Universidad Nacional de Córdoba

Facultad de Ciencias Agropecuarias

Escuela para Graduados

HERRAMIENTAS ESTADÍSTICAS PARA EL MONITOREO Y USO DE LA VARIABILIDAD ESPACIAL DEL RENDIMIENTO Y PROPIEDADES DE SUELO INTRALOTE

Ing. Agr. Mariano Augusto Córdoba

Tesis

Para optar al Grado Académico de

Doctor en Ciencias Agropecuarias

Febrero, 2014

HERRAMIENTAS ESTADÍSTICAS PARA EL MONITOREO Y USO DE LA VARIABILIDAD ESPACIAL DEL RENDIMIENTO Y PROPIEDADES DE SUELO INTRALOTE

Mariano Augusto Córdoba

Comisión Asesora de Tesis

Director: Ing. Agr. (PhD) Mónica Balzarini

Asesores: Ing. Agr. (PhD) José L. Costa

Ing. Agr. (Dr) Edgar A. Rampoldi

Tribunal Examinador de Tesis

Ing. Agr. (Dr) Cecilia Bruno .....

Ing. Agr. (Dr) Gabriel Espósito .....

Ing. Agr. (Dr) Edgar A. Rampoldi .....

Presentación formal académica

Marzo 2014

Facultad de Ciencias Agropecuarias

Universidad Nacional de Córdoba

AGRADECIMIENTOS

Esta tesis ha sido posible gracias a numerosas personas e instituciones que me han brindado su apoyo a lo largo de los últimos cinco años.

En primer lugar quisiera agradecr a mi directora de tesis, Dra. Monica Balzarini por su constante y paciente seguimiento y asistencia compartiendo su tiempo de manera generosa durante el desarrollo del presente trabajo, por su generosidad al brindarme la oportunidad de recurrir a su capacidad y experiencia científica permitiéndome crecer tanto profesional como personalmente.

A la Dra. Cecilia Bruno, por guiarme constantemente, por su ayuda desinteresada, sus consejos, paciencia y motivación.

A los Profesores Julio Di Rienzo, Laura González y Margot Tablada de la Cátedra de Estadística y Biometría por estar siempre a disposición para atender mis consultas. A mi compañeros y amigos becarios por su ánimo y compañía.

A los investigadores que facilitaron sus datos para ser utilizados en esta ilustración; al Dr. José Luis Costa y Dr. Nahuel Peralta del INTA Balcarce.

A la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba por brindar un espacio de trabajo.

Al Consejo de Investigación Científica y Tecnológica por permitir llevar a cabo este trabajo de investigación a través del otorgamiento de la beca.

Agradezco a los Miembros del Comité Evaluador por aceptar gentilmente formar parte del tribunal examinador y por dedicar su valioso tiempo a la revisión de este trabajo.

A mi familia por permitirme recorrer este camino libremente acompañando mis logros

A mis amigos, incondicionales por su aliento y por que siempre están.

A todos MUCHAS GRACIAS.

A Maria Marta, mi hijo Octavio
y mis padres

RESUMEN

La agricultura de precisión (AP) debe ser comprendida como un concepto moderno de gestión agrícola basado en el uso de variables georreferenciadas tanto de características de suelo y topografia como de rendimientos. El óptimo uso del gran volumen de datos derivado de maquinarias pecisas depende fuertemente de las capacidades para explorar y analizar datos de complejas interacciones que subyacen los resultados productivos en cada sitio. En esta tesis se propone un protocolo integrado para procesar datos espaciales de sitio en AP. El análisis de la estructura espacial de variables de suelo y endimiento es investigado desde un enfoque interdisciplinario, que incluye perspectivas agronómicas y estadística-computacionales. Mediante la revisión y comparación del desempeño de métodos estadísticos usados para detectar y caracterizar variabilidad espacial se recomiendan y proponen estrategias de análisis para comprender y manejar mejor la variabilidad espacial del rendimiento. Se desarrolló un método para delimitar clases de sitios intralote, denominado KM-sPC, basado en análisis multivariado restringido espacialmente. El nuevo método es potente para detectar estructuración espacial, mejorando la delimitación de zonas de manejo (ZM). Los conglomerados de sitio formados dentre de cada lote tuvieron mayores diferencias de rendimiento que los obtenidos sin la restricción espacial. Se propusieron también herramientas, basadas en Modelos Lineales Mixtos (MLM), para analizar la estabilidad temporal de la variación espacial. Los resultados mostraron que la variación espacial de carcateristicas de suelo no es permanente. Finalmente, se propuso una estrategia de análisis de ensayos de fertilización sitio-específica utilizando MLM de clasificación. El procedimiento basado en la delimitación de ZM y luego el ajuste de un MLM con efectos fijos de tratamiento, zona, interacción tratamiento-zona y efecto aleatorio de bloque dentro de cada zona, mas correlación espacial entre parcelas resultó el mas recomendable para comparar efectos de tratamientos condicionados a las zonas delimitadas.

Palabras clave: agricultura de precisión, modelos mixtos, análisis multivariado espacial, análisis espacio-temporal.

ABSTRACT

Precision agriculture (PA) must be understood as a modern agriculture management concept based on the use of georeferenced variables both topography and soil characteristics and yields. Optimal use of the large volume of data derived from machinery of PA depends heavily on capabilities to explore and to analyze information on complex interactions underlying yields at each field site. The aim of this work was to develop an integrated methodology to process spatial data on yield, soil, and terrain, in the PA context. The analysis of spatial structure of yield and soil variables is investigated with an interdisciplinary approach, including both the methodological and agronomical perspectives. Through performance comparison of several statistical methods used to characterize spatial structure and its temporal stability, we propose analytic strategies to spatially analyze data variability. We propose and evaluate a method (KM-sPC) based on spatially constrained multivariate analysis to management zone (MZ) delineation. Using KM-sPC the degree of spatial structure detected is increased, improving the management zone delineation. The method delineated management classes with the largest differences in yield. The proposed algorithm was integrated into an automated protocol for management zone delineation. A tool set based on Mixed Linear Models (MLM) is proposed to analyze temporal stability of spatial variation in soil properties. The results showed that variation of soil variables were not permanent, producing significant changes to MZ delineation through years. Finally, we propose a statistical strategy for analyzing site-specific fertilization trials using a classification MLM. The procedure first involves the MZ delineation, and then the fitting of a MLM with fixed effects of treatment, treatment*MZ interaction, and random effect of block within each zone. The procedure was effective to compare treatment effects conditional on MZ in the context of spatial correlation between plots.

Key words: precision agriculture, mixed models, spatial multivariate analysis, spatio-temporal analysis.

.

TABLA DE CONTENIDOS

INTRODUCCIÓN GENERAL 1

OBJETIVO GENERAL 8

OBJETIVOS ESPECÍFICOS 8

CAPÍTULO I. HERRAMIENTAS ESTADÍSTICAS PARA EL ANÁLISIS DE DATOS ESPACIALES: APROXIMACIÓN UNIVARIADA 9

INTRODUCCIÓN 9

MATERIALES Y MÉTODOS 11

Datos 11
Procedimientos estadísticos para el análisis de variabilidad espacial 13
Índices de autocorrelación espacial 13
Geoestadística 16
Modelos lineales mixtos 21
Predicción y construcción de mapas 26

RESULTADOS 27

Cálculo del índice de Moran y de Geary 27
Implementación del análisis basado en semivariograma 29
Ajuste de un MLM a datos espaciales 33
Mapeo de variabilidad espacial de variables de suelo y rendimiento 36

DISCUSIÓN 37

CONCLUSIÓN 39

CAPÍTULO II. APROXIMACIÓN MULTIVARIADA EN EL ANÁLISIS DE DATOS ESPACIALES 40

INTRODUCCIÓN 40

MATERIALES Y MÉTODOS 43

Procedimientos de análisis multivariado 43
Análisis de *cluster fuzzy k-means* 43
Análisis de Componentes Principales sin y con restricción espacial 48
Implementación de algoritmos multivariados 50

Criterios de comparación 51

RESULTADOS 52

Clasificación de sitios intralote vía *cluster fuzzy k-means* 52
Clasificación de sitios intralote vía PCA y MULTISPATI-PCA 53

DISCUSIÓN 59

CONCLUSIÓN 61

CAPÍTULO III. PROPUESTA PARA LA CLASIFICACIÓN DE SITIOS INTRALOTE . 62

INTRODUCCIÓN 62

MATERIALES Y MÉTODOS 65

Datos 65
Propuesta algorítmica para la clasificación de sitios intralote 67
Evaluación del algoritmo KM-sPC 68
Procedimientos comparados con datos experimentales 68
Evaluación mediante simulación 68

RESULTADOS 70

DISCUSIÓN 81

CONCLUSIÓN 84

CAPÍTULO IV. PROTOCOLO DE ANÁLISIS PARA LA DELIMITACIÓN DE ZONAS DE MANEJO INTRALOTE 85

INTRODUCCIÓN 85

MATERIALES Y MÉTODOS 85

Datos 85
Protocolo 86
Conversión de coordenadas espaciales 86
Eliminación de outliers 87
Eliminación de inliers 88
Interpolación espacial 90
Clasificación de sitios 92
Delimitación de zonas de manejo 93
Validación de las zonas de manejo 95

RESULTADOS 96

Ilustración de la aplicación del protocolo 96
Paso 1-Conversión de coordenadas espaciales 96
Paso 2-Eliminación de outliers 97
Paso 3-Eliminación de inliers 99
Paso 4-Interpolación espacial de los datos 102
Paso 5-Delimitación de clases de sitios 106
Paso 7-Delimitación de zonas de manejo 108
Paso 8-Validación de zonas de manejo 114

CAPÍTULO V. ANÁLISIS TEMPORAL DE LA VARIABILIDAD ESPACIAL DE PROPIEDADES DE SUELOS INTRALOTE 119

INTRODUCCIÓN 119

MATERIALES Y MÉTODOS 122

Datos 122
Estrategia de análisis 122
Resumen de los pasos para implementar el algoritmo propuesto 125

RESULTADOS 126

CONCLUSIÓN 131

CAPÍTULO VI. MODELOS MIXTOS PARA EL ANÁLISIS DE ENSAYOS DE FERTILIZACIÓN SITIO-ESPECÍFICA 133

INTRODUCCIÓN 133

MATERIALES Y MÉTODOS 138

Datos 138
Análisis de ensayo bajo un modelo de clasificación 139

RESULTADOS 140

CONCLUSIÓN 148

CONCLUSIONES 150

BIBLIOGRAFÍA 155

ANEXO 1 169

ANEXO 2 171

ANEXO 3 175

ANEXO 4 179

LISTA DE TABLAS

Tabla 1.1. Índices de autocorrelación espacial en variables de suelo y rendimiento. 28

Tabla 1.2. Modelos teóricos de semivariogramas. Funciones de semivariograma para el modelo exponencial, esférico y gaussiano. C0=2, C=10 y R=200.....20

Tabla 1.3. Estadística descriptiva para las propiedades de suelo y rendimiento de un lote con mediciones georreferenciadas en 664 sitios.....29

Tabla 1.4. Modelos de regresión lineal múltiple para evaluar tendencia a gran escala.....31

Tabla 1.5. Estimaciones (WLS) de los parámetros de un semivariograma exponencial ajustado a variables de suelo y rendimiento en un lote con mediciones georreferenciadas en 664 sitios.....31

Tabla 1.6. Criterios de información sobre ajustes de MLM de correlación espacial para variables edáficas (N=664 sitios).....33

Tabla 1.7. Criterios de información sobre ajustes de MLM de correlación espacial para variables de rendimiento (N=664 sitios).....33

Tabla 1.8. Test del cociente de verosimilitud (LRT) basada en los estimadores ML para evaluar tendencia a gran escala.....35

Tabla 1.9. Estimaciones de los parámetros del MLM ajustado para datos espaciales de suelo y rendimiento.....35

Tabla 1.10. Kriging Simple, Ordinario y Universal. 26

Tabla 2.1. Selección del número de clases de la partición de sitios intralote a partir del análisis de cluster *fuzzy k-means*.....52

Tabla 2.2. Autovalores, varianza espacial e índices de Moran de las componentes principales generados a partir de MULTISPATI-PCA sobre cuatro variables de suelo y dos de rendimiento. 53

Tabla 2.3. Autovalores (varianza) e índices de Moran de las componentes principales generadas a partir del PCA sobre cuatro variables de suelo y dos de rendimiento. 53

Tabla 2.4. Estimaciones de los parámetros del MLM ajustado para las componentes principales del PCA (CP) y MULTISPATI-PCA (sPC).....57

Tabla 3.1. Estadística descriptiva de variables de suelo y rendimiento en tres lotes agrícolas monitoreados intensivamente. 71

Tabla 3.2. Coeficientes de correlación entre variables de suelo y rendimiento para tres lotes en producción agrícola.....72

Tabla 3.3. Estadística descriptiva para las tres primeras componentes principales generadas con los análisis de componentes principales (PCA) y MULTISPATI-PCA aplicados a bases de datos multivariados de tres lotes agrícolas. 74

Tabla 3.4. Autovectores (ponderaciones de variables) de los análisis de componentes principales (PCA) y MULTISPATI-PCA. Se subrayan los coeficientes más importantes. 75

Tabla 3.5. Rendimientos promedios para dos clases de manejo delimitadas por los siguientes métodos de clasificación: *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), *cluster fuzzy k-means* sobre componentes principales (KM-PC) y *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC). 77

Tabla 3.6. Promedios de variables de suelo para dos clases de sitios delimitadas por los siguientes métodos de clasificación: *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), *cluster fuzzy k-means* sobre componentes principales (KM-PC) y *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC). 79

Tabla 3.7. Porcentaje de las simulaciones que identifican clases de sitios con diferencias estadísticamente significativas entre rendimiento para tres procedimientos de clasificación: *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), *cluster fuzzy k-means* sobre componentes principales (KM-PC) y *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC). 81

Tabla 5.1. Escala de valoración del índice de kappa 124

Tabla 5.2. Media, coeficiente de variación (CV), valores mínimos (Min.) y máximos (Max.) de variables de suelo para un mismo lote agrícola en tres años. 127

Tabla 5.3. Índice Kappa para evaluar la concordancia de la zonificación obtenida en tres años de captura de mediones de propiedades de suelo. 130

Tabla 6.1. Criterios de selección de modelos para el Lote 1. 140

Tabla 6.2. Criterios de selección de modelos para el Lote 2. 141

Tabla 6.3. Criterios de selección de modelos para el Lote 3. 142

Tabla 6.4. Criterios de selección de modelos para el Lote 4. 143

Tabla 6.5. Criterios de selección de modelos para el Lote 5. 144

Tabla 6.6. Criterios de selección de modelos para el Lote 6. 145

LISTA DE FIGURAS

Fig. 1.1. Red de conexión calculada mediante la distancia Euclidea. Se consideran sitios vecinos a aquellos separados a una distancia menor a 30 m entre sitios georreferenciados con datos del archivo *CasoI.txt*. 28

Fig. 1.2. Semivariograma esférico. Se representan los tres parámetros que lo definen: el rango, el umbral y el efecto pepita. 18

Fig. 1.3. Distribución de frecuencias para variables de suelo y rendimiento de un lote con 664 mediciones georreferenciadas. CE30: conductividad eléctrica aparente a 30 cm de profundidad, CE90: conductividad eléctrica aparente a 90 cm de profundidad, E: elevación, Pe profundidad tosca, Sj: rendimiento de soja; Tg: rendimiento de trigo. 30

Fig. 1.4. Semivariogramas empíricos (círculos) y teóricos (línea partida) obtenidos a partir de análisis geoestadístico. a) Conductividad eléctrica aparente a 30 cm de profundidad, b) conductividad eléctrica aparente a 90 cm de profundidad, c) Elevación, d) profundidad de tosca, e) rendimiento de soja, f) rendimiento de trigo. 32

Fig. 1.5. Mapas de variabilidad espacial de variables de suelo y rendimiento obtenidos mediante la interpolación por kriging ordinario utilizando parámetros del semivariograma estimados con goestadística clásica. a) Conductividad eléctrica aparente a 30 cm de profundidad, b) Conductividad eléctrica aparente a 90 cm de profundidad, c) Elevación, d) profundidad de tosca, e) rendimiento de soja, f) rendimiento de trigo. 36

Fig. 1.6. Mapas de variabilidad espacial de variables de suelo y rendimiento obtenidos mediante la interpolación por kriging ordinario utilizando parámetros del semivariograma estimados con modelos lineales mixtos. a) Conductividad eléctrica aparente a 30 cm de profundidad, b) Conductividad eléctrica aparente a 90 cm de profundidad, c) Elevación, d) profundidad de tosca, e) rendimiento de soja, f) rendimiento de trigo. 37

Fig. 2.1. Mapa con clases delimitadas: dos (izquierda), tres (centro) y cuatro (derecha) clases. 52

Fig. 2.2. Representación gráfica de los dos primeros ejes del PCA (a) y MULTISPATI-PCA (b) y sus autovalores. Se muestra la correlación entre las variables y entre estas y las componentes principales. 54

Fig. 2.3. Mapas de la PC1 del PCA (a) y sPC1 del MULTISPATI-PCA (b). 55

Fig. 2.4. Mapas de la PC2 del PCA (a) y sPC2 del MULTISPATI-PCA (b) 55

Fig. 2.5. Mapas obtenidos por interpolación (Kriging) de la PC1 del PCA (a) y sPC1 del MULTISPATI-PCA (b). 57

Fig. 2.6. Mapas obtenidos por interpolación (Kriging) de la PC2 del PCA (a) y sPC2 del MULTISPATI-PCA (b). 58

Fig. 2.7. Mapas de variabilidad espacial de variables de suelo y rendimiento obtenidos mediante la interpolación por kriging ordinario utilizando parámetros del semivariograma estimados con modelos lineales mixtos. a) Conductividad eléctrica aparente a 30 cm de profundidad, b) Conductividad eléctrica aparente a 90 cm de profundidad, c) Elevación, d) profundidad de tosca, e) rendimiento de soja, f) rendimiento de trigo. 58

Fig. 3.1. Fuzziness Performance Index (FPI, círculos) y Normalized Classification Entropy (NCE, cuadrados) para tres métodos de clasificación: a) *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC)..... 76

Fig. 3.2. Gráficos Box-plots de la distribución de CE30 dentro de clases delimitadas por tres métodos de clasificación: *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), (b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y (c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC)..... 78

Fig. 3.3. Clase 1 (blanco) y clase 2 (negro) del lote 1 delimitadas por tres métodos de clasificación: (a) *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), (b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y (c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC). 80

Fig. 3.4. Clase 1 (blanco) y clase 2 (negro) del lote 2 delimitadas por tres métodos de clasificación: (a) *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), (b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y (c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC). 80

Fig. 3.5. Clase 1 (blanco) y clase 2 (negro) del lote 3 delimitadas por tres métodos de clasificación: (a) *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), (b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y (c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC). 80

Fig. 4.1. Histograma y Box-plot de datos de conductividad eléctrica aparente a 30 cm de profundidad previo a la eliminación de *outliers.* 98

Fig. 4.2. Histograma y Box-plot de datos CE30 luego de la eliminación de *outliers.* 99

Fig. 4.3. Gráfico de dispersión de Moran de la variable CE30. 100

Fig. 4.4. Semivariograma empírico de la variable CE30. 103

Fig. 4.5. Semivariograma empírico (puntos) y teórico (línea) de la variable CE30. 104

Fig. 4.6. Grilla de predicción de 10 × 10 m. 105

Fig. 4.7. Mapa de interpolación espacial de la variable CE30. 105

Fig. 4.8. Gráfico Biplot del análisis MULTISPATI-PCA. 107

Fig. 4.9. Mapa con dos clases de manejo intralote. 110

Fig. 4.10. Mapa con tres clases de manejo intralote. 110

Fig. 4.11. Mapa con cuatro clases de manejo intralote. 111

Fig. 4.12. Mapa con dos clases de manejo intralote previo a la aplicación del filtro de la mediana. 112

Fig. 4.13. Mapa con dos clases de manejo intralote luego de aplicar un filtro de la mediana de 3 × 3 píxeles. 112

Fig. 4.14. Mapa con dos clases de manejo intralote luego de aplicar un filtro de la mediana de 5 × 5 píxeles. 113

Fig. 4.15. Mapa con dos clases de manejo intralote luego de aplicar un filtro de la mediana de 7× 7 píxeles. 113

Fig. 4.16. Mapa con dos zonas de manejo intralote. 118

Fig. 5.1. Distribuciones de frecuencias de datos de MO, P, pH y CE en tres años (2005 (verde), 2008 (azul), y 2011 (rojo). 126

Fig. 5.2. Tendencia temporal en promedios de MO, P, pH y CE. Letras diferentes indican diferencias estadísticamente significativas ($p<0.05$). 128

Fig. 5.3. Gráficos de dispersión de la variabilidad temporal en función de la media temporal de cada variable de suelo. 129

Fig. 5.4. Zonas de manejo para cada año obtenidas con datos de cuatro variables de suelo. 130

Fig. 5.5. Zonas de manejo para cada año obtenidas con datos de cuatro variables de suelo y particionadas en función de la inestabilidad temporal de cada variable. Sitios con desviación estándar temporal por encima del tercer cuartil de la desviación temporal (P(75)) son más inestables. 131

Fig. 6.1. Rendimientos promedios de acuerdo a dosis de nitrógeno y zona de manejo. Letras diferentes indican diferencias estadísticamente significativas ($p<0.05$). 147

Fig. 6.2. Análisis de compoentes principales de variables de sitio. CE90: Conductividad eléctrica aparente a 90 cm de profundidad, Pe: profundidad de tosca, E: elevación. 148

LISTA DE SIMBOLOS Y ABREVIATURAS

AIC: Criterio de información de Akaike

BIC: Criterio bayesiano de Schwarz

C: Varianza estructural (varianza espacialmente estructurada)

$C + C_0$: Varianza umbral (total)

C_0: Varianza nugget

CP: Componente principal

CV: Coeficiente de variación

EE: Error estándar

ESDA: Análisis exploratorio de datos espaciales

Esf: Esférico

Exp: Exponencial

IM: Índice de Moran

IG: Índice de Geary

ML: Máxima verosimilitud

MLM: Modelo lineal mixto

MO: Material orgánica

MULTISPATI- PCA: Análisis de componentes principales restringido espacialmente

Nug: Nugget

OK: Kriging Ordinario

PCA: Análisis de componentes principales

R: Rango

REML: Máxima verosimilitud restringida

R_p: Rango práctico

RSV: Variabilidad estructural relativa $\left(\frac{C_0}{C_0+C}\right)$

SK: Kriging Simple

$S(x)$: Campo aleatorio

UK: Kriging Universal

WLS: Mínimos cuadrados ponderados

ZM: zona de manejo

CM: clase de manejo

AP: Agricultura de Precisión

CE30: Conductividad eléctrica aparente a 30 cm de profundidad

CE90: Conductividad eléctrica aparente a 90 cm de profundidad

Pe: Profundidad efectiva del suelo o de tosca

E: Elevación

Tg: Rendimiento de trigo

Sj: Rendimiento de soja

P: Fósforo

CE: Conductividad eléctrica aparente

sPC: Componentes Principales espaciales,

DGPS: Sistema de posicionamiento global diferencial

LRT: Test de razón de verosimilitud

RB: Modelo de bloques aleatorios

SP: Modelo de correlación espacial

RB+SP: Modelo de bloques aleatorios más modelo de correlación espacial

FPI: fuzziness performance index

NCE: normalized classification entropy

KM-sPC: Análisis de cluster fuzzy k-means sobre componentes principales espaciales

KM-SV: Análisis de cluster fuzzy k-means sobre variables de suelo originales

KM-PC: Análisis de cluster fuzzy k-means sobre componentes principales

MZA: Management Zone Analyst

RV: complemento de la varianza relativa

L1: Lote 1

L2: Lote 2

L3: Lote 3

MA: media ajustada

Ii: Índice autocorrelación espacial Local de Moran

DE: Desviación estándar

LI: Límite inferior

LS: Límite superior

DBCA: Diseño en bloques completos al azar

INTRODUCCIÓN GENERAL

Las investigaciones en agricultura deben orientarse al desarrollo y aplicación de tecnologías que incrementen las fuentes primarias de alimentos. La alimentación de la incrementada población mundial requiere, cada vez más, de un sistema de agricultura sostenible. Las mayores escalas de producción agrícola, así como el cambio climático, el incremento en el costo de la tierra y la necesidad de bajar el nivel de insumos destinados a la producción agropecuaria, plantean fuertes motivaciones para la adaptación a la innovación tecnológica. Los pronosticados cambios en temperaturas y precipitaciones hacen pensar que la sostenibilidad de la agricultura demandará una mayor capacidad para monitorear detalladamente los cultivos y adaptarse rápidamente a escenarios variables (FAO, FIDA y PMA., 2012). Este nuevo escenario de la agricultura obliga a los productores a aumentar la eficiencia productiva a través de la utilización de nuevas prácticas agrícolas, entre ellas las relacionadas a la agricultura de precisión (AP) (Hörbe *et al*., 2013; Oliver, 2013).

En ambientes agrícolas, donde frecuentemente la disponibilidad de agua y de nutrientes son limitantes del cultivo, la producción de granos depende en gran medida del tipo de suelo; en particular de su capacidad para retener agua y nutrientes para luego ponerlos a disponibilidad del cultivo. Uno de los requerimientos centrales de la AP es la obtención de zonas de manejo (ZM) definidas por factores limitantes del rendimiento, que luego podrán ser manejadas de acuerdo a sus propiedades intrínsecas. Según Fraisse *et al*. (2001) y Plant *et al*. (2001), los principales requerimientos que las ZM deben cumplir para ser consideradas como tales son: a) las diferencias de rendimientos entre ZM debe ser mayor que las diferencias dentro de las ZM; y b) los factores limitantes de rendimiento dentro de la ZM deben ser los mismos. La clasificación de sitios en ZM debería ser analizada en el tiempo, ya que estas zonas no son necesariamente estáticas y posiblemente varíen ante cambios en las prácticas de manejo de los productores y también como consecuencia de factores aleatorios como los climáticos. Pueden ser necesarios varios años de datos para entender mejor las interacciones entre la variabilidad espacial en las propiedades de los suelos y la productividad de los cultivos (Bongiovanni *et al*., 2006). La

definición de clases manejo intralote, basadas en la capacidad del suelo, adquiere mayor importancia en sistemas productivos de secano (Taylor *et al.*, 2007).

Diferentes fuentes de datos pueden utilizarse para delimitar clases de manejo intralote. Los ejemplos en la literatura van desde utilizar el conocimiento que tiene el productor de su lote (Fleming *et al*., 2000; Hörbe *et al.*, 2013), utilizar mapas de rendimiento (Diker *et al*., 2004; Flowers *et al*., 2005; Hörbe *et al*., 2013), basarse en cartas de suelo (Franzen *et al*, 2002; Dillon *et al*., 2005), el uso de imágenes satelitales (Zarco-Tejada *et al*., 2005; Pedroso *et al*., 2010), la implementación de sensores proximales de suelo (Halcro *et al*., 2003; Cowin y Lesch 2005; Hemmat y Adamchuk, 2008; Ben-Dor *et al*., 2009; Piikki *et al*., 2013) y combinaciones de estas fuentes (Fleming *et al*., 2004; Koch *et al*., 2004; Schepers *et al*., 2004; Taylor *et al*., 2007; Gregoret *et al*., 2011; Esposito *et al*., 2012). Sin embargo, no es sencillo determinar cúal fuente puede ser mas conveniente para delimitar zonas de manejo en un lote en particular. Así mismo, la selección de la fuente de datos dependerá de la capacidad del productor de contar con una u otra tecnologías de captura de datos. Actualmente, excepto casos específicos, se recomienda utilizar sensores proximales a partir de los cuales se puedan obtener mediciones no solo del rendimiento de cada sitio sino también de factores potencialmente limitantes del rendimiento (Taylor *et al*., 2007). La utilización de monitores de rendimiento, en la Argentina, es una práctica cada vez más frecuente entre los productores. Se estima que, a mayo de 2013, 8915 cosechadoras contaban con monitores de rendimiento, razón por la cual los equipos en funcionamiento en la actualidad, alcanzarían para monitorear el rendimiento de aproximadamente el 70% de la superficie cultivable del país (Bragachini *et al*., 2011; Melchiore *et al*., 2013). Los mapas de rendimiento contienen información sobre la integración de efectos de procesos químicos, físicos y biológicos, que bajo ciertas condiciones climáticas permiten monitorear los patrones espaciales de la productividad de los cultivos indicando dónde es necesario variar los *inputs* del sistema para mejorar la productividad (Long, 1998; Hörbe *et al*., 2013; Rodrigues *et al*., 2013). Los *inputs* requeridos para optimizar la productividad de los cultivos y minimizar impactos sobre los ambientes son mejor determinados cuando se conocen los factores que provocan los patrones de variación espacial intralote de los rendimientos. Los mapas de rendimiento no son suficientes para obtener información que permita distinguir las distintas fuentes de variación y, consecuentemente, no proveen una guía clara para el manejo cuando no son

acompañados de mediciones relativas a las propiedades del suelo, factores de manejo y variables climáticas (Van Uffelen *et al*., 1997; Taylor *et al*., 2007; Rodrigues *et al*., 2013). Cada factor que afecta la variación intralote de los rendimientos debe estar también caracterizado espacialmente.

En los últimos años junto a las cosechadoras precisas hubo un incremento del uso de sensores proximales que captura datos de conductividad eléctrica aparente (CE) y otras propiedades edáficas y del terreno. Las mediciones de CE responden fuertemente al contenido de arcilla y humedad del suelo en suelos no salinos (Taylor *et al*., 2007). Por ello el uso de la CE ha demostrado ser eficaz en la delimitación de zonas de manejo en diferentes países donde se desarrolla la agricultura extensiva de precisión (Li *et al*., 2007; Taylor *et al*., 2007; Corwin y Lesch, 2010; Moral *et al*., 2010; Arno *et al*., 2011), incluyendo la región Pampeana Argentina (Peralta y Costa, 2013; Peralta *et al*., 2013; Simón *et al*., 2013). Dado que para georreferenciar los datos capturados por sensores de CE se utilizan GPS de alta precisión, es posible recolectar simultáneamente datos de otra fuente de información como la elevación. Esta propiedad topográfica influye en el movimiento del agua y el desarrollo del suelo a nivel del sitio intralote por lo tanto, es un indicadora de variabilidad espacial del rendimiento. Otra propiedad de suelo que suele utilizarse para agricultura de precisión, es la profundidad efectiva del suelo (Peralta *et al*., 2012). Esta influye principalmente en la capacidad de almacenaje del suelo y en su distribución espacial, generando variabilidad espacial de los rendimientos de los cultivos.

En Argentina, desde fines de la década del 90 hasta la actualidad, se destacó el rol del INTA en la difusión, experimentación y capacitación en AP. Sin embargo aún existen limitantes para la adopción y el uso de esta tecnología. Esto se puede atribuir principalmente a la falta de especialización y capacitación, el alto costo de la tecnología y servicios, y las dudas de los productores sobre el retorno económico de la inversión en tecnologías de AP (Melchiori *et al*., 2013).

La mejora en la productividad de los cultivos a través de la adopción de AP es más factible cuando se conoce la magnitud y las condiciones bajo las cuales los patrones espaciales de los rendimientos intralote se mantienen temporalmente o son estables (Blackmore *et al*., 2003). En síntesis, la AP debe ser comprendida como un concepto moderno de gestión agrícola que se basa en el uso de variables georreferenciadas o

regionalizadas tanto de características de los sitios intralote, como lo son propiedades del suelo, topografía, nutrientes, como de los rendimientos en cada uno de esos sitios.

Dado que la AP se basa en información regionalizada, es necesario contemplar en el análisis de los datos espaciales capturados, el fenómeno de autocorrelación espacial positiva (Anselin, 2001). Este tipo de autocorrelación, esperable en datos espaciales de características de sitio, genera dependencias entre las observaciones según la distancia espacial que existe entre los sitios desde donde se realizan las mediciones. Nuevas tecnologías de información son necesarias para contemplar la variabilidad espacial de las características de sitio y rendimiento en el espacio productivo (lote). El uso de la información obtenida, desde regiones vecinas a un sitio determinado, permite no sólo mejorar la calidad de la información en ese sitio sino también modelar la variabilidad espacial de las características de interés.

El óptimo uso del gran volumen de datos derivado de maquinarias de agricultura de precisión y de sistemas de sensores depende fuertemente de las capacidades para explorar los datos espaciales que de ellos se obtienen (Bullock y Lowenberg-DeBoer, 2007). La covariación espacial de las propiedades del sitio y el rendimiento de los cultivos puede ser estudiada a través de una gran diversidad de métodos y modelos estadísticos. Los más usados son los modelos geoestadísticos clásicos (Cressie, 1993; Schabenberger y Gotway, 2004; Webster y Oliver, 2007). La aplicación explícita de la geoestadística en la agricultura de precisión se produjo en el año 1988 (Mulla y Hammond, 1988). Los objetivos de su trabajo fueron mapear los patrones de variación espacial de P y K, determinar la naturaleza y el alcance de esta variación y la intensidad de muestreo necesaria para identificar los principales patrones del suelo. Mulla y Hammond (1988) recomendaron que si el suelo es variable los agricultores debieran evitar aplicaciones uniformes. Burgess y Webster (1980), Miller *et al*. (1988) y Webster y Oliver (1989) utilizaron geoestadística para el análisis de datos de suelo. Su objetivo fue cuantificar la estructura espacial en la variación mediante el uso de variogramas y utilizar sus parámetros para realizar predicciones con kriging para obtener mapas de variación espacial de propiedades de suelo. Aunque su trabajo fue anterior al de Mulla y Hammond (1988) y se relacionó explícitamente con la agricultura, no se basó explícitamente en el concepto moderno de la AP. Además del uso de variogramas e interpolaciones tipo Kriging, el análisis de estructura espacial fue, desde los comienzo, realizado mediante el uso de

índices de autocorrelación, como el índice de Moran (Moran, 1948). En las últimas décadas, surgieron nuevos desarrollos de modelos estadísticos contemporáneos, entre los que se destacan los modelos lineales mixtos (Demidenko, 2004; West *et al*., 2007) que constituyen herramientas innovadoras para el tratamiento de datos correlacionados, particularmente espacial y/o temporalmente (Rodrigues *et al*., 2013). Varios problemas de estimación de niveles medios y variabilidad de propiedades de suelo y de rendimiento han sido tratados empleando modelos mixtos (Kravchenko *et al*., 2005; 2006; Gili, 2013). La estimación se hace por métodos basados en la función de verosimilitud, ya sea máxima verosimilitud (ML) o, alternativamente, máxima verosimilitud restringida (REML) (Searle *et al*., 1992; Gili, 2013).

Los métodos antes mencionados han sido frecuentemente utilizados en el contexto univariado, es decir para el análisis de una única variable registrada en cada sitio del lote en estudio. Actualmente, los métodos de análisis multivariado (Johnson y Wichern, 2007; Rencher y Christensen, 2012), principalmente *clusters* y componentes principales, resultan apropiados para la visualización y exploración simultanea de datos de varias variables regionalizadas. Sin embargo, la mayoría de la técnicas multivariadas, no han sido desarrolladas explícitamente para manejar datos espaciales por lo cual la combinación de la información multivariada obtenida desde cada sitio del lote y su posición usualmente se realiza *a posteriori* del análisis. Un caso distinto es el de la técnica multivariada MULTISPATI-PCA (Dray *et al*., 2008) que fue diseñada para contemplar la característica espacial del dato *a priori* del análisis.

El presente trabajo de investigación trata, en primer lugar y en virtud de la multiplicidad de enfoques y análisis estadísticos disponibles, el problema de la caracterización de la estructura espacial uni y multivariada de rendimientos y características de sitios intralote, desde una perspectiva estadístico-metodológica. Bajo esta dimensión, los primeros Capítulos se focalizan en el análisis de datos de georreferenciados y discuten aspectos de la aplicación de métodos estadísticos disponibles utilizando datos de rendimiento y suelo, medidos intensivamente, dentro de lotes en producción de cultivo de grano de la región pampeana. En segundo lugar, se propone y evalúa un nuevo método multivariado para la delimitación de clases de manejo (CM) que considera la naturaleza espacial de los datos; dicha propuesta está orientada a manejar las correlaciones espaciales entre las distintas características de sitio durante el proceso de delimitación de CM, de

manera tal de incrementar las diferencias de rendimientos entre las CM delineadas y disminuir la variabilidad de rendimiento dentro de las mismas. El método propuesto es integrado en un protocolo recomendado para la delimitación de zonas de manejo. Posteriormente, se abordan modelos que permiten incluir la dimensión serial de las mediciones que se realizan año tras año en cada lote y la cuantificación de la estabilidad temporal de la variabilidad espacial usada en la delimitación de zonas de manejo. Finalmente, se analizan modelos alternativos para el análisis de ensayos comparativo de dosis de fertilización según características de suelo y terreno. Esta sección se incorpora debido a que en el contexto de la AP, no solo es necesario delimitar las ZM sino también identificar cuáles son los factores limitantes del rendimiento y ensayar posibles manejos sitio-específicos.

Así, el Capítulo I consiste en la revisión e ilustración usando un conjunto de datos reales, de metodologías de análisis estadístico para detectar y caracterizar la estructura espacial del rendimiento y variables de sitio de manera univariada. A partir de estos procedimientos se ilustra la obtención de predicciones sitio-especifica y mapas de variabilidad espacial para cada variable. En el Capítulo II se exponen tres técnicas multivariadas (análisis de *cluster*, análisis de componentes principales y análisis de componentes principales restringido espacialmente) candidatas para el análisis de variación espacial en el contexto de observaciones multivariadas, es decir cuando existen registros para varias variables en cada sitio, y se espera que estas variables estén correlacionadas. Para facilitar la comprensión de la importancia del uso de técnicas multivariadas específicamente desarrolladas para datos espaciales, se realizó un análisis comparativo de los resultados obtenidos, para datos de AP, con la implementación de un PCA clásico y de una versión restringida espacialmente (MULTISPATI-PCA, Dray *et al*., 2008). Los resultados de los Capítulos I y II muestran cuán diferentes son las interpretaciones agronómicas que deben derivarse, en un estudio particular de datos espaciales, dependiendo del análisis utilizado. En el Capítulo III se considera la autocorrelación espacial en cada una de las variables de sitio, vía MULTISPATI-PCA, con la finalidad de generar variables sintéticas (Componentes Principales espaciales, sPC) que combinan las mediciones originales con la información espacial, y que luego son usadas para la clasificación de sitios bajo un algoritmo de conglomeración (*fuzzy k-means*). La estrategia de análisis propuesta se denominó KM-sPC. Para evaluar el desempeño estadístico de KM-

sPC, se realizaron sobre las bases de datos de tres lotes en producción, análisis de *cluster* usando como *input* las variables de suelos originalmente medidas y ambos tipos de componentes principales, restringidas y no restringidas espacialmente. La comparación del desempeño estadístico alcanzado en la clasificación de sitios intralote, fue también evaluada mediante la simulación de campos espaciales aleatorios con estructura de datos multidimensional. En el Capítulo IV se presenta un protocolo diseñado para el análisis estadístico de datos de sitio intralote con la finalidad de delimitar zonas homogéneas que podrían potencialmente ser utilizadas como zonas de manejo para agricultura sitio-especifica. El Capítulo V integra el estudio de variabilidad temporal de la variación espacial de variables de sitio, particularmente características de suelo. Se propone una metodología estadística para estimar tendencia temporal de estas variables y construir mapas de variabilidad espacio-temporal. Finalmente, en el Capítulo VI se exponen y comparan vía distintos criterios de comparación, modelos lineales mixtos para el análisis de ensayos de fertilización sitio específico. En todos los Capítulos se han ilustrados el desarrollo de aplicaciones clásicas y más novedosas a partir de conjuntos de datos provenientes de lotes de la región pampeana de Argentina donde se desarrollan cultivos extensivos (Soja, Maíz, Trigo) bajo AP. Para la mayoría de las aplicaciones implementadas en esta tesis se han documentado los *script* en R (R Core Team, 2013) que permiten trabajar datos de AP desde su pre-procesamiento hasta la delimitación de zonas de manejo y la evaluación de prácticas de manejo sitio-específicas.

OBJETIVO GENERAL

Desarrollar herramientas estadísticas para el análisis de la variación espacial y temporal de variables de suelo y rendimientos que faciliten la identificación y comparación de zonas de manejo intralote.

OBJETIVOS ESPECÍFICOS

1. Comparar, desde su aplicación en escenarios de procesos espaciales continuos registrados en agricultura de precisión, la información obtenida desde distintos métodos estadísticos para datos espaciales univariados: índices de autocorrelación espacial, semivariogramas y modelos mixtos de covarianza residual espacial.

2. Ilustrar, desde su aplicación en escenarios de procesos espaciales continuos multivariados asociados a la caracterización multidimensional de sitios intralote, la información obtenida desde métodos multivariados.

3. Proponer una nueva metodología para la clasificación multivariada de sitios intralote que contemple la autocorrelación espacial entre las variables de sitio.

4. Desarrollar un protocolo de análisis estadístico recomendable para la delimitación de zonas de manejo intralote en el contexto de agricultura de precisión.

5. Proponer una metodología de análisis para la clasificación de sitios intralote en función de la variación espacio-temporal de variables de suelo.

6. Realizar recomendaciones respecto a la modelación estadística contemporánea orientada al análisis de ensayos de fertilización sitio-específica.

CAPÍTULO I

HERRAMIENTAS ESTADÍSTICAS PARA EL ANÁLISIS DE DATOS ESPACIALES: APROXIMACIÓN UNIVARIADA

INTRODUCCIÓN

La variabilidad espacial en las propiedades o factores determinantes de la producción en sistemas agrícolas, ha sido observada por los productores desde los inicios de la agricultura (Bullock *et al*., 2007). Pero, sólo recientemente se dispone de la tecnología necesaria para cuantificar, comprender y manejar esa variabilidad. Las características del suelo y del cultivo varían en el espacio y en el tiempo siendo posible caracterizar esa variación a través de observaciones repetidas sobre el mismo lote tanto sea en el espacio como en el tiempo. Con la modernización de las prácticas agrícolas y la aplicación de la agricultura de precisión (AP) surgen nuevos desafíos respecto a la generación de tecnologías de información que permitan contemplar las variaciones y correlaciones esperables entre datos estructurados espacialmente. La variación espacial de los rendimientos es el resultado de una compleja interacción o covariabilidad de factores biológicos, edáficos, topográficos, antropogénicos o de manejo y climáticos (Bongiovanni *et al*., 2006). La mayoría de estos factores también se encuentran estructurados en el espacio y exhiben, por tanto, variabilidad espacial con menor o mayor grado de covariabilidad con el rendimiento logrado en cada sitio.

La variabilidad espacial de las propiedades edáficas, es el producto de factores formadores que operan e interactúan en una escala temporal y espacial continua (Trangmar *et al.*, 1985). La variabilidad inherente (no ocasionada por acciones antrópicas) resulta del producto de la interacción entre el clima, el material parental, la topografía, el material biológico y el tiempo cronológico actuante en la formación del suelo, mientras que la

variabilidad de origen antrópico es inducida por el manejo histórico y actual de los suelos, principalmente, a través de los aportes o pérdidas de carbono y otros nutrientes. El reconocimiento de la importancia de la variabilidad espacial en el uso del suelo ha llevado al estudio de la heterogeneidad del mismo, desde la escala global a la micro escala o escala fina. Esta variabilidad espacial genera correlaciones entre las observaciones de una misma variable registrada repetidamente en el espacio y por tanto los datos no pueden tratarse estadísticamente como datos independientes. Con datos de suelo, la autocorrelación espacial es típicamente positiva, es decir observaciones más cercanas en el espacio son más parecidas que observaciones más distantes (McBratney y Pringle, 1997; Mouazen *et al*., 2003; Iqbal *et al*., 2005; Gili, 2013). Bajo estructuras de autocorrelación espacial, los modelos de análisis estadísticos clásicos para estimar diferencias entre condiciones, resultan inapropiados (Schabenberger y Pierce, 2002).

Numerosas estrategias estadísticas para el análisis de correlaciones han sido propuestas en el estudio de variabilidad espacial. Los primeros índices formales para detectar la presencia de autocorrelación espacial se deben a Moran (Moran, 1948) y Geary (Geary, 1954). Otros procedimientos para modelar autocorrelación espacial se basan en el ajuste de datos en función de lo observado en muestras vecina (Papadakis, 1937; Webster, 1973; Vieira *et al.*, 1981; Wilkinson *et al.*, 1983), o en la modelación de funciones de variogramas (Cressie, 1985) así como en el uso de modelos que contemplan las correlaciones espaciales en la especificación de las varianzas y covarianzas de los términos error aleatorio (Zimmerman, 1991; Gilmour *et al*., 1997; Schabenberger y Pierce, 2002; Schabenberger y Gotway, 2004).

Una forma simple de explorar y detectar variabilidad espacial es agrupar las observaciones disponibles en función de la distancia que las separa en el espacio y analizar las diferencias entre observaciones como función de la distancia espacial entre ellas. Así se originaron los procedimientos clásicos basados en las funciones denominadas semivariogramas y correlogramas (Cressie, 1985, 1993). Tradicionalmente, estas funciones han sido estimadas a partir de procedimientos estadísticos basados en estimaciones por mínimos cuadrados ordinarios y/o mínimos cuadrados ponderados (Diggle y Ribeiro Jr, 2007; Gili, 2013). El desempeño de estos procedimientos geoestadísticos clásicos se sustenta fuertemente no sólo en el supuesto de errores con distribución normal sino también en el de procesos espaciales estacionarios (sin cambios sistemáticos).

Actualmente, la modelación de la variabilidad espacial se está realizando mediante la aplicación de modelos estadísticos más flexibles. Las aproximaciones más modernas se enmarcan dentro del marco teórico de los modelos lineales mixtos (MLM) (Schabenberger, 2006; West *et al*., 2007; Gbur *et al*., 2012). La estimación de parámetros en estos modelos se hace por métodos basados en la verosimilitud, máxima verosimilitud (ML) o máxima verosimilitud restringida (REML) (Searle *et al*., 1992; Gili, 2013), que permiten contemplar la correlación entre los datos.

En el presente Capítulo se describen, analizan e ilustran los principales procedimientos estadísticos que permiten describir la variabilidad espacialmente estructurada de características medidas sobre sitios georreferenciados dentro de un lote bajo AP. Las herramientas usadas en este Capítulo son de naturaleza univariada, es decir aplicables a las mediciones registradas a través de los sitios del lote para una única variable. Se provee de rutinas del software R (R Core Team, 2013) (Anexo 1 y 2) desde las cuales se pueden implementar cada uno de las aproximaciones descriptas. El conjunto de datos utilizado proviene del trabajo a campo realizado por investigadores del grupo Calidad y Manejo de Suelo y Agua de la EEA-INTA Balcarce. El Capítulo constituye una revisión sobre métodos estadísticos univariados actualmente usados para el análisis de datos espaciales en AP. La ilustración, en un conjunto de datos reales, permite comparar los procedimientos de análisis haciendo énfasis en los distintos tipos de conclusiones agronómicas que se pueden derivar desde su aplicación.

MATERIALES Y MÉTODOS

DATOS

Se analizaron datos provenientes de un lote de 65,4 ha en producción de trigo y soja de segunda, ubicado al sudeste pampeano de la provincia de Buenos Aires, Argentina. El clima de esta región es subhúmedo-húmedo, según índice hídrico de Thornthwaite (Burgos y Vidal, 1951), con una precipitación de 880 mm por año y una temperatura media anual de 13.3°C. Los suelos predominantes de esta región pertenecen al orden de los Molisoles, gran grupo Argiudoles o Paleudoles, desarrollados sobre sedimentos loéssicos, bajo

régimen údico-térmico. El sitio experimental esta principalmente constituido por la serie Azul (fina, mixta, térmica, Paleudol Petrocalcico) (SAGyP-INTA ,1989).

Se compilaron valores georreferenciados de conductividad eléctrica aparente (CE) (mS m^{-1}) en dos profundidades 0-30 cm (CE30) y 0-90 cm (CE90), Elevación (m), profundidad del suelo (tosca) (Pe) (cm) y rendimiento de soja (Sj) (t ha^{-1}) y trigo (Tg) (t ha^{-1}). Los valores de CE fueron tomados utilizando un sensor (Veris 3100, Division of Geoprobe Systems, Salina, KS) que utiliza el principio de la inducción electromagnética. El sensor Veris 3100 recorrió el lote en una serie de transectas paralelas espaciados a intervalos de 15 a 20 m, debido a que una separación de más de 20 m genera errores de medición (Farahani y Flynn, 2007). El instrumento fue calibrado, según las instrucciones del fabricante, antes de la recolección de los datos. Los datos de CE fueron simultáneamente georreferenciados con un DGPS (Trimble R3, Trimble Navegation Limited, USA) con una exactitud de medición submétrica y configurado para tomar la posición del satélite cada segundo. Los datos de elevación del terreno también se midieron con un DGPS y se procesaron para obtener una precisión vertical de entre 3 y 5 cm aproximadamente. Las mediciones de profundidad de tosca se realizaron utilizando un penetrómetro hidráulico (Gidding) acoplado a un DGPS en una grilla regular de 30 m. Para cuantificar el rendimiento en grano del cultivo se utilizó un monitor de rendimiento acoplado a un equipo de cosecha conectados a un DGPS.

Previo a los análisis que se ilustran, los datos fueron sometidos a un procedimiento de depuración vía la construcción de gráficos box-plots y gráficos de dispersión del índice de Moran Local (Anselin, 1996) para la identificación de valores extremos (outliers e inliers) (ver Capítulo IV). Debido a las diferentes resoluciones espaciales de las variables medidas, se impuso una grilla de 30 m × 30 m y los valores medidos de cada variable se asignaron a los nodos de la grilla más cercanos en el espacio. Cuando se tenía más de un dato para un nodo se asignó al mismo el promedio de las mediciones. La matriz de datos resultante estuvo conformada por n=664 sitios (filas) y p=5 variables (columnas). La base de datos (*CasoI.txt*) se encuentra disponible en:
https://drive.google.com/file/d/0B_8UVonay55COUNrTy1zempab3c/edit?usp=sharing

PROCEDIMIENTOS ESTADÍSTICOS PARA EL ANÁLISIS DE VARIABILIDAD ESPACIAL

ÍNDICES DE AUTOCORRELACIÓN ESPACIAL

La autocorrelación espacial mide la correlación lineal entre los valores de una variable en una determinada posición con valores de la misma variable en otras posiciones en el espacio. Permite evaluar si una variable tiende a asumir valores similares en unidades geográficamente cercanas (Anselin, 2001). Una propiedad de los datos autocorrelacionados espacialmente es que los valores no son aleatorios en el espacio, sino que están relacionados entre sí y la magnitud de esa correlación depende de las distancias que los separan (Lee y Wong, 2001). La autocorrelación espacial puede presentarse con valores positivos o negativos; existe autocorrelación positiva cuando valores similares de una variable aleatoria tienden a aglomerarse en el espacio, es decir valores más cercanos son más parecidos; por otra parte, la autocorrelación negativa se presenta cuando las unidades geográficas de observación más cercanas tienden a tener valores opuestos o más distintos.

El análisis exploratorio de datos espaciales (ESDA, por sus siglas en inglés) se realiza a partir de un conjunto de técnicas utilizadas para describir y visualizar distribuciones espaciales, detectar patrones de asociación espacial y aglomeraciones, así como para sugerir regímenes espaciales u otras formas de heterogeneidad espacial. La autocorrelación espacial puede ser medida en términos de su intensidad; una autocorrelación espacial positiva fuerte significa que los valores de la variable en sitios cercanos geográficamente, están altamente relacionados o son muy parecidos y, consecuentemente, emergen aglomeraciones espaciales de los datos. En otros casos la distribución de la variable de interés puede presentar una autocorrelación débil, o incluso mostrar un patrón de dispersión espacial aleatorio. Para cuantificar la magnitud de la estructuración espacial de una variable existen índices entre los que se encuentran el índice de Moran (Moran, 1948) y el índice de Geary (Geary, 1954).

El cálculo del índice o coeficiente de Moran de autocorrelación espacial en un espacio continuo requiere la definición de una matriz de ponderación espacial. Ésta suele tener elementos binarios (0/1) para indicar cuáles son las observaciones que pertenecen al

vecindario de cada dato, o cuáles son las observaciones "conectadas" con cada dato. No obstante, también pueden tener elementos continuos que pueden ser entendidos como un coeficiente de continuidad que mide el grado de conexión entre cada par de datos.

Para calcular el índice de Moran se mide la variable de interés en el sitio *i*-ésimo y se compara su valor con el valor promedio de la variable en los sitios de su vecindario. La expresión del índice es:

$$MI = \frac{N\sum_i \sum_j \mathbf{W}_{i,j}(X_i - \bar{X})(X_j - \bar{X})}{(\sum_i \sum_j \mathbf{W}_{i,j})\sum_i (X_i - \bar{X})^2} \qquad (1.1)$$

donde N es el número total de observaciones, X_i es el valor de la variable en un sitio particular (posición *i*) y X_j es el valor de la variable en otro sitio (posición *j*). El elemento W_{ij} de la matriz de ponderaciones $\mathbf{W}$, es el peso aplicado a la comparación de las observaciones en la posición *i* y la posición *j*. Usualmente, para el cálculo del Índice de Moran se utilizan redes de conexión que derivan en un matriz $\mathbf{W}$ binaria, es decir compuesta por ceros y unos (si la posición *j* es adyacente a la posición *i*, el término $_{ij}$ recibe un peso de 1 y si no, de 0). Otra posibilidad para construir la matriz $\mathbf{W}$ es relacionar los elementos con la distancia *d* entre las posiciones de manera inversamente proporcional, es decir: $W_{ij} = 1/d_{ij}$.

La red de vecindarios también puede ser definida en función de la distancia Euclídea considerando puntos vecinos a aquellos contiguos ubicados entre un límite inferior y superior, previamente preestablecido. Otra red de conexión es la obtenida por el método de triangulación de Delaunay (Lee y Schachter, 1980), recomendado para construir gráficos de vecindario cuando las entidades se encuentran distribuidas en forma homogénea en el espacio. La red de conexión de Gabriel (Gabriel y Sokal, 1969) es un subconjunto de la red de Delaunay que no incluye conexiones periféricas. Las redes de conexión también pueden ser adaptadas manualmente pudiéndose excluir contactos entre sitios cercanos o incluir relaciones entre sitios lejanos, siguiendo criterios agronómicos

como por ejemplo la existencia de fertilizaciones o controles realizados dentro de un lote en sitios puntuales.

El índice de Moran varía entre –1 y 1. Cuando la autocorrelación es alta, el coeficiente será cercano a –1 o 1. Un valor cercano a 1 indica una alta autocorrelación positiva, mientras que valores cercanos a –1 indican autocorrelación negativa. Un valor de cero significa que no existe un patrón espacial o que la dispersión de las observaciones en el espacio es completamente aleatoria. El índice de Moran es una estadística global que considera los valores de todas las observaciones, el cual sugiere numéricamente la existencia de aglomeraciones espaciales.

Para evaluar la estructura local de autocorrelación espacial se puede utilizar el índice Moran Local (*Ii*) (Anselin, 1995). El *Ii* es básicamente el índice de Moran aplicado a cada sitio individualmente, que da idea del grado de similitud o diferencia entre el valor de la observación en un sitio determinado con respecto al valor de los sitios vecinos. Los valores positivos de *Ii* se corresponden con agrupamiento (*clusters*), mientras que los valores negativos se corresponden con valores extremos o atípicos (*outliers* o *inliers* espaciales), indicando para la observación que su valor rompe con la tendencia observada en sus vecinos (Anselin, 1995).

El Índice de Geary, es similar al índice de Moran, pero en su numerador no mide la interacción a través del producto cruzado de las desviaciones con respecto a la media, sino que expresa la magnitud de las desviaciones entre observaciones en las diferentes localizaciones. La expresión del índice es:

$$GI = \frac{(N-1)\sum_i \sum_j W_{i,j}(X_i - X_j)^2}{2(\sum_i \sum_j W_{i,j})\sum_i (X_i - \bar{X})^2} \qquad (1.2)$$

El valor índice de Geary se encuentra en el intervalo [0,2]. Si no hay autocorrelación espacial, el valor esperado de *GI* es 1. Valores del índice entre 1 y 2 indican autocorrelación espacial negativa, y entre 0 y 1 autocorrelación espacial positiva. Este índice se relaciona inversamente con el índice de Moran, es decir valores más cercanos a 0 sugieren autocorrelaciones positivas más fuerte. Al enfatizar las diferencias

entre pares de observaciones más que la covariación entre ellos, el índice de Geary no provee una inferencia agronómica idéntica como el índice de Moran.

Para evaluar la significancia estadística de estos índices es posible utilizar procedimientos del tipo Monte Carlo (Babai *et al*., 1995). Las ubicaciones son permutadas para obtener la distribución de los índices bajo la hipótesis nula de distribución aleatoria. El índice de Moran y de Geary pueden calcularse en R mediante las librerías "spdep" (Bivand *et al*., 2013a) y "ape" (Paradis *et al*., 2004).

GEOESTADÍSTICA

El término geoestadística, fue acuñado por Matheron para describir su trabajo referente a problemas de predicción espacial sobre variables regionalizadas. Estas ideas fueron desarrolladas extensamente e independientemente de la corriente principal de la estadística. Las herramientas que se usan en la práctica de la geoestadística, reflejan su origen independiente y esto genera diferencias considerables, como por ejemplo, el hecho de que la inferencia es frecuentemente de naturaleza *ad hoc*, con modelos estocásticos no declarados explícitamente. Debido a esto, el poco uso de los métodos basados en verosimilitud, que son centrales en la estadística moderna. Diggle *et al.* (1998) usaron el término geoestadística, basada en modelos, para describir el tratamiento de datos georreferenciados desde la aplicación de métodos estadísticos formales bajo un modelo estocástico asumido explícitamente.

La teoría de variables regionalizadas, bajo la cual se definen los semivariogramas, adopta una perspectiva estocástica de los procesos espaciales. Según esta, cada dato es una realización de un proceso aleatorio, por lo cual existe una distribución de probabilidad asociada al mismo. Uno de los supuestos más importantes que se asume es que las distribuciones de probabilidad asociadas a cada sitio son normales y tienen la misma media y varianza. Estos últimos supuestos se conocen como estacionariedad de primer y segundo orden, respectivamente. Una forma de corroborar que se cumpla el supuesto de estacionariedad de primer orden es realizando regresiones de la variable repuesta con las coordenadas geográficas del sitio. En el caso de encontrar una relación lineal significativa, es decir una tendencia longitudinal o latitudinal, se intenta descontar esa tendencia y

trabajar con los residuos. De esta manera, la tendencia promedio (cambios de la media) se remueve y la autocorrelación espacial se estudia en la porción aleatoria de la variable representada por los residuos.

Bajo este enfoque, un primer paso para analizar la presencia de autocorrelación espacial en un continuo es construir un semivariograma empírico; el semivariograma es el punto de partida de la geoestadística. La función semivariograma de un proceso estacionario, denotado por $\gamma(\mathbf{s}_i - \mathbf{s}_j)$, es sólo función de la diferencia entre las coordenadas $(\mathbf{s}_i - \mathbf{s}_j)$ y puede expresarse como:

$$\gamma(\mathbf{s}_i - \mathbf{s}_j) = \gamma(\mathbf{h}) = \frac{1}{2}\left\{\text{Var}\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]\right\} \tag{1.3}$$

donde **h** es la distancia espacial entre los puntos muestrales $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$ que se suponen sobre un espacio continuo. Entre otros factores a tener en cuenta para ajustar el semivariograma se encuentra el tamaño de muestra con el que se estima cada semivarianza; comúnmente se recomienda que la estimación se realice con al menos 30 pares de puntos. La distribución de los puntos en el espacio determinará para qué *lags* esto es posible. Si el semivariograma no solo depende de la longitud del vector h sino también de la dirección del vector entonces el semivariograma es anisotrópico (Shabenberger y Gotway, 2004).

Los parámetros del semivariograma son: la varianza nugget o efecto pepita (C_0), la varianza estructural (C) y el rango (R). La asíntota es llamada el umbral del semivariograma o C y el lag o distancia h en el cual la meseta es alcanzada, es R. Observaciones $Z(s_i)$ y $Z(s_j)$ para las cuales $||\, Z(s_i) - Z(s_j)|| \geq R$ son incorrelacionadas. Cuando el semivariograma alcanza la meseta asintóticamente (como en el caso de los modelos de semivarianza exponencial), se define un rango práctico (R_P). Este parámetro representa la distancia en el cual la semivarianza alcanza el 95% de la varianza umbral o total. Puede ocurrir que el semivariograma no alcance la meseta y esto frecuentemente es debido a que el proceso no es estacionario o a que el proceso es estacionario de segundo orden pero el lag más grande para el cual el semivariograma puede ser estimado es más chico que R (problema de grilla).

En la práctica el semivariograma empírico $\hat{\gamma}(h)$ puede no pasar a través del origen. La ordenada al origen del semivariograma representa a C_0 , por lo tanto $C_0 = \lim_{h\to 0} \hat{\gamma}(h) \neq 0$. Este parámetro representa la suma de errores aleatorios o no espaciales, así como errores asociados con la variabilidad espacial a escalas más finas que la usada para realizar las mediciones (Schlesinger *et al.*, 1996). Un alto valor de C_0 indica que la mayoría de la variación ocurre en distancias cortas o fue debida a errores de medición (Schlesinger *et al.*, 1996).

La varianza umbral se obtiene sumando las varianzas antes mencionadas $(C_0 + C)$ y es la varianza de observaciones independientes, es decir observaciones que fueron tomadas a mayor distancia que R.

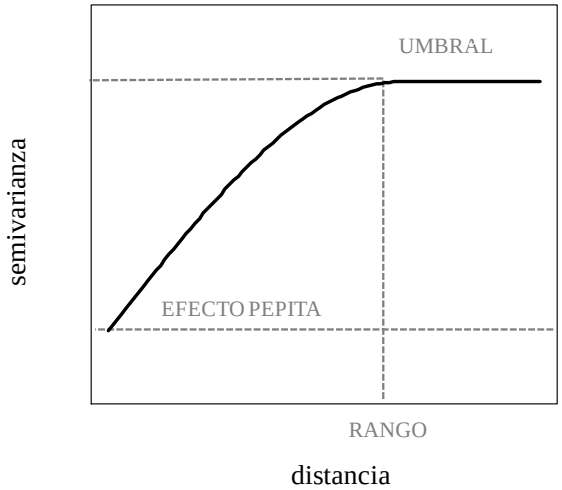


Fig. 1.2. Semivariograma esférico. Se representan los tres parámetros que lo definen: el rango, el umbral y el efecto pepita.

Una medida común del grado de estructura espacial, en este contexto, es la varianza estructural relativa (RSV).

$$RSV = \left(\frac{C}{C+C_0}\right) \times 100\% \qquad (1.4)$$

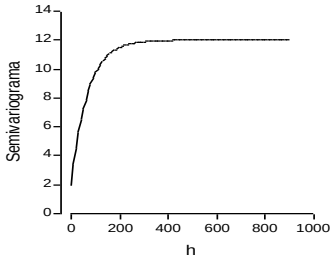
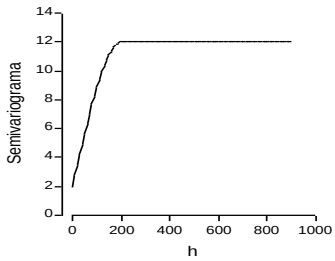
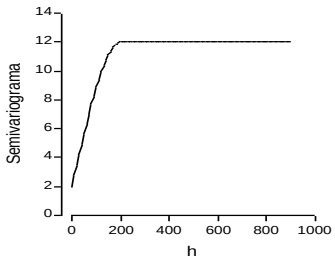
Un valor alto de RSV indica que las predicciones geoestadísticas serán más eficientes que aquellas obtenidas con métodos de predicción que ignoran la información espacial ya que existe una mayor estructuración espacial. Zimback (2001) establece que el

grado de dependencia en función del RSV entre muestras puede ser clasificado como: ≤ 25% bajo, entre 25% y 75% medio y de ≥ 75% alto. También se puede usar el cociente $\frac{C_0}{C_0+C}$ y en función de este se define estructura espacial fuerte cuando el cociente es ≤ 25%, intermedia si el mismo se encuentra entre 25% y 75% y débil si el mismo es mayor al 75%.

El semivariograma empírico $\hat{\gamma}(h)$, es un estimador insesgado de $\gamma(h)$, pero provee solo estimaciones de un conjunto finito de distancias. Para obtener estimaciones de $\gamma(h)$, en cualquier distancia arbitraria, al semivariograma empírico debe ajustársele un modelo teórico. Por ello, el análisis geoestadístico cumple los siguientes dos pasos: 1) obtención del semivariograma empírico y 2) ajuste de un modelo teórico de semivariograma al semivariograma empírico. Las funciones que sirven como modelos de semivariograma deben ser condicionalmente definidas positivas. Existen distintos modelos teóricos para funciones semivariogramas, entre los que se encuentran el modelo nugget, el lineal, el esférico, el gaussiano y el exponencial (Tabla 1.2). El semivariograma de un proceso de ruido blanco (modelo nugget), donde las $Z(s_i)$ se comportan como muestras aleatorias, todas con igual media y varianza sin correlación entre ellas. Un modelo sin estructura espacial con sólo efecto nugget, es comúnmente un modelo apropiado si la menor distancia de muestreo en los datos es mayor que el rango del proceso espacial.

El modelo esférico es uno de los más populares entre los modelos de semivariograma. Tiene dos características principales: un comportamiento lineal cerca del origen y el hecho de que a la distancia R el semivariograma encuentra la meseta y después de esta se mantiene llano. El modelo exponencial se aproxima a la meseta del semivariograma (C) asintóticamente cuando $\|h\| \to \infty$. En la parametrización mostrada en la Tabla 1.2, el parámetro R es el rango práctico del semivariograma. Frecuentemente el modelo puede encontrarse en una parametrización donde el exponente es $-\|h\|/R$. Entonces el R_p corresponde a $3R$. Para el mismo rango y meseta de un modelo esférico, el modelo exponencial alcanza el rango más rápidamente, es decir, a menor distancia que el modelo esférico.

Tabla 1.2. Modelos teóricos de semivariogramas. Funciones de semivariograma para el modelo exponencial, esférico y gaussiano. $\mathbf{C_0}$=2, **C**=10 y **R**=200.

Modelo Exponencial	
	$\gamma(h) = \begin{cases} C_0 & h = 0 \\ C_0 + C\left\{1 - exp\left\{-3\dfrac{h}{R}\right\}\right\} & h \neq 0 \end{cases}$
Modelo Esférico	
	$\gamma(h) = \begin{cases} C_0 & h = 0 \\ C_0 + C\left\{\dfrac{3}{2}\dfrac{h}{r} - \dfrac{1}{2}\left(\dfrac{h}{R}\right)^3\right\} & h \neq 0 \end{cases}$
Modelo Gaussiano	
	$\gamma(h) = \begin{cases} C_0 & h = 0 \\ C_0 + C\left\{1 - exp\left\{-3\dfrac{h^2}{R^2}\right\}\right\} & h \neq 0 \end{cases}$

En particular, es importante notar que si se realiza un análisis basado en semivariogramas y se pretende comparar los parámetros de los semivariogramas obtenidos bajo distintas condiciones, la utilización de modelos teóricos diferentes resulta poco útil. Hay que tener en cuenta que, por ejemplo, los rangos del modelo esférico y el exponencial no son directamente comparables. El modelo esférico es el único que tiene un umbral verdadero, ya que tanto el modelo exponencial como el gaussiano alcanzan el umbral de forma asintótica, o lo que es lo mismo, no lo alcanzan nunca y el modelo lineal ni siquiera

tiene umbral. En consecuencia, los rangos no son directamente equivalentes entre modelos. En este caso, es más conveniente elegir un único modelo para las condiciones a comparar.

Los modelos de semivariograma son no lineales a excepción del modelo solo pepita (nugget); y para la minimización de estas funciones se requiere métodos no lineales. El método de ajuste por mínimos cuadrados ponderados (WLS) es común en la práctica. Se debe tener presente que es un método aproximado. El tamaño del conjunto de datos a partir del cual el modelo de semivariograma es ajustado depende del número *k* de clases *lag* que se elija. Los valores de las clases de *lag* en las cuál el número de pares no es mayor a 30 debieran ser removidos si se ajusta el semivariograma por mínimos cuadrados. Journel y Huijbregts (1978) recomiendan solo usar *lags* menores a la mitad del máximo *lag* en el conjunto de datos.

MODELOS LINEALES MIXTOS

Numerosos problemas de estimación de niveles medio y variabilidad de variables edáficas y de rendimientos medidos espacialmente han sido ya tratados empleando modelos lineales mixtos (MLM) (Hong *et al*., 2005; Kravchenko *et al*., 2005; Kravchenko *et al*., 2006; Griffin, 2010; Lawes *et al*., 2012; Rodrigues *et al*., 2013). La mayor ventaja de los MLM es la generalidad en la inferencia luego de modelar la autocorrelación espacial entre las observaciones (Gili, 2013). La estimación de los parámetros de varianza y covarianza (también parámetros del semivariograma) puede realizarse, en este marco de trabajo, de manera simultánea a la de aquellos parámetros relacionados a la estructura de media del proceso (tendencias a gran escala en una o más dimensiones). La estimación de parámetros en estos modelos se hace por métodos basados en la verosimilitud: máxima verosimilitud (ML) o por máxima verosimilitud restringida (REML) (Searle *et al*., 1992). En este contexto se ajusta un modelo directamente a los datos, y no a las semivarianzas como en la geoestadística clásica. El MLM ajustado puede contemplar la correlación espacial mediante la incorporación de efectos aleatorios desde los cuales se inducen correlaciones espaciales entre subconjuntos de datos o, alternativamente, con la especificación explícita de la estructura de correlación espacial en la matriz de varianzas-covarianzas de los términos de error del modelo. En este último caso, la estructura de

covariación espacial se define considerando que la misma es función de la distancia entre la separación de las observaciones y las más utilizadas para datos de suelo son las funciones espacial esférica, exponencial y gaussiana (Schabenberger y Gotway, 2004). Para aplicar MLM y lograr estimadores basados en la función de verosimilitud es necesario que la variable de interés presente distribución normal. Patterson y Thompson (1971) desarrollaron el método de REML, que ajusta por los grados de libertad de los efectos fijos (estructura de medias) antes de estimar los componentes de varianza. Estas estimaciones REML son recomendadas para obtener componentes de varianza en el contexto de los MLM. No obstante la estimación ML es preferible para las etapas del proceso de modelización en la que se comparen o se evalúen parámetros de la estructura de medias del MLM (efectos fijos).

La ecuación matricial para el modelo lineal mixto es:

$$y = X\beta + Zu + e \tag{1.5}$$

donde y es un vector de observaciones, X es una matriz de valores de variables independientes en el caso del modelo de regresión o la matriz de diseño en el caso del modelo de análisis de la varianza, β es el vector de parámetros (o efectos fijos), e es el vector de errores, Z es la matriz de incidencia de los efectos aleatorios y u, vector de efectos aleatorios que usualmente se asume distribuido $N(0, \mathrm{G})$. Sobre el vector e se supone distribución $N(0, \mathrm{R})$, e es definido como $e = y - E(y|u) = y - (X\beta + Zu)$. La matriz R es modelada como $R = \sigma^2 I$ cuando se considera que los términos de error (generalmente asociados a la parcela en experimentación agrícola) son independientes y tienen la misma varianza σ^2. Los términos aleatorios en u se suponen independientes de los términos aleatorios en e. Resumiendo matricialmente los supuestos usuales sobre la esperanza y la varianza de las componentes aleatorias, se tiene que:

$$E\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \; y \; Var\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$$

Cuando se asume distribución normal para el vector de observaciones, la función de densidad queda completamente determinada por el vector de valores esperados y la matriz de varianzas y covarianzas. La matriz de varianzas y covarianzas de **y** está dada por:

$$V(y) = V = V(X\beta + Zu + e) \qquad (1.6)$$
$$= ZV(u)Z' + V(e)$$
$$= ZGZ' + R$$

Los supuestos clásicos de independencia y homogeneidad de varianzas para los términos aleatorios del modelo lineal general (muestreo ideal) se flexibilizan en el marco del modelo mixto general. Tanto la estructura de correlaciones como la presencia de varianzas heterogéneas puede ser especificada a través de la modelación de las matrices de covarianza G y R (Balzarini *et al*., 2004). Esta característica hace a los modelos mixtos muy interesantes para el análisis de datos de suelos que ya se conocen presentan alta heterogeneidad y variabilidad de tipo espacial.

Al ajustar distintos modelos a un mismo conjunto de datos, es necesario utilizar criterios para la comparación de los ajustes. Una de las herramientas utilizadas para la selección de modelos se conoce como criterios de información. Los criterios de información proporcionan una forma de evaluar el ajuste de un modelo basado en el logaritmo de su valor de verosimilitud (log-likelihood), después de aplicar una función de penalización debida a la cantidad de parámetros que se estima en el ajuste del modelo. Una característica clave de los criterios de información es que proporcionan una forma de comparar cualquier dos modelos ajustados con el mismo set de datos. La interpretación de los criterios se basa en que un menor valor del criterio indica un "mejor" ajuste. El criterio de información de Akaike (AIC) se puede calcular sobre la base (ML o REML) del log-likelihood, $l(\beta,\theta)$, de la siguiente manera (Akaike, 1973):

$$AIC = -2l(\beta,\theta) + 2p \qquad (1.7)$$

En 1.7, p representa el número total de parámetros que estima el modelo. El criterio AIC en efecto "penaliza" el ajuste de un modelo (*i.e.* la verosimilitud) en función del número de parámetros que estima.

El criterio de información de Bayes (BIC) también es comúnmente utilizado y se calcula de la siguiente manera:

$$BIC = -2l(\beta,\theta) + p \times \ln(n) \tag{1.8}$$

El BIC aplica una penalización mayor para los modelos con más parámetros respecto a AIC, ya que multiplica el número de parámetros estimandos por el logaritmo natural de n, siendo n el número total de observaciones utilizadas en la estimación del modelo.

Guerin y Stroup (2000) compararon el desempeño de AIC y BIC en la selección del modelo de covarianzas y concluyeron que AIC tiende a dar lugar a un modelo más complejo, pero con un mejor control de tasa de error de Tipo I que el criterio BIC. Por lo tanto, si el interés está centrado en los efectos del tratamiento, AIC sería un mejor criterio. Sin embargo, Gurka (2006), sugiere que no hay un criterio de información que se destaque como el mejor para ser utilizado cuando se seleccionan MLM.

Otra de las herramientas estadísticas utilizadas para la selección de modelos, en el contexto de los MLM, es la prueba del cociente de verosimilitud (LRT, Likelihood Ratio Tests) (West *et al*., 2007). Esta se basa en una prueba de hipótesis que se formula en el contexto de dos modelos anidados. El modelo más general, abarca tanto la hipótesis nula y alternativa, es denominado modelo de referencia. El segundo modelo, más simple, satisface la hipótesis nula y se denomina modelo anidado. La única diferencia entre estos dos modelos es que el modelo de referencia contiene todos los parámetros, mientras que el modelo anidado (hipótesis nula) no contiene aquellos que se suponen podrían ser iguales a cero.

La prueba LRT puede emplearse para probar hipótesis acerca de los parámetros de covarianza y también para parámetros de efectos fijos en el contexto de los MLM. El estadístico LRT se calcula como se muestra en la siguiente ecuación:

$$-2\log(\frac{L_{anidado}}{L_{referencia}}) = -2\log(L_{anidado}) - (-2\log(L_{referencia})) \sim \chi^2_{gl} \qquad (1.9)$$

donde $L_{anidado}$ y $L_{referencia}$ se refiere al valor de la función de verosimilitud evaluada en las estimaciones ML o REML de los parámetros en el modelo anidado y de referencia, respectivamente. La teoría indica que el estadístico LRT se aproxima asintóticamente a una distribución χ^2, en el que el número de grados de libertad (gl) se obtiene de la diferencia de parámetros entre el modelo de referencia y el anidado.

Utilizando la ecuación 1.9, pueden probarse hipótesis acerca de los parámetros de los MLM. Si el estadístico LRT es suficientemente grande, hay evidencias para rechazar la hipótesis nula (modelo anidado) y por lo tanto preferir el modelo de referencia. Si los valores de verosimilitud de los dos modelos están muy cerca, el estadístico LRT será pequeño y la evidencia es a favor del modelo anidado (hipótesis nula).

Al probar hipótesis acerca de parámetros de covarianza en un MLM, la estimación REML debe ser utilizada tanto para el modelo de referencia como para el anidado. La estimación REML ha demostrado que reduce el sesgo en las estimaciones ML de los parámetros de covarianza (Morrell, 1998). Se supone que los modelos anidado y de referencia tienen el mismo conjunto de parámetros de efectos fijos, pero diferentes conjuntos de parámetros de covarianza. Cuando se prueba la significancia de un efecto aleatorio, la distribución del estadístico LRT es dependiente del tipo de comparación que se realiza.

Por ejemplo, cuando debe probarse que un único efecto aleatorio (o en el caso de modelos espaciales el rango o el nugget) deba o no mantenerse en un modelo. No se prueba directamente la hipótesis acerca del efecto aleatorio sino que en lugar de ello, se pone a prueba si la varianza correspondiente es igual a cero y el modelo reducido es un modelo sin efecto aleatorio, es decir, ya no un MLM. La distribución de la estadístico de la prueba LRT en este caso es una mezcla de distribuciones χ^2 (χ^2_0 y χ^2_1), cada una con peso 0.5 (Verbeke y Molenberghs, 2000). También en el caso de que los modelos que se comparan difieran en dos efectos aleatorios y se requiera probar si uno de ellos se puede omitir, es necesario comprobar no solo si la varianza del efecto es igual a cero sino

también se requiere probar si la covarianza es igual a cero. Aquí también, la distribución asintótica del estadístico LRT bajo hipótesis nula es una mezcla de distribuciones χ^2 (χ_1^2 y χ_2^2), cada una con un peso de 0.5. Para otros tipos de comparaciones que involucran más de dos efectos aleatorios el LRT se distribuye directamente como una χ^2 con *gl* igual a la diferencia en el número de parámetros entre ambos modelos.

PREDICCIÓN Y CONSTRUCCIÓN DE MAPAS

El método de interpolación llamado Kriging (Webster y Oliver, 2007) provee una solución al problema de predicción basado en un modelo continuo de variación espacial estocástica. En su formulación original una predicción basada en Kriging de una variable aleatoria en una ubicación, en particular, fue simplemente realizado mediante una combinación lineal de promedios ponderados de los datos observados en los vecinos de la posición de interés (Kriging lineal). Los pesos o ponderaciones son asignados a partir de los datos muestrales del vecindario y son estimados de manera tal que se minimice la varianza de la predicción y que las estimaciones sean insesgadas.

Las técnicas de Kriging son métodos para predecir $Z(x_0)$ basados en de supuestos acerca del modelo espacial y requerimientos acerca del predictor $p(Z;s_0)$: 1) $p(Z;s_0)$ es una combinación lineal de los valores observados $Z(s_1), \ldots, Z(s_n)$, 2) $p(Z;s_0)$ es insesgado en el sentido que E[$p(Z;s_0)$]= $E[Z(s_0)]$ y 3) $p(Z;s_0)$ minimiza el error de predicción cuadrático medio.

Existen tres métodos básicos de Kriging se distinguen respecto a la estructura de medias del modelo espacial $Z(s) = \mu(s) + e(s)$. Éstos se resumen en la Tabla 1.10.

Tabla 1.10. Kriging Simple, Ordinario y Universal.

Método	$\mu(s)$	e(s)
Kriging Simple (SK)	$\mu(s)$ es conocida	Estacionaridad de segundo orden
Kriging Ordinario (OK)	$\mu(s) = \mu$, μ es desconocida	Estacionaridad de segundo orden
Kriging Universal (UK)	$\mu(s) = x'\beta$, β es desconocido	Estacionaridad de segundo orden

Los predictores de Kriging pueden ser calculados para cada ubicación para la cual se desee hacer la predicción. Aunque sólo el vector de covarianzas entre $Z(s_0)$ y $Z(s)$ debe ser recalculado cada vez que s_0 cambia, el problema computacional es considerable. Una solución es considerar para la predicción de $Z(s_0)$ sólo datos de puntos dentro de la vecindad de s_0, esto es conocido como "*kriging neighborhood"* o Kriging local. El Kriging local esencialmente asigna pesos $\lambda(s_0) = 0$ para todo los puntos s_i fuera de la zona en la que se quiere predecir. Aún cuando el mejor estimador lineal insesgado es aquel obtenido con todos los datos, los predictores con Kriging local no son los mejores, pero son frecuentemente usados por cuestiones computacionales.

Una decisión a tomar a la hora de realizar la predicción espacial es la selección del tipo de interpolación: puntual o por bloques. Mientras que la interpolación puntual se estima el valor de la variable en un sitio del espacio, en la interpolación por bloques esta estimación se corresponde con la media de un área predeterminada que rodea a ese sitio. La elección del método dependerá del objetivo concreto del estudio, pero en la mayoría de los casos la interpolación por bloques (que produce un "suavizado" de las estimas) correlaciona mejor con los valores verdaderos, siendo generalmente más exacta que la interpolación puntual (Isaaks y Srivastava 1989). La posibilidad de trabajar con bloques circulares o irregulares a la hora de realizar estimas utilizando kriging está implementada en la librería "gstat" (Pebesma, 2004) del software R.

Una vez que se ha hecho la predicción en un conjunto de puntos diferentes de los muestrales por vía de Kriging, se puede elaborar un mapa de contorno para obtener una representación de la variable de interés. En el caso de los mapas de contorno, se divide el área de estudio en una grilla y se hace la estimación en cada uno de los nodos de la misma.

RESULTADOS

CÁLCULO DEL ÍNDICE DE MORAN Y DE GEARY

Su aplicación al estudio de variabilidad espacial en AP puede encontrarse en diversos trabajos (Lambert *et al*., 2004; Roel y Plant, 2004; Mzuku *et al*., 2005). Para

determinar la matriz de ponderadores espaciales, para el caso de estudio, se definieron los vecindarios para cada sitio mediante una red de conexión construida en base a la distancia Euclidea. Se consideraron sitios vecinos a aquellos contiguos ubicados hasta 30 m de distancia (Fig. 1.1). Para este procedimientos se utilizó la librería "spdep" del software R (Bivand *et al*., 2013a).

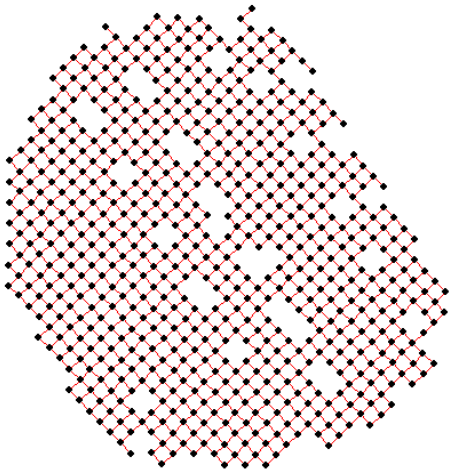


Fig. 1.1. Red de conexión calculada mediante la distancia Euclidea. Se consideran sitios vecinos a aquellos separados a una distancia menor a 30 m entre sitios georreferenciados.

Tabla 1.1. Índices de autocorrelación espacial en variables de suelo y rendimiento.

Variable	Índice de Moran		Índice de Geary	
	MI	p-valor	GI	p-valor
Suelo				
CE30, mS m^{-1}	0.34	0.001	0.66	<0.001
CE90, mS m^{-1}	0.21	0.001	0.79	<0.001
E, m	0.46	0.001	0.53	<0.001
Pe, cm	0.27	0.001	0.73	<0.001
Rendimiento				
Sj, t ha^{-1}	0.40	0.001	0.59	<0.001
Tg, t ha^{-1}	0.65	0.001	0.35	<0.001

CE30: conductividad eléctrica aparente a 30 cm de profundidad, CE90: conductividad eléctrica aparente a 90 cm de profundidad, E: elevación, Pe profundidad tosca, Sj: rendimiento de soja; Tg: rendimiento de trigo.

Todas las variables analizadas mostraron autocorrelación espacial significativa y positiva (Tabla 1.1), tanto con el índice de Moran global (MI) como con el de Geary (GI). Para ambos índices, la variable de suelo y terreno con mayor autocorrelación global positiva fue la elevación (E) y entre los rendimientos, el más estructurado espacialmente fue el de trigo.

IMPLEMENTACIÓN DEL ANÁLISIS BASADO EN SEMIVARIOGRAMA

Se realizó, un análisis descriptivo para todas las variables. Se calcularon media, mediana, desvío estándar y coeficiente de variación (CV%) (Tabla 1.3). Para estudiar la distribución de cada variable se graficaron los histogramas para cada una de ellas. Adicionalmente se verificó el cumplimiento del supuesto de estacionariedad a través de un análisis de regresión por mínimos cuadrados ordinarios usando las coordenadas como variables regresoras (Giraldo, 2003). Se obtuvo el semivariograma empírico y sobre este se ajustaron, por WLS, los modelos exponencial y esférico. El cuadrado medio del error (CME) fue el criterio usado para la selección del mejor modelo. Para ambos se calculó la RVS como medida del grado de estructuración espacial. El análisis geoestadístico fue realizado con la librería "geoR" (Ribeiro Jr. y Diggle, 2001) del software R.

La estadística descriptiva muestra que la variable E presentó la menor variabilidad relativa (CV=1.27), Si bien E muestra poca variabilidad en los lotes pampeanos dedicados al cultivo extensivo de grano, los CV de CE y de Pe fueron altos. Para todas las variables medidas, la similitud encontrada entre media y mediana sugiere que las distribuciones podrían considerarse como simétricas.

Tabla 1.3. Estadística descriptiva para las propiedades de suelo y rendimiento de un lote con mediciones georreferenciadas en 664 sitios.

Variables	Media	Mediana	DE	CV (%)	Min.	Max.
Suelo						
CE30, mS m^{-1}	31.12	31.22	7.16	23.01	15.90	58.00
CE90, mS m^{-1}	30.40	30.15	5.84	19.22	13.77	54.60
E, m	141.77	141.88	1.80	1.27	135.19	146.18
Pe, cm	73.32	71.00	21.97	29.97	20.00	101.00
Rendimiento						
Sj, t ha^{-1}	1.79	1.78	0.34	19.17	1.06	2.76
Tg, t ha^{-1}	3.76	3.67	0.62	16.62	2.33	5.66

CE30: conductividad eléctrica aparente a 30 cm de profundidad, CE90: conductividad eléctrica aparente a 90 cm de profundidad, E: elevación, Pe profundidad tosca, Sj: rendimiento de soja; Tg: rendimiento de trigo.

Como puede observarse en la Fig. 1.3 todas las variables presentaron una distribución aproximadamente normal. El estudio de las tendencias a gran escala se realizó a través de regresiones lineales (Tabla 1.4). Si bien, todas las variables excepto CE30 y Pe

presentaron tendencias estadísticamente significativas, los R^2 de estos ajustes fueron bajos por lo que se decidió trabajar con los datos originalmente relevados sin necesidad de remover las tendencia sistemáticas. Algunos autores sugieren que cuando el coeficiente de determinación es menor al 20% esta decisión no es objetable (Kerry y Oliver, 2004; Alesso *et al*., 2012). En la Fig. 1.4 se muestran los semivariogramas empíricos y los teóricos ajustados para cada variable (modelo exponencial).

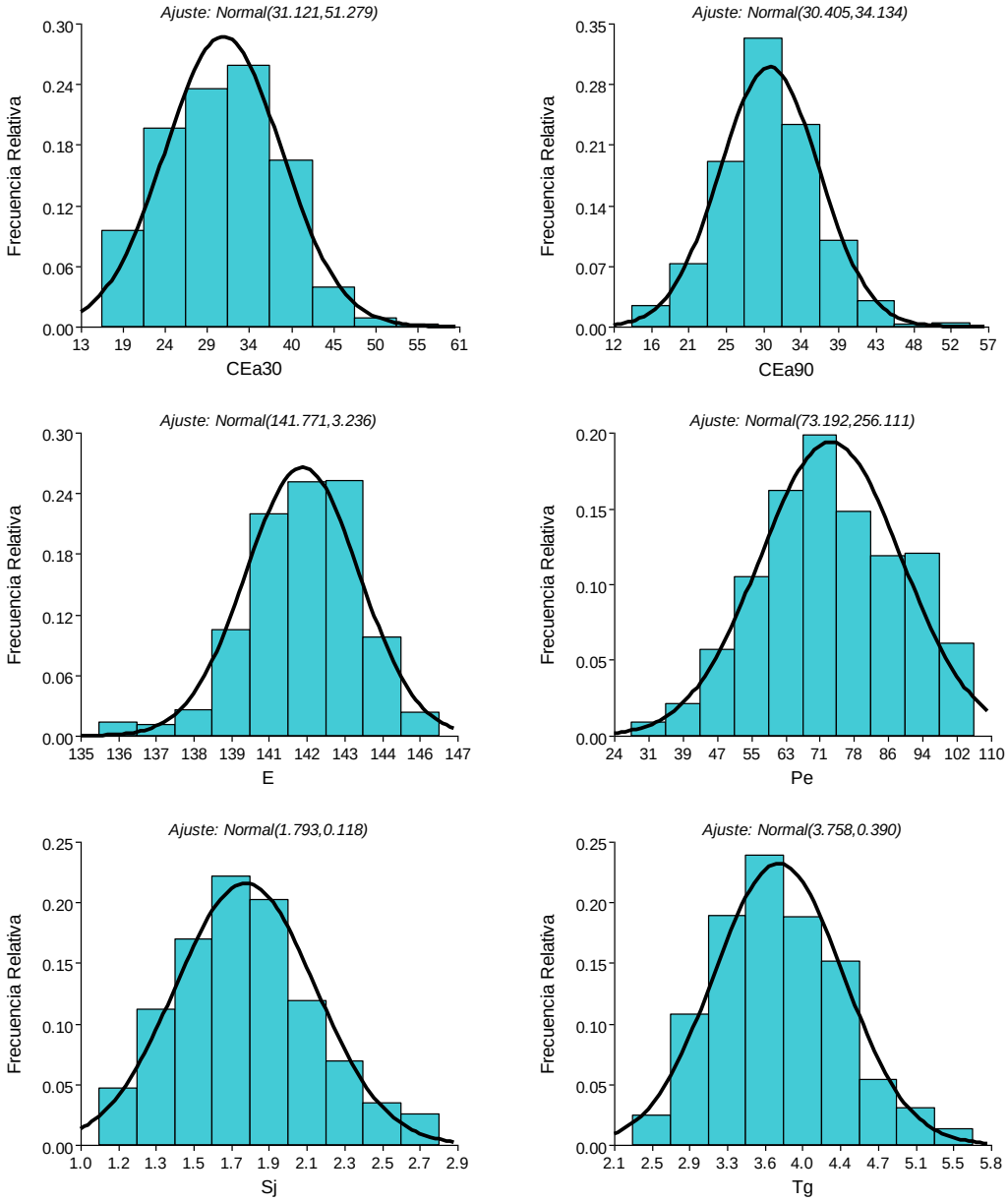


Fig. 1.3. Distribución de frecuencias para variables de suelo y rendimiento de un lote con 664 mediciones georreferenciadas. CE30: conductividad eléctrica aparente a 30 cm de profundidad, CE90: conductividad eléctrica aparente a 90 cm de profundidad, E: elevación, Pe profundidad tosca, Sj: rendimiento de soja; Tg: rendimiento de trigo.

Tabla 1.4. Modelos de regresión lineal múltiple para evaluar tendencia a gran escala.

Variables	Modelo Estimado	p-valor	R^2
Suelo			
CE30, mS m^{-1}	-25862055.6 + 4.6 X + 4.5 Y - 8.0E-07 XY	0.9044	0.02
CE90, mS m^{-1}	696831400.4 – 125.0 X -120.2 Y + 2.2E-05 XY	0.0001	0.03
E, m	116279243.7 – 20.9 X – 20.1 Y + 3.6E-06 XY	0.0290	0.05
Pe, cm	-1104056068.3 + 198.0 X +190.4 Y- 3.4E-05 XY	0.0961	0.01
Rendimiento			
Sj, t ha^{-1}	-61726862.4 + 11.1 X + 10.6 Y - 1.9E-06 XY	0.0001	0.10
Tg, t ha^{-1}	59905799.7 – 10.7 X -10.3 Y + 1.9E-06 XY	0.0014	0.03

CE30: conductividad eléctrica aparente a 30 cm de profundidad, CE90: conductividad eléctrica aparente a 90 cm de profundidad, E: elevación, Pe profundidad tosca, Sj: rendimiento de soja; Tg: rendimiento de trigo.

El resumen de las estimaciones de los parámetros obtenidos a partir de los semivariogramas teóricos ajustados se muestra en la Tabla 1.5. Las variables CE30, E y Tg presentaron una estructura espacial fuerte si se considera el cociente $\frac{C_0}{C_0+C}$. Mientras que en CE90, Pe y Sj la estructura espacial fue intermedia. Los rangos prácticos estimados presentaron valores entre los 55 y 300 m.

Tabla 1.5. Estimaciones (WLS) de los parámetros de un semivariograma exponencial ajustado a variables de suelo y rendimiento en un lote con mediciones georreferenciadas en 664 sitios.

Variable	Modelo	C_0	C	$\frac{C_0}{C_0+C}$	R_p
Suelo					
CE30, mS m^{-1}	Exp.	0	51.46	0	55
CE90, mS m^{-1}	Exp.	24.27	9.12	0.73	175
E, m	Exp.	0	3.25	0	68
Pe, cm	Exp.	0	483.98	0	51
Rendimiento					
Sj, t ha^{-1}	Exp.	0	0.11	0	108
Tg, t ha^{-1}	Exp.	0	0.40	0	191

CE30: conductividad eléctrica aparente a 30 cm de profundidad, CE90: conductividad eléctrica aparente a 90 cm de profundidad, E: elevación, Pe profundidad tosca, Sj: rendimiento de soja; Tg: rendimiento de trigo.
C_0: varianza nugget, C: varianza estructural, R_p: rango práctico.

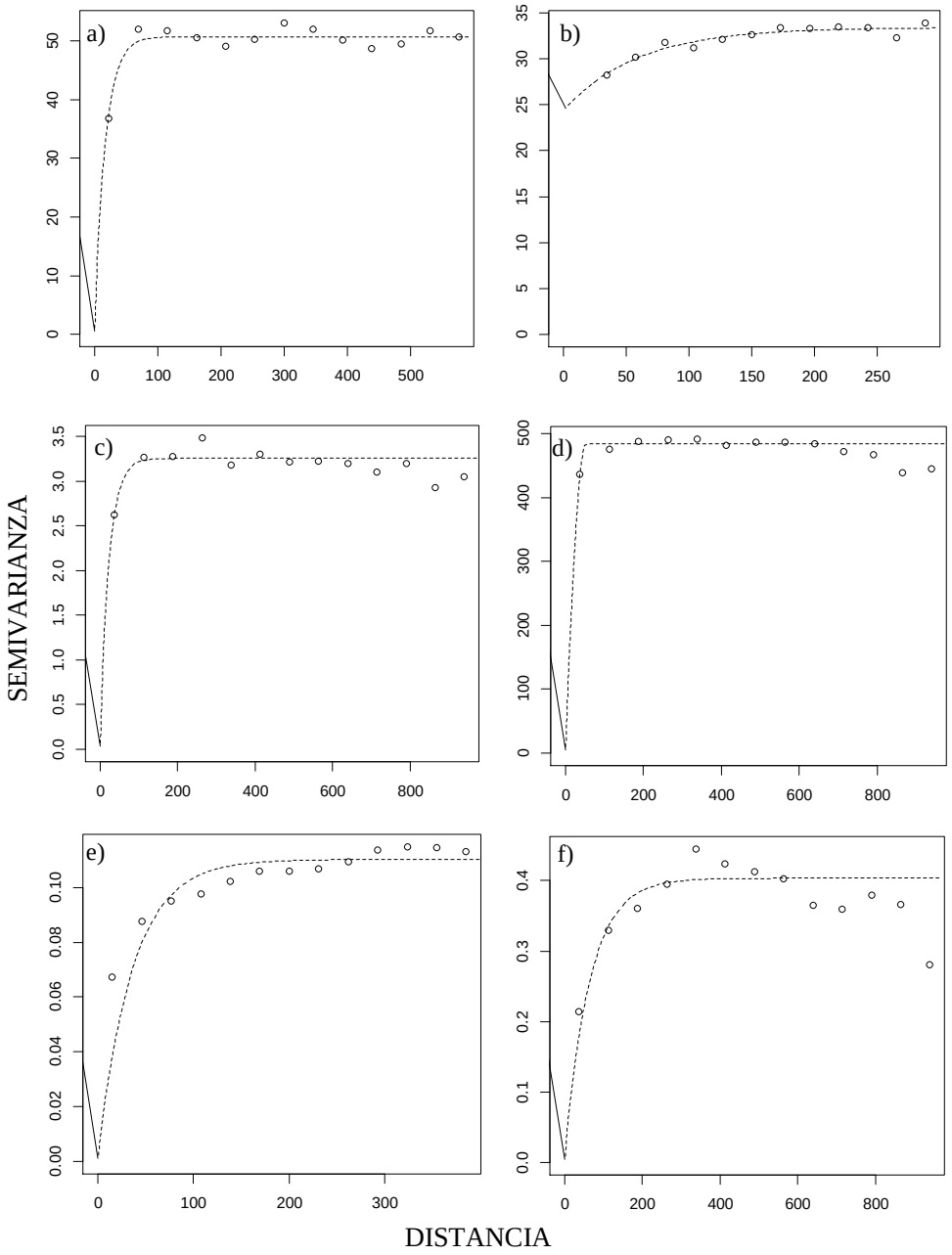


Fig. 1.4. Semivariogramas empíricos (círculos) y teóricos (línea partida) obtenidos a partir de análisis geoestadístico. a) Conductividad eléctrica aparente a 30 cm de profundidad, b) conductividad eléctrica aparente a 90 cm de profundidad, c) Elevación, d) profundidad de tosca, e) rendimiento de soja, f) rendimiento de trigo.

AJUSTE DE UN MLM A DATOS ESPACIALES

Para cada variable se compararon los ajustes obtenidos (vía REML) de los modelos de correlación espacial exponencial y esférico (ambos con y sin efecto nugget), incorporando también el modelo de errores independientes (nugget). Estos ajustes se hicieron para un modelo de estructura de medias del tipo efecto fijo de latitud y longitud y su interacción a los fines de descontar, en caso de que exista, tendencia a gran escala. Para la selección de modelos se usaron los criterios AIC, BIC y la prueba de cociente de verosimilitud (LRT) basada en los estimadores REML de los parámetros de varianza y covarianza. Una vez seleccionado el modelo de correlación espacial, se procedió a comparar los modelos con y sin efecto fijo de las coordenadas espaciales. Para esta comparación se utilizó la prueba LRT, pero basada en estimaciones ML. Todos los análisis fueron realizados con la librería "geoR" (Ribeiro Jr. y Diggle, 2001) del software R

Los criterios de información sobre los ajustes de MLM alternativos para variables de suelo y rendimiento se presentan en la Tabla 1.6 y Tabla 1.7 respectivamente. Para las variables CE30, elevación, Sj y Tg el modelo de correlación espacial exponencial proveyó el mejor ajuste según el AIC y BIC. Estos criterios de información también coincidieron en la selección del modelo esférico para la variable Pe. Sin embargo, en CE90 el AIC favorece la selección del modelo de correlación exponencial con efecto nugget mientras que BIC al modelo exponencial sin efecto nugget. Para esta variable se realizó la prueba LRT que sugirió que el modelo con efecto nugget es el mejor. Así mismo, en todas las variables el ajuste de la tendencia a gran escala no fue estadísticamente significativo (Tabla 1.8).

El resumen de las estimaciones de los parámetros obtenidos a partir de los MLM seleccionados para cada variable se muestra en la Tabla 1.9. Las variables CE30, E, Pe, Sj y Tg presentaron una estructura espacial fuerte. Mientras que en CE90 la estructura espacial fue intermedia. Los rangos fluctuaron entre los 50 y 215 m.

Tabla 1.6. Criterios de información sobre ajustes de MLM de correlación espacial para variables edáficas (N=664 sitios).

Modelo	CE30		CE90		E		Pe	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
1. Exponencial	4363.1	4390.1	4205.7	4228.2	2481.0	2503.5	5950.7	5973.2
2. Esférico	4377.8	4400.3	4215.2	4237.7	2483.0	2510.0	5915.5	5937.9
3. Exponencial + Nugget	4379.0	4406.0	4201.3	4228.3	2485.3	2507.8	5952.7	5979.7
4. Esférico + Nugget	4377.0	4399.0	4215.0	4242.0	2481.7	2508.6	5917.5	5944.4
5. Errores independientes	4517.1	4535.1	4256.7	4274.7	2674.0	2692.0	6008.2	6026.1

CE30: conductividad eléctrica aparente a 30 cm de profundidad, CE90: conductividad eléctrica aparente a 90 cm de profundidad, E: elevación, Pe profundidad tosca.
Se subraya el mejor modelo seleccionado por cada criterio de información.

Tabla 1.7. Criterios de información sobre ajustes de MLM de correlación espacial para variables de rendimiento (N=664 sitios).

Modelo	Sj		Tg	
	AIC	BIC	AIC	BIC
1. Exponencial	322.2	344.7	766.7	789.2
2. Esférico	377.5	399.9	779.5	802.0
3. Exponencial + Nugget	324.2	351.2	768.7	795.7
4. Esférico + Nugget	349.3	376.3	782.5	809.5
5. Errores independientes	505.3	523.3	1296.0	1314.0

Sj: rendimiento de soja; Tg: rendimiento de trigo.
Se subraya el mejor modelo seleccionado por cada criterio de información.

Tabla 1.8. Test del cociente de verosimilitud (LRT) basada en los estimadores ML para evaluar tendencia a gran escala.

Variable	Modelo	Test	p-valor
Suelo			
CE30, mS m^{-1}	1. Exponencial + X + Y		
	2. Exponencial	1 vs. 2	0.4117
CE90, mS m^{-1}	1. Exponencial Nugget + X + Y		
	2. Exponencial Nugget	1 vs. 2	0.5429
E, m	1. Exponencial + X + Y		
	2. Exponencial	1 vs. 2	0.0945
Pe, cm	1. Esférico + X + Y		
	2. Exponencial	1 vs. 2	0.2746
Rendimiento			
Sj, t ha^{-1}	1. Exponencial + X + Y		
	2. Exponencial	1 vs. 2	0.3301
Tg, t ha^{-1}	1. Exponencial + X + Y		
	2. Exponencial	1 vs. 2	0.8799

CE30: conductividad eléctrica aparente a 30 cm de profundidad, CE90: conductividad eléctrica aparente a 90 cm de profundidad, E: elevación, Pe profundidad de tosca, Sj: rendimiento de soja; Tg: rendimiento de trigo.

Tabla 1.9. Estimaciones de los parámetros del MLM ajustado para datos espaciales de suelo y rendimiento.

Variable	Modelo	Estructura de varianzas y covarianzas‡				Estructura de medias
		C_0	C	$\frac{C_0}{C_0 + C}$	R_p	Media General
Suelo						
CE30, mS m^{-1}	Exponencial	0	53.4	0	93	31.1
CE90, mS m^{-1}	Exponencial	22.2	12.2	0.64	157	29.0
E, m	Exponencial	0	3.3	0	107	141.8
Pe, cm	Esférico	0	457.6	0	50	73.4
Rendimiento						
Sj, t ha^{-1}	Exponencial	0	0.1	0	98	1.8
Tg, t ha^{-1}	Exponencial	0	0.4	0	215	3.7

CE30: conductividad eléctrica aparente a 30 cm de profundidad, CE90: conductividad eléctrica aparente a 90 cm de profundidad, E: elevación, Pe profundidad de tosca, Sj: rendimiento de soja; Tg: rendimiento de trigo.
‡C_0: varianza nugget, C: varianza estructural, R_p: rango práctico.

MAPEO DE VARIABILIDAD ESPACIAL DE VARIABLES DE SUELO Y RENDIMIENTO

A partir de los parámetros del semivariograma estimados mediante geoestadística clásica y mediante los parámetros de los modelos lineales mixtos, se realizó la interpolación mediante Kriging puntual y se obtuvieron mapas de contorno para cada variable. Los mapas generados presentan una variabilidad espacial similar en ambos métodos de estimación (Fig. 1.5 y 1.6).

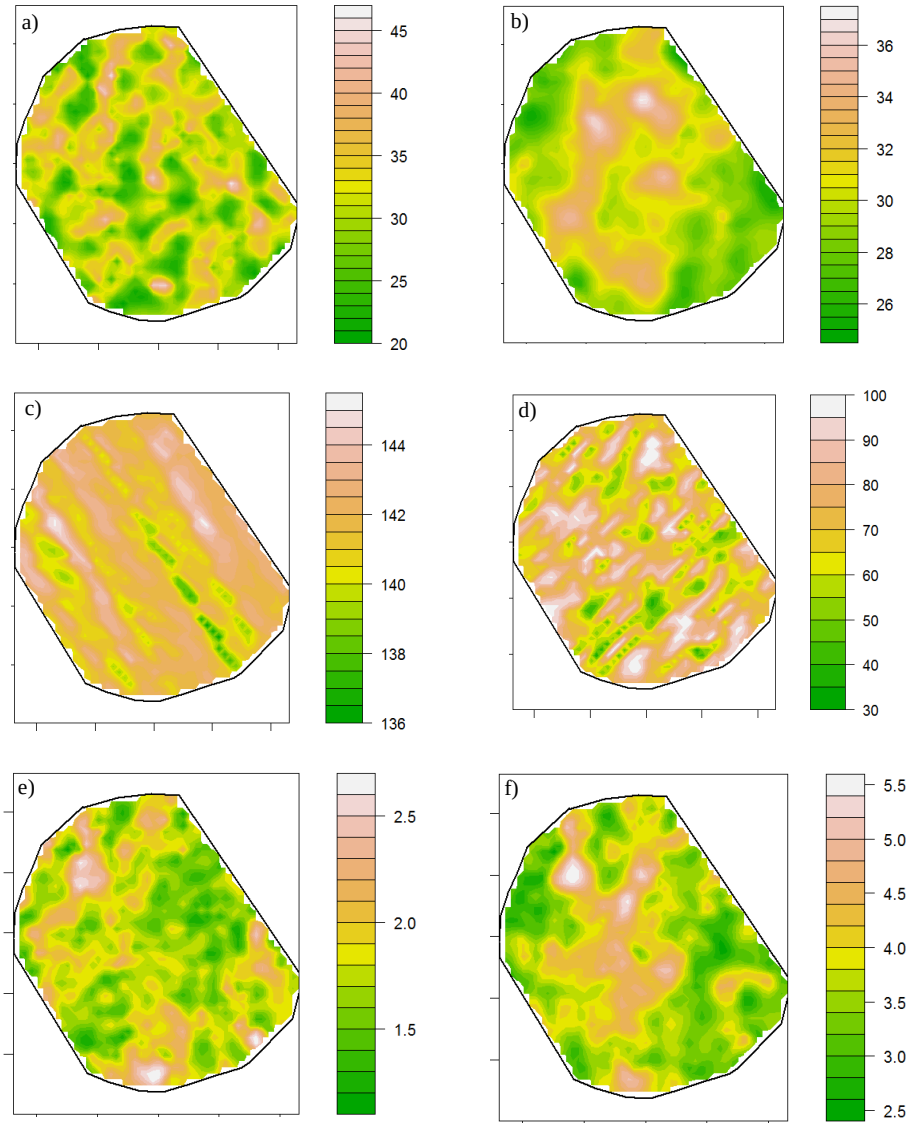


Fig. 1.5. Mapas de variabilidad espacial de variables de suelo y rendimiento obtenidos mediante la interpolación por kriging ordinario utilizando parámetros del semivariograma estimados con goestadística clásica. a) Conductividad eléctrica aparente a 30 cm de profundidad, b) Conductividad eléctrica aparente a 90 cm de profundidad, c) Elevación, d) profundidad de tosca, e) rendimiento de soja, f) rendimiento de trigo.

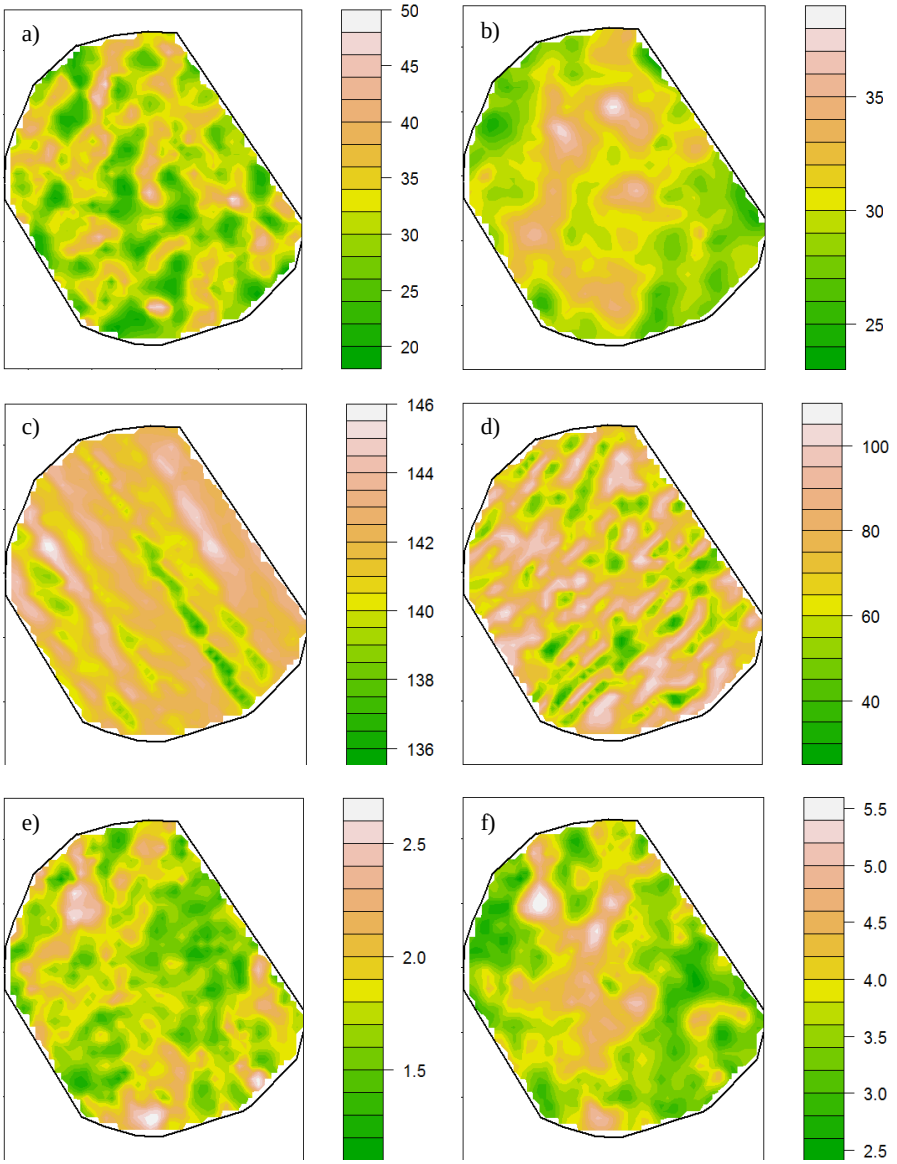


Fig. 1.6. Mapas de variabilidad espacial de variables de suelo y rendimiento obtenidos mediante la interpolación por kriging ordinario utilizando parámetros del semivariograma estimados con modelos lineales mixtos. a) Conductividad eléctrica aparente a 30 cm de profundidad, b) Conductividad eléctrica aparente a 90 cm de profundidad, c) Elevación, d) profundidad de tosca, e) rendimiento de soja, f) rendimiento de trigo.

DISCUSIÓN

Todas las variables analizadas presentaron autocorrelación positiva y significativa. La tendencia de los datos a ser correlacionados positivamente, es una importante característica de los datos espaciales y es omnipresente entre las observaciones

provenientes de mediciones realizadas con sensores de variables de suelos y de rendimiento. En el análisis de variabilidad espacial realizado para cada variable por separado mediante técnicas geoestadísticas y mediante MLM, mostró que para todas las variables, excepto CE90 la estructura espacial es de magnitud fuerte. Esto es coincidente con lo detectado con los índices de autocorrelación espacial, de Moran y Geary, donde la CE90 tuvo la menor autocorrelación espacial.

Las estimaciones de los parámetros del semivariograma mediante WLS y MLM fueron similares. Los rangos variaron entre los 50 m en Pe y 215 m para la variable Tg. El rango del semivariograma influye en la interpolación de los datos, los sitios ubicados a distancias mayores al rango tienen la mínima capacidad predictora. Si el rango se amplía, estos puntos tendrían mayor peso en la interpolación. Una estructura espacial fuerte y con un rango amplio genera mapas de variabilidad con zonas grandes y contiguas como sucede en la variable CE90 y Tg. Cuando el rango fue más bajo los mapas de variabilidad presentaron una mayor fragmentación o estructura de "parches", como sucedió con las variables Pe y CE30.

Aún cuando la diferencia en las estimaciones logradas con la aproximación basada técnicas geoestadísticas o con MLM no fue grande, la utilización de MLM presenta claras ventajas. Cuando se trabaja con técnicas geoestadísticas es necesario realizar en la etapa exploratoria de los datos el ajuste de regresiones para evaluar las tendencias a gran escala. En caso de que la tendencia fuese significativa, será necesario descontar la tendencia y trabajar con los residuos del modelo de regresión. Mientras que usando MLM, se puede modelar la correlación espacial y la tendencia a gran escala en un solo paso. En esta estrategia las coordenadas espaciales se incorporan en la estructura de medias del modelo, permitiendo que en el término de error aleatorio se elimine el sesgo producido por la tendencia a gran escala. De esta forma puede ser modelada la correlación espacial sin el "ruido" producido por esa tendencia.

Para la selección de los modelos, al trabajar en el contexto de los MLM, existen diferentes herramientas como los criterios de información y pruebas estadísticas formales como el test del cociente de verosimilitud. Cuando se utilizan técnicas geoestadísticas, la selección de modelos se basa solo en el cálculo del cuadrado medio del error., no existiendo el mismo nivel de desarrollo teórico como para utilizar pruebas alternativas.

Adicionalmente, en el contexto de los MLM, es posible obtener las medias ajustadas por modelo para hacer una interpretación agronómica más sencillas de los parámetros obtenidos. Las medias ajustadas podrían diferir respecto a las media de la variable sin el ajuste por modelo. Por ejemplo, utilizando la varianza umbral ($C+C_o$) puede obtenerse la desviasión estándar de los datos y junto a la media ajustada calcular el CV como $\sqrt{C+Co}/\text{Media}$. La variable Pe tuvo el mayor CV (30%) mientras que la E presentó la menor variabilidad relativa (1%). Si bien E mostró poca variabilidad en los lotes pampeanos dedicados al cultivo extensivo de grano, los CV de CE y Pe fueron altos.

CONCLUSIÓN

Existe una amplia variedad de métodos estadísticos para el estudio de variabilidad espacial y la aplicación de más de uno de éstos, sobre un mismo caso, puede ser necesaria ya que no existe un único método que pueda determinar todas las características de importancia de la estructura espacial. Los resultados de cada método no son independientes entre sí, sino que aportan información complementaria. Los índices de autocorrelación de Moran y Geary, son una estadística global que considera los valores de todas las observaciones. Detectada la existencia de aglomeraciones espaciales en torno a valores superiores o inferiores a la media general, se puede utilizar técnicas que permitan modelar las dependencias espaciales. En este Capítulo las dependencias espaciales de las propiedades del sitio y de los rendimientos de los cultivos fueron abordadas a través de modelos geoestadísticos clásicos y MLM de covarianza espacial. Aún cuando las diferencias en las estimaciones logradas con ambas aproximaciones no fueran grandes, la utilización del MLM presentó claras ventajas. Los métodos descriptos permitieron detectar la existencia de estructura espacial en el rendimiento y propiedades de suelo, pero, solo con la construcción de mapas de variabilidad espacial se pudo visualizar el patrón espacial subyacente.

CAPÍTULO II

APROXIMACIÓN MULTIVARIADA EN EL ANÁLISIS DE DATOS ESPACIALES

INTRODUCCIÓN

En las últimas décadas se ha impulsado el desarrollo y la utilización de diversas nuevas tecnologías para agricultura de precisión que permiten capturar diferentes tipos de datos espaciales, *i.e.* datos de diferentes variables asociados a una localización en el espacio para diferentes sitios dentro del lote. Los monitores de rendimiento acoplados a los equipos de cosecha y sensores de diferentes tipos, permiten recabar datos espaciales georreferenciados produciendo gran cantidad de datos para diferentes variables en áreas relativamente pequeñas y en cortos períodos de tiempo (Roel y Terra, 2007). La disponibilidad de datos para una gran cantidad de sitios dentro de un mismo lote, no sólo para variables de cultivo sino también para otras variables como son las topográficas (elevación, profundidad de tosca) y las de suelo (conductividad eléctrica aparente), genera la necesidad de contar con técnicas de análisis de naturaleza multivariada. Es más, si un objetivo del análisis es delimitar zonas de manejo (ZM) para el manejo sitio específico del cultivo, éstas debieran ser uniformes en sentido multivariado, es decir, considerando varias variables simultáneamente.

La mayoría de los métodos multivariados, utilizados en trabajos sobre AP para la delimitación de ZM, se basan en algoritmos de agrupamiento no supervisados (Vrindts *et al*., 2005). Fraisse *et al*. (2001) utilizaron para la delimitación de ZM algoritmos de *clustering* jerárquicos, mientras que otros autores propusieron utilizar el algoritmo de *clustering* no jerárquico *k-means* (Taylor *et al* 2003; Whelan y McBratney 2003; Hornung *et al*., 2006; Ortega y Santibañez, 2007). El algoritmo no jerárquico difuso *fuzzy k-means*, es, actualmente, uno de los más usados en trabajos de AP (Li *et al*., 2007; Arno *et al*., 2011; Davatgar *et al*., 2012). Contrariamente al algoritmo *k-means* u otros métodos

determinísticos de agrupamiento, como ISODATA (Tou y Gonzalez, 1974), en los que cada observación sólo puede pertenecer a un único cluster, los métodos de clasificación basados en la teoría difusa (por ejemplo *fuzzy k-means*), permiten que cada observación pueda asignarse a más de un *cluster*, con diferentes grados de pertenencia para cada *cluster*.

La expansión del uso del enfoque de conglomerados difusos se ha acelerado con el lanzamiento de software, de fácil acceso como FuzME (Minasny y McBratney, 2002) y Management Zone Analyst (MZA) (Fridgen *et al*., 2004), donde el método se puede aplicar sin dificultades. A pesar de la teoría subyacente, en varias aplicaciones de *fuzzy k-means* en AP, se observó que el algoritmo puede presentar el inconveniente de una alta fragmentación de las zonas. Esta fragmentación podría deberse a que el algoritmo de agrupación no tiene en cuenta la información espacial asociada a cada observación (Ping y Dobermann 2003; Frogbrook y Oliver 2007). Frogbrook y Oliver (2007) y Milne *et al*. (2012) propusieron introducir la restrincción espacial a través de parámetros del variograma co-regionalizado o del variograma de la componente principal, variable sintética que resume la información en las variables originales. Para formar clases más contiguas y reducir la fragmentación de las ZM delimitadas, también se pueden aplicar filtros espaciales a la clasificación resultante de un método de clasificación (Lark, 1998; Ping y Dobermann 2003; Galarza *et al*., 2013).

El análisis de las covariaciones o correlaciones entre variables es otro aspecto clave a tener en cuenta en estudios de AP de naturaleza multivariada, ya que estas permiten comprender las interacciones que subyacen a los rendimientos. No obstante, es importante remarcar que la estructura de covariación reflejada por un análisis multivariado clásico puede verse afectada por los patrones espaciales subyacentes en los datos. La variación espacial de propiedades del sitio y el rendimiento de los cultivos ha sido estudiada a través de geoestadística (Oliver, 2010) y mediante la estimación de modelos lineales mixtos (Gbur *et al*., 2012), que se basan en el concepto de autocorrelación, pero típicamente estos análisis se realizan variable por variable. Los índices de autocorrelación espacial, como el de Moran (Moran, 1948) y el de Geary (Geary, 1954), fueron los primeros usados para medir y analizar el grado de dependencia entre observaciones de una misma variable ubicadas en diferentes posiciones de un contexto geográfico. Estos análisis que se abordan variable a variable (univariados) dificultan la interpretación de la variabilidad conjunta, es

decir, aquella causada por las relaciones entre variabilidad del rendimiento y de otras variables (Córdoba *et al*., 2011). La teoría de las variables co-regionalizadas (Schabenberger y Pierce, 2002) ofrece una alternativa para analizar covariación espacial de dos variables georreferenciadas. El uso de estas aproximaciones en aplicaciones de AP es limitado. Desarrollos futuros de aplicaciones en este contexto teórico así como en el de variabilidad multivariada (Pebesma, 2004) seguramente, encontrarán en la AP numerosas implementaciones.

Dos diferentes objetivos pueden surgir cuando analizamos un conjunto de datos georeferenciados que además es multivariado. Por un lado, se puede querer resumir la variabilidad de los atributos entre sitios en unos pocos componentes. Por otro lado, se puede querer revelar los patrones espaciales existentes o índices que combinan las múltiples características de sitio. Una solución al primer problema es usar el Análisis de Componentes Principales (PCA, Pearson, 1901) ya ampliamente difundido en AP (Schepers *et al*., 2004). Mientras que el segundo objetivo puede ser abordado mediante la evaluación de la autocorrelación espacial de las nuevas variables originadas (componentes principales) (Wartenberg, 1985).

Las componentes principales (PC) son apropiadas sólo para resumir variabilidad y no está diseñado para revelar patrones espaciales. Entonces, es necesario usar una metodología que resuma la variabilidad y revele estructuras espaciales al mismo tiempo; existen hoy métodos que abarcan estos dos objetivos. Dray *et al*. (2008), proponen un método de análisis multivariado que incorpora la información espacial previo al análisis multivariado, el método es conocido como MULTISPATI-PCA. Éste se basa en el PCA pero incorpora la restricción dada por los datos espaciales mediante el cálculo del índice de Moran antes de obtener las PC. El objetivo es encontrar variables sintéticas independientes que optimicen el producto de la varianza total y el coeficiente de Moran. Para delimitar los vecindarios, MULTISPATI-PCA utiliza una matriz de pesos espaciales determinando cuáles y cuántas observaciones cercanas a cada sitio deben ser consideradas para el cálculo del índice de Moran. Este análisis permite estudiar las relaciones entre las variables considerando su estructura espacial. La técnica ha mostrado ser eficiente en estudios de ecología y suelos realizados a escala macrogeográfica (Dray *et al*., 2008; Arrouays *et al*., 2011) y a escala de lote (Gili, 2013), pero no ha sido evaluada en el contexto de AP:

El objetivo del presente Capítulo es ilustrar e interpretar agronómicamente, la aplicación de tres técnicas multivariadas en datos de AP: análisis de *cluster fuzzy k-means*, PCA y MULTISPATI-PCA. Los resultados obtenidos con la implementación de PCA y con MULTISPATI-PCA, son comparados detalladamente ya que ambos comparten la misma naturaleza del algoritmo base, *i.e.* PCA no restringido y restringido espacialmente.

MATERIALES Y MÉTODOS

PROCEDIMIENTOS DE ANÁLISIS MULTIVARIADO

ANÁLISIS DE CLUSTER FUZZY K-MEANS

Existen tres matrices primarias que participan en el análisis *fuzzy k-means*. La primera de ellas es la matriz de datos a clasificar ($\mathbf{X}$). Usualmente en AP, los datos de la matriz $\mathbf{X}$ incluye n observaciones cada una con p variables de suelo y rendimiento. La segunda matriz ($\mathbf{V}$), consta de los k centroides correspondientes a cada *cluster* localizado en el espacio de los atributos definido por las p variables. La tercera, es la matriz de pertenencia difusa ($\mathbf{U}$), que contiene los valores o asignaciones parciales de cada una de las n observaciones en cada uno de los k *clusters* o conglomerados, limitada por la restricción que se muestra en (2.1), debiéndose cumplir ésta para cualquier $i = 1,\ldots,n$ y para cualquier $j = 1,\ldots,k$:

$$u_{ij} \in [0,1] \forall_{i,j} \text{ y } \quad \sum_{j=1}^{k} u_{ij} = 1, \forall_j \tag{2.1}$$

La partición difusa óptima de los datos es la que minimiza la función objetivo j_m igual a la suma ponderada de las distancias cuadráticas entre las observaciones y los centroides de cada *cluster*:

$$j_m(U,v)=\sum_{i=1}^{n}\sum_{j=1}^{k}\left(u_{ij}\right)^m\left(d_{ij}\right)^2 \tag{2.2}$$

donde m es el coeficiente de ponderación difuso ($1 \leq m < \infty$) cuya función es controlar el grado de solapamiento que se establece entre los *clusters* y $(d_{ij})^2$ es el cuadrado de la distancia en el espacio de los atributos entre el punto *i* y la clase centroide *j*, que se puede calcular de la siguiente manera:

$$(d_{ij})^2 = ||x_i - v_j||^2 = (x_i - v_j)^T \mathbf{A}(x_i - v_j) \tag{2.3}$$

donde x_i es la observación *i-ésima* de la matriz de datos $\mathbf{X}$, v_j el centroide del *cluster j*, y $\mathbf{A}$ es la matriz de pesos definida positiva ($p \times p$) que induce norma por el producto interno. La matriz de ponderación $\mathbf{A}$ define un procedimiento de normalización de la distancia. El resultado representa la distancia entre dos puntos o vectores en un espacio vectorial lineal.

Fridgen *et al*. (2004) aconsejan tomar $\mathbf{A} = \mathbf{I}$ (matriz identidad), únicamente cuando las variables sean estadísticamente independientes y presenten la misma varianza. La métrica obtenida es, por tanto, la distancia Euclídea entre la observación *i-ésima* y el centroide. En el caso de que las varianzas de las variables sean distintas, es recomendable estandarizar las variables mediante la utilización de una matriz diagonal ($\mathbf{A} = \mathbf{D}$) cuyos términos sean precisamente las varianzas de las variables en estudio o bien trabajar con las variables previamente estandarizadas. Finalmente, la tercera posibilidad es tomar $\mathbf{A} = \mathbf{S}$ (matriz de varianzas y covarianzas de $\mathbf{X}$), con lo que la métrica resultante es la distancia de Mahalanobis. Se utiliza esta distancia cuando las variables de clasificación no solo muestran varianzas distintas sino que están correlacionadas entre sí. Mientras que el algoritmo iterativo *fuzzy k-means* siempre converge a un mínimo local de J_m a partir de un determinado $\mathbf{U}$ inicial, una aleatorización diferente de $\mathbf{U}$ podría dar lugar a un mínimo local diferente (Xie y Beni; 1991; Bezdek, 1981).

El algoritmo difuso *fuzzy k-means* utiliza un proceso iterativo para la obtención del par ($\mathbf{U}$, $\mathbf{V}$) que hace óptima la partición difusa de los datos $\mathbf{X}$. La estructura del algoritmo (Bezdek, 1981) se muestra a continuación.

1. Se elige el número de grupos o *clusters* k, con $2 \leq k \leq n$.

2. Se fija el valor del exponente difuso m, con $1 < m < \infty$.

3. Se selecciona una medida apropiada de similaridad o distancia d_{ij}.

4. Se selecciona el valor del criterio de convergencia (finalización) del algoritmo.

5. Se selecciona el número máximo de iteraciones, $l_{\max}$.

6. Se inicializa la matriz $\mathbf{U}^0$ con valores aleatorios y según la restricción especificada en (2.1).

7. En las sucesivas iteraciones l= 1,2,3…, se recalculaba $\mathbf{V}^l$ (matriz de centroides) a partir de $\mathbf{U}^{(l-1)}$, utilizando la siguiente expresión:

$$v_j = \frac{\sum_{i=1}^{n} (u_{ij})^m x_j}{\sum_{i=1}^{n} (u_{ij})^m}, \quad 1 \leq j \leq k \tag{2.4}$$

8. La minimización de (2.2) mediante el método iterativo de Picard hace posible el cálculo (actualización) de $\mathbf{U}^l$ a partir de la matriz actualizada $\mathbf{V}^l$, según:

$$u_{ij} = \left[\sum_{j=1}^{k} \left(\frac{d_{ij}}{d_{lj}} \right)^{2/(m-1)} \right]^{-1}, \quad i = 1,...,n \;\; j = 1,...,k \tag{2.5}$$

9. Se interrumpe el algoritmo cuando se alcanzaba el número máximo de iteraciones (l_{max}), o cuando $||\mathbf{U}^l - \mathbf{U}^{(l-1)}|| \leq \varepsilon$; en otro caso, se volvía al paso 7.

10. Se computaban finalmente los índices para validar los *cluster*.

Para evaluar la clasificación conseguida con un determinado número de grupos, existen diferentes índices como el coeficiente de partición (o fuzziness performance index-FPI, Bezdek, 1981), el índice de entropía de la clasificación (o normalized classification

entropy-NCE, Bezdek, 1981), el índice de Xie-Beni (Xie y Beni, 1991) y el de Fukuyama-Sugeno (Fukuyama y Sugeno, 1989), entre otros.

El coeficiente de partición (CP) mide el grado de separación o solapamiento (grado de *fuzziness*) entre los grupos formados. Se considera que mientras menos difusa es la partición, mejor es la clasificación. Su cálculo resulta de la siguiente expresión:

$$CP(U) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^2}{n} \tag{2.6}$$

En este caso el óptimo se da al maximizar CP y equivale a una clasificación en la que cada observación pertenece a un único *cluster*. El mínimo se da cuando cada observación pertenece, con el mismo grado, a cada *cluster* (mayor incertidumbre).

La entropía de la partición (EP) estima la cantidad de desorganización creada por la partición difusa de la matriz de datos $\mathbf{X}$ con un número específico de *clusters*. Para este índice los valores de EP próximos a 0 son indicativos de una mejor clasificación, es decir, con mayor grado de organización.

$$EP(U) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij} \log(u_{ij})}{n} \tag{2.7}$$

En el índice de Xie-Beni (XB) se incorpora a v y X. Este índice prefiere particiones cuya distancia intra-*cluster* sea mínima y la distancia inter-*cluster* máxima:

$$XB(U, v, X) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^2 \, || x_i - v_j ||^2}{n(\min_{j \neq j'} || v_j - v_{j'} ||^2)} \tag{2.8}$$

El índice XB es considerado como una medida de compacidad. Un valor pequeño de XB, representa una clasificación donde los grupos son compactos y bien separados. Por consiguiente, la mejor partición se obtiene mediante la minimización de XB.

El índice Fukuyama-Sugeno (FS) está formado por la diferencia entre la medida de compacidad y la medida de separación entre los centroides de los grupos y la media de todos los centroides. El mínimo de FS corresponde a una partición difusa con clases compactas y bien separadas.

$$FS(U,v,X)=\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}^{m}(||x_i-v_j||^2-||v_j-\bar{v}||^2) \tag{2.9}$$

En AP los indicadores más utilizados son CP y EP. Con el tiempo se ha ido generando un gran número de funciones de validez para la clasificación difusa que podrían ser mejores en determinados contextos. Es importante considerar que para un conjunto de datos, los índices no son necesariamente consistentes entre sí e incluso pueden contradecirse, es decir, sugerir diferentes números de *cluster* como partición óptima. Una solución es obtener un único índice que resuma los anteriores (Galarza *et al*., 2013). En los índices mencionados excepto para el coeficiente de partición (CP), un valor menor del índice implica una mejor clasificación. Por ello, se recalcula CP como CP*= 1/CP para que el valor mínimo en todos los índices represente la mejor elección. Adicionalmente, se normalizan los valores de los índices entre 0 y 1 dividiendo cada valor por el máximo alcanzado por el índice en las diferentes clasificaciones. Luego, se calcula la distancia Euclídea para cada clasificación utilizando los valores de los índices normalizados y se selecciona la clasificación con menor valor de este nuevo índice. Los índices aportan información sobre cúal podrá ser la clasificación óptima. En AP la selección final de la cantidad de *cluster* debe seguir una relación de compromiso entre lo sugerido por los índices y el criterio agronómico.

ANÁLISIS DE COMPONENTES PRINCIPALES SIN Y CON RESTRICCIÓN ESPACIAL

El análisis de componentes principales (PCA) permite identificar las variables que explican la mayor parte de la variabilidad total contenida en los datos, explorar las correlaciones entre variables y reducir la dimensión del análisis al combinar las variables originales en nuevas variables sintéticas. Este análisis, encuentra para los datos una base ortogonal de tal manera que el primer eje del nuevo espacio considera la dirección de mayor variación de los datos originales, proporcionando un conjunto de autovectores y sus correspondientes autovalores. Los autovectores contienen los coeficientes de ponderación para construir las combinaciones lineales, que indican la importancia relativa de las variables para explicar la variabilidad entre las observaciones en cada eje. Las combinaciones lineales obtenidas con PCA se llaman componentes principales (PC), son ortogonales y en conjunto explican toda la variabilidad de los datos originales. Existen tantas PC posibles de formar como variables originales existan. La primera componente (PC1) explica la mayor parte de la variación total en el conjunto de datos y la segunda (PC2), la mayor parte de la variabilidad remanente o no explicada por la PC1, y así sucesivamente.

Los resultados del PCA se pueden visualizar en un gráfico denominado Biplot (Gabriel, 1971) el cual permite representar en un plano óptimo para el estudio de variabilidad, las diferencias entre sitios, la correlación entre variables y las variables que mejor explican las principales variaciones. La incorporación de la información geográfica o la característica espacial de los datos puede realizarse a posteriori del PCA mediante la asignación de los valores de las componentes a cada uno de los sitios georreferenciados o bien ajustando semivariogramas (Schabenberger y Pierce, 2002) a las PC. Una ventaja de la utilización de variables sintéticas para mapeo es que se colapsa la caracterización multidimensional de las observaciones, permitiendo la construcción de mapas sintéticos de variabilidad espacial. Esta técnica permite visualizar el patrón de la variabilidad espacial y explorar gráficamente la estructura espacial de las variables analizadas. También se puede estudiar la presencia de autocorrelación espacial en las PC utilizando estadísticos de autocorrelación univariados como el índice de Moran (Moran, 1948) o el de Geary (Geary, 1954). Estos índices son utilizados para medir y analizar el grado de dependencia entre observaciones en un contexto geográfico (Cliff y Ord, 1973).

Los datos multivariados son generalmente registrados en una matriz $\mathbf{X}$ con n filas (observaciones) y p columnas (variables). El diagrama de dualidad provee un marco teórico que define la estructura de numerosos métodos de análisis multivariado usando tres matrices $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. La teoría del diagrama de dualidad incluye métodos estándar como el análisis de componentes principales (PCA) y la extensión de éste a datos espaciales. Se considera la matriz de datos (originales o transformados) $\mathbf{X}$ (nxp) como parte del triplete $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$, donde $\mathbf{Q}$ (pxp) y $\mathbf{D}$ (nxn) son usualmente matrices simétricas definidas positivas usadas como métricas de relaciones o distancias.

Para la realización del PCA restringido espacialmente, denominado MULTISAPTI-PCA, es necesario primero definir cómo la información espacial es introducida en el análisis. En MULTISAPTI - PCA, la detección de la estructura espacial se realiza a través del índice de Moran (MI). Esta aproximación entonces requiere que los sitios vecinos sean definidos. Esto en general se consigue por la construcción de una red de conexión (también llamada gráficos de vecinos) la cual usa un criterio objetivo para definir que entidades son vecinas y cuáles no. Existen diferentes opciones o alternativas metodológicas para definir los vecindarios que dependen de los diferentes tipos de muestreo presente en los datos (grilla regular, irregular o transectas) (Bivand, 2008). Para muestreos irregulares los métodos se basan en el gráfico de Gabriel (Gabriel y Sokal, 1969), la triangulación de Delaunay (Lee y Schachter, 1980), los vecinos más cercanos (Cover y Hart, 1967) y la distancia Euclidea entre otros.

Una vez que la red de conexión es definida, la información espacial es almacenada en una matriz de conexión binaria $\mathbf{C}$ (en la cual $c_{ij} = 1$ si las unidades espaciales $i\ y\ j$ son vecinas o $c_{ij} = 0$ en caso contrario), la cual es simétrica y sus filas y columnas corresponden a la misma entidad biológica (como una matriz de distancias). Esta matriz de conectividad $\mathbf{C}$ en general es escalada para obtener la matriz de pesos espaciales $\mathbf{W}$. La matriz $\mathbf{W}$ es una representación matemática de la disposición geográfica de los puntos en la región (Bivand, 2008). Los pesos espaciales reflejan a priori la ausencia $(w_{ij} = 0)$, presencia $(w_{ij} = 1)$ o intensidad $(w_{ij} > 0)$ de la relación espacial entre las ubicaciones de interés. Una vez que los pesos espaciales han sido definidos, el índice de autocorrelación MI es computado. El método MULTISPATI-PCA introduce una matriz de pesos espaciales estandarizada por fila $(\mathbf{W})$ mediante un el análisis del triplete estadístico $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. La

matriz $\widetilde{\mathbf{X}} = \mathbf{WX}$ está compuesta por los promedios ponderados de los valores de los vecinos de acuerdo a la matriz de conexión espacial, es llamada matriz lagged. Las dos tablas $\mathbf{X}$ y $\widetilde{\mathbf{X}}$ tienen las mismas columnas (variables) y las mismas filas (observaciones). El análisis MULTISPATI-PCA consiste en el análisis de este par de tablas ($\widetilde{\mathbf{X}}$ y $\mathbf{X}$) mediante un análisis de coinercia (Dray *et al.*, 2003). Para establecer la significación estadística de la estructura espacial de la tabla **X**, puede usarse un procedimiento basado en permutación. El estadístico usado es igual a $traza(\mathbf{X}^T\mathbf{DWXQ})$. El p-valor es computado por comparación del valor observado a aquellos obtenidos por permutación de las filas de **X**.

MULTISPATI-PCA maximiza el producto escalar entre una combinación lineal de las variables originales y una combinación lineal de variables lagged (Saby *et al*., 2009). La ventaja de MULTISPATI-PCA respecto al PCA es que las componentes principales espaciales del MULTISPATI-PCA (sPC) maximizan la autocorrelación espacial entre los sitios. Por lo tanto, las sPC del MULTISPATI-PCA muestran fuertes estructuras espaciales sobre los primero pocos ejes (Arrouays *et al*., 2011).

IMPLEMENTACIÓN DE ALGORITMOS MULTIVARIADOS

Se trabajó con datos provenientes de un lote en producción (65,4 ha) ubicado al sudeste bonaerense de la República Argentina, con información de 672 sitios dentro del lote. Se compilaron valores georreferenciados de conductividad eléctrica aparente (CE) en dos profundidades 0-30 cm (CE30) y 0-90 cm (CE90), Elevación (E), profundidad de tosca (Pe) y rendimiento de soja (Sj) y trigo (Tg). Para mayor nivel de detalle sobre la obtención de estas mediciones referirse a la descripción de los datos utilizados en el Capítulo I.

Para realizar los análisis estadísticos se utilizó el software R (R Core Team, 2013). La librería "e1071" (Meyer *et al*., 2013) se utilizó para realizar el análisis de cluster *fuzzy k-means* y para el cálculo de los índices que permiten seleccionar el número de cluster óptimo. Como medida de similitud se utilizó la distancia de Mahalanobis. Otras opciones de configuración utilizadas fueron: número máximo de iteraciones=300 y criterio de convergencia=0.0001. El exponente difuso se fijó en el valor convencional de 1,30 (Odeh et al., 1992), mientras que el número mínimo y máximo de clases fue de tres y cuatro, respectivamente.

La librería “ade4” (Chessel *et al*., 2004) fue empleada para el PCA, y MULTISPATI-PCA, para este último también se usó la librería “spdep” (Bivand *et al*., 2013a). La red de vecindarios fue definida en función de la distancia Euclídea considerando puntos vecinos a aquellos contiguos ubicados entre los 0 a 35 m de distancia. Para ello se utilizó la función de *dnearneigh* del paquete "spdep". Finalmente para la obtención de mapas por interpolación se utilizó la librería “geoR” (Ribeiro Jr. y Diggle, 2001). Las sentencias de los análisis se encuentran en el Anexo 3.

CRITERIOS DE COMPARACIÓN

Los resultados obtenidos por PCA y MULTISPATI-PCA sobre los datos ilustración, se compararon en términos de la varianza explicada por las PC. Para el PCA las varianzas son iguales a los autovalores asociados a cada PC, mientras que para MULTISPATI-PCA las varianzas, reflejadas en los autovalores de las sPC, son equivalentes a la varianza espacial (no varianza total) o varianza corregida por la presencia de autocorrelación. También se analizó la presencia de autocorrelación en las PC de ambos métodos con el índice de Moran. Se comparó la pérdida de la inercia (varianza espacial *vs.* varianza total) y el aumento de la información espacial (índice de Moran del MULTISPATI-PCA *vs.* índice de Moran del PCA). La representación gráfica de las dos primeras componentes principales del MULTISPATI-PCA y el PCA se utilizó para evaluar la covariación de las variables CE30, CE90, E, Pe, Sj y Tg. La predicción Kriging (Schabenberger y Pierce, 2002) aplicado sobre semivariogramas de la PC1 del PCA y la sPC1 del MULTISPATI-PCA, fue utilizada para obtener mapas de variabilidad espacial multivariada.

RESULTADOS

CLASIFICACIÓN DE SITIOS INTRALOTE VÍA CLUSTER FUZZY K-MEANS

En el mapa de la Fig. 2.1 se presentan los mapas resultantes de la clasificación de los sitios realizada a partir de las variables de suelo y rendimiento mediante el análisis de *cluster fuzzy k-means*. Se observa que la clasificación con tres y cuatro *clusters* producen alta fragmentación de las clases delimitadas.

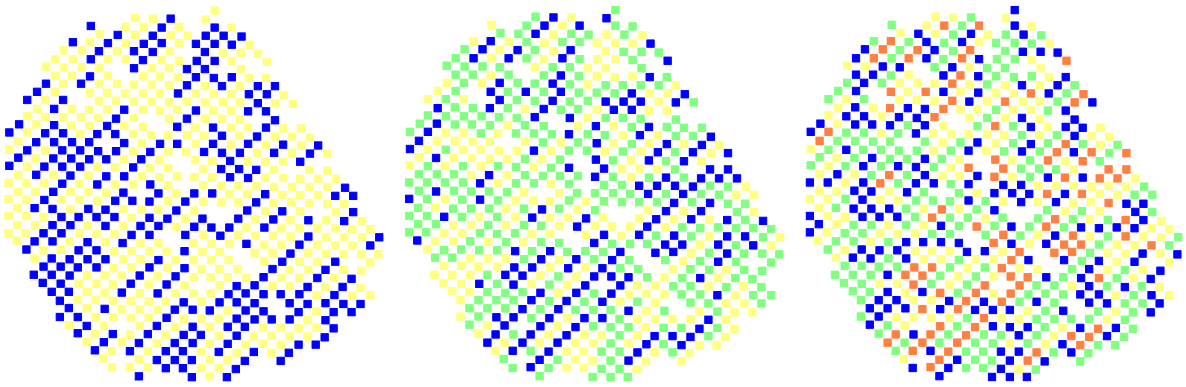


Fig. 2.1. Mapa con clases delimitadas: dos (izquierda), tres (centro) y cuatro (derecha) clases.

Los índices utilizados para la selección número óptimo de clases no fueron coincidentes en la indicación del número de *cluster* a retener. El Coeficiente de Partición y el índice de Fucuyama-Sugeno sugieren que la partición óptima se encuentra con cuatro clases mientras que para el índice de Entropía de la clasificación y el de Xie-Beni el óptimo sugerido, es de dos clases. Un índice resumen que contiene información de cada uno de los índices previamente calculados, indicó que la partición óptima es de dos clases.

Tabla 2.1. Selección del número de clases de la partición de sitios intralote a partir del análisis de cluster *fuzzy k-means*.

Índice[‡]	2 clases	3 clases	4 clases
Coeficiente de Partición	1.050	1.070	1.110
Entropía de Partición	0.0811	0.118	0.169
Xie-Beni	2.19E-04	3.20E-04	6.17E-04
Fukuyma-Sugeno	-2.33E+05	-2.78E+05	-3.03E+05
Índice Resumen	1.5019	1.767	2.168

[‡]Para cada índice se indica el número óptimo de clases sugerido subrayando el mejor valor del índice.

CLASIFICACIÓN DE SITIOS INTRALOTE VÍA PCA Y MULTISPATI-PCA

En las Tablas 2.2 y 2.3 se presentan los varianzas y autocorrelación de cada una de las componentes principales generadas a partir de las propiedades del suelo y rendimiento por el análisis MULTISPATI-PCA y PCA, respectivamente. Los resultados muestran que con MULTISPATI-PCA se explica una menor proporción de la varianza acumulada en los tres primeros ejes respecto a PCA (1.816 vs. 1.890 para el eje 1, 0.996 vs. 1.116 para el eje 2, 0.936 vs. 0.999 para el eje 3). En ambos análisis los 6 ejes presentan autocorrelación espacial significativa ($p<0.001$). No obstante el índice de Moran sugiere que la autocorrelación aumenta cuando se usa MULTISPATI-PCA (0.532 vs 0.490 para el eje 1, 0.467 vs 0.255 para el eje 2, 0.392 vs. 0.388 para el eje 3). Mientras que a nivel del eje 4, 5 y 6 este comportamiento fue inverso.

Tabla 2.2. Autovalores, varianza espacial e índices de Moran de las componentes principales generados a partir de MULTISPATI-PCA sobre cuatro variables de suelo y dos de rendimiento.

Eje	Autovalores	Varianza espacial	Proporción‡	Proporción acumulada	Índice de Moran
1	0.966	1.816	0.303	0.303	0.532
2	0.465	0.996	0.166	0.469	0.467
3	0.367	0.936	0.156	0.625	0.392
4	0.252	0.945	0.158	0.782	0.267
5	0.147	0.596	0.099	0.882	0.246
6	0.130	0.711	0.118	1.000	0.183

‡Proporción de varianza total y covarianza espacial explicada por cada eje.

Tabla 2.3. Autovalores (varianza) e índices de Moran de las componentes principales generadas a partir del PCA sobre cuatro variables de suelo y dos de rendimiento.

Eje	Autovalores (varianza)	Proporción	Proporción acumulada	Índice de Moran
1	1.890	0.315	0.315	0.490
2	1.116	0.816	0.501	0.255
3	0.999	1.484	0.668	0.388
4	0.972	2.313	0.830	0.328
5	0.558	3.236	0.923	0.466
6	0.464	4.236	1.000	0.324

La Fig. 2.2 muestra la representación gráfica de los dos primeros ejes del PCA y MULTISPATI-PCA y sus autovalores (gráfico de barras). Las barras de color negro corresponden a la cantidad de ejes seleccionados que fueron utilizados para la representación gráfica e interpretación de la variabilidad subyacente, en este caso las dos primeras componentes principales. La altura de cada barra representa la proporción de la variabilidad total reflejada por cada componente principal. Así, para MULTISPATI-PCA analizar las dos primeras sPC aporta suficiente información para el análisis. Mientras que para el PCA la selección del número de PC a tener en cuenta para la interpretación de la variabilidad es menos trivial ya que, la tercera y cuarta PC explican, aproximadamente, la misma proporción de variabilidad que la segunda PC. Por simplicidad, se interpretan aquí las primeras dos dimensiones. La variación en el eje 1, para ambos métodos, fue impulsada principalmente por las variables Tg, Sj y CE30 (variables de mayor proyección sobre el eje de las abscisas).

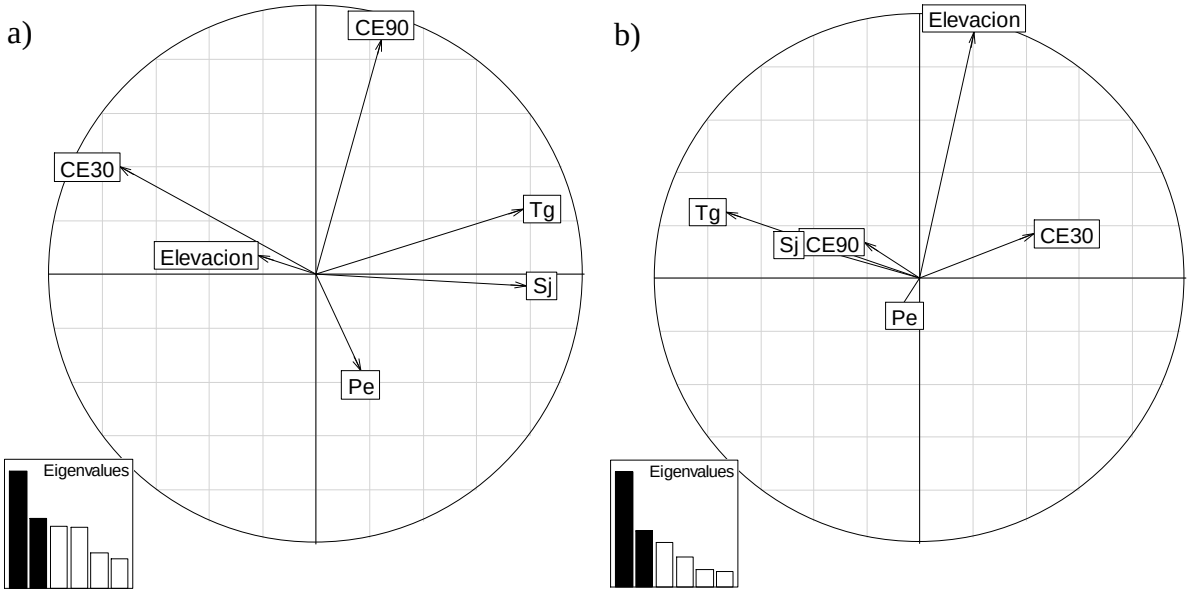


Fig. 2.2. Representación gráfica de los dos primeros ejes del PCA (a) y MULTISPATI-PCA (b) y sus autovalores. Se muestra la correlación entre las variables y entre estas y las componentes principales.

El reposicionamiento de CE90 sobre el primer eje de MULTISPATI-PCA para indicar la presencia de correlación positiva entre los rendimientos y la CE90, produjo un cambio de ponderación de las variables sobre la sPC2 (eje de ordenadas) haciendo que ésta quede más correlacionada con la E. La CE30 en ambos análisis se correlacionó en forma

negativa con los rendimientos. Ambos rendimientos (trigo y soja) presentaron una mayor correlación utilizando MULTISPATI-PCA. La Pe tuvo la menor correlación con los rendimientos, pero en ambos análisis esta correlación positiva señala que aquellos lugares con mayor profundidad de tosca o con menos impedimentos físicos, fueron los de mayor rendimiento.

Las Fig. 2.3 y 2.4 muestran los mapas de la primer y segunda componente principal del PCA y MULTISPATI-PCA, respectivamente Estos fueron obtenidos asignando los valores de las componentes a los sitios del lote, utilizando para ello las coordenadas X e Y. En estos mapas los diferentes tamaños de los símbolos utilizados (cuadrados) representan diferentes valores absolutos de las variables sintéticas: sitios con cuadrados negros y grandes son bien diferenciados de los sitios con cuadros blancos y grandes, mientras que observaciones representadas con pequeños cuadrados son menos diferenciadas entre ellas.

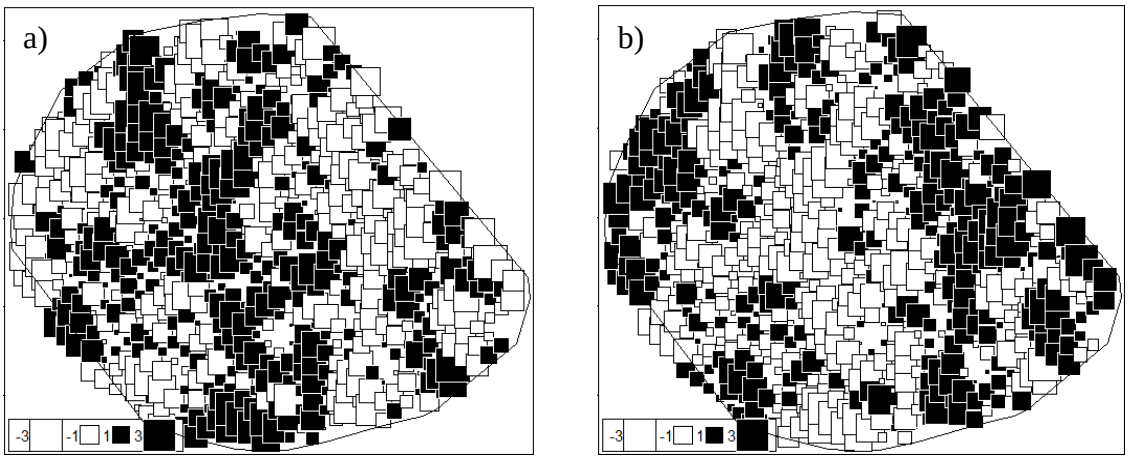


Fig. 2.3. Mapas de la PC1 del PCA (a) y sPC1 del MULTISPATI-PCA (b).

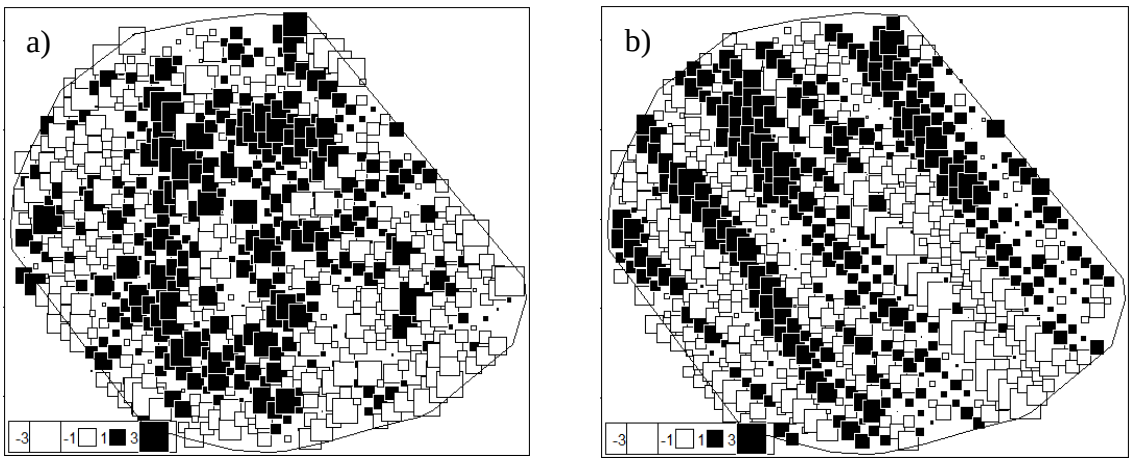


Fig. 2.4. Mapas de la PC2 del PCA (a) y sPC2 del MULTISPATI-PCA (b)

Como era de esperar por la estructura de correlación de las observaciones, donde las mismas variables lideran la PC1 y sPC1, los mapas de la primer componente de ambos métodos son muy similares, aunque en sentido inverso, es decir para PCA los sitios con alto valor de la PC1 corresponden a sitios con los valores más bajos de la sPC1 del MULTISPATI-PCA. Puede observarse que los sitios con altos valores positivos (cuadrados negros) o negativos (cuadrados blancos) para PC1 y sPC1 respectivamente, corresponden a sectores del lote donde se obtuvo un mayor valor de Sj, Tg y que presentan menores valores de CE30 (Fig. 2.3). Adicionalmente, en el mapa de la sPC1 los cuadrados blancos poseen también mayores valores de CE90. En los mapas de la PC2 para el PCA se representa principalmente la variación de la CE90, siendo los cuadrados negros sitios con alto valor de CE90. Mientras que en la sPC2 de MULTISPATI-PCA la variabilidad está representada por la E, una variable no bien explicada por PCA. Los cuadrados blancos corresponden a sitios que presentan mayor E.

La estructura espacial se manifiesta por el hecho de que puntos del mismo color presentan agrupamiento espacial. Al comparar la Fig. 2.3 a) y la Fig. 2.3 b), se observa que en la sPC1 (MULTISPATI-PCA) las zonas están más definidas, el mapa es más suavizado. Lo mismo se observa al comparar los mapas de la componente dos (Fig. 2.4). Estas dos componentes fueron las que tuvieron el mayor valor de autocorrelación espacial (Tabla 2.2 y 2.3)

A partir de las componentes principales obtenidas del PCA y MULTISPATI-PCA se estudio la presencia de estructura espacial. En la Tabla 2.4 se observan los valores de las estimaciones de los parámetros del semivariograma ajustados por MLM. Para sPC1, sPC2 y PC1 el modelo que mejor explico la variabilidad espacial fue el esférico, siendo su estructura espacial de magnitud fuerte ($0.25 < C_0/C_0+C$). Si se compara sPC1 con PC1 la magnitud de estructuración espacial es algo mayor en la variable sintética obtenida con MULTISPATI-PCA. La PC2 mostró una estructura espacial exponencial de magnitud intermedia ($0.25 < C_0/C_0+C < 0.75$).

Tabla 2.4. Estimaciones de los parámetros del MLM ajustado para las componentes principales del PCA (CP) y MULTISPATI-PCA (sPC).

	Modelo	C_0	C	C_0/C_0+C	R_P
CP1	Esférico	0.06	1.85	0.03	85
CP2	Exponencial	0.60	1.28	0.32	152
sPC1	Esférico	0.03	1.85	0.02	93
sPC2	Esférico	0	0.83	0	57

C_0: varianza nugget, C: varianza estructural, R_p: rango práctico.

En las Fig.s 2.5 y 2.6, pueden observarse los mapas variabilidad multivariados obtenidos como resultado de la interpolación por kriging para la PC1 y PC2 del PCA y sPC1 y sPC2 del MULTISPATI-PCA. Si se comparan los mapas obtenidos de cada variable con el mapa obtenido a partir de las variables sintéticas (Fig. 2.7), se observan las correlaciones descriptas anteriormente. Por ejemplo, la sCP1 presenta correlación positiva con CE30 y negativa con los rendimientos y CE90, por lo tanto, puede observarse que los sectores con valores más altos de las componentes coinciden con los sitios donde existen altos valore de CE30, mientras que los sitios que presentan valores negativos de las componentes, están asociados con aquellos de mayor rendimiento. A nivel de la sPC2 la similitud es más evidente con la variable E mientras que para PC2 su mapa de variabilidad espacial es similar al de CE90.

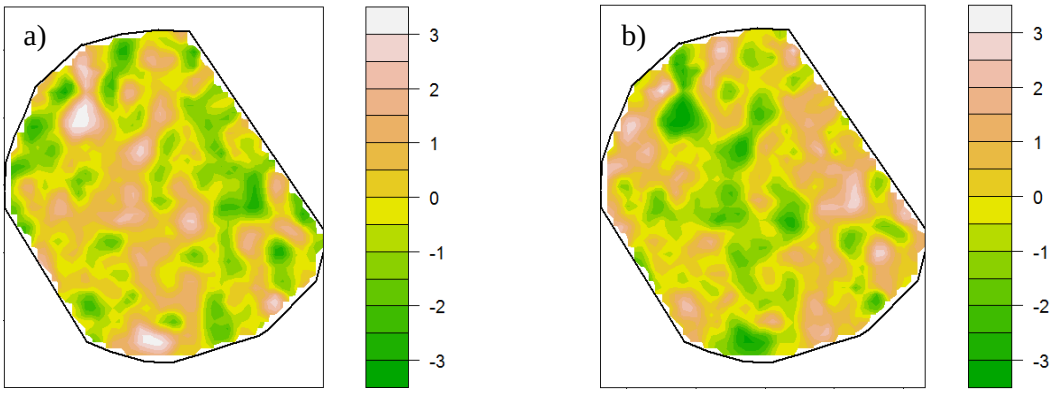


Fig. 2.5. Mapas obtenidos por interpolación (Kriging) de la PC1 del PCA (a) y sPC1 del MULTISPATI-PCA (b).

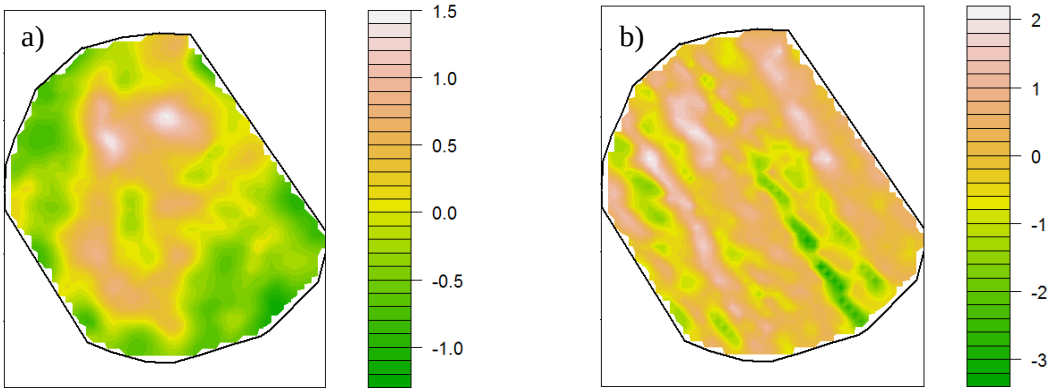


Fig. 2.6. Mapas obtenidos por interpolación (Kriging) de la PC2 del PCA (a) y sPC2 del MULTISPATI-PCA (b).

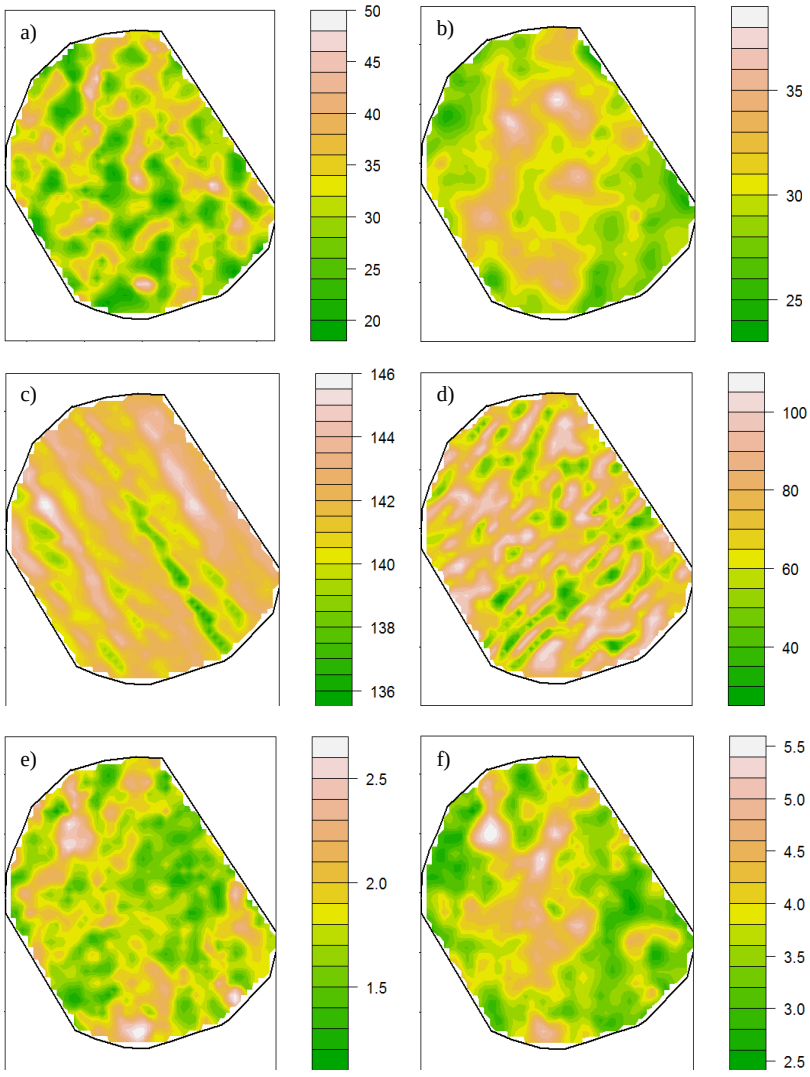


Fig. 2.7. Mapas de variabilidad espacial de variables de suelo y rendimiento obtenidos mediante la interpolación por kriging ordinario utilizando parámetros del semivariograma estimados con modelos lineales mixtos. a) Conductividad eléctrica aparente a 30 cm de profundidad, b) Conductividad eléctrica aparente a 90 cm de profundidad, c) Elevación, d) profundidad de tosca, e) rendimiento de soja, f) rendimiento de trigo.

DISCUSIÓN

La agricultura de precisión está demandando nuevas técnicas de análisis, entre las que se encuentran combinaciones de análisis geoestadísticos y multivariados para capturar la naturaleza de la observación espacial multivariada (Saby *et al*., 2009; Arrouays *et al*., 2011; Moral *et al*., 2010; Milne *et al*., 2012; Diacono *et al*., 2012). En este trabajo se aplicaron tres métodos para analizar datos espaciales multivariados (*cluster fuzzy k-means*, PCA y MULTISPATI-PCA) en una base de datos derivada del uso de maquinarias de agricultura de precisión, como son monitores de rendimientos acoplados a cosechadoras de granos y sensores de CE, medición de profundidad de tosca y elevación. Estas variables de suelo impactan altamente en los rendimientos de cultivos de granos (Corwin y Lesch, 2010), como soja y trigo, en la región pampeana argentina y por lo tanto podrían, junto con el rendimiento constituir indicadores sensibles a las diferencias en terreno para favorecer la delimitación de zonas homogéneas y el posterior manejo de precisión de la producción.

Cuando se realizó el análisis de *cluster fuzzy k-means* los indicadores del número de *cluster* óptimos no coincidieron. En este tipo de situaciones donde los índices se contradicen, en la selección del número de clases óptimo, es de gran utilidad el uso de un índice resumen que contenga información de cada uno de los índices calculados. Para esta ilustración el índice resumen indicó que el óptimo era de dos clases. Esto es coincidente con una selección que se haría de acuerdo a lo practicable por el productor, ya que con dos clases se produce menor fragmentación de las clases delimitadas. Sin embargo, aún con la selección de dos clases, será necesario lograr una mayor continuidad espacial a la clasificación lograda por el algoritmo *fuzzy k-means*. Lotes con grandes zonas con límites coherentes son más fáciles de manejar diferencialmente que lotes con numerosas zonas pequeñas y de forma irregular. Obtener zonas espacialmente estructuradas y con diferencias agronómicas significativas entre clases es uno de los requisitos para aplicar manejos diferenciales. (Taylor *et al*., 2007). Para ello, se pueden utilizar diferentes métodos que permitan reducir la fragmentación de las clases.

Utilizando MULTISPATI-PCA fue posible realizar un PCA que incorpore las coordenadas espaciales de los sitios monitoreados para considerar la posible presencia de autocorrelación espacial y ajustar las estimaciones de varianza del proceso monitoreado.

Con estos ajustes espaciales se produjo un reposicionamiento de CE90 sobre el primer eje de MULTISPATI-PCA (eje de mayor importancia del análisis). Es importante resaltar el peso que MULTISPATI-PCA le da a la CE90, ya que esta variable es frecuentemente utilizada en remplazo de la CE30 debido a que las mediciones de CE90 son más estables en el tiempo (Veris Technologies, 2001; Sudduth *et al*., 2003; Peralta *et al*., 2013). Adicionalmente, se produjo un cambio de ponderación de las variables sobre la sPC2 haciendo que ésta quede más correlacionada con la E, permitiendo así analizar la variabilidad del rendimiento desde otra dimensión distinta a la de la CE. La covariación entre los rendimientos de soja y trigo fue aún mayor que la detectada por PCA. Desde un punto de vista agronómico se supone que es más probable que ambas variables presenten una alta correlación. Se trata de rendimientos correspondientes a la misma campaña agrícola. Aun cuando el nivel de concordancia entre las dos primeras componentes principales de ambos métodos puede ser significativo y/o alto, como se ha reportado en otras aplicaciones del PCA con restricción espacial respecto del PCA clásico (Arrouays *et al*., 2011; Dray *et al*., 2008; Gili *et al*., 2013), es de esperar diferencias entre los ordenamientos de los puntos o sitios de muestreos.

Los resultados observados sugieren que los mapas de variabilidad espacial logrados por estas técnicas pueden ser diferentes debido a que en MULTISPATI-PCA la varianza es corregida por la autocorrelación espacial. La incorporación de las coordenadas espaciales *a priori* del análisis hizo que la selección del número de componentes principales para la interpretación de la variabilidad fuese más clara que en PCA clásico, resultado también publicado por otros autores que trabajaron en escalas espaciales mayores o regionales y no a nivel intralote (Dray y Jombart, 2011). El grado de estructuración espacial fue mayor con MULTISPATI-PCA que con PCA no restringido espacialmente. Esto se ve reflejado en los mapas de las variables sintéticas donde la estructura espacial es más clara con sPC1 y sPC2 que con PC1 y PC2, respectivamente. El mapeo de las componentes principales espaciales permite una mejor visualización del patrón de la variabilidad espacial. El suavizado logrado en los mapas generó zonas más definidas, lo cual constituye un aspecto clave para la delimitación de zonas de manejo en agricultura sitio-específica. La fragmentación observada en los mapas obtenidos con MULTISPATI-PCA fue menor a la obtenida mediante el algoritmo de *cluster fuzzy k-means.*

CONCLUSIÓN

Los resultados mostraron que la incorporación de la autocorrelación espacial antes de construir las componentes principales permite detectar relaciones subyacentes en el ensamblaje de variables originales que no fueron tenidas en cuenta si la estructura espacial fuera incorporada *a posteriori*. MULTISPATI-PCA permitió estudiar la distribución espacial de los rendimientos y propiedades de suelo con menor cantidad de componentes principales que PCA. Los mapas derivados de las componentes principales espaciales permitieron una mejor visualización del patrón de variabilidad espacial que el obtenido con los análisis sin restricción espacial. El método MULTISPATI-PCA constituye una herramienta estadística promisoria no sólo para el el mapeo de la variabilidad conjunta de variables de suelo y rendimiento capturadas dentro de lotes monitoreados con maquinarias precisas sino también para la delimitación de zonas homogéneas en sentido multivariado.

CAPÍTULO III

PROPUESTA PARA LA CLASIFICACIÓN DE SITIOS INTRALOTE

INTRODUCCIÓN

El conocimiento de la variabilidad del rendimiento dentro de lotes en producción, es esencial para el manejo sitio específico, uno de los objetivos de la agricultura de precisión (AP) orientado a optimizar el uso de los insumos agrícolas. Las nuevas tecnologías en maquinarias agrícolas asociadas a la AP proporcionan la oportunidad de medir, con mayor precisión, la variabilidad espacial en el rendimiento cosechado y en numerosas variables de sitio intralote. El punto de partida para aplicar el manejo sitio-específico es delimitar zonas de manejo (ZM) dentro de los lotes. Estas subregiones constituyen áreas con características similares, tales como textura, topografía, estado hídrico y niveles de nutrientes del suelo (Moral *et al*., 2010). Por lo tanto, bajo el mismo tratamiento, se espera que las subregiones delimitadas presenten diferencias en rendimiento.

Una subregión del lote es considerada como perteneciente a una ZM si un tratamiento en particular puede aplicarse a esta región. El término ZM se usa típicamente cuando se trata de regiones espacialmente contiguas (Taylor *et al*., 2007). El agricultor generalmente quiere identificar ZM con localizaciones espacialmente coherentes (Milne *et al*., 2012). Para realizar la delimitación de ZM generalmente se comienza con la aplicación de un algoritmo de clasificación de sitios intralote según valores de variables de suelo y/o rendimiento. Para ello, es necesario la aplicación de algoritmos de clasificación multivariados que contemplen la naturaleza espacial de los datos. Las propiedades físicas y químicas del suelo y la topografía son frecuentemente utilizadas para delimitar ZM. La conductividad eléctrica aparente (CE) a distintas profundidades es también muy utilizada ya que su variabilidad ha sido reportada como indicador de la distribución espacial de otras propiedades de suelo (Corwin *et al*., 2006; Corwin y Lesch, 2010; Moral *et al*., 2010;

Rodríguez-Pérez *et al*., 2011; Peralta *et al*., 2013). La elevación del terreno así como otras propiedades topográficas (pendiente, curvatura, índices de humedad), las cuales pueden ser estimadas con modelos digitales de elevación, también proveen importante información para la clasificación de sitios intralote. Estas variables también han sido mencionadas como indicadoras de variabilidad espacial del rendimiento ya que afectan directamente el crecimiento y desarrollo de los cultivos. El flujo y acumulación del agua en diferentes posiciones del terreno, así como por la redistribución de partículas minerales del suelo y materia orgánica a través de la erosión y deposición del suelo, son factores a los que se atribuye parte de la variabilidad espacial en el rendimiento (Pachepsky *et al*., 2001).

En suelos de la Pampa húmeda y suhúmeda Argentina; la profundidad del horizonte petrocálcico conocido localmente como "tosca" (Buschiazzo, 1986), es otra variable que suele ser usada en la clasificación de sitios (Peralta *et al*., 2012). La profundidad efectiva del suelo, dada por la tosca, interviene en la distribución espacial del agua en el suelo y en su capacidad de almacenaje, lo que impacta diferentemente sobre la disponibilidad de agua y nutrientes accesibles para los cultivos. Los mapas de rendimientos de campañas previas también suministran información para la clasificación de sitios. No obstante, éstos deben analizarse simultáneamente con el patrón de variabilidad dado por otras variables, ya que usar sólo mapas de rendimiento para la clasificación, introduce el sesgo dado por variaciones asociadas a cambios menos predecibles (climáticos) (Corwin y Lesch, 2010).

Respecto a los métodos para la clasificación de los sitios intralote, varios algoritmos del análisis de *cluster* no jerárquicos como *k-means* (Anderberg, 1973), han sido usados en AP (Stafford *et al*., 1998). El análisis de *cluster* agrupa sitios similares dentro de distintas clases llamadas "*clusters*" en el espacio *p*-dimensional de los atributos o variables de sitio medidas sobre cada sitio intralote. Los métodos de clasificación basados en la teoría difusa (Burrough, 1989) permiten que cada observación pueda ser asignada a más de un *cluster*, con diferentes grados de pertenencia y han sido adoptados en AP ya que permiten contemplar la variación continua de las variables del suelo. Si bien los métodos de agrupamiento difuso asignan a cada sitio un grado de pertenencia a cada uno de los *cluster* de sitios conformados, en la práctica, es necesario asignar cada sitio a una clase única y esto se hace en función del valor de pertenencia máxima en un proceso llamado "defuzzificación" (Guastaferro *et al*., 2010). El algoritmo de *cluster fuzzy k-means* ha sido ampliamente utilizado para identificar potenciales zonas de manejo en agricultura de

precisión (Boydell y McBratney, 2002; Li *et al*., 2007; Davatgar *et al*., 2012; Peralta *et al*., 2013). Sin embargo, este algoritmo es no restringido espacialmente y suele aplicarse usando como *inputs* variables de suelo o, alternativamente, combinaciones lineales de éstas que tienen asociada información espacial.

El Análisis de Componentes Principales (PCA) es comúnmente usado para construir Componentes Principales (PC) con variables de suelo en AP (Schepers *et al*., 2004; Li *et al*., 2007; Xin- Zhong *et al*., 2009). La implementación del PCA se realiza para reducir el número de variables originales disponibles para la clasificación y resumir la variabilidad de múltiples mediciones en unas pocas variables sintéticas. La primera variable sintética, PC1, contiene las principales señales de la variabilidad conjunta de todas las mediciones realizadas sobre el mismo sitio, y la última CP es comúnmente asociada a "ruido" o variabilidad espuria.

Los análisis de conglomerado no restringidos espacialmente, como *k-means* o *fuzzy k-means*, no incluyen referencia respecto a la posición geográfica de los sitios en los que se registran las variables. Algunas propuestas de delimitación de ZM intentan restringir espacialmente el algoritmo de *cluster* (Ping y Dobermann, 2003; Frogbrook y Oliver, 2007; Milne *et al*., 2012) pero estos no han sido ampliamente adoptados (Pedroso *et al*., 2010). Oliver y Webster (1989) utilizaron el variograma para incorporar la restricción espacial en el algoritmo *k-means*. Los autores modificaron la matriz de disimilitud que contiene las distancias entre cada par de puntos de muestreo antes de la clasificación. Bourgault *et al*. (1992) exploró la técnica y demostró que la clasificación podría ser suavizada con más fuerza si las similitudes se modifican con las covarianzas en vez de usar variogramas (Milne *et al*., 2012). En el enfoque propuesto por Lark (1998), a diferencia de los métodos anteriores, el suavizado espacial se realiza después de la clasificación de los sitios, también basada en el algoritmo *fuzzy k-means*. Los valores de pertenencia de cada observación a cada *cluster* son objeto de un suavizado espacial cuando se trabaja con datos espaciales.

Dray *et al*. (2008) proponen una forma no de *cluster* sino de PCA que incorpora la información espacial previo a la conformación de las variables sintéticas, el método es conocido como MULTISPATI-PCA. La restricción dada por los datos espaciales, se incorpora mediante el índice de Moran para medir la dependencia o correlación espacial

entre las observaciones en un sitio y el promedio de observaciones multivariadas en el vecindario de ese sitio. MULTISPATI-PCA ha resultado provechoso en estudios de suelos realizados a escala macrogeográfica (Arrouays *et al*., 2011), pero no registra su aplicación en la escala de sitios intralote en AP.

En este capítulo se propone utilizar las componentes principales espaciales obtenidas con MULTISPATI-PCA como *input* del análisis de *cluster fuzzy k-means* para la identificación de clases de sitios intralote. El nuevo método, denominado KM-sPC, se diferencia de los propuestos por Oliver y Webster (1989) y Bourgault *et al*. (1992), ya que la autocorrelación espacial es tenida en cuenta antes de la obtención de la matriz de distancias para aplicar el algoritmo *fuzzy k-means*. La hipótesis que subyace a la metodología propuesta es que la incorporación de la autocorrelación espacial a través del PCA espacial aplicado sobre variables del suelo y terreno producirá clases de sitios que contiene menos "ruido" , *i.e.* más homogéneas y contiguas, que las derivadas del algoritmo *fuzzy k-means* no restringido espacialmente. En el presente capítulo se ilustra y evalúa la metodología propuesta como herramienta para la identificación de clases de sitio intralote que pueden luego ser usados para delimitar zonas para manejo sitio-específico.

MATERIALES Y MÉTODOS

DATOS

Los datos fueron recolectados en tres lotes agrícolas ubicados al sudeste pampeano de la provincia de Buenos Aires, Argentina. Conforman el archivo *datosL3.txt* que se encuentra disponible en https://drive.google.com/file/d/0B_8UVonay55COUNrTy1zempab3c/edit?usp=sharing. La región Pampeana es una vasta planicie de alrededor 50 Mha y es considera como una de las áreas más adecuadas para la producción de cultivos de granos en el mundo (Satorre y Slafer, 1999). El clima de esta región es subhúmedo-húmedo, según índice hídrico de Thornthwaite (Burgos y Vidal, 1951), con una precipitación de 880 mm por año y una temperatura media anual de 13,3°C. Los suelos predominantes de esta región pertenecen al orden de los Molisoles, gran grupo Argiudoles o Paleudoles, desarrollados sobre sedimentos loéssicos, bajo régimen údico-térmico. El sitio

experimental esta principalmente constituido por la serie Azul (fina, mixta, térmica, Paleudol Petrocalcico) (SAGyP-INTA ,1989). Los lotes, denominados L1, L2, y L3 con una superficie de 65.4, 60.8 y 72 ha, respectivamente, fueron cultivados con trigo en el 2007 y soja en el 2008.

Se registraron mediciones georreferenciadas de conductividad eléctrica aparente (CE) en dos profundidades 0-30 cm (CE30) y 0-90 cm (CE90), elevación, profundidad de tosca (Pe) y rendimiento de soja (Sj) y trigo (Tg). Los valores de CE fueron tomados utilizando un sensor (Veris 3100, Division of Geoprobe Systems, Salina, KS). El sensor Veris 3100 recorrió el lote en una serie de transectas paralelas espaciados a intervalos de 15 a 20 m, debido a que una separación de más de 20 m genera errores de medición (Farahani y Flynn, 2007). Los datos de CE fueron simultáneamente georreferenciados con un DGPS (Trimble R3, Trimble Navegation Limited, USA) con una exactitud de medición submétrica y configurado para tomar la posición del satélite cada segundo. Los datos de elevación del terreno también se midieron con un DGPS y se procesaron para obtener una precisión vertical de entre 3 y 5 cm aproximadamente. Las mediciones de profundidad de tosca se realizaron utilizando un penetrómetro hidráulico (Gidding) acoplado a un DGPS en una grilla regular de 30 m. Para cuantificar el rendimiento en grano del cultivo se utilizó un monitor de rendimiento acoplado a un equipo de cosecha conectados a un DGPS.

Los datos de suelo y rendimiento fueron sometidos a procedimientos de depuración vía la construcción de gráficos box-plots para la identificación de valores extremos. Así se excluyeron los valores que se encontraban fuera del intervalo media ± 4 desvíos estándares. También se depuraron respecto *inliers* usando el índice de autocorrelación local de Moran (Anselin, 1995) y su diagrama de dispersión (Anselin, 1996). Debido a las diferentes resoluciones espaciales y posiciones geográficas usados para la medición de las diferentes variables, se llevó toda la información a una grilla de 30 m × 30 m. Los datos medidos se asignaron a los nodos de la grilla más cercanos en el espacio. Cuando se tenía más de un dato, para el nodo de una variable, se asignó al mismo el promedio de sus mediciones. De esta forma, se obtuvo un conjunto de mediciones con igual georreferenciación para cada lote. El número final de sitios con datos fue de 664, 676 y 378 para los lotes 1 (L1), 2 (L2) y 3 (L3), respectivamente.

PROPUESTA ALGORÍTMICA PARA LA CLASIFICACIÓN DE SITIOS INTRALOTE

La propuesta para la delimitación de clases de manejo se basa en los análisis MULTISPATI-PCA y *cluster fuzzy k-means* descriptos en la literatura del análisis multivariado (Dray *et al*., 2008; Fridgen *et al*., 2004; Bezdek, 1981). Los datos a clasificar deben estar debidamente pre-procesados para eliminar valores extremos. Además de las variables registradas, la base de datos debe incluir las coordenadas espaciales de cada punto de datos. Las coordenadas geográficas son generalmente convertidas a coordenadas cartesianas. Esto permite que las distancias se muestren como absolutas (metros) en lugar de distancias relativas (grados). La etapa de pre-procesamiento se puede realizar utilizando cualquier Sistemas de Información Geográfica (SIG).

El próximo paso del algoritmo es aplicar MULTISPATI-PCA a las variables de suelo y terreno (CE30, CE90, Elevación y Pe) y obtener las componentes principales espaciales (sPC). El análisis puede ser realizado con los paquetes "ade4" (función *multispati*, Chessel *et al*., 2004) y "spdep" (Bivand *et al*., 2013a) del software R (R Core Team, 2013). La red de vecindarios es definida en función de la distancia Euclídea considerando puntos vecinos a aquellos contiguos ubicados entre los 0 a 70 m de distancia. Para ello se utilizó la función *dnearneigh* del paquete "spdep". Usando el paquete "ade4", el software R devuelve un objeto de clase *multispati*, que contiene varios elementos, entre ellos las sPC. Un conjunto reducido de estas variables sintéticas resultantes, que explican una gran cantidad de la variación total (≥ 70%), son posteriormente usadas como *input* del análisis de *cluster fuzzy k-means*.

Finalmente se realiza la aplicación del análisis de *cluster fuzzy k-means* usando las componentes principales espaciales como variables en las que se basa la clasificación. Así la matriz de datos utilizada en el análisis *fuzzy k-means* incluye las *n* observaciones cada una con $a < p$ componentes principales espaciales. La distancia Euclídea se utiliza como medida de similitud en la función de optimización del *fuzzy k-means*, ya que las componentes principales son independientes y se estandarizan cuando se realiza el análisis MULTISPATI-PCA, por lo tanto sus varianzas no difieren. El exponente difuso se fija en el valor convencional de 1,30 (Odeh *et al*., 1992). Alternativamente, el algoritmo *fuzzy k-means* puede ser implementado desde otros software como MZA (Fridgen *et al*., 2004) o FuzMe (Minasny y McBratney, 2002) que además de trabajar con la distancia Euclídea

permiten utilizar las distancias de Mahalanobis o Diagonal que son apropiadas cuando las variables no son estadísticamente independientes y/o presentan varianzas diferentes.

El coeficiente de partición (conocido también como fuzziness performance index, FPI) y la entropía de clasificación normalizada (normalized classification entropy, NCE) (Odeh *et al*., 1992) se pueden utilizar para determinar el número óptimo de *clusters*. Este se obtiene cuando ambos índices se reducen al mínimo, lo que representa el menor solape entre los grupos (FPI) o el mayor grado de organización (NCE) como consecuencia del proceso de agrupación de los datos (Fridgen *et al*., 2004). Para ejecutar el nuevo algoritmo denominado KM-sPC, se utilizaron los *scripts* desarrollados en el software R (R Core Team, 2013) (Anexo 4).

EVALUACIÓN DEL ALGORITMO KM-SPC

PROCEDIMIENTOS COMPARADOS CON DATOS EXPERIMENTALES

Para cada lote, L1, L2 y L3, tres algoritmos diferentes fueron llevados a cabo simultáneamente para la delimitación de clases de sitios. En primer lugar la clasificación usando *fuzzy k-means* en MZA (Fridgen *et al*., 2004) se realizó con las variables del suelo medidas originalmente (KM-SV). Debido a que las mismas presentaban varianzas desiguales y covarianzas no nulas, se utilizó como medida de similitud la distancia de Mahalanobis. El mismo análisis de *cluster* también se realizó utilizando como *input* las componentes principales provenientes del PCA, no restringido espacialmente, sobre las variables del suelo (KM-PC) y también sobre las componentes principales espaciales obtenidas previa aplicación de MULTISPATI-PCA (KM-sPC). En los análisis de componentes principales se usaron para la clasificación vía *fuzzy k-means*, las tres primeras variables sintéticas. Los resultados obtenidos de la aplicación de cada uno de estos tres algoritmos fueron comparados estadísticamente y agronómicamente.

EVALUACIÓN MEDIANTE SIMULACIÓN

Se realizó una simulación geoestadística como una manera adicional para evaluar el método propuesto. Con los datos simulados se llevaron a cabo comparaciones simultáneas

para evaluar la performance relativa de KM-sPC respecto a KM-SV y KM-PC. La simulación geoestadística permitió generar realizaciones, igualmente probables, de un proceso descrito por funciones de dependencia espacial. Se utilizó un campo aleatorio que simula un proceso estocástico gaussiano por medio de un variograma exponencial y se simularon valores de CE30, CE90, elevación (E), profundidad de tosca (Pe) y rendimiento usando parámetros distribucionales del orden de los observados en los datos de campo o experimentales. Se simularon 100 realizaciones para cada variable con el paquete "geoR" (Ribeiro Jr. y Diggle, 2001) del software R. La simulación se realizó de manera que las variables presentaran una componente espacial individual y una componente común para generar correlación espacial entre ellas (Diggle y Ribeiro Jr., 2007). Los tres métodos para delimitar clases de manejo descriptos arriba se aplicaron en cada simulación y se evaluó la repetitividad de los resultados.

Tanto para los datos experimentales como para los simulados, se realizó posteriormente un análisis de varianza (ANAVA) sobre los datos de rendimiento y suelo con la finalidad de evaluar las particiones o clases obtenidas bajo los diferentes algoritmos. Se compararon las diferencias entre los rendimientos medios de las clases delimitadas por cada método, a través de un modelo lineal mixto (MLM) para datos correlacionados espacialmente ya que el método propuesto tiene en cuenta la correlación espacial en las variables del suelo, pero no contempla la correlación de los datos de rendimiento. La prueba de comparación de medias LSD de Fisher se aplicó sobre las medias ajustadas por el MLM de las clases de sitio. Bajo autocorrelación positiva, se espera que las diferencias entre las medias de las clases ajustadas sean generalmente más pequeñas que las diferencias entre las medias no ajustadas. La correlación espacial entre los datos de rendimiento se modeló por medio de una función exponencial de los términos de error (Schabenberger y Pierce, 2002). Se comparó no sólo la magnitud de las diferencias en rendimiento entre las clases de manejo delimitadas sino también la varianza residual (DE) dentro de las clases y el EE de la diferencia entre las medias de las clases.

Adicionalmente, para comparar la eficiencia de los diferentes métodos en la explicación de varianza del rendimiento en cada año *j*, se utilizó el complemento de la varianza relativa (Webster y Oliver, 1990):

$$RV_j = 1 - S_W^2 / S_T^2 \quad (3.1)$$

donde S_W^2 es la varianza dentro de la clase y S_T^2 es la varianza total, ambas estimadas por el análisis de la varianza para un año en particular *j*. RV*j* es una medida de la proporción de la varianza explicada por la clasificación. Una clasificación perfecta daría como resultado un valor de la varianza dentro de la clase de cero y un RV*j* de 1 (Ping y Dobermann, 2003). Para facilitar la interpretación de los resultados estadísticos, se calcularon las correlaciones entre las variables del suelo y el rendimiento. La misma se evaluó utilizando el paquete "SpatialPack" de software de R que permite contemplar la autocorrelación espacial entre las variables (Osorio *et al*., 2012). El test-t modificado por la correlación espacial, calcula la correlación producto-momento de Pearson entre dos variables y prueba su significancia mediante el procedimiento de Clifford *et al*. (1989). Adicionalmente, se calculó el Índice de Moran (Moran, 1948) para las variables del suelo originales y las variables sintéticas derivadas del PCA y MULTISPATI-PCA.

RESULTADOS

En la Tabla 3.1 se presentan las medidas descriptivas para las variables de suelo y rendimiento medidas en cada sitio. Puede observarse que la elevación (E) fue la variable de suelo que presentó menor variabilidad relativa y mayor correlación espacial en los tres lotes evaluados. Si bien E mostró poca variabilidad en estos lotes pampeanos dedicados al cultivo extensivo de grano, los CV de CE y Pe fueron altos. La autocorrelación espacial para las variables del suelo fue estadísticamente significativa ($p < 0{,}001$), con excepción de Pe en L3. El amplio rango de variación de Pe en L3 (20 a 150 cm) sugiere que este lote tendría los mayores cambios espaciales para la variable Pe. Los valores reportados sugieren que las variables de suelo, solas o combinadas en variables sintéticas, podrían ser usadas para mapear variabilidad espacial intralote.

Tabla 3.1. Estadística descriptiva de variables de suelo y rendimiento en tres lotes agrícolas monitoreados intensivamente.

Lote	Variables	Media	Mediana	CV	Min	Max	MI[b]
L1	Suelo[a]						
	CE30, mS m^{-1}	31.12	31.22	23.01	15.90	58.00	0.34*
	CE90, mS m^{-1}	30.40	30.15	19.21	13.77	54.60	0.21*
	Elevación, m	141.77	141.88	1.27	135.19	146.18	0.46*
	Profundidad tosca, cm	73.32	71.00	29.96	20.00	101.00	0.27*
	Rendimiento						
	Trigo 2007, t ha^{-1}	3.76	3.67	16.49	2.33	5.66	0.64*
	Soja 2008, t ha^{-1}	1.79	1.78	18.99	1.06	2.76	0.40*
L2	Suelo						
	CE30, mS m^{-1}	30.20	30.81	25.20	8.87	51.38	0.65*
	CE90, mS m^{-1}	28.86	28.60	23.74	8.33	50.04	0.66*
	Elevación, m	140.60	140.18	3.24	128.71	156.25	0.67*
	Profundidad tosca, cm	75.96	75.83	12.51	48.72	112.43	0.32*
	Rendimiento						
	Trigo 2007, t ha^{-1}	3.95	4.00	15.19	2.21	5.87	0.49*
	Soja 2008, t ha^{-1}	1.88	1.88	19.15	0.89	3.02	0.52*
L3	Suelo						
	CE30, mS m^{-1}	25.45	25.42	21.65	13.83	40.94	0.10*
	CE90, mS m^{-1}	25.72	26.03	18.55	11.52	39.07	0.12*
	Elevación, m	146.03	146.15	1.11	141.97	150.01	0.47*
	Profundidad tosca, cm	63.96	60.00	36.34	20.00	150.00	-0.02
	Rendimiento						
	Trigo 2007, t ha^{-1}	3.43	3.45	12.54	1.90	5.24	0.36*
	Soja 2008, t ha^{-1}	2.40	2.36	18.33	0.82	3.67	0.46*

[a] CE30: conductividad eléctrica aparente a 30 cm, CE90: conductividad eléctrica aparente a 90 cm.
[b] Índice de Moran. *Estadísticamente significativo (α=0.05).

Tabla 3.2. Coeficientes de correlación entre variables de suelo y rendimiento para tres lotes en producción agrícola.

Variables	CE30			CE90			E			Pe			Tg		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
Soil															
CE30	1	1													
CE90	0.11	0.69*	0.30*	1	1										
E	0.13	0.74*	0.03	-0.05	0.81*	-0.07	1	1							
Pe	-0.05	-0.26	-0.26*	-0.02	-0.21	0.11	-0.04	-0.20	-0.05	1	1				
Rendimiento															
Tg (2007)	-0.39*	-0.55*	-0.17	0.24*	-0.57*	0.10	-0.07	-0.68*	-0.05	0.03	0.23	0.14	1	1	
Sj (2008)	-0.44*	-0.52*	-0.24*	0.13	-0.57*	-0.04	-0.03	-0.70*	0.14	0.11	0.08	0.17*	0.42*	0.51*	0.10

[a]CE30: conductividad eléctrica aparente a 30 cm, CE90: conductividad eléctrica aparente a 90 cm, E: elevación, Pe: profundidad de tosca, Tg: rendimiento de trigo 2007, Sj: rendimiento de soja 2008, L1: Lote 1, L2: Lote 2, L3: Lote 3.

*Estadísticamente Significativo (α= 0.05) de acuerdo al test-t modificado para contemplar la autocorrelación espacial (con corrección por multiplicidad, según criterio de Bonferroni).

En la Tabla 3.2 se muestra la significancia de los coeficientes de correlación lineal entre las variables originales. Se observa una correlación significativa y positiva entre los rendimientos (r=0.42) para L1. En el lote L1 la CE30 fue la variable de suelo que más se correlacionó con los rendimientos tanto de soja (r=-0.44) y como de trigo (r=-0.39). Mientras que en L2, la elevación fue la variable de mayor correlación con los rendimientos. En L3 la variable CE30 se correlacionó con el rendimiento, pero sólo en el cultivo de soja.

En la Tabla 3.3 se presentan las varianzas y los coeficientes de autocorrelación de las tres primeras variables sintéticas generadas a partir del PCA y MULTISPATI-PCA. Los coeficientes de autocorrelación espacial (MI) de las tres primeras variables sintéticas generadas con PCA y MULTISPATI-PCA fueron estadísticamente significativos. MULTISPATI-PCA produjo una primer componente principal espacial (sPC1) de menor varianza y mayor correlación espacial respecto a la primer componente principal (PC1) del PCA no restringido espacialmente. La misma tendencia se observó con la segunda componente, pero sólo en L1. Los índices de autocorrelación, en la primera componente, aumentaron con el uso de MULTISPATI-PCA respecto a PCA.

En la Tabla 3.4 se presentan los autovectores que representan los coeficientes con que cada variable original de suelo fue ponderada para conformar las combinaciones lineales que produjeron los componentes principales. Al construir la primera componente, la variable E (variable de mayor autocorrelación) fue la que tuvo mayor peso en el primer componente espacial (Tabla 3.4). Sin embargo, en L1 y L3, la conductividad eléctrica aparente fue la variable que más contribuyó con la construcción de la primera componente principal del PCA, y en menor medida también fue importante en la PC1 de L2. La CE90 y E fueron las variables más importantes en la explicación de la variación a nivel del PC2, mientras que en el sPC2 la variable que recibió el peso más alto fue CE30 (Tabla 3.4). Estos cambios en los pesos con que cada variable es ponderada para construir los PC y las sPC, sumado a los diferentes MI de las variables originales (Tabla 3.1), permiten explicar los aumentos y disminuciones en los índices de autocorrelación de las componentes principales.

Tabla 3.3. Estadística descriptiva para las tres primeras componentes principales generadas con los análisis de componentes principales (PCA) y MULTISPATI-PCA aplicados a bases de datos multivariados de tres lotes agrícolas.

Lote	Variables	Varianza	Proporción	Proporción Acumulada	MI[a]
L1	PCA				
	PC1	1.18	0.29	0.29	0.37*
	PC2	1.05	0.26	0.55	0.28*
	PC3	0.97	0.24	0.79	0.28*
	MULTISPATI-PCA				
	sPC1	1.13	0.28	0.28	0.47*
	sPC2	0.83	0.21	0.49	0.34*
	sPC3	1.00	0.25	0.74	0.27*
L2	PCA				
	PC1	2.59	0.65	0.65	0.69*
	PC2	0.91	0.23	0.88	0.37*
	PC3	0.32	0.08	0.96	0.34*
	MULTISPATI-PCA				
	sPC1	2.56	0.64	0.64	0.71*
	sPC2	0.90	0.22	0.86	0.36*
	sPC3	0.34	0.8	0.94	0.31*
L3	PCA				
	PC1	1.34	0.34	0.34	0.05*
	PC2	1.15	0.29	0.63	0.17*
	PC3	0.96	0.24	0.87	0.38*
	MULTISPATI-PCA				
	sPC1	1.02	0.25	0.25	0.47*
	sPC2	1.21	0.30	0.55	0.12*
	sPC3	0.77	0.19	0.74	0.11*

[a] Índice de Moran, *Estadísticamente Significativo (α= 0.05).

Tabla 3.4. Autovectores (ponderaciones de variables) de los análisis de componentes principales (PCA) y MULTISPATI-PCA. Se subrayan los coeficientes más importantes.

Lote	Variables	CE30	CE90	E	Pe
L1	PCA				
	PC1	0.70	0.33	0.52	-0.37
	PC2	-0.09	-0.79	0.60	-0.03
	PC3	0.27	0.10	0.23	0.92
	MULTISPATI-PCA				
	sPC1	0.50	-0.07	0.85	-0.11
	sPC2	0.81	-0.29	-0.48	0.14
	sPC3	-0.09	-0.08	0.17	0.98
L2	PCA				
	PC1	0.55	0.56	0.57	-0.24
	PC2	-0.06	-0.16	-018	-0.97
	PC3	0.81	-0.55	-0.21	0.08
	MULTISPATI-PCA				
	sPC1	-0.56	-0.58	-0.58	0.10
	sPC2	0.17	-0.06	-0.28	-0.94
	sPC3	0.80	-0.48	-0.24	0.25
L3	PCA				
	PC1	0.75	0.51	0.04	-0.41
	PC2	-0.04	0.60	-0.47	0.65
	PC3	0.01	0.25	0.88	0.40
	MULTISPATI-PCA				
	sPC1	-0.04	0.13	-0.99	-0.01
	sPC2	-0.50	-0.80	-0.08	-0.31
	sPC3	0.86	-0.50	-0.10	-0.08

[a]CE30: conductividad eléctrica aparente a 30 cm, CE90: conductividad eléctrica aparente a 90 cm, E: elevación, Pe: profundidad de tosca.

En la Fig. 3.1 se representaron los valores de FPI y NCE y el número óptimo de clases sugerido para cada método de clasificación (KM-SV, KM-PC, KM-sPC), en el lote 1. El número óptimo de zonas se determina cuando FPI y NCE alcanzan el valor mínimo. La expectativa inicial de obtener un resultado coincidente para ambos índices no se cumplió. Por lo tanto, como recomienda Lark y Stafford, (1997), se seleccionó el menor número de clases que en este caso fue dos.

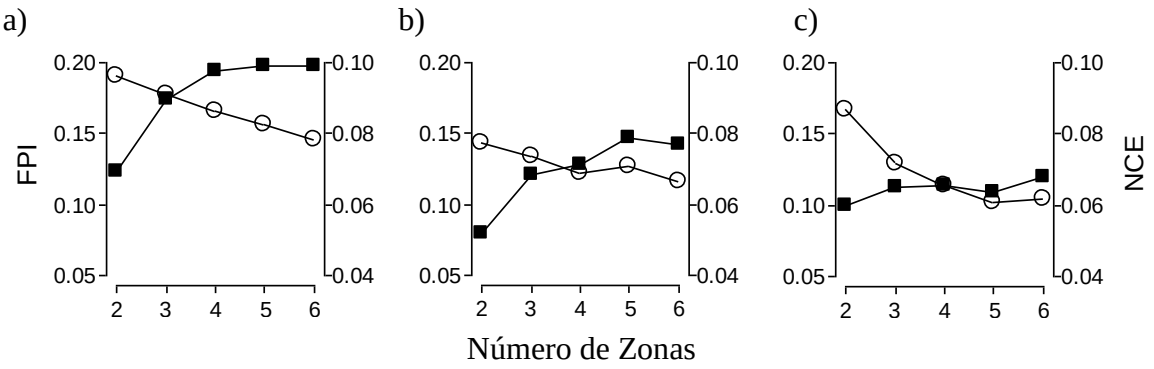


Fig. 3.1. Fuzziness Performance Index (FPI, círculos) y Normalized Classification Entropy (NCE, cuadrados) para tres métodos de clasificación: a) *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC).

En la Tabla 3.5 se presentan los resultados del ANAVA usado para comparar los rendimientos promedios entre las dos clases definidas por cada método. En L1 KM-sPC, fue el único método, entre los comparados, que delimitó clases de sitios con diferencias estadísticamente significativas de rendimiento tanto para soja como para trigo, es decir que el método generó clases por atributos de suelo que se correlacionaron con una productividad diferencial. KM-sPC permitió identificar *clusters* no sólo con mayores diferencias entre las medias ajustadas de rendimiento sino también con menor variabilidad residual dentro de cada zona y consecuentemente menor EE para la comparación de medias (Tabla 3.5). Las diferencias de rendimiento promedio entre las clases delimitadas fueron indicativas de condiciones del suelo que tienen una influencia diferencial sobre el rendimiento del cultivo.

Para L2, KM-sPC también tuvo un buen desempeño para la delimitación de clases. Las diferencias diferencias de rendimiento entre las clases delimitadas fueron mayores que con KM-PC. Mientras que en L3 los tres métodos mostraron diferencias estadísticamente significativas en el rendimiento para los dos cultivos. La alta variabilidad no sólo en CE, sino también en el Pe explica las diferencias entre las clases delimitadas. Sin embargo, KM-sPC generó clases con mayores diferencias de rendimiento y varianzas residuales más bajas para ambos cultivos. El mismo patrón que se observó a partir de las diferencias de medias entre las clases se obtuvo utilizando el coeficiente de *RV*. El método KM-sPC tuvo el mayor valor de *RV* en los tres lotes analizados (Tabla 3.5). Ambos indicadores calculados a partir de los valores de rendimiento validan el desempeño del método KM-

sPC para delimitar clases de sitios intralote. Las diferencias en rendimiento entre las clases delimitadas por cualquiera de los métodos son más pequeñas de lo esperado que cuando se usan métodos de comparación clásicos ya que han sido ajustados por las correlaciones espaciales. Sin embargo, estas diferencias fueron estadísticamente significativas aún después del ajuste de las correlaciones.

Tabla 3.5. Rendimientos promedios para dos clases de manejo delimitadas por los siguientes métodos de clasificación: *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), *cluster fuzzy k-means* sobre componentes principales (KM-PC) y *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC).

Lote	Método	Estimaciones	Rendimiento 2007		Rendimiento 2008	
			Zona I	Zona II	Zona I	Zona II
			t ha^{-1}			
L1	KM-SV	MA†	3.691a*	3.680a	1.792a	1.814a
		EE††	0.125	0.125	0.035	0.036
		RV‡	0.0004		0.0042	
	KM-PC	MA	3.672a	3.702a	1.781a	1.827a
		EE	0.125	0.125	0.035	0.036
		RV	0.0003		0.0099	
	KM-sPC	MA	3.563a	3.805b	1.675a	1.910b
		EE	0.111	0.111	0.031	0.030
		RV	0.1277		0.1741	
L2	KM-SV	MA	3.816a	4.102b	1.814a	1.941b
		EE	0.062	0.063	0.043	0.043
		RV	0.2048		0.2164	
	KM-PC	MA	3.992a	3.921a	1.867a	1.882b
		EE	0.073	0.074	0.053	0.052
		RV	0.0188		0.0074	
	KM-sPC	MA	3.773a	4.137b	1.791a	1.961b
		EE	0.060	0.059	0.040	0.040
		RV	0.2318		0.2634	
L3	KM-SV	MA	3.365a	3.478b	2.250a	2.363b
		EE	0.074	0.075	0.115	0.116
		RV	0.0066		0.0497	
	KM-PC	MA	3.366a	3.466b	2.256a	2.342b
		EE	0.076	0.076	0.115	0.116
		RV	0.0001		0.0395	
	KM-sPC	MA	3.364a	3.480b	2.248a	2.367b
		EE	0.073	0.075	0.113	0.114
		RV	0.0076		0.0566	

† Media Ajustada. †† Error Estándar. ‡ RV*j*, varianza entre clases relativa a la media general. *Letras diferentes indican diferencias estadísticamente significativas (P <0.05) entre clases respecto al rendimiento.

Las diferencias en las variables del suelo entre las clases delimitadas, muestran que la clasificación obtenida a partir del algoritmo KM-sPC fue la mayor, mostrando mayores diferencias en CE30 entre las clases y error estándar más bajos del rendimiento dentro de la clase (L1 y L3) (Tabla 3.6). Particularmente, en L1 el uso de KM-sPC delimitó clases con altas diferencias en CE30 (Fig. 3.2). La variable CE30 estuvo bien representada en las primeras componentes principales espaciales y tuvo mayor variabilidad que la elevación. Por ello se supone que CE lideró el proceso de clasificación. Debido a la correlación significativa entre CE30 y el rendimiento de los cultivos (Tabla 3.2), las diferencias en rendimiento entre las clases delimitadas se muestran mejor con el nuevo algoritmo propuesto.

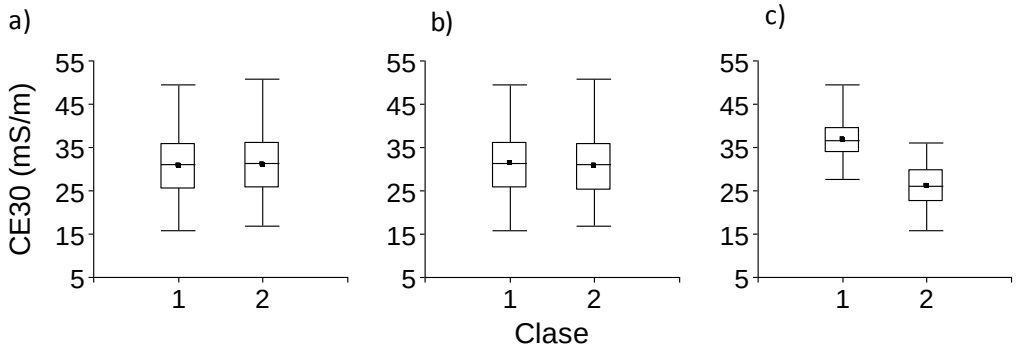


Fig. 3.2. Gráficos Box-plots de la distribución de CE30 dentro de clases delimitadas por tres métodos de clasificación: *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), (b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y (c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC).

En el L1, los mapas de KM-SV y KM-PC fueron similares pero en ambos casos los límites de la clasificación variaron de manera más abrupta que con KM-sPC. El nuevo algoritmo mostró bordes de forma más "suave" en las clases delimitadas (Fig. 3.3). En L2, se eliminaron algunos puntos pequeños ("manchas"), dentro de cada zona usando KM-sPC (Fig. 3.4), este tipo de suavizado logrado con el uso de las sPC es importante desde un punto de vista agronómico, ya que el algoritmo se aplica para delimitar zonas de manejo para la aplicación del manejo sitio-específico. Sin embargo, para este lote, las diferencias en el rendimiento entre las zonas no fueron altas, debido a la baja variabilidad de las variables del suelo (Tabla 3.1). En L3, las diferencias visuales en los mapas multivariados, obtenidos con los diferentes métodos, fueron pequeñas a causa de la alta diferencia en las variables del suelo originales (Fig. 3.5) entre las clases conformadas.

Tabla 3.6. Promedios de variables de suelo para dos clases de sitios delimitadas por los siguientes métodos de clasificación: *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), *cluster fuzzy k-means* sobre componentes principales (KM-PC) y *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC).

Lote	Método	Estimaciones	CE30[a]		CE90[b]		E[c]		Pe[d]	
			Zona I	Zona II	Zona I	Zona II	Zona I	Zona II	Zona I	Zona II
			mS m^{-1}				m		cm	
L1	KM-SV	MA[†]	30.95	31.28	30.37	29.53	141.93	141.56	57.46	93.96
		EE[††]	0.93	0.94	0.52	0.54	0.13	0.13	0.70	0.77
	KM-PC	MA	31.23	30.94	30.21	29.73	141.85	141.67	57.41	94.63
		EE	0.93	0.94	0.52	0.54	0.13	0.13	0.82	0.89
	KM-sPC	MA	36.64	26.34	29.10	30.80	142.04	141.53	70.06	76.46
		EE	0.33	0.32	0.53	0.51	0.13	0.13	1.52	1.46
L2	KM-SV	MA	33.94	26.74	31.17	27.44	141.94	139.99	76.25	75.74
		EE	0.67	0.68	0.88	0.88	0.77	0.77	0.91	0.92
	KM-PC	MA	27.76	33.38	30.71	28.79	141.13	141.20	76.50	75.56
		EE	1.24	1.24	1.41	1.41	1.07	1.07	0.87	0.85
	KM-sPC	MA	34.09	26.79	30.95	27.83	142.31	139.59	76.95	75.05
		EE	0.70	0.70	0.95	0.95	0.68	0.68	0.92	0.92
L3	KM-SV	MA	27.69	23.24	25.38	26.29	145.32	146.43	57.53	70.20
		EE	0.55	0.56	0.50	0.51	0.29	0.29	1.68	1.67
	KM-PC	MA	26.68	24.56	25.71	25.90	145.11	146.62	57.26	69.62
		EE	0.72	0.72	0.51	0.50	0.21	0.22	1.78	1.66
	KM-sPC	MA	27.80	23.05	25.34	26.35	145.29	146.44	59.49	68.45
		EE	0.54	0.55	0.50	0.52	0.29	0.30	1.69	1.72

[a]CE30: conductividad eléctrica aparente a 30 cm, [b]CE90: conductividad eléctrica aparente a 90 cm, [c]E: elevación, [d]Pe: profundidad de tosca.
[†] Media Ajustada. [††]Error Estándar.

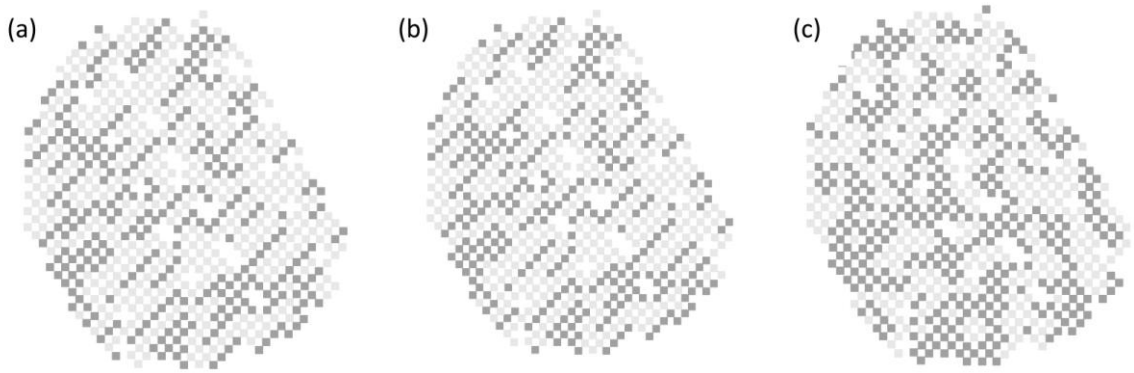


Fig. 3.3. Clase 1 (blanco) y clase 2 (negro) del lote 1 delimitadas por tres métodos de clasificación: (a) *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), (b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y (c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC).

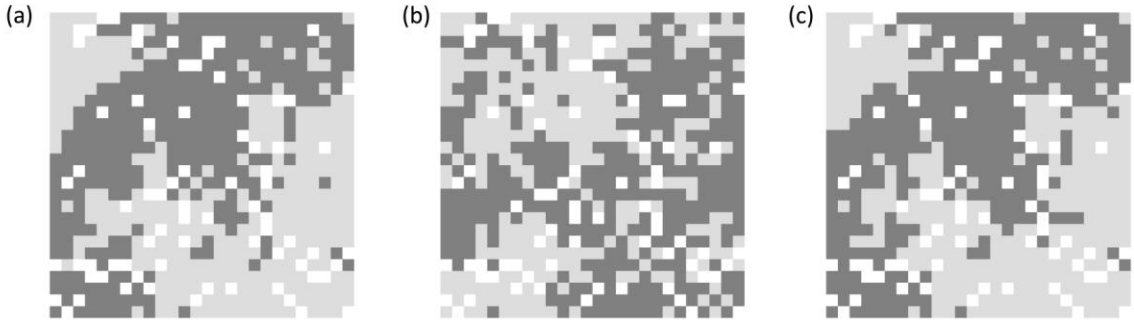


Fig. 3.4. Clase 1 (blanco) y clase 2 (negro) del lote 2 delimitadas por tres métodos de clasificación: (a) *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), (b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y (c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC).

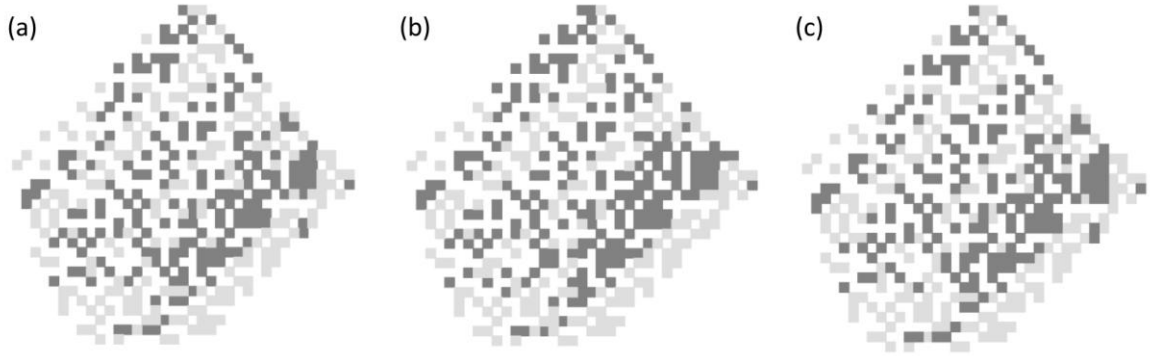


Fig. 3.5. Clase 1 (blanco) y clase 2 (negro) del lote 3 delimitadas por tres métodos de clasificación: (a) *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), (b) *cluster fuzzy k-means* sobre componentes principales (KM-PC) y (c) *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC).

En la Tabla 3.7 se presentan indicadores (obtenidos por simulación) del desempeño relativo de los tres procedimientos de delimitación de clases comparados. El método KM-sPC detectó diferencias estadísticamente significativas entre rendimientos de las clases delimitadas en un 81% de las base de datos simuladas, mientras que KM-PC y KM-SV en el 79 y 67%, respectivamente. Además KM-sPC en el 39 % de las simulaciones identificó clases de sitios con la mayor magnitud en las diferencias de medias de rendimiento. En el 43% de los análisis realizados sobre datos simulados, KM-sPC generó zonas con la menor variabilidad residual.

Tabla 3.7. Porcentaje de las simulaciones que identifican clases de sitios con diferencias estadísticamente significativas entre rendimiento para tres procedimientos de clasificación: *cluster fuzzy k-means* sobre variables de suelo originales (KM-SV), *cluster fuzzy k-means* sobre componentes principales (KM-PC) y *cluster fuzzy k-means* sobre componentes principales espaciales (KM-sPC).

Método	Diferencias en rendimiento (%)	> Magnitud de la diferencia (%)	< Variabilidad dentro de las clases (%)
	—————————————————**%**—————————————————		
KM-SV	67	27	21
KM-PC	79	34	36
KM-sPC	81	39	43

DISCUSIÓN

En AP, cuando se delimitan zonas de manejo a partir de variables de suelo es deseable que las zonas sean realmente diferentes porque serán sometidas a manejos diferenciales. Una forma de corroborar la magnitud de sus diferencias es a través de la evaluación de la significancia de la diferencia entre las medias de rendimiento. Las variables del suelo suelen variar en el espacio en forma gradual más que abrupta y por lo tanto, es difícil de definir límites espaciales de las áreas delimitadas sin ambigüedades (Guastaferro *et al*., 2010). La autocorrelación espacial positiva, habitual en este tipo de datos espaciales, puede ser explotada para mejorar el rendimiento de los métodos de clasificación difuso (Arno *et al*., 2011). En particular, el método *fuzzy k-means* asigna a cada punto un valor de pertenencia parcial a cada una de las *k* clases. Este método *fuzzy k-*

means, es frecuentemente utilizado para la delimitación de ZM en agricultura de precisión, y está disponible en varios software de acceso libre. En este capítulo se mostró que la combinación del análisis de componentes principales espacial con el algoritmo de clasificación *fuzzy k-means* resultó una estrategia para la delimitación de clases de sitios que potencialmente podrían convertirse en zonas de manejo diferencial dentro del lote

El método MULTISPATI-PCA (Dray *et al*., 2008) fue utilizado para capturar la autocorrelación espacial de las variables utilizadas como *input* del análisis de *cluster* difuso *k-means*. Las combinaciones obtenidas a partir de MULTISPATI-PCA generaron más variabilidad contigua respecto a la obtenida con el PCA no restringido espacialmente. En esta ilustración la autocorrelación espacial en variables de suelo y terreno resultó más fuerte que la reportada en los estudios ecológicos de Dray *et al*. (2008), donde ya se enfatiza la necesidad de considerar las correlaciones espaciales en la estimación de variabilidad. Usando MULTISPATI-PCA también es posible identificar las variables que mejor representan la variabilidad espacial global y aquellas de mayor autocorrelación espacial. Por el contrario, PCA puede fallar en la detección de estructuración espacial si ésta no está asociada con la variabilidad más pronunciada (Jombart *et al*., 2008), ya que PCA, la componente principal explica la varianza entre las observaciones más que autocorrelación. Este tema fue discutido en el contexto del análisis de datos ecológicos por Thioulouse *et al*. (1995), quien se basó en un trabajo de Wartenberg (1985) para evaluar la significancia estadística de estructuras espaciales multivariadas.

Para ejecutar MULTISPATI-PCA es necesario definir una red de vecindarios espaciales a partir de diferentes matrices de conectividad. Las opciones seleccionada dependen del esquema de muestreo (grilla regular o irregular) (Bivand, 2008). Otros enfoques en la delimitación de clases de manejo, al igual que el método propuesto por Oliver y Webster (1989), utilizan un variograma para identificar el vecino de cada sitio. Sin embargo, la matriz de conectividad tiene una relación más directa con el vecindario que el variograma.

La partición difusa en el espacio de las componente espaciales se obtuvo estableciendo un valor del coeficiente difuso de 1.3, como es habitualmente usado en aplicaciones de agricultura de precisión (Guastaferro *et al*., 2010; Arno *et al*., 2011). No hay evidencia teórica o computacional para distinguir un coeficiente difuso óptimo, sin

embargo, debido a la suavización que produce MULTISPATI-PCA, se supone que es mayor, o por lo menos igual, al coeficiente utilizado cuando se usan, como variables de la clasificación, las variables del suelo originales.

El uso de los índices FPI y NCE para la determinación del número de *cluster* óptimo sólo proporciona una medida estadística relacionada a la partición de los sitios, pero no tiene en cuenta si el resultado es agronómicamente aceptable. Sin bien no existen reglas fijas para la elección del tamaño y la forma de las zonas de manejo y la limitación suele estar dada por la habilidad del productor de poder manejarlas individualmente (dimensiones y capacidades del parque de maquinaria, características físicas del lote y patrón de trabajo de la maquinaria) (Roel y Terra, 2006), siempre interesan zonas cuyas diferencias se expresen en los rendimientos.

Cuando KM-sPC fue aplicado a conjuntos de datos experimentales, el algoritmo mostró resultados que son consistentes con otros trabajos respecto a la importancia de la CE y la elevación del suelo como variables de peso en la delimitación de áreas intralote (Fridgen *et al*., 2004; Taylor *et al.*, 2007; Peralta *et al*., 2011). La variabilidad en la CE y elevación puede ser explotada con los algoritmos de agrupamiento para generar clases con diferente potencial de rendimiento. En los conjuntos de datos del ejemplo, donde los lotes son relativamente planos y la variabilidad de la elevación fue más pequeña entre las variables del suelo, sólo el enfoque KM-sPC fue sensible a la variabilidad en la elevación. Esta variable del terreno, con autocorrelación alta y significativa, recibió altos pesos para la construcción de las componentes espaciales. El mapeo con KM-sPC mostró una zonificación más contigua. Para KM-sPC las diferencias en rendimiento entre las clases delimitadas fue la más alta, principalmente debido a las mayores diferencias en CE30 entre las clases respecto a las clase obtenidas con los otros dos métodos de clasificación utilizados en este estudio.

Como fuera mencionado en un trabajo anterior, este estudio muestra que la identificación de las zonas intralote y su tamaño está fuertemente relacionada no sólo a la existencia de variabilidad intralote (Li *et al*., 2007), sino también a los procedimientos de clasificación utilizados. El análisis de *cluster fuzzy k-means* (no restringido espacialmente) ha sido ampliamente utilizado tanto sobre variables originales como en componentes principales (Moral *et al*., 2010; Davatgar *et al*., 2012) pero no son los enfoques más

adecuados bajo autocorrelación espacial. La necesidad de tener en cuenta la estructura espacial de las variables incluidas para la subdivisión del lote ha sido reconocido por varios autores (Pringle *et al*., 2003; Tozer y Isbister, 2007; Pedroso *et al*., 2010). Sin embargo, la selección de los parámetros de suavizado de tal procedimiento puede afectar fuertemente el patrón de suavizado de las clases. En el algoritmo propuesto, la continuidad espacial se logró mediante un filtrado dado por los componentes principales espaciales.

CONCLUSIÓN

El análisis de *cluster fuzzy k-means* que incluye una dimensión espacial a través del uso del análisis de componentes principales espaciales (MULTISPATI-PCA) como variables de clasificación, mejoró el desempeño del análisis de *cluster* no restringido espacialmente en la clasificación sitios intralote. A su vez, KM-sPC aplicado sobre variables del suelo y terreno delimitó clases de manejo con las mayores diferencias en rendimiento respecto a los métodos de clasificación sin restriccion espacial. El procedimiento propuesto es adecuado para grandes conjuntos de datos multivariados y no requiere el ajuste previo de un modelo de variograma.

CAPÍTULO IV

PROTOCOLO DE ANÁLISIS PARA LA DELIMITACIÓN DE ZONAS DE MANEJO INTRALOTE

INTRODUCCIÓN

En el presente Capítulo se propone un protocolo para la implementación de técnicas estadísticas recomendadas como apropiadas para asignar mayor rigor científico al proceso de delimitación de zonas de manejo intralote. El protocolo puede ser implementado utilizando los script del software libre R (R Core Team, 2013) desarrollados en esta tesis. Para ilustrar su aplicación, la delimitación de zonas de manejo se realizó utilizando datos de rendimiento, conductividad eléctrica aparente, elevación y profundidad efectiva de un lote comercial cultivado bajo agricultura de precisión.

MATERIALES Y MÉTODOS

DATOS

Se trabajó con datos provenientes de un lote en producción (90 ha) bajo agricultura continua con rotaciones de cultivos anuales, ubicado en el sudeste bonaerense de la Argentina. Se dispone de datos de conductividad eléctrica aparente a los 30 y 90 cm de profundidad (CE30 y CE90) que fueron tomados utilizando un sensor (Veris 3100, Division of Geoprobe Systems, Salina, KS). El equipo fue configurado para tomar posición satelital cada segundo y montado a una camioneta pick-up para recorrer el lote en dirección a los surcos de siembra en transectas paralelas distanciadas entre 15 y 20 m. La velocidad

promedio de avance fue entre 15 y 20 km/h. simultáneamente se registraron datos de elevación con un DGPS (Trimble R3, Trimble Navegation Limited, USA) y se procesaron para obtener una precisión vertical de entre 3 y 5 cm aproximadamente. Las mediciones de la profundidad del suelo (horizonte petrocálcico o tosca) en cada sitio fueron realizadas con un muestreador manual acoplado a un GPS (Juno ST; Trimble Navigation Ltd,, EEUU). Este muestreo fue realizado en grilla de 30×30 m. Para cuantificar los datos de rendimiento de cultivo (trigo) se utilizó un monitor de rendimiento acoplado a una cosechadora.

Se tomaron 8 puntos de muestreo de suelos georreferenciadas dentro de cada zona zonas de manejo previamente delimitadas. Cada punto de muestreo consistió en tres submuestras, centradas en las zonas delimitadas para evitar muestrear sitios de transición. Las muestras de suelo fueron tomadas a una profundidad de 90 cm, utilizando un barreno de accionamiento hidráulico de 5 cm diámetro (Machine Co. Giddings, Windsor, CO). En cada sitio la capa de suelo (0-90 cm) fue mezclada para homogeneizar la muestra y, por tanto, sea representativa de la profundidad analizada. El contenido SOM sólo se midió en el estrato de 0-30 cm (Barbieri *et al*., 2009). Las muestras fueron recogidas en bolsas de plástico y en laboratorio fueron secadas en estufa a 60 ºC con circulación forzada de aire por un tiempo de 10 a 16 horas. Se molieron y tamizaron por una malla de 2 mm. Posteriormente, se determinó la distribución del tamaño de partículas por el método de Bouyoucos (Dewis y Freitas, 1970), el contenido de MO por el método de digestión húmeda de Walkley y Black (1934), el contenido de $N-NO_3^-$ fue determinado por método colorimétrico de ácido 2,4 fenoldisulfónico (Bremner, 1965).

PROTOCOLO

CONVERSIÓN DE COORDENADAS ESPACIALES

Un paso inicial en el análisis de datos espaciales es convertir las coordenadas geográficas en coordenadas cartesianas UTM (Universal Transverse Mercator). Esto permite que las distancias entre los sitios o puntos desde donde se leen los datos se

expresen como distancias absolutas (metros) en vez de distancias relativas (grados). La mayoría del software GIS (Geographic Information System) tiene la capacidad para realizar dicha transformación de coordenadas. En R, la librería "rgdal" (Bivand *et al*., 2013b) cuenta con la función *spTransform* que permite hacer transformación de sistemas de coordenadas.

ELIMINACIÓN DE OUTLIERS

Los *outliers*, o valores atípicos, son observaciones con valores que se encuentran fuera del patrón general del conjunto de datos. La eliminación de los *outliers*, previo al análisis, es fundamental para garantizar que las decisiones agronómicas tomadas sean las correctas. Los *outliers* se pueden eliminar fácilmente a través de un proceso de dos pasos.

Paso 1: El conjunto de datos primero debe limitarse dentro de un rango de variación razonable. Este paso, en general, se utiliza para depurar valores de rendimiento. Los valores máximos y mínimos pueden obtenerse desde la distribución de los datos y, consecuentemente, difieren dependiendo del tipo de cultivo y ubicación geográfica del lote. Los valores extremos también pueden obtenerse desde información histórica de los rendimientos del lote o desde el conocimiento del productor.

Paso 2: La teoría estadística establece que en una distribución normal, aproximadamente el 99% de los valores se encuentran dentro de 2,5 desvíos estándar de la media. En este paso se determinan para todo el conjunto de datos, previamente acotados en su rango, la media y la desviación estándar (DE). Luego se identifican los valores de rendimiento que se encuentran fuera de la media ± 2,5 DE. Estos son considerados como valores atípicos por defecto (Taylor *et al*., 2007). El diagrama de cajas (Box-plot), es un instrumento gráfico de la estadística descriptiva que permite una visualización más detallada y concisa de la distribución de los datos, indicando la presencia de *outliers* cuando estos existen. En ocasiones en que los datos de monitores de rendimiento son sesgados negativamente, como resultado de malas lecturas tales como las causadas por cosechar con cabezal a medio llenar o transitar con el cabezal hacia abajo sobre áreas cosechadas, puede justificarse una reducción del límite inferior (por ejemplo, media-1,5

DE) (Taylor *et al*., 2007). Finalmente, antes de la eliminación de los *outliers*, los mismos deben ser graficados utilizando las coordenadas para visualizarlos. Esto permite identificar si los datos seleccionados para ser eliminados representan un efecto real, por ejemplo sitios que pertenecen a una zona de bajo rendimiento dentro del lote, o por el contrario se relacionan a errores aleatorios de lectura.

ELIMINACIÓN DE INLIERS

La aplicación de los pasos 1 y 2 elimina los extremos del conjunto de datos, pero no se ocupa de los valores extremos locales (*inliers* espaciales). Los *inliers* son datos que difieren significativamente de su vecindario pero se sitúan dentro del rango general de variación del conjunto de datos. La identificación y remoción de los *inliers* es más dificultosa comparada con la de los *outliers*. Sin embargo, el proceso de eliminación de *outliers* de dos pasos descrito anteriormente ha demostrado ser eficaz también para la limpieza de los *inliers* (Taylor *et al*., 2007). Algunos software de AP como Yield Editor (Sudduth y Drummond, 2007) o el desarrollado por Sun *et al*. (2012) fueron diseñados para el diagnostico y limpieza de datos de rendimiento de cosechadoras y son eficaces en la eliminación tanto de *outliers* como de *inliers*. Llevar registros de diagnósticos de la cosecha puede ser de utilidad para identificar cuándo datos erróneos (*inliers* u *outliers*) son susceptibles de ser recolectado. Los datos erróneos pueden aparecer, por ejemplo, cuando la cosechadora está girando, cuando la velocidad es demasiado baja o cuando no se cosecha con el ancho del cabezal completo (Taylor *et al*., 2007).

Una herramienta estadística diseñada específicamente para identificar *inliers* es el índice autocorrelación espacial Local de Moran (I_i) (Anselin, 1995). Dado un grupo de datos que pertenecen a diferentes vecindarios, el I_i es básicamente un índice de Moran aplicado a cada vecindario individualmente y que da idea del grado de similitud o diferencia entre el valor de una observación respecto al valor de sus vecinos. La fórmula del índice de autocorrelación espacial Local de Moran es la siguiente:

$$I_i = \frac{X_i - \bar{X}}{\sigma^2} \sum_{j=1, j\neq 1}^{n} \left[w_{ij}(X_j - \bar{X}) \right] \qquad (4.1)$$

donde x_i es el valor de la variable X en la posición i; $\bar{x}$ y σ^2 es la media y varianza de X, respectivamente; x_j es el valor de la variable X en todos los otros sitios (donde $j \neq i$); w_{ij} es el peso espacial entre las ubicaciones i y j. El I_i se puede estandarizar, por lo que su nivel de significación puede ser evaluado en base a una distribución normal estándar. Los valores positivos de I_i se corresponden con agrupamiento espacial de valores similares (ya sean altos o bajos) (autocorrelación positiva), mientras que un valor de I_i negativo indica un agrupamiento de valores diferentes (por ejemplo, un sitio con valor bajo de la variable se encuentra rodeado de vecinos con valores altos) (autocorrelación negativa). En ambas instancias, el valor *p* para un índice determinado debe ser lo suficientemente pequeño para considerar el valor estadísticamente diferente.

Anselin (1996) propuso para visualizar el I_i un diagrama de dispersión que constituye una herramienta visual útil para el análisis exploratorio de datos espaciales ya que permite evaluar la similitud de un valor observado respecto a sus observaciones vecinas. El eje horizontal se basa en los valores de las observaciones mientras que en el eje vertical se representa el retardo espacial de la variable que se está analizando. Adicionalmente, se puede ajustar y añadir a este diagrama modelos de regresión lineal o no lineal. El diagrama de dispersión de Moran es una herramienta gráfica intuitiva para evaluar y representar el grado de autocorrelación espacial así como la presencia de valores atípicos (Anselin, 1996).

Las funciones *localmoran* y *moranplot* de la librería "spdep" (Bivand *et al*., 2013a) del software R, permiten calcular el I_i y realizar el gráfico de dispersión de Moran para identificar *inliers*. Aplicando la función *localmoran* se obtiene el I_i y su significancia estadística para cada observación. La función *moraplot* además de realizar el diagrama de dispersión ajusta un modelo de regresión lineal y calcula una serie de estadísticos de diagnostico. Los datos que se alejen de la recta de 45° sugieren sitios que presentan un valor de autocorrelación espacial que es diferente a la de su vecindario. Los criterios de diagnósticos son los siguientes:

Distancia de Cook: esta medida evalúa el cambio que se produce en la estimación del parámetro de regresión, cuando se elimina cada observación, es decir, evalúa la influencia de una observación sobre la estimación de los coeficientes de regresión. La estrategia que sigue es obtener la estimación de los parámetros del modelo con y sin esa

observación. Aquellas que presenten un gran impacto sobre el modelo ajustado se denominan observaciones influyentes.

Leverage: mide la distancia de un caso individual respecto a la media de la variable independiente. Un caso con alto leverage está lejos del centro o la mediana de la variable regresora(s). El valor promedio para Leverage es *p/N*, donde *p* es el número de predictores incluyendo a la constante y *N* número de casos. Casos con alto leverage son potencialmente casos influyentes.

DFFITS: este índice mide la influencia sobre la predicción, de la eliminación de cada observación y se calcula para todas las observaciones.

DFBETAS: es una medida normalizada del efecto de las observaciones en la estimación de los coeficientes de regresión. Existe un DFBETA para cada coeficiente de regresión, incluyendo la ordenada al origen.

COVRATIO: mide el efecto de cada una de las observaciones sobre la matriz de varianzas y covariancias de la estimación de los parámetros.

La función *moranplot* calcula estos índices para cada observación y considera a una observación como influyente si al menos uno de los índices de diagnóstico la detecta como tal.

INTERPOLACIÓN ESPACIAL

Después de la limpieza de los datos es necesario estimar los valores de la variable de interés en aquellos lugares donde no se ha muestreado la variable. Esto se puede realizar utilizando técnicas de interpolación espacial. La mayoría del software utilizado en AP como Farm Work (Trimble Navigation Limited), SMS (Ag Leader Technologies, Inc.) o SSToolBox (SST Development Group, Inc.), permiten realizar esta interpolación y elaborar bases de datos con información sobre cada variable en cada uno de los sitios intralote. Estos mapas son útiles para visualizar errores sistemáticos y también estocásticos (Whelan *et al.*, 2001). Cuando se cosecha un lote por más de un año, difícilmente el mismo sitio georreferenciado sea medido en los diferentes años. Las mediciones de suelo y rendimiento de los cultivos usualmente no se realizan en los mismos sitios o con la misma

resolución espacial. Esto hace que sea importante la etapa de interpolación ya que a través de ésta es posible combinar datos de diferentes fuentes (por ej. diferentes sensores o diferentes años).

Antes de realizar la interpolación en la base de datos, es necesario conformar una grilla artificial regular, dispuesta idealmente sobre el lote del la misma manera para todas las fuentes de información diferente. El espaciado de la grilla debe reflejar el nivel de detalle requerido, la capacidad de procesamiento de los datos y la capacidad de software estadístico para analizarlo. Una cuadricula de 5 × 5 m suele recomendarse para sistemas extensivos ya que se aproxima a la mitad de ancho de operación usual de las maquinarias utilizadas por los productores. Este tamaño de cuadrícula genera 400 puntos por ha. Sin embargo, la potencia de procesamiento y análisis puede ser problemática en lotes grandes debido al gran número de interpolaciones requeridas. Utilizando una grilla de 10 × 10 m se reducen los problemas mencionados manteniendo una resolución del mapa adecuada para la visualización de datos, el análisis y la aplicación de prácticas de manejo agronómicas en AP. La visualización de las grillas usadas para la recolección de los datos de cada variable también provee una nueva instancia para detectar posibles errores en las bases de datos, usualmente errores de tipeo.

La predicción espacial puede llevarse a cabo utilizando diferentes técnicas estadísticas de interpolación. La selección del método, dependerá del número y la densidad de los sitios con datos recolectados. La utilización de técnicas geoestadísticas como kriging ordinario en bloque (Webster y Oliver, 2007), suele ser la más recomendad en AP. Sin embargo, cuando la base de datos es muy grande y la densidad de datos es alta otros métodos no geoestadísiticos como el de interpolación por el método del vecino más cercano (Sibson, 1981) o el inverso de la distancia (Franke, 1982), pueden utilizarse. En general, los resultados de la interpolación son de buena calidad en AP debido a la gran cantidad de información inicial, razón por la cual la selección de uno u otro método no sería crucial.

Existe un gran número de software que permiten hacer el grillado y la predicción espacial intralote. Algunos de ellos son Surfer (Golden Software, Inc.), Vesper (Minasny, *et al*., 2005), SGeMS (Remy *et al*., 2009), software GIS como ArcGis (ESRI, Inc.) e Idrisi (Eastman, 2009). En el software R las librerías “geoR” (Ribeiro Jr. y Diggle, 2001) y

"gstat" (Pebesma, 2004) son las más utilizadas para análisis geoestadísticos incluyendo interpolaciones.

CLASIFICACIÓN DE SITIOS

Luego de realizar la predicción espacial se recomienda mapear los datos interpolados utilizando software que permita la representación geográfica de los datos. El mapeo de los valores predichos permite identificar si existen patrones inusuales ("ruido") remanentes, errores en la interpolación, o efectos de manejo y/o ambientales inusuales. Si existen errores de interpolación o "ruidos" remanentes, los datos deben ser nuevamente depurados y re-interpolados.

Para delimitar clases de manejo, desde una perspectiva multivariada, las diferentes capas de información interpoladas (capas de distintas variables) deben ser colocadas en un solo archivo. Para ello los archivos individuales deben concatenarse horizontalmente utilizando las columnas correspondientes a las coordenadas X e Y como criterios de concatenación. Esto asegura que los datos de diferentes fuentes de información queden correctamente unidos en el espacio geográfico.

Cualquier factor de clasificación asociado a efectos de manejo (por ejemplo, parte del lote con doble cultivo, o con un manejo diferencial intencional como pulverización o fertilización) o efectos ambientales (por ejemplo, heladas, insectos, animales, o daño por enfermedad) que es reconocido a priori del análisis, puede ser descontado del análisis de variabilidad espacial a través de modelación estadística.

Una vez que se concatena y depura la información de distintas variables, se puede realizar un análisis de *cluster* del tipo *fuzzy k-means*. En esta tesis, se propone aplicar el análisis de *cluster* sobre componentes principales espaciales de las variables de sitio (algoritmo propuesto en el Capítulo 3). Los software FuzME (Minasny y McBratney, 2002) o MZA (Fridge *et al*., 2004) fueron especialmente diseñados para realizar el análisis de *cluster* utilizando el algoritmo *fuzzy k-means*. Mientras que en el software R este análisis se implementa con las librerías "cluster" (Maechler *et al*., 2013) y "e1071" (Meyer *et al*., 2013).

Para realizar el análisis de *cluster fuzzy k-means* con la librería “e1071” se utiliza la función *cmeans*. Algunas de sus opciones de configuraciones son las siguientes: número máximo de iteraciones, número de *cluster* y exponente difuso. Odeh *et al*. (1992) sugiere utilizar un valor del exponente de 1.30. Otra opción de configuración es la selección de la distancia de similitud incluida en la función de optimización. La distancia Euclídea (opción por defecto) utiliza, para el cálculo de la distancia entre cada observación y el centroide, el cuadrado medio error, mientras que la distancia Manhattan utiliza el error medio absoluto. Los resultados de la clasificación no supervisada, debieran someterse a control, utilizando conocimiento local de un "experto", para garantizar que el agrupamiento sea agronómicamente válido. Otros métodos de clasificación podrían utilizarse en este paso, por ejemplo *fuzzy k-meas* sobre las variables de sitio originales.

DELIMITACIÓN DE ZONAS DE MANEJO

Un aspecto importante de la clasificación de sitios es determinar el número óptimo de *cluster* para describir suficientemente bien la variabilidad espacial multivariada. Utilizando el conocimiento "experto" del productor o asesor suele hacerse una determinación subjetiva del número de clases. Además, diferentes índices estadísticos como el coeficiente de partición y entropía de la clasificación (conocidos también como fuzziness performance index-FPI y normalized classification entropy-NCE) (Bezdek, 1981) son frecuentemente utilizados en estudios de agricultura de precisión para determinar el número óptimo de clases de manejo. Otros índices tales como Xie-Beni (Xie y Beni, 1991), Fukuyama-Sugeno (Fukuyama y Sugeno, 1989), exponente de proporción (Windham, 1981) también podrían ser utilizados en AP. Si bien la utilización de los índices aporta información sobre cuál podría ser la mejor clasificación de los datos, frecuentemente estos no coinciden en la recomendación sobre el número óptimo de clases. Este hecho genera un inconveniente para la toma de decisiones. Para disminuir este problema, se pueden modificar los resultados de los índices y obtener un único índice que resuma los anteriores (Galarza *et al*., 2013).

La determinación de diferencias significativas entre las clases es otra forma para la selección del número de clases. Sin embargo, presenta el inconveniente de que la

autocorrelación entre puntos de datos invalida la utilización de análisis clásicos como ANAVA, ya que el supuesto de independencia de los residuos del modelo es violado. Sin embargo, algunos enfoques estadísticos se han propuesto para proporcionar apoyo a las decisiones (Fridgen *et al*., 2000; Taylor *et al*., 2007). Un enfoque más moderno es la utilización de Modelos Lineales Mixtos (MLM) los cuales permiten evaluar las clases contemplando la correlación espacial entre los sitios del lote. Sin embargo, cuando la base de datos es grande estos métodos pueden tener problemas de convergencia debido a las dimensiones de la matriz utilizadas en las estimaciones. Su utilización se espera que sea de mayor utilidad en etapas posteriores de la aplicación del protocolo.

No se debe dejar la decisión del número de clases como una decisión puramente estadística. Los índices aportan objetivamente una idea clara acerca de cúal podrá ser la clasificación óptima, aunque la selección final de la cantidad de grupos debe seguir una relación de compromiso entre lo sugerido por los índices y lo realmente practicable por la maquinaria de dosificación variable, la que no tiene posibilidad de alcanzar cambios instantáneos. Por lo tanto, cuando se selecciona el número óptimo de clases de manejo debe considerarse la capacidad del productor para manejar las clases. Lotes con grandes zonas con límites coherentes son más fáciles de manejar diferencialmente que lotes con numerosas zonas pequeñas y de forma irregular. Obtener zonas espacialmente estructuradas y con diferencias agronómicas significativas entre clases es uno de los requisitos para aplicar manejos diferenciales. En la práctica suelen utilizarse de dos a cuatro clases de manejo por lote (Taylor *et al*., 2007).

Determinado el número de clases de manejo en que será dividido el lote es necesaria la delimitación, precisa y factible, de las mismas. Para formar clases más contiguas que las producidas por cualquier algoritmo de *cluster* y reducir la fragmentación de las clases a los fines de delimitar zonas de manejo, se recomienda aplicar filtros espaciales sobre la clasificación resultante del método de clasificación (Ping y Dobermann 2003; Lark, 1998; Galarza *et al*., 2013). Para aplicar los filtros espaciales se trabaja con la base de datos resultante de la clasificación como un archivo de imagen. Se aplican filtros estadísticos que son filtros espaciales no lineales cuya respuesta se basa en el ordenamiento (ranking) de los pixeles contenidos en una porción de la imagen (máscara). Las máscaras utilizadas pueden tener diferentes tamaño 3×3, 5×5 ó 7×7 pixeles y, a continuación, se sustituye el valor del pixel central con el valor que resulte del ordenamiento (Gonzalez y

Woods, 2008). El más conocido de estos filtros es el filtro de mediana, el cual reemplaza el valor del píxel central por la mediana de los valores del vecindario de ese píxel (el valor original del píxel es incluido en el cálculo de la mediana). La librería "raster" (Hijmans, 2013) del software R permite aplicar el filtro de la mediana y otros como el de la media, moda, mínimo o máximo (Gonzalez y Woods, 2008).

VALIDACIÓN DE LAS ZONAS DE MANEJO

Una vez creadas las zonas de manejo, se requieren realizar algunas pruebas adicionales de verificación. Los datos de los sensores más usados en la actualidad (de rendimiento o datos de suelo) no suelen dar mediciones directas de los factores determinantes del rendimiento, como deficiencias de nutrientes, toxicidades, pH del suelo, u otras propiedades del suelo. Por ello, se recomienda un muestreo de suelo para valuar las zonas. El mismo se lleva a cabo utilizando un muestreo aleatorio estratificado usando las zonas de manejo potenciales como los estratos. Se imponen restricciones sobre la asignación aleatoria de los puntos de muestreo para evitar muestrear los límites de las zonas. Este proceso tiene como objetivo garantizar que las zonas de transición no se tomarán muestras y que los factores determinantes del rendimiento en cada zona pueden ser examinadas por separado. Para esto último, se recomienda tomar no menos de tres muestras por cada zona de manejo. Si es económicamente factible, se pueden tomar más muestras para proporcionar resultados más robustos. El procedimiento de muestreo debe contemplar la toma de muestras del suelo superficial y del subsuelo. La elección de las propiedades del suelo que se medirán debe basarse en el conocimiento local sobre cuáles son las posibles propiedades de suelo que podrían afectar el rendimiento y que puede diferir entre zonas del lote. El proceso de estratificación de la variación intralote, usando el proceso de delimitación de zonas de manejo, ayuda a reducir al mínimo el número de muestras necesarias para caracterizar las tendencias del suelo y los factores determinantes del rendimiento dentro del lote y entre las zonas (Taylor *et al*., 2007).

Para determinar las diferencias de las propiedades de suelo analizadas entre las zonas de manejo, se pueden utilizar modelos lineales mixtos que permiten contemplar la correlación espacial que pueden presentar las observaciones (Gbur *et al*., 2012). La

conjunción del conocimiento de los datos, la experiencia y el conocimiento del productor, permitirá determinar los regímenes de manejo específicos de la zona, y/o determinar los niveles de tratamiento para los experimentos basados en zonas de manejo.

RESULTADOS

ILUSTRACIÓN DE LA APLICACIÓN DEL PROTOCOLO

En esta sección se muestra cómo aplicar el protocolo propuesto utilizando el software R. Los comandos se escriben en azul y las salidas en rojo. Previo a la realización de los análisis, las siguientes librerías deben ser instaladas y cargadas:

```
install.packages ("spdep","rgdal","geoR","gstat","ade4","e1071","raster")
library(spdep)
library(rgdal)
library(geoR)
library(gstat)
library(ade4)
library(e1071)
library(raster)
library(nlme)
```

En la primer parte de la ilustración del protocolo (pasos 1 a 4) se utilizarán datos CE30. Estos pasos deberían repetirse para todas las variables medidas que se utilizarán para la delimitación de zonas de manejo. La base de datos inicial requiere de al menos tres columnas, las primeras dos identifican las coordenadas espaciales bidimensionales (x e y) y la tercera corresponde a la variable medida (por ejemplo CE30). Para la carga de los datos se requiere especificar la ubicación del archivo, en este ejemplo la base se denominará “datos”.

```
datos<-read.table("C:\\Users\\....\\datos.txt", header = TRUE)
```

PASO 1-CONVERSIÓN DE COORDENADAS ESPACIALES

Previo a la conversión de las coordenadas espaciales se requiere determinar cuáles son las columnas que identifican las coordenadas del archivo y cuál es su sistema de

referencia. En este ejemplo, la base “datos” posee coordenadas geográficas. Se muestran las primeras 10 filas.

```
coordinates(datos) <- ~x+y
proj4string(datos) <- CRS("+proj=longlat + datum=WGS84")
```

```
           x         y    CE30
1    -59.13236 -37.91546 27.8
2    -59.13241 -37.91550 26.1
3    -59.13246 -37.91554 22.4
4    -59.13251 -37.91558 20.0
5    -59.13256 -37.91562 23.6
6    -59.13261 -37.91566 29.0
7    -59.13265 -37.91570 26.6
8    -59.13270 -37.91574 25.2
9    -59.13275 -37.91578 23.2
10   -59.13279 -37.91582 21.9
```

Asignado su sistema de referencia se realiza la conversión de las coordenadas geográficas a UTM (Universal Transverse Mercator). Requiere especificar la faja o zona (que para este ejemplo corresponde a la zona 21, sur) y el elipsoide (WGS84).

```
datos <- spTransform(datos, CRS("+proj=utm +zone=21 +south +ellps=WGS84
+datum=WGS84 "))
```

Luego se extraen los datos con las coordenadas transformadas y se ordenan las columnas ya que la transformación alteró el orden de las mismas.

```
datos <- as.data.frame(datos)
datos <- datos[,c(2,3,1)]

datos
```

```
          x       y    CE30
1    312558.9 5801421 27.8
2    312554.9 5801416 26.1
3    312550.7 5801412 22.4
4    312546.5 5801407 20.0
5    312542.2 5801402 23.6
6    312538.0 5801398 29.0
7    312533.8 5801393 26.6
8    312529.8 5801389 25.2
9    312525.8 5801384 23.2
10   312521.9 5801380 21.9
```

PASO 2-ELIMINACIÓN DE OUTLIERS

Previo a la depuración se realiza un análisis exploratorio que incluye medidas resumen y gráficos (histograma y box-plot) (Fig. 4.1).

```
summary(datos$CE30)

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.80   18.40   22.60   23.84   28.10   47.90

par(mfrow=c(1,2))
hist(datos$CE30,col='green',nclass=20,main="Histograma",ylab='Frecuencia
Relativa',xlab='CE30 (mS/m)')
boxplot(datos$CE30,col='green',ylab='CE30 (mS/m)',main="Box-Plot")
```

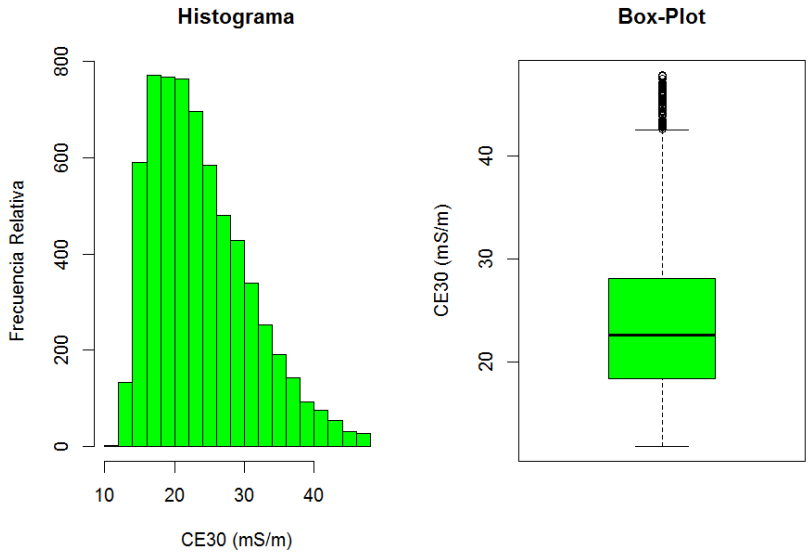


Fig. 4.1. Histograma y Box-plot de datos de conductividad eléctrica aparente a 30 cm de profundidad previo a la eliminación de *outliers*.

El gráfico box-plot muestra la presencia de algunos valores atípicos que se ubican muy por encima de la media de los datos. Esto genera asimetría en la distribución de los datos. Para eliminar los *outliers* se necesita primero calcular la media y la desviación estándar (DE). Luego se calculan dos límites, inferior (LI) y superior (LS), que posteriormente serán utilizados para eliminar los datos que se ubique por fuera de estos límites.

```
Media <- mean(datos$CE30)
DE <- sd(datos$CE30)
LI <- Media-2.5*DE; LI
[1] 6.303319

LS <- Media+2.5*DE; LS
[1] 41.38409
```

En esta sentencia se seleccionan los datos que se encuentran entre la media ± 2,5 DE y se vuelve a repetir el análisis exploratorio utilizando la base de datos sin los *outliers* (Fig. 4.2). Para el caso de estudio, en la etapa de la depuración se eliminaron 139 casos lo que representa un 2% del total de sitios (n=6425) con mediciones.

```
datos$CE30[LS<datos$CE30|datos$CE30<LI] <-NA
datos <- subset(na.omit(datos),select=c(x,y,CE30))

summary(datos$CE30)

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
11.8    18.3    22.4    23.4    27.7    41.3

par(mfrow=c(1,2))
hist(datos$CE30,col='green',nclass=20,main="Histograma",ylab='Frecuencia
Relativa',xlab='CE30 (mS/m)')

boxplot(datos$CE30,col='green',ylab='CE30 (mS/m)',main="Box-Plot")
par(mfrow=c(1,1))
```

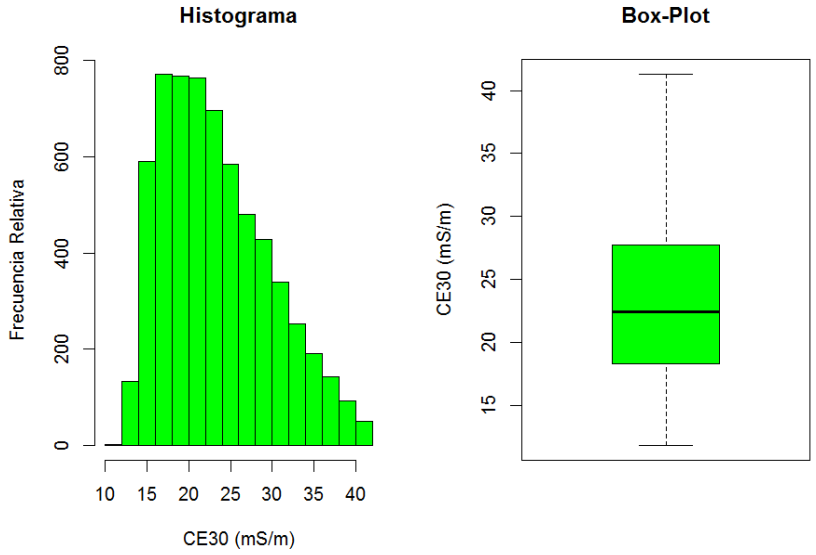


Fig. 4.2. Histograma y Box-plot de datos CE30 luego de la eliminación de *outliers*.

PASO 3-ELIMINACIÓN DE INLIERS

El primer paso para el análisis de autorrelación espacial local es la definición de una matriz de ponderación espacial. Esta puede ser representada en forma gráfica (como gráficos de vecindarios), donde los nodos corresponden a los sitios del lote y los bordes a pesos espaciales no nulos. Existen diferentes opciones o alternativas metodológicas para

definir los vecindarios. En este ejemplo la red de vecindarios fue definida en función de la distancia Euclídea considerando puntos vecinos a aquellos contiguos ubicados entre los 0 a 20 m de distancia.

```
cord <- coordinates(datos[,1:2])
gri <- dnearneigh(cord,0,20)
lw2 <- nb2listw(gri, style = "W")
```

Definida la matriz de ponderación especial se procede a realizar el gráfico de Moran para identificar los *inliers*. La Fig. 4.3 muestra el ajuste de una regresión lineal entre el valor de autocorrelación de cada sitio (CE30) y la autocorrelación de sus respectivos vecindarios (Spatially Lagged CE30). Aquellos datos que se alejen de la recta de 45° son sitios que presentan un valor de la variable no correlacionada con la de su vecindario.

```
MP<-moran.plot(datos$CE30,lw2,quiet=T,labels=F,zero.policy=F,xlab="CE30",
ylab="CE30 Spatially Lagged")
```

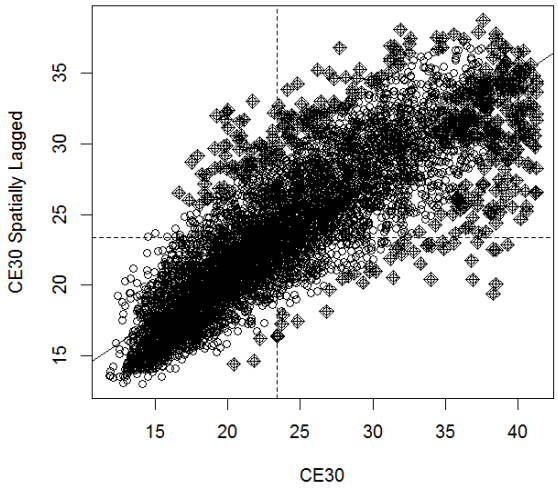


Fig. 4.3. Gráfico de dispersión de Moran de la variable CE30.

Para extraer los puntos influyentes de la regresión, estos son identificados mediante diferentes estadísticos de diagnóstico como leverage (hat), distancia de Cook (cook.d), Covratio (cov.r), DFFITS (dffit) y DFBETAS (dfb.1_ para la ordenada al origen y dfb.x para la pendiente). Un punto se determina como influyente si al menos uno de los criterios así lo considera. Se muestran los primeros 10 datos considerados como influyentes.

```
summary(MP)
```

```
Potentially influential observations of
       lm(formula = wx ~ x) :

     dfb.1_ dfb.x dffit   cov.r   cook.d hat
41     0.00   0.00  0.00    1.00_*  0.00   0.00
91    -0.01   0.01  0.01    1.00_*  0.00   0.00
92    -0.01   0.02  0.02    1.00_*  0.00   0.00
128    0.01   0.00  0.03    1.00_*  0.00   0.00
142    0.04  -0.03  0.05    1.00_*  0.00   0.00
148    0.01   0.00  0.03    1.00_*  0.00   0.00
166    0.08  -0.09 -0.10_*  1.00_*  0.01   0.00_*
170   -0.01   0.02  0.04    1.00_*  0.00   0.00
174    0.00   0.01  0.01    1.00_*  0.00   0.00
176    0.00   0.00  0.00    1.00_*  0.00   0.00_*
```

Otra forma de identificar los *inliers* es calculando el Índice de Moran Local (*Ii*) y su significancia estadística (ajustando los valores-p por el criterio de Bonferroni). Sitios con valores *Ii* negativos y significativamente distintos de cero, son considerados *inliers*.

```
ML <- localmoran(datos$CE30,lw2,p.adjust.method="bonferroni",alternative
="less")
```

En las sentencias siguientes, se unen a la base de datos, previamente depurada por los outliers, los valores de *Ii* observado (Ii), su valor esperado (E.Ii), la varianza (Var.Ii) y el valor-p (Pr(z < 0)). Además se adiciona la tabla con los datos influyentes del gráfico de Moran donde para cada criterio de diagnóstico se determina si el dato es influyente (FALSE) o no (TRUE).

```
IMl<-printCoefmat(data.frame(ML,row.names=datos$Casos),check.names=FALSE)
IMl
```

```
           Ii            E.Ii       Var.Ii         Z.Ii      Pr(z < 0)
1   -0.0605676695 -0.000159109 0.3331406 -0.1046609537          1
2    0.0034078404 -0.000159109 0.2498157  0.0071365297          1
3   -0.0462422071 -0.000159109 0.1998208 -0.1030911467          1
4   -0.2084667032 -0.000159109 0.1664908 -0.5105167539          1
5    0.0072995205 -0.000159109 0.1664908  0.0182794839          1
6    0.0138882784 -0.000159109 0.1664908  0.0344271011          1
7    0.0326351798 -0.000159109 0.1664908  0.0803716923          1
8    0.0190967155 -0.000159109 0.1664908  0.0471918512          1
9   -0.0040322027 -0.000159109 0.1664908 -0.0094921131          1
10  -0.0414254607 -0.000159109 0.1664908 -0.1011348819          1
```

```
Influ=MP$is.inf ; Influ
```

```
      dfb.1_ dfb.x dffit cov.r cook.d  hat
1     FALSE FALSE FALSE FALSE  FALSE FALSE
2     FALSE FALSE FALSE FALSE  FALSE FALSE
3     FALSE FALSE FALSE FALSE  FALSE FALSE
4     FALSE FALSE FALSE FALSE  FALSE FALSE
5     FALSE FALSE FALSE FALSE  FALSE FALSE
6     FALSE FALSE FALSE FALSE  FALSE FALSE
7     FALSE FALSE FALSE FALSE  FALSE FALSE
8     FALSE FALSE FALSE FALSE  FALSE FALSE
9     FALSE FALSE FALSE FALSE  FALSE FALSE
10    FALSE FALSE FALSE FALSE  FALSE FALSE
```

```
datos <- data.frame(datos,Iml,Influ) ; datos
```

Luego de identificar los *inliers* se procede a eliminarlos. Primero se eliminan los *inliers* detectados con el gráfico de Moran. Se espera que en este paso se eliminen la mayor cantidad de datos *inliers*.

```
datos1<- datos[datos$dfb.1_ == FALSE & datos$dfb.x == FALSE & datos$dffit
== FALSE & datos$cov.r == FALSE & datos$cook.d == FALSE & datos$hat ==
FALSE,]
```

Luego, se eliminan los datos con *Ii* negativos y estadísticamente significativos ($p<0.05$).

```
datos2 <- as.matrix(datos1)
datos2 <- subset(datos1,datos1[,4] > 0 | datos1[,8]>0.05 )
```

La nueva base de datos (“datos2”) cuenta con 5836 casos, es decir, se eliminaron 450 casos (7% de los datos) respecto a la base sin outliers.

PASO 4-INTERPOLACIÓN ESPACIAL DE LOS DATOS

El siguiente paso es establecer una grilla regular y realizar la interpolación espacial. La misma grilla será utilizada para cada variable. De esta forma todas las variables tendrán las mismas ubicaciones espaciales para los sitios donde se realizará la predicción. Para llevar a cabo este paso se implementan técnicas de análisis geoestadístico utilizando las librerías “gstat” y “geoR”. Primero, es necesario realizar el ajuste del semivariograma empírico y teórico (Fig. 4.4 y 4.5).

```
coordinates(datos2) <- ~x+y
CE30vario <- variogram(CE30~1, datos2, cutoff=380)

plot(CE30vario,main="CE30",xlab="Distancia",ylab="Semivarianza")
```

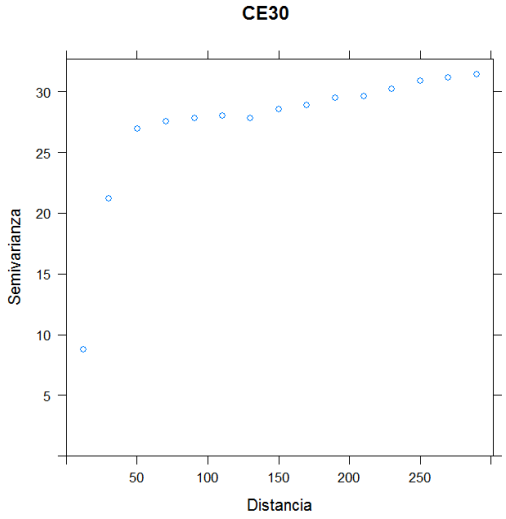


Fig. 4.4. Semivariograma empírico de la variable CE30.

El modelo elegido para representar la función de correlación espacial en el caso de estudio fue el esférico. La máxima semivarianza encontrada entre pares de puntos (21.32 $(mS/m)^2$) se conoce como sill (psill) y representa la varianza de los datos bajo independencia. El rango (range) es la distancia (88.67 m) a la que la semivarianza deja de aumentar, indicando la interdistancia a partir de la cual un par de datos se podría considerar como no correlacionados espacialmente. El nugget (Nug) es la varianza no explicada por las interdistancias entre sitios (9.34 $(mS/m)^2$); este representa la varianza residual observada en cada sitio o en entornos menores a los de la mínima distancia de muestreo y se calcula como el intercepto de la función del semivariograma.

```
CE30fitvariog <- fit.variogram(fit.method=1,CE30vario, vgm(25, "Sph",
80,10))

CE30fitvariog
```

```
model    psill    range
1   Nug  9.33839  0.00000
2   Sph 21.31372 88.67145
```

```
plot(CE30vario,CE30fitvariog,xlab="Distancia",ylab="Semivarianza")
```

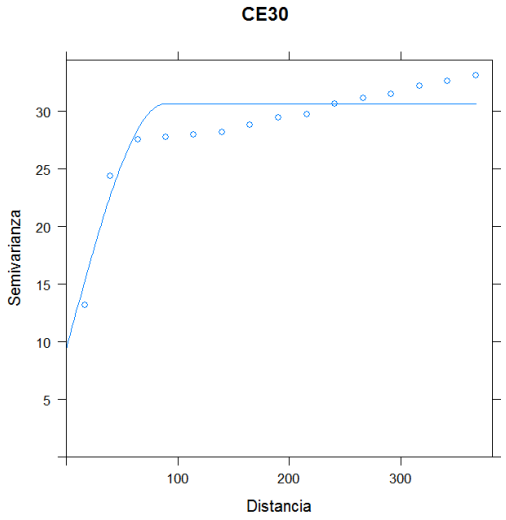


Fig. 4.5. Semivariograma empírico (puntos) y teórico (línea) de la variable CE30.

Los siguientes comandos realizan la interpolación utilizando kriging ordinario en bloque. Para ello se necesita cargar una base de datos con los puntos georreferenciados que conforman el polígono del lote, es decir, representan los límites del lote. En este ejemplo el archivo se denomina “bordes”.

```
bordes<-read.table("C:\\Users\\...\\bordes.txt", header = TRUE)
```

Luego se confecciona una grilla regular sobre el polígono (Fig. 4.6) para poder realizar las interpolaciones dentro de los límites del lote (Fig. 4.7). La dimensión de la misma es de 10 × 10 m.

```
gr<-pred_grid(bordes, by=10)
gri <- polygrid(gr,bor=bordes)
plot(gri,col = "red", pch = 10, cex = 0.2,xlab="X",ylab="Y")
gridded(gri) = ~Var1+Var2
```

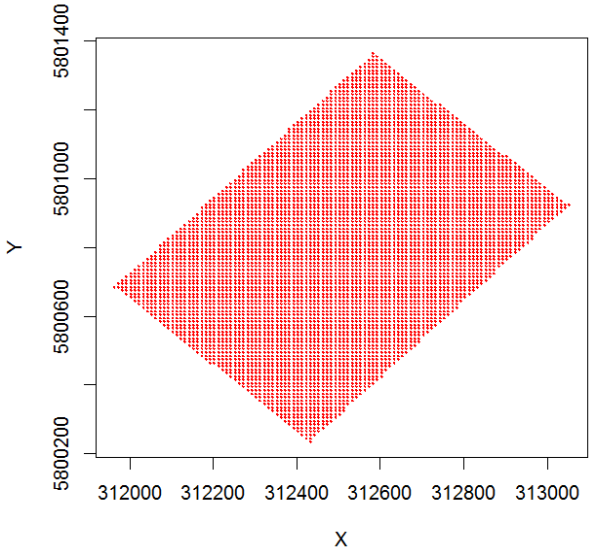


Fig. 4.6. Grilla de predicción de 10 × 10 m.

```
CEKg <- krige(CE30~1, datos2, gri, model = CE30fitvariog, block =
c(40,40))

spplot(CEKg["var1.pred"], main = "Mapa de variabilidad
espacial",col.regions=terrain.colors(100))
```

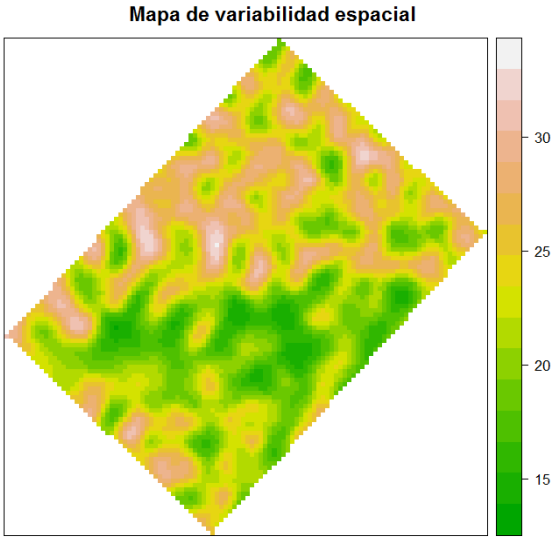


Fig. 4.7. Mapa de interpolación espacial de la variable CE30.

Realizada la interpolación se procede a extraer los datos predichos para cada uno de los nodos de la grilla de predicción. Debido a que se produce un cambio en el nombre de las columnas del archivo generado los mismos son cambiados utilizando la función *names*.

```
PredCE30 <- as.data.frame(CEKg)
PredCE30 <- PredCE30[,1:3]
names(PredCE30)[1]<-paste("x")
names(PredCE30)[2]<-paste("y")
names(PredCE30)[3]<-paste("CE30")
```

En el mapa de variabilidad espacial se observan dos zonas definidas por valores altos (parte superior) y bajos (parte inferior) de CE30 (Fig. 4.7).

PASO 5-DELIMITACIÓN DE CLASES DE SITIOS

Se necesita que todas las variables hayan sido procesadas desde el paso inicial hasta la interpolación con la misma grilla de predicción. Luego se realiza la concatenación de las diferentes bases de datos obtenidas. Con los valores predichos de CE30 (PredCE30), CE90 (PredCE90), Pe (PredPe), elevación (PredElev) y rendimiento de trigo (PredRtoTg) se utilizó la función *cbind* para la concatenación. Para PredCE30, se extrajeron las tres primeras columnas correspondientes a las coordenadas y valores predichos, mientras que para las restantes sólo se extrajeron los valores predichos de cada variable (columna 3).

```
Pred <- cbind(PredCE30[,1:3], PredCE90[,3], PredElev[,3],PredPe[,3],
PredRtoTg[,3]);Pred
```

```
           x       y     CE30 PredCE90.3 PredElev.3  PredPe.3 PredRtoTg.3
1   312432.8 5800234 26.02600   28.50741   160.4142 -78.07533    3.734030
2   312422.8 5800244 26.61192   28.14623   160.4164 -77.21150    3.731334
3   312432.8 5800244 24.97033   27.95664   160.4286 -78.65952    3.725866
4   312412.8 5800254 27.24498   27.56873   160.4184 -75.93873    3.727748
5   312422.8 5800254 26.03662   27.48603   160.4275 -76.89577    3.715309
6   312432.8 5800254 24.61090   27.26781   160.4379 -78.04122    3.700570
7   312442.8 5800254 23.34762   26.89906   160.4490 -78.70945    3.689491
8   312402.8 5800264 27.12211   26.76693   160.4079 -75.61594    3.727860
9   312412.8 5800264 26.70546   26.89592   160.4244 -75.50337    3.712381
10  312422.8 5800264 25.64192   26.75352   160.4340 -76.31566    3.690306
```

Para delimitar las zonas de manejo se utilizó el algoritmo propuesto en el Capítulo 3. Para ello se realizó un Análisis de Componentes Principales espacial (MULTISPATI-PCA) utilizando las librerías “spdep” y “ade4”. Luego, las componentes principales espaciales resultantes del MULTISPATI-PCA fueron utilizadas como *input* del análisis de *cluster fuzzy k-means*. Para este último análisis se utilizó la librería “e1071” que también permite obtener índices para la selección del número óptimo de clases. Para aplicar el

algoritmo se requiere primero calcular la matriz de ponderación espacial en forma similar a la realizada en el Paso 3.

```
cord <- coordinates(Pred[,1:2])
gri <- dnearneigh(cord,0,10)
lw2 <- nb2listw(gri, style = "W")
```

Luego se realizó un Análisis de Componentes Principales (PCA) clásico y posteriormente sobre las componentes generadas por PCA, se aplicó MULTISPATI-PCA. El gráfico obtenido (biplot) permite estudiar la estructura de correlación entre las variables utilizadas para la delimitación de zonas (Fig. 4.8). Las variables CE30, CE90 y Pe se encuentran correlacionas positivamente y son las más importantes en la explicación de la variabilidad espacial a nivel de la primer eje (sPC1, eje horizontal). Mientras que RtoTg y Elevación se correlacionan negativamente y presentan mayor importancia en la sPC2. El gráfico de autovalores (barras) sugiere dos estructuras principales a nivel de sPC1 y sPC2, siempre la sPC1 explica la mayor parte de la variabilidad de los datos seguida por sPC2, sPC3, y así sucesivamente.

```
pca2 <- dudi.pca(Pred[,3:7], center=T,scannf = FALSE,  nf = 5)
ms2 <- multispati(pca2, lw2, scannf = F, nfposi = 5)
s.arrow(ms2$c1,xax = 1, yax = 2, clabel = 1)
add.scatter.eig(ms2$eig, xax = 1, yax = 2, posi = "bottomleft", ratio =
0.2)
```

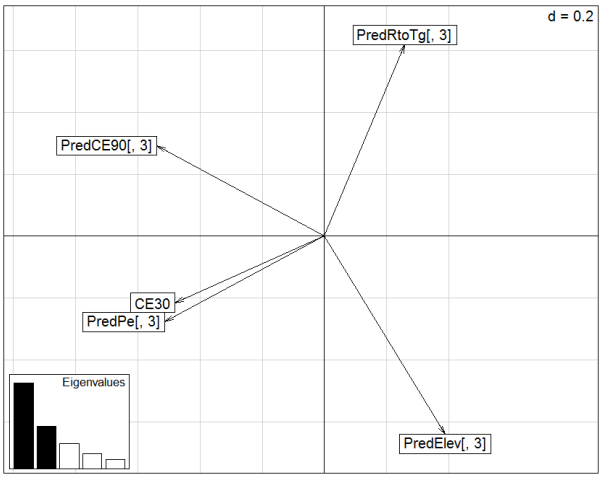


Fig. 4.8. Gráfico Biplot del análisis MULTISPATI-PCA.

Realizado el análisis MULTISPATI-PCA se procede a extraer las sPC para unirlas a la base de datos del análisis y utilizarlas posteriormente como *input* del análisis de *cluster*

fuzzy k-means. El archivo resultante (PredAM) estará conformado por los datos de las coordenadas, los valores predichos y las sPC.

```
CS <- ms2$li[,1:5]
PredAM <- cbind(Pred,CS); PredAM
```

Para realizar el análisis de *cluster fuzzy k-means* se necesita determinar las sPC que se utilizarán como *input*. En este caso se seleccionaron las columnas 8 a 10 que corresponden a la sPC1, sPC 2 y sPC 3. Luego se determinó el número de *cluster* a formar. En este ejemplo se utilizaron 2, 3 y 4 *clusters*. Otras opciones de configuración son el número de iteraciones=100; método=cmeans (opción para usar el algoritmo *fuzzy*) y exponente difuso m=1.3.

```
CM2<-cmeans(PredAM[,8:10],2,100,method="cmeans",m=1.3)
CM3<-cmeans(PredAM[,8:10],3,100,method="cmeans",m=1.3)
CM4<-cmeans(PredAM[,8:10],4,100,method="cmeans",m=1.3)


CM22<-as.data.frame(CM2$cluster)
CM33<-as.data.frame(CM3$cluster)
CM44<-as.data.frame(CM4$cluster)
```

PASO 7-DELIMITACIÓN DE ZONAS DE MANEJO

Delimitadas las clases de manejo se necesita determinar cuál es el número óptimo de clases. En este ejemplo se debe seleccionar entre las dos (I2CM), tres (I3CM) y cuatro (I4CM) clases conformadas. Para ello se utilizaron los siguientes índices: Xie-Beni, coeficiente de partición, entropía de clasificación y Fukuyama-Sugeno. En todos los índices, excepto el coeficiente de partición, el número de clases óptimo se obtiene cuando los índices tienen el menor valor. Para hacer que la interpretación del coeficiente de partición sea igual a los otros índices, se divide a 1 por el valor del índice.

```
I2CM<-fclustIndex(CM2,PredAM[,8:10],index=c("xie.beni","fukuyama.sugeno",
"partition.coefficient", "partition.entropy"))

I3CM<-fclustIndex(CM3,PredAM[,8:10],index=c("xie.beni","fukuyama.sugeno",
"partition.coefficient", "partition.entropy"))

I4CM<-fclustIndex(CM4,PredAM[,8:10],index=c("xie.beni","fukuyama.sugeno",
"partition.coefficient", "partition.entropy"))

Indices0 <- cbind(I2CM,I3CM,I4CM)
XieBeni <-Indices0[1,]
FukSug <-Indices0[2,]
```

```
CoefPart_1 <-Indices0[3,]
CoefPart <- 1/CoefPart_1
EntrPart <-Indices0[4,]

Indices <- as.data.frame(rbind(XieBeni,FukSug,CoefPart,EntrPart))
Indices

                  I2CM          I3CM          I4CM
XieBeni   5.506347e-05  9.456612e-05  9.789067e-05
FukSug   -1.055221e+04 -1.218075e+04 -1.338847e+04
CoefPart  1.091993e+00  1.167154e+00  1.217946e+00
EntrPart  1.416388e-01  2.547190e-01  3.248154e-01
```

En este ejemplo la mayoría de los índices muestran que el número de clases a seleccionar, siguiendo un criterio estadístico, es de dos clases. Puede suceder que ninguno de los índices coincida con otro en el número óptimo de clases. Para facilitar la toma de decisiones se recomienda calcular un índice resumen para cada clasificación. Este nuevo índice puede ser la distancia Euclidea de los valores de los índices previamente normalizados por su valor máximo a través de las diferentes clasificaciones.

```
XieBeniMax<-max(Indices[1,])
FukSugMax<-max(Indices[2,])
CoefPartMax<-max(Indices[3,])
EntrPartMax<-max(Indices[4,])

XieBeniN<- XieBeni/XieBeniMax
FukSugN<- FukSug/FukSugMax
CoefPartN<- CoefPart/CoefPartMax
EntrPartN<-EntrPart/EntrPartMax

IndicesN <- as.data.frame(rbind(XieBeniN,FukSugN,CoefPartN,EntrPartN))
IndicesN2 <- (IndicesN)^2

Indice2CM <- sqrt(sum(IndicesN2[,1]))
Indice3CM <- sqrt(sum(IndicesN2[,2]))
Indice4CM <- sqrt(sum(IndicesN2[,3]))
```

Nuevamente, los valores de los índices resumen muestran que debiera seleccionarse dos clases de manejo. La clasificación con dos clases de manejo presenta grandes zonas con límites más coherentes respecto a la clasificación con 3 y 4 clases que presentan varias zonas pequeñas y de forma irregular (Fig. 4.9, 4.10 y 4.11). La selección final de la cantidad de grupos de acuerdo a lo sugerido por los índices indica que el número óptimo de clases de manejo es de dos.

```
Indice2CM
Indice3CM
Indice4CM
```

```
[1] 1.520007
[1] 1.949105
[1] 2.147047
```

```
base00 <- cbind(PredAM[,1:2],CM22,CM33,CM44)
coordinates(base00) = ~x+y
gridded(base00)=TRUE
spplot(base00["CM2$cluster"],col.regions=terrain.colors(100),colorkey= F)
```

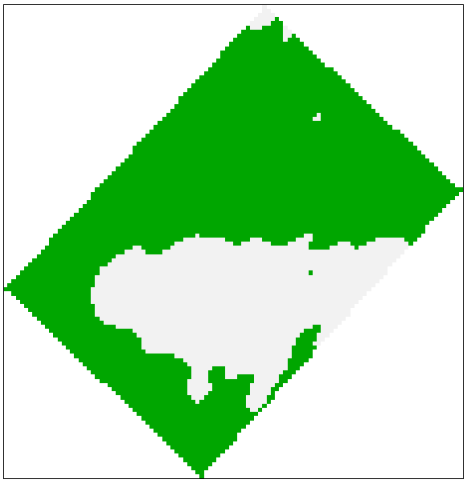


Fig. 4.9. Mapa con dos clases de manejo intralote.

```
spplot(base00["CM3$cluster"],col.regions=terrain.colors(100),colorkey= F)
```

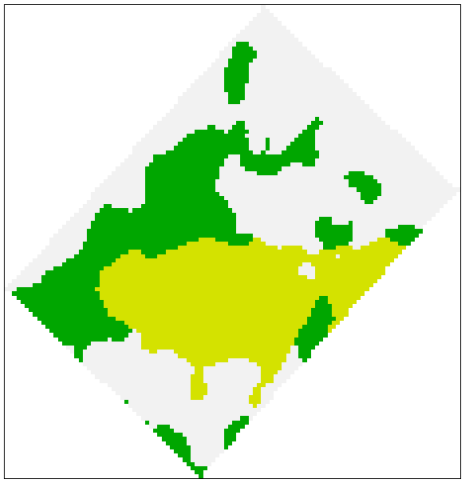


Fig. 4.10. Mapa con tres clases de manejo intralote..

```
spplot(base00["CM4$cluster"],col.regions=terrain.colors(100),colorkey= F)
```

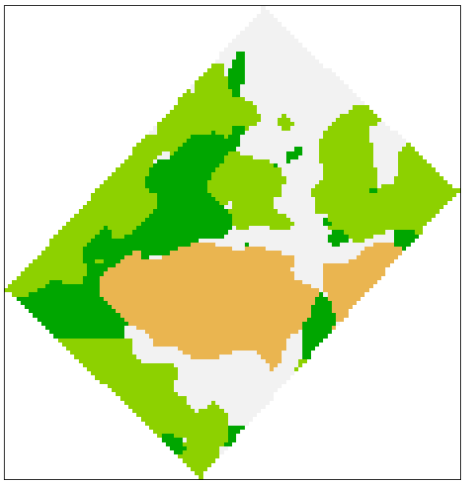


Fig. 4.11. Mapa con cuatro clases de manejo intralote.

Determinado el número de clases en que será dividido el lote es necesario delimitar zonas más contiguas que las producidas y reducir la fragmentación que produce la clasificación a los fines de delimitar zonas de manejo. Para ello, se aplicó un filtro espacial a la clasificación resultante. La librería "raster" se utilizó para aplicar el filtro de la mediana mediante la función *focal*. Es necesario convertir la base de datos en un archivo de imagen, por lo tanto previo a la aplicación del filtro se realizó la conversión de los datos.

```
base0 <- cbind(PredAM[,1:2],CM22)
names(base0)[3]<-paste("Zona")

base1 <- base0
coordinates(base1) = ~x+y
gridded(base1)=TRUE

base2 <- raster(base1)
plot(base2)
```

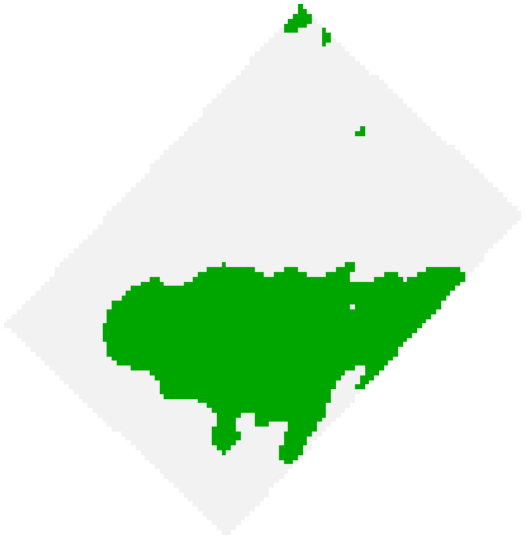


Fig. 4.12. Mapa con dos clases de manejo intralote previo a la aplicación del filtro de la mediana.

El filtro de la mediana reemplaza el valor del píxel central por la mediana de los valores del vecindario de ese píxel. Las mascaras que definen el tamaño de los vecindarios (números de pixel) pueden tener diferentes dimensiones. En este ejemplo se probaron máscaras de 3×3, 5×5 y 7×7 pixeles.

```
median3x3 <- focal(base2, fun=median,w=3)
plot(median3x3)
```

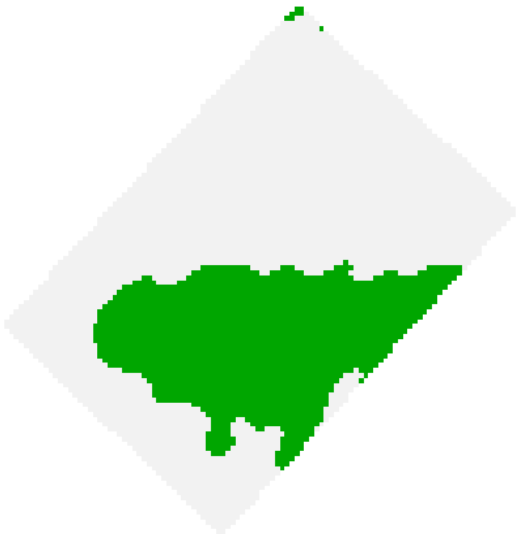


Fig. 4.13. Mapa con dos clases de manejo intralote luego de aplicar un filtro de la mediana de 3 × 3 píxeles.

```
med5x5 <- focal(base2, fun=median,w=5)
plot(med5x5)
```

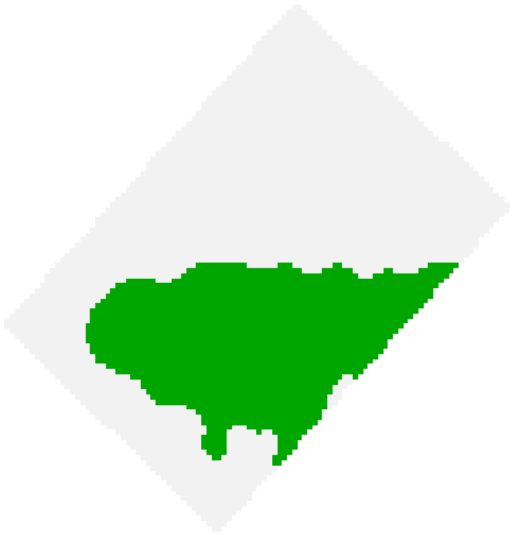


Fig. 4.14. Mapa con dos clases de manejo intralote luego de aplicar un filtro de la mediana de 5 × 5 píxeles.

```
med7x7 <- focal(base2, fun=median,w=7)
plot(med7x7)
```

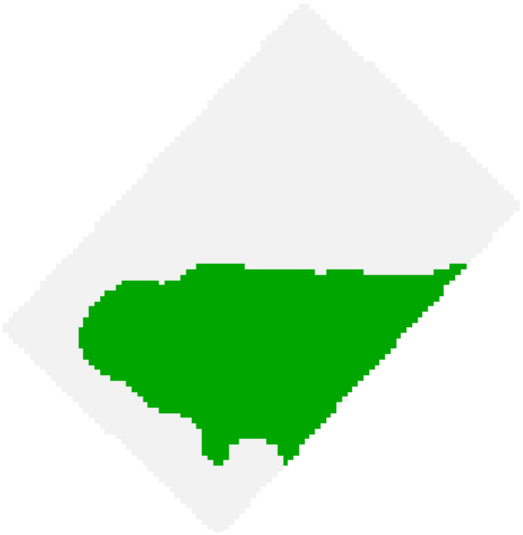


Fig. 4.15. Mapa con dos clases de manejo intralote luego de aplicar un filtro de la mediana de 7× 7 píxeles.

En este ejemplo, el filtro de 7 × 7 píxeles (Fig. 4.14) resultó ser más adecuado que los de 3 × 3 (Fig. 4.12) y 5 × 5 (Fig. 4.13) para lograr una zonificación de bordes menos abruptos y sin fragmentación dentro de las zonas. Finalmente, luego de la aplicación del filtro de la mediana la información de las potenciales zonas de manejo es extraída.

```
base3 <- as.data.frame(med7x7,xy=T)
base3 <- na.omit(base3)
names(base3)[3]<-paste("Zona")

Basefinal<-subset(merge(base3,PredAM,by=c("x","y"),all=T),
select=c(x,y,CE30,CE90,Elev,Pe,RtoTg,Zona))
Basefinal <- na.omit(Basefinal)
```

PASO 8-VALIDACIÓN DE ZONAS DE MANEJO

Se utilizaron datos de MO para la validación de la zonificación mediante el contraste de medias de zonas. Para ello se utilizó un modelo lineal mixto con efecto fijo de zona y errores correlacionados espacialmente. La comparación debiera realizarse con todas las propiedades de suelo medidas en el muestreo de validación y las funciones de correlación espacial podría ser diferentes para las diferentes variables. En esta ilustración, se ajustaron funciones de correlación espacial exponencial, gaussiana y esférica con y sin efecto nugget. Para la selección de los modelos se utilizó el criterio de información de Akaike (AIC) y el test de la razón de verosimilitud (Likelihood Ratio Test, LRT). El modelo de correlación espacial seleccionado, se comparó con el modelos de errores independientes. Para realizar los análisis de este paso del protocolo, se utilizó la librería "nlme". A continuación se procede a cargar el archivo con los datos del muestreo, posteriormente se ajustan modelos y finalmente se selecciona el mejor modelo.

```
Muestreo<-read.table("C:\\Users\\.....\\Muestreo.txt", header = TRUE)
```

```
        X       Y     Zon  mo      nitrato  arcilla
1   312594.6 5801202    1 4.42531  8.40574 33.9095
2   312595.4 5801053    1 4.37601  9.50813 33.7904
3   312668.6 5800985    1 4.27021  9.05652 32.7791
4   312486.5 5801081    1 4.53078 10.07180 34.4820
5   312523.4 5800973    1 4.41935  9.96068 32.3697
6   312743.7 5800767    2 4.18944  9.22031 30.9247
7   312598.2 5800756    2 4.22738  9.17578 30.8015
8   312342.3 5800921    1 4.33593 10.88710 31.1396
9   312526.1 5800676    2 3.96259  9.26113 31.0442
10 312599.4 5800607    2 4.27692  9.34704 33.0203
11 312416.4 5800704    2 3.82355  9.49310 30.3777
12 312380.4 5800664    2 3.89093 10.18570 30.8277
13 312418.0 5800555    2 4.11455 11.08390 32.8783
14 312234.9 5800653    2 4.29290 12.06710 32.1374
15 312089.9 5800641    1 4.85092 14.07280 32.8819
```

Ajuste de modelo de correlación espacial exponencial.

```
modelo.001_mo_REML<-gls(mo~1+Zona,correlation=corExp
(form=~as.numeric(as.character(X))+as.numeric(as.character(Y)),metric="eu
clidean" ,nugget=FALSE) ,method="REML" ,na.action=na.omit ,data=Muestreo)
```

Ajuste de modelo de correlación espacial exponencial con efecto nugget.

```
modelo.002_mo_REML<-gls(mo~1+Zona,correlation=corExp
(form=~as.numeric(as.character(X))+as.numeric(as.character(Y)),metric="eu
clidean",nugget=TRUE),method="REML",na.action=na.omit,data=Muestreo)
```

Ajuste de modelo de correlación espacial gaussiana.

```
modelo.003_mo_REML<-gls(mo~1+Zona ,correlation=corGaus
(form=~as.numeric(as.character(X))+as.numeric(as.character(Y)),metric="eu
clidean" ,nugget=FALSE),method="REML" ,na.action=na.omit ,data=Muestreo)
```

Ajuste de modelo de correlación espacial gaussiana con efecto nugget.

```
modelo.004_mo_REML<-gls(mo~1+Zona ,correlation=corGaus
(form=~as.numeric(as.character(X))+as.numeric(as.character(Y)),metric="eu
clidean",nugget=TRUE),method="REML",na.action=na.omit,data=Muestreo)
```

Ajuste de modelo de correlación espacial esférica.

```
modelo.005_mo_REML<-gls(mo~1+Zona,correlation=corSpher (form=~as.numeric
(as.character(X))+as.numeric(as.character(Y)),metric="euclidean",nugget=F
ALSE),method="REML",na.action=na.omit,data=Muestreo)
```

Ajuste de modelo de correlación espacial esférica con efecto nugget.

```
modelo.006_mo_REML<-gls(mo~1+Zona,correlation=corSpher(form=~as.numeric
(as.character(X))+as.numeric(as.character(Y)),metric="euclidean",nugget=T
RUE),method="REML",na.action=na.omit,data=Muestreo)
```

Ajuste de modelo de errores independientes.

```
modelo.007_mo_REML<-gls(mo~1+Zona ,method="REML" ,na.action=na.omit
,data=Muestreo)
```

En la selección del modelo de correlación espacial, el AIC indica que el modelo con función de correlación espacial gasussiana es el mejor para estos datos, tanto entre los modelos sin nugget como para los modelos con efecto nugget.

```
AICmod1 <- AIC(modelo.001_mo_REML)
AICmod3 <- AIC(modelo.003_mo_REML)
AICmod5 <- AIC(modelo.005_mo_REML)

AICmod1
AICmod3
AICmod5

[1] -1.515517
[1] -4.058434
[1] -1.220597

AICmod2 <- AIC(modelo.002_mo_REML)
AICmod4 <- AIC(modelo.004_mo_REML)
AICmod6 <- AIC(modelo.006_mo_REML)

AICmod2
AICmod4
AICmod6

[1] 0.4844829
[1] -2.060023
[1] -0.4010228
```

En la comparación de los modelos de correlación espacial gaussiana con y sin nugget debido a que la hipótesis implica comprobar que una componente de la varianza (nugget) es igual a cero, se utilizó el estadístico LRT01 (ver Capítulo 6), siendo el valor *p* de la prueba la mitad del valor informado en la distribución chi-cuadrado con un grado de libertad. La hipótesis nula que se somete a prueba con LRT es que el modelo más parsimonioso (con menos parámetros) es el adecuado, en este caso sería el modelo sin efecto nugget. Los resultados de la prueba indican que no se rechaza la hipótesis nula por lo tanto el modelo sin efecto nugget es el que presenta mejor ajuste. Mayor detalle de la selección de modelos de covarainza puede encontrarse en el Capítulo 6.

```
LRT0 <- anova(modelo.003_mo_REML, modelo.004_mo_REML);LRT0

                Model df   AIC   BIC  logLik  Test  L.Ratio   p-value
modelo.003_mo_REML  1  4 -4.058 -1.502 6.029
modelo.004_mo_REML  2  5 -2.060  1.135 6.030 1 vs 2 0.001589  0.9682
```

Finalmente se compara el modelo de correlación espacial seleccionado vs. el modelo de errores independientes (iid). En este ejemplo el modelo no contiene efecto nugget por lo tanto la diferencia respecto al modelo iid es solo de un parámetro, el rango. Por lo tanto, se utiliza para comparar los modelos la mitad del valor *p* informado por la prueba LRT que provee el software. Cuando el mejor modelo de correlación espacial contiene un efecto nugget la determinación del modelo con el mejor ajuste se debiera basar

en los valores de AIC. Los resultados obtenidos en este caso, muestran que el modelo de correlación espacial gaussiana es mejor que el de errores independientes.

```
LRT1 <- anova(modelo.003_mo_REML, modelo.007_mo_REML);LRT1

                   Model df       AIC       BIC   logLik   Test  L.Ratio
p-value
modelo.003_mo_REML     1  4 -4.058434 -1.502205 6.029217
modelo.007_mo_REML     2  3  2.925729  4.842901 1.537136 1 vs 2 8.984163
0.0027


summary(modelo.003_mo_REML)


Generalized least squares fit by REML

  Model: mo ~ 1 + Zona

 Data: Muestreo

        AIC       BIC   logLik
  -4.058434 -1.502205 6.029217

Correlation Structure: Gaussian spatial correlation

 Formula: ~as.numeric(as.character(X)) + as.numeric(as.character(Y))

 Parameter estimate(s):
   range
176.5513


Coefficients:

                Value Std.Error   t-value p-value
(Intercept)  4.797588 0.1643261 29.195530  0.0000
Zona        -0.297806 0.1040640 -2.861755  0.0126


Correlation:
     (Intr)
Zona -0.867


Standardized residuals:

       Min         Q1        Med         Q3        Max
-1.8607719 -0.8864720 -0.3808446  0.2064385  1.7265838

Residual standard error: 0.2033711
Degrees of freedom: 16 total; 14 residual
```

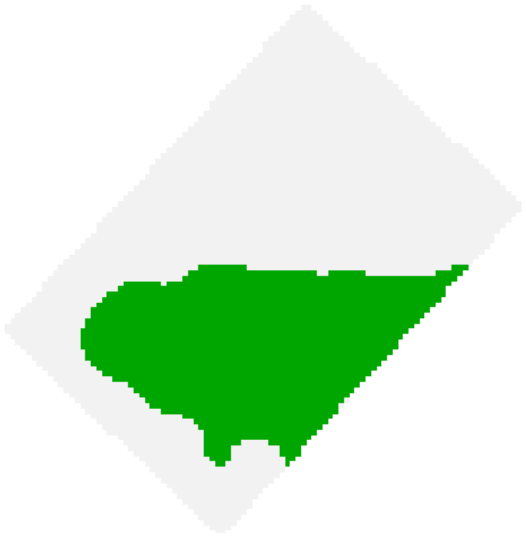


Fig. 4.16. Mapa con dos zonas de manejo intralote.

Los resultados muestran que existen diferencias estadísticamnte significativas entre las zonas delimitadas en cuanto al contenido de MO. La zona 1 (gris) presenta un contenido promedio de 4,80% mientras que para la zona 2 (verde) el valor promedio de MO es de 4,50% (Fig. 4.15).

CAPÍTULO V

ANÁLISIS TEMPORAL DE LA VARIABILIDAD ESPACIAL DE PROPIEDADES DE SUELOS INTRALOTE

INTRODUCCIÓN

En los inicios de la década del 90’ cuando se comenzaron a utilizar los monitores de rendimiento, se esperaba que siempre las zonas del lote que presentaban altos o bajos rendimientos mantuvieran esta característica a través del tiempo. Esto era debido a la suposición de que las características del suelo se comportaban de la misma manera cada año. La variable medida en un determinado sitio del lote tendría valores relativamente menores/mayores que la del sitio de otra zona del mismo lote mientras no se aplican tratamientos para cambiar esa condición (Blackmore *et al*., 2003).

Esta idea fue más fuertemente sostenida para características de suelo que para el rendimiento. La alta variabilidad intranual de los rendimientos ocasionada por variaciones climáticas aleatorias, reforzó aún más la característica de “permanente” o “sistemática” dada a las zonificaciones producidas a partir de variables de suelo. Esto llevó a que se realizaran prácticas de manejo de suelo y fertilidad diferenciadas, según dichas zonificaciones, para “controlar” la variabilidad espacial de la producción. Por ello, para manejar la variabilidad espacial surgió el concepto de zonas homogéneas; éstas se interpretan como conjuntos de sitios dentro del lote con una misma combinación de factores limitantes de la producción en su principal factor no aleatorio, el suelo, y por tanto con una misma demanda de tratamiento o manejo para uniformar “hacia arriba” los rendimientos (Vrindts *et al*., 2005). Whelan y McBratney (2000) sostenían que el manejo de la variabilidad de los rendimientos a una resolución espacial fina representaría una mejora en la producción respecto al manejo uniforme de un lote. También sostenían que la

oportunidad de aplicar el manejo sitio-específico de los cultivos debiera ser planteada como la hipótesis alternativa de la denominada "hipótesis nula" de la agricultura de precisión: dada la elevada variabilidad temporal del rendimiento de los cultivos, el manejo uniforme del lote es la estrategia óptima y de menor riesgo. Por lo tanto, para que sea posible aplicar el manejo sitio-específico es necesario que el patrón de la distribución espacial del rendimiento se mantenga estable en el tiempo. Si esta premisa no se cumple, probablemente el manejo diferencial del cultivo no conducirá a los resultados previstos (Cook y Bramley, 1998).

Las variables de suelo medidas en un momento de tiempo (trasversalmente) son utilizadas, conjuntamente con los rendimientos, para la delimitación de zonas homogéneas en cada ciclo agrícola. La zonificación puede realizarse en estos casos de forma multivariada, es decir contemplando todas las variables simultáneamente (Yan *et al*., 2007) o variable a variable. Los mapas de rendimiento y los mapas de suelo a escala de lote proveen indicadores de riesgos asociados con la producción y constituyen herramientas claves para implementar el manejo sitio-específico en agricultura de precisión. Cuando los datos de suelo se obtienen para cada sitio en varios años, la zonificación podría depender no sólo de la variabilidad espacial o variabilidad entre sitios sino también de la variabilidad temporal (variabilidad interanual) de las mediciones. Luego, la comparación de los valores de una determinada variable, tomada durante varios años es posible sólo si se comparan los valores correspondientes a las mismas localizaciones de referencia. Por ello, las series de mapas de rendimiento y mapas de suelo, resultante de procesos de interpolación espacial llevado a cabo año tras año, constituyen una herramienta apropiada para analizar la estabilidad temporal de la zonificación (Blackmore *et al*., 2003). Boydell y McBratney (2002) recomendaron la utilización de un mínimo de 5 años (±2 años) consecutivos de información para la delimitación de zonas estables de cosecha en lotes de algodón. Para otros cultivos y ambientes climáticos este número puede ser distinto, aunque el mínimo para tal análisis deberá ser de tres años. Cuando se usan datos de suelo generalmente éstos son de un año y por ello, los cambios temporales de la distribución espacial son más comunes en el análisis de datos de rendimientos que en el análisis de datos de suelo. Sin embargo, la variabilidad temporal en las variables de suelo puede impactar las zonificaciones y consecuentemente las predicciones de rendimientos para años futuros.

La tendencia promedio de los cambios de las variables medidas subsecuentemente en distintos años puede ser analizada con modelos estadísticos contemporáneos como los Modelos Lineales Mixtos (MLM) (Gbur *et al*., 2012). Estos modelos permiten contemplar las correlaciones esperadas no sólo entre datos obtenidos en distintos sitios separados espacialmente dentro del mismo lote, sino también entre los obtenidos en un mismo sitio a través del tiempo. Lark y Stafford (1997), Shearer (2001), Blackmore *et al*. (2003), Diker *et al*. (2004) y Marques da Silva (2006), desarrollaron distintas metodologías para el análisis de la variabilidad temporal de datos georreferenciados y en el contexto de la zonificación intralote.

Blackmore (20003) y Marques da Silva (2006) propusieron una metodología de análisis basada en la información que aportan los mapas de rendimiento anuales. La combinación adecuada de un mapa de tendencia espacial (calculado a partir de la media aritmética del rendimiento intralote en cada sitio para los distintos años de la serie) y de un mapa de estabilidad temporal (cuyos valores de sitio son ahora los desvío estándar del rendimiento para los años de la serie estudiada) permite la diferenciación de zonas de producción estable en el tiempo, ya sean de alta o baja productividad, y zonas cuya producción es temporalmente inestable. Un método similar es el que utiliza Shearer (2001) en viñas. En este caso, los mapas de rendimiento que se compararon fueron estandarizados (media cero y varianza uno) para que las diferencias interanuales del rendimiento no influyan en la interpretación de los mismos. El método propuesto por Diker *et al*. (2004) utiliza la información de varios años de rendimiento para la delimitación de zonas de distinto potencial productivo. Empleado inicialmente en maíz, las zonas se diferencian entre sí por el número de veces (número de años de una serie de tres) que el rendimiento se sitúa por encima de la media intralote anual. Ninguno de estos análisis contempla simultáneamente las correlaciones esperadas por el proceso espacio-temporal subyacente.

En este capítulo proponemos un algoritmo de análisis de datos basados en la combinación de MLM que permitan contemplar la correlación temporal y espacial en los datos de variables de suelo. Las predicciones derivadas de estos modelos para varias variables sobre una grilla de sitios dentro del lote, son luego usadas para clasificar los distintos sitios de acuerdo al valor promedio de la variable y su variabilidad interanual. Esta propuesta metodológica podría ser usada para construir zonificaciones que no solo

consideren la variabilidad espacial de las variables de suelo sino también su estabilidad temporal.

MATERIALES Y MÉTODOS

DATOS

Para ilustrar la metodología propuesta, se analizaron datos de un estudio llevado a cabo en un campo ubicado en el departamento Rio Seco de la provincia Córdoba, Argentina. El área de relevamiento de datos pertenece a la región semiárida y se caracteriza por una topografía de sierra ondulada y una precipitación anual de 700 mm con veranos calurosos. Se analizó la tendencia temporal en un periodo de 12 años (con mediciones cada tres años: 2005, 2008 y 2011) de las variables MO (%), P (mg kg^{-1}), pH, y CE (dS m^{-1}). Los datos experimentales fueron recolectados en 12 lotes (post-desmonte) de 160 a 220 has cada uno, abarcado una superficie total de 2.240 ha bajo agricultura continua con rotaciones de cultivos anuales (Soja, Maíz y Trigo). Se obtuvieron entre 170 y 280 muestras de suelo en cada año; las muestras fueron georreferenciadas con un DGPS. Los datos fueron pre-procesados para la eliminación de errores identificables mediante el uso de gráficos box-plot.

ESTRATEGIA DE ANÁLISIS

Para cada variable se evaluó la tendencia interanual promedio del área con un MLM con efecto fijo de año y de lote, un término aleatorio de sitio y estructura de correlación espacial para los errores. El modelo con correlación espacial exponencial (con y sin efecto nugget) basado en el cálculo de distancias entre los sitios de muestreo georreferenciados fue comparado con el modelo de errores independientes mediante la prueba de cociente de verosimilitud (LRT) basada en los estimadores REML (máxima verosimilitud restringida) de los parámetros de varianza y covarianza (Pinheiro y Bates, 2004). Ajustado el modelo de correlación espacial, se evaluó la reducción del modelo (eliminación de efecto lote)

mediante la prueba LRT basada en estimadores ML (máxima verosimilitud). Las diferencias encontradas entre año fueron evaluadas con la prueba LSD de Fisher (α=0.05) para las medias ajustadas por la estructura de varianza y covarianza subyacente. Luego el mismo procedimiento fue realizado para cada variable analizada para cada año de muestreo. Posteriormente, se realizaron predicciones en sitios no muestreados en una grilla de 60 m × 60 m utilizando el método de kriging ordinario. Para ello se utilizaron los parámetros del semivariograma que fueron estimados con el MLM seleccionado. Para calcular la variabilidad temporal de cada variable interpolada, en un determinado sitio o punto muestral, se usó la siguiente expresión sobre los valores predichos por el modelo ajustado para cada sitio:

$$\delta_i = \sqrt{\frac{\sum_{t=05,08,11} (Y_{t,i} - \bar{Y}_t)^2}{3}} \qquad (5.1)$$

donde δ_i es la desviación estándar temporal del sitio *i*; *t* es el año (2005, 2008 y 2011); $Y_{t,i}$ es el valor de la variable de suelo predicho por el modelo para el año *t* en el sitio *i*; $\bar{Y}_t$ es la media de la variable para todo el lote en el año *t*.

La ecuación refleja la desviación del dato de cada sitio con respecto a la media temporal de la variable. Por definición, la desviación estándar temporal (DET) presenta un valor bajo, si un determinado sitio del lote presenta un valor que está siempre cerca de la media general de la variable (Blackmore *et al*., 2003). Este sitio es definido como estable en términos temporales, mientras que sitios que tienen un valor de la variable que a veces se aproximan a la media y a veces se alejan de ella son considerados sitios inestables.

El dato de la DET, obtenido con la función descripta en (5.1), fue representado gráficamente en función de la media temporal de la variable en el sitio, usando una grilla de 60 m × 60 m. La media de la variable de suelo a través de los años fue trazada como referencia vertical del gráfico para identificar los sitios que se ubican por encima de la media. Mientras que el tercer cuartil de la desviación estándar temporal, denotado como P(75), fue trazado como referencia horizontal del gráfico con el objeto de identificar los sitios más inestables *i.e.* DET> P(75).

Usando los datos espaciales de las cuatro variables de suelo se delimitaron dos zonas homogéneas en cada año mediante el análisis de *cluster*. Para evaluar la concordancia entre las clasificaciones de cada año se cálculo el índice de concordancia kappa de Cohen (k) (Cohen, 1960).

$$k = \frac{p_o - p_e}{1 - p_e} \tag{5.2}$$

donde p_o es la proporción de observaciones dentro de las celdas diagonales de la tabla de contingencia y p_e son las proporciones esperadas en dicha diagonal bajo la hipótesis nula de no asociación entre las clasificaciones. Dicho índice mide el grado de concordancia entre dos variables categóricas en una escala de 0 a 1. A mayor cantidad de observaciones en la diagonal de la tabla de contingencia entre dos clasificaciones, mayor es el grado de concordancia y por lo tanto el índice tiende a 1 (máximo acuerdo). Por el contrario, valores de kappa cercanos a 0 indican independencia entre ambas clasificaciones (total desacuerdo). Landis y Koch (1977) proponen la siguiente escala de valoración del k (Tabla 5.1). Adicionalmente, se obtuvieron los errores estándar y los intervalos de confianza de forma no-paramétrica para el índice k, mediante la técnica de remuestreo Bootstrap (Efron y Tibshirani, 1993).

Tabla 5.1. Escala de valoración del índice de kappa

kappa	Grado de acuerdo
< 0.00	sin acuerdo
>0.00 – 0.20	insignificante
0.21 – 0.40	discreto
>0.41 – 0.60	moderado
0.61 – 0.80	sustancial
0.81 – 1.00	casi perfecto

Finalmente, cada zona fue particionada nuevamente según la variabilidad temporal de cada variable. De esta manera se obtuvieron mapas espacio-temporales delimitando cuatro clases de sitios: ZM1-DET> P(75), ZM1-DET< P(75), ZM2-DET> P(75) y ZM2-DET< P(75).

RESUMEN DE LOS PASOS PARA IMPLEMENTAR EL ALGORITMO PROPUESTO

Paso 1: ajustar un MLM para datos longitudinales (con correlación temporal) que contemple la correlación espacial entre las observaciones de los distintos sitios tomadas en un mismo año. Se presenta el código del programa SAS(Versión 9.1) para un ajuste de este tipo donde la correlación temporal queda inducida a través de la incorporación de un efecto aleatorio de sitio y la correlación espacial es ajustada con un modelo espacial exponencial entre las observaciones de un mismo año. Los parámetros del modelo espacial podrían cambiar entre los años. La sintaxis de los comandos Proc Mixed para el ajuste de un MLM con correlación espacial y temporal en SAS (Versión 9.1) es:

```
proc mixed  data=cai; class lote  ano  sitio; model mo= ano lote
lote*ano/outpredm=predichos;
lsmeans ano/pdiff;  random sitio;  repeated/type=sp(exp) (x y) subject=ano; run;
```

Paso 2: con los parámetros del modelo ajustado (en este caso rango=0.2, sill=0.1 y nugget=0.14) realizar una predicción en una grilla regular utilizando el método de kriging. El código del programa SAS para realizar un kriging del tipo ordinario es:

```
proc krige2d  data= predichos ; outest=kriged; coordinates  xcoord=x  ycoord=y;
grid griddata= predichos   xcoord= x  ycoord= y; predict  var= z;
model  form= exponential  range= 0.2  scale= 0.1  nugget= 0. 14;
run;
```

Paso 3: aplicar la ecuación de la DET dada en (5.1) sobre los valores predichos para la variable de interés en cada punto de la grilla.

Paso 4: usando los valores predichos de las variables analizadas realizar un análisis de cluster para delimitar zonas homogéneas. El código del programa SAS para realizar un análisis de *cluter k-means* se muestra a continuación. En este paso otros métodos de clasificación, como el propuesto en el Capítulo III, pueden utilizarse.

```
proc fastclus  data=caipred  maxc=2  maxiter=10  out=cluster;
var  mo  fosforo  ph  ce;
run;
```

Paso 5: particionar las zonas delimitadas según los percentiles (P75) de la desviación estándar temporal de cada variable.

Paso 6: representar gráficamente las zonificaciones obtenidas.

RESULTADOS

En el caso de estudio usado como ilustración, el modelo de correlación espacial ajustó mejor que el modelo de errores independientes para todas las variables (LRT, $p<0.05$). El efecto de lote fue también significativo para MO y P (LRT $p<0.05$). Las distribuciones de valores de todas las variables, excepto CE, fueron relativamente simétricas, con mayores cambios en la media que en la varianza a través de los años (Fig. 5.1).

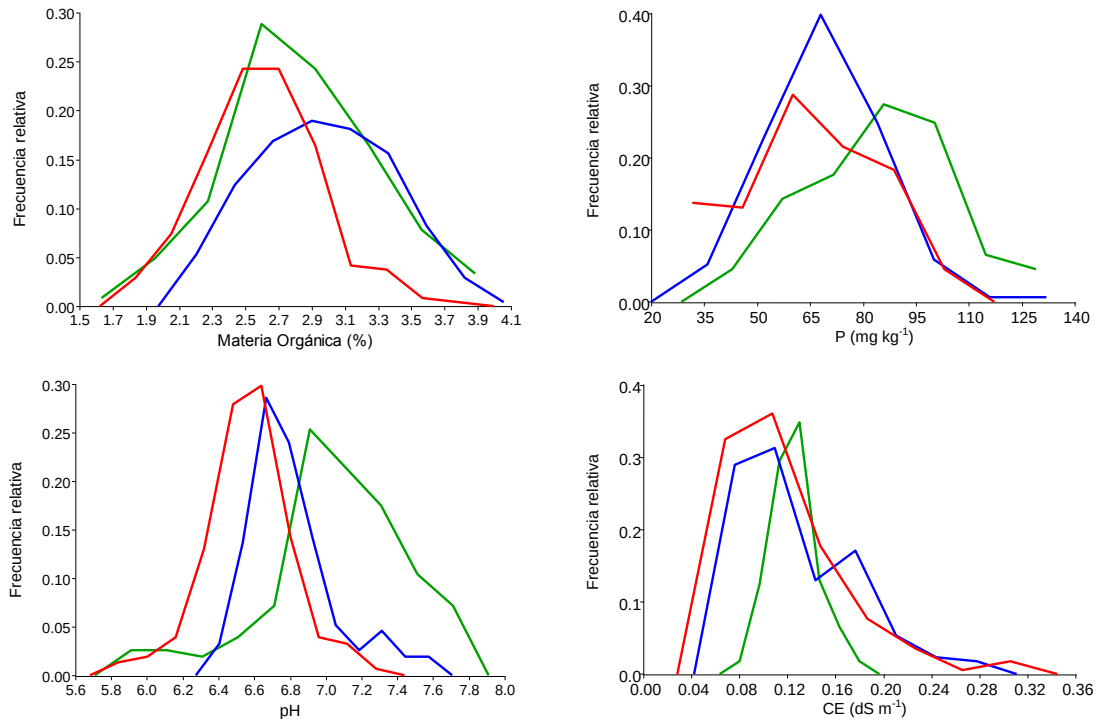


Fig. 5.1. Distribuciones de frecuencias de datos de MO, P, pH y CE en tres años (2005 (verde), 2008 (azul), y 2011 (rojo).

La comparación de los valores promedios, de cada variable, entre los diferentes años de estudio (Tabla 5.2), indicó decrecimientos significativos y progresivos en P y pH (Fig. 5.2). La caída de la MO fue significativa pero en el noveno año ($P<0.05$), mientras que los valores de CE permanecieron sin cambios temporal significativos y fueron los de mayor variabilidad espacial dentro de cada año. La variable P también presentó alta varaibilidad espacial. Estos cambios podrían indicar sitios con mayor/menor riesgo de obtención de rendimientos pobres, debido a correlaciones detectadas entre la P y el rendimiento (Fu *et al*., 2010).

Tabla 5.2. Media, coeficiente de variación (CV), valores mínimos (Min.) y máximos (Max.) de variables de suelo para un mismo lote agrícola en tres años.

Variable	Año	Media	CV	Min.	Max.
MO	2005	2.88	17.50	1.47	4.69
	2008	2.98	14.42	2.00	4.40
	2011	2.61	13.69	1.73	3.89
P	2005	85.54	24.15	35.54	136.21
	2008	70.13	22.56	27.83	139.95
	2011	64.02	32.79	24.30	115.65
pH	2005	7.06	5.51	5.81	7.81
	2008	6.80	3.64	6.30	7.74
	2011	6.57	3.58	5.76	7.36
CE	2005	0.13	15.22	0.07	0.19
	2008	0.13	40.77	0.05	0.32
	2011	0.12	41.19	0.04	0.33

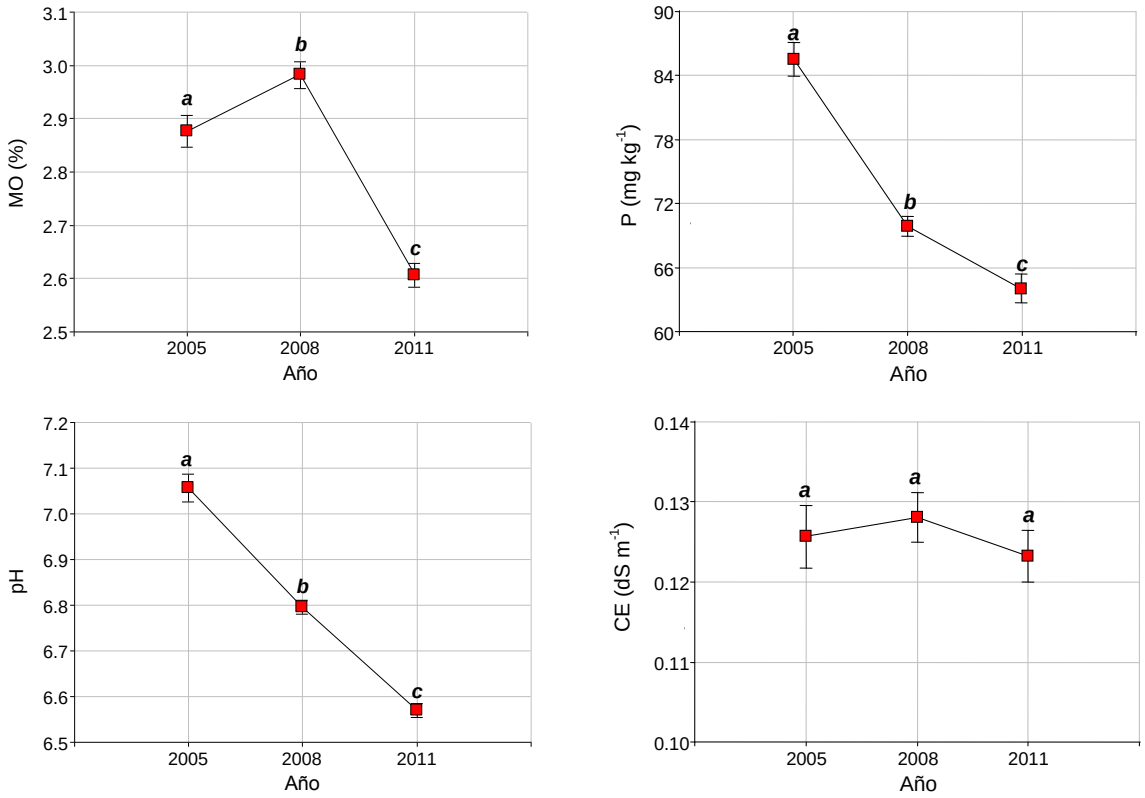


Fig. 5.2. Tendencia temporal en promedios de MO, P, pH y CE. Letras diferentes indican diferencias estadísticamente significativas ($p<0.05$).

Los gráficos de dispersión de la variabilidad temporal en función de la media temporal de cada variable (Fig. 5.3), muestran que para MO, los sitios más inestables fueron los que tenían valores más alejado del promedio temporal (2.83%) y que la DET de estos sitios fue relativamente alta (muy por encima del P(75)). Para CE, los sitios más inestables fueron los de valores mayores a la media (0.126 dS/m).

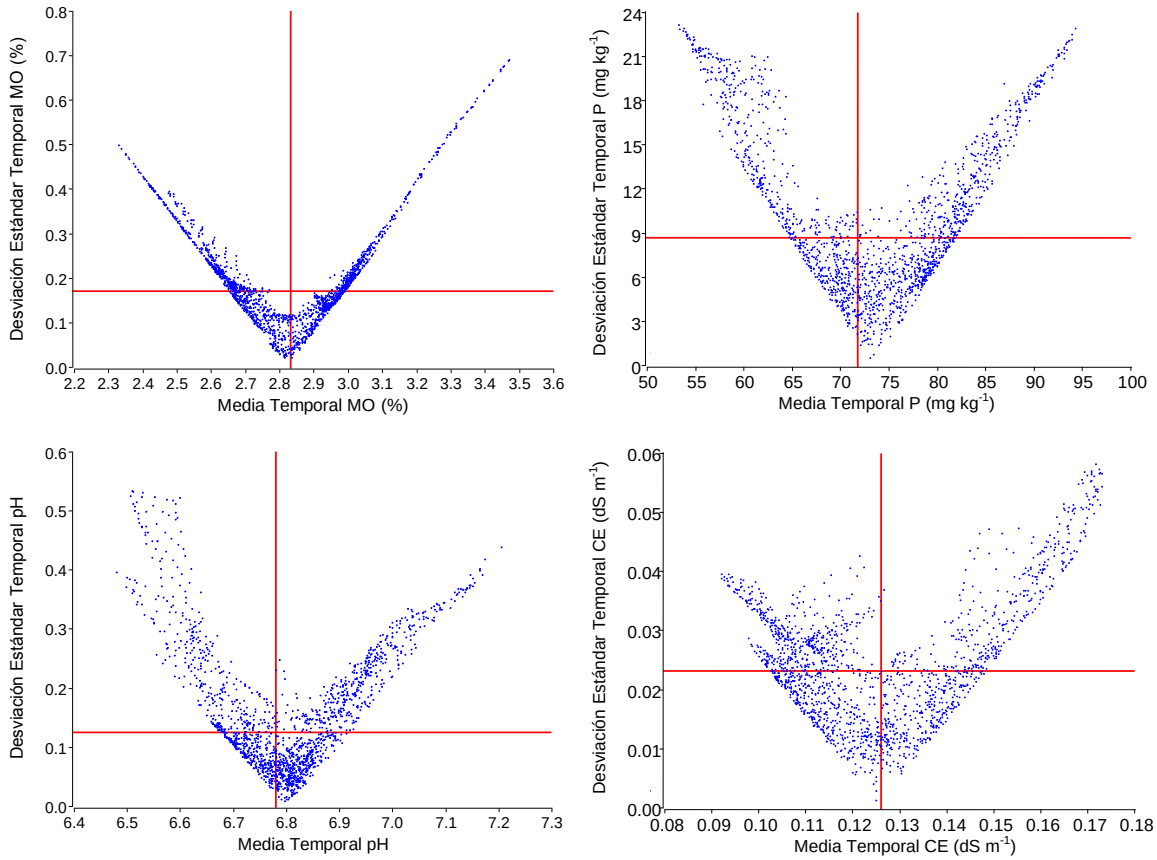


Fig. 5.3. Gráficos de dispersión de la variabilidad temporal en función de la media temporal de cada variable de suelo.

Los mapas de la Figura 5.4, muestran las zonificaciones logradas con los datos de las cuatro variables en cada año. En el análisis de *cluster*, se delimitaron dos zonas homogéneas (ZM) cada año. En la Tabla 5.3 pueden observarse los valores del índice Kappa que permiten medir el grado de concordancia entre las zonificaciones obtenidas en los tres años evaluados. En la comparación de la zonificación de los años 2005 y 2008 se obtuvo el mayor nivel de concordancia (k=0.40) mientras que en la comparación de los años 2008/2011 y 2005/2011 el nivel de acuerdo fue similar (k=0.12). Según la escala de Landis y Koch (1977), estos valores de k corresponden a un grado de acuerdo insignificante y moderado, respectivamente. En las tres comparaciones, con un nivel de confianza del 95%, los índices Kappa fueron mayores a cero (sus intervalo de confianza no incluye el cero).

Tabla 5.3. Índice Kappa para evaluar la concordancia de la zonificación obtenida entre tres años de captura de mediones de propiedades de suelo.

Años comparados	Índice Kappa	Error Estándar	LI (95%)	LS (95%)
2005 - 2008	0.41	0.023	0.35	0.45
2008 - 2011	0.12	0.024	0.07	0.17
2005 - 2011	0.12	0.024	0.07	0.17

LI: límite inferior, LS: límite superior.

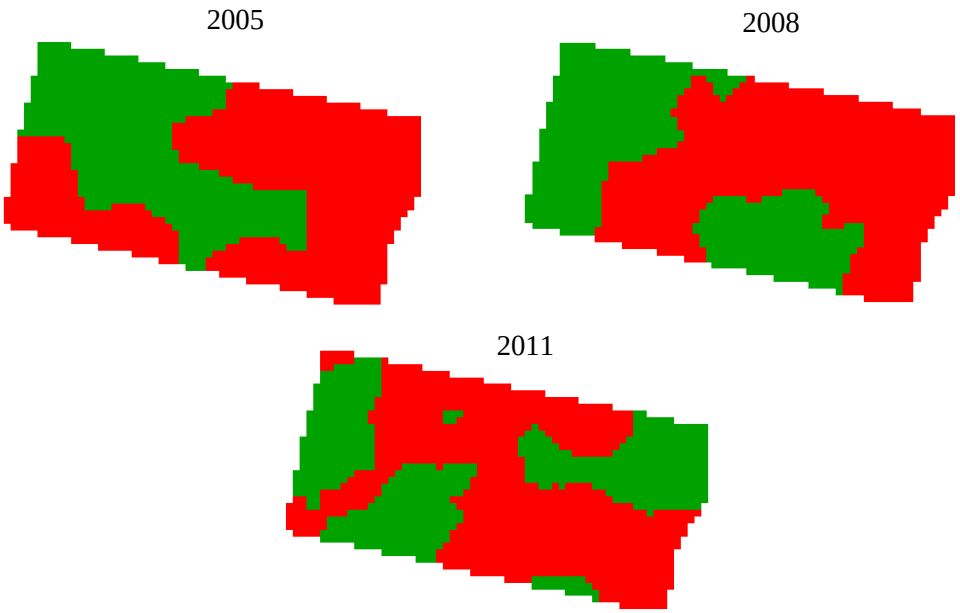


Fig. 5.4. Zonas de manejo para cada año obtenidas con datos de cuatro variables de suelo.

Los mapas de la Figura 5.5, muestran las zonificaciones logradas con los datos de las cuatro variables en cada año y su partición en relación a la inestabilidad temporal de cada variable. La ZM 1 se caracterizó por presentar mayores contenidos de MO, P y CE y menor pH. Esta ZM mostró mayor proporción de sitios clasificados como inestables.

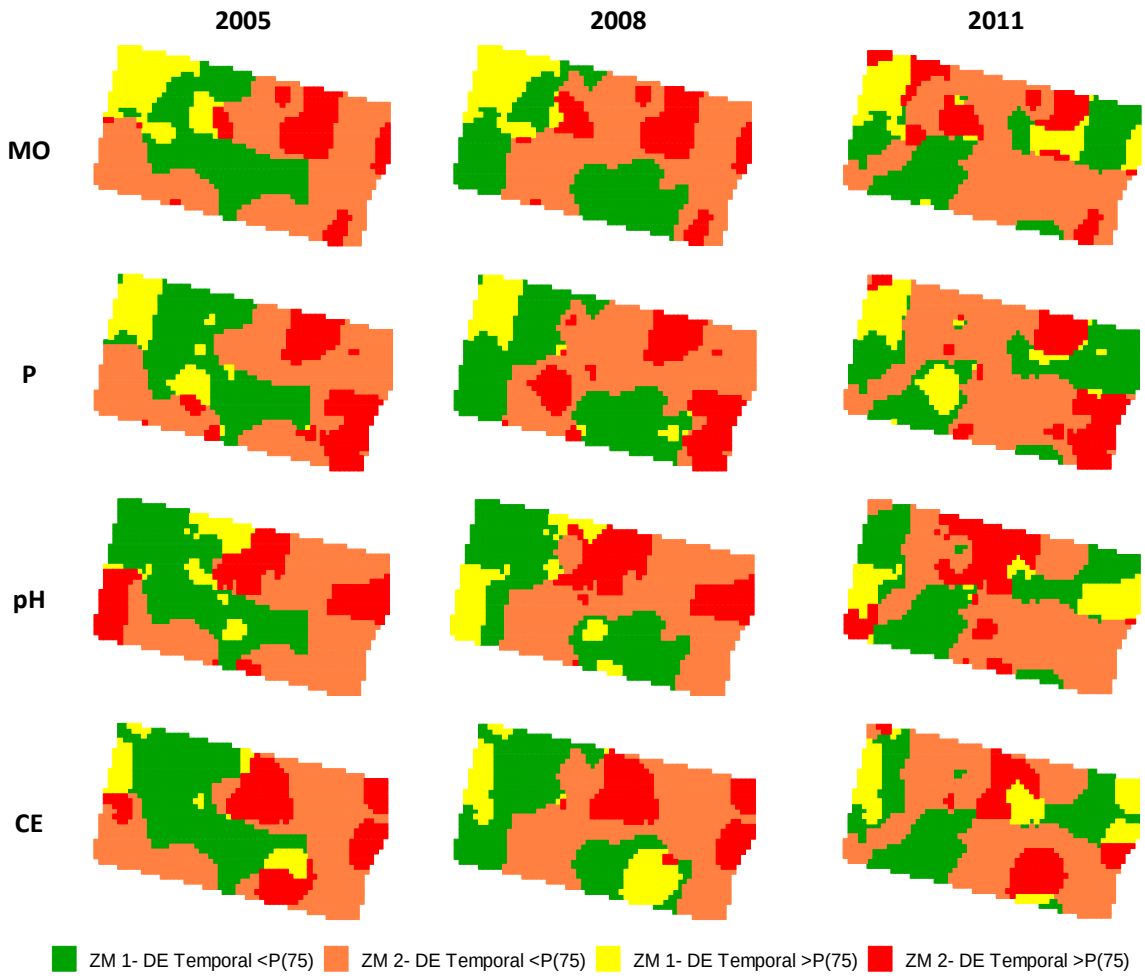


Fig. 5.5. Zonas de manejo para cada año obtenidas con datos de cuatro variables de suelo y particionadas en función de la inestabilidad temporal de cada variable. Sitios con desviación estándar temporal por encima del tercer cuartil de la desviación temporal (P(75)) son más inestables.

CONCLUSIÓN

El algoritmo propuesto a diferencias de otros métodos de identificación de zonas de manejo clásico, permite construir nuevas clasificaciones dentro de cada zona contemplando las correlaciones esperadas entre datos obtenidos tanto en un mismo sitio a través del tiempo como en distintos sitios separados espacialmente dentro del mismo lote. Los resultados de su aplicación mostraron que la variación espacial en las características de

suelo no fue permanente, produciéndose significativos cambios en la delimitación de zonas homogéneas a través de los años. Por ello, el manejo sitio-específico debiera realizarse acorde a las condiciones o zonificaciones emergentes en cada año. Será necesario también evaluar si utilizando variables de sitio que pueden medirse en forma intensiva (conductividad eléctrica aparente, elevación y profundidad del suelo) y cuyos patrones de distribución espacial no cambian o cambien menos en el tiempo con a finalidad de delimitar zonas que sean mas repetibles en el tiempo, aún bajo distintas condiciones de suelo.

CAPÍTULO VI

MODELOS MIXTOS PARA EL ANÁLISIS DE ENSAYOS DE FERTILIZACIÓN SITIO-ESPECÍFICA

INTRODUCCIÓN

En la última década, la adopción de tecnologías de Agricultura de Precisión (AP) por parte de los agricultores argentinos ha tenido un crecimiento sostenido (Melchiore *et al*., 2013). La AP permite implementar estrategias de manejo sitio-específico para abordar la variabilidad espacio-temporal de la producción agrícola (Pierce y Nowak, 1999) que favorecerán incrementos en rentabilidad al ajustar, por ejemplo, las dosis de fertilizantes según el tipo de suelo y otras condiciones del ambiente (Gregoret *et al*., 2006).

La aplicación de fertilizantes nitrogenados en el cultivo de cereales es una estrategias de manejo relevante ya que el nitrógeno (N) es uno de los nutrientes con importante respuesta sitio-específica (Bongiovanni *et al*., 2007). Para optimizar la fertilización con N es necesario comprender cómo zonas o sitios específicos dentro del lote, responden a la aplicación diferencial de fertilizante. Para ello, se llevan a cabo experimentos *in situ* usando maquinarias de AP ya que las propiedades del suelo y del terreno se miden y georreferencian en cada sitio del lote. Luego, diferentes tratamientos de fertilización son esparcidos en todo el lote, en parcelas más pequeñas bajo un diseño experimental particular. Uno de los diseños más usados es el de bloques completos al azar (DBCA) (Balzarini *et al*., 2012) altamente replicado, es decir, con numerosos bloques que permiten una alta repetición de cada tratamiento (dosis de fertilizante) dentro de un mismo lote. Finalmente, las cosechadores equipadas con monitores de rendimiento y GPS son utilizadas para proporcionar datos de cada sitio. Estos datos deben ser capturados, pre-procesados para eliminar casos “raros” y analizados estadísticamente con el objetivo de determinar qué efecto tuvo un tratamiento particular en el rendimiento del cultivo dentro de cada clase de sitios, previamente delimitada.

El análisis puede realizarse con un modelo lineal de ANAVA para evaluar la significancia de la diferencia entre tratamientos o bien con un modelo lineal de regresión para evaluar a través de una función, la respuesta del cultivo en relación a la dosis usada. Tanto el modelo de ANAVA como el de regresión lineal, bajo la teoría de modelo lineal general (MLG), suponen errores independientes (Balzarini *et al*., 2012). Sin embargo, los datos espaciales generados por las tecnologías de AP están correlacionados, impidiendo el uso de modelos estadísticos convencionales, ya que el supuesto de independencia de los datos es difícil de sostener. En AP, es frecuente encontrar correlación espacial en las observaciones y en los residuales de los modelos estadísticos utilizados en los análisis de datos (Hong *et al*., 2005). Está bien documentado que las propiedades físicas del suelo (Davidoff y Selim, 1988), la humedad (Achouri y Gifford, 1984) y el contenido de nutrientes del suelo (Webster y Nortcliff, 1984), varían espacialmente. Por tanto, es de esperar que las observaciones de rendimiento del cultivo también estén correlacionadas espacialmente, e incluso puedan tener diferentes varianzas según el sector del lote al que pertenecen (Scharf y Alley, 1993).

El DBCA es ampliamente utilizado en ensayos agrícolas. La estratificación, o bloqueo de parcelas, es una técnica usada para controlar los efectos de variación entre estratos de unidades experimentales. El análisis de DBCA supone que dentro de los bloques los errores del modelo son independientes e idénticamente distribuidos (iid) y tiene la misma varianza. Sin embargo, los ensayos en AP se llevan a cabo en una escala espacial grande (varias hectáreas), por tanto resulta difícil asegurar la homogeneidad de las parcelas que conforman el bloque que muchas veces se delimitan por cuestiones de logística. Las parcelas que conforman el estrato más próximas pueden ser más similares que las más distantes, generando variabilidad espacial dentro del bloque. Debido a la existencia de variabilidad espacial intrabloque, el ANAVA del DBCA no siempre elimina los sesgos introducidos por las parcelas, en las comparaciones de efectos de tratamientos (Casanoves *et al*., 2005). Cuando la variabilidad espacial está presente en una escala que el bloqueo no puede modelar por sí solo, puede resultar útil introducir funciones de correlación entre los errores. Otro rasgo característico en los experimentos de AP es que dentro de cada parcela pueden registrarse numerosos datos de sitio o "submuestras" que en la práctica se denominan "pseudoreplicas". Estas observaciones que pertenecen a la misma parcela no

pueden considerarse independientes. El uso de modelos lineales mixtos (MLM) (West *et al*., 2007) permite la modelación de éstas y otros tipos de dependencias.

Los principios asociados con el diseño y el análisis estadístico de los ensayos en AP han evolucionado rápidamente para hacer frente a las violaciones de los supuestos de los análisis estadísticos clásicos. Los primeros ejemplos de diseños específicos incluyen al "tablero de ajedrez" (Cook y Bramley, 1998) el cual se basa en aplicar todos los tratamientos a través del lote en un arreglo matricial. El objetivo de éste y otros diseños (Li *et al*, 2001; Bishop y Lark, 2006; Willers *et al.*, 2008; Panten *et al*., 2010) es establecer un ensayo capaz de explorar la respuesta de un tratamiento particular a variaciones de suelo e interacciones locales. En otras palabras, en lugar de eliminar variabilidad espacial del experimento, tales diseños trataron de hacer uso de esta variabilidad como una herramienta experimental. El DBCA profusamente replicado persigue la misma idea. Los bloques pueden distribuirse sin restricción en todo el lote o dentro de zonas previamente delimitadas.

Diferentes procedimientos estadísticos que contemplan la variación espacial entre parcelas de ensayos a campo, han sido propuestos para el análisis de datos correlacionados espacialmente. Desde métodos que realizan el ajuste de medias de tratamientos en función de lo observado en las parcelas vecinas más cercanas (Papadakis, 1937) hasta el uso de modelos que contemplan las correlaciones espaciales en términos del error ajustando medias de tratamientos (Mead, 1971; Besag, 1974; Besag, 1977; Ripley, 1981). Gilmour *et al*. (1997) particionan la variabilidad espacial, entre parcelas de un ensayo, en variabilidad espacial local y global. La variabilidad espacial local hace referencia a las diferencias entre parcelas a pequeña escala, donde se contemplan las variaciones intra-bloque.

Actualmente, la modelación de la estructura espacial a partir de funciones de distancia puede realizarse en el contexto de los MLM (Zimmerman y Harville 1991; Gilmour *et al*., 1997; Cullis *et al*., 1998) donde además de contemplar la estructura de correlación entre observaciones provenientes de distintas parcelas y la correlación de observaciones dentro de una misma parcela, es posible modelar heterogeneidad de varianza residual. Brownie y Gumpertz (1997) investigaron la validez y solidez de los análisis de modelos de correlación espacial y concluyeron que el modelado de las correlaciones podría lograr incrementos sustanciales en la precisión comparado con el análisis DBCA bajo

MLG, cuando los errores están correlacionados espacialmente. Cuando la correlación espacial de los errores entre las observaciones espacialmente georeferenciadas es contemplada, se obtiene una estimación más eficiente de las comparaciones de tratamientos (Casanoves *et al*., 2005).

En el análisis de ensayos en AP la modelación en el marco de los MLM, ha cobrado popularidad (Hong *et al*., 2005; Kravchenko *et al*., 2005; Griffin, 2010; Lawes *et al*., 2012; Rodrigues *et al*., 2013). Los MLM permiten manejar las correlaciones entre observaciones mediante la incorporación de variables aleatorias o mediante la modelación directa de la matriz de covarianzas residual (Balzarini *et al*., 2002). En los ensayos de fertilización que suelen realizarse bajo DBCA profusamente replicados, la inclusión de un efecto aleatorio del bloque puede ser suficiente para modelar la correlación entre parcelas provenientes de un mismo bloque. Igualmente, la correlación de las observaciones provenientes de una misma parcela, se podría modelar mediante la inclusión de un efecto aleatorio de la parcela si se trabaja con los datos de sitio en lugar de promedios de parcela (Balzarini *et al*., 2004)

Cuando la correlación espacial de los errores se presenta a escala de parcelas intrabloque y depende de la distancia de separación entre parcelas (*i.e*. no es constante), el efecto de bloque no es suficiente en la contabilización de la heterogeneidad espacial intrabloque. En este caso, el uso de una función de correlación espacial puede ser beneficioso. Esta función puede aplicarse a los errores de un modelo que incluye el efecto del bloque o sobre un modelo donde el efecto de bloque es eliminado.

En el análisis de correlación espacial para modelar la covariación entre errores de observaciones registradas en diferentes parcelas, es necesario elegir un modelo teórico para la relación entre correlación y distancia. Se puede elegir entre las funciones lineal, esférica, exponencial y Gaussiana, entre otras. Todas con o sin la inclusión de variación que se produce en una escala menor a la distancia entre parcelas (nugget). Existe una relación directa entre estas funciones y la teoría de las variables regionalizadas basadas en semivariogramas (Cressie, 1993) (Capítulo I). Las funciones de correlación para modelos estacionarios de segundo orden pueden ser isotrópicas o anisotrópicas. Las primeras son idénticas en cualquier dirección (sólo dependen de la magnitud de las distancias) mientras que las segundas permiten diferentes valores de parámetros para diferentes direcciones. Los modelos anisotrópicos han demostrado ser eficientes en ensayos agrícolas con arreglos

rectangulares y alta variabilidad en la dirección del bloqueo (Smith *et al*., 2001; Casanoves *et al*., 2005). Sin embargo, en la literatura referida a ensayos en AP correlación espacial es modelada usualmente con modelos espaciales isotrópicos.

La selección de uno u otro modelo de correlación espacial pueden resultar en diferentes valores de *p* para los efectos de tratamiento y diferentes estimaciones de sus medias. Dado que los modelos estadísticos que permiten contemplar la correlación espacial son numerosos (se pueden optar por diferentes funciones de correlación espacial, modelos con y sin efecto nugget, modelos isotrópicos o anisotrópicos y modelos con y sin efecto de bloque), la selección del modelo más apropiado para un conjunto de datos particular es una etapa crucial. Esto puede ser realizado a través de criterios de información como AIC y BIC o mediante la prueba del cociente de verosimilitud (West *et al*., 2007) (Capítulo I).

Mas allá de la selección de un modelo de correlación espacial, en esta tesis se propone que las estrategias de modelado para explorar la respuesta a los tratamientos y sus interacciones con el sitio dentro del lote, pueden ser de dos grandes tipos: 1- análisis de ensayo considerando zonas de manejo previamente delimitadas (modelo de clasificación); 2- análisis de relaciones sitio-específicas entre variables del suelo y la respuesta al tratamiento (modelo de regresión). La estrategia 1 aborda el análisis en dos etapas, primero requiere de la delimitación de zonas de manejo (ZM), *i.e.* áreas con características similares, tales como textura, topografía, estado hídrico y niveles base de nutrientes del suelo (Moral *et al*., 2010) y luego se realizan las comparaciones de tratamientos dentro de cada zona incluyendo la interacción zona × tratamiento en el contexto de un MLM de ANAVA con correlación espacial de los errores. La estrategia 2 se realiza en una sola etapa e implica el planteo de un modelo de regresión del rendimiento específico de cada sitio con las variables de sitio (suelo y/o terreno). En el presente Capítulo se interpretan resultados obtenidos mediante la implementación de la estrategia 1 para el análisis de seis ensayos de fertilización sitio-específica.

MATERIALES Y MÉTODOS

DATOS

Se analizaron datos de seis ensayos en lotes de 30 a 70 hectáreas cada uno, ubicados al sudeste pampeano de la provincia de Buenos Aires, Argentina. Los ensayos se realizaron para evaluar el efecto de diferentes dosis de nitrógeno aplicadas en cultivos de cereales de invierno. En cinco de los ensayos el cultivo evaluado fue trigo y en el restante cebada. El nitrógeno se aplicó en dosis bajas, medias y altas en parcelas de 60 × 30 m en bloques contiguos abarcando la totalidad de la superficie de cada lote. Antes de aplicar los tratamientos de fertilización, se tomaron mediciones intensivas de conductividad eléctrica aparente a 90 cm de profundidad (CE90), elevación y profundidad de tosca (Pe). La medición de la CE90 se realizó con un sensor (Veris 3100, Division of Geoprobe Systems, Salina, KS) que utiliza el principio de la inducción electromagnética. El sensor Veris 3100 recorrió el lote en una serie de transectas paralelas espaciados a intervalos de 15 a 20 m, debido a que una separación de más de 20 m genera errores de medición (Farahani y Flynn, 2007). El instrumento fue calibrado, según las instrucciones del fabricante, antes de la recolección de los datos. Los datos de conductividad eléctrica aparente fueron simultáneamente georreferenciados con un DGPS (Trimble R3, Trimble Navegation Limited, USA) con una exactitud de medición submétrica y configurado para tomar la posición del satélite cada segundo. Los datos de elevación del terreno también se midieron con un DGPS y se procesaron para obtener una precisión vertical de entre 3 y 5 cm aproximadamente. Las mediciones de Pe se realizaron utilizando un penetrómetro hidráulico (Gidding) acoplado a un DGPS en una grilla regular de 30 m. Para cuantificar el rendimiento en grano de los cultivos se utilizó un monitor de rendimiento acoplado a un equipo de cosecha conectados a un DGPS. El registro de la localización espacial de los datos de rendimiento se configuró para tomar la medición con un intervalo de 3 s. Debido a las diferentes resoluciones espaciales y posiciones geográficas donde las variables fueron medidas, los datos de suelo y rendimiento fueron promediados y asignados a cada una de los centroides de parcelas del ensayo utilizando software GIS (ArcGis 9.3.1).

ANÁLISIS DE ENSAYO BAJO UN MODELO DE CLASIFICACIÓN

La estrategia propuesta consiste en delimitar, en una primera etapa, ZM utilizando las variables de sitio (por ejemplo CE90, elevación y Pe). El protocolo de delimitación de ZM propuesto en el Capítulo IV, fue utilizado en la implementación de la estrategia realizada en este Capítulo. Dentro de cada ZM se seleccionan todos los bloques que quedan clasificados sin ambigüedad dentro de la misma ZM o se plantean los bloques si estos no existieran. Para comparar los tratamientos, los datos de rendimiento son analizados usando distintas modificaciones del siguiente modelo básico (6.1).

$$y_{ijk} = \mu + T_i + Z_j + B(Z)_{k(j)} + TZ_{(ij)} + \varepsilon_{ijk} \qquad (6.1)$$

donde y_{ijk} representa el rendimiento observado bajo la dosis de fertilizante *i*, zona de manejo *j*, bloque *k*; μ es la media general de la respuesta; T_i es el efecto (fijo) de la dosis de fertilizante con $i=1,..,t$; Z_j es el efecto (fijo) de la zona de manejo con $j=1,..,z$; $B(Z)_{k(j)}$ es el efecto aleatorio del bloque dentro de la zona de manejo con $k=1,..,b$; $TZ_{(ij)}$ es el efecto de la interacción entre la dosis de fertilizante *i* y la zona de manejo *j* y ε_{ijk} es el término de error aleatorio asociado a la observación y_{ijk} que se considera correlacionado bajo los diferentes modelos de covarianza:

- Solo modelo de bloques aleatorios (RB).
- Solo modelo de correlación espacial (SP).
- Modelo de bloques aleatorios más modelo de correlación espacial (RB+SP).

Para los modelos SP, en la presente implementación, se evaluaron las funciones de correlacion exponencial, gaussiana y esférica sin efecto nugget. Estos siete modelos (RB, SP(Exp), SP(Gau), SP(Esf), RB+SP(Exp), RB+SP(Gau), RB+SP(Esf)), se ajustaron con varianzas residuales homogéneas y con varianzas heterogéneas para las diferentes ZM.

RESULTADOS

En función de la información provista en las tablas 6.1 a 6.6, en los lotes L1, L3, L5 y L6 el modelo con efecto aleatorio de bloque fue suficiente para modelar la correlación espacial. En dos de estos lotes (L1 y L3) fue necesario diferenciar las varianzas residuales dentro de cada zona *i.e.* los modelo de varianzas residuales heterogéneas fueron mejores a los que suponen homogeneidad de varianza. En los dos lotes restantes (L2 y L4) los modelos de correlación espacial ajustaron las correlaciones espaciales subyacentes mejor que el modelo de bloques aleatorios. En ninguno de los seis lotes evaluados el modelo que incluía ambas estrategias de modelado de la correlación espacial (BA+SP) fue el mejor.

Tabla 6.1. Criterios de selección de modelos para el Lote 1.

	Modelos	AIC	BIC	LRT
1	RB	-19.90	-1.60	
2	RB_H	-23.15	-1.52	1 vs. 2*
3	RB+SP(Exp)	-17.90	2.06	
4	RB+SP(Exp)_H	-21.15	2.14	3 vs. 4*
5	RB+SP(Gau)	-17.90	2.06	
6	RB+SP(Gau)_H	-21.15	2.14	5 vs. 6*
7	RB+SP(Esf)	-17.90	2.06	
8	RB+SP(Esf)_H	-21.15	2.14	7 vs. 8*
9	SP(Exp)	-2.56	15.73	
10	SP(Exp)_H	-0.57	21.06	9 vs. 10
11	SP(Gau)	-11.97	6.33	
12	SP(Gau)_H	-8.87	12.76	11 vs. 12
13	SP(Esf)	-6.09	12.21	
14	SP(Esf)_H	-0.19	21.43	13 vs. 14

*Estadísticamente Significativo (α= 0.05). Para AIC y BIC un menor valor indica un mejor ajuste del modelo. RB: bloques aleatorios, RB_H: bloques aleatorios y varianza residual heterogénea, RB+SP (Exp): bloques aleatorios y correlación espacial exponencial, RB+SP (Exp)_H: bloques aleatorios, correlación espacial exponencial y varianza residual heterogénea, RB+SP (Gau): bloques aleatorios y correlación espacial gaussiana, RB+SP (Exp): bloques aleatorios, correlación espacial gaussiana y varianza residual heterogénea, RB+SP (Esf): bloques aleatorios y correlación espacial esférica, RB+SP (Esf): bloques aleatorios, correlación espacial esférica y varianza residual heterogénea.

Tabla 6.2. Criterios de selección de modelos para el Lote 2.

	Modelos	AIC	BIC	LRT
1	RB	20.68	32.65	
2	RB_H	22.67	36.14	1 vs. 2
3	RB+SP(Exp)	22.68	36.15	
4	RB+SP(Exp)_H	24.67	39.64	3 vs. 4
5	RB+SP(Gau)	22.68	36.15	
6	RB+SP(Gau)_H	24.67	39.64	5 vs. 6
7	RB+SP(Esf)	22.68	36.15	
8	RB+SP(Esf)_H	24.68	39.64	7 vs. 8
9	SP(Exp)	20.39	32.37	
10	SP(Exp)_H	22.38	35.85	9 vs. 10
11	SP(Gau)	22.53	34.50	
12	SP(Gau)_H	24.52	37.99	11 vs. 12
13	SP(Esf)	22.46	34.43	
14	SP(Esf)_H	24.45	37.92	13 vs. 14

*Estadísticamente Significativo (α= 0.05). Para AIC y BIC un menor valor indica un mejor ajuste del modelo. RB: bloques aleatorios, RB_H: bloques aleatorios y varianza residual heterogénea, RB+SP (Exp): bloques aleatorios y correlación espacial exponencial, RB+SP (Exp)_H: bloques aleatorios, correlación espacial exponencial y varianza residual heterogénea, RB+SP (Gau): bloques aleatorios y correlación espacial gaussiana, RB+SP (Exp): bloques aleatorios, correlación espacial gaussiana y varianza residual heterogénea, RB+SP (Esf): bloques aleatorios y correlación espacial esférica, RB+SP (Esf): bloques aleatorios, correlación espacial esférica y varianza residual heterogénea.

Tabla 6.3. Criterios de selección de modelos para el Lote 3.

	Modelos	AIC	BIC	LRT
1	RB	220.59	241.35	
2	RB_H	213.28	236.64	1 vs. 2*
3	RB+SP(Exp)	222.59	245.94	
4	RB+SP(Exp)_H	215.28	241.23	3 vs. 4*
5	RB+SP(Gau)	222.59	245.94	
6	RB+SP(Gau)_H	215.28	241.23	5 vs. 6*
7	RB+SP(Esf)	222.59	245.94	
8	RB+SP(Esf)_H	215.28	241.23	7 vs. 8*
9	SP(Exp)	220.67	241.43	
10	SP(Exp)_H	213.50	236.86	9 vs. 10*
11	SP(Gau)	220.67	241.43	
12	SP(Gau)_H	213.31	236.67	11 vs. 12*
13	SP(Esf)	220.67	241.43	
14	SP(Esf)_H	213.50	236.86	13 vs. 14*

*Estadísticamente Significativo (α= 0.05). Para AIC y BIC un menor valor indica un mejor ajuste del modelo. RB: bloques aleatorios, RB_H: bloques aleatorios y varianza residual heterogénea, RB+SP (Exp): bloques aleatorios y correlación espacial exponencial, RB+SP (Exp)_H: bloques aleatorios, correlación espacial exponencial y varianza residual heterogénea, RB+SP (Gau): bloques aleatorios y correlación espacial gaussiana, RB+SP (Exp): bloques aleatorios, correlación espacial gaussiana y varianza residual heterogénea, RB+SP (Esf): bloques aleatorios y correlación espacial esférica, RB+SP (Esf): bloques aleatorios, correlación espacial esférica y varianza residual heterogénea.

Tabla 6.4. Criterios de selección de modelos para el Lote 4.

	Modelos	AIC	BIC	LRT
1	RB	122.45	139.59	
2	RB_H	123.97	143.25	1 vs. 2
3	RB+SP(Exp)	124.45	143.74	
4	RB+SP(Exp)_H	125.97	147.40	3 vs. 4
5	RB+SP(Gau)	124.45	143.74	
6	RB+SP(Gau)_H	125.97	147.40	5 vs. 6
7	RB+SP(Esf)	124.43	143.71	
8	RB+SP(Esf)_H	125.95	147.38	7 vs. 8
9	SP(Exp)	122.45	139.60	
10	SP(Exp)_H	124.02	143.31	9 vs. 10
11	SP(Gau)	122.23	139.37	
12	SP(Gau)_H	123.75	143.03	11 vs. 12
13	SP(Esf)	122.41	139.56	
14	SP(Esf)_H	123.99	143.28	13 vs. 14

*Estadísticamente Significativo (α= 0.05). Para AIC y BIC un menor valor indica un mejor ajuste del modelo. RB: bloques aleatorios, RB_H: bloques aleatorios y varianza residual heterogénea, RB+SP (Exp): bloques aleatorios y correlación espacial exponencial, RB+SP (Exp)_H: bloques aleatorios, correlación espacial exponencial y varianza residual heterogénea, RB+SP (Gau): bloques aleatorios y correlación espacial gaussiana, RB+SP (Exp): bloques aleatorios, correlación espacial gaussiana y varianza residual heterogénea, RB+SP (Esf): bloques aleatorios y correlación espacial esférica, RB+SP (Esf): bloques aleatorios, correlación espacial esférica y varianza residual heterogénea.

Tabla 6.5. Criterios de selección de modelos para el Lote 5.

	Modelos	AIC	BIC	LRT
1	RB	3.43	16.74	
2	RB_H	5.33	20.31	1 vs. 2
3	RB+SP(Exp)	5.43	20.40	
4	RB+SP(Exp)_H	7.33	23.97	3 vs. 4
5	RB+SP(Gau)	5.43	20.40	
6	RB+SP(Gau)_H	7.33	23.97	5 vs. 6
7	RB+SP(Esf)	5.40	20.37	
8	RB+SP(Esf)_H	7.31	23.94	7 vs. 8
9	SP(Exp)	15.79	29.10	
10	SP(Exp)_H	15.16	30.14	9 vs. 10
11	SP(Gau)	19.41	32.72	
12	SP(Gau)_H	18.89	33.86	11 vs. 12
13	SP(Esf)	19.03	32.33	
14	SP(Esf)_H	13.94	28.91	13 vs. 14*

*Estadísticamente Significativo (α= 0.05). Para AIC y BIC un menor valor indica un mejor ajuste del modelo. RB: bloques aleatorios, RB_H: bloques aleatorios y varianza residual heterogénea, RB+SP (Exp): bloques aleatorios y correlación espacial exponencial, RB+SP (Exp)_H: bloques aleatorios, correlación espacial exponencial y varianza residual heterogénea, RB+SP (Gau): bloques aleatorios y correlación espacial gaussiana, RB+SP (Exp): bloques aleatorios, correlación espacial gaussiana y varianza residual heterogénea, RB+SP (Esf): bloques aleatorios y correlación espacial esférica, RB+SP (Esf): bloques aleatorios, correlación espacial esférica y varianza residual heterogénea.

Tabla 6.6. Criterios de selección de modelos para el Lote 6.

	Modelos	AIC	BIC	LRT
1	RB	19.46	30.95	
2	RB_H	21.38	34.96	1 vs. 2
3	RB+SP(Exp)	21.46	34.00	
4	RB+SP(Exp)_H	23.38	38.01	3 vs. 4
5	RB+SP(Gau)	21.46	34.00	
6	RB+SP(Gau)_H	23.38	38.01	5 vs. 6
7	RB+SP(Esf)	21.46	34.00	
8	RB+SP(Esf)_H	23.38	38.01	7 vs. 8
9	SP(Exp)	20.44	31.93	
10	SP(Exp)_H	21.68	35.26	9 vs. 10
11	SP(Gau)	21.26	32.75	
12	SP(Gau)_H	22.53	36.11	11 vs. 12
13	SP(Esf)	20.63	32.12	
14	SP(Esf)_H	21.77	35.35	13 vs. 14

*Estadísticamente Significativo (α= 0.05). Para AIC y BIC un menor valor indica un mejor ajuste del modelo. RB: bloques aleatorios, RB_H: bloques aleatorios y varianza residual heterogénea, RB+SP (Exp): bloques aleatorios y correlación espacial exponencial, RB+SP (Exp)_H: bloques aleatorios, correlación espacial exponencial y varianza residual heterogénea, RB+SP (Gau): bloques aleatorios y correlación espacial gaussiana, RB+SP (Exp): bloques aleatorios, correlación espacial gaussiana y varianza residual heterogénea, RB+SP (Esf): bloques aleatorios y correlación espacial esférica, RB+SP (Esf): bloques aleatorios, correlación espacial esférica y varianza residual heterogénea.

Bajo los modelos seleccionados se valuaron los efectos de tratamiento, ZM e interacción tratamiento × ZM. Para los lotes L1, L3, L4 y L5 el efecto de la interacción tratamiento × ZM resultó significativo ($p<0.05$) indicando que la respuesta a la fertilización no es la misma en las diferentes ZM. En el L2 y L6, donde la interacción no fue estadísticamente significativa, tampoco se detectaron diferencias entre las zonas. Si bien el efecto de la dosis de fertilización fue significativo, sugiriendo una respuesta casi lineal a los incrementos de N (80, 120 y 170 kg N/ha), las diferencias fueron equivalentes en las ZM delimitadas.

Analizando los cuatro de los seis casos donde el manejo sitio-específico podría ser beneficioso (*i.e.* casos con interacción tratamiento × ZM significativa) se observa que en el L1 las mayores diferencias entre las tres dosis evaluadas se presentaron en la ZM que tuvo el menor rendimiento promedio. Esta zona tuvo un incremento del rendimiento de 410 kg/ha entre la dosis más baja (80 kg N/ha) y la dosis más alta (170 kg N/ha) esto representa un 12.5% de aumento del rendimiento debido a la aplicación de la fertilización. La ZM3 fue la que tuvo el mayor rendimiento promedio. En esta zona el efecto de la fertilización fue significativo pero menor al de la ZM1. El incremento del rendimiento entre la dosis más baja y el promedio de las dos dosis más altas (no presentaron diferencia significativas entre ellas) fue de 240 kg/ha (6%).

Para el L3 los tratamientos no presentaron diferencias significativas en la ZM de más bajo rendimiento. Mientras que en la ZM2 las dosis de N media (125 kg N/ha) y alta (170 kg N/ha) fueron estadísticamente diferentes entre ellas y respecto a la dosis baja (97 kg N/ha). En esta zona la diferencia de rendimiento entre la dosis baja y media fue de 914 kg/ha (28%). Pero, cuando se utilizó la dosis más alta se produjo un menor incremento del rendimiento (558 kg/ha, 17%).

En el L4 la ZM de menor rendimiento promedio, tanto con 100 kg N/ha como con 200 kg N/ha se observaron diferencias respecto al testigo (0 kg N/ha). El incremento de rendimiento por aplicar fertilizante fue de 1280 kg/ha en promedio. Mientras que en la ZM2 el incremento de rendimiento cuando se aplicaron 100 kg N/ha fue de 951 kg/ha (19%) y, cuando se aplicaron 200 kg N/ha el aumento del rendimiento fue menor (565 kg/ha, 11%).

En el L5, el incremento de la dosis de fertilizante se evidenció siempre con aumentos de rendimiento en la ZM1 (de menor rendimiento promedio), mientras que en la ZM2 las dosis de 125 y 170 kg N/ha no difirieron entre ellas. En la ZM1 el incremento del rendimiento cuando se utilizan dosis medias y altas de fertilizante fueron de 512 kg/ha (14%) y 687 kg/ha (18%) respectivamente. Mientras que en la ZM2 el aumento del rendimiento entre la dosis más baja y el promedio de las dos dosis más altas fue de 899 kg/ha (22%).

Para caracterizar las ZM de los lotes en los que la zonificación se relacionó con diferencias de rendimiento estadísticamente significativa (L1, L3, L4 y L5), se realizó un análisis de componentes principales (PCA) utilizando las variables de sitio (CE90, elevación y Pe) (Fig. 6.3). En el L1 y en el L5, la ZM de rendimiento promedio más alto se caracterizó por presentar mayor profundidad del suelo (tosca) y menor elevación, es decir, se corresponde con los sectores más bajos del lote. Mientras que en el L3 la ZM de menor rendimiento promedio, se caracteriza por presentar menor valor de CE90 y mayor variabilidad topográfica. En el L4 la ZM de menor rendimiento, también presenta alta variabilidad de la Pe, pero también esta variabilidad se produce con la CE90.

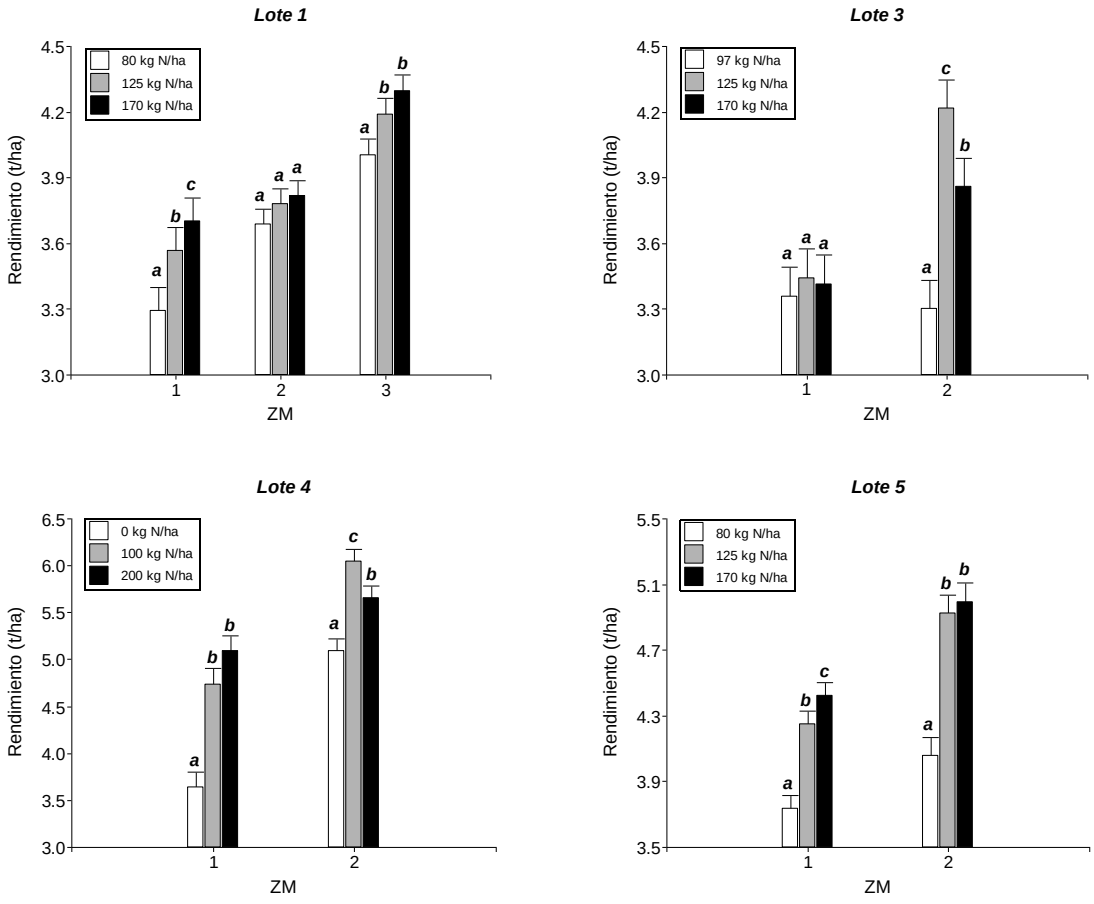


Fig. 6.1. Rendimientos promedios de acuerdo a dosis de nitrógeno y zona de manejo. Letras diferentes indican diferencias estadísticamente significativas ($p<0.05$).

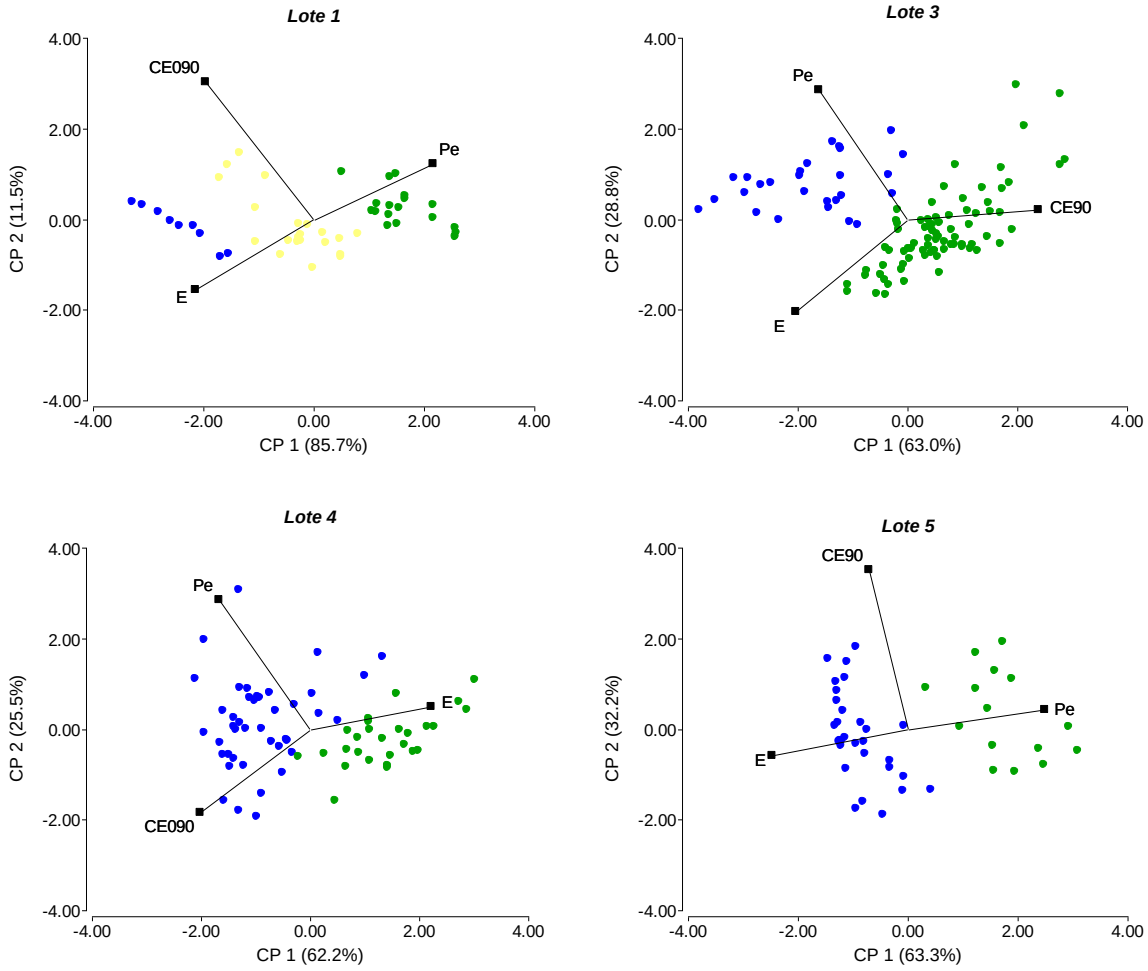


Fig. 6.2. Análisis de compoentes principales de variables de sitio. CE90: Conductividad eléctrica aparente a 90 cm de profundidad, Pe: profundidad de tosca, E: elevación.

CONCLUSIÓN

La estrategia de modelación realizada en el contexto en el contexto de un MLM de ANAVA mostró que en cuatro de seis ensayos la inclusión de un efecto de bloques aleatorio fue suficiente para modelar la correlación espacial entre parcelas provenientes de un mismo bloque (correlación inducida). Mientras que en los dos lotes restantes la correlación espacial estuvo presente a una escala que bloqueo no pudo controlar, por tanto, resultó más beneficioso el uso de una función de correlación espacial (correlación explícita). Los resultados de los modelos seleccionados mostraron que en cuatro de los seis ensayos sería posible la aplicación diferencial de nitrógeno, mientras que en los dos

ensayos restantes será necesario profundizar si las variables utilizadas en la delimitación de zonas de manejo son las adecuadas para rescatar los principales factores que generan variabilidad espacial del rendimiento del cultivo.

CONCLUSIONES

En los últimos años se han difundido, en Argentina, la utilización de tecnologías de agricultura de precisión (AP) que permiten medir y, consecuentemente manejar diferenciadamente, la reconocida variabilidad espacial intralote tanto de propiedades del sitio (suelo, terreno, nutrientes) como de los rendimientos. Junto al incremento del número de cosechadoras con tecnologías de mapeo del rendimiento hubo un aumento en el uso de sensores proximales que captura datos de conductividad eléctrica aparente y otras propiedades edáficas y del terreno. Uno de los requerimientos centrales de la AP es la obtención de zonas de manejo (ZM) definidas por factores limitantes del rendimiento, que luego podrán ser manejadas de acuerdo a sus propiedades intrínsecas. Para la delimitación de ZM se recomienda utilizar sensores proximales a partir de los cuales se puedan obtener mediciones no solo del rendimiento de cada sitio sino también de los factores potencialmente limitantes del rendimiento. De esta forma los fertilizantes nitrogenados y otros insumos, pueden ser aplicados en forma variable lo que permite maximizar la productividad, la sustentabilidad del sistema y la protección del medio ambiente.

El óptimo uso de la información espacial disponible, la cual es cada vez más abundante, demanda nuevos desarrollos estadístico-computacionales. Los datos espaciales, georreferenciados o regionalizados presentan autocorrelación espacial, característica que les confiere una estructura de dependencia que puede ser explicada a través de modelos cuyos parámetros contienen información significativa para el manejo de los sistemas agrícolas. Por el contrario los modelos lineales clásicos, para datos independientes, son inapropiados para datos correlacionados espacialmente. Por ello, las dependencias espaciales de las propiedades del sitio y de los rendimiento de los cultivos han sido profusamente abordadas a través de modelos geoestadísticos clásicos y, últimamente, a través de modelos estadísticos contemporáneos entre los que se destacan los modelos lineales mixtos (MLM) de covarianza espacial. Aún cuando las diferencias en las estimaciones logradas con la aproximación basada en técnicas geoestadísticas o con MLM no fueran grandes, la utilización de MLM presenta claras ventajas. Utilizando MLM es posible modelar la correlación espacial y la tendencia a gran escala en un solo paso, mientras que cuando se utilizan técnicas geoestadísticas es necesario, previo a realizar el

análisis espacial, evaluar la presencia de tendencia a gran escala mediante el ajuste de modelos de regresión. Luego, podría ser necesario trabajar con los residuales de esa regresión. Para la selección de modelos cuando se trabaja en el contexto de los MLM es posible utilizar diferentes herramientas estadísticas para elegir el modelo de mejor ajuste, mientras que con las técnicas geoestadísticas la selección no tiene tanta base estadística. Adicionalmente, con los MLM es posible obtener medias ajustadas por el modelo de correlación e incluso por heterocedasticidades, mientras que en los análisis geoestadisticos clases no solo se trabaja con residuos sino que frecuentemente se suponen varianza homogéneas. Estas medias pueden resultar diferentes a las obtenidas para la variable sin los ajustes.

Los métodos antes mencionados han sido frecuentemente utilizados en el contexto univariado. No obstante, en la actualidad los datos disponibles son multivariados ya que usualmente, se registran varias variables de cada sitio. Las técnicas multivariadas facilitan la interpretación de complejas relaciones entre variables, reducen la dimensión de la base de datos para mapear la variabilidad espacial, permiten detectar estructuras y revelan nuevas relaciones espaciales que pueden no ser evidentes cuando las variables de sitio se analizan individualmente. Los análisis multivariados de *cluster fuzzy-kmeans* y componentes principales (PCA), resultan apropiados para la visualización y exploración simultanea de datos de varias variables regionalizadas. Sin embargo, estas técnicas, no han sido desarrolladas explícitamente para manejar datos espaciales. Un caso distinto es el de la técnica multivariada MULTISPATI-PCA que fue diseñada para contemplar las relaciones entre las variables y su estructura espacial (autocorrelación). Este se basa en el PCA pero incorpora la restricción espacial mediante el cálculo del índice de autocorrelación espacial de Moran. En esta tesis cuando se utilizó el análisis de *cluster fuzzy-kmeans* los índices para la selección del número óptimo de clases no fueron coincidentes. En estos casos, la utilización de un índice que resuma la información de varios índices mostró ser de gran utilidad. Por otra parte, la clasificación obtenida mostró clases fragmentadas que requerirán de un suavizado para la obtención de potenciales zonas de manejo. En la comparación de MULTISPATI-PCA vs. PCA, utilizando el análisis restringido espacialmente la selección del número de componentes principales para la interpretación de la variabilidad fue más clara. Los resultados mostraron que la incorporación de la autocorrelación espacial en el análisis permite detectar relaciones

subyacentes que no serían tenidas en cuenta si la estructura espacial fuera incorporada a *posteriori*. El grado de estructuración espacial fue mayor con MULTISPATI-PCA que con PCA no restringido espacialmente. Esto se evidenció en los mapas de las variables sintéticas donde la estructura espacial fue más clara con las componentes principales espaciales de MULTISPAI-PCA. El suavizado logrado en los mapas generó zonas más definidas o menos fragmentadas, lo cual constituye un aspecto clave para la delimitación de zonas de manejo en agricultura sitio-específica. Este aspecto constituyó el antecedente para el desarrollo de una nueva metodológica para la clasificación multivariada de sitios intralote que se denominó KM-sPC. El algoritmo propuesto, utiliza el análisis de *cluster fuzzy k-means* e incluye una dimensión espacial a través de la utilización de las componentes principales espaciales (MULTISPATI-PCA) como variables de clasificación. Los resultados de su aplicación tanto en datos experimentales como simulados, mostraron que se mejoró el desempeño del método de agrupación no espacial en la formación de clases dentro del lote. A su vez, KM-sPC aplicado sobre variables del suelo y del terreno delimitaron clases de sitios con las mayores diferencias en el rendimiento respecto a los métodos sin restricciones espaciales aplicado en los mismos datos. KM-sPC no sólo maximizó las diferencias en rendimiento entre las clases delimitadas sino que también permitió identificar cuáles son las variables de mayor contribución en la explicación de la variabilidad espacial. El procedimiento propuesto es adecuado para grandes conjuntos de datos multivariados y no requiere un ajuste previo de un modelo de variograma. En esta tesis se desarrolló un protocolo diseñado para el análisis estadístico de datos de sitio intralote con la finalidad de delimitar zonas homogéneas que podrían potencialmente ser utilizadas como zonas de manejo para agricultura sitio-especifica. Se presentan los scripts para realizar el pre-procesamiento de los datos hasta la delimitación de zonas de manejo y la evaluación de prácticas de manejo sitio-específicas.

Un aspecto importante respecto a la delimitación de zonas de manejo es que los datos de suelo pueden obtenerse por sitio en varios años, por lo cual la zonificación podría depender no sólo de la variabilidad espacial sino también de la variabilidad temporal de las mediciones. La tendencia promedio de los cambios de las variables medidas subsecuentemente en distintos años puede también ser analizada con MLM. Los modelos espacio-temporales resultan promisorios para la obtención de estimaciones de productividad más precisas, ya que integran mayor cantidad de información que los

modelos aplicados para análisis de datos transversales o de una única campaña agrícola. La combinación de técnicas multivariadas con técnicas geoestadísticas clásicas y modelos espacio-temporal surge como la mejor alternativa metodológica para comprender mejor series temporales de datos espaciales en AP. La complementación de los métodos y modelos mencionados es indispensable no sólo para la identificación de ZM, sino también para dilucidar y manejar los mecanismos implícitos en la generación de rendimientos en cada uno de las zonas delimitadas. Se propuso una metodología de análisis para la clasificación de sitios intralote en función de la variación espacio-temporal de variables de suelo. El algoritmo se basó en el análisis de la tendencia interanual promedio de sitios mediante un MLM, la estimación de la varianza temporal por sitio para cada variable y la delimitación de clases de sitios considerando tanto la variabilidad espacial como la estabilidad de la misma. Para ilustrar su aplicación, se analizó la tendencia interanual promedio y la variabilidad temporal de MO, P, pH y CE en lotes de la región semiárida, abarcando una superficie de 2.240 ha bajo agricultura intensiva. Los resultados mostraron que la variación espacial en las características de suelo no es permanente, produciéndose significativos cambios en la delimitación de zonas homogéneas a través de los años. Por ello, el manejo sitio-específico debiera realizarse acorde a las condiciones o zonificaciones emergentes en cada año y/o utilizando en la zonificación variables de sitio cuyo patrones espaciales se mantegan estables en el tiempo.

En el contexto de la AP, no solo es necesario delimitar las ZM sino también identificar cuáles son los factores limitantes del rendimiento y ensayar posibles manejos sitio-específicos. Por ello, se realizan recomendaciones respecto a la modelación estadística contemporánea orientada al análisis de ensayos de fertilización sitio-específica. La estrategia de modelación se realiza en el contexto de un modelo de clasificación (ANAVA) para comparar el efecto de la fertilización precisa en seis ensayos de fertilización nitrogenada. El procedimiento incluye primero la delimitación de ZM y luego el ajuste de un MLM con efectos fijos de tratamiento, zona e interacción tratamiento-zona y efecto aleatorio de bloque dentro de cada zona. En cuatro de seis ensayos la inclusión de un efecto de bloques aleatorio fue suficiente para modelar la correlación espacial entre parcelas provenientes de un mismo bloque (correlación inducida). Mientras que en los dos lotes restantes la correlación espacial estuvo presente a una escala que bloqueo no pudo controlar (*i.e.* presencia de heterogeneidad intrabloque), por tanto, resultó más beneficioso

el uso de una función de correlación espacial (correlación explícita). Los resultados de los modelos seleccionados mostraron que en cuatro de los seis ensayos sería posible la aplicación diferencial de nitrógeno, mientras que en los dos ensayos restantes será necesario profundizar si las variables utilizadas en la delimitación de zonas de manejo son las adecuadas para rescatar los principales factores que generan variabilidad espacial del rendimiento del cultivo.

Finalmente, caben destacar que futuras líneas de investigación podrían estar orientada al desarrollo nuevos métodos estadísticos para la delimitación de ZM a partir de múltiples variables de sitio, que incorporen árboles de regresión para obtener valores umbrales de las variables de suelo georreferenciadas con potencialidad para predecir rendimientos y análisis de *cluster* de los sitios con algoritmos restringidos espacialmente. La teoría de las variables co-regionalizadas ofrece una alternativa, que no ha sido explotada en esta tesis, para analizar covariación espacial de dos variables georreferenciadas. En cuanto a la modelación estadística, sería interesante evaluar las estimaciones de parámetros de modelos que permitan predecir la productividad potencial sitio-específica, obtenidos en el marco de los modelos lineales mixtos espacio-temporales, así como la capacidad predictiva de funciones desarrolladas en algunos tipos de suelos.

BIBLIOGRAFÍA

Achouri M. and Gifford G.F. 1984. Spatial and seasonal variability of field measured infiltration rates on a rangeland site in Utah. J. Range Manage. 37: 451–455.

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle, in 2nd International Symposium on Information Theory and Control, Petrov, E.B.N. and Csaki, F., (ed.), pp. 267.

Alesso C.A., Pilatti M.A., Imhoff S. y Grilli M. 2012. Variabilidad espacial de atributos químicos y físicos en un suelo de la pampa llana santafesina. Ciencia del Suelo 30(1): 85–93.

Anderberg M.R. 1973. Cluster analysis for applications. Academic Press, Inc., New York.

Anselin L. 1995. Local indicators of spatial association – LISA. Geographical Analysis, 27: 93-115.

Anselin L. 1996. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. En Fischer M., Scholten H., and Unwin D., (ed.), Spatial analytical perspectives on GIS, p. 111-125. Taylor and Francis, London.

Anselin L. 2001. Spatial Effects in Econometric Practice in Environmental and Resource Economics. Am. J. Agric. Econ.83 (3): 705–710.

Arno J., Martinez-Casasnovas J.A., Ribes-Dasi M. and Rosell J.R. 2011. Clustering of grape yield maps to delineate site-specific management zones. Span. J. Agric. Res. 9: 721–729.

Arrouays D., Saby N.P.A., Thioulouse J., Jolivet C., Boulonne L. and Ratié C. 2011. Large trends in French topsoil characteristics are revealed by spatially constrained multivariate analysis. Geoderma 161, 107–114.

Babai L., Cooperman G., Finkelstein L., Luks E. and Seress Á. 1995. Fast Monte Carlo algorithms for permutation groups. J. Comp. Syst. Sci. 50(2): 296-308.

Balzarini M. 2002. Applications of Mixed Models in Plant Breeding. En: Quantitative Genetics, Genomics, and Plant Breeding. Kang, M.S. (ed.) CABI Publishing.

Balzarini M., Di Rienzo J., Tablada M., Gonzalez L., Bruno C., Córdoba M., Robledo W. y Casanoves F. 2012. Estadística y Biometría. Ilustraciones del uso de InfoStat en problemas de Agronomía. Ed. Brujas. 402 pp.

Balzarini M., Macchiavelli R. y Casanoves F. 2004. Aplicaciones de Modelos Mixtos en Agricultura y Forestería. Curso de Capacitacion Centro Agronomico Tropical de Investigación y Enseñanza - CATIE, 210 p.

Balzarini M., Teich I., Brun C. and Peña A. 2011. Making genetic biodiversity measurable: a review of statistical multivariate methods to study variability at gene level. Rev. FCA UNCUYO 43(1): 261-275.

Barbieri P., Echeverría H. and Sainz Rozas H. 2009. Nitrates in soil at planting or tillering as a diagnostic of the nitrogenated nutrition in wheat in the Southeastern Pampas. Soil Sci. 27: 41–47.

Ben-Dor E., Chabrillat S., Dematte J.A.M., Taylor G.R., Hill J., Whiting M.L. and Sommer S. 2009. Using imaging spectroscopy to study soil properties. Remote Sens. Environ. 113: 38–55.

Besag J. E. 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). J. Roy. Statist. Soc. Ser. B. 36: 192–225.

Besag J.E. 1977. Errors-in-variables estimation for Gaussian lattice schemes. J. Roy. Statist. Soc. Ser. B. 39: 73–78.

Bezdek J.C. 1981. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Bishop T.F. and Lark R.M. 2006. The geostatistical analysis of experiments at the landscape-scale. Geoderma, 133(1-2): 87–106.

Bivand R. 2008. Implementing representations of space in economic geography. J. Reg. Sci. 48(1): 1–27.

Bivand R., Keitt T. and Rowlingson B. 2013b. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.8-10. http://CRAN.R-project.org/package=rgdal

Bivand R., with contributions by Altman M., Anselin L., Assunção R., Berke O., Bernat A., Blanchet G., Blankmeyer E., Carvalho M., Christensen B., Chun Y., Dormann C., Dray S., Halbersma R., Krainski E., Legendre P., Lewin-Koh N., Li H., Ma J., Millo G., Mueller W., Ono H., Peres-Neto P., Piras G., Reder M., Tiefelsdorf M. and Yu D. 2013a. spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.5-56. http://CRAN.R-project.org/package=spdep.

Blackmore B.S., Godwin R.J. and Fountas S. 2003. The analysis of spatial and temporal trends in yield map data over six years. Biosyst. Eng. 84(4): 455–466.

Bongiovanni R., Montovani E.C., Best S. y Roel, A. 2006. Agricultura de precisión: Integrando conocimientos para una agricultura moderna y sustentable. PROCISUR/ IICA, Montevideo, 246 pp.

Bongiovanni R.G., Robledo C.W. and Lambert D.M. 2007. Economics of site-specific nitrogen management for protein content in wheat. Comp. Electron. Agric. 58: 13–24.

Bourgault G., Marcotte D. and Legendre P. 1992. The multivariate (co)variogram as a spatial weighting function in classification methods. Math. Geol. 24: 463–478.

Boydell B. and McBratney A.B, 2002. Identifying Potential Within-Field Management Zones from Cotton-Yield Estimates. Precis. Agric. 3: 9–23.

Bragachini M., Méndez A. y Vélez J.P. 2011. Argentina, un referente mundial en tecnología de Agricultura de Precisión. INTA, Manfredi, 6 pp.

Bremner J.M. 1965. Inorganic forms of nitrogen. In: Black, C.A. (Ed.), Methods of soil analysis, Madison, Wisconsin, pp. 1179–1237.

Brownie C. and M.L. Gumpertz. 1997. Validity of spatial analyses for large field trials. J. Agric. Biol. Environ. Stat. 2: 1–23.

Bullock D. S., Kitchen N.,Bullock D. G. 2007. Multidisciplinary Teams: A Necessity for Research in Precision Agriculture Systems. Crop Sci. 47:1765–1769.

Bullock D.S. and Lowenberg-DeBoer J. 2007. Using Spatial Analysis to Study the Values of Variable Rate Technology and Information. JAE 58(3): 517–535.

Burgess T.M. and Webster R. 1980. Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. J. Soil Sci. 31: 315–331.

Burgos J.J. and Vidal A.L., 1951. The climates of the Argentine republic according to the new Thornthwaite classification. Ann. Assoc. Am. Geogr. 41: 237–263.

Burrough P.A. 1989. Fuzzy mathematical methods for soil survey and land evaluation. J. Soil Sci. 40, 477–492.

Buschiazzo D. 1986. Estudio sobre la tosca. Parte I: Evidencias de un movimiento descendente del carbonato en base a la interpretación de características macro y geomorfológicas. Ciencia del Suelo 4: 55-65.

Casanoves F., Baldessari J., and Balzarini M. 2005. Evaluation of multi-environmental trials of peanut (Arachis hypogaea L.) cultivars. Crop Sci. 45:18-26

Chessel D., Dufour A.B. and Thioulouse J. 2004. The ade4 package-I- One-table methods. R News 4: 5–10.

Cliff A.D. and Ord J.K. 1973. Spatial autocorrelation. Pion, London.

Clifford P., Richardson S. and Hémon D. 1989. Assessing the significance of the correlation between two spatial processes. Biometrics 45: 123–134.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20: 37-46.

Cook S.E. and Bramley R.G.V. 1998. Precision agriculture - opportunities, bene-fits and pitfalls of site-specific crop management in Australia. Aust. J. Exp. Agric. 38: 753-763.

Córdoba M., Bruno C., Costa J. y Balzarini M. 2011. Geoestadística multivariada. En agricultura de precisión. 40° Jornadas Argentinas de Informática, 3° Congreso Argentino de AgroInformática. Córdoba. p. 32-42.

Corwin D.L. and Lesch S.M. 2005. Apparent soil electrical conductivity measurements in agriculture, Comp. Electron. Agric. 46: 11–43.

Corwin D.L. and Lesch S.M. 2010. Delineating site-specific management units with proximal sensors, En: Margaret, O. (Ed.), Geostatistical Applications in Precision Agriculture. Springer, New York, pp. 139–165.

Corwin D.L., Lesch S.M., Oster J.D., Kaffka S.R. 2006. Monitoring management-induced spatio–temporal changes in soil quality through soil sampling directed by apparent electrical conductivity. Geoderma 131: 369–387.

Cover T. M. and Hart P.E. 1967. Nearest Neighbor Pattern Classification. IEEE Trans. Inf. Theory. 13: 21–27.

Cressie N.A.C. 1993. Statistics for Spatial Data Revised Edition. John Wiley and Sons, New York, 900 pp.

Cressie N.A.C.1985. Fitting variogram models by weighted least squares. Math. Geol. 17:563-586.

Cullis B.R., Gogel B.J., Verbyla A.P. and Thompson R. 1998. Spatial analysis of multi-environment early generation trials. Biometrics 54: 1–18.

Davatgar N., Neishabouri M.R., Sepaskhah A.R. 2012. Delineation of site specific nutrient management zones for a paddy cultivated area based on soil fertility using fuzzy clustering. Geoderma 173-174: 111–118.

Davidoff B. and Selim H.M. 1988. Correlation between spatially variable soil moisture content and soil temperature. Soil Sci. 145: 1–10.

Demidenko E. 2004. Mixed Models: Theory and Applications. John Wiley and Sons, New Jersey.

Dewis J. and Freitas F. 1970. Métodos físicos y químicos de análisis de suelos y aguas. FAO. Boletín sobre Suelos Nº 10, 252p.

Diacono M., Castrignanò A., Troccoli A., De Benedetto D., Basso B. and Rubino P. Spatial and temporal variability of wheat grain yield and quality in a Mediterranean environment: A multivariate geostatistical approach. 2012. F. Crop Res. 131:49–62.

Diggle P.J. and Ribeiro Junior P.J. 2007. Model-based geostatistics. Springer, New York, 228 pp.

Diggle P.J., Moyeed R.A. and Tawn J.A. 1998. Model-based geostatistics (with discussion). Applied Statistics 47: 299–350.

Diker K., Heermann D.F. and Brodahl M.K. 2004. Frequency analysis of yield for delineating yield response zones. Precis. Agric. 5: 435-444.

Dillon C.R., Saghaian S., Salim J. and Kanakasabai M. 2005. Optimal water storage location and management zone delineation under variable subsurface drip irrigation. p. 959–965. In J.V. Stafford (ed.) Precision agriculture ’05: Proc. 5th Conf. on Precision Agriculture, Uppsala, Sweden, 8–11 June 2005. Wageningen Academic Publishers,Wageningen, The Netherlands.

Dray S. and Jombart T. 2011. Revisiting Guerry's data: Introducing spatial constraints in multivariate analysis. Ann. Appl. Stat. 5(4): 2265–2687

Dray S., Chessel D. and Thioulouse J. 2003. Co-Inertia Analysis and the Linking of Ecological Data Tables. Ecology. 84: 3078–3089.

Dray S., Saïd S. and Débias F. 2008. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. J. Veg. Sci. 19:45-56.

Eastman J. R. 2009. IDRISI 16: The Taiga Edition (Worcester, MA: Clark University).

Efron B. and Tibshirani R. 1993. Bootstrap methods for standard errors, confidence intervals, and other methods of statistical accuracy. Stat. Sci., 1(1): 54-77.

Espósito G., Robledo W., Bongiovanni R., Ruffo M., Diez E. y Balboa G. 2012. Análisis del efecto año sobre la dosis variable de nitrógeno en maíz. XIX Congreso Lantinoamericano y XXIII Congreso Argentino y Latino Americano de la Ciencia del Suelo. Mar del Plata. Buenos Aires. Argentina.

FAO, FIDA y PMA. 2012. El estado de la inseguridad alimentaria en el mundo 2012. El crecimiento económico es necesario pero no suficiente para acelerar la reducción del hambre y la malnutrición. Roma, FAO.

Fleming K.L., Heermann D.F., Westfall D.G. 2004. Evaluating Soil Color with Farmer Input and Apparent Soil Electrical Conductivity for Management Zone Delineation. Agron. J. 96: 1581–1587

Fleming K.L., Westfall D.G., Wiens D.W. and Brodahl M.C. 2000. Evaluating farmer defined management zones for variable rate fertilizer application. Precis. Agric. 2:201–215.

Flowers M., Weisz R. and White J.G. 2005. Yield-Based Management Zones and Grid Sampling Strategies. Agron. J. 97:968–982.

Fraisse C., Sudduth K. y Kitchen N. 2001. Delineation of site-specific management zones by unsupervised classification of topographic attributes and soil electrical conductivity. Transactions of the ASAE. 44: 155–166.

Franke R. 1982. Scattered Data Interpolation: Test of Some Methods, Math. Comp. 38: 181–200

Franzen D.W., Hopkins D.H., Sweeney M.D., Ulmer M.K. and Halvorson A.D. 2002. Evaluation of soil survey scale for zone development of site-specific nitrogen management. Agron. J. 94: 381–389.

Fridgen J.J., Kitchen N.R. and Sudduth K.A. 2000. Variability of soil and landscape attributes within sub-field management zones. 16 p. In P.C. Robert, R.H. Rust, and W.E. Larson (ed.) Precision Agriculture: Proc. 5th Int. Conf. on Precision Agriculture, ASA, Madison, WI.

Fridgen J.J., Kitchen N.R., Sudduth K.A., Drummond S.T., Wiebold W.J. and Fraisse C.W. 2004. Management Zone Analyst (MZA): Software for Subfield Management Zone Delineation. Agron. J. 96: 100–108.

Frogbrook Z.L. and Oliver M.A. 2007. Identifying management zones in agricultural fields using spatially :constrained classification of soil and ancillary data. Soil Use Manage. 23: 40–51.

Fu W., Tunney H. and Zhang C. 2010. Spatial variation of soil nutrients in a dairy farm and its implications for site-specific fertilizer application. Soil Till. Res. 106(2): 185–193.

Fukuyama Y. and Sugeno M. 1989. A new method of choosing the number of clusters for the fuzzy c-means method, Proc. 5th Fuzzy Syst. Symp., p. 247-250, 1989

Gabriel K.R. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika 58: 453–467.

Gabriel K.R. and Sokal R.R. 1969. A New Statistical Approach to Geographic Variation Analysis. Syst. Zool. 18: 259-278.

Galarza R., Mastaglia N., Albornoz E.M. y Martínez C.E. 2013. Identificación automática de zonas de manejo en lotes productivos agrícolas. V Congreso Argentino de Agroinformática (CAI) - 42da. JAIIO, Córdoba.

Gbur E.E., Stroup W.W., McCarter K.S., Durham S., Young L.J., Christman M., West M. and Kramer M. 2012. Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences, ASA, CSSA, SSSA, Madison, WI.

Geary R. 1954. The contiguity ratio and statistical mapping. The Incorporated Statistician. 5: 115-145.

Gili A.A. 2013. Modelación de la variación espacial de variables edáficas y su aplicación en el diseño de planes de muestreo de suelos. Tesis Doctorado, Universidad Nacional de Córdoba, Córdoba, Argentina. 193 pp.

Gilmour A.R., Thompson R, Cullis B.R. and Verbyla A.P. 1997. Accounting for natural and extraneous variation in the analysis of field experiments. J. Agric. Biol. Envir. S. 2: 269:273.

Giraldo H.R. 2003. Introducción a la Geoestadística. Bogotá: Universidad Nacional de Colombia, 94 pp.

Gonzalez R.C. and Woods R. 2008. Digital Image Processing. Pearson Prentice Hall, Upper Saddle River New Jersey.

Gregoret M.C., Dardanelli J., Bongiovanni R. and Díaz-Zorita M. 2006. Modelo de respuesta sitio-específica del maíz al nitrógeno y agua edáfica en un Haplustol. Ciencia del Suelo 24: 147–159.

Gregoret M.C., Díaz Zorita M., Dardanelli J. and Bongiovanni R.G. 2011. Regional model for nitrogen fertilization of site-specific rainfed corn in haplustolls of the central Pampas, Argentina. Precis. Agric. 12: 831–849.

Griffin T.W. 2010. The Spatial Analysis of Yield Data. En: Margaret, O. (ed.), Geostatistical Applications in Precision Agriculture. Springer, New York, pp. 89–116.

Guastaferro F., Castrignanò A., Benedetto D., Sollitto D., Troccoli A. and Cafarelli B. 2010. A comparison of different algorithms for the delineation of management zones. Precis. Agric. 11: 600–620.

Guerin L. and Stroup W.W. 2000. A simulation study to evaluate PROC MIXED analysis of repeated measures data. Proc. Ann. Conf. Applied Stat. Agric., 12th, Manhattan, KS. 30 Apr.–2 May 2000. Kansas State Univ., Manhattan, KS.

Gurka M.J. 2006. Selecting the Best Linear Mixed Model under REML. The American Statistician, 60(1):19.

Halcro G., Corstanje R. and Mouazen A.M. 2013. Site-specific land management of cereal crops based on management zone delineation by proximal soil sensing. In Precision agriculture'13 (pp. 475-482). Wageningen Academic Publishers.

Hemmat R. and Adamchuk V.I. 2008. Sensor systems for measuring soil compaction: Review and analysis. Comp. Electron. Agric. 63: 89–103.

Hijmans R.J. 2013. raster: raster: Geographic data analysis and modeling. R package version 2.1-49. http://CRAN.R-project.org/package=raster

Hong N., White J.G., Gumpertz M.L. and Weisz R. 2005. Spatial analysis of precision agriculture treatments in randomized complete blocks. Guidelines for covariance model selection. Agron. J. 97: 1082–1096.

Hörbe T.A.N., Amado T.J.C., Ferreira A.O. and Alba P.J. 2013. Optimization of corn plant population according to management zones in Southern Brazil. Precis. Agric.14: 450–465.

Hornung A., Khosla R., Reich R., Inman D. and Westfall D.G. 2006. Comparison of site-specific management zones: Soil-color-based and yield-based. Agron. J., 98, 407–415.

Iqbal J., Thomasson J.A., Jenkins J.N., Owens P.R. and Whisler F.D. 2005. Spatial variability analysis of soil physical properties of alluvial soils. Soil Sci. Soc Am J. 69: 1338–1350.

Isaaks E.H. and Srivastava R.M. 1989. An Introduction to Applied Geostatistics. Oxford Univ. Press, New York, 561 pp.

Johnson R.A. and Wichern D.W. 2007. Applied Multivariate Statistical Analysis. Prentice Hall. New York.

Jombart T., Devillard S., Dufour A.B. and Pontier D. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101: 92–103.

Journel A.G. and Huijbregts C.J. 1978. Mining geostatistics. Academic Press, Inc., London, UK.

Kerry R. and Oliver M.A. 2004. Average variograms to guide soil sampling. Int. J. Appl. Earth Obs. 5:307–325.

Koch B., Khosla R., Frasier W.M., Westfall D.G. and Inman D. 2004. Economic feasibility of variable-rate nitrogen application utilizing site-specific management zones. Agron. J. 96:1572–1580.

Kravchenko A.N., Robertson G.P., Hao X. and Bullock D.G. 2006. Management Practice Effects on Surface Total Carbon. Agron. J. 98: 1559–1568.

Kravchenko A.N., Robertson G.P., Thelenand K.D. and Harwood R.R. 2005. Management, topographical, and weather effects on spatial variability of crop grain yields. Agron. J. 97:514–523.

Lambert D.M., Lowenberg-Deboer J. and Bongiovanni R. 2004. A Comparison of Four Spatial Regression Models for Yield Monitor Data: A Case Study from Argentina. Precis. Agric. 5: 579–600.

Landis J.R. and Koch G.G. 1977. The measurement of observer agreement for categorical data. Biometrics 33: 159-174.

Lark R. and Stafford J. 1997. Classification as a first step in the interpretation of temporal and spatial variation of crop yield. Ann. Appl. Biol. 130: 111–121.

Lark R.M. 1998. Forming spatially coherent regions by classification of multi-variate data: an example from the analysis of maps of crop yield. Int. J. Geogr. Inf. Sci. 12: 83–98.

Lawes R.A., and Bramley R.G.V. 2012. A simple method for the analysis of on-farm strip trials. Agron. J. 104(2): 371–377.

Lee D.T., Schachter B.J. 1980. Two algorithms for constructing a Delaunay triangulation. Int. J. Comput. Inf. Sci. 9: 219–242.

Lee J. and Wong D.W.S. 2001 Statistical Analysis with Arcview GIS. John Wile and Son, New York, 192 pp.

Legendre P. and Legendre L. 1998. Numerical Ecology, 2nd ed. Elsevier Science, Amsterdam, 853 pp.

Li H., Lascano R.J., Barnes E.M., Booker J., Wilson L.T., Bronson K.F. and Segarra E. 2001. Multispectral refl ectance of cotton related to plant growth, soil water and texture, and site elevation. Agron. J. 93: 1327–1337.

Li Y., Shi Z., Li F. and Li H.Y. 2007. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. Comp. Electron. Agric. 56: 174–186.

Long D.S. 1998. Spatial autoregression modeling of site-specific wheat yield. Geoderma, 85: 181–197.

Maechler M., Rousseeuw P., Struyf A., Hubert M. and Hornik K. 2013. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4.

Marques Da Silva, J.R. 2006. Analysis of the Spatial and Temporal Variability of Irrigated Maize Yield. Biosyst. Eng. 94: 337–349.

McBratney A.B. and Pringle M.J. 1997. Spatial variability in soil: Implications for precision agriculture. p. 3–31. En J.V. Stafford (ed.) Precision agriculture '97. Vol. I: Spatial variability in soil and crop. BIOS Sci. Publ., Oxford, UK.

Mead R. 1971. Models for interplant competition in irregularly spaced population. In: Statistical Ecology, Patil G.P., Pielou E.C. and Waters W.E. (Eds.). Pensilvania State University Press, State College, PA, pp: 13-22.

Melchiori R., Albarenque S. y Kemerer A. 2013. Estado de la adopción de la agricultura de precisión en argentina. 12º Curso Internacional de Agricultura de Precisión. INTA Manfredi, Córdoba.

Meyer D., Dimitriadou E., Hornik K., Weingessel A. and Leisch F. 2013. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-2/r455. http://R-Forge.R-project.org/projects/e1071/

Miller M.P., Singer M.J. and Nielsen D.R. 1988. Spatial variability of wheat yield and soil properties on complex hills. Soil Sci. Soc. Am. J. 52: 1133–1141.

Milne A.E., Webster R., Ginsburg D. and Kindred D. 2012. Spatial multivariate classification of an arable field into compact management zones based on past crop yields. Comp. Electron. Agric. 80: 17–30.

Minasny B., McBratney A.B. and Whelan B.M. 2005. VESPER version 1.62. Australian Centre for Precision Agriculture, McMillan Building A05, The University of Sydney, NSW 2006. (http://www.usyd.edu.au/su/agric/acpa)

Minasny B., McBratney A.B., 2002. FuzME version 3.0, Australian Centre for Precision Agriculture, The University of Sydney, Australia.

Moral F.J., Terrón J.M. and Marques da Silva J.R. 2010. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. Soil Till. Res. 106, 335–343.

Moran P. 1948. The interpretation of statistical maps. J. Roy. Stat. Soc. B Method. 10, 243–251.

Morrell C.H. 1998. Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. Biometrics 54: 1560–1568.

Mouazen A.M., Dumont K., Maertens K. and Ramon H. 2003. Two-dimensional prediction of spatial variation in topsoil compaction of a sandy loam field-based on measured horizontal force of compaction sensor. Soil Till. Res. 74:91–102

Mulla D. J. and Hammond M.W. 1988. Mapping soil test results from large irrigation circles. En J. S. Jacobsen (Ed.), Proceedings of the 39th Annual Far West Regional Fertilizer Conference (pp. 169–171). Pasco, WA: Agricultural Experimental Station Technical Paper No. 8597.

Mzuku M., Khosla R., Reich R., Inman D., Smith F. and MacDonald L. 2005. Spatial variability of measured soil properties across site-specific management zones. Soil Sci. Soc. Am. J. 69:1572–1579.

Odeh I.O.A., Chittleborough D.J. and McBratney A.B. 1992. Soil Pattern Recognition with Fuzzy-c-means: Application to Classification and Soil-Landform Interrelationships. Soil Sci. Soc. Am. J. 56: 505.

Oliver M.A. 2010. Geostatistical applications for precision agriculture. Springer, New York.

Oliver M.A. 2013. Precision agriculture and geostatistics: How to manage agriculture more exactly. Significance, 10(2): 17–22.

Oliver M.A. and Webster R. 1989. A geostatistical basis for spatial weighting in multivariate classification. Math. Geol. 21: 15–35.

Ortega R.A. and Santibanez O.A., 2007. Determination of management zones in corn (Zea mays L.) based on soil fertility. Comp. Electron. Agric. 58: 49–59.

Osorio F., Vallejos R., Cuevas F. 2012. SpatialPack: Package for analysis of spatial data. R package version 0.2. http://CRAN.R-project.org/package=SpatialPack.

Pachepsky Y.A., Timlin D.J. and Rawls W.J. 2001. Soil Water Retention as Related to Topographic Variables. Soil Sci. Soc. Am. J. 65: 1787.

Panten K., Bramley R.G.V., Lark R.M. and Bishop T.F.A. 2010. Enhancing the value of field experimentation through whole-of-block designs. Precis. Agric. 11:198–213.

Papadakis J.S. 1937. Methode statistique pour des experiencessur champ. Institut d'Amelioration des Plantes aThessaloniki. Bullettin, 23, 30pp.

Paradis E.; Claude J. and Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20(2): 289–290.

Patterson H.D. and Thompson R. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554.

Pearson K. 1901. On lines and planes of closest fit to systems of points in space. Philos. Mage. 2: 559–572.

Pebesma E.J. 2004. Multivariable geostatistics in S: the gstat package. Comp. Geosci. 30: 683-691.

Pedroso M., Taylor James, Tisseyre B., Charnomordic B. and Guillaume S. 2010. A segmentation algorithm for the delineation of agricultural management zones. Comp. Electron. Agric. 70: 199–208.

Peralta N., Costa J.L., Castro Franco M. y Balzarini M. 2012. Delimitación de zonas de manejo con modelos de elevación digital y profundidad de suelo. Interciencia 38(6) 418–424.

Peralta N., Franco Castro M. y Costa J. 2011. Relación espacial entre variables de sitio y rendimiento para la delimitación de zonas de manejo mediante el uso de herramientas informáticas. En: Mendarozqueta, A.R. de, Marciszack, M.M., Groppo, M.A. (Eds.), III Congreso Argentino De Agroinformatica. Córdoba, pp. 58–69.

Peralta N.R. and Costa J.L. 2013. Delineation of management zones with soil apparent electrical conductivity to improve nutrient management. Comp. Electron. Agric. 99: 218–226.

Peralta N.R., Costa J.L., Balzarini M. and Angelini H. 2013. Delineation of management zones with measurements of soil apparent electrical conductivity in the southeastern pampas. Can. J. Soil Sci. 93: 205–218.

Pereira L.N., Coelho P.S. 2012. Small area estimation using a spatio-temporal linear mixed model. REVSTAT–Statistical Journal 10(3): 285–308.

Pierce F.J. and Novak P. 1999. Aspects of precision agriculture. Adv. Agron. 67: 1–85.

Piikki, K., Söderström, M., & Stenberg, B. 2013. Sensor data fusion for topsoil clay mapping. Geoderma 199: 106–116.

Ping J.L. and Dobermann A. 2003. Creating Spatially Contiguous Yield Classes for Site-Specific Management. Agron. J. 95: 1121.

Pinheiro J., Bates D., DebRoy S. and Sarkar D. and the R Development Core Team. 2013. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-109.

Pinheiro J.C. and Bates D.M. 2004. Mixed-Effects Models in S and S-PLUS. Springer, New York. 530 pp.

Plant R.E. 2001. Site Specific Management: the application of information technology to crop production. Comp. Electron. Agric. 30: 9–29.

Pringle M., McBratney A., Whelan B. and Taylor J.A. 2003. A preliminary approach to assessing the opportunity for site-specific crop management in a field, using yield monitor data. Agric. Syst. 76: 273–292.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Remy N., Boucher A. and Wu J. 2009. Applied geostatistics with SGeMS. Cambridge University Press, New York.

Rencher A.C. and Christensen W.F. 2012. Methods of multivariate analysis. John Wiley & Sons, Inc., Hoboken, New Jersey, 800 pp.

Ribeiro Junior P.J. and Diggle P.J. 2001. geoR: a package from geostatistical analysis. RNEWS 1(2):15–18.

Ripley B. D. 1981. Spatial Statistics, Wiley, New York.

Rodrigues M.S., Corá J.E., Castrignanò A., Mueller T.G. and Rienzi E. 2013. A Spatial and Temporal Prediction Model of Corn Grain Yield as a Function of Soil Attributes. Agron. J. 105(6): 1878–1887.

Rodríguez-Pérez J.R., Plant R.E., Lambert J.-J. and Smart D.R. 2011. Using apparent soil electrical conductivity (ECa) to characterize vineyard soils of high clay content. Precis. Agric. 12: 775–794.

Roel A. and Plant R.E., 2004. Factors Underlying Yield Variability in Two California Rice Fields. Agron. J. 96:1481–1494.

Roel A. and Terra J. 2006. Muestreo de suelos y factores limitantes del rendimiento, in: Bongiovanni, R. (Ed.), Agricultura De Precisión: Integrando Conocimientos Para Una Agricultura Moderna y Sustentable. PROCISUR/IICA, Montevideo, pp. 65–80.

Saby N.P.A., Thioulouse J., Jolivet C.C., Ratié C., Boulonne L., Bispo A., Arrouays D. 2009. Multivariate analysis of the spatial patterns of 8 trace elements using the French soil monitoring network data. Sci. Total Environ. 407, 5644–5652.

SAGyP (Secretaría de Agricultura, Ganadería y Pesca de la Nación Argentina)-INTA (Instituto Nacional de Tecnología Agropecuaria). 1989. Mapa de Suelos de Provincia de Buenos Aires. Escala 1: 500000. Proyecto PNUD Arg. 85/019. Buenos Aires.

SAS Institute Inc. 2004. SAS STAT User's Guide, Version 9.1, Cary, NC, USA.

Satorre E.H. and Slafer G.A. 1999. Wheat production systems of the Pampas. En: Satorre, E.M. and Slafer, G.A. (Ed.), Wheat Ecology and Physiology of Yield Determination. The Haworth Press Inc., New York, pp. 333–348.

Schabenberger O. 2006. Generalized linear Mixed Models in Agriculture: Theory and Applications. Full-Day Workshop. Kansas State University Conference on Applied Statistics in Agriculture, 151 pp.

Schabenberger O. and Gotway C. A. 2004. Statistical Methods for Spatial Data Analysis. Taylor and Francis. Chapman and Hall/CRC, 488pp.

Schabenberger O. and Pierce F. 2002. Contemporary Statistical Models for the Plant and Soil Sciences. Taylor and Francis. CRC Press, 738 pp.

Scharf P.C. and Alley M.M. 1993. Accounting for spatial yield variability in field experiments increases statistical power. Agron. J. 85: 1254–1256.

Schepers A.R., Shanahan J.F., Liebig M.A., Schepers J.S., Johnson S.H. and Luchiari Jr A. 2004. Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. Agron. J. 96:195–203.

Schlesinger W.H., Raikes J.A., Hartley A.E. and Cross A.E. 1996. On the spatial pattern of soil nutrients in desert ecosystems. Ecology 77 (2), 364–374.

Searle S.R., Casella G. and McCullogh C.E. 1992. Variance Components. Wiley, New York.

Shearer J. 2001. DGPS yield monitoring to assist in managing vineyard variabi-lity. En: Proceedings of the 11th Australian Wine Industry Technical Conferen-ce, Adelaide, p. 9-13.

Sibson R. 1981. A Brief Description of Natural Neighbor Interpolation, Interpreting Multivariate Data, V. Barnett editor, John Wiley and Sons, New York, p. 21–36.

Simón M., Peralta N. y Costa J.L. 2013. Relación entre la conductividad eléctrica aparente con propiedades del suelo y nutrientres. Ciencia del suelo, 31(1): 45–55.

Smith A.B., Cullis B.R. and Thompson R. 2001. Analyzing varietyby environment data using multiplicative mixed models and adjustmentsments for spatial field trend. Biometrics 57: 1138–1147.

Stafford J.V., Lark R.M. and Bolam H.C. 1998. Using yield maps to regionalize fields into potential management units. En: Robert, P.C. (Ed.), Precision Agriculture. 4th Proc. Int. Conf., St. Paul, MN, 19–22 July 1998. ASA, CSSA and SSSA, Madison, WI, pp. 225–237.

Sudduth K.A. and Drummond S.T. 2007. Yield Editor. Agron. J. 99(6): 1471–1482.

Sudduth K.A., Kitchen N.R., Bollero G.A., Bullock D.G. and Wiebold W.J. 2003. Comparison of electromagnetic induction and direct sensing of soil electrical conductivity. Agron. J. 95: 472–482.

Sun W., Whelan B., McBratney A.B. and Minasny B. An integrated framework for software to provide yield data cleaning and estimation of an opportunity index for site-specific crop management. Precis. Agric., 14: 376–391.

Taylor J., Wood G., Earl R. and Godwin R. 2003. Soil factors and their influence on within-crop variability. Part ii: Spatial analysis and determination of management zones. Biosystems Engineering 84(4): 441–453.

Taylor J.A., McBratney A.B. and Whelan B.M. 2007. Establishing Management Classes for Broadacre Agricultural Production. Agron. J. 99: 1366–1376.

Thioulouse J., Chessel D. and Champely S. 1995. Multivariate analysis of spatial patterns: a unified approach to local and global structures. Environ. Ecol. Stat. 2: 1–14.

Tou J.T. and Gonzalez R.C. 1974. Pattern recognition principles. Addison-Wesley, Reading, MA.

Tozer P.R. and Isbister B.J. 2007. Is it economically feasible to harvest by management zone? Precis. Agric. 8: 151–159.

Trangmar B.B., Yost R.S. and Uehara G. 1985. Application to geostatistics to spatial studies of soil properties. Advances in Agronomy 38: 45-94.

Van Uffelen C.G.R., Verhagen J. and Bouma J. 1997. Comparison of simulated crop yield patterns for site-specific management. Agric. Syst. 54: 207–222.

Verbeke G. and Molenberghs G. 2000. Linear Mixed Models for Longitudinal Data, Springer-Verlag, Berlin.

Veris Technologies. 2001. Frequently asked questions about soil electrical conductivity. Veris Technologies, Salina. KS. http://www.veristech.com (acceso 9/02/14).

Vieira S.R., Nielsen D.R. and Biggar J.W. 1981. Spatial variability of field-measured infiltration rate. Soil Sci. Soc. Am. J. 45: 1040–1048.

Vrindts E., Mouazen A., Reyniers M., Maertens K., Maleki M. and Ramon H. 2005. Management zones based on correlation between soil compaction, yield and crop data. Biosys. Engin. 92(4): 419–428.

Walkley A. and Black I.A. 1934. An examination of Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Sci. 37: 29–37.

Wartenberg D. 1985. Multivariate spatial correlation: a method for exploratory geographical analysis. Geogr Anal. 17: 263–83.

Webster R. 1973. Automatic soil boundary location for transect data. Mathematical Geology 5(1): 27–37.

Webster R. and Nortcliff S. 1984. Improved estimation of micronutrients in hectare plots of the Sonning series. J. Soil Sci. 35: 667–672.

Webster R. and Oliver M.A. 1989 Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional probability. J. Soil Sci. 40: 497–512

Webster R. and Oliver M.A. 1990. Statistical Methods in Soil and Land Resource Survey. Oxford University Press, New York.

Webster R. and Oliver M.A. 2007. Geostatistics for environmental scientists, 2nd edn. John Wiley and Sons,, Chichester UK.

West T.B., Welch K.B., and Galecki A.T. 2007. Linear mixed models: a practical guide using statistical software. Chapman & Hall/CRC, Boca Raton, 339 pp.

Whelan B. and McBratney A. B. 2003. Definition and interpretation of potential management zones in Australia. In Proceedings of the 11th Australian Agronomy Conference. Geelong, Australia: Australian Society of Agronomy.

Whelan B.M. and McBratney A.B. 2000. The "null hypothesis" of precision agriculture management. Precis. Agric. 2: 265–279.

Whelan B.M., McBratney A.B. and Minasny B. 2001. Vesper: Spatial prediction software for precision agriculture. p. 139–144. In G. Grenier and S. Blackmore (ed.) ECPA 2001: Proc. 3rd European Conf. on Precision Agriculture, Montpellier, France, 18–20 June. agro-Montpellier ENSAM, Montpellier, France.

Wilkinson G.N., Eckert S.R., Hncock T.W. and Mayo O. 1983. Nearest neighbor analysis whit field experiments. J. Roy. Statist. Soc. Ser. B. 45: 151–178

Willers J.L., Milliken G.A., Jenkins J.N., O'Hara C.G., Gerard P.D., Reynolds D.B., Boykin D.L., Good P.V. and Hood K.B. 2008. Defining the experimental unit for the design and analysis of site-specific experiments in commercial cotton fields. Agric. Syst. 96: 237–249.

Windham M.P. Cluster validity for fuzzy clustering algorithms. 1981. Fuzzy Sets Syst. 5: 177–185.

Xie L.X. and Beni G. 1991. Validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 3(8): 841–847.

Xin-Zhong W., Guo-Shun L., Hong-Chao H., Zhen-Hai W., Qing-Hua L., Xu-Feng L., Wei-Hong H. and Yan-Tao L. 2009. Determination of management zones for a tobacco field based on soil fertility. Comp. Electron. Agric. 65: 168–175.

Yan L., Zhou S., Feng L. and Hong-Yi L. 2007. Delineation of site-specific ma-nagement zones using fuzzy clustering analysis in a coastal saline land. Comp. Electron. Agric. 56: 174–186.

Zarco-Tejada P.J., Ustin S. L. and Whiting M.L. 2005. Temporal and spatial relationships between within-field yield variability in cotton and high-spatial hyperspectral remote sensing imagery. Agron. J. 97(3): 641–653.

Zimback C.R.L. 2001. Analise espacial de atributos químicos de solos para fins de mapeamento da fertilidade do solo. Tese (Livre-Docência) Faculdade de Ciências Agronômicas, Universidade Estadual Paulista. Botucatu.

Zimmerman D.L. 1991. A random field approach to the analysis of field-plot experiments and other spatial experiments. Biometrics 47: 223–239.

ANEXO 1

Códigos R. Cálculo de los índices de autocorrelación espacial de Moran y de Geary

Carga de la librería.

```
library(spdep)
```

Lectura de datos.

Se utilizará la base de datos correspondiente al archivo *CasoI.txt* utilizado en el Capítulo I. Las coordenadas se encuentran en las dos primeras columnas de la base.

```
sa <- read.table("C:\\Documents\\...\\CasoI.txt", header = TRUE)
cord <- coordinates(sa[,1:2])
```

Obtención de la matriz de pesos espaciales.

Para definir los vecindarios se utiliza la la distancia Euclídea considerando sitios vecinos a aquellos contiguos ubicados entre los 0 a 70 m de distancia. Para ello, se utiliza la función "dnearneigh". Para construir la matriz de pesos espaciales se emplea la estandarización por fila (style = "W").

```
gri <- dnearneigh(cord,0,35)
lw2 <- nb2listw(gri, style = "W")
plot(gri,cord,col = "red", pch = 20, cex = 1)
```

Calculo del MI y GI a partir de la matriz de pesos espaciales obtenida en el paso anterior. La significancia de los índices para cada variable se obtiene a través del método de Monte- Carlo basado en 999 permutaciones .

```
i.moran <- lapply(sa[,3:8], moran.mc, lw2, 999)
i.moran

i.geary <- lapply(sa[,3:8], geary.test, lw2, 999)
i.geary
```

ANEXO 2

Códigos de R. Estudio de la variabilidad espacial en datos gerreferenciados utilizando geoestadística y Modelos Lineales Mixtos

Se presentan los códigos de R para estudiar la variabilidad espacial a través de la aplicación de geoestadística y modelos lineales mixtos (MLM). Se plantea el ajuste de un modelo de correlación espacial exponencial isotrópico. Para la ilustración se utiliza la variable conductividad eléctrica aparente a 30 cm de profundidad (CE30). La base de datos corresponde al archivo *CasoI.txt* utilizado en el Capítulo I. También se necesita cargar el archivo *bordes.txt* el cual posee las coordenadas que delimitan el lote.

Geoestadística

Carga de la librería geoR.

```
library(geoR)
```

Lectura de datos.

```
bordes <- read.table("C:\\ Documents\\....\\CasoI.txt", header = TRUE)
ce30=as.geodata(CasoI, coords.col = 1:2, data.col = 3)
bordes <- read.table("C:\\ Documents\\....\\bordes.txt", header = TRUE)
```

Obtención de las semivarianzas sin tener en cuenta la presencias de tendencias espaciales.

```
semiv <- variog(ce30)
```

Obtención de las semivarianzas para estudiar la presencia de tendencias espaciales, en este caso se va a evaluar la presencia de una tendencia líneas en función de las coordenadas espaciales X e Y.

```
semiv_t <- variog(ce30, trend="1st")
```

Comparación del semivariograma con y sin tendencia.

```
par(mfrow=c(1,2))
plot(semiv, main = "Sin Tendencia”)
plot(semiv_t, main = "Con Tendencia")
```

Ajuste de un modelo exponencial al semivariograma empírico a través del método de mínimos cuadrados ponderados (WLS).

```
wls <-variofit(semiv,ini=c(50,100),nug=10, cov.model="exp",messages=T)
summary(wls)
```

Obtención de gráfico con el semivariograma empírico y del ajuste obtenido por WLS.

```
par(mfrow = c(1,1))
plot(bin, main = expression(paste( "Estimación WLS")))
lines(wls, lty = 2)
```

Utilización del método de interpolación de Kriging ordinario y obtención del mapa de variabilidad espacial

```
gr<-pred_grid(bordes, by=20)
KC<-krige.control(obj.model=wls,type.krige="ok")
pred<-krige.conv(ce30, loc=gr, krige=KC, bor=bordes)
contour(pred,filled=TRUE, color=terrain.colors)
title(main = "CE30")
```

Modelos lineales mixtos

Carga de la librería geoR.

```
library(geoR)
```

Lectura de datos.

```
bordes <- read.table("C:\\ Documents\\....\\CasoI.txt", header = TRUE)
ce30=as.geodata(CasoI, coords.col = 1:2, data.col = 3)
bordes <- read.table("C:\\ Documents\\....\\bordes.txt", header = TRUE)
```

Ajuste de los modelos sin tendencia espacial y con tendencia lineal en función de las coordenadas espaciales. Se utiliza el método de estimación máxima verosimilitud restringida (REML).

```
Modelo1 <- likfit(ce30, ini=c(30, 200), lik.method = "REML")
summary(Modelo1)

Modelo2 <- likfit(ce30, ini=c(30, 200), lik.method = "REML", trend="1st")
summary(Modelo2)
```

Utilización del método de interpolación kriging ordinario y obtención de mapa de variabilidad espacial.

```
gr<-pred_grid(bordes, by=20)
KC<-krige.control(obj.model=Modelo1,type.krige="ok")
pred<-krige.conv(ce30, loc=gr, krige=KC, bor=bordes)
contour(pred,filled=TRUE, color=terrain.colors)
title(main = "CE30")
```

ANEXO 3

Códigos R. Estudios multivariados de datos georreferenciados

Se presentan los códigos de R para realizar análisis multivariados en datos georrefrenciados. Se consideran dos alternativas de análisis:

Alternativa 1: Uso del método de análisis de componentes principales clásico (PCA)
Alternativa 2: Uso del método de componentes principales con restricción espacial (MULTISPATI- PCA)

Luego a partir de cualquiera de estas dos alternativas se puede proceder a través de las componentes principales obtenidas al estudio de la variabilidad espacial y al mapeo de las mismas con los códigos presentados en el ANEXO I y II.

Alternativa 1

Lectura de datos.

Se utiliza la base de datos del archivo *CasoI.txt* utilizado en el Capítulo I.

```
sa <- read.table("C:\\Documents\\...\\CasoI.txt", header = TRUE)
cord <- coordinates(sa[,1:2])
```

Análisis de componentes principales

```
pca2 <- dudi.pca(sa[,3:8], center=TRUE,scannf = FALSE,  nf = 6)
```

Obtención de biplots con las componentes obtenidas

```
scatter(pca2, xax = 1, yax = 2)
scatter(pca2, xax = 1, yax = 3)
```

Autovectores

```
pca2$c1
```

Autovalores

```
pca2$eig
```

Porcentaje de explicación de cada eje

```
pca2eig/sum(pca2$eig) * 100
```

Gráficos de correlación entre la variable y el eje (CP1 y CP2)

```
s.corcircle(pca2$co, clabel = 1.1)
add.scatter.eig(pca2$eig, xax = 1, yax = 2, posi = "bottomleft", ratio =
0.2)
```

Gráficos de correlación entre la variable y el eje (CP1 y CP3)

```
s.corcircle(pca2$co,xax = 1, yax = 3, clabel = 1.1)
add.scatter.eig(pca2$eig, xax = 1, yax = 3, posi = "bottomleft", ratio =
0.2)
```

Alternativa 2

Carga de las librerías.

```
library(ade4)
library(spdep)
```

Lectura de datos.

```
sa <- read.table("C:\\Documents\\...\\CasoI.txt", header = TRUE)
cord <- coordinates(sa[,1:2])
```

Análisis de componentes principales restringido espacialmente

Para realizar este análisis es necesario primero obtener una matriz de pesos espaciales (lw2), la obtención de la misma es similar a la detalla en el ANEXO I.

```
gri <- dnearneigh(cord,0,35)
lw2 <- nb2listw(gri, style = "W")
plot(gri,cord,col = "red", pch = 20, cex = 1)
```

```
ms2 <- multispati(pca2, lw2, scannf = F, nfposi = 6)
ms2
```

Obtención del biplot, junto con el diagrama de autovalores (sPC1 y sPC2).

```
s.arrow(ms2$c1,xax = 1, yax = 2, clabel = 1)
```

```
add.scatter.eig(ms2$eig, xax = 1, yax = 2, posi = "bottomleft", ratio =
0.2)
```

Obtención del biplot, junto con el diagrama de autovalores (sPC1 y sPC2).

```
s.arrow(ms2$c1,xax = 1, yax = 3, clabel = 1)
add.scatter.eig(ms2$eig, xax = 1, yax = 3, posi = "topright", ratio =
0.2)
```

Comparación de los resultados obtenidos a través de la Alternativa 1 y 2

```
summary(ms2)
```

ANEXO 4

Códigos R. Protocolo para la delimitación de zonas de manejo

Carga de librerías.

```
install.packages ("spdep","rgdal","geoR","gstat","ade4","e1071","raster")
library(spdep)
library(rgdal)
library(geoR)
library(gstat)
library(ade4)
library(e1071)
library(raster)
library(nlme)
```

Carga de datos

```
datos <-read.table("C:\\Users\\...\\datos.txt", header = TRUE)
```

Conversión de coordenadas espaciales

```
coordinates(datos) <- ~x+y
proj4string(datos) <- CRS("+proj=longlat + datum=WGS84")

datos <- spTransform(datos, CRS("+proj=utm +zone=21 +south +ellps=WGS84
+datum=WGS84 "))

datos <- as.data.frame(datos)
datos <- datos[,c(2,3,1)]
datos
```

Eliminación de outliers

```
summary(datos$CE30)
par(mfrow=c(1,2))
hist(datos$CE30,col='green',nclass=20,main="Histograma",ylab='Frecuencia
Relativa',xlab='CE30 (mS/m)')
boxplot(datos$CE30,col='green',ylab='CE30 (mS/m)',main="Box-Plot")

Media <- mean(datos$CE30)
DE <- sd(datos$CE30)
LI <- Media-2.5*DE
LS <- Media+2.5*DE

datos$CE30[LS<datos$CE30|datos$CE30<LI] <-NA
datos <- subset(na.omit(datos),select=c(x,y,CE30))

summary(datos$CE30)
par(mfrow=c(1,2))
hist(datos$CE30,col='green',nclass=20,main="Histograma",ylab='Frecuencia
Relativa',xlab='CE30 (mS/m)')
boxplot(datos$CE30,col='green',ylab='CE30 (mS/m)',main="Box-Plot")
par(mfrow=c(1,1))
```

Eliminación de inliers

```
cord <- coordinates(datos[,1:2])
gri <- dnearneigh(cord,0,20)
```

```
lw2 <- nb2listw(gri, style = "W")

MP <- moran.plot(datos$CE30,
lw2,quiet=T,labels=F,zero.policy=F,xlab="CE30", ylab="CE30 Spatially
Lagged")
summary(MP)

ML <- localmoran(datos$CE30,lw2,p.adjust.method="bonferroni",alternative
="less")

Influ <- MP$is.inf ; Influ

IMl <- printCoefmat(data.frame(ML,
row.names=datos$Casos),check.names=FALSE)
IMl
datos <- data.frame(datos,IMl,Influ); datos

datos1 <- datos[datos$dfb.1_ == FALSE & datos$dfb.x == FALSE &
datos$dffit == FALSE
& datos$cov.r == FALSE & datos$cook.d  == FALSE & datos$hat == FALSE, ]

datos2 <- as.matrix(datos1)
datos2 <- subset(datos1,datos1[,4] > 0 | datos1[,8]>0.05 )
```

Interpolación espacial de los datos

```
coordinates(datos2) <- ~x+y
CE30vario <- variogram(CE30~1, datos2, cutoff=380)
plot(CE30vario,main="CE30",xlab="Distancia",ylab="Semivarianza")

CE30fitvariog <- fit.variogram(fit.method=1,CE30vario, vgm(25, "Sph",
80,10))
CE30fitvariog
plot(CE30vario,CE30fitvariog,main="CE30",xlab="Distancia",ylab="Semivaria
nza")

bordes <-read.table("C:\\Users\\...\\bordes.txt", header = TRUE)

gr<-pred_grid(bordes, by=10)
gri <- polygrid(gr,bor=bordes)
plot(gri,col = "red", pch = 10, cex = 0.2,xlab="X",ylab="Y")
gridded(gri) = ~Var1+Var2

CEKg <- krige(CE30~1, datos2, gri, model = CE30fitvariog, block =
c(40,40))

spplot(CEKg["var1.pred"], main = "Mapa de variabilidad
espacial",col.regions=terrain.colors(100))

PredCE30 <- as.data.frame(CEKg)
PredCE30 <- PredCE30[,1:3]
names(PredCE30)[1]<-paste("x")
names(PredCE30)[2]<-paste("y")
names(PredCE30)[3]<-paste("CE30")
```

Delimitación de clases de sitios

```
Pred <- cbind(PredCE30[,1:3], PredCE90[,3], PredElev[,3],PredPe[,3],
PredRtoTg[,3])
Pred

pca2 <- dudi.pca(Pred[,3:7], center=T,scannf = FALSE,  nf = 5)
#scatter(pca2, xax = 1, yax = 2,clab.r=0.4, clab.c=0.9)

cord <- coordinates(Pred[,1:2])
gri <- dnearneigh(cord,0,10)
lw2 <- nb2listw(gri, style = "W")

ms2 <- multispati(pca2, lw2, scannf = F, nfposi = 5)

s.arrow(ms2$c1,xax = 1, yax = 2, clabel = 1)
add.scatter.eig(ms2$eig, xax = 1, yax = 2, posi = "bottomleft", ratio =
0.2)

CS <- ms2$li[,1:5]
PredAM <- cbind(Pred,CS) ;PredAM
```

Delimitacion de clases de manejo

```
CM2<-cmeans(PredAM[,8:10],2,100,method="cmeans",m=1.3)
CM3<-cmeans(PredAM[,8:10],3,100,method="cmeans",m=1.3)
CM4<-cmeans(PredAM[,8:10],4,100,method="cmeans",m=1.3)

CM22<-as.data.frame(CM2$cluster)
CM33<-as.data.frame(CM3$cluster)
CM44<-as.data.frame(CM4$cluster)
```

Delimitacion de zonas de manejo

```
I2CM <- fclustIndex(CM2,PredAM[,8:10], index=c("xie.beni",
"fukuyama.sugeno",
"partition.coefficient", "partition.entropy"))

I3CM <- fclustIndex(CM3,PredAM[,8:10], index=c("xie.beni",
"fukuyama.sugeno",
"partition.coefficient", "partition.entropy"))

I4CM <- fclustIndex(CM4,PredAM[,8:10], index=c("xie.beni",
"fukuyama.sugeno",
"partition.coefficient", "partition.entropy"))

Indices0 <- cbind(I2CM,I3CM,I4CM)

XieBeni <-Indices0[1,]
FukSug <-Indices0[2,]
CoefPart_1 <-Indices0[3,]
CoefPart <- 1/CoefPart_1
EntrPart <-Indices0[4,]

Indices <- as.data.frame(rbind(XieBeni,FukSug,CoefPart,EntrPart))
Indices
```

```
XieBeniMax<-max(Indices[1,])
FukSugMax<-max(Indices[2,])
CoefPartMax<-max(Indices[3,])
EntrPartMax<-max(Indices[4,])

XieBeniN<- XieBeni/XieBeniMax
FukSugN<- FukSug/FukSugMax
CoefPartN<- CoefPart/CoefPartMax
EntrPartN<-EntrPart/EntrPartMax

IndicesN <- as.data.frame(rbind(XieBeniN,FukSugN,CoefPartN,EntrPartN))
IndicesN2 <- (IndicesN)^2

Indice2CM <- sqrt(sum(IndicesN2[,1]))
Indice3CM <- sqrt(sum(IndicesN2[,2]))
Indice4CM <- sqrt(sum(IndicesN2[,3]))

Indice2CM
Indice3CM
Indice4CM

base00 <- cbind(PredAM[,1:2],CM22,CM33,CM44)
coordinates(base00) = ~x+y
gridded(base00)=TRUE
spplot(base00["CM2$cluster"],col.regions=terrain.colors(100),colorkey= F)

spplot(base00["CM3$cluster"],col.regions=terrain.colors(100),colorkey= F)

spplot(base00["CM4$cluster"],col.regions=terrain.colors(100),colorkey= F)

base0 <- cbind(PredAM[,1:2],CM22)
names(base0)[3]<-paste("Zona")

base1 <- base0
coordinates(base1) = ~x+y
gridded(base1)=TRUE

base2 <- raster(base1)
plot(base2)

median3x3 <- focal(base2, fun=median,w=3)
plot(median3x3)

med5x5 <- focal(base2, fun=median,w=5)
plot(med5x5)

med7x7 <- focal(base2, fun=median,w=7)
plot(med7x7)

base3 <- as.data.frame(med7x7,xy=T)
base3 <- na.omit(base3)
names(base3)[3]<-paste("Zona")

Basefinal <- subset(merge(base3,PredAM,by =c( "x","y"), all = T),
select=c(x,y,CE30,CE90,Elev,Pe,RtoTg,Zona))
Basefinal <- na.omit(Basefinal)
```

Validación de zonas de manejo

```
Muestreo <-read.table("C:\\Users\\...\\Muestreo.txt", header = TRUE)

modelo.001_mo_REML<-gls(mo~1+Zona
,correlation=corExp(form=~as.numeric(as.character(X))+as.numeric(as.chara
cter(Y)),metric="euclidean",nugget=FALSE),method="REML",na.action=na.omit
,data=Muestreo)

modelo.002_mo_REML<-gls(mo~1+Zona
,correlation=corExp(form=~as.numeric(as.character(X))+as.numeric(as.chara
cter(Y)),metric="euclidean",nugget=TRUE),method="REML",na.action=na.omit
,data=Muestreo)

modelo.003_mo_REML<-gls(mo~1+Zona
,correlation=corGaus(form=~as.numeric(as.character(X))+as.numeric(as.char
acter(Y)),metric="euclidean",nugget=FALSE),method="REML",na.action=na.omi
t,data=Muestreo)

modelo.004_mo_REML<-gls(mo~1+Zona
,correlation=corGaus(form=~as.numeric(as.character(X))+as.numeric(as.char
acter(Y)),metric="euclidean",nugget=TRUE),method="REML",na.action=na.omit
,data=Muestreo)

modelo.005_mo_REML<-gls(mo~1+Zona
,correlation=corSpher(form=~as.numeric(as.character(X))+as.numeric(as.cha
racter(Y)),metric="euclidean",nugget=FALSE),method="REML",na.action=na.om
it,data=Muestreo)

modelo.006_mo_REML<-gls(mo~1+Zona
,correlation=corSpher(form=~as.numeric(as.character(X))+as.numeric(as.cha
racter(Y)),metric="euclidean",nugget=TRUE),method="REML",na.action=na.omi
t,data=Muestreo)

modelo.007_mo_REML<-gls(mo~1+Zona,method="REML",na.action=na.omit
,data=Muestreo)

AICmod1 <- AIC(modelo.001_mo_REML)
AICmod3 <- AIC(modelo.003_mo_REML)
AICmod5 <- AIC(modelo.005_mo_REML)

AICmod1
AICmod3
AICmod5

AICmod2 <- AIC(modelo.002_mo_REML)
AICmod4 <- AIC(modelo.004_mo_REML)
AICmod6 <- AIC(modelo.006_mo_REML)

AICmod2
AICmod4 AICmod6

LRT <- anova(modelo.003_mo_REML, modelo.004_mo_REML);LRT

LRT1 <- anova(modelo.003_mo_REML, modelo.007_mo_REML);LRT1

summary(modelo.003_mo_REML)
```