datos.

Rendimiento: Se refiere al resultado obtenido de la producción de un cultivo.

Requerimiento: Es una necesidad documentada sobre el contenido, forma o funcionalidad de un producto o servicio.

Restricción: Toda condición que debe cumplir un programa o función dentro del mismo.

Ruta absoluta: Es aquella que parte del directorio raíz. El directorio raíz es el primer directorio en una jerarquía.

Sistema de Información Geográfico: Sistema, generalmente basado en computadoras, para la entrada, almacenaje, recuperación, análisis y muestra de datos geográficos. La base de datos está usualmente compuesta de mapas como representaciones espaciales llamadas capas. Estas capas pueden contener información de un número de atributos incluyendo la topografía del terreno, el uso de la tierra, posición de la tierra, rendimiento de los cultivos, dosis de aplicación de insumos y niveles de nutrientes del suelo.

Sistema de Posicionamiento Global: Red de satélites que son diseñados para ayudar a determinar la posición de un radio receptor en latitud, longitud y altitud. Tecnología que es usada en Agricultura de Precisión.

Software: Se denomina software, programática, equipamiento lógico o soporte lógico a todos los componentes intangibles de una computadora.

Tecnología: Aplicación de los conocimientos científicos a las actividades humanas, con el propósito de hacer más eficiente y eficaz la producción de bienes y servicios.

Tecnología de Dosis Variable: Es la herramienta que permite la implementación de decisiones de manejo en Manejo Sitio Específico. Permitiendo aplicar a cada lote las dosis necesarias de insumos según las características del mismo.

Variabilidad espacial: Es la variación de los elementos en el espacio. En el caso de la agricultura de precisión, uno de los elementos más utilizados para evaluar la variabilidad espacial es el mapa de rendimiento.

Variabilidad temporal: Es el resultado de comparar un determinado número de mapas del mismo terreno a través de los años.

Variable: Es un elemento que puede adquirir o ser sustituido por un valor.

Varianza: Medida estadística que muestra la variabilidad de un valor. A mayor varianza, mayores variaciones con respecto al promedio y en consecuencia, mayor volatilidad.

Zafra: Época, tiempo o período de gran intensidad laboral en que transcurre el desarrollo del cultivo.

Zona: Es una extensión del territorio cuyos límites están determinados por razones administrativas, características del suelo, económicas, políticas, etc.; pero además implica grandes divisiones de la superficie de la tierra.

Zonas de manejo: Áreas de un predio con una combinación única de factores potenciales limitantes del rendimiento. Puede ser delineada y manejada similarmente para optimizar recursos, ingresos o minimizar impactos ambientales.

Anexo 1: Estado del Arte: Agricultura de Precisión.

Se entrega en un documento aparte.

Anexo 2: Estado del Arte: Data Mining

Se entrega en un documento aparte.

Anexo 3: Formato de Archivos

Archivo de Entrada (.ZF)

Ejemplo de un archivo de entrada .ZF

```
[PDT-SIGA: Zone File]
Rows: 3
Cols: 4
CoordX: xxxxx.xxxxxx
CoordY: yyyyyy.yyyyyy
CellSize: 10

[Variables]
VarQty=2
Var1="Texto1"
Var2="Texto2"

[Cells]
valueVar1_11; valueVar2_11
valueVar1_12; valueVar2_12
valueVar1_13; valueVar2_13
valueVar1_14; valueVar2_14
valueVar1_21; valueVar2_21
valueVar1_22; valueVar2_22
valueVar1_23; valueVar2_23
valueVar1_24; valueVar2_24
valueVar1_31; valueVar2_31
valueVar1_32; valueVar2_32
valueVar1_33; valueVar2_33
valueVar1_34; valueVar2_34
```

Archivo de Generación de Zonas (.ZF)

Ejemplo de un archivo de salida .ZF

[PDT-SIGA: Zone File]
Rows: 3
Cols: 4
CoordX: xxxxx.xxxxxx
CoordY: yyyyyy.yyyyyy
CellSize: 10

[Variables]
VarQty=2
Var1="Texto1"
Var2="Texto2"

[Cells]
valueZone_11
valueZone_12
valueZone_13
valueZone_14
valueZone_21
valueZone_22
valueZone_23
valueZone_24
valueZone_31
valueZone_32
valueZone_33
valueZone_34

Archivo de Caracterización de Zonas (.ZAF)

Ejemplo de un archivo de salida .ZAF:

[PDT-SIGA: Zone Attribute File]
ZoneFile: archivo.ZF
ZoneLayer: 1
ZoneQty: 3

[Performance Indexes]
PIndexQty: 2
PIndex1: FPI; 0.333
PIndex2: NCE; 0.434

[Attributes]
AttributeQty: 3

[Attr1: PH]
Attr1:Zone1: valorMax; valorMin; valorMedia; valorDS
Attr1:Zone2: valorMax; valorMin; valorMedia; valorDS
Attr1:Zone3: valorMax; valorMin; valorMedia; valorDS

[Attr2: Conductividad Electrica]
Attr2:Zone1: valorMax; valorMin; valorMedia; valorDS
Attr2:Zone2: valorMax; valorMin; valorMedia; valorDS
Attr2:Zone3: valorMax; valorMin; valorMedia; valorDS

[Attr3: Elevacion]
Attr3:Zone1: valorMax; valorMin; valorMedia; valorDS
Attr3:Zone2: valorMax; valorMin; valorMedia; valorDS
Attr3:Zone3: valorMax; valorMin; valorMedia; valorDS

Archivo de Estadísticas

Ejemplo de archivo de salida conteniendo las estadísticas

[Statistics File]
ZoneFile: archivo.ZF

[Attributes]
AttributeQty: 3

Attr1: PH; valorMax; valorMin; valorMedia; valorDS
Attr1: Conductividad Eléctrica; valorMax; valorMin; valorMedia; valorDS
Attr3: Elevación; valorMax; valorMin; valorMedia; valorDS

[Variance-Covariance Matrix]
value11; valur12; value13
value21; valur22; value23
value31; valur32; value33

[Correlation Matrix]
value11; valur12; value13
value21; valur22; value23
value31; valur32; value33

Anexo 4: Seudo-código

A continuación se presentan las estructuras de datos utilizadas para la implementación:

```
internal class Datos
//Singleton conteniendo la info levantada del archivo de entrada
     private static Datos instance;
     private int numColumnas;
      private int numFilas;
      private string coordenadaX;
      private string coordenadaY;
      private int tamCel;
      private IList<Variable> variables;
      private string archEntrada;


internal class Variable
     private string nombre;
     private double[] datos;
     private double maximo;
     private double minimo;
     private double media;
     private double desviacionEstandar;
     private int tope;


internal class SubZona
     private int zona;
     private ICollection<Elemento> puntos;


internal class Elemento
     private int x;
     private int y;


internal class InfoZona
     public struct Zona_Variable
     {
       public ArrayList valores;
     }
     private int numZona;
     private ArrayList datos_zona; //arreglo de Zona_Variable
```

Negocio

```
CalcularEstadísticas(archivo_entrada, archivo_salida)
Begin
   List<Variable> var = AccesoADatos.CargarDatos(archivo_entrada)
   MC[,] = Calculos.CalcularMatrizCorrelacion(var)
   MVC[,] = Calculos.CalcularMartrizVarianzaCovarianza(var)
   AccesoADatos.GenerarArchivoCalculoEstadisticas(var, MC, MVC,
archivo_salida)
End

GenerarZonas(archivo_entrada, archivo_salida, cant_zonas,
tam_min_zona, distancia)
Begin
   AccesoADatos.CargarDatos(archivo_entrada)
   U[,] = Algoritmo.FuzzyCMeans(c, distancia)
   índices[] = Algoritmo.CalcularIndices(U, cant_zonas)
   vector_zonas[] = Calculos.GenerarVectorZonas(U, cant_zonas)
   AccesoADatos.GenerarArchivoZF(vector_zonas, archivo_salida)
   AccesoADatos.GenerarArchivoZAF(vector_zona, índices, archivo_salida)
   //Post procesamiento
   Z[,] = Calculos.GenerarMatrizZonas(U, cant_zonas)
   PostProcesamiento(Z, cant_zonas, tam_min_zona)
   vector_zonas[] = Calculos.GenerarVectorZonas(Z)
   AccesoADatos.GenerarArchivoZF(vector_zonas, archivo_salida)
End

PostProcesamiento(Z, cant_zonas, tam_min_zona)
Begin
   subZonas = Calculos.ObtenerSubZonasMenorTamMin(Z, tam_min_zona);
   Mientras (subZonas != null)
      SubZona s = Calculos.obtenerSubzonaMenorTam(subZonas);
      Calculos.ProcesarSubZonas(Z, s, c);
      subZonas = Calculos.ObtenerSubZonasMenorTamMin(Z, tam_min_zona);
    Fin Mientras
End
```

Cálculos

```
double[, ] CalcularMatrizVarianzaCovarianza(List<Variable> v)
Begin
  vc[,] = 0 //inicializo matriz de p x p en 0
  For i = 1 to p Begin
    For j = 1 to p Begin
      vc[i,j] = calcular_sp(vi, vj) / (n-1)
    End
   End
   return vc
End

double calcular_sp(v1, v2)
Begin
   sp, s1, s2 = 0
   For i = 1 to n Begin
     sp = sp + (v1[i] * v2[i])
     s1 = s1 + v1[i]
     s2 = s2 + v2[i]
   End
      return (sp – ((s1*s2)/n))
End

double[,] CalcularMatrizCorrelacion(List<Variable> v)
Begin
  mc[,] = 0 // inicializo matriz de p x p en 0
  vc[,] = CalcularMatrizVarianzaCovarianza(v1, …, vp)
  For i = 1 to p
    For j = 1 to p
       mc = vc[i,j] / (vi.getDesvEstandar() * vj.getDesvEstandar())
    End
  End
  return mc
End
```

```
List<SubZona> ObtenerSubZonasMenorTamMin(zonas, tam_min_zona)
//Genera un lista de sub-zonas cuyo tamaño es menor que tam_min_zona
Begin
  List<SubZona> subZonas;
  bool[,] visitados = new bool[filas, cols];
  Inicializo_visitados_en_false()
  For i = 1 to filas
    For j = 1 to cols
      Si (!visitados[i, j])
        visitados[i, j] = true
        SubZona s = new SubZona()
        s.setZona(Z[i, j])
        s.addPunto(i, j)
        ady = obtengoAdyacentes(i, j)
        // adyacentes de [i,j] = [i,j-1]; [i-1,j]; [i,j+1]; [i+1,j];
        [i-1,j-1]; [i-1,j+1]; [i+1,j-1]; [i+1,j+1]
        Mientras (ady != null)
          Elemento e = first(ady)
          Si (Z[e.getX(), e.getY()] == s.getZona())
                s.addPunto(e)
                append(ady, obtengoAdyacentes(e.getX(), e.getY())
                visitados[e.getX(), e.getY()] = true
          Fin Si
        Fin Mientras
        subZonas.Add(s)
      Fin Si
    End
  End

  //me quedo con las subzonas que no cumplen con el tamaño mínimo
  List<SubZona> subZonasNoTamMin
  Para cada SubZona s en subZonas
    tam = s.getCantPuntos() * (tamCelda * tamCelda);
    Si (tam < tam_min_zona) subZonasNoTamMin.Add(s);
  Fin Para

  return subZonasNoTamMin
End


char ObtenerNorma(v1, …, vp)
Begin
  double[,] mvc = CalcularMatrizVarianzaCovarianza(v1, …, vp)
  Si (covarianzas ≈ 0 y varianzas iguales)
    return 'E'
  Sino Si (covarianzas ≈ 0)
    return 'D'
  Sino
    return 'M'
  Fin Si
End
```

```
Bool TieneDatos(índice, v1, …, vp)
Begin
  Para cada Variable v en v1, …, vp
    Si (v[indice] == NaN)
      return false
    Fin Si
  Fin Para
  return true
End
```

```
List<SubZona> ProcesarSubZona(Z[,], s, cant_zonas)
//actualiza la matriz de zonas reasignando la subzona dada (s) a la
zona con mayor borde circundante
Begin
  elemZona = vector de tamaño cant_zonas //cuenta la cantidad de
bordes circundantes a cada zona
  Inicializo elemZona en 0
  Para cada Elemento e de la SubZona s
    adyacentes = obtenerAdyacentes(e)
    //adyacentes de [i,j] = [i,j-1]; [i-1,j]; [i,j+1]; [i+1,j];
    Para cada Elemento ady en adyacentes
      zona = Z[ady.getX(), ady.getY()];
      Si (zona != s.getZona())
        elemZona[zona]++;
      Fin Si
    Fin Para
    zonaCircundante = obtenerZonaMayorBordeCircundante(elemZona);
    //Reasigno elementos a la zona obtenida
    Para cada Elemento e de la SubZona s)
      Z[e.getX(), e.getY()] = zonaCircundante
    Fin Para
  Fin Para
End
```

```
double[] GenerarVectorZonas(U[,], cant_zonas)
//genera un vector de tamaño n con la zona a la que pertenece cada
punto a partir de la matriz de pertenencia
Begin
  maxIndice = 0;
  maxValor = 0;
  For i = 1 to n Begin
    maxIndice = 0;
    maxValor = U[i, 0];
    For j = 1 to cant_zonas Begin
      Si (U[i, j] > maxValor)
        maxValor = U[i, j];
        maxIndice = j;
      Fin Si
    End
    vector[i] = maxIndice;
  End
  return vector;
End
```

Algoritmo

```
double[,] FuzzyCMeans(cant_zonas, distancia)
//devuelve la matriz de pertenencia
Begin
  Obtengo parámetros: max_iter, exp_fuzzy, stop
  p = cantidad de variables
  n = cantidad datos = cantidad de elementos de una variable)
  Si not (distancia == 'E' or distancia == 'D')
       Distancia = Calculos.ObtenerNorma()

  U = new double[n, c]; //U = Matriz de pertenencia
  Inicializo matriz de pertenencia con valores aleatorios entre 0 y 1
  V = new double[c, p]; //V = Matriz de Centroides
  D = new double[n, c]; //D = Matriz de Distancias
  Inicializo D y V con 0

  V = CalcularCentroides(U);
  D = CalcularDistancias(V, distancia);
  U_actual = CalcularMatrizDePertenencia(D);
  iter = 0;
  Mientras ((iter < max_iter) and (|U_actual - U| > stop))
    U = U_actual;
    V = CalcularCentroides(U);
    D = CalcularDistancias(V, distancia);
    U_actual = CalcularMatrizDePertenencia(D);
    iter++;
  Fin Mientras
  return U_actual;
End
```

```
double[,] CalcularCentroides(U, V)
Begin
  For i = 1 to c Begin
    For j = 1 to p Begin
      s1 = 0
      s2 = 0
      For k = 1 to n Begin
        s1 = s1 + U[k,i]^m * v_j[k]
        s2 = s2 + U[k,i]^m
       End
       V[i,j] = s1/s2
    End
  End
  return V
End
```

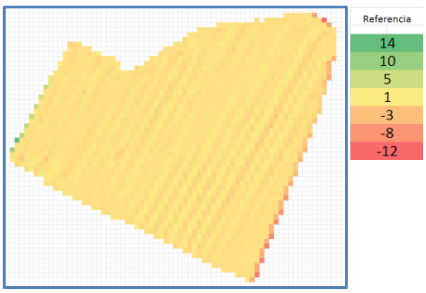
```
double[,] CalcularMatrizDePertenencia(D)
Begin
  For i = 1 to n Begin
    For j = 1 to c Begin
      s = 0;
      For k = 1 to c Begin
        s = s + (D[i,j]/D[i,k])^2/(m-1)
      End
      U[i,j] = s^-1
    End
  End
  return U
End


double[,] CalcularDistancias(V, distancia)
Begin
  Si (distancia = 'E') //Euclideana
    For i = 1 to n Begin
      For j = 1 to c Begin
        d = 0
        For k = 1 to p Begin
          d = d + (v_k[i] - V[j,k])^2
        End
        D[i,j] = √d
    End
  Sino Si distancia = 'D') //Diagonal
    For i = 1 to n Begin
      For j = 1 to c Begin
        d = 0
        For k = 1 to p Begin
          A[k] = V[j,k]
        }
        d = A * matriz_diagonal * A^T
        D[i,j] = √d
      End
    End
  Sino // Mahalanobis
    ∑^-1 = inversa(CalcularMatrizVarianzaCovarianza(v_1, …, v_p))
    For i = 1 to n Begin
      For j = 1 to c Begin
        a = 0
        For k = 1 to p Begin
          a[k] = (v_k[i]-V[j,k])
        End
      End
      D[i,j] = √(a * ∑^-1 * a^T)
    End
  Fin Si
  return D
End
```

```
double CalcularIndices(U[,], cant_zonas)
Begin
   f = 0
   For i = 1 to n
      For j = 1 to cant_zonas Begin
         f = f + U[i,j]^2
      End
   End
   f = f / n
   FPI = 1 - cant_zonas/(cant_zonas-1)[1 - f]

   h = 0
   For i = 1 to n Begin
      For j = 1 to cant_zonas Begin
         h = h + U[i,j]*log_a U[i,j]
       End
    End
    //El valor de a utilizado fue 2
    h = h / n
    NCE = h / (1-(cant_zonas/n))

    return FPI, NCE
End
```

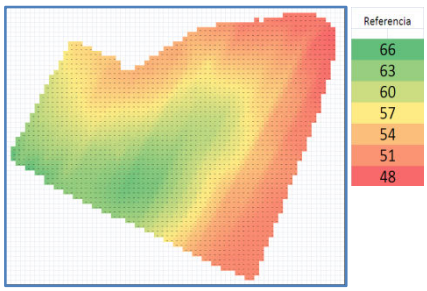
Anexo 5: Datos de prueba

Potrero 1

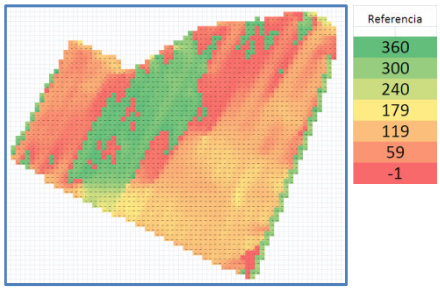
Curvatura



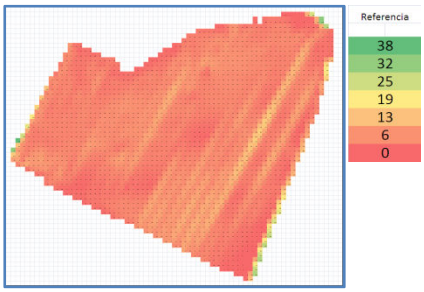
Elevación



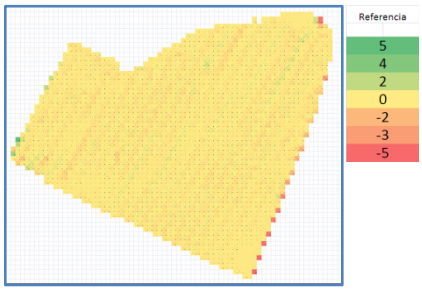
Orientación



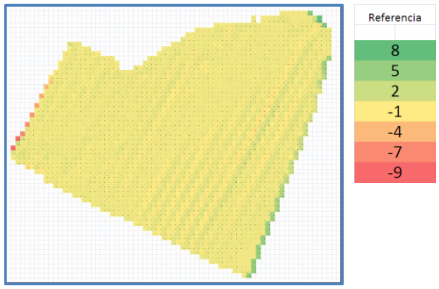
Pendiente



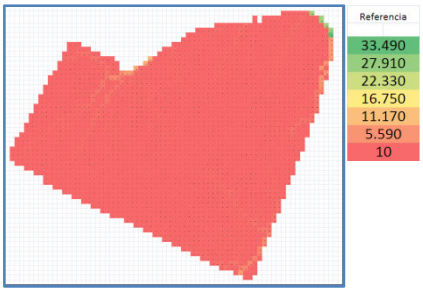
Plano de curvatura



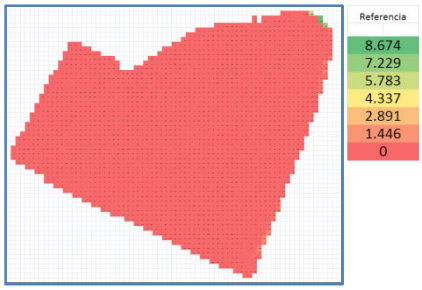
Perfil de Curvatura



Área específica de cuenca

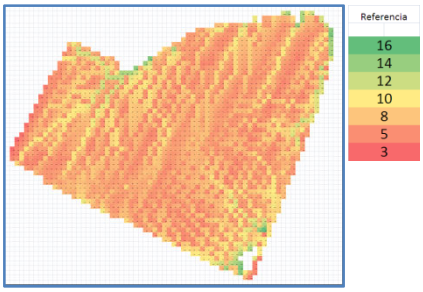


Índice Topográfico SPI

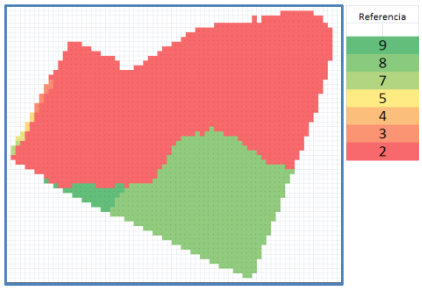


Índice Topográfico STCI

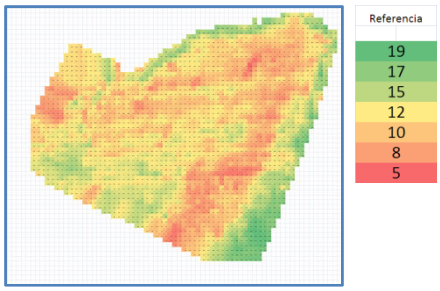
Indice topográfico WI



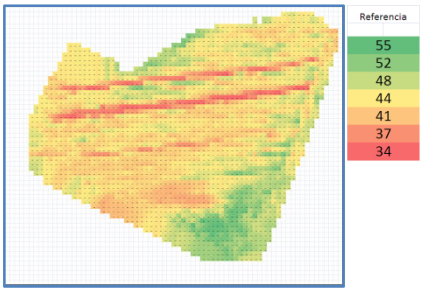
Cuenca



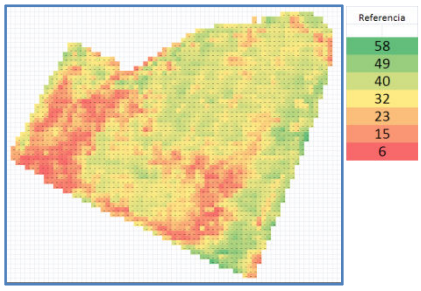
Rendimiento normalizado 2006 trigo



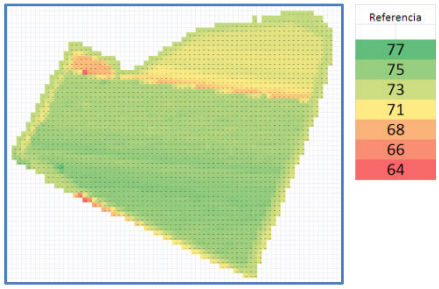
Humedad normalizada 2006 trigo



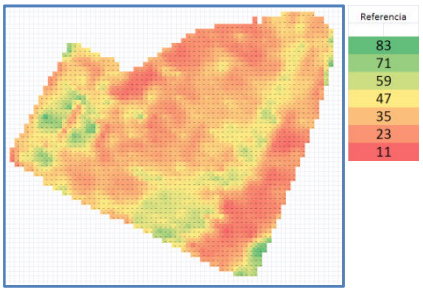
Rendimiento normalizado 2007 maíz



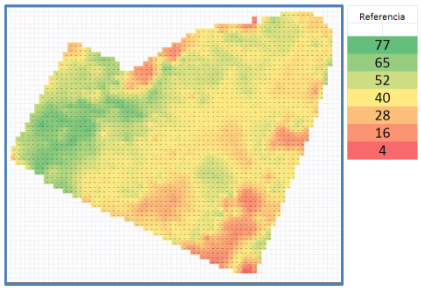
Humedad normalizada 2007 maíz



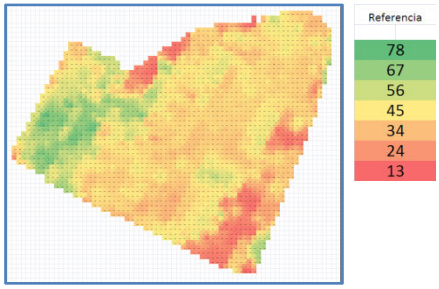
CE a 30cm



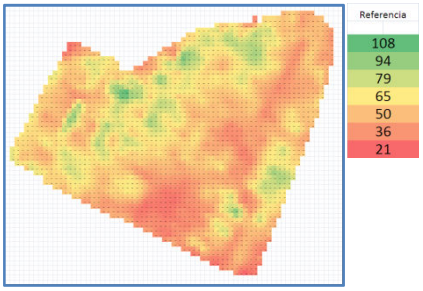
CE de 30 a 90 cm



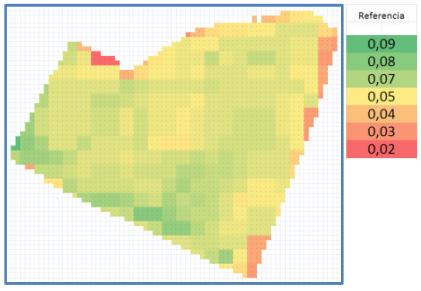
CE a 90 cm



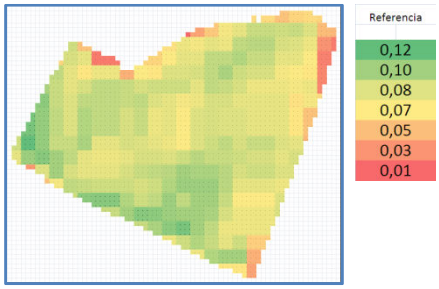
CE Relación 90-30



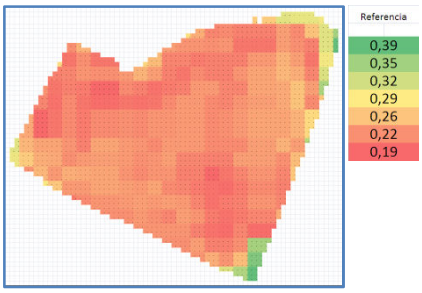
Reflectancia banda 2



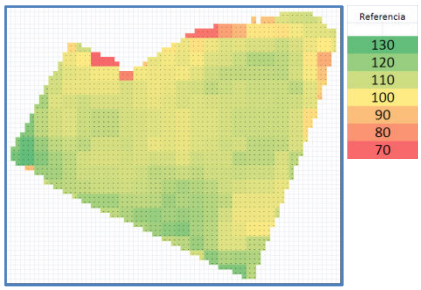
Reflectancia banda 3



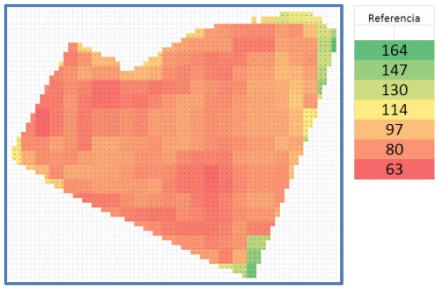
Reflectancia banda 4



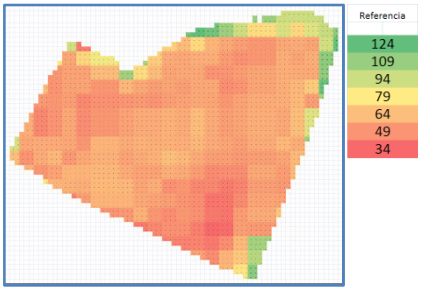
Tasseled cap banda 1



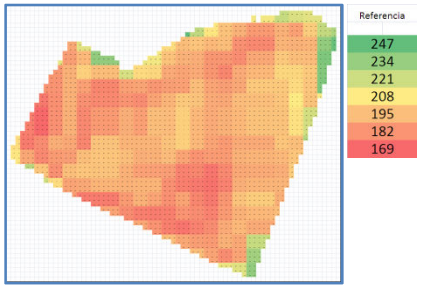
Tasseled cap banda 2



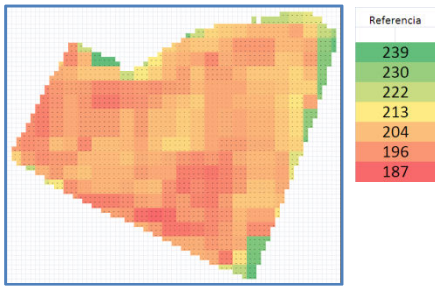
Tasseled cap banda 3



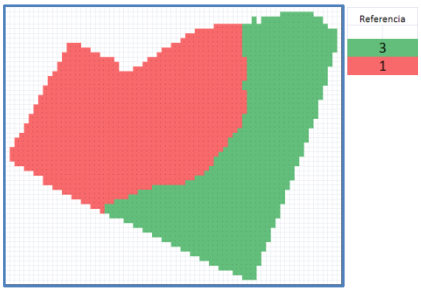
Índice NDVI



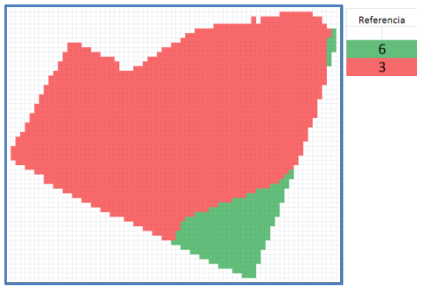
Índice GNDVI



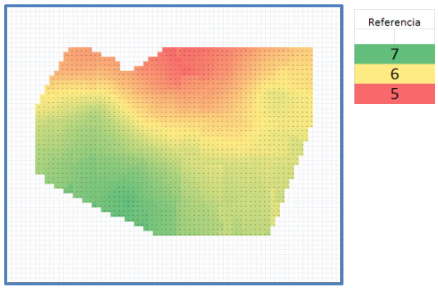
Índice CONEAT



Drenaje

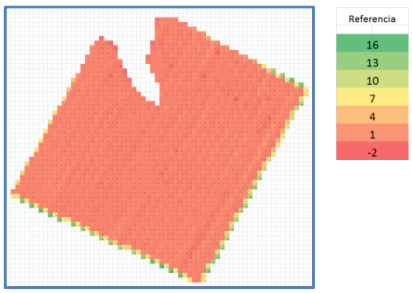


Muestreo de Pbray

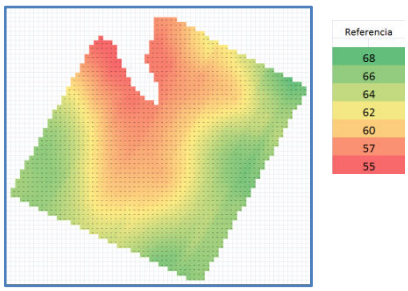


Potrero 6

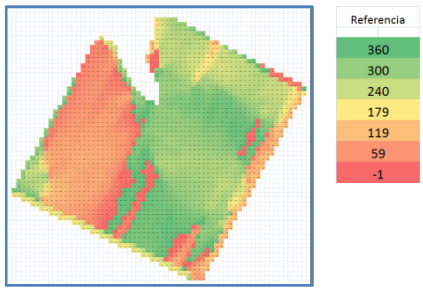
Curvatura



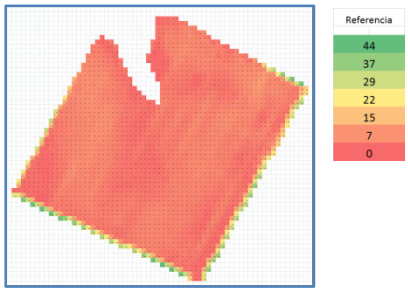
Elevación



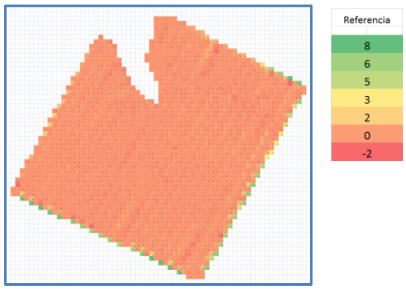
Orientación



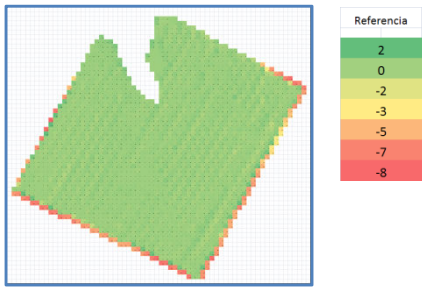
Pendiente



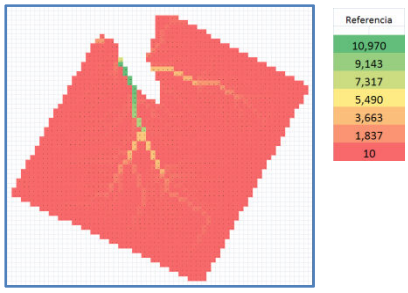
Plano de curvatura



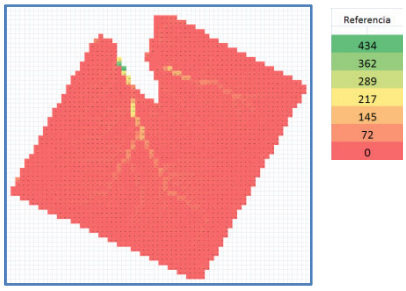
Perfil de Curvatura



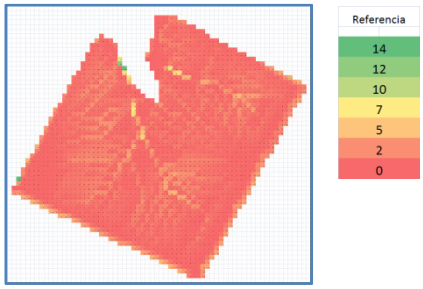
Área específica de cuenca



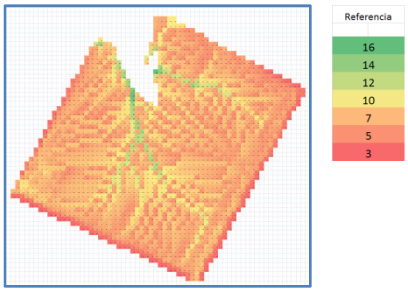
Índice Topográfico SPI



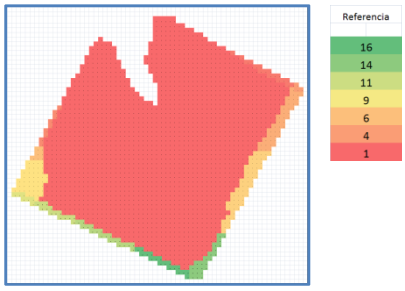
Índice Topográfico STCI



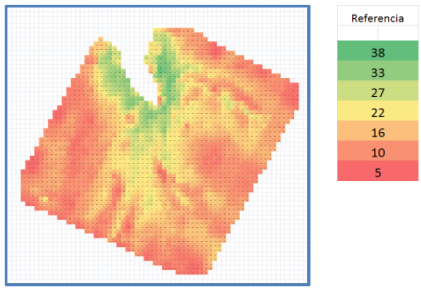
Indice topográfico WI



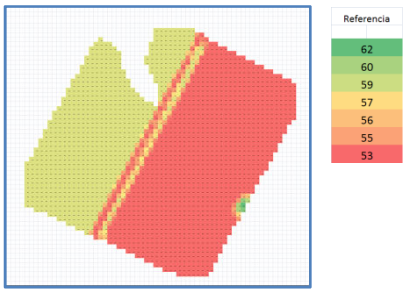
Cuenca



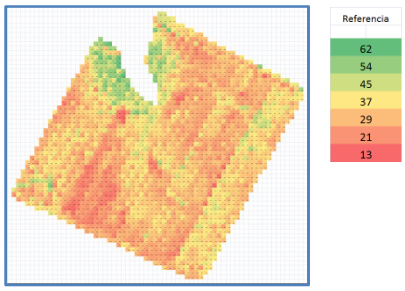
Rendimiento normalizado 2006 maíz



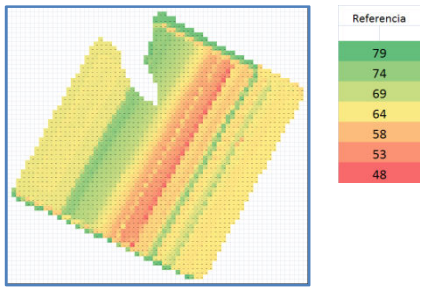
Humedad normalizada 2006 maíz



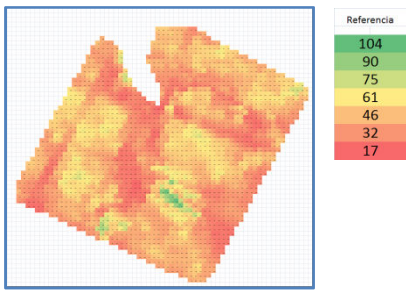
Rendimiento normalizado 2007 soja



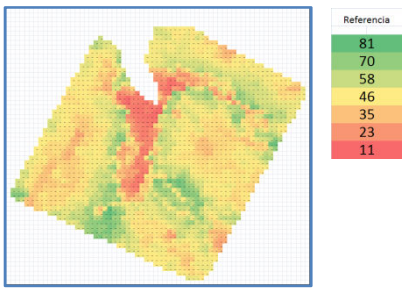
Humedad normalizada 2007 soja



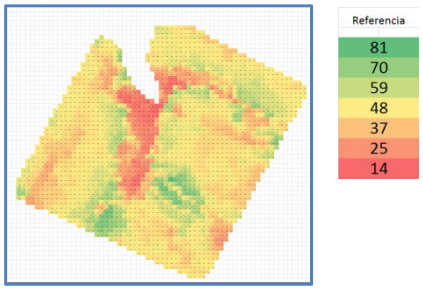
CE a 30cm



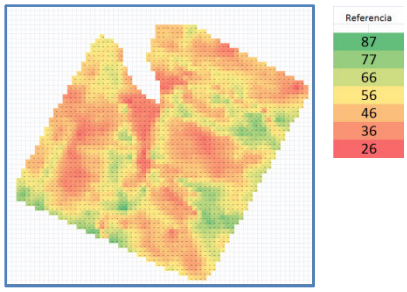
CE de 30 a 90 cm



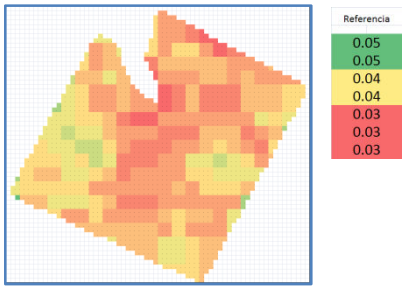
CE a 90 cm



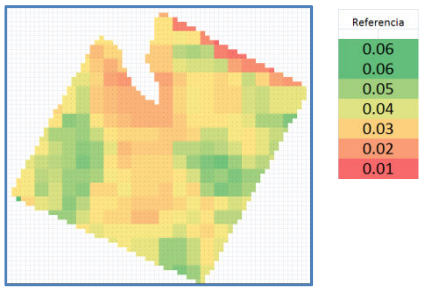
CE Relación 90-30



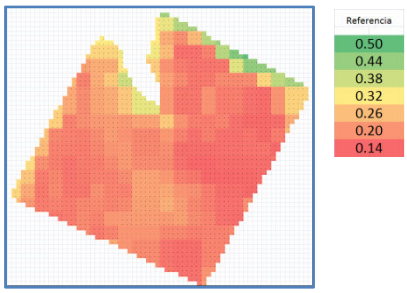
Reflectancia banda 2



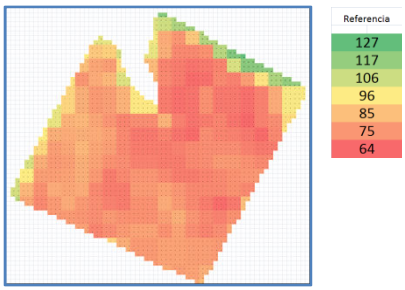
Reflectancia banda 3



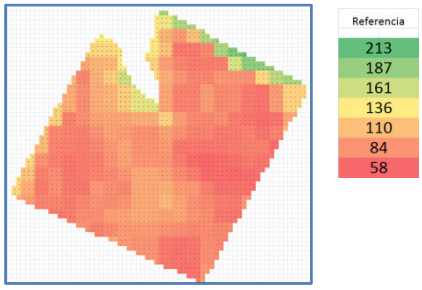
Reflectancia banda 4



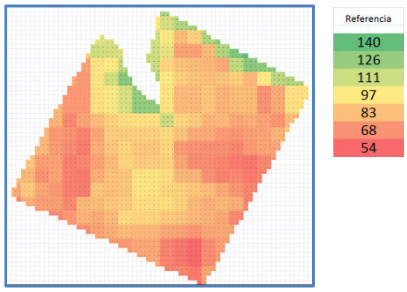
Tasseled cap banda 1



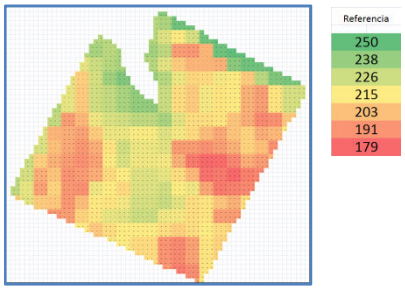
Tasseled cap banda 2



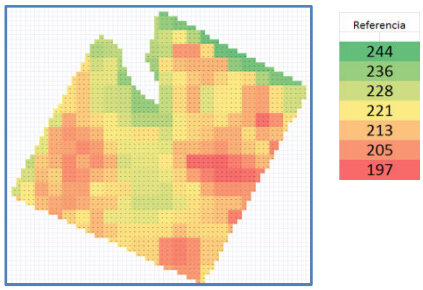
Tasseled cap banda 3



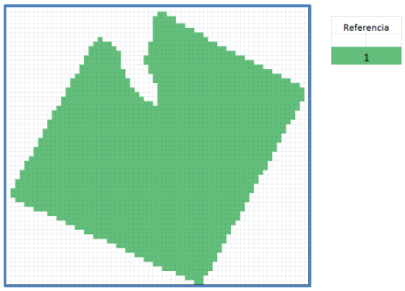
Índice NDVI



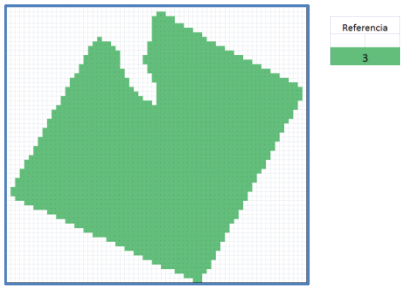
Índice GNDVI



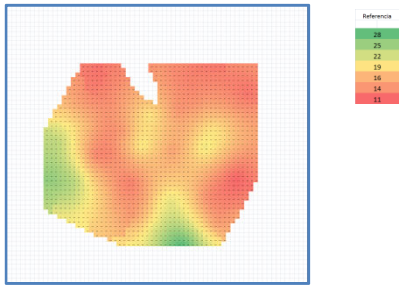
Índice CONEAT



Drenaje



Muestreo de Pbray



Anexo 6: Análisis Potrero 6

En esta sección se presenta un análisis de las pruebas realizadas para el Potrero 6, el cual pretende demostrar que los resultados obtenidos son coherentes con los datos de entrada.

Las variables utilizadas son:

Nº	Variable	Descripción
1	a_curvature(Banda 1)	Curvatura
2	a_dem_corr(Banda 1)	Elevación
3	a_orientac(Banda 1)	Orientación
4	a_pendiente(Banda 1)	Pendiente
5	a_plan(Banda 1)	Plano de curvatura
6	a_profile(Banda 1)	Perfil de curvatura
7	a_sca(Banda 1)	Área específica de cuenca
8	a_spi(Banda 1)	Índice topográfico SPI
9	a_stci(Banda 1)	Índice topográfico STCI
10	a_wi(Banda 1)	Índice topográfico WI
11	a_basin(Banda 1)	Cuenca
12	maizrend(Banda 1)	Rendimiento normalizado 2006 maíz
13	humrend(Banda 1)	Humedad normalizada 2006 maíz
14	sojarend(Banda 1)	Rendimiento normalizado 2007 soja
15	sojahum(Banda 1)	Humedad normalizada 2007 soja
16	ec30(Banda 1)	CE a 30 cm
17	ec30a90(Banda 1)	CE de 30 a 90 cm
18	ec90(Banda 1)	CE a 90 cm
19	ec9030(Banda 1)	CE relación 90-30
20	20060127_reflect.img - Layer_2(Banda 1)	Reflectancia banda 2
21	20060127_reflect.img - Layer_3(Banda 1)	Reflectancia banda 3
22	20060127_reflect.img - Layer_4(Banda 1)	Reflectancia banda 4
23	20060127_reflect_tc_1a3.img - Layer_1(Banda 1)	Tasseled cap banda 1
24	20060127_reflect_tc_1a3.img - Layer_2(Banda 1)	Tasseled cap banda 2
25	20060127_reflect_tc_1a3.img - Layer_3(Banda 1)	Tasseled cap banda 3
26	20060127_ndvi_rec.img(Banda 1)	Índice NDVI
27	20060127_gndvi_rec.img(Banda 1)	Índice GNDVI
28	raster_cone1(Banda 1)	Índice CONEAT
29	drenaje_Clip.img(Banda 1)	Drenaje
30	pbray_pot6(Banda 1)	Muestreo de PBray (set) 2007

Figura 19: Variables Potrero 6.

Primero se realizó el cálculo de estadísticas.

En la Figura 20 se muestra la matriz de correlación coloreada según el siguiente criterio:

	1 < r < 0,7	Existe correlación positiva
	-1 < r < -0,7	Existe correlación negativa

VAR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	1	0,27	-0,08	0,92	0,84	-0,94	-0,08	-0,07	0,38	-0,48	0,57	-0,20	-0,03	0,01	0,07	-0,10	0,04	0,00	0,16	0,10	0,07	0,05	0,14	0,03	-0,08	-0,03	-0,02	NeuN	Inf	0,03
2	0,27	1	0,03	0,27	0,20	-0,27	-0,26	-0,23	-0,09	-0,42	0,37	-0,72	-0,34	-0,26	-0,10	0,09	0,14	0,15	0,13	0,43	0,57	-0,30	0,05	-0,33	-0,61	-0,53	-0,51	NeuN	-Inf	0,39
3	-0,08	0,03	1	-0,07	-0,07	0,07	0,05	0,05	0,00	0,04	-0,06	0,18	-0,51	0,00	-0,27	-0,02	-0,01	-0,01	0,01	-0,30	-0,12	-0,04	-0,14	-0,01	0,06	0,03	0,07	NeuN	Inf	-0,02
4	0,92	0,27	-0,07	1	0,73	-0,88	-0,10	-0,04	0,49	-0,52	0,60	-0,26	-0,02	-0,04	0,04	0,00	0,10	0,08	0,13	0,15	0,11	0,04	0,15	0,02	-0,11	-0,05	-0,05	NeuN	Inf	0,00
5	0,84	0,20	-0,07	0,73	1	-0,60	-0,08	-0,07	0,27	-0,39	0,43	-0,16	-0,04	0,02	0,10	-0,08	0,02	-0,01	0,13	0,07	0,06	0,03	0,10	0,02	-0,06	-0,03	-0,03	NeuN	Inf	0,02
6	-0,94	-0,27	0,07	-0,88	-0,60	1	0,07	0,05	-0,38	0,45	-0,56	0,20	0,03	-0,01	-0,05	0,09	-0,05	-0,01	-0,15	-0,10	-0,07	-0,05	-0,14	-0,04	0,08	0,03	0,02	NeuN	-Inf	-0,04
7	-0,08	-0,26	0,05	-0,10	-0,08	0,07	1	0,85	0,51	0,57	-0,07	0,24	0,12	0,19	0,08	-0,04	-0,12	-0,12	-0,09	-0,11	-0,18	0,16	0,06	0,16	0,23	0,18	0,18	NeuN	NeuN	-0,07
8	-0,07	-0,23	0,05	-0,04	-0,07	0,05	0,85	1	0,72	0,48	-0,05	0,20	0,10	0,17	0,05	0,01	-0,05	-0,04	-0,07	-0,08	-0,15	0,12	0,04	0,13	0,19	0,15	0,14	NeuN	-Inf	-0,08
9	0,38	-0,09	0,00	0,49	0,27	-0,38	0,51	0,72	1	0,27	0,22	0,03	0,07	0,08	0,05	0,06	0,06	0,07	0,01	0,03	-0,06	0,07	0,05	0,07	0,08	0,08	0,07	NeuN	-Inf	-0,10
10	-0,48	-0,42	0,04	-0,52	-0,39	0,45	0,57	0,48	0,27	1	-0,35	0,37	0,12	0,14	0,03	-0,07	-0,12	-0,12	-0,07	-0,17	-0,20	0,03	-0,13	0,05	0,22	0,16	0,14	NeuN	-Inf	-0,09
11	0,57	0,37	-0,06	0,60	0,43	-0,56	-0,07	-0,05	0,22	-0,35	1	-0,27	-0,03	-0,02	0,03	-0,15	0,06	0,00	0,22	0,14	0,18	-0,06	0,07	-0,08	-0,20	-0,13	-0,11	NeuN	NeuN	0,01
12	-0,20	-0,72	0,18	-0,26	-0,16	0,20	0,24	0,20	0,03	0,37	-0,27	1	0,12	0,38	0,02	-0,28	-0,22	-0,27	0,00	-0,44	-0,51	0,15	-0,19	0,19	0,49	0,41	0,38	Inf	-Inf	-0,22
13	-0,03	-0,34	-0,51	-0,02	-0,04	0,03	0,12	0,10	0,07	0,12	-0,03	0,12	1	0,15	0,50	-0,04	-0,16	-0,15	-0,14	0,25	-0,03	0,20	0,22	0,17	0,13	0,18	0,13	NeuN	Inf	0,13
14	0,01	-0,26	0,00	-0,04	0,02	-0,01	0,19	0,17	0,08	0,14	-0,02	0,38	0,15	1	0,04	-0,26	-0,29	-0,31	0,00	-0,10	-0,27	0,30	0,20	0,31	0,35	0,30	0,29	NeuN	Inf	-0,11
15	0,07	-0,10	-0,27	0,04	0,10	-0,05	0,08	0,05	0,05	0,03	0,03	0,02	0,50	0,04	1	-0,08	-0,09	-0,11	-0,03	0,16	0,01	0,22	0,28	0,20	0,12	0,08	0,07	NeuN	Inf	0,12
16	-0,10	0,09	-0,02	0,00	-0,08	0,09	-0,04	0,01	0,06	-0,07	-0,15	-0,28	-0,04	-0,26	-0,08	1	0,42	0,67	-0,55	0,17	0,14	0,05	0,15	0,03	-0,08	-0,07	-0,05	NeuN	NeuN	0,18
17	0,04	0,14	-0,01	0,10	0,02	-0,05	-0,12	-0,05	0,06	-0,12	0,06	-0,22	-0,16	-0,29	-0,09	0,42	1	0,93	0,44	0,08	0,08	-0,09	-0,05	-0,09	-0,11	-0,07	-0,09	NeuN	Inf	-0,04
18	0,00	0,15	-0,01	0,08	-0,01	-0,01	-0,12	-0,04	0,07	-0,12	0,00	-0,27	-0,15	-0,31	-0,11	0,67	0,93	1	0,19	0,12	0,10	-0,06	0,00	-0,07	-0,11	-0,08	-0,09	NeuN	Inf	0,02
19	0,16	0,13	0,01	0,13	0,13	-0,15	-0,09	-0,07	0,01	-0,07	0,22	0,00	-0,14	0,00	-0,03	-0,55	0,44	0,19	1	-0,07	-0,02	-0,17	-0,22	-0,16	-0,08	-0,05	-0,08	NeuN	Inf	-0,20
20	0,10	0,43	-0,30	0,15	0,07	-0,10	-0,11	-0,08	0,03	-0,17	0,14	-0,44	0,25	-0,10	0,16	0,17	0,08	0,12	-0,07	1	0,68	-0,21	0,17	-0,28	-0,50	-0,51	-0,63	NeuN	Inf	0,24
21	0,07	0,57	-0,12	0,11	0,06	-0,07	-0,18	-0,15	-0,06	-0,20	0,18	-0,51	-0,03	-0,27	0,01	0,14	0,08	0,10	-0,02	0,68	1	-0,67	-0,28	-0,73	-0,82	-0,93	-0,87	NeuN	-Inf	0,25
22	0,05	-0,30	-0,04	0,04	0,03	-0,05	0,16	0,12	0,07	0,03	-0,06	0,15	0,20	0,30	0,22	0,05	-0,09	-0,06	-0,17	-0,21	-0,67	1	0,88	1,00	0,85	0,84	0,84	NeuN	-Inf	-0,07
23	0,14	0,05	-0,14	0,15	0,10	-0,14	0,06	0,04	0,05	-0,13	0,07	-0,19	0,22	0,20	0,28	0,15	-0,05	0,00	-0,22	0,17	-0,28	0,88	1	0,84	0,52	0,54	0,55	NeuN	NeuN	0,12
24	0,03	-0,33	-0,01	0,02	0,02	-0,04	0,16	0,13	0,07	0,05	-0,08	0,19	0,17	0,31	0,20	0,03	-0,09	-0,07	-0,16	-0,28	-0,73	1,00	0,84	1	0,87	0,88	0,88	NeuN	NeuN	-0,10
25	-0,08	-0,61	0,06	-0,11	-0,06	0,08	0,23	0,19	0,08	0,22	-0,20	0,49	0,13	0,35	0,12	-0,08	-0,11	-0,11	-0,08	-0,50	-0,82	0,85	0,52	0,87	1	0,89	0,89	NeuN	NeuN	-0,28
26	-0,03	-0,53	0,03	-0,05	-0,03	0,03	0,18	0,15	0,08	0,16	-0,13	0,41	0,18	0,30	0,08	-0,07	-0,07	-0,08	-0,05	-0,51	-0,93	0,84	0,54	0,88	0,89	1	0,96	NeuN	NeuN	-0,16
27	-0,02	-0,51	0,07	-0,05	-0,03	0,02	0,18	0,14	0,07	0,14	-0,11	0,38	0,13	0,29	0,07	-0,05	-0,09	-0,09	-0,08	-0,63	-0,87	0,84	0,55	0,88	0,89	0,96	1	NeuN	NeuN	-0,15
28	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	Inf	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	-Inf
29	Inf	-Inf	Inf	Inf	Inf	-Inf	NeuN	-Inf	-Inf	-Inf	NeuN	-Inf	Inf	Inf	Inf	NeuN	Inf	Inf	Inf	Inf	-Inf	-Inf	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	NeuN	-Inf
30	0,03	0,39	-0,02	0,00	0,02	-0,04	-0,07	-0,08	-0,10	-0,09	0,01	-0,22	0,13	-0,11	0,12	0,18	-0,04	0,02	-0,20	0,24	0,25	-0,07	0,12	-0,10	-0,28	-0,16	-0,15	-Inf	-Inf	1

Figura 20: Matriz de correlación para el Potrero 6.

Por lo tanto, podemos ver que las variables que tienen una correlación positiva para el criterio tomado son:

1	Curvatura	Y	4	Pendiente
1	Curvatura	Y	5	Plano de curvatura
4	Pendiente	Y	5	Plano de curvatura
7	Área específica de cuenca	Y	8	Índice topográfico SPI
8	Índice topográfico SPI	Y	9	Índice topográfico STCI
17	CE de 30 a 90 cm	Y	18	CE a 90 cm
22	Reflectancia banda 4	Y	23	Tasseled cap banda 1
22	Reflectancia banda 4	Y	24	Tasseled cap banda 2
22	Reflectancia banda 4	Y	25	Tasseled cap banda 3
22	Reflectancia banda 4	Y	26	Índice NDVI
22	Reflectancia banda 4	Y	27	Índice GNDVI
23	Tasseled cap banda 1	Y	24	Tasseled cap banda 2
24	Tasseled cap banda 2	Y	25	Tasseled cap banda 3
24	Tasseled cap banda 2	Y	26	Índice NDVI

24	Tasseled cap banda 2	Y	27	Índice GNDVI
25	Tasseled cap banda 3	Y	26	Índice NDVI
25	Tasseled cap banda 3	Y	27	Índice GNDVI
26	Índice NDVI	Y	27	Índice GNDVI

Figura 21: Variables con correlación positiva para el Potrero 6

Y las variables que tienen una correlación negativa son:

1	Curvatura	Y	6	Perfil de curvatura
2	Elevación	Y	12	Rendimiento normalizado 2006 maíz
4	Pendiente	Y	6	Perfil de curvatura
21	Reflectancia banda 3	Y	24	Tasseled cap banda 2
21	Reflectancia banda 3	Y	25	Tasseled cap banda 3
21	Reflectancia banda 3	Y	26	Índice NDVI
21	Reflectancia banda 3	Y	27	Índice GNDVI

Figura 22: Variables con correlación negativa para el Potrero 6.

Estos resultados son coherentes con las imágenes que se presentan de las variables (*ver: Anexo 5: Datos de prueba*). Es decir, para las variables que presentan una correlación positiva, puede observarse que los colores de una variable se corresponden con la otra, ambas tienen valores altos o bajos en los mismos puntos. Para las variables que tienen una correlación negativa ocurre lo inverso, cuando una variable tiene un color, la otra variable tiene el color opuesto para los mismos puntos.

Ahora analizaremos la matriz de varianza-covarianza para establecer la medida de similitud a utilizar.

A continuación se muestran los resultados obtenidos:

4,29	1,69	-18,35	10,79	1,57	-2,72	-157,05	-2,87	0,80	-1,69	3,21	-2,75	-0,17	0,21	0,83	-2,38	0,86	-0,03	3,53	0,00	0,00	0,01	2,96	1,74	-2,55	-0,98	-0,45	0,00	0,00	0,22
1,69	9,30	10,20	4,75	0,56	-1,13	-739,28	-14,73	-0,29	-2,18	3,08	-14,43	-2,50	-6,12	-1,71	3,33	4,90	4,58	4,04	0,01	0,02	-0,06	1,46	-25,22	-28,07	-24,67	-14,65	0,00	0,00	3,98
-18,35	10,20	13.444,68	-43,56	-6,90	11,47	5.880,81	124,02	0,42	7,58	-18,34	137,43	-140,37	-0,53	-166,21	-28,50	-12,06	-14,30	13,28	-0,15	-0,14	-0,30	-171,25	-37,84	112,34	47,62	75,22	0,00	0,00	-7,98
10,79	4,75	-43,56	32,46	3,76	-7,03	-534,50	-5,16	2,86	-5,11	9,40	-9,86	-0,30	-1,80	1,35	-0,24	6,44	4,71	7,97	0,00	0,01	0,01	8,80	2,92	-9,50	-4,51	-2,43	0,00	0,00	0,01
1,57	0,56	-6,90	3,76	0,82	-0,76	-66,59	-1,34	0,25	-0,61	1,07	-0,95	-0,08	0,11	0,49	-0,92	0,15	-0,12	1,26	0,00	0,00	0,00	0,92	0,50	-0,81	-0,46	-0,21	0,00	0,00	0,05
-2,72	-1,13	11,47	-7,03	-0,76	1,96	90,43	1,53	-0,55	1,08	-2,14	1,80	0,09	-0,10	-0,34	1,46	-0,71	-0,10	-2,27	0,00	0,00	0,00	-2,04	-1,23	1,74	0,52	0,25	0,00	0,00	-0,17
-157,05	-739,28	5.880,81	-534,50	-66,59	90,43	890.287,34	16.841,30	499,01	914,11	-169,96	1.511,06	274,44	1.351,24	419,23	-499,36	-1.256,68	-1.129,34	-924,75	-0,45	-1,54	9,04	597,64	3.802,84	3.327,71	2.550,60	1.575,44	0,00	0,00	-225,18
-2,87	-14,73	124,02	-5,16	-1,34	1,53	16.841,30	444,02	15,65	17,49	-3,08	27,98	4,98	26,86	5,16	2,60	-11,75	-7,79	-14,89	-0,01	-0,03	0,16	8,70	65,58	59,86	47,25	27,74	0,00	0,00	-5,72
0,80	-0,29	0,42	2,86	0,25	-0,55	499,01	15,65	1,06	0,47	0,63	0,23	0,17	0,60	0,28	0,68	0,71	0,74	0,14	0,00	0,00	0,01	0,56	1,80	1,24	1,26	0,64	0,00	0,00	-0,35
-1,69	-2,18	7,58	-5,11	-0,61	1,08	914,11	17,49	0,47	2,95	-1,63	4,11	0,49	1,89	0,30	-1,47	-2,29	-2,18	-1,29	0,00	0,00	0,00	-2,37	2,07	5,71	4,17	2,22	0,00	0,00	-0,52
3,21	3,08	-18,34	9,40	1,07	-2,14	-169,96	-3,08	0,63	-1,63	7,45	-4,78	-0,21	-0,41	0,45	-4,80	1,89	-0,01	6,47	0,00	0,00	-0,01	2,10	-5,08	-8,32	-5,35	-2,82	0,00	0,00	0,13
-2,75	-14,43	137,43	-9,86	-0,95	1,80	1.511,06	27,98	0,23	4,11	-4,78	42,91	1,94	19,10	0,65	-22,32	-16,37	-18,01	0,15	-0,01	-0,03	0,06	-13,22	30,66	48,27	40,64	23,39	0,00	0,00	-4,85
-0,17	-2,50	-140,37	-0,30	-0,08	0,09	274,44	4,98	0,17	0,49	-0,21	1,94	5,75	2,77	6,42	-1,18	-4,33	-3,67	-3,48	0,00	0,00	0,03	5,58	10,04	4,86	6,45	2,90	0,00	0,00	1,02
0,21	-6,12	-0,53	-1,80	0,11	-0,10	1.351,24	26,86	0,60	1,89	-0,41	19,10	2,77	59,01	1,68	-23,67	-25,07	-24,96	-0,31	0,00	-0,02	0,14	16,11	58,18	40,23	35,12	21,03	0,00	0,00	-2,78
0,83	-1,71	-166,21	1,35	0,49	-0,34	419,23	5,16	0,28	0,30	0,45	0,65	6,42	1,68	29,24	-5,45	-5,54	-5,94	-1,86	0,00	0,00	0,07	15,46	26,16	9,40	6,45	3,57	0,00	0,00	2,09
-2,38	3,33	-28,50	-0,24	-0,92	1,46	-499,36	2,60	0,68	-1,47	-4,80	-22,32	-1,18	-23,67	-5,45	144,12	57,23	82,83	-70,26	0,01	0,02	0,03	19,01	9,15	-13,72	-11,96	-6,08	0,00	0,00	7,33
0,86	4,90	-12,06	6,44	0,15	-0,71	-1.256,68	-11,75	0,71	-2,29	1,89	-16,37	-4,33	-25,07	-5,54	57,23	129,42	109,61	52,52	0,00	0,01	-0,06	-5,41	-25,42	-18,12	-12,63	-9,80	0,00	0,00	-1,60
-0,03	4,58	-14,30	4,71	-0,12	-0,10	-1.129,34	-7,79	0,74	-2,18	-0,01	-18,01	-3,67	-24,96	-5,94	82,83	109,61	107,03	20,73	0,01	0,01	-0,04	-0,12	-17,18	-17,27	-13,04	-8,98	0,00	0,00	0,71
3,53	4,04	13,28	7,97	1,26	-2,27	-924,75	-14,89	0,14	-1,29	6,47	0,15	-3,48	-0,31	-1,86	-70,26	52,52	20,73	111,96	0,00	0,00	-0,11	-23,81	-40,88	-13,35	-8,55	-7,65	0,00	0,00	-6,96
0,00	0,01	-0,15	0,00	0,00	0,00	-0,45	-0,01	0,00	0,00	0,00	-0,01	0,00	0,00	0,00	0,01	0,00	0,01	0,00	0,00	0,00	0,00	0,01	-0,03	-0,03	-0,03	-0,03	0,00	0,00	0,00
0,00	0,02	-0,14	0,01	0,00	0,00	-1,54	-0,03	0,00	0,00	0,00	-0,03	0,00	-0,02	0,00	0,02	0,01	0,01	0,00	0,00	0,00	0,00	-0,03	-0,17	-0,12	-0,13	-0,08	0,00	0,00	0,01
0,01	-0,06	-0,30	0,01	0,00	0,00	9,04	0,16	0,01	0,00	-0,01	0,06	0,03	0,14	0,07	0,03	-0,06	-0,04	-0,11	0,00	0,00	0,00	0,56	1,49	0,77	0,77	0,48	0,00	0,00	-0,01
2,96	1,46	-171,25	8,80	0,92	-2,04	597,64	8,70	0,56	-2,37	2,10	-13,22	5,58	16,11	15,46	19,01	-6,41	-0,12	-23,81	0,01	-0,03	0,56	108,49	217,51	80,94	84,68	54,05	0,00	0,00	3,96
1,74	-25,22	-37,84	2,92	0,50	-1,23	3.802,84	65,58	1,80	2,07	-5,08	30,66	10,04	58,18	26,16	9,15	-25,42	-17,18	-40,88	-0,03	-0,17	1,49	217,51	613,92	326,38	328,87	205,53	0,00	0,00	-8,08
-2,55	-28,07	112,34	-9,50	-0,81	1,74	3.327,71	59,86	1,24	5,71	-8,32	48,27	4,86	40,23	9,40	-13,72	-18,12	-17,27	-13,35	-0,03	-0,12	0,77	80,94	326,38	226,92	201,95	127,07	0,00	0,00	-14,19
-0,98	-24,67	47,62	-4,51	-0,46	0,52	2.550,60	47,25	1,26	4,17	-5,35	40,64	6,45	35,12	6,45	-11,96	-12,63	-13,04	-8,55	-0,03	-0,13	0,77	84,68	328,87	201,95	229,54	137,61	0,00	0,00	-8,02
-0,45	-14,65	75,22	-2,43	-0,21	0,25	1.575,44	27,74	0,64	2,22	-2,82	23,39	2,90	21,03	3,57	-6,08	-9,80	-8,98	-7,65	-0,03	-0,08	0,48	54,05	205,53	127,07	137,61	89,76	0,00	0,00	-4,54
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,22	3,98	-7,98	0,01	0,05	-0,17	-225,18	-5,72	-0,35	-0,52	0,13	-4,85	1,02	-2,78	2,09	7,33	-1,60	0,71	-6,96	0,00	0,01	-0,01	3,96	-8,08	-14,19	-8,02	-4,54	0,00	0,00	10,97

Figura 23: Matriz de varianza-covarianza para el Potrero 6

Como puede verse, los valores de varianza son bastante diferentes. En cuanto a la covarianza, los elementos coloreados que no están en la diagonal, indican la existencia de valores positivos y negativos distintos de cero. Por lo tanto, es coherente, que la

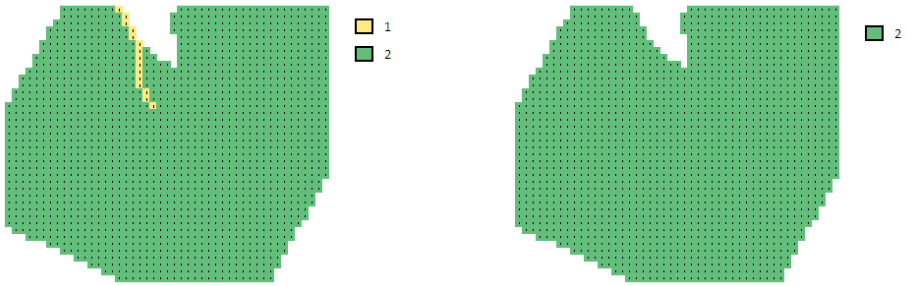
operación que calcula la norma a utilizar devuelva como resultado la distancia de Mahalanobis para este caso. Esto concuerda con lo planteado en (33) para el tipo de variables utilizadas.

Luego, se generaron las zonas de manejo estableciendo el tamaño mínimo de zona en 10.000 m^2.

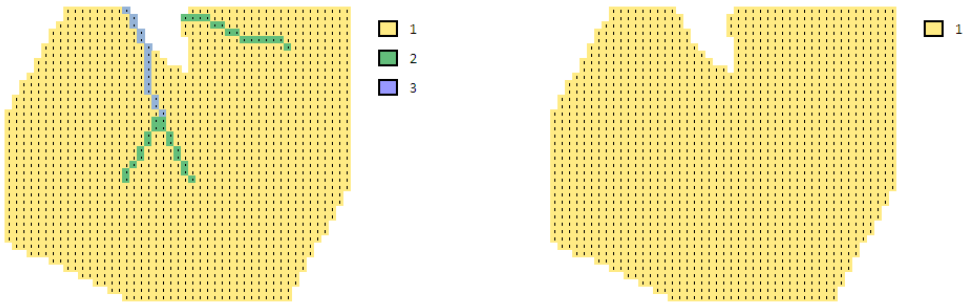
En las imágenes presentadas a continuación, pueden verse las zonas generadas antes y después de aplicar el post-procesamiento. También se presentan los índices de performance obtenidos para cada caso. Es importante recordar, que los índices son calculados antes de realizar el post-procesamiento.

Distancia Euclideana

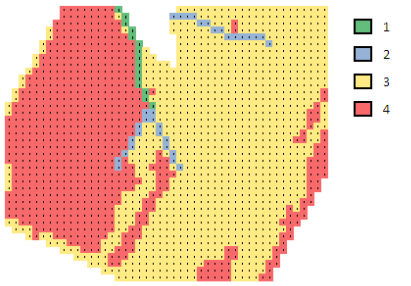
Cantidad de zonas = 2

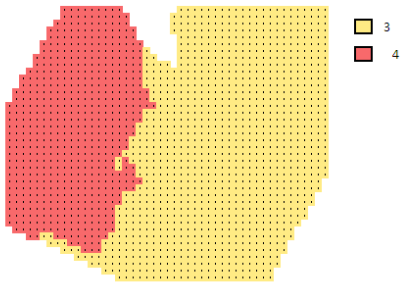


Cantidad de zonas = 3

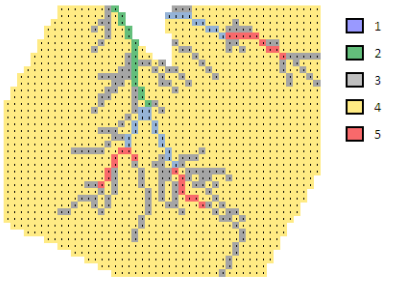


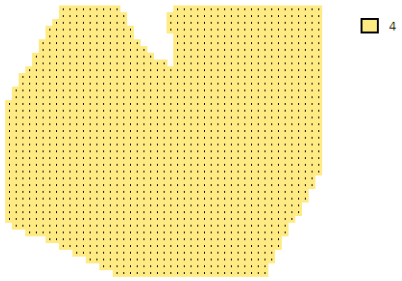
Cantidad de zonas = 4



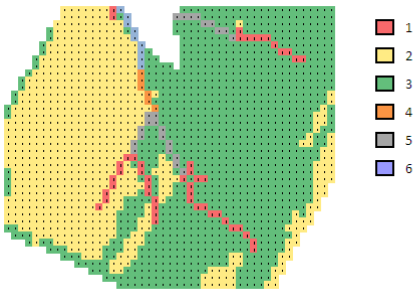


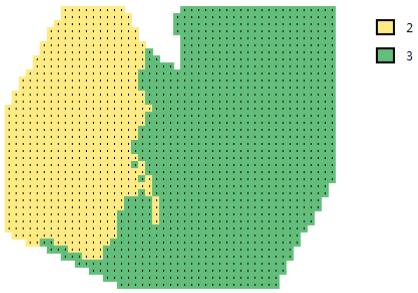
Cantidad de zonas = 5





Cantidad de zonas = 6





Índices de Performance

Los índices de performance obtenidos para estos casos se muestran en la Figura 24.

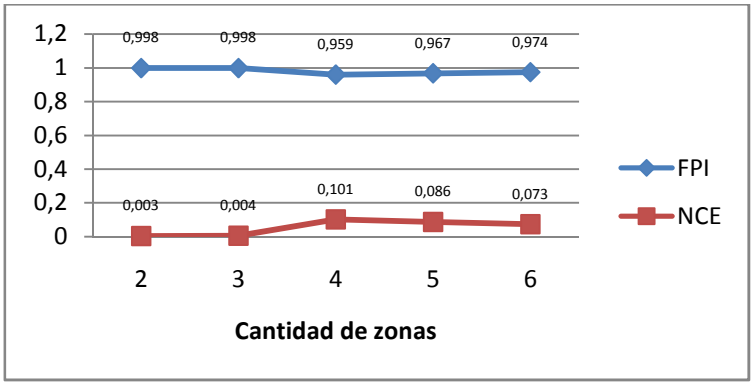
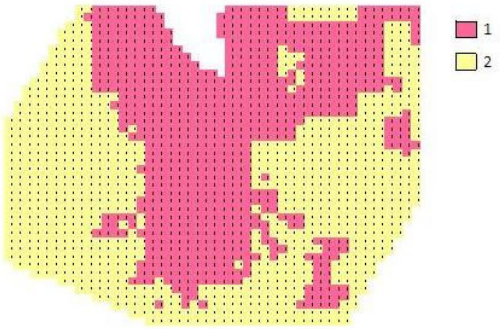


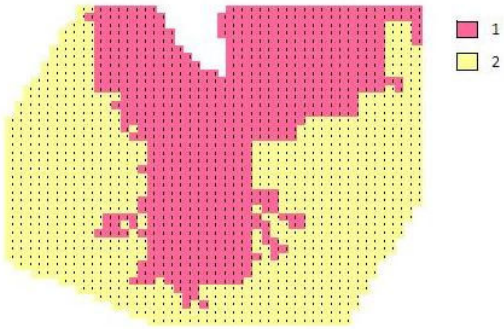
Figura 24: Índices de performance FPI y NCE calculados para el Potrero 6 utilizando la distancia Euclideana.

En este caso, no está muy claro cuál es la cantidad óptima de zonas. Esto confirma que la distancia Euclideana no es la mejor opción debido a las características presentadas por las variables.

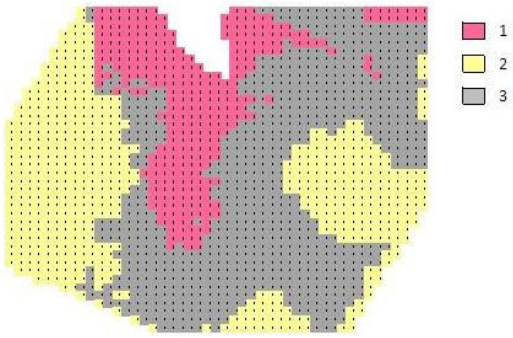
Distancia Diagonal

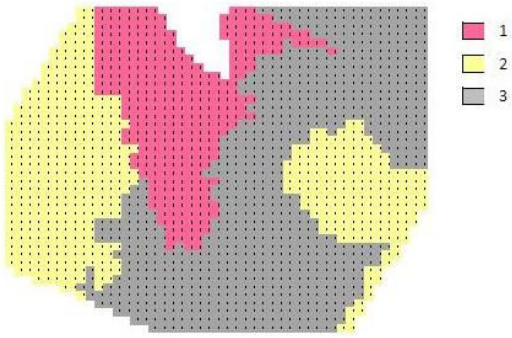
Cantidad de zonas = 2



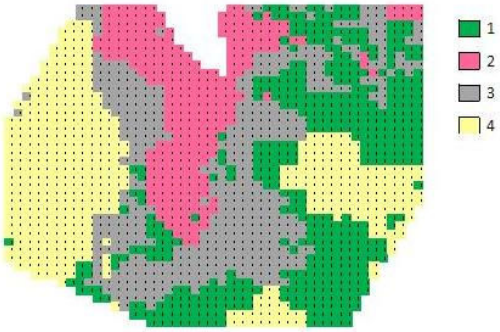


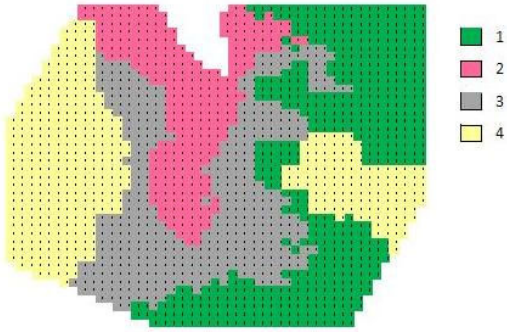
Cantidad de zonas = 3



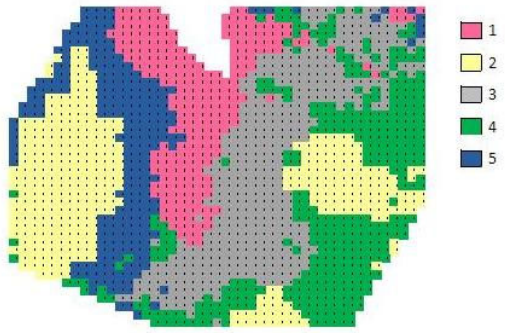


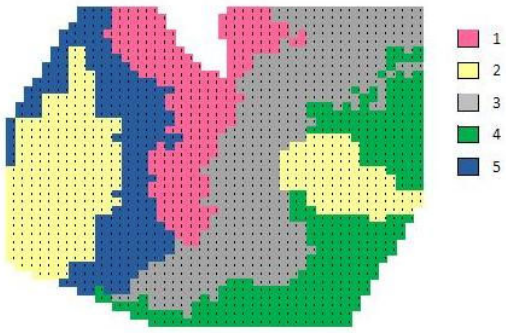
Cantidad de zonas = 4



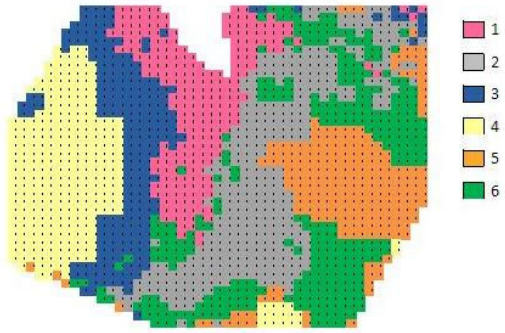


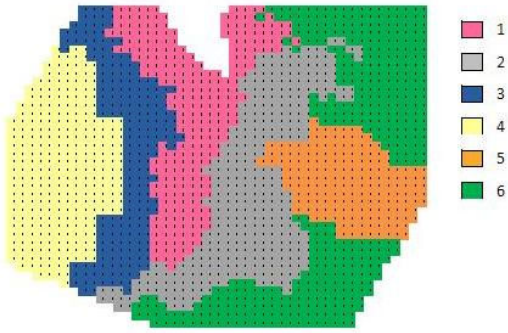
Cantidad de zonas = 5





Cantidad de zonas = 6





Índices de Performance

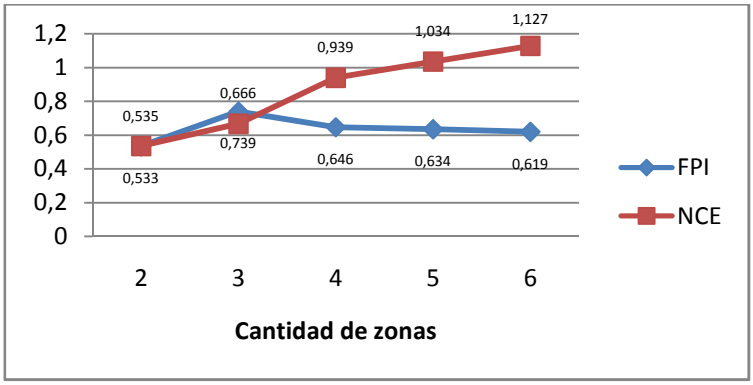


Figura 25: Índices de performance FPI y NCE calculados para el Potrero 6 utilizando la distancia Diagonal

En este caso, la mejor clasificación corresponde a la generación de 2 zonas de manejo.

Distancia de Mahalanobis

Para este caso no fue posible realizar pruebas, dado que para los datos correspondientes al Potrero 6, no existe la matriz inversa de la matriz de varianza-covarianza, que como se explicó anteriormente, se utiliza durante el algoritmo para calcular la matriz de distancias.