Análisis de la Variabilidad Espacial en Lotes Agrícolas

Manual de Buenas Prácticas

Programa PROTRI

UNC Universidad Nacional de Córdoba

FCA Facultad de Ciencias Agropecuarias

Secretaría de Ciencia y Tecnología | Ministerio de Industria, Comercio, Minería y Desarrollo Científico Tecnológico | Gobierno de la Provincia de Córdoba

BALZARINI MÓNICA
(COMP.)

Análisis de la variabilidad espacial en lotes agrícolas.

Manual de buenas prácticas.

CÓRDOBA, MARIANO
BRUNO, CECILIA
AGUATE, FERNANDO
TABLADA, MARGOT
BALZARINI, MÓNICA

DISEMINACIÓN CIENTÍFICA Y TRANSFERENCIA DE RESULTADOS DE INVESTIGACIÓN, PROMOVIDAS POR EL MINISTERIO DE CIENCIA Y TECNOLOGÍA DE LA PROVINCIA DE CÓRDOBA

La cita bibliográfica para el presente documento

Córdoba M, Bruno C, Aguate F, Tablada M, Balzarini M. 2014. Análisis de la variabilidad espacial en lotes agrícolas. Manual de Buenas Prácticas. Ed. Balzarini, M. Eudecor. Córdoba, Argentina.

Análisis de la variabilidad espacial en lotes agrícolas. Manual de Buenas Prácticas Agrícolas / Mónica Balzarini ... [et.al.]. - 1a ed. – Córdoba: Eudecor, 2014. 119 p. ; 25x17 cm.
ISBN 978-987-1536-66-5
1. Agronomía. 2. Geoestadística. 3. Biometría. I. Balzarini, Mónica
CDD 630.7

1° Edición

Primera Impresión

Impreso en Argentina

ISBN: 978-987-1536-66-5

Queda hecho el depósito que prevé la ley 11.723

PRÓLOGO

En las últimas décadas se ha impulsado el desarrollo y la utilización de nuevas tecnologías para la agricultura que permiten capturar diferentes tipos de datos espaciales, *i.e.* datos de diferentes variables asociados a una localización en el espacio para diferentes sitios, incluso a escala de lote. La variabilidad espacial es clave en la agricultura de precisión (AP). La AP debe ser comprendida como una forma de gestión agrícola basada en el uso de información sobre variables georreferenciadas tanto de características de suelo y topografía como de rendimientos. El óptimo uso del gran volumen de datos, derivado de maquinarias precisas, depende fuertemente de las capacidades para explorar y analizar datos de sitios donde subyacen complejas interacciones.

En esta publicación se propone un protocolo integrado para procesar variables de sitio en lotes agrícolas. El análisis de la estructura espacial de variables de suelo y rendimiento es investigado desde un enfoque interdisciplinario, que incluye perspectivas agronómicas y estadística-computacionales. En la cátedra de Estadística y Biometría de la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba, hemos investigado sobre el desempeño de métodos estadísticos para detectar y caracterizar variabilidad espacial. Aquí presentamos estrategias de análisis de datos para caracterizar la variabilidad espacial en lotes agrícolas. Usando el protocolo propuesto se ilustra la obtención de predicciones espaciales y mapas

de variabilidad espacial de propiedades de suelo medidas intensivamente sobre el terreno en un lote de trigo bajo rotación. Se presenta también una guía de buenas prácticas para la delimitación de zonas orientadas al manejo sitio-específico basado en datos de rendimiento y suelo.

El texto ha sido desarrollado proveyendo los comandos necesarios para utilizar estas herramientas analíticas en el intérprete de R (R Core Team, 2014) del software InfoStat (Di Rienzo *et al*., 2014). Se presenta un tutorial con los comandos de programa que permiten trabajar datos de AP desde su pre-procesamiento hasta la delimitación de zonas de manejo. Los métodos presentados aquí no son exhaustivos, existen numerosas estrategias analíticas que podrían ser implementadas para el análisis de datos georreferenciados, pero los incluidos en el protocolo propuesto constituyen una guía de buenas prácticas para su análisis de datos. Esperamos que su implementación sea provechosa para el desarrollo agropecuario y que, a partir de su uso, surjan nuevas preguntas que generen un medio propicio para explorar los desafíos y oportunidades de la modelación estadística en agricultura por ambientes.

Los autores.

TABLA DE CONTENIDOS

PARTE I.

ANÁLISIS DE DATOS GEORREFERENCIADOS

TÉCNICAS EXPLORATORIAS PARA DATOS ESPACIALES

TRANSFORMACIÓN DE COORDENADAS GEOGRÁFICAS

En geodesia un *datum* es un conjunto de puntos de referencia en la superficie terrestre en base a los cuales las medidas de la posición son tomadas bajo un modelo asociado de la forma de la tierra (elipsoide de referencia) para definir el sistema de coordenadas geográficas. Dado que diferentes *datum* tienen diferentes radios y puntos centrales, un punto puede tener coordenadas diferentes, existiendo cientos de *datum* de referencia. Para Sudamérica el *datum* comúnmente utilizado es WGS84 (World Geodetic System 84). Éste es el *datum* estándar por defecto para coordenadas en los dispositivos GPS comerciales. Los usuarios de GPS deben chequear el *datum* utilizado ya que un error puede suponer una traslación de las coordenadas de varios cientos de metros.

Por una cuestión de practicidad, proyectamos este sistema de coordenadas geodésicas (expresados en grados, minutos y segundos) a otro sistema de coordenadas cartesianas (pasar de un modelo 3D a uno 2D) llamado sistema de proyección, típicamente UTM (*Universal Transverse Mercator*). Esta transformación 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) lo que facilita los cálculos de distancia y superficie. Por ello, un paso inicial en el análisis de datos espaciales es convertir las coordenadas geográficas en coordenadas cartesianas (UTM). 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*.,

2014) cuenta con la función *spTransform* que permite hacer transformación de sistemas de coordenadas.

ESTUDIO DE LA DISTRIBUCIÓN DE LA VARIABLE Y ELIMINACIÓN DE DATOS RAROS

DISTRIBUCIÓN DE LA VARIABLE

Un paso importante en el análisis exploratorio de los datos geoestadísticos continuos, es determinar el cumplimiento del supuesto de normalidad. Para ello, puede realizarse una estadística descriptiva que incluye la elaboración de gráficos de distribución de frecuencias y medidas resumen (media, mediana y coeficiente de asimetría) de las variables en análisis. Se considera que una distribución de frecuencias es simétrica y está próxima a la normalidad cuando la media y la mediana tienen valores muy próximos entre sí y el coeficiente de asimetría es inferior a 1.

OUTLIERS

Los *outliers*, o valores atípicos, son observaciones con valores que se encuentran fuera del patrón general o distribución del conjunto de datos. La eliminación de los *outliers*, previo al análisis, es fundamental para garantizar que las decisiones tomadas a partir del análisis sean las correctas. Los *outliers* se pueden eliminar fácilmente a través de un proceso donde se complementan distintas técnicas y teorías: 1) el conjunto de datos se limita dentro de un rango de variación razonable donde los valores máximos y mínimos se obtienen desde el conocimiento previo de su distribución, 2) para el conjunto de datos de una

variable, se calcula la media ($\bar{x}$) y la desviación estándar (SD) y se identifican los valores que se encuentran fuera de la media ± 3 SD. Según conocimiento teórico, se conoce que el 89% de los datos se encontrarán entre la media ± 3 SD cualquiera sea la distribución de la variable. En ocasiones en que los datos de monitores de rendimiento son sesgados como resultado de procesos no aleatorios tales como malas lecturas, cosechadoras funcionando a medio llenar o con el cabezal hacia abajo sobre áreas cosechadas, puede justificarse una modificación de estos límites (Taylor *et al*., 2007). Antes de la eliminación de los *outliers*, los mismos deben ser graficados utilizando coordenadas espaciales para visualizarlos. De esta manera será posible identificar si los datos seleccionados para ser eliminados indican algún efecto sistemático, por ejemplo sitios que pertenecen a una zona de bajo rendimiento dentro del lote, o si por el contrario se relacionan a errores aleatorios de lectura.

INLIERS

La aplicación de los pasos descriptos anteriormente, 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. Existen herramientas estadísticas diseñadas específicamente para identificar *inliers.* Tal es el caso del índice autocorrelación espacial local de Moran (IMl) (Anselin, 1995). Dado un grupo de datos que pertenecen a diferentes vecindarios, el IMl es aplicado a cada dato individualmente y 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:

$$IMl_i = \frac{x_i - \bar{X}}{\sigma^2} \sum_{j=1, j\neq 1}^{n} \left[w_{ij}(x_j - \bar{x}) \right] \quad (1)$$

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.

Para el cálculo del Índice de Moran se utilizan redes de conexión que derivan en un matriz de ponderación espacial binaria (W), es decir compuesta por ceros y unos ya que si la posición j es adyacente a la posición i, el término w_{ij} recibe un peso de 1 y si no, de 0. Otra posibilidad para construir la matriz W es relacionar los elementos con la distancia d entre los sitios de manera inversamente proporcional, es decir: $W_{ij} = 1/d_{ij}$. Así, valores muy cercanos en el espacio tendrán mayor peso o coeficiente de ponderación. Existen diferentes opciones o alternativas metodológicas para definir el tamaño y la forma de los vecindarios (Dray *et al*., 2006). En el protocolo propuesto aquí, la red de vecinos (vecindario) es definida utilizando la distancia Euclídea. Se considera puntos vecinos a aquellos contiguos ubicados entre un rango de distancia definido por un límite inferior y un límite superior, previamente preestablecido.

El IMl se puede estandarizar y su nivel de significación puede ser evaluado en base a una distribución normal estándar. Los valores positivos del IMl se corresponden con agrupamiento espacial de valores similares (ya sean altos o bajos) (autocorrelación positiva), mientras que un valor de IMl 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).

Para determinar la significancia estadística IMl, se calcula el valor-*p* asociado a la prueba de hipótesis que establece que la correlación de la información de un sitio con la de sus vecinos es nula. El valor-*p* para un índice determinado debe ser lo suficientemente pequeño para considerar el valor en cuestión como un *inlier* (rechazar la hipótesis nula). Dado que se realiza una prueba de hipótesis para cada uno de los puntos espaciales, se recomienda el ajuste de los valores-*p* por el criterio de Bonferroni (Bland y Altman, 1995). De no ajustase los valores-*p* por multiplicidad, algún IMl podría resultar significativo solo por azar (falsos positivos).

Anselin (1996) propuso para visualizar el IMl un diagrama de dispersión 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.

Las funciones *localmoran* y *moranplot* de la librería "spdep" (Bivand, 2014) del software R, permiten calcular el IMl y realizar el gráfico de dispersión de Moran para identificar *inliers*. Aplicando la función *localmoran* se obtiene el IMl y su significancia estadística para cada sitio. La función *moranplot* además de realizar el diagrama de dispersión ajusta un modelo de regresión lineal y calcula una serie de estadísticos de diagnóstico. 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: *Distancia de Cook*, *Leverage*, *DFFITS*, *DFBETAS* y *COVRATIO* (Drapper y Smith, 1998). 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. Se recomienda que los primeros *inliers* a ser removidos sean los detectados con el IMl (datos con IMl negativo y estadísticamente significativo) posteriormente, se construye el gráfico de dispersión de Moran.

ILUSTRACIÓN DEL PROTOCOLO DE ANÁLISIS EXPLORATORIO

A continuación, se muestran los procedimientos para realizar el análisis exploratorio utilizando datos de conductividad eléctrica aparente a los 30 cm de profundidad (CE30). La descripción de la base de datos usada (*CE30.txt*) se encuentra en Anexo I. La base de datos requiere al menos tres columnas, las primeras dos identifican las coordenadas espaciales bidimensionales (X e Y) y la tercera corresponde a la variable medida.

CONVERSIÓN DE COORDENADAS ESPACIALES

Para convertir las coordenadas geográficas en coordenadas cartesianas UTM se requiere especificar la "faja" o "zona". En este ejemplo corresponde a la zona 21, sur y elipsoide WGS84. Luego se extraen los datos con las coordenadas transformadas. La salida obtenida se muestra en la Figura 1.

	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

	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

Figura 1. Variable conductividad eléctrica aparente (CE30) georreferenciada en coordenadas (X, Y) geográficas (izquierda) y en coordenadas cartesianas (derecha).

ESTUDIO DE LA DISTRIBUCIÓN DE LA VARIABLE Y ELIMINACIÓN DE OUTLIERS

En el histograma de la Figura 2 se observa asimetría derecha en la distribución de los datos. La asimetría también puede advertirse con los estadísticos de posición, dado que la media (23.84 mS m^{-1}) es mayor que la mediana (22.60 mS m^{-1}) y el coeficiente de asimetría de 0.82. En el gráfico box-plot se observan valores extremos de la variable que se encuentran por encima de la media + 3 SD. En la Figura 3 se presenta el histograma y box-plot luego de la eliminación de los *outliers*. Para la variable en análisis, se eliminaron durante la depuración 48 casos que representan un 1% del total de sitios (n=6425) con mediciones.

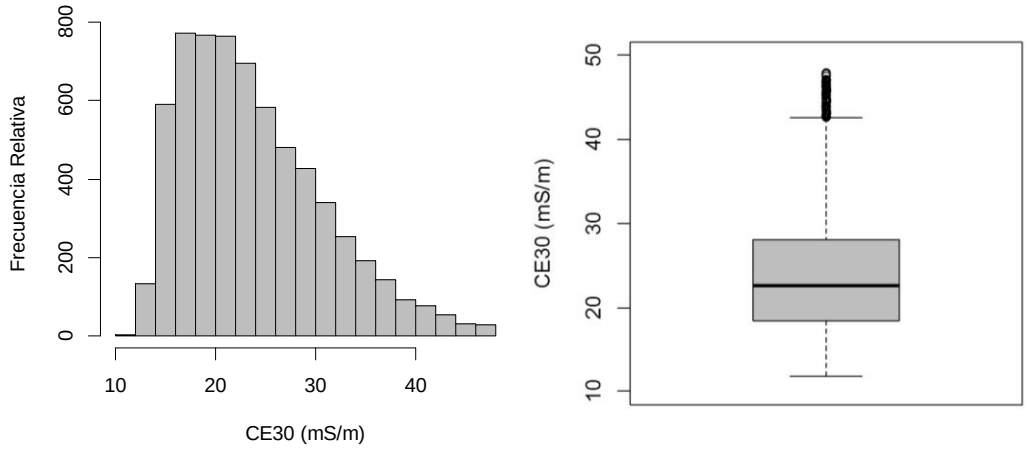


Figura 2. Histograma (izquierda) y box-plot (derecha) de datos de conductividad eléctrica aparente a 30 cm de profundidad (CE30) previo a la eliminación de *outliers*. Coeficiente de asimetría=0.82.

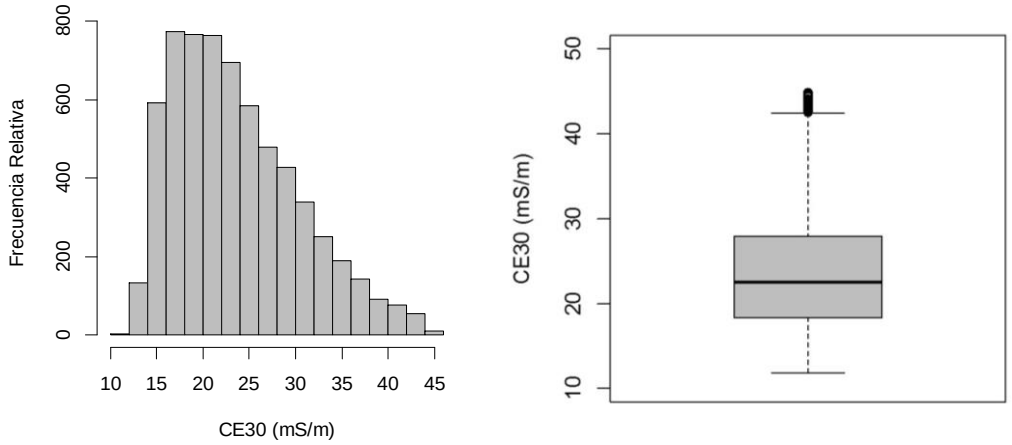


Figura 3. Histograma (izquierda) y box-plot (derecha) de datos de conductividad eléctrica aparente a 30 cm de profundidad (CE30) luego de la eliminación de *outliers*. Coeficiente de asimetría=0.72.

ELIMINACIÓN DE INLIERS

Se consideraron como puntos vecinos a aquellos puntos contiguos ubicados entre los 0 y 25 m de distancia. Calculando el índice local de Moran y su significancia estadística (ajustando los valores-*p* por el criterio de Bonferroni), se identificaron 12 potenciales *inliers* (Figura 4).

Caso	Ii	E.Ii	Var.Ii	Z.Ii	Pr(z<0)
1098	-1.401	-0.000157	0.050	-6.276	<0.0001
1962	-3.033	-0.000157	0.055	-12.885	<0.0001
2077	-2.651	-0.000157	0.100	-8.389	<0.0001
5548	-1.187	-0.000157	0.055	-5.045	<0.0001
6362	-0.944	-0.000157	0.077	-3.406	0.0046

Figura 4. Valores del índice de Moran local observado (Ii), su valor esperado (E.Ii), la varianza (Var.Ii) y el valor-p (Pr(z < 0)) de los primeros 5 *inliers* detectados para la variable conductividad eléctrica aparente a 30 cm de profundidad.

La Figura 5 muestra el gráfico de dispersión de Moran y el ajuste de un modelo de regresión. Los puntos influyentes de la regresión son identificados usando diferentes estadísticos de diagnóstico como DFBETAS (dfb.1_ para la ordenada al origen y dfb.x para la pendiente), DFFITS (dffit), Covratio (cov.r), distancia de Cook (cook.d) y leverage (hat). Un punto se determina como influyente si al menos uno de los estadísticos así lo considera (datos marcado con * en la salida de R) (Figura 6). En la Figura 5 los puntos negros con forma romboidal fueron identificados como influyentes y se los considera como *inliers*.

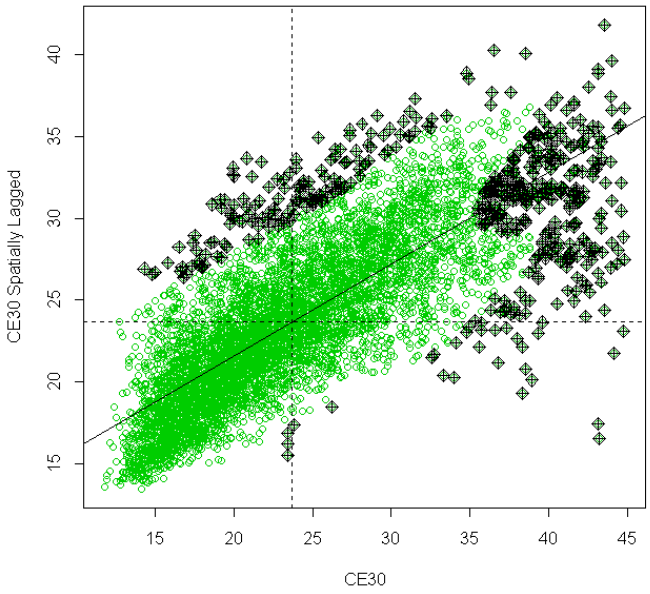


Figura 5. Gráfico de dispersión de Moran para la variable conductividad eléctrica aparente a 30 cm de profundidad (CE30). Puntos negros representan *inliers*.

Caso	dfb.1_	dfb.x	dffit	cov.r	cook.d	hat
1098	0.10	-0.12	-0.13_*	1.00_*	0.01	0.00
1962	0.19	-0.22	-0.23_*	0.99_*	0.03	0.00_*
2077	0.18	-0.21	-0.22_*	0.99_*	0.02	0.00_*
5548	0.10	-0.12	-0.13_*	1.00_*	0.01	0.00_*
6362	0.09	-0.10	-0.11_*	1.00_*	0.01	0.00

Figura 6. Criterios de diagnósticos: DFBETAS (dfb.1_ para la ordenada al origen y dfb.x para la pendiente), DFFITS (dffit), Covratio (cov.r), distancia de Cook (cook.d) y leverage (hat) de los primeros 5 *inliers* detectados para la variable conductividad eléctrica aparente a 30 cm de profundidad.

Luego de identificar los *inliers* se procede a eliminarlos. Primero se eliminan los *inliers* detectados con el índice de Moran y posteriormente los identificados por el gráfico de Moran. La nueva base de datos cuenta con 5910 casos, es decir, se eliminaron 467 casos (7% de los datos) respecto a la base sin *outliers*.

MODELACIÓN Y PREDICCIÓN DE VARIABILIDAD ESPACIAL

ÍNDICE 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 ser medida a nivel global 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 entre sí 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 (sin autocorrelación).

Para cuantificar la magnitud de la estructuración espacial en una variable, existen estadísticos como el índice global de Moran (IM) (Moran, 1948). El IM 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 próximo a cero significa que no existe un patrón espacial o que la dispersión de las observaciones en el espacio es completamente aleatoria. Para calcular el índice de Moran (MI) 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. El cálculo del IM al igual que el del índice local de Moran requiere la definición de una matriz de ponderación espacial. Para este paso la red de vecinos puede ser definida utilizando la distancia Euclídea.

Para evaluar la significancia estadística del IM se utilizan procedimientos de simulación del tipo Monte Carlo. Las ubicaciones son permutadas para obtener la distribución de los índices bajo la hipótesis nula de distribución aleatoria. El índice de Moran se obtiene en R utilizando la librería

"spdep" (Bivand, 2014) que provee no solo el valor del índice sino también la significancia estadística.

MODELACIÓN DE LA VARIABILIDAD ESPACIAL

La teoría de variables regionalizadas define funciones para modelar variabilidad espacial denominados semivariogramas (Cressie, 1993; Matheron, 1971). Bajo este marco teórico, cada dato espacial es una realización de un proceso aleatorio y existe una distribución de probabilidad asociada al mismo. Para procesos continuos asume que estas distribuciones de probabilidad son normales y tienen la misma media y varianza (estacionariedad de primer y segundo orden). Una forma de verificar el supuesto de igual media es realizando regresiones de la variable repuesta con las coordenadas geográficas del sitio. En el caso de encontrar una relación significativa con alguna coordenada, es decir una tendencia longitudinal o latitudinal, se recomienda descontar esa tendencia trabajando con los residuos del modelo de regresión para analizar la variabilidad espacial no asociada a tendencias sistemáticas que se reconocen *a priori*.

Bajo este enfoque geoestadístico, el primer paso para analizar variabilidad espacial es construir un semivariograma empírico. La función semivariograma de un proceso estacionario, denotado por $\gamma\left(\mathbf{s}_i-\mathbf{s}_j\right)$, es función de la diferencia entre las coordenadas $\left(\mathbf{s}_i-\mathbf{s}_j\right)$ y puede expresarse como:

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

donde **h** es la distancia espacial entre las observaciones $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$ que se suponen sobre un espacio continuo y la función $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ es también conocida como semivarianza (mitad de la varianza de las diferencias entre observaciones separadas en el espacio por una distancia o lag **h**).

Los parámetros de la función semivariograma son: la varianza *nugget* o efecto pepita (C_0), la varianza estructural (C) o "partial sill" y el rango (R). C_0 es la ordenada al origen del semivariograma. Este parámetro representa la suma de errores aleatorios o no espaciales, o de errores asociados con la variabilidad espacial a escalas más finas que la usada para realizar las mediciones. Un alto valor de C_0 indica que la mayoría de la variación ocurre en distancias más cortas que la mínima distancia que separa dos observaciones en la grilla de estudio. La asíntota (C) es también llamada umbral del semivariograma. La varianza umbral se obtiene sumando las varianzas antes mencionadas ($C_0 + C$) y es la varianza de observaciones independientes. El rango es el lag o distancia h en el cual la asíntota es alcanzada. Observaciones $Z(\boldsymbol{s}_i)$ y $Z(\boldsymbol{s}_j)$ para las cuales su interdistancia es mayor al rango R se consideran no correlacionadas (Figura 7). Cuando el semivariograma alcanza la meseta asintóticamente (semivariograma 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. Esto frecuentemente se produce cuando el proceso tiene tendencias en la media o cuando el *lag* más grande para el cual el semivariograma puede ser estimado es menor que R (problema de tamaño de grilla).

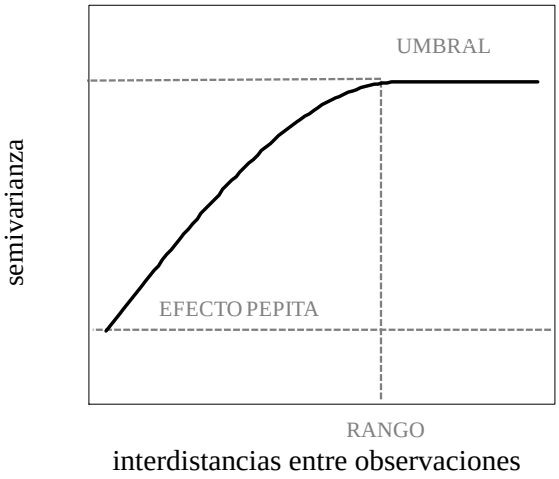


Figura 7. Semivariograma. Se representan los tres parámetros que lo definen: rango, umbral y efecto pepita.

Una medida del grado de estructuración espacial, que suele ser usada en casos donde las estimaciones de los parámetros del semivariograma se realizan con bajo error, es la varianza estructural relativa (RSV).

$$RSV = \left(\frac{\mathrm{C}}{\mathrm{C}+C_0}\right) \times 100\% \qquad (3)$$

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. 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 ≥ 75% alto.

Para obtener estimaciones de la función semivariograma para cualquier interdistancia perteneciente al dominio espacial estudiado, sobre el semivariograma empírico se ajusta un modelo teórico de semivariograma. Las funciones que sirven como modelos de semivariograma deben ser condicionalmente definidas positivas. Existen distintos modelos teóricos para funciones semivariogramas. Los más usados son: modelo exponencial, modelo esférico y el modelo gaussiano (Tabla 1).

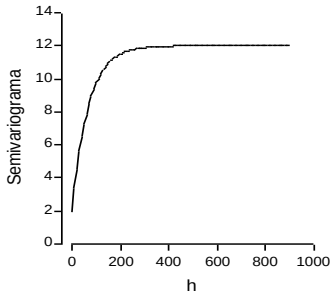
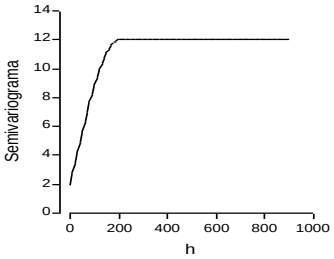
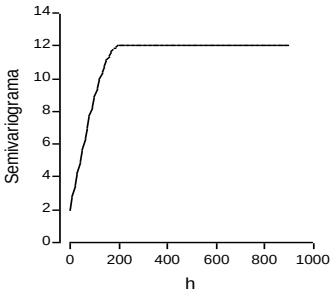
El semivariograma de un modelo sólo *nugget* es también conocido como semivariograma de un proceso de ruido blanco, donde las observaciones se comportan como muestras aleatorias, con igual media y varianza sin correlación espacial. El modelo sin estructura espacial suele ser el modelo de mejor ajuste cuando la menor distancia de muestreo en los datos es mayor que el rango del proceso espacial subyacente (problema de grilla).

El modelo esférico tiene dos características principales: un comportamiento lineal cerca del origen y el hecho que a la distancia R el semivariograma encuentra la meseta. Por el contrario, el modelo exponencial se aproxima a la meseta del semivariograma (C) de manera asintótica. En la parametrización mostrada en la Tabla 1, 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 rango práctico 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.

Este semivariograma exhibe un comportamiento cuadrático cerca del origen y modela correlaciones de rango corto, que son usualmente, más altas que las de otro modelo de media constante con el mismo rango práctico. La

diferencia entre el semivariograma gaussiano y el exponencial es el exponente cuadrado. El modelo gaussiano es el más continuo cerca del origen. El rango práctico suele parametrizarse como $\sqrt{3}$ R.

Tabla 1. Funciones de semivariograma para el modelo, exponencial, esférico y gaussiano. Con C_0=2, C=10 y R=200.

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

Es importante notar que cuando 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 de los citados que tiene un umbral verdadero, ya que tanto el modelo exponencial como el gaussiano alcanzan el umbral de forma asintótica.

El tamaño del conjunto de datos a partir del cual el modelo de semivariograma es ajustado depende del número de *lags* 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. Journel y Huijbregts (1978) recomiendan usar *lags* menores a la mitad del máximo *lag* del conjunto de datos.

Los modelos de semivariograma son no lineales. A diferencia de los modelos lineales donde el método de estimación por mínimos cuadrados garantiza una solución óptima y estable, para los modelos no lineales la optimalidad del método de ajuste depende no sólo del modelo, sino también de las características de los datos que están siendo ajustados. La optimización no lineal es un tópico complejo y existen varios métodos para su alcance. Entre estos el método de mínimos cuadrado ponderados (WLS) suele ser el elegido para la estimación de funciones de semivariograma.

Alternativamente, la estimación de los parámetros del semivariograma puede hacerse simultáneamente a la estimación de tendencia para la media bajo la teoría de los modelos lineales mixtos (MLM) (West *et al*., 2007) con métodos de estimación basados en verosimilitud. La estimación de los parámetros de varianza y covarianza del MLM, que son 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 que reflejan tendencias a gran escala en una o más dimensiones. La estimación de parámetros se puede realizar por los métodos de máxima verosimilitud (ML) (Searle *et al*., 1992) o por máxima verosimilitud restringida (REML) (Patterson y Thompson, 1971). Bajo este marco teórico se ajusta un modelo directamente sobre los datos y no sobre las semivarianzas como en la geoestadística clásica. La estimación REML de la estructura de covariación espacial se define considerando que la misma es función de la distancia entre la separación de las observaciones. Las más utilizadas para datos de suelo son las funciones espacial esférica, exponencial y gaussiana (Schabenberger y Gotway, 2004) los que contienen parámetros relacionados a los del semivariograma del mismo nombre.

Al ajustar distintos MLM a un mismo conjunto de datos, es necesario utilizar criterios para la comparación de los ajustes, para ello se usan criterios de información. En el marco de los MLM, los criterios de información se basan en el logaritmo de la verosimilitud residual (log-reslikelihood) y aplican una función de penalización debida a la cantidad de parámetros del modelo ajustado. Un menor valor del criterio indica un "mejor" ajuste. Los criterios de información de Akaike (AIC) (Akaike, 1973) y el criterio bayesiano (BIC) (Schwarz, 1978) son los más usados en la selección de un MLM.

Otra de las herramientas estadísticas utilizadas para la selección de modelos 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 en sus parámetros. Utiliza el valor de la función de verosimilitud evaluada en las estimaciones ML o REML

de los modelos comparados. El modelo más general o con más parámetros, abarca tanto la hipótesis nula y alternativa, es denominado modelo de referencia. El segundo modelo, más simple, satisface la hipótesis nula (parámetros igual a cero) 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 no contiene aquellos que se suponen podrían ser iguales a cero. Si el estadístico LRT es suficientemente grande, hay evidencias para rechazar el modelo reducido y preferir el modelo de referencia o modelo más parametrizado. Si los valores de verosimilitud de los dos modelos están muy cerca, el estadístico LRT será pequeño sugiriendo evidencia a favor del modelo reducido.

Aun cuando la diferencia en las estimaciones logradas tanto con la aproximación basada en técnicas geoestadísticas como con los MLM puede ser poca, 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 que la tendencia fuese significativa, será necesario descontar la tendencia y trabajar con los residuos del modelo. 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 esa tendencia.

En el contexto de los MLM, es posible obtener las medias ajustadas por el modelo de correlación espacial para representar las tendencias espaciales. Las medias ajustadas podrían diferir respecto a las media de la variable sin ajustar.

Utilizando la varianza umbral ($C+C_o$) puede obtenerse la desviación estándar de los datos y junto a la media ajustada calcular un coeficiente de variación como:

$$CV = \frac{\sqrt{C + Co}}{Media} \times 100$$

PREDICCIÓN Y MAPEO DE LA VARIABILIDAD ESPACIAL

La técnica utilizada en geostadística para realizar interpolaciones espaciales y poder predecir los valores de la variable en sitios no muestreados se denomina kriging. El método de kriging proporciona el mejor estimador lineal para el valor de la variable en un sitio, suministrando además un error de estimación conocido como varianza de kriging, que depende del modelo de variograma obtenido y de las localizaciones de los datos originales. La varianza kriging brinda la posibilidad de analizar la calidad de las estimaciones.

El método kriging se basa en el conocimiento del comportamiento de la variable en el espacio; la covarianza entre cualquier punto muestral y un punto cuyo valor debe predecirse, decrece a medida que la distancia entre la observación muestral y el punto aumenta. Una función usada para modelar este fenómeno es la función inverso de la distancia (Gallardo y Maestre, 2008). Las distancias usadas en las técnicas de kriging son distancias estadísticas, en contraste con las distancias geométricas utilizadas en otros métodos de interpolación espacial. El método kriging evita muestras redundantes, ponderando de forma distintas muestras que están muy cerca entre sí y proceden de la misma región que muestras que estén en lados opuestos al punto que se quiere asignar un valor por interpolación. Los parámetros del semivariograma elegido tienen importancia a la hora de asignar ponderadores a

las muestras que rodean el punto a interpolar. El rango del semivariograma también influye en la interpolación espacial. Los puntos que se encuentran respecto al punto a predecir, a una distancia superior al valor del rango, tienen mínimo impacto sobre la capacidad predictora.

Entre los métodos de interpolación espacial que utilizan todos los datos simultáneamente se destacan los métodos de kriging ordinario, simple y universal. En el kriging ordinario la media de la variable es estimada localmente. En caso de conocer la media de la variable, hecho que raramente ocurre, se utiliza el kriging simple. En el kriging universal la media es estimada y se incluye también la influencia de una tendencia espacial de los datos.

La predicción asignada a los puntos incógnita puede realizarse de manera puntual (kriging puntual) o definiendo bloques (kriging en bloques) (Webster y Oliver, 2007). La interpolación puntual es la estimación del valor de la variable en el punto incógnita, mientras que la interpolación por bloques estima la media de puntos de un área predeterminada que rodea al punto incógnita. La interpolación por bloques (que produce un "suavizado" de las estimaciones) suele correlacionar mejor con los valores verdaderos (Isaaks y Srivastava 1989).

ILUSTRACIÓN DE MODELACIÓN Y PREDICCIÓN ESPACIAL

Se utilizarán los datos de conductividad eléctrica aparente a los 30 cm de profundidad (CE30) (Anexo 1) disponibles en la base de datos *datos2.txt*, la cual se obtuvo luego de realizar análisis exploratorio sobre la base *CE30.txt*.

CÁLCULO DEL ÍNDICE DE MORAN

Para la conformación de la matriz de ponderadores espaciales se definieron los vecindarios de cada sitio mediante una red de conexión construida en base a la distancia Euclidea. Se consideraron sitios vecinos a aquellos contiguos ubicados hasta 25 m de distancia. Los resultados del cálculo del IM muestran que la variable CE30 presentó autocorrelación espacial significativa y positiva (0.58, p=0.01) (Figura 8).

```
Monte-Carlo simulation of Moran's I

data:  datos2$CE30
weights: lw_1
number of simulations + 1: 1000

statistic = 0.5829, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater
```

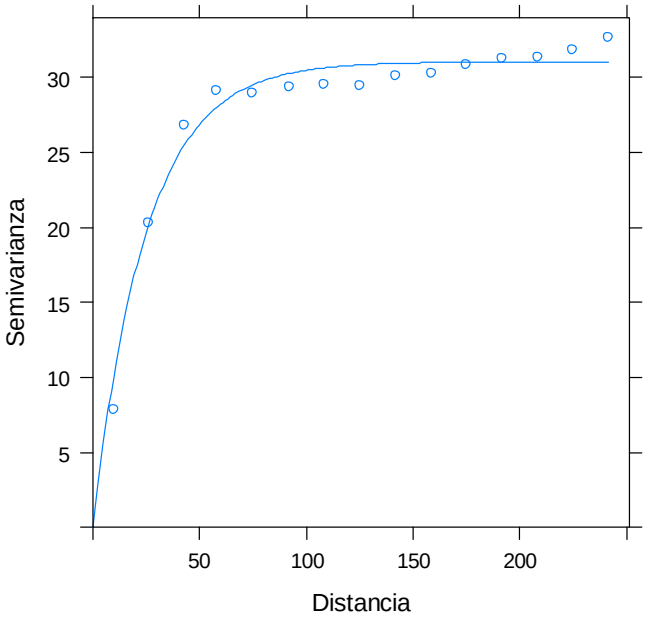
Figura 8. Índice de autocorrelación espacial de Moran y su significancia estadística para la variable conductividad eléctrica aparente (CE30). Salida del software R.

AJUSTE DE UN MODELO DE VARIABILIDAD ESPACIAL

Se obtuvo el semivariograma empírico y sobre éste se ajustó, por WLS, los modelos de semivariograma exponencial y esférico. La suma de cuadrado del error (SCE) fue el criterio usado para la selección del mejor modelo. Para ambos se calculó la RSV como medida del grado de estructuración espacial capturado por el modelo. El análisis geoestadístico fue realizado con la librería "gstat" (Pebesma, 2004) del software R.

El modelo de mejor ajuste (menor SCE) para CE30 fue el exponencial. En la Figura 9 se muestran el semivariograma empírico y el semivariograma

teórico ajustado, así como las estimaciones de los parámetros para la variable CE30 y el modelo exponencial. La variable CE30 presentó una estructura espacial fuerte.



	model	psill	range
1	Nug	0.00000	0.00000
2	Exp	31.05177	24.91754

Figura 9. Semivariograma empírico (puntos) y teórico (línea) de la variable conductividad eléctrica aparente a 30 cm de profundidad (CE30). Abajo se presenta la salida del software R que contiene los parámetros del semivariograma teórico ajustado: Nugget (C_o=0), Sill (C=31.05) y Range (24.92m) o Rango Practico (R_p=24.92m×3). Nota: bajo la columna "psill", para la fila Nugget, se debe leer el valor C_o.

AJUSTE DE UN MODELO LINEAL MIXTO A DATOS ESPACIALES

Cuando se realiza el ajuste de un MLM a grandes bases de datos los métodos pueden tener problemas de convergencia debido a las dimensiones de la matriz residual utilizada en las estimaciones. Una sugerencia es realizar el ajuste de los modelos sobre una muestra de los datos originales. En esta ilustración se trabajó con el 10% de los datos (n=500) seleccionados al azar.

Se compararon los ajustes obtenidos (vía REML) de los modelos de correlación espacial exponencial y esférico (ambos sin y con efecto *nugget*), incorporando también el modelo de errores independientes correlación espacial nula (sólo *nugget*). Estos ajustes se hicieron para un modelo de medias que incluyó 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 espaciales se usaron los criterios AIC y BIC. 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, basada en estimaciones máximo verosímil (ML). Así, para el modelo de correlación espacial seleccionado se compararon modelos con y sin tendencia a gran escala para las coordenadas (X, Y) (Tabla 2). Los análisis fueron realizados con la librería “geoR” (Ribeiro Jr. y Diggle, 2001) del software R.

Los índices AIC y BIC muestran que el modelo de correlación espacial exponencial proveyó el mejor ajuste (Tabla 2). La prueba LRT que compara los modelos con y sin tendencia, no fue estadísticamente significativa ($p>0.05$), por lo que se seleccionó el modelo más sencillo, es decir sin tendencia espacial (Tabla 3).

Tabla 2. Criterios de información sobre ajustes de MLM de correlación espacial para variable conductividad eléctrica aparente a 30 cm de profundidad (CE30).

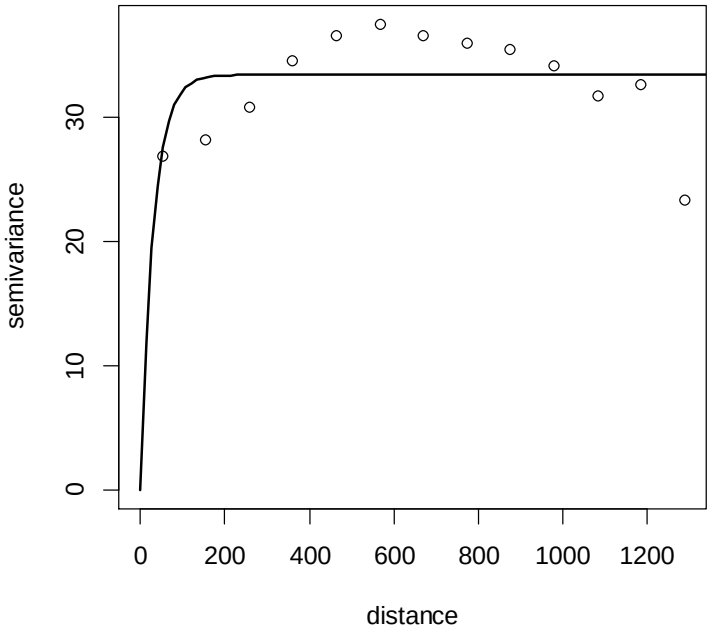
Modelo	AIC	BIC
1. Exponencial + Nugget	2992	3017
2. Esférico + Nugget	3006	3032
3. Exponencial	2990	3011
4. Esférico	3066	3088
5. Errores independientes	3159	3176

Tabla 3. Test del cociente de verosimilitud (LRT) basada en los estimadores ML para evaluar tendencia a gran escala.

Modelo	Loglik	Test	L. Ratio	p-valor
1. Exponencial + X + Y	-1499.6			
2. Exponencial	-1502.3	1 vs. 2	5.40	0.9999

Los semivariogramas empírico y teórico ajustados y las estimaciones de los parámetros obtenidos a partir del MLM seleccionado (exponencial) se muestran en la Figura 10. Utilizando la varianza umbral ($C+C_o$) y la media ajustada (denominada como beta en la salida del R) se determinó que el coeficiente de variación para estos datos es del 25%.

Los parámetros del semivariograma fueron similares entre una (WLS) y otra (REML) estrategia de estimación. A partir de estos parámetros estimados, se realizó la interpolación mediante Kriging en bloque y se obtuvieron mapas de contorno para cada variable. La dimensión del bloque utilizada fue de 40m × 40m. Los mapas generados muestran variabilidad espacial (Figura 11). La zona más oscura se corresponde con el área de menor CE30.




```
Summary of the parameter estimation
-----------------------------------
Estimation method: restricted maximum likelihood

Parameters of the spatial component:
   correlation function: exponential
     (estimated) variance parameter sigmasq (partial sill) =  33.43
     (estimated) cor. fct. parameter phi (range parameter) =  30.58
Parameter of the error component:
      (fixed) nugget = 0
```

Figura 10. Semivariograma empírico (puntos) y teórico (línea) de la variable conductividad eléctrica aparente a 30 cm de profundidad (CE30). Abajo se presenta la salida del software R con los parámetros del semivariograma teórico ajustado mediante Modelos Lineales Mixtos. Nugget (0), Sill (C=33.43) y Range (30.58m) o Rango Practico (R_p=30.58m×3).

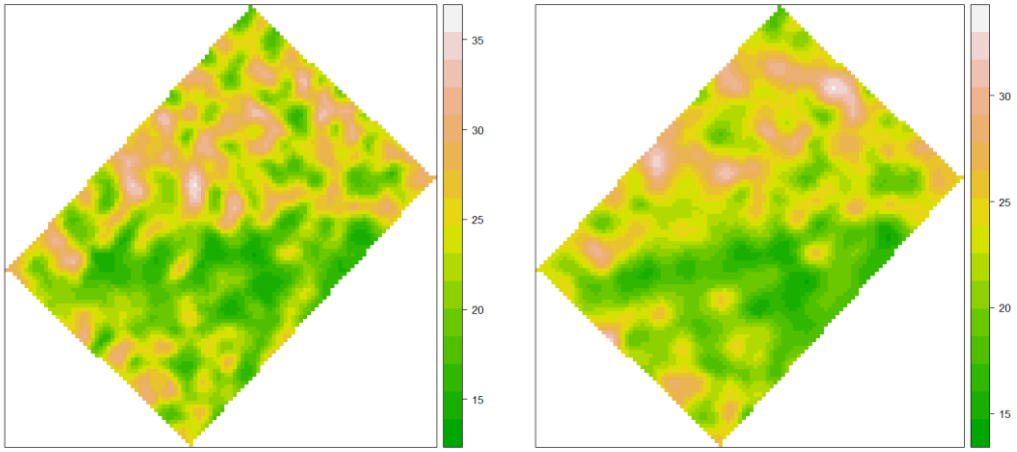


Figura 11. Mapas de variabilidad espacial de conductividad eléctrica aparente a 30 cm de profundidad obtenidos mediante interpolación por kriging ordinario utilizando parámetros del semivariograma estimados con goestadística clásica (izquierda) y con modelos lineales mixtos (derecha).

La estructura espacial en estos datos de CE30 es de magnitud fuerte. Bajo estructuras espaciales fuertes y con rangos amplios se genera mapas de variabilidad con zonas contiguas grandes y contiguas. Con menor rango, los mapas de variabilidad presentan mayor fragmentación o estructura de "parches".

TUTORIAL PARA EL PROCESAMIENTO DE DATOS ESPACIALES

AMBIENTE DE TRABAJO DE INFOSTAT Y SU INTERFAZ CON R

InfoStat (http://www.infostat.com.ar/), es un programa estadístico desarrollado en el ambiente Windows que ofrece una interfaz avanzada para el manejo de datos basada en el difundido concepto de planilla electrónica. Permite importar y exportar bases de datos en formato texto, dbase y Excel, entre otros. Posee rápido acceso a herramientas para el manejo de datos como por ejemplo editar fórmulas, transformar, clasificar y categorizar variables. Las capacidades de copia y pegado permiten trasladar fácilmente tablas, resultados y gráficos a otras aplicaciones Windows.

InfoStat ofrece distintas herramientas para explorar su información de manera sencilla, intuitiva y amigable. Al abrir InfoStat, se visualizará una barra de herramientas localizada en la parte superior de la ventana del programa, la que contiene los siguientes menús: **Archivo**, **Edición**, **Datos**, **Resultados**, **Estadísticas**, **Gráficos**, **Ventanas**, **Aplicaciones**, **Ayuda** y **[R]**. El menú **[R]** es el vínculo a la interfaz con R y solo aparece cuando InfoStat, R y el programa que los vincula (statconnDCOM) han sido correctamente instalados (ver "¿Cómo instalar R?" en el menú **Ayuda** de InfoStat).

R Project (http://www.r-project.org/), más conocido como R, es un lenguaje de programación libre y gratuito que ha sido desarrollado principalmente para análisis estadístico. Permite generar algoritmos (conjunto

de instrucciones) utilizando funciones o comandos para obtener determinados procesamientos de datos. R es un lenguaje orientado a objetos. Esto significa que R lee, genera y trabaja sobre objetos creados por el usuario, o que lee desde otros ambientes. El ambiente o entorno de trabajo es aquel en el que se incluyen todos los objetos relacionados con un trabajo específico. Mediante su interfaz con R, InfoStat permite ejecutar y modificar estos algoritmos de manera más sencilla facilitando el manejo de los objetos, su procesamiento y visualización.

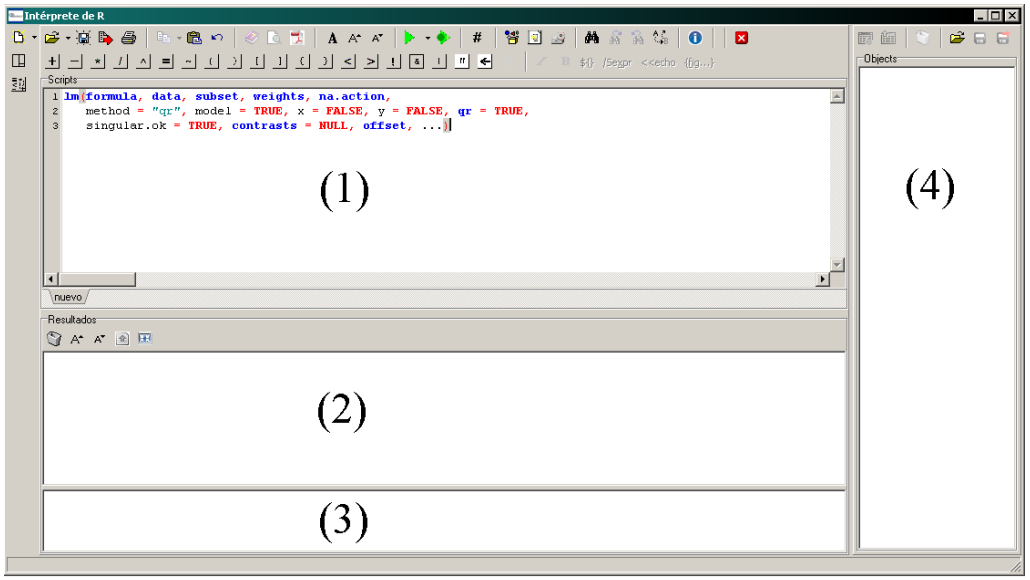
Para ajustar, por ejemplo, una modelo lineal en R se debe invocar la función `lm()`, que tiene la siguiente sintaxis:

```
>lm(formula, data, subset, weights, na.action, method = "qr",
model=TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE,
contrasts = NULL, offset, ...)
```

Cada término entre los paréntesis y delimitado por comas es un argumento de la función. Por ejemplo, el argumento `formula` debe indicar cuál es la variable dependiente y cuál/es la/s independiente/s del modelo lineal a ajustar, el argumento `data` indica cuál es el objeto (la base de datos) sobre el que se ajustará el modelo. La función `help()`, cuyo argumento será el nombre de alguna función, permite acceder a información sobre funciones disponibles.

El intérprete, disponible en InfoStat, permite trabajar con R desde InfoStat, escribiendo y ejecutando **scripts**. Un **script** o código es una secuencia ordenada de instrucciones, ingresadas como texto editable. Para acceder al intérprete de R en InfoStat se hace click sobre el menú **[R]**. Con las teclas *Ctrl+R* se pude ejecutar el renglón del **script** donde el cursor esté posicionado,

las flechas verdes en el centro del margen superior de la pantalla también sirven para ejecutar **scripts**. La zona de trabajo se presenta en cuatro paneles:



1) **Editor de Scripts**, donde se editan las instrucciones en forma de texto.

2) **Resultados o Salidas**, donde se despliegan los resultados en forma de texto. En caso de superar el tamaño del panel, aparecerá una barra de navegación en el margen derecho. Los resultados gráficos se abrirán en una nueva ventana.

3) **Mensajes**, información adicional sobre paquetes correctamente instalados y mensajes de error se despliegan en este panel.

4) **Objetos**, una lista con todos los objetos creados se muestran en este panel. Las opciones de este panel permiten manipular objetos, levantar y/o exportar bases de datos al entorno de InfoStat, eliminar objetos, cambiar su nombre, etc.

Se pueden cargar **scripts** previamente escritos o bien escribir nuevos **scripts**. El texto se destaca con distintos colores **palabras clave** (azul) de R, **números** (violeta), símbolos (rojo), **palabras reservadas** (rojo trazo grueso) y **comentarios** (verde).

Los objetos de R pueden ser de diferentes clases según la información que contienen. Una forma sencilla para consultar la clase de un objeto es invocando la función `class()`. Existen muchas clases de objetos. Por ejemplo, un `dataframe` es una clase de objeto que representa una lista de vectores y/o matrices de la misma longitud. Una tabla de datos de InfoStat se exporta al intérprete de R como un objeto des tipo `dataframe`.

PROTOCOLO DE ANÁLISIS DE VARIABILIDAD ESPACIAL

En esta sección se expone un tutorial para implementar los métodos del protocolo a través del intérprete de R en InfoStat. El código será presentado en `color azul` y los resultados en `color rojo`. Todas las herramientas estadísticas usadas están contenidas en "paquetes" de R que se instalan con las sentencias presentadas a continuación.

INSTALACIÓN Y CARGA DE PAQUETES

Utilizando la siguiente sentencia, el intérprete de R intentará descargar los paquetes desde la web. Para ello es necesario contar con conexión a internet.

```
install.packages
("spdep","rgdal","geoR","gstat","ade4","e1071","sampling","nlme"
,"lsmeans")
```

Los paquetes también pueden descargarse desde el menú **[R]** opción "instalar paquete desde la WEB". Una vez instalados, quedarán guardados de forma permanente en la carpeta de R del ordenador. Por lo tanto, no es necesario que se reinstalen cada vez que se deseen usar. La carga de los paquetes es realizada corriendo el siguiente comando:

```
library(spdep)
library(rgdal)
library(geoR)
library(gstat)
library(ade4)
library(e1071)
library(sampling)
library(nlme)
library(lsmeans)
```

CARGA DE DATOS

Para cargar una base de datos desde InfoStat puede utilizarse el botón correspondiente del margen superior del panel de objetos o utilizar la función `read.table()`. Esta función permite abrir distintos tipos de archivos entre ellos aquellos de extensión *.txt*. El siguiente ejemplo crea un objeto llamado "datos" de clase `dataframe` que se carga desde un archivo de texto (*.txt*). El argumento `header=TRUE` indica que la primera fila de los datos contiene los nombres de las columnas y `sep="\t"` indica que están separados por tabulaciones.

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

Para visualizar el contenido de un objeto, basta con escribir su nombre. En la ventana de resultados se despliega el contenido del objeto.

```
datos
```

```
      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
```

CONVERSIÓN DE COORDENADAS GEOGRÁFICAS

Los siguientes comandos permiten transformar las coordenadas geográficas de los datos a coordenadas cartesianas UTM. La función `coordinates()` transforma el `dataframe` en un objeto de datos espaciales e indica al software que las columnas "x" e "y" son coordenadas espaciales. Esta transformación es necesaria para correr funciones estadísticas que solo trabajan sobre objetos de datos espaciales (clase `SpatialPixelsDataFrame`).

```
coordinates(datos) <- ~x+y
```

La función `CRS`, *Coordinate Reference System*, tiene una variedad de argumentos que permiten hacer referencia a diferentes sistemas de proyecciones. La proyección `longlat` es utilizada en esta aplicación. Esta proyección solo admite valores de longitud mayores a -180 y menores a 180 y valores de latitudes mayores a -90 y menores a 90. El *datum* especificado será WGS84.

```
proj4string(datos) <- CRS("+proj=longlat +datum=WGS84")
```

La función `spTransform` permite convertir las coordenadas. Cundo se realiza la transformación del sistema de proyección geográfico a cartesiano, es necesario indicar a que zona o faja pertenecen los datos bajo análisis. Al igual

que en la sentencia anterior, se debe indicar el *datum* y *elipsoide* que en ambos casos corresponde a WGS84.

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

Para trabajar con otras funciones que no pertenecen a los paquetes de estadística espacial se necesita objetos de clase `dataframe`. Por lo tanto, es necesario aplicar la sentencia `as.data.frame` para convertir el objeto "datos" a un `dataframe`. Luego se ordenan las columnas para el objeto resultante contenga primero las columnas de las coordenadas seguido por la variable medida.

```
datos <- as.data.frame(datos)
datos <- datos[,c("x","y","CE30")]
datos
```

En la ventana de resultados se visualizan los datos con las coordenadas ya transformadas.

```
      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
```

ESTUDIO DE LA DISTRIBUCIÓN DE LA VARIABLE Y ELIMINACIÓN DE DATOS RAROS

DISTRIBUCIÓN DE LA VARIABLE

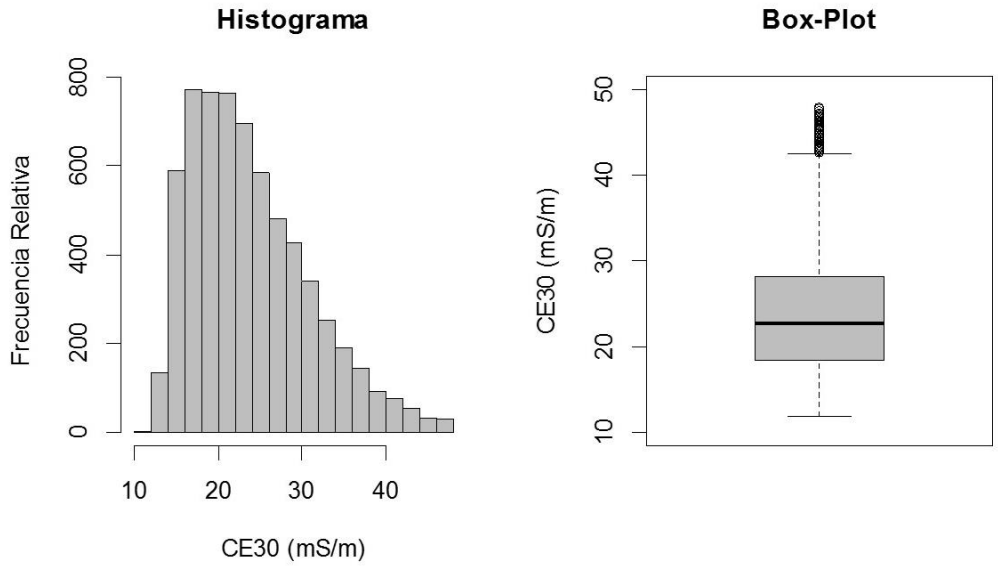
En un `dataframe`, una forma sencilla para obtener medidas resumen de una variable es con la función `summary()`. Se utiliza `$` para hacer referencia a una columna particular dentro de un objeto.

```
summary(datos$CE30)
```

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

Las funciones `hist()` y `boxplot()` realizan gráficos de histogramas y box-plots, respectivamente. Sus múltiples argumentos permiten la edición de cada gráfico. La función `par()` permite dividir la ventana gráfica de R, en el siguiente ejemplo se divide la ventana gráfica de R en dos columnas y una fila.

```
par(mfrow=c(1,2))

hist(datos$CE30,col='grey',nclass=20,main="Histograma",ylab='Fre
cuencia Relativa',xlab='CE30 (mS/m)')
boxplot(datos$CE30,col='grey',ylab='CE30(mS/m)',main="Box-
Plot",ylim = c(10, 50))
```



La función `skewness()` calcula la asimetría. Existen 3 fórmulas para su cálculo (por defecto usa la tipo 3). Para mayor información, se puede utilizar `help()` sobre la función `skewness()`.

```
Asimetria <- skewness(datos$CE30) ; Asimetria
```

```
[1] 0.820089
```

OUTLIERS

Las siguientes instrucciones calculan y crean objetos para la media, el DE y los límites superior (media +3DE) e inferior (media -3DE) con los que pueden detectarse los *outliers*.

```
Media <- mean(datos$CE30)
DE <- sd(datos$CE30)
LI <- Media-3*DE
LS <- Media+3*DE
```

El uso de `[ ]` hace referencia a posiciones dentro de un objeto; el símbolo `|` es un operador lógico para "or". La siguiente instrucción asigna un "dato faltante" o `NA` a aquellos datos dentro de la columna "CE30" del `dataframe` que son mayores al límite superior o menores al límite inferior.

```
datos$CE30[LS<datos$CE30|datos$CE30<LI] <-NA
```

La función `subset()` selecciona datos que cumplen con cierta condición lógica. Una forma de eliminar los datos `NA` del `dataframe` utilizando esta función es la siguiente:

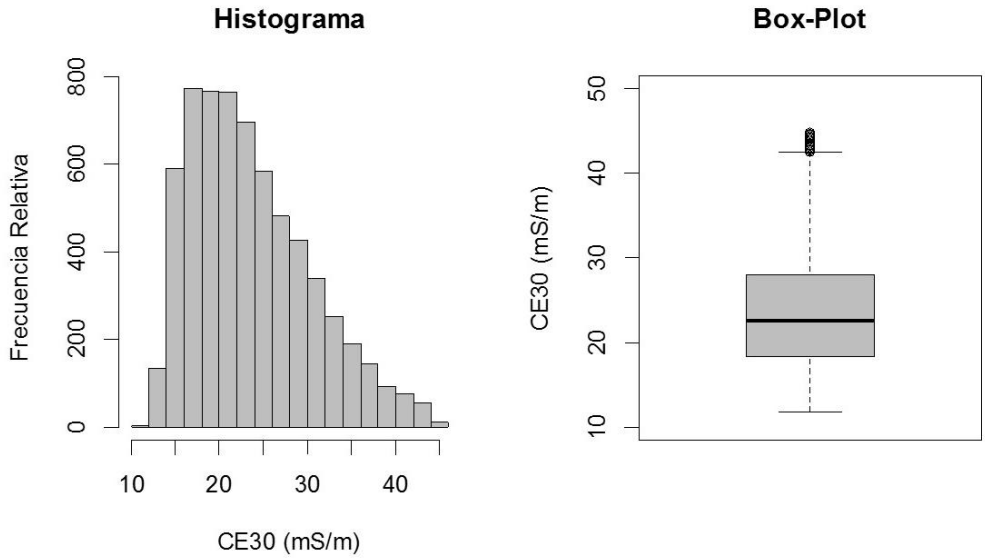
```
datos <- subset(na.omit(datos),select=c(x,y,CE30))
```

Para ver el impacto de la eliminación de *outliers* pueden obtenerse nuevamente las medidas resumen, histograma, box-plot y coeficiente de asimetría.

```
summary(datos$CE30)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
11.80   18.30   22.50   23.67   28.00   44.80
par(mfrow=c(1,2))
```

```
hist(datos$CE30,col='grey',nclass=20,main="Histograma",ylab='Fre
cuencia Relativa',xlab='CE30 (mS/m)')
boxplot(datos$CE30,col='grey',ylab='CE30(mS/m)',main="Box-
Plot",ylim = c(10, 50))
```




```
Asimetria <- skewness(datos$CE30); Asimetria
[1] 0.720552
```

INLIERS

La identificación y eliminación de *inliers* requiere pasos previos. Para crear una matriz de ponderación espacial, se creará un objeto que contenga las coordenadas espaciales. Los corchetes utilizados en la función `coordinates()` indican que la primera y segunda columna dentro del objeto "datos" son las que contienen las coordenadas. La función `dnearneigh` se utiliza para identificar el vecindario de cada sitio utilizando la distancia Euclidea. En este ejemplo, se consideran datos vecinos a aquellos que se encuentran a una distancia Euclidea de 0 a 25 m. La función `nb2listw` transforma el objeto *gri* que contiene las distancias a una matriz de pesos estandarizados por filas (`style = "W"`).

```
cord <- coordinates(datos[,1:2])
gri <- dnearneigh(cord,0,25)
lw <- nb2listw(gri, style = "W")
```

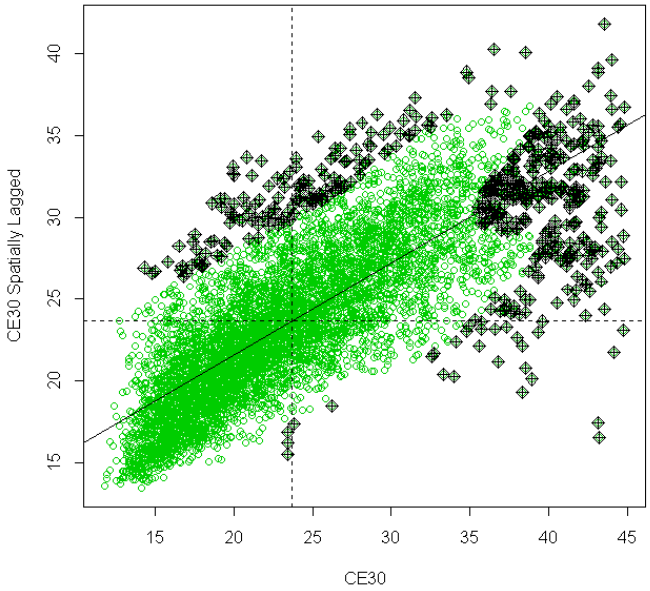
La función `localmoran()` calcula el índice local de Moran que permite identificar potenciales *inliers*. También permite el ajuste de los valores-*p* por el criterio de Bonferroni.

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

```
     Ii            E.Ii         Var.Ii      Z.Ii          Pr(z < 0)
1  -5.857111e-02 -0.0001568381 0.24980645 -1.168738e-01 1.000000e+00
2  -1.191147e-02 -0.0001568381 0.24980645 -2.351837e-02 1.000000e+00
3  -5.130661e-02 -0.0001568381 0.16648538 -1.253590e-01 1.000000e+00
4  -1.717996e-01 -0.0001568381 0.14267936 -4.544069e-01 1.000000e+00
5  -2.226416e-03 -0.0001568381 0.12482484 -5.857757e-03 1.000000e+00
```

El gráfico de Moran permite la identificación de puntos influyentes. La función `moran.plot()` construye el gráfico y devuelve los estadísticos de diagnóstico para cada punto.

```
MP <- moran.plot(datos$CE30, col=3,
lw,quiet=T,labels=F,col=3,zero.policy=F,xlab="CE30", ylab="CE30
Spatially Lagged")
```



Para visualizar en una tabla los puntos potencialmente influyentes y sus estadísticos de diagnóstico puede usarse la función `summary()`.

```
summary(MP)
```

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

     dfb.1_ dfb.x dffit   cov.r   cook.d hat
41   -0.01   0.01  0.02    1.00_*  0.00   0.00
91    0.00   0.01  0.01    1.00_*  0.00   0.00
92   -0.01   0.01  0.01    1.00_*  0.00   0.00
142   0.03  -0.02  0.04    1.00_*  0.00   0.00
148   0.01   0.00  0.03    1.00_*  0.00   0.00
```

Desde el objeto `MP` puede extraerse una matriz de valores lógicos (verdadero/falso) para los estadísticos diagnóstico que identifican un punto como influyente o no.

```
Influ <- MP$is.inf ; Influ
```

Los siguientes `dataframe` pueden concatenarse verticalmente en un mismo objeto: *Datos* (datos obtenidos luego de la eliminación de *outliers*), *ML*

(objeto con valores del índice local de Moran), *Influ* (objeto que identifica cada dato como influyente o no).

```
datos0 <- data.frame(datos,ML,Influ)
```

Posteriormente procedemos a eliminar los datos con Índice de Moran Local negativo y estadísticamente significativos (p<0.05).

```
datos1 <- subset(datos0,datos0[,4] > 0 | datos0[,8]>0.05)
```

La columna `Ii` contiene el Índice de Moran Local y es la cuarta columna de izquierda a derecha en el `dataframe datos0`. La octava columna contiene los valores-*p*. Por ello se utiliza la referencia `datos0[,4]` y `datos0[,8]`. Además note el operador lógico "or" que indica que extraiga los datos que cumplen con las dos condiciones, Índice de Moran Local negativo y estadísticamente significativo.

Existen varias formas de eliminar las filas de la tabla que fueron identificadas como *inliers* con la función `moran.plot`. Una alternativa es usando sentencias lógicas con los operadores "==" y "&" que significan igualdad lógica y "and" respectivamente.

```
datos2 <- datos1[datos1$dfb.1_ == FALSE & datos1$dfb.x == FALSE
& datos1$dffit == FALSE & datos1$cov.r == FALSE & datos1$cook.d
== FALSE & datos1$hat == FALSE, ]
```

La sentencia anterior instruye al software para que cree un objeto llamado `datos2` a partir de las filas del objeto `datos1` cuyas columnas `dfb.1`, `dfb.x`, `dffit`, `cov.r`, `cook.d` y `hat` son simultáneamente iguales a `FALSE`.

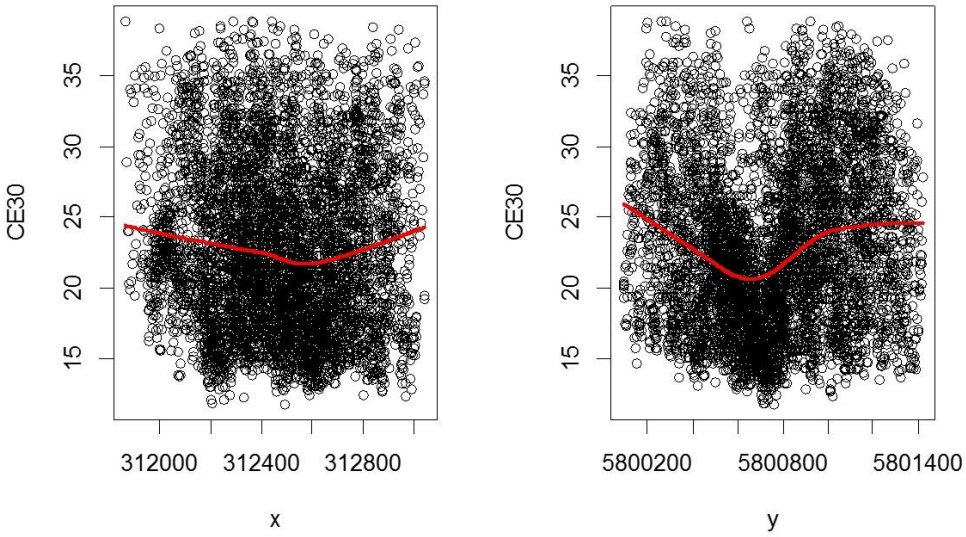
DETECCIÓN DE TENDENCIAS ESPACIALES

Para visualizar tendencias espaciales graficamos la variable en estudio en función de las coordenadas. Si se desea desplegar los gráficos para la coordenada `x` e `y` en una misma ventana gráfica, se puede particionar la ventana en una fila y dos columnas utilizando la siguiente función:

```
par(mfrow=c(1,2))
```

La función `plot()` permite realizar gráficos de dispersión. Además, puede adicionarse una línea de suavizado *lowess* con la función `lines()`. Esta última, realiza el ajuste sobre una ventana gráfica preexistente.

```
plot(datos2$x,datos2$CE30,xlab="x",ylab="CE30")
lines (lowess(datos2$x,datos2$CE30), col = "red", lwd=3)

plot(datos2$y,datos2$CE30,xlab="y",ylab="CE30")
lines (lowess(datos2$y,datos2$CE30), col = "red", lwd=3)
```



Mediante un modelo lineal de regresión, puede ajustarse la tendencia entre la variable en estudio y las coordenadas. Si la tendencia lineal resulta significativa, debería descontarse trabajando con los residuos del modelo, que se obtienen con la función `residuals()`.

```
regresion <-lm(formula = CE30 ~ x + y , data = datos2, na.action
= na.omit)
```

La siguiente línea despliega una tabla resumen del modelo:

```
summary(regresion)

Call:
lm(formula = CE30 ~ x + y, data = datos2, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-11.6080  -4.5743  -0.9379   4.0866  17.3154

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.729e+04  1.594e+03 -10.850   <2e-16 ***
x           -3.072e-03  3.407e-04  -9.017   <2e-16 ***
y            3.150e-03  2.828e-04  11.141   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.783 on 5907 degrees of freedom
Multiple R-squared:  0.02363, Adjusted R-squared:  0.0233
F-statistic: 71.47 on 2 and 5907 DF,  p-value: < 2.2e-16
```

En este caso, si bien los valores-*p* dan significativos ($p<0.05$), se decidió trabajar con la variable original porque el coeficiente de determinación del modelo acusa un ajuste pobre (0.023).

CÁLCULO DEL ÍNDICE DE MORAN

Con las funciones trabajadas anteriormente, se recalcula la matriz de ponderación espacial para el objeto `datos2` depurado en pasos previos.

```
cord_1 <- coordinates(datos2[,1:2])
gri_1 <- dnearneigh(cord_1,0,25)
lw_1 <- nb2listw(gri_1, style = "W")
```

Para realizar el cálculo del Índice de Moran y determinar su significancia estadística mediante simulación Monte Carlo, se utiliza `moran.mc()`. Se debe especificar la variable en estudio, la lista con los pesos de las ponderaciones espaciales y el número de simulaciones.

```
i.moran <- moran.mc(datos2$CE30, listw=lw_1, nsim=999)
i.moran

Monte-Carlo simulation of Moran's I

data:  datos2$CE30
weights: lw_1
number of simulations + 1: 1000

statistic = 0.5829, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater
```

Estos resultados permiten concluir que existe autocorrelación espacial positiva (0.5829) y que esta es estadísticamente significativa (p=0.001).

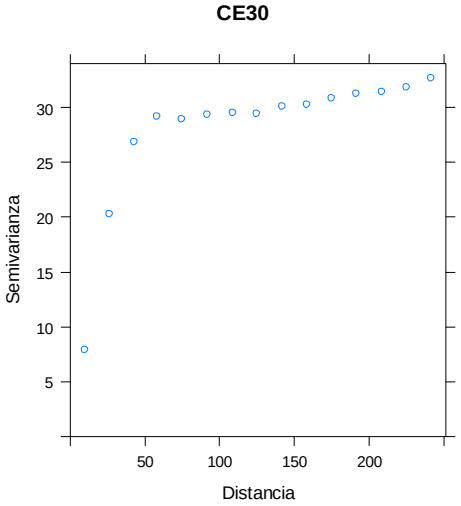
IMPLEMENTACIÓN DEL ANÁLISIS BASADO EN SEMIVARIOGRAMA

Las semivariogramas empíricos pueden obtenerse usando la función `variogram()`. Esta tiene múltiples argumentos, entre ellos una fórmula, un objeto de datos espaciales y una distancia de separación hasta donde los pares de puntos son incluidos en la estimación de semivarianza. Para realizar el análisis es necesario transformar el objeto `dataframe` en uno de clase `SpatialPointsDataFrame`, para ello se ejecuta lo siguiente:

```
coordinates(datos2) <- ~x+y
```

Luego se usa `variogram()` y `plot()` para el ajuste y la presentación grafica del semivariograma empírico.

```
CE30vario <- variogram(CE30~1, datos2, cutoff=250)
plot(CE30vario,main="CE30",xlab="Distancia",ylab="Semivari
anza")
```



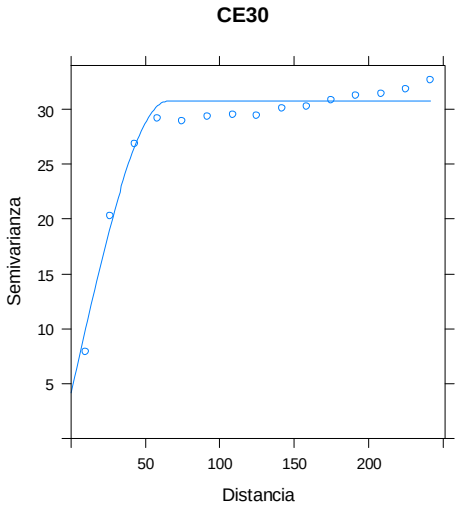
Se ajusta un modelo de semivariograma teórico sobre el semivariograma empírico usando `fit.variogram()` y `vgm()`. La función `vgm()` ajusta el modelo teórico, sus argumentos indican el tipo de modelo a ajustar y los parámetros de ajuste (*partial sill*, rango y efecto *nugget*). Estos parámetros iniciales son de referencia y pueden obtenerse a partir del semivariograma empírico. Cambiar los parámetros modifica la suma de cuadrados del error (SCE). A continuación se ajusta un modelo esférico, con los valores 25, 80 y 10 como parámetros iniciales para estimar el *sill*, *range* y *nugget*, respectivamente.

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

```
  model     psill    range
1   Nug  4.146725  0.00000
2   Sph 26.612088 65.43791
```

El semivariograma empírico y teórico ajustado (esférico) puede graficarse de la siguiente manera:

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



El modelo que mejor ajusta será el de menor SCE. La función `attr()` devuelve atributos de un objeto y puede usarse para consultar la SCE del modelo esférico.

```
attr(Esf_wls, 'SSErr')
```

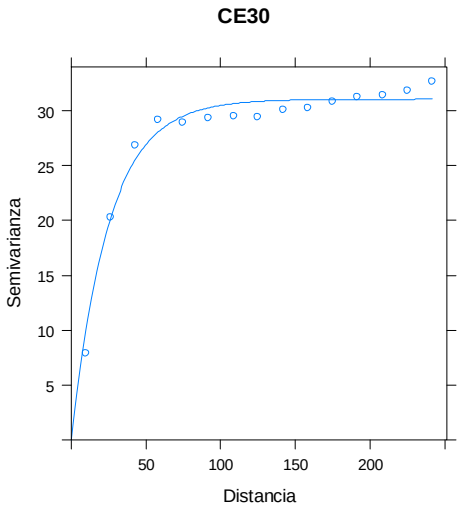
```
[1] 4084724
```

Repitiendo el procedimiento para un modelo exponencial (especificado en la función `vgm()`) se obtiene un menor SCE indicando mejor ajuste.

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

```
  model    psill    range
1   Nug  0.00000  0.00000
2   Exp 31.05177 24.91754
```

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




```
attr(Exp_wls, 'SSErr')
```

```
[1] 2924423
```

AJUSTE DE MODELOS LINEALES MIXTOS A DATOS ESPACIALES

Para evitar problemas de convergencia por la gran dimensionalidad de las bases de datos, se tomó una muestra aleatoria de n=500 sobre la base de datos original de n=5910, usando la función `sample()`. Dado que el proceso de toma de muestras es aleatorio, algunos resultados pueden variar si se repite el procedimiento.

```
datos3 <- datos2[sample(1:nrow(datos2), 500, replace =FALSE)
,1:3]
```

En este ejemplo cargamos una base de datos que proviene de un muestreo previamente realizado. La misma se denomina *datos3*.txt.

```
datos3 <-read.table   ("C: \\.... \\datos3.txt", header = TRUE)
```

Las funciones que permiten el ajuste de MLM a datos espaciales usando la librería geoR, requieren trabajar con objetos de clase `geodata`. Para cambiar el objeto a esta clase es necesario ubicar las columnas del archivo que contienen las coordenadas y las que contienen los datos. Puede visualizar los nombres de columnas en el orden que figuran en el objeto ejecutando el siguiente comando:

```
colnames(as.data.frame(datos3))
```

```
[1] "x"    "y"    "CE30"
```

En este caso las coordenadas se encuentran en la primera y segunda columna y la variable respuesta en la 3ra columna.

```
datos4 <- as.geodata(as.data.frame(datos3), coords.col = 1:2,
data.col = 3)
```

La función `likfit()` realiza estimaciones basadas en verosimilitud para datos normales tanto por máxima verosimilitud (ML) o por máxima verosimilitud restringida (REML). Por defecto ajusta un modelo para la función de correlación de tipo exponencial con efecto *nugget*. Otros tipos de modelos deben especificarse con el argumento `cov.model`. Con el argumento `ini` se especifican los valores iniciales para los parámetros de covarianza (*partial sill* y rango). El argumento `trend` permite sumar al modelo la estacionariedad espacial. La tendencia a través de las coordenadas puede especificarse mediante una fórmula o bien como un polinomio de primer grado (“1st”) o de segundo (“2nd”). A continuación se presentan los modelos ajustados:

Modelo exponencial con efecto *nugget* y tendencia de primer orden.

```
Exp_Nug <- likfit(datos4, ini=c(25, 80), lik.method = "REML",
trend= "1st")
```

Modelo esférico con efecto *nugget* y tendencia de primer orden.

```
Esf_Nug <- likfit(datos4, ini=c(25, 80), lik.method = "REML",
trend= "1st", cov.model= "sph")
```

Modelo exponencial sin efecto *nugget*, con tendencia de primer orden.

```
Exp <- likfit(datos4, ini=c(25, 80), lik.method = "REML",
trend="1st",fix.nugget = TRUE )
```

Modelo esférico sin efecto *nugget*, con tendencia de primer orden.

```
Esf <- likfit(datos4, ini=c(25, 80), lik.method = "REML",
trend="1st", cov.model= "sph",fix.nugget = TRUE )
```

Los modelos se compararon a través del criterio de información de Akaike (AIC).

```
Exp_Nug$AIC
Esf_Nug$AIC
Exp$AIC
Esf$AIC

[1] 2991.576
[1] 3006.427
[1] 2989.576
[1] 3066.637
```

También utilizando el criterio de información de Bayes (BIC).

```
Exp_Nug$BIC
Esf_Nug$BIC
Exp$BIC
Esf$BIC
```

```
[1] 3016.864
[1] 3031.715
[1] 3010.649
[1] 3087.71
```

Con los criterios de información deduce que el modelo exponencial sin efecto *nugget* es el mejor modelo. La comparación de este modelo con tendencia de primer orden vs. el modelo sin tendencia debe hacerse estimando por ML ya que la tendencia modela la estructura de medias. El procedimiento se muestra a continuación.

Modelo exponencial con tendencia.

```
Exp_t <- likfit(datos4, ini=c(25, 80), lik.method = "ML",
trend="1st",fix.nugget = TRUE )
```

Modelo exponencial sin tendencia.

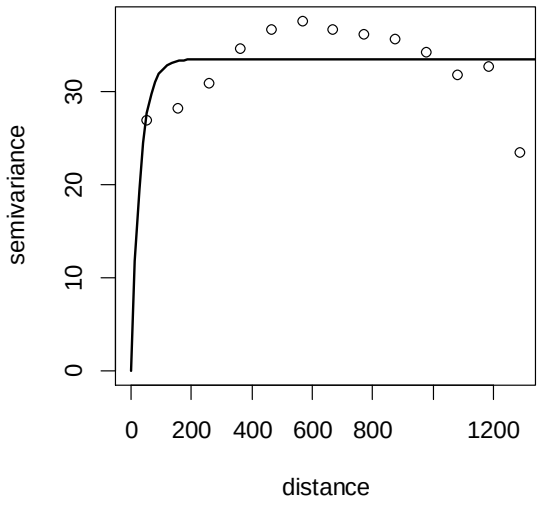
```
Exp_st <- likfit(datos4, ini=c(25, 80), lik.method = "ML",
fix.nugget = TRUE )

Exp_t$AIC
Exp_st$AIC

[1] 3009.19
[1] 3010.587

Exp_t$BIC
Exp_st$BIC

[1] 3030.263
[1] 3023.231
```

En este caso la tendencia adiciona dos parámetros al modelo, uno para cada coordenada. Por lo tanto, se deben utilizar dos grados de libertad para el test de cociente de verosimilitud.

```
ratio_t <- -2*(Exp_st$loglik-Exp_t$loglik)
LRT_t <- 1-pchisq(ratio, df=2); LRT

[1] 0.9999999
```

Los resultados expuestos sugieren seleccionar al modelo exponencial sin efecto *nugget* y sin tendencia como el mejor modelo. El semivariograma empírico y ajustado con este modelo se muestran a continuación:

```
modelo <- likfit(datos4, ini=c(25, 80), lik.method = "REML",
fix.nugget = TRUE )
plot(variog(datos4))
lines(modelo, lwd = 2)
```



Los parámetros estimados con MLM pueden consultarse con la función `summary()`.

```
summary(modelo)

Summary of the parameter estimation
-----------------------------------
Estimation method: restricted maximum likelihood

Parameters of the mean component (trend):
   beta
23.0927

Parameters of the spatial component:
   correlation function: exponential
      (estimated) variance parameter sigmasq (partial sill) =
33.43
      (estimated) cor. fct. parameter phi (range parameter)  =
30.58
   anisotropy parameters:
      (fixed) anisotropy angle = 0  ( 0 degrees )
      (fixed) anisotropy ratio = 1

Parameter of the error component:
      (fixed) nugget = 0

Transformation parameter:
      (fixed) Box-Cox parameter = 1 (no transformation)

Practical Range with cor=0.05 for asymptotic range: 91.61424

Maximised Likelihood:
   log.L n.params      AIC      BIC
 "-1499"      "3"   "3004"   "3017"

non spatial model:
   log.L n.params      AIC      BIC
 "-1589"      "2"   "3181"   "3190"

Call:
likfit(geodata = datos4, ini.cov.pars = c(25, 80), fix.nugget =
TRUE,
    lik.method = "REML")
```

MAPEO DE LA VARIABILIDAD ESPACIAL

Para el mapeo de la variabilidad espacial se confeccionarán una grilla de predicción donde se realiza el *kriging*. Las dimensiones de la grilla se establecerán mediante un polígono de los límites del lote. El archivo *bordes.txt* tiene los puntos georreferenciados de cada arista del polígono. La última fila se repite para cerrar el polígono.

```
Bordes <-read.table("C:/Bordes.txt", header = TRUE)
Bordes
```

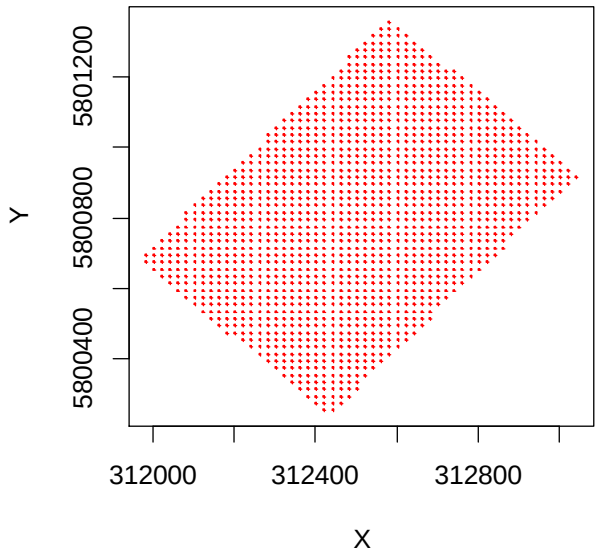
```
         x       y
1 311962.8 5800684
2 312432.8 5800234
3 313052.8 5800924
4 312582.8 5801364
5 311962.8 5800684
```

La primera función hace una grilla cuadrada especificando 20 metros como la distancia entre los puntos que la conforman. La segunda función recorta el polígono en la grilla.

```
gr <- pred_grid(Bordes, by=20)
gri <- polygrid(gr, bor=Bordes)
```

Para la grilla se usa la función `plot()`.

```
plot(gri,col = "red", pch = 10, cex = 0.2,xlab="X",ylab="Y")
```

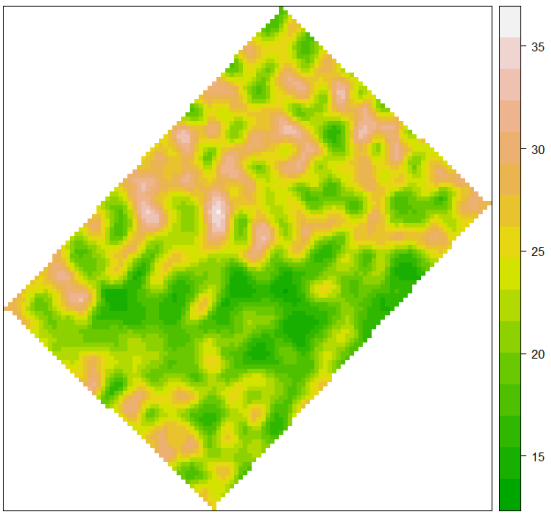


La función `krige()` realiza interpolación *kriging* univariada y simulaciones condicionales mediante diferentes métodos de predicción. En este caso, se presenta la interpolación por *ordinary kriging* (especificado con la fórmula `CE30~1`) en bloques de 40m × 40m con el modelo de semivariograma exponencial estimado con geoestadística clásica.

```
gridded(gri) = ~Var1+Var2
coordinates(datos2) <- ~x+y
Kg_wls <- krige(CE30~1, datos2, gri, model = Exp_wls, block =
c(40,40))
```

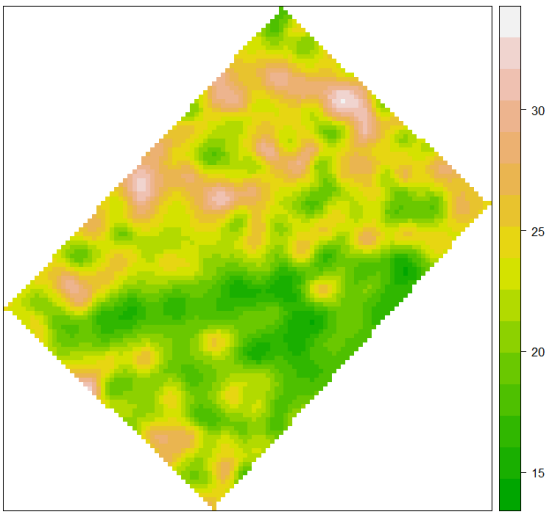
La función `spplot()`se utiliza para graficar las predicciones de la interpolación espacial.

```
spplot(Kg_wls["var1.pred"], col.regions=terrain.colors(100))
```



La interpolación también puede hacerse utilizando los parámetros de *partial sill*, rango y efecto *nugget* del semivariograma exponencial estimado con MLM.

```
coordinates(datos3) <- ~x+y
m <- vgm(33.43, "Exp", 30.58,0)
Kg_mlm <- krige(CE30~1, datos3, gri, model = m, block =
c(40,40))
spplot(Kg_mlm["var1.pred"], col.regions=terrain.colors(100))
```



Las predicciones con sus coordenadas se extraen del objeto creado por `krige()`. Las dos primeras columnas de este objeto contienen las coordenadas y la tercera columna, por lo general llamada `"var1.pred"`, las predicciones.

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

PARTE II.

APLICACIÓN EN AGRICULTURA POR AMBIENTES

DELIMITACIÓN DE ZONAS DE MANEJO

Las nuevas tecnologías en maquinarias agrícolas asociadas a la agricultura de precisión (AP) proporcionan la oportunidad de medir la variabilidad espacial en el rendimiento cosechado y en numerosas variables de sitio a escala fina o al interior del lote. Una forma difundida para el uso de esta información en AP es la clasificación de los sitios del lote en conjuntos de sitios homogéneos respecto a las variables medidas. Los resultados de la clasificación son luego usados para delimitar zonas o aéreas contiguas dentro del lote para el manejo sitio-especifico (Oliver, 2013). Las zonas de manejo (ZM) son usualmente áreas con características de suelo similares como textura, topografía, estado hídrico y niveles de nutrientes (Moral *et al*., 2010). Luego, diferentes fuentes de datos pueden ser usadas para delimitar ZM intralote. Las propiedades físicas y químicas del suelo son las más utilizadas, seguidos por atributos del paisaje y propiedades del cultivo como los índices de vegetación (Khosla *et al*., 2010). En ambientes agrícolas, donde la baja disponibilidad de agua y de nutrientes son los principales limitantes del cultivo, la producción depende en gran medida del tipo de suelo. Por ello, la definición de ZM intralote, basadas en propiedades de suelo, es ampliamente utilizada.

Determinada las fuentes de información a utilizar, el primer paso para la clasificación de sitios es la eliminación de datos raros o atípicos (*outliers* e *inliers*) que se puede realizar como se presentó en la sección de técnicas exploratorias para datos espaciales. Dado que será necesario combinar datos de diferentes fuentes de información y posiblemente diferentes resoluciones espaciales, se requerirá disponer los datos de las distintas variables de manera

tal de posibilitar el análisis multivariado espacial. A través de interpolación espacial se pueden realizar re-escalados de las mediciones originales que permite asociar los datos de las diferentes variables relevadas a cada sitio de una grilla común a todas las características medidas.

Sobre la base de datos depurada y re-escalada se aplican diferentes algoritmos del análisis de *cluster* (Stafford *et al*., 1998). Un algoritmo común en AP es el método no jerárquico *fuzzy k-means* (Bezdek, 1981). Sin embargo, en varias aplicaciones de *fuzzy k-means*, se observó que el algoritmo puede presentar el inconveniente de una alta fragmentación de las zonas porque, como la mayoría de los algoritmos de clasificación, ignora la naturaleza espacial de los datos georreferenciados (Ping y Dobermann, 2003; Frogbrook y Oliver, 2007). Para mitigar este problema, nosotros recomendamos realizar un análisis de *cluster* del tipo *fuzzy k-means* pero sobre las componentes principales obtenidas desde un análisis de componentes principales espacial (Córdoba *et al*., 2013). El método propuesto al incorporar la información espacial facilita la obtención de clases más contiguas reduciendo la fragmentación de las ZM delimitadas.

Los coeficientes de partición y de entropía de la clasificación (conocidos también como *fuzziness performance index*-FPI y *normalized classification entropy*-NCE) (Bezdek, 1981) son frecuentemente utilizados 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 calculados con tal fin. En esta ilustración se combinarán los índices obteniendo una métrica que resume la información brindada por cada índice (Galarza *et al*., 2013).

Para formar clases más contiguas y reducir la fragmentación también se recomienda aplicar filtros espaciales sobre la clasificación resultante (Ping y Dobermann 2003; Lark, 1998; Galarza *et al*., 2013). Sobre los datos resultantes de la clasificación, aplicamos filtros espaciales no lineales (Arce, 2005) basados en el ordenamiento de los pixeles contenidos en una porción de la imagen (máscara). Las máscaras utilizadas pueden tener diferentes tamaños 3×3, 5×5, 7×7 o *n*×*n* pixeles. Posteriormente se sustituye el valor del pixel central con el valor que resulta del ordenamiento. Un filtro ampliamente difundido es el filtro de la 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).

Es recomendable disponer de un muestreo de suelo y/o rendimientos para validar las ZM delimitadas. El mismo se puede llevar a cabo usando un muestreo aleatorio estratificado siendo las ZM usadas como estratos. Usualmente, se imponen restricciones sobre la asignación aleatoria de los puntos de muestreo para evitar muestrear los límites de las zonas. Se recomienda tomar no menos de tres muestras por cada ZM. 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.

Dado que los datos generados por las tecnologías de AP están espacialmente correlacionados no es posible el uso de modelos de ANAVA para datos independientes para evaluar las diferencias de medias entre zonas. Un enfoque alternativo para la validación de las ZM es utilizar Modelos Lineales Mixtos (MLM) (West *et al*., 2007) que permiten modelar la estructura de correlación espacial subyacente produciendo medias y errores estándares

ajustados. A continuación se presenta un tutorial "paso a paso" para implementar el protocolo de delimitación de ZM con funciones de R en el intérprete de InfoStat (Di Rienzo *et al*., 2014). Seguidamente, se presenta otra herramienta para implementar el análisis. Se trata del menú "Estadística Espacial". Con este menú puede implementarse el análisis directamente desde InfoStat sin necesidad de recurrir a un ambiente de programación.

ILUSTRACIÓN DEL ANÁLISIS USANDO EL INTÉRPRETE DE R EN INFOSTAT

Para delimitar zonas de manejo es necesario que las variables a utilizar hayan sido interpoladas con la misma grilla de predicción. Es decir, que cada punto de la grilla tendrá un dato para cada variable predicha. En el siguiente ejemplo se realizó el procedimiento de interpolación con mediciones de conductividad eléctrica aparente en dos profundidades 0-30 cm (CE30) y 0-90 cm (CE90), elevación, profundidad de suelo (Pe) y rendimiento de trigo (Tg) (archivo *Pred.txt*, ver Anexo I). Para cada variable se realizó un análisis exploratorio y la predicción espacial para el re-escalado usando una grilla común a todas ellas de 10 × 10 m. Para la validación de las zonas de manejo se utilizó la base *Muestreo.txt* la cual contiene el registro de 8 puntos de muestreo de suelos dentro de cada zona de manejo delimitada. Las variables de suelo relevadas para la validación fueron MO, contenido de nitratos y %Arcilla.

Una vez que se realizael re-escalado de cada variable, se tiene un objeto para cada variable con igual número de filas y columnas que pueden unirse en un único objeto usando la función `cbind()`. Para `PredCE30` se extraen las 3 primeras columnas correspondiente a las coordenadas y valores predichos,

mientras que para las restantes sólo se extraen los valores predichos de cada variable (columna 3) considerando que, si se utilizó la misma grilla de predicción, las coordenadas de cada `dataframe` deberían ser las mismas. Se recomienda mantener clara la nomenclatura de cada variable, teniendo en cuenta que el software es *case-sensitive* (sensible a mayúsculas y minúsculas). A tal efecto, se renombraron las columnas. A continuación se muestras los códigos de R para hacer el procedimiento de concatenación, pero para la ejemplificación se carga y utiliza una base de datos que previamente fue concatenada.

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

names(Pred)[3]<-paste("CE30")
names(Pred)[4]<-paste("CE90")
names(Pred)[5]<-paste("Elev")
names(Pred)[6]<-paste("Pe")
names(Pred)[7]<-paste("Tg")
```

Carga de archivo con variables re-escaladas y concatenadas.

```
Pred <-read.table ("C: \\Users \\mbalzarini \\Desktop \\Libro \\
Pred.txt", header = TRUE)
Pred
```

```
     x       y        CE30     CE90     Elev       Pe       Tg
1    312432.8 5800234 25.80810 28.50741 160.4142  -78.07533 3.734030
2    312422.8 5800244 26.19117 28.14623 160.4164  -77.21150 3.731334
3    312432.8 5800244 25.26266 27.95664 160.4286  -78.65952 3.725866
4    312412.8 5800254 26.28254 27.56873 160.4184  -75.93873 3.727748
5    312422.8 5800254 25.68631 27.48603 160.4275  -76.89577 3.715309
```

Para delimitar las zonas de manejo se realizó primero un Análisis de Componentes Principales espacial (MULTISPATI-PCA) utilizando las librerías “spdep” (Bivand, 2014) y “ade4” (Chessel *et al*., 2004). Luego, las componentes principales espaciales resultantes del MULTISPATI-PCA fueron utilizadas como *input* del análisis de *cluster fuzzy k-means*. El exponente difuso

se fijó en valor convencional de 1.30 (Odeh *et al*., 1992) y la distancia de similitud incluida en la función de optimización fue la Euclidea. Para realizar el análisis de *cluster fuzzy k-means* se utilizó la librería "e1071" (Meyer *et al*., 2014) 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 punto para el cálculo del índice de Moran. Luego, se realizó un Análisis de Componentes Principales (PCA) clásico y posteriormente sobre las componentes generadas por PCA, se aplicó MULTISPATI-PCA.

La función `dudi.pca()` permite realizar un PCA sobre objetos de clase `dataframe`. Sus argumentos indican, las variables con las que se realizará el PCA, un valor lógico (`TRUE` o `FALSE`) indicando si debe o no centrarse por la media, un valor lógico para la realización o no del gráfico y la cantidad de ejes guardados, que coincide con la cantidad de variables utilizadas en el análisis.

```
pca <- dudi.pca(Pred[,3:7], center=T,scannf = FALSE, nf = 5)
```

Para transformar un PCA en un PCA espacial (MULTISPATI-PCA) se debe calcular, como ya se hizo anteriormente con los índices de Moran, la red de vecindarios y la matriz de ponderación espacial.

```
cord_2 <- coordinates(Pred[,1:2])
gri_2 <- dnearneigh(cord_2,0,25)
lw_2 <- nb2listw(gri_2, style = "W")
```

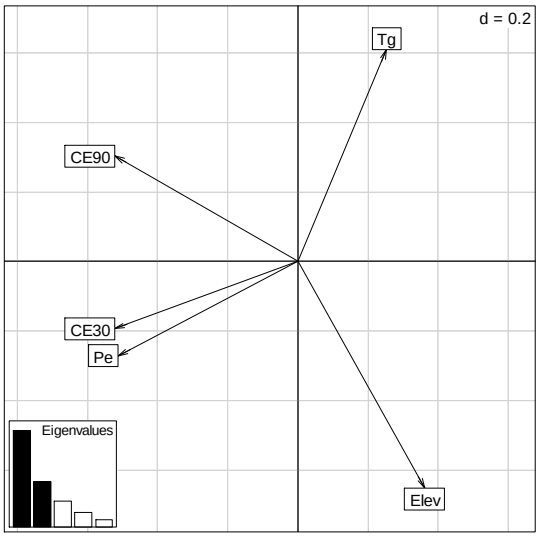
La función `multispati()` permite realizar el MUlTISPATI-PCA.

```
ms2 <- multispati(pca, lw_2, scannf = F, nfposi = 5)
```

Para realizar un gráfico que muestre las correlaciones entre las variables se puede usar la función `s.arrow()`. En este gráfico de traza un vector para

cada variable en el espacio definido por las componentes principales que se seleccionen. En este caso de estudio, la función utiliza la primera componente para graficar el eje horizontal y la segunda componente para el eje vertical. Para adicionar un gráfico de barras con los autovalores puede usarse el argumento `add.scatter.eig()`.

```
s.arrow(ms2$c1,xax = 1, yax = 2, clabel = 1)
add.scatter.eig(ms2$eig, xax = 1, yax = 2, posi = "bottomleft",
ratio = 0.2)
```



El gráfico obtenido del MULTISPATI-PCA permite además estudiar la estructura de correlación entre las variables utilizadas para la delimitación de zonas. 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 Tg y Elev 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.

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*. La función `multispati()` almacena las sPC en la posición `$li` dentro de los objetos creados. La siguiente sentencia crea un nuevo objeto con la unión de las columnas con las coordenadas, las predicciones para cada variable y las sPC.

```
PredAM <- cbind(Pred,ms2$li)
```

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 que corresponden a la sPC1, sPC2 y sPC3, de esta forma una gran cantidad de la variabilidad total es contemplada (≥70%) en el análisis. 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.

```
MC_2<-cmeans(PredAM[,8:10],2,100,method="cmeans",m=1.3)
MC_3<-cmeans(PredAM[,8:10],3,100,method="cmeans",m=1.3)
MC_4<-cmeans(PredAM[,8:10],4,100,method="cmeans",m=1.3)
```

Delimitadas las clases de manejo, se necesita determinar cuál es el número óptimo de clases. En este ejemplo se debe seleccionar entre dos, tres y cuatro clases conformadas. Para ello se utilizaron los siguientes índices: Xie-Beni, coeficiente de partición, entropía de clasificación y Fukuyama-Sugeno. Estos índices serán calculados para 2, 3 y 4 clases de manejo respectivamente, utilizando la función `fclustIndex()`. 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 utiliza el valor inverso del índice. Luego se confección una tabla con los índices obtenidos.

```
I2CM <- fclustIndex(MC_2,PredAM[,8:10], index=c("xie.beni",
"fukuyama.sugeno","partition.coefficient","partition.entropy" ))

I3CM <- fclustIndex(MC_3,PredAM[,8:10], index=c("xie.beni",
"fukuyama.sugeno","partition.coefficient","partition.entropy" ))

I4CM <- fclustIndex(MC_4,PredAM[,8:10], index=c("xie.beni",
"fukuyama.sugeno","partition.coefficient","partition.entropy" ))

Indices0 <- cbind(I2CM,I3CM,I4CM)
rownames(Indices0)<-c("XieBeni", "FukSug", "CoefPart",
"EntrPart")
Indices0[3,]<-1/Indices0[3,]
Indices0
```

```
                  I2CM          I3CM          I4CM
XieBeni    4.832910e-05  9.371783e-05  1.088984e-04
FukSug    -1.201711e+04 -1.317559e+04 -1.440002e+04
CoefPart   1.080456e+00  1.158522e+00  1.198528e+00
EntrPart   1.257312e-01  2.428802e-01  3.018947e-01
```

En este ejemplo la mayoría de los índices, excepto Fukuyama-Sugeno, 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.

```
IndicesN<-as.data.frame (rbind (Indices0[1,]/max(Indices0[1,])
,Indices0[2,]/max(Indices0[2,]) ,Indices0[3,]/max(Indices0[3,])
,Indices0[4,]/max(Indices0[4,])))
```

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

Indice2CM; Indice3CM; Indice4CM

[1] 1.477527
[1] 1.877321
[1] 2.10616
```

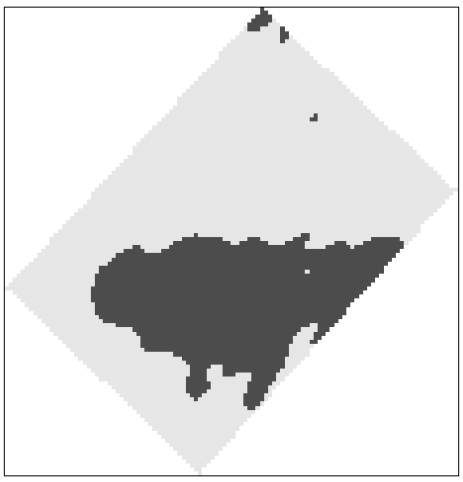
Los valores de los índices resumen sugieren que debiera seleccionarse dos clases de manejo. Así mismo, la clasificación con dos clases de manejo presenta grandes zonas con límites más coherentes respecto a la clasificación con 3 y 4 clases que presentan varias zonas pequeñas y de forma irregular.

Para realizar mapas de las clases de manejo, primero se debe extraer los datos de las clases delimitadas con el algoritmo *fuzzy k-means* y luego se grafican con la función `spplot()`. La información sale de objetos ya creados anteriormente, que se unen en un nuevo objeto llamado `base0` junto con las coordenadas de los puntos.

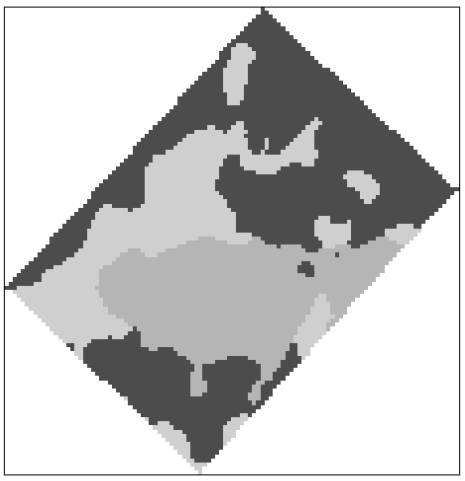
```
MC_22 <-as.data.frame(MC_2$cluster)
MC_33 <-as.data.frame(MC_3$cluster)
MC_44 <-as.data.frame(MC_4$cluster)
base0 <- cbind(PredAM[,1:2],MC_22,MC_33,MC_44)
```

La función `spplot()` trabaja sobre objetos de clase espacial (clase `SpatialPixelsDataFrame`) donde están previamente definidas las coordenadas y un valor lógico que indica si éstas provienen de una grilla ordenada. Las columnas de este tipo de objetos están en forma de lista, en consecuencia se debe acceder a cada columna utilizando [] como en los ejemplos siguientes.

```
coordinates(base0) <- ~x+y
gridded(base0) <- TRUE
spplot(base0["MC_2$cluster"],col.regions=gray.colors(2),
colorkey= F)
```



```
spplot(base0["MC_3$cluster"],col.regions=gray.colors(10),
colorkey=F)
```



```
spplot(base0["MC_4$cluster"],col.regions=gray.colors(4),
colorkey= F)
```

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 ZM. Para ello, se aplica el filtro espacial de la mediana aplicando la función `smooth`. Este funciona remplaza el valor del píxel central por la mediana de los valores del vecindario de ese píxel. Las máscaras que definen el tamaño de los vecindarios (números de pixel) pueden tener diferentes dimensiones. En este ejemplo se probaron máscaras (ventanas) de 3×3, 5×5 y 7×7 pixeles. Previo a la aplicación del filtro es necesario convertir la base de datos en un archivo del tipo matriz, utilizando para ello la función `obtainM`.

Función para obtener la matriz.

```
obtainM <- function(mytable){
  x<-as.numeric(names(table(mytable$x)))
  y<-as.numeric(names(table(mytable$y)))
  myframe    <-    matrix(1:(length(x)*length(y)),    length(x),
length(y))
  position<-function(pos) {
    col=as.integer((pos-1)/nrow(myframe))+1
```

```
    row=pos-((nrow(myframe)*col)-nrow(myframe))
    myindex=which(mytable$x==x[row]                                    &
mytable$y==y[col],arr.ind=T)
    if(length(myindex)==0) return(NA) else mytable[myindex,3]
  }
 thematrix<-as.matrix(apply(myframe,c(1,2),position))
 rownames(thematrix)<-x
 colnames(thematrix)<-y
 thematrix}
```

Función filtro de la mediana.

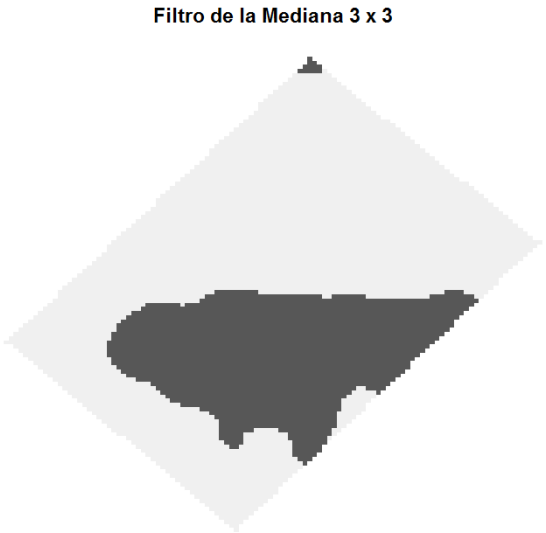
```
smooth <-function(mytable,mywindow){
  newtable<-
matrix(1:(dim(mytable)[1]*dim(mytable)[2]),dim(mytable)[1],dim(m
ytable)[2])
      vecinity<-function(pos) {
        col=as.integer((pos-1)/nrow(newtable))+1
        row=pos-((nrow(newtable)*col)-nrow(newtable))
        if (is.na(mytable[row,col])) NA else{
        myrow1<-ifelse(row-mywindow<1,1,row-mywindow)
        mycol1<-ifelse(col-mywindow<1,1,col-mywindow)
        myrow2<-
ifelse(row+mywindow>dim(newtable)[1],row,row+mywindow)
        mycol2<-
ifelse(col+mywindow>dim(newtable)[2],col,col+mywindow)

        neighbor<-
na.omit(as.vector(mytable[myrow1:myrow2,mycol1:mycol2]))
        round(median(neighbor),digits=0)
      }}
      as.matrix(apply(newtable,c(1,2),vecinity))}
```

Aplicación de la función `obtainM`.

```
base0 <- cbind(PredAM[,1:2],MC_22)
datafilter <- obtainM(base0)
```

Aplicación del filtro de la mediana dimensión 3×3.

```
smoot3x3 <- smooth(datafilter,3)
image(smoot5x5, main= "Filtro de la Mediana 3 x 3",axes = FALSE,
xlab="",ylab="",col=palette(c("grey94","grey34")))
```



Aplicación del filtro de la mediana dimensión 5 x 5.

```
smoot5x5 <- smooth(datafilter,5)
image(smoot7x7, main= "Filtro de la Mediana 5 x 5",axes = FALSE,
xlab="",ylab="",col=palette(c("grey94","grey34")))
```

Aplicación del filtro de la mediana dimensión 7 x 7.

```
smoot7x9 <- smooth(datafilter,7)
image(smoot9x9, main= "Filtro de la Mediana 7 x 9",axes = FALSE,
xlab="",ylab="",col=palette(c("grey94","grey34")))
```

En este ejemplo, el filtro de 7 × 7 píxeles resultó ser más adecuado que los filtros 3×3 y 5×5 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 la cantidad de ZM potenciales es extraída. Para obtener la base de datos final, se utiliza la función `MtoT` para transformar la matriz "suavizada" por el filtro de la mediana en una tabla de datos. Luego, se une horizontalmente la base con las zonas de manejo y la base con las variables predichas.

```
MtoT <- function(mymatrix){
position <- function(ij){
data.frame(x=rownames(mymatrix)[ij[1]],y=colnames(mymatrix)[ij[2
]],z=mymatrix[ij[1],ij[2]])
}
myindex                                                    <-
which(!is.na(mymatrix),arr.ind=T);rownames(myindex)=NULL
b <- apply(myindex,1,position)
b <- do.call("rbind",b);b}

base1 <- as.data.frame(smoot9x9)
base2 <- MtoT(base1)
base2[order(base2[,1], base2[,2]),]
PredMA[order(PredMA[,1], PredMA[,2]),]
Finalbase <- cbind(PredMA[,1:6],base2[,3])
names(Finalbase)[7]<-paste("Zona")
```

	x	y	CE30	CE90	Elev	Pe	Tg	Zona
1	312432.8	5800234	25.8081	28.5074	160.414	-78.0753	3.73403	2
2	312422.8	5800244	26.1912	28.1462	160.416	-77.2115	3.73133	2
3	312432.8	5800244	25.2627	27.9566	160.427	-78.6595	3.72587	2
4	312412.8	5800254	26.2825	27.5687	160.418	-75.9387	3.72775	2
5	312422.8	5800254	25.6863	27.4860	160.428	-76.8958	3.71531	2
6	312432.8	5800254	24.8235	27.2678	160.438	-78.0412	3.70057	2

VALIDACIÓN DE ZONAS DE MANEJO

Se utilizaron datos de MO para la validación de la zonificación mediante el contraste de medias de MO para cada zona. Para ello se utilizó un MLM con efecto fijo de zona y errores correlacionados espacialmente. La comparación debe realizarse con todas las propiedades de suelo medidas en el muestreo de validación y las funciones de correlación espacial podrían 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” (Pinheiro *et al*., 2014). A continuación se procede a cargar el archivo con los datos del muestreo, posteriormente se ajustan modelos y finalmente se seleccionó el mejor modelo. Los modelos para MO se ajustan con efecto fijo de zona y errores correlacionados espacialmente.

Carga de datos.

```
Muestreo <-read.table("C:/muestreo.txt", header = TRUE; Muestreo
```

```
          X       Y Zona      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
```

Aplicando la función `gls()` se ajustará un modelo de correlación espacial exponencial. Los argumentos declarados en esta función consisten en una fórmula para el modelo, la declaración de una estructura de correlación, que hacer en caso de valores faltantes y sobre qué datos trabajar.

Modelo de correlación espacial exponencial

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

Modelo de correlación espacial exponencial con efecto *nugget*.

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

Modelo de correlación espacial gaussiana.

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

Modelo de correlación espacial gaussiana con efecto *nugget*.

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

Modelo de correlación espacial esférica.

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

Modelo de correlación espacial esférica con efecto *nugget*.

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

Modelo de errores independientes.

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

Ajustado los diferentes modelos se procede a realizar la selección del modelo de mejor ajuste usando el Criterio de Información de Akaike (AIC).

```
AICmod1 <- AIC(modelo.001_mo_REML)
AICmod2 <- AIC(modelo.002_mo_REML)
AICmod3 <- AIC(modelo.003_mo_REML)
AICmod4 <- AIC(modelo.004_mo_REML)
AICmod5 <- AIC(modelo.005_mo_REML)
AICmod6 <- AIC(modelo.006_mo_REML)
AICmod7 <- AIC(modelo.007_mo_REML)
```

```
AICmod1
AICmod2
AICmod3
AICmod4
AICmod5
AICmod6
AICmod7

[1] -1.515517
[1] 0.4844829
[1] -4.058434
[1] -2.060023
[1] -1.220597
[1] -0.4010228
[1] 2.925729
```

En la selección del modelo de correlación espacial, el AIC indica que el modelo con función de correlación espacial gaussiana es el mejor para estos datos, tanto entre los modelos sin *nugget* como para los modelos con efecto *nugget*. Así mismo, este modelo también fue mejor que el de errores independientes.

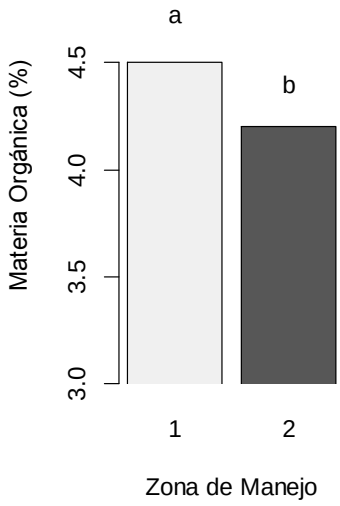
Los resultados del modelo seleccionado muestran que existen diferencias estadísticamente significativas entre las zonas delimitadas en cuanto al contenido de MO. La zona 1 (gris claro) presenta un contenido promedio de 4,50% mientras que para la zona 2 (gris oscuro) el valor promedio de MO es de 4,20%.

```
summary(modelo.003)

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)  4.499783 0.09048247 49.73099  0.0000
Zona2       -0.297806 0.10406395 -2.86176  0.0126

Correlation:
      (Intr)
Zona2 -0.424

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.8607720 -0.8864720 -0.3808446  0.2064385  1.7265838
Residual standard error: 0.2033711
Degrees of freedom: 16 total; 14 residual
```

```
MOMedia <- summary(lsmeans(modelo.003_mo_REML,"Zona")); MOMedia

Zona   lsmean         SE df lower.CL upper.CL
 1    4.499783 0.09048247 14 4.305717 4.693848
 2    4.201977 0.10502233 14 3.976727 4.427228

Confidence level used: 0.95
```

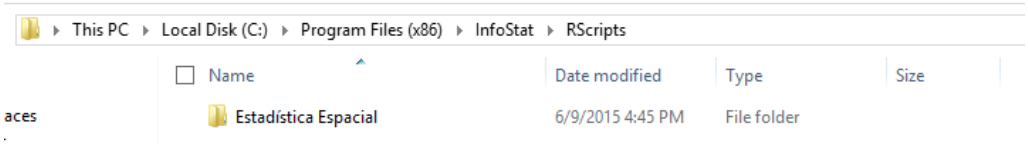
```
attach(MOMedia)
MOMedia <-by(lsmean,Zona,mean)
so  <-  barplot(MOMedia,xlab="Zona  de  Manejo",  ylab="Materia
Orgánica (%)",col=c("grey94","grey34")
,ylim=c(3,max(MOMedia+MOMedia*0.1)),xpd=F)
letters = c("a","b")
text(x=so,y=MOMedia+MOMedia*0.05,label=letters,cex = 1)
```


a
b
Materia Orgánica (%)
4.5
4.0
3.5
3.0
1
2
Zona de Manejo

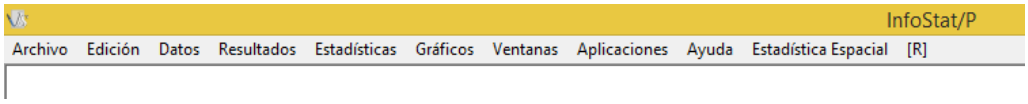
IMPLEMENTACIÓN DEL PROTOCOLO A TRAVÉS DEL MENÚ "ESTADÍSTICA ESPACIAL" EN INFOSTAT

INSTALACIÓN DEL MENÚ ESTADÍSTICA ESPACIAL EN INFOSTAT

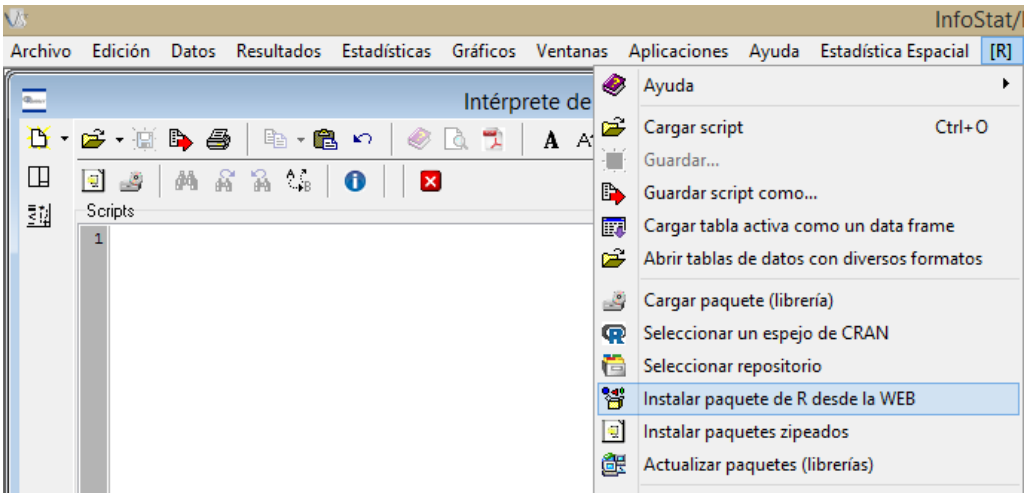
Para instalar el menú "Estadística Espacial" es necesario que el software R haya sido instalado previamente y vinculado a InfoStat (ver "¿Cómo instalar R?" en el menú **Ayuda** de InfoStat). Posteriormente, se debe copiar la carpeta "Estadística Espacial" dentro de la carpeta "Rscripts" ubicada en el directorio donde InfoStat fue instalado.



Al reiniciar InfoStat, el menú se agrega automáticamente en la barra de herramientas como muestra la siguiente figura:

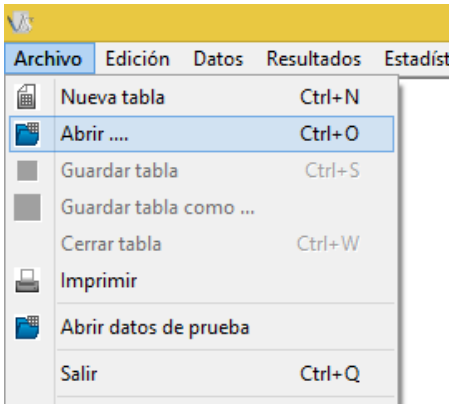


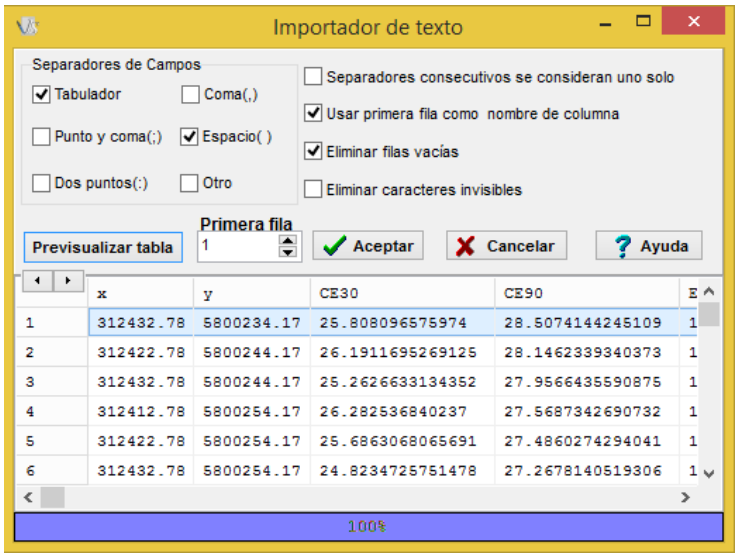
Las rutinas realizadas desde el menú "Estadística Espacial" están programadas en R y, por lo tanto, requieren determinados paquetes para ejecutarse. InfoStat solicitará estos paquetes cuando los necesite e intentará descargarlos desde la web cuando se ejecute cada rutina. En caso de que InfoStat no logre tener acceso a internet, los paquetes pueden ser manualmente instalados desde el menú **[R]**, tanto desde la web como desde un archivo *.zip*. La siguiente figura ilustra parte del menú **[R]**.



DELIMITACIÓN DE ZONAS DE MANEJO CON INFOSTAT

Paso 1. *Abrir la base de datos.* Los datos están contenidos en el archivo *Pred.txt*, para abrirlo utilice el comando abrir dentro del menú archivo. Como el archivo está en un formato de texto, se desplegará la ventana "Importador de texto", donde puede especificarse la separación de campos y otros atributos del archivo que posibilitan su correcta lectura.





La base de datos desplegada contiene las coordenadas y las variables en estudio, descriptas anteriormente. La siguiente figura muestra los primeros 19 registros.

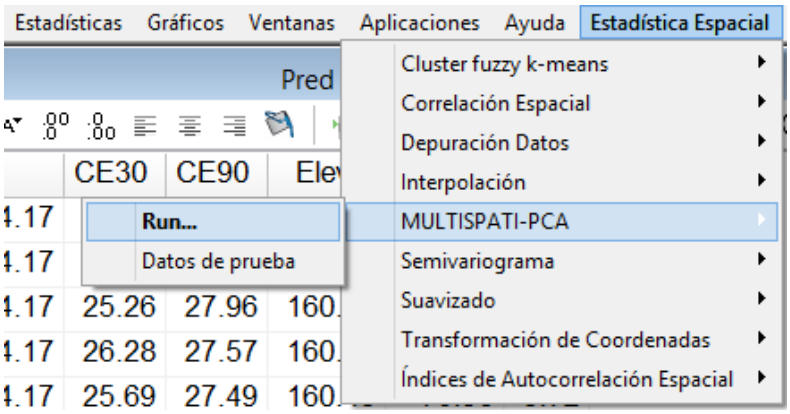
Pred

Caso	x	y	CE30	CE90	Elev	Pe	Tg
1	312432.78	5800234.17	25.81	28.51	160.41	-78.08	3.73
2	312422.78	5800244.17	26.19	28.15	160.42	-77.21	3.73
3	312432.78	5800244.17	25.26	27.96	160.43	-78.66	3.73
4	312412.78	5800254.17	26.28	27.57	160.42	-75.94	3.73
5	312422.78	5800254.17	25.69	27.49	160.43	-76.90	3.72
6	312432.78	5800254.17	24.82	27.27	160.44	-78.04	3.70
7	312442.78	5800254.17	23.89	26.90	160.45	-78.71	3.69
8	312402.78	5800264.17	25.75	26.77	160.41	-75.62	3.73
9	312412.78	5800264.17	25.62	26.90	160.42	-75.50	3.71
10	312422.78	5800264.17	25.23	26.75	160.43	-76.32	3.69
11	312432.78	5800264.17	24.53	26.57	160.44	-77.17	3.67
12	312442.78	5800264.17	23.82	26.28	160.45	-77.80	3.66
13	312452.78	5800264.17	23.17	25.91	160.46	-78.16	3.66
14	312392.78	5800274.17	24.61	26.27	160.39	-77.48	3.73
15	312402.78	5800274.17	24.93	26.40	160.41	-76.33	3.72
16	312412.78	5800274.17	24.90	26.37	160.43	-75.74	3.69
17	312422.78	5800274.17	24.64	26.20	160.44	-76.02	3.67
18	312432.78	5800274.17	24.17	26.01	160.45	-76.52	3.65
19	312442.78	5800274.17	23.68	25.76	160.46	-77.20	3.64
20	312452.78	5800274.17	23.19	25.47	160.47	-78.04	3.64

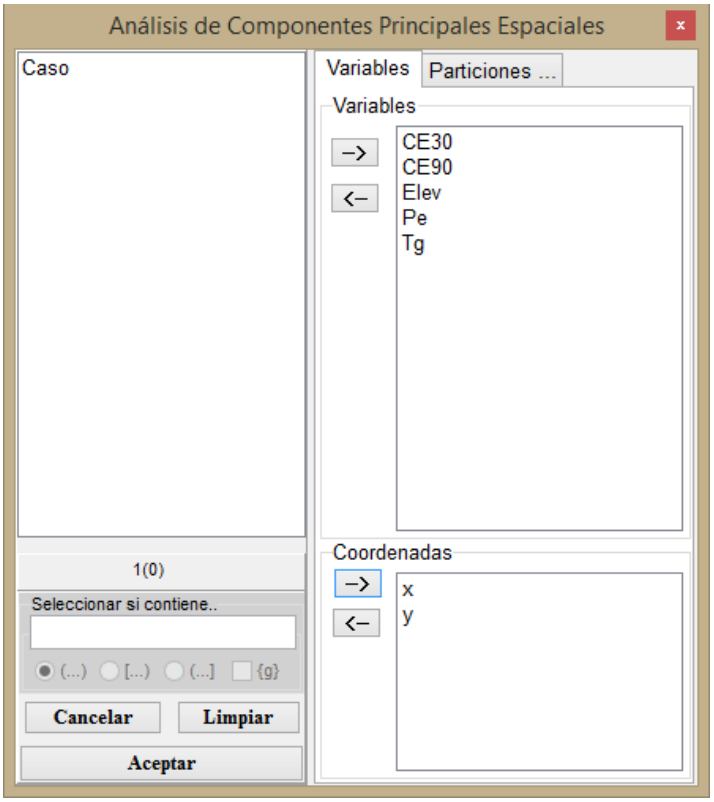
Real | Registros: 5982*7 | n=1 Suma = 312433 Media = 312433 D.E. = 0 Min = 312432.78 Max = 312432.78 P05 = 312432.7

Para obtener esta tabla se realizó previamente la interpolación a una misma grilla de predicción. Este proceso ya fue descripto en secciones anteriores y puede realizarse desde el submenú "Estadística Espacial>Interpolación".

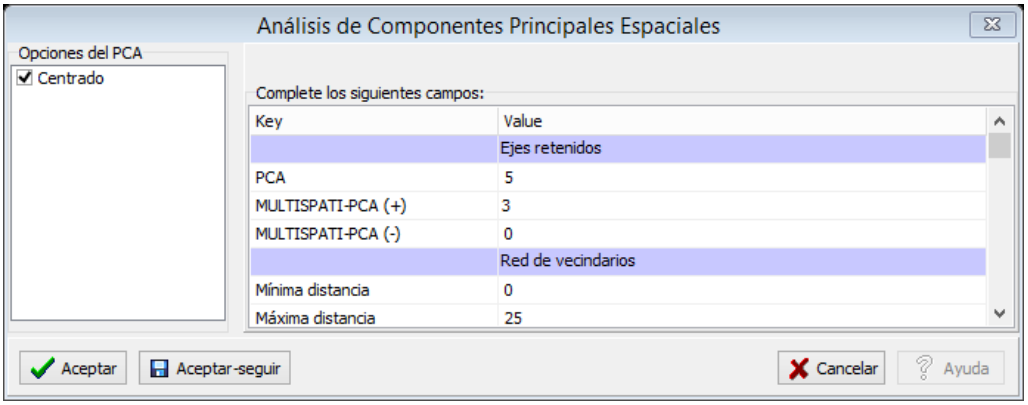
Paso 2. *Componentes principales espaciales*. El Análisis de Componentes Principales espacial puede realizarse desde el sub-menú "MULTISPATI-PCA" en InfoStat.



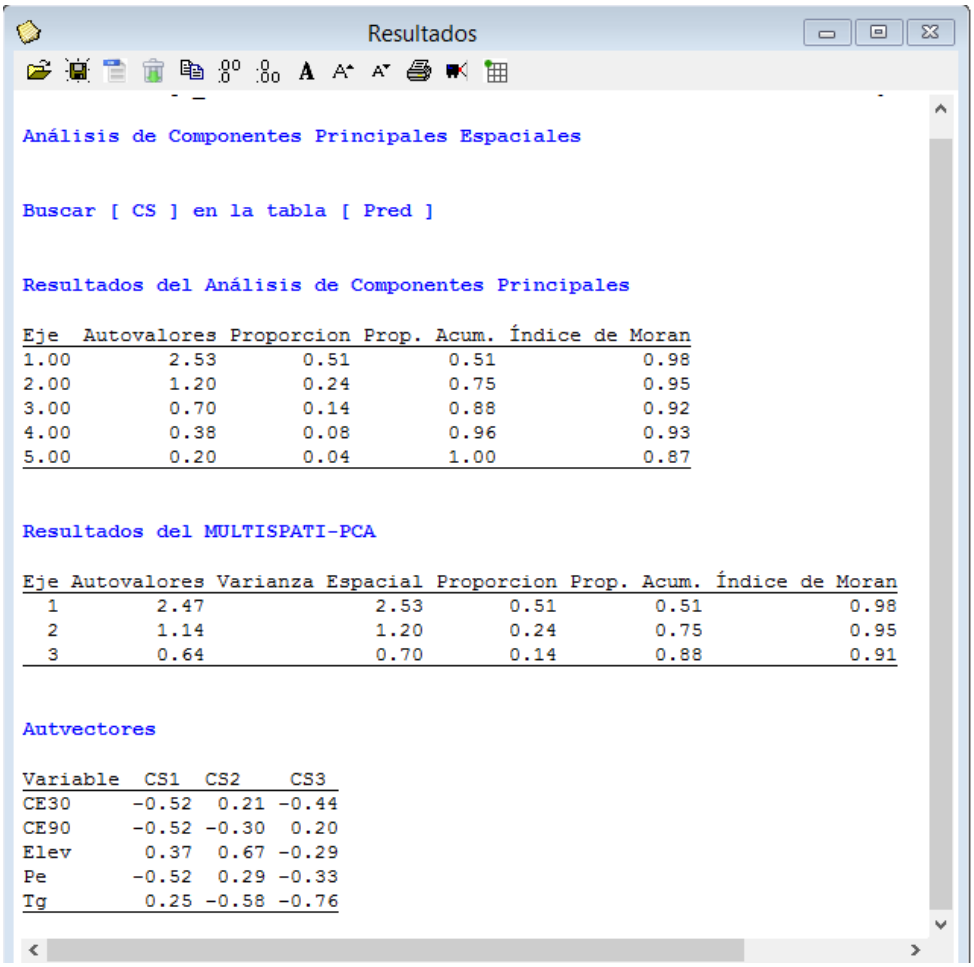
En la ventana desplegada, se indica al software las variables con las que se realizará el PCA y las coordenadas.



La siguiente ventana, permite modificar la opción de centrado, especificar la cantidad de ejes calculados y retenidos y las distancias de la red de vecindarios.



Luego del proceso de cálculo, se desplegarán ventanas gráficas de R y la ventana “Resultados” de InfoStat. En la tabla de datos se adicionan nuevas columnas con los ejes retenidos.

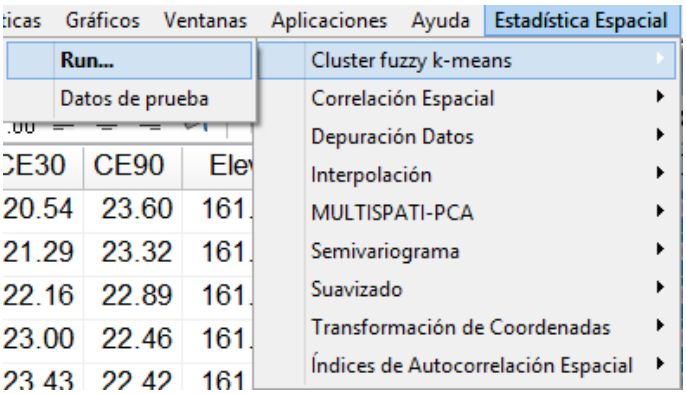

Resultados
Análisis de Componentes Principales Espaciales
Buscar [CS] en la tabla [Pred]
Resultados del Análisis de Componentes Principales
Eje Autovalores Proporcion Prop. Acum. Índice de Moran
1.00 2.53 0.51 0.51 0.98
2.00 1.20 0.24 0.75 0.95
3.00 0.70 0.14 0.88 0.92
4.00 0.38 0.08 0.96 0.93
5.00 0.20 0.04 1.00 0.87
Resultados del MULTISPATI-PCA
Eje Autovalores Varianza Espacial Proporcion Prop. Acum. Índice de Moran
1 2.47 2.53 0.51 0.51 0.98
2 1.14 1.20 0.24 0.75 0.95
3 0.64 0.70 0.14 0.88 0.91
Autvectores
Variable CS1 CS2 CS3
CE30 -0.52 0.21 -0.44
CE90 -0.52 -0.30 0.20
Elev 0.37 0.67 -0.29
Pe -0.52 0.29 -0.33
Tg 0.25 -0.58 -0.76

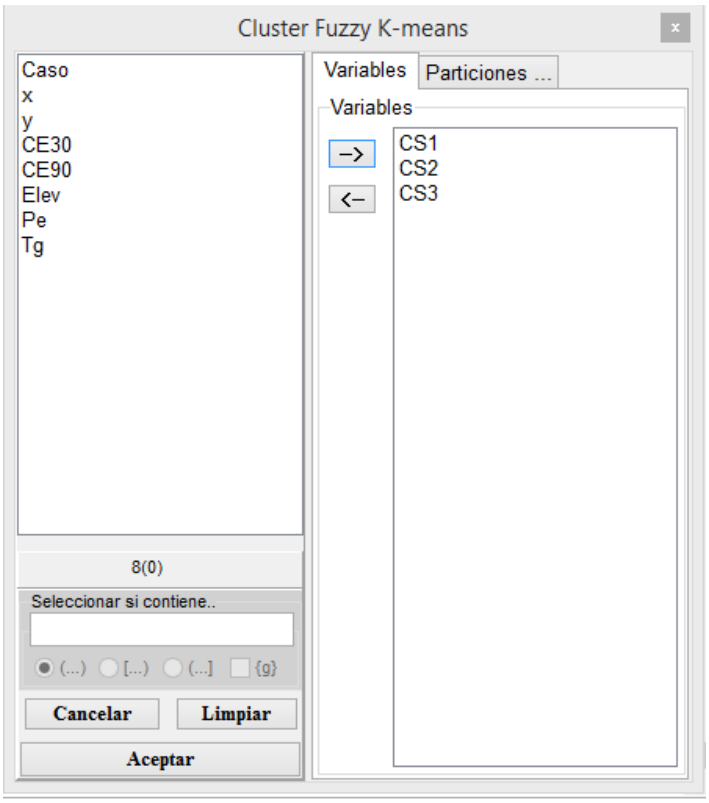
Pred

Caso	x	y	CE30	CE90	Elev	Pe	Tg	CS1	CS2	CS3
5964	312572.78	5801324.17	20.54	23.60	161.17	-75.94	3.80	0.67	0.60	-0.49
5965	312582.78	5801324.17	21.29	23.32	161.23	-76.37	3.80	0.65	0.78	-0.63
5966	312592.78	5801324.17	22.16	22.89	161.30	-77.82	3.78	0.68	0.99	-0.77
5967	312602.78	5801324.17	23.00	22.46	161.33	-78.55	3.77	0.67	1.17	-0.89
5968	312612.78	5801324.17	23.43	22.42	161.38	-80.02	3.76	0.66	1.27	-0.92
5969	312622.78	5801324.17	23.32	22.50	161.42	-81.38	3.74	0.70	1.32	-0.85
5970	312562.78	5801334.17	20.02	23.72	161.18	-74.25	3.77	0.66	0.67	-0.33
5971	312572.78	5801334.17	20.17	23.73	161.16	-74.69	3.77	0.62	0.65	-0.32
5972	312582.78	5801334.17	20.69	23.45	161.24	-75.16	3.76	0.66	0.83	-0.43
5973	312592.78	5801334.17	21.52	22.98	161.30	-76.12	3.75	0.69	1.04	-0.59
5974	312602.78	5801334.17	22.54	22.49	161.34	-77.73	3.75	0.69	1.21	-0.76
5975	312612.78	5801334.17	23.41	22.22	161.38	-79.29	3.74	0.67	1.36	-0.89
5976	312572.78	5801344.17	19.91	23.87	161.24	-74.08	3.74	0.65	0.79	-0.25
5977	312582.78	5801344.17	20.05	23.61	161.27	-74.54	3.74	0.72	0.86	-0.29
5978	312592.78	5801344.17	20.69	23.14	161.30	-75.24	3.74	0.76	1.00	-0.43
5979	312602.78	5801344.17	21.69	22.67	161.34	-76.80	3.74	0.76	1.17	-0.61
5980	312582.78	5801354.17	19.51	23.49	161.30	-74.32	3.74	0.85	0.90	-0.23
5981	312592.78	5801354.17	19.79	23.20	161.31	-75.28	3.74	0.90	0.96	-0.29
5982	312582.78	5801364.17	19.35	23.42	161.34	-74.33	3.73	0.92	0.95	-0.23

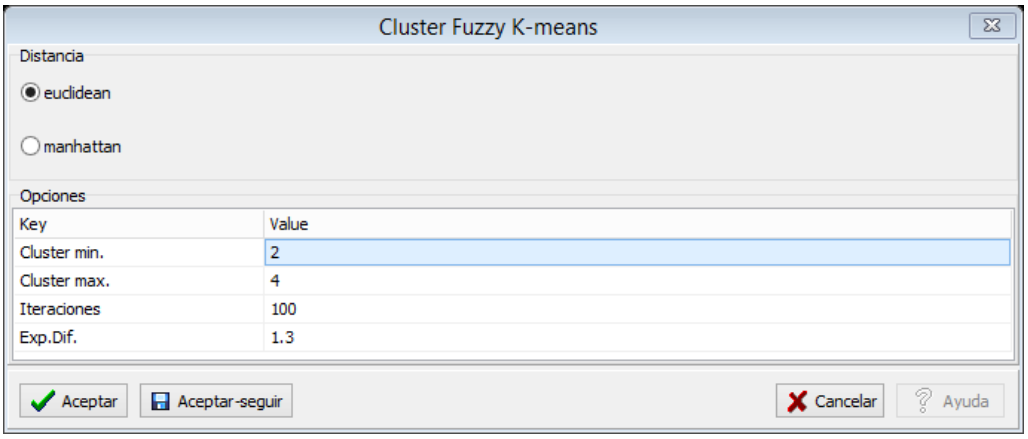
Real | Registros: 5982*10

Paso 3. *Cluster fuzzy k-means*. Las siguientes figuras ilustran el análisis de *cluster fuzzy k-means* con las componentes principales espaciales calculadas en el Paso 2.

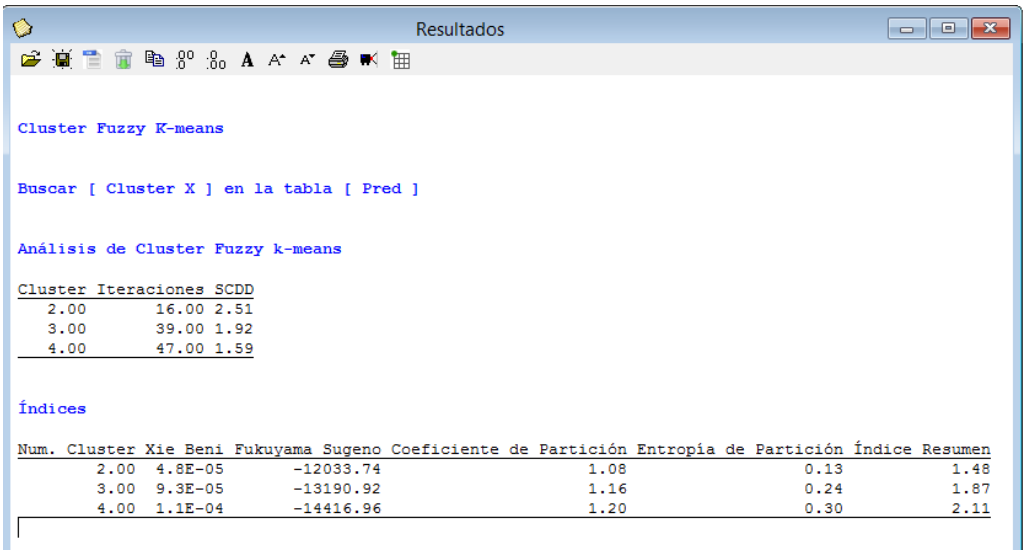




La medida de distancia a utilizar, el número de clusters a probar, la cantidad de iteraciones y el exponente difuso son argumentos que pueden modificarse en la ventana “Cluster Fuzzy K-means”.



La ventana "Resultados" de InfoStat, muestra los valores de los índices para cada número de clases. En la tabla de datos se adicionan columnas con las clases de manejo calculadas (X1, X2, X3).

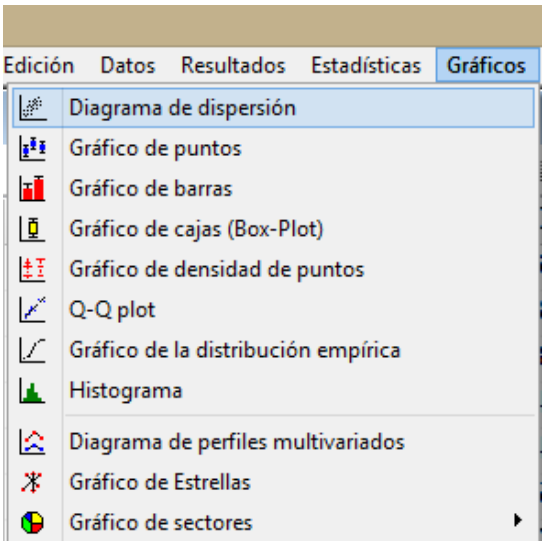


Pred

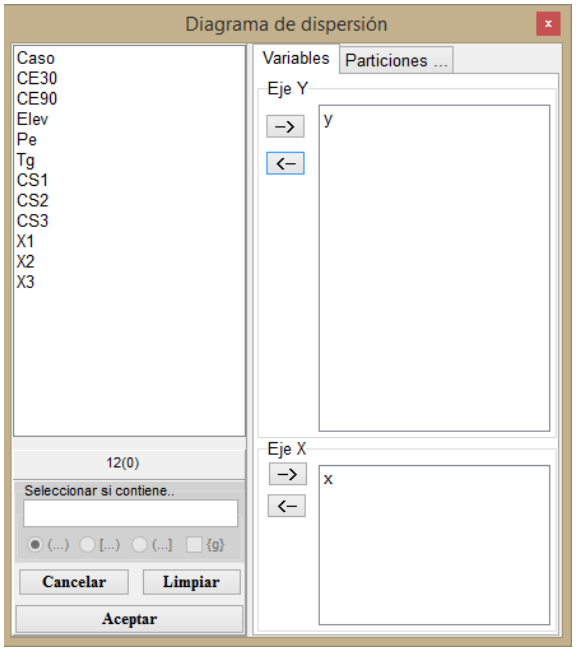
Caso	x	y	CE30	CE90	Elev	Pe	Tg	CS1	CS2	CS3	X1	X2	X3
5964	312572.78	5801324.17	20.54	23.60	161.17	-75.94	3.80	0.67	0.60	-0.49	1.00	3.00	2.00
5965	312582.78	5801324.17	21.29	23.32	161.23	-76.37	3.80	0.65	0.78	-0.63	2.00	3.00	2.00
5966	312592.78	5801324.17	22.16	22.89	161.30	-77.82	3.78	0.68	0.99	-0.77	2.00	3.00	2.00
5967	312602.78	5801324.17	23.00	22.46	161.33	-78.55	3.77	0.67	1.17	-0.89	2.00	3.00	2.00
5968	312612.78	5801324.17	23.43	22.42	161.38	-80.02	3.76	0.66	1.27	-0.92	2.00	3.00	2.00
5969	312622.78	5801324.17	23.32	22.50	161.42	-81.38	3.74	0.70	1.32	-0.85	2.00	3.00	2.00
5970	312562.78	5801334.17	20.02	23.72	161.18	-74.25	3.77	0.66	0.67	-0.33	1.00	3.00	2.00
5971	312572.78	5801334.17	20.17	23.73	161.16	-74.69	3.77	0.62	0.65	-0.32	2.00	3.00	2.00
5972	312582.78	5801334.17	20.69	23.45	161.24	-75.16	3.76	0.66	0.83	-0.43	1.00	3.00	2.00
5973	312592.78	5801334.17	21.52	22.98	161.30	-76.12	3.75	0.69	1.04	-0.59	1.00	3.00	2.00
5974	312602.78	5801334.17	22.54	22.49	161.34	-77.73	3.75	0.69	1.21	-0.76	2.00	3.00	2.00
5975	312612.78	5801334.17	23.41	22.22	161.38	-79.29	3.74	0.67	1.36	-0.89	2.00	3.00	2.00
5976	312572.78	5801344.17	19.91	23.87	161.24	-74.08	3.74	0.65	0.79	-0.25	1.00	3.00	2.00
5977	312582.78	5801344.17	20.05	23.61	161.27	-74.54	3.74	0.72	0.86	-0.29	1.00	3.00	2.00
5978	312592.78	5801344.17	20.69	23.14	161.30	-75.24	3.74	0.76	1.00	-0.43	1.00	3.00	2.00
5979	312602.78	5801344.17	21.69	22.67	161.34	-76.80	3.74	0.76	1.17	-0.61	1.00	3.00	2.00
5980	312582.78	5801354.17	19.51	23.49	161.30	-74.32	3.74	0.85	0.90	-0.23	1.00	3.00	2.00
5981	312592.78	5801354.17	19.79	23.20	161.31	-75.28	3.74	0.90	0.96	-0.29	1.00	3.00	2.00
5982	312582.78	5801364.17	19.35	23.42	161.34	-74.33	3.73	0.92	0.95	-0.23	1.00	3.00	2.00

Real | Registros: 5982*13

Paso 4. *Visualización*. Con la opción "Diagrama de dispersión" del menú "Graficos", se puede visualizar el mapa con las clases de manejo delimitadas previamente.

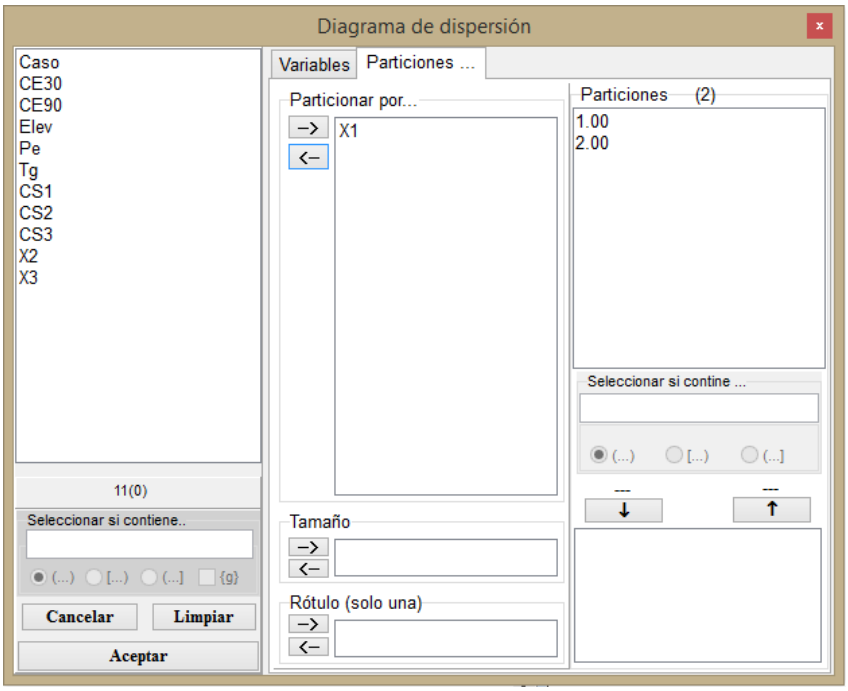


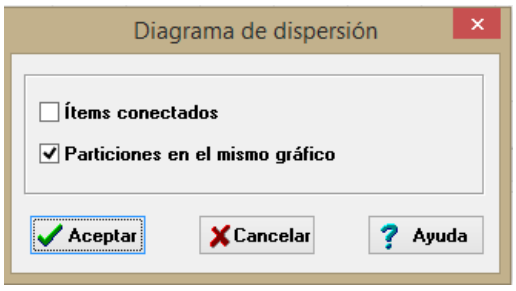
En la venta de selección de variables se debe colocar en los casilleros "X" e "Y" las respectivas coordenadas espaciales.



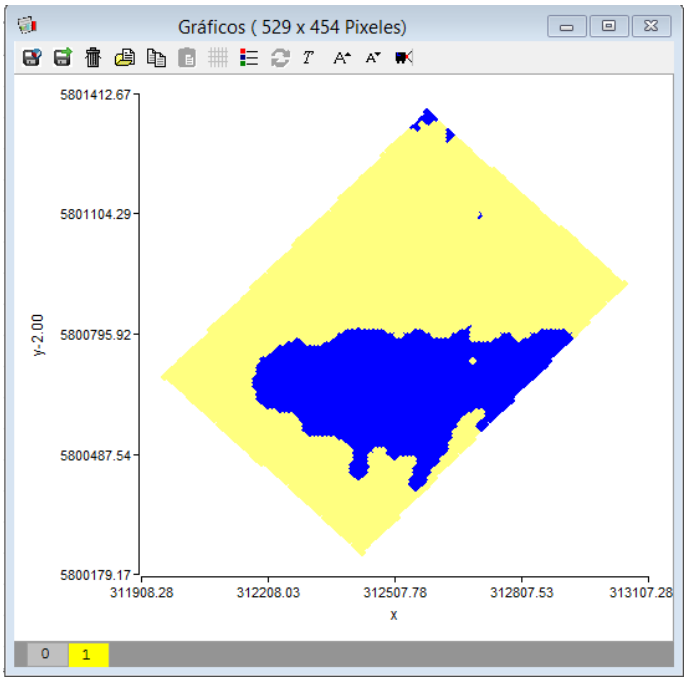
En la solapa "Particiones" se debe indicar una de las clases de manejo. Accionamos "Aceptar" y en la siguiente ventana se deja marcado "Particiones

en el mismo gráfico" para que las categorias de las clases de manejo se desplieguen en un mismo gráfico.

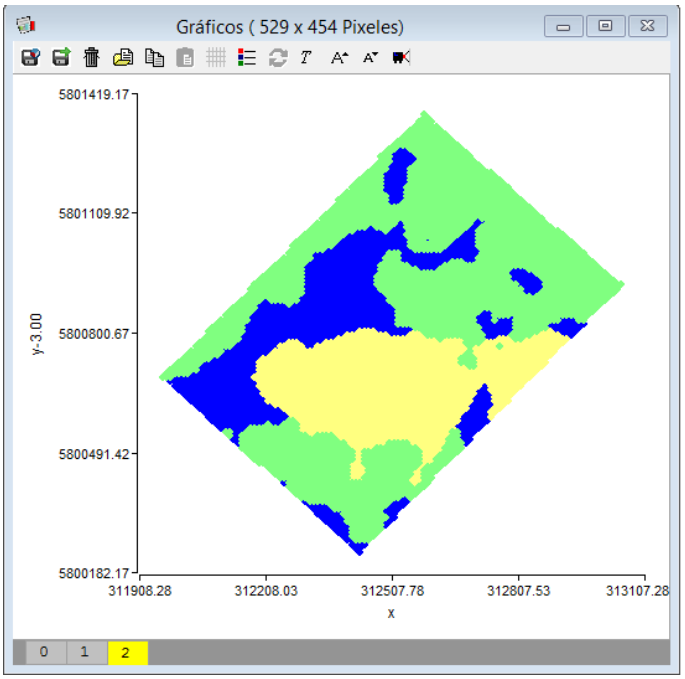


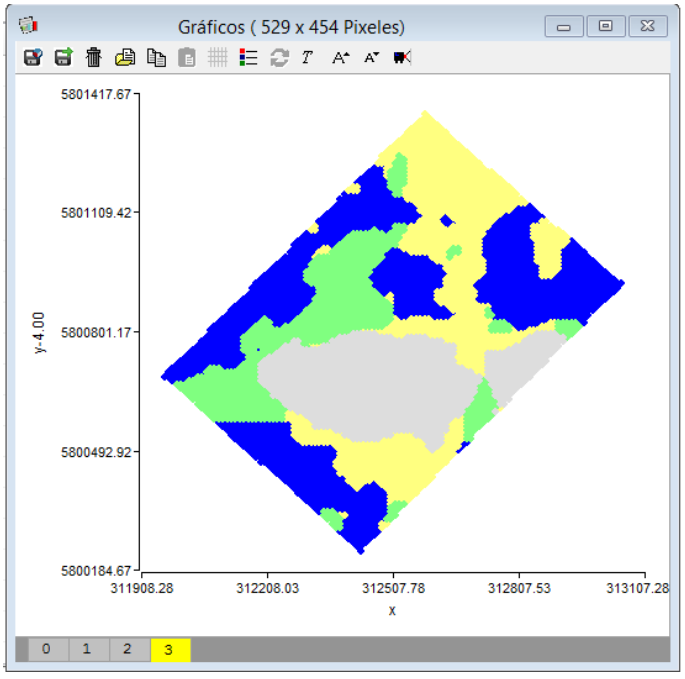


La ventana de gráficos de InfoStat muestra el mapa de las clases de manejo. Este gráfico puede editarse desde las herramientas gráficas que aparecen al hacer clic sobre el mapa. En este ejemplo, se puso un tamaño de punto igual a 6.

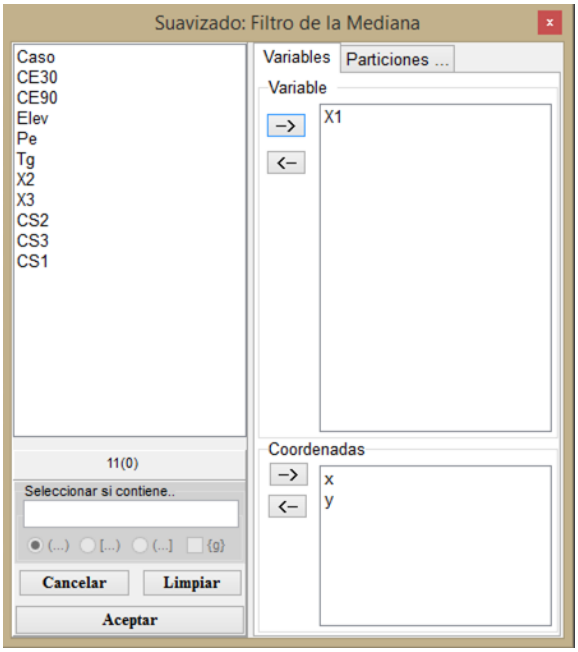


El paso 4 puede repetirse con 3 y 4 clases de manejo, especificadas una por una desde la solapa “Particiones”. Los mapas obtenidos se muestran a continuación.

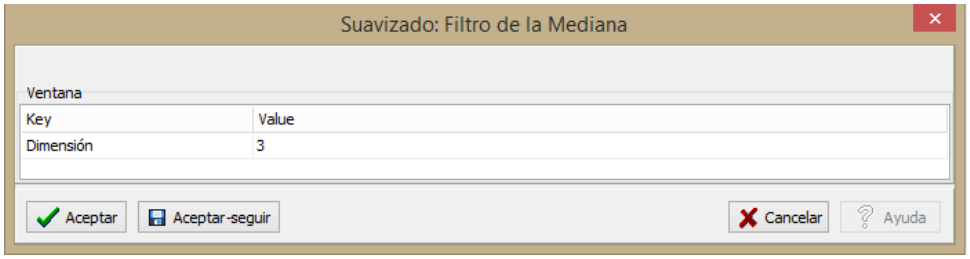

Gráficos (529 x 454 Pixeles)
5801419.17
5801109.92
5800800.67
5800491.42
5800182.17
y-3.00
311908.28
312208.03
312507.78
312807.53
313107.28
x
0
1
2


Gráficos (529 x 454 Pixeles)
5801417.67
5801109.42
5800801.17
5800492.92
5800184.67
y-4.00
311908.28
312208.03
312507.78
312807.53
313107.28
x
0
1
2
3

Paso 5. *Suavizado*. Seleccionar el menú "Estadística Espacial", sub-menú "Suavizado". Colocar la variable X1, que contiene la clase a la que pertenece cada punto, en el cuadro "Variable" y las coordenadas x e y en el casillero "Coordenadas".



La dimensión del filtro se especifica desde la ventana "Suavizado Filtro de la Mediana", en este ejemplo se usa un valor de 3. El resultado se presenta es un mapa y una nueva tabla con las coordenadas y la clase suavizada.



Suavizado: Filtro de la Mediana

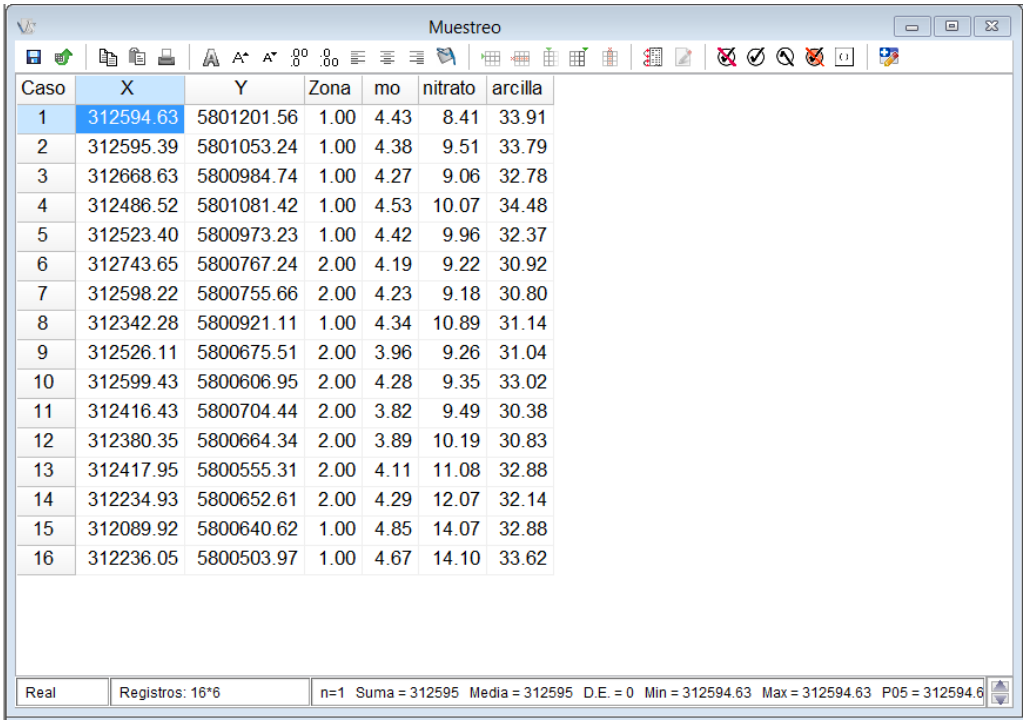
Clasificación Suavizada

Caso	x	y	Suavizado
1	312432.78	5800234.17	2.00
2	312422.78	5800244.17	2.00
3	312432.78	5800244.17	2.00
4	312412.78	5800254.17	2.00
5	312422.78	5800254.17	2.00
6	312432.78	5800254.17	2.00
7	312442.78	5800254.17	2.00
8	312402.78	5800264.17	2.00
9	312412.78	5800264.17	2.00
10	312422.78	5800264.17	2.00
11	312432.78	5800264.17	2.00
12	312442.78	5800264.17	2.00
13	312452.78	5800264.17	2.00
14	312392.78	5800274.17	2.00
15	312402.78	5800274.17	2.00
16	312412.78	5800274.17	2.00
17	312422.78	5800274.17	2.00
18	312432.78	5800274.17	2.00
19	312442.78	5800274.17	2.00
20	312452.78	5800274.17	2.00
21	312462.78	5800274.17	2.00
22	312282.78	5800284.17	2.00

Real | Registros: 5982*3 | n=5982 Suma = 10188.0 Media = 1.703

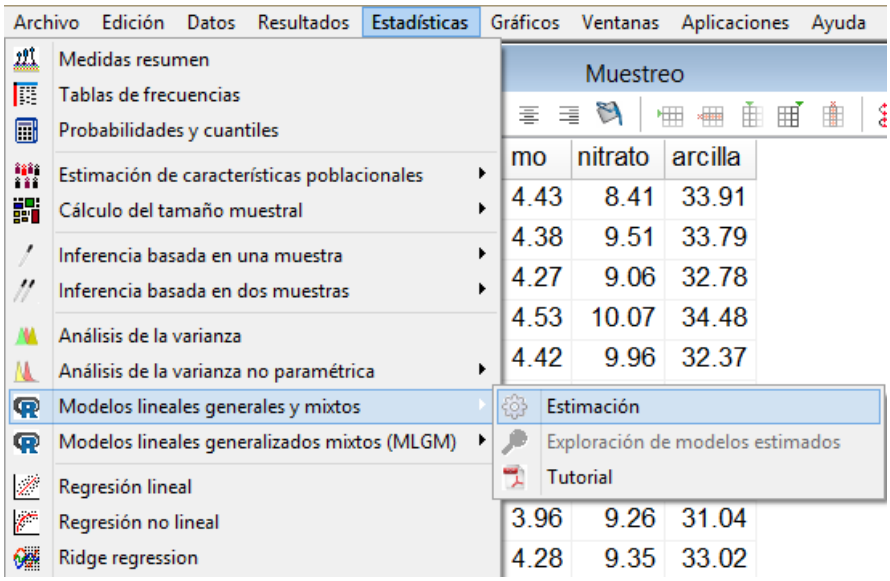
VALIDACIÓN DE ZONAS DE MANEJO

Paso 6. *Validación*. El archivo con los datos *Muestreo.txt* puede abrirse como se indicó en el paso 1.

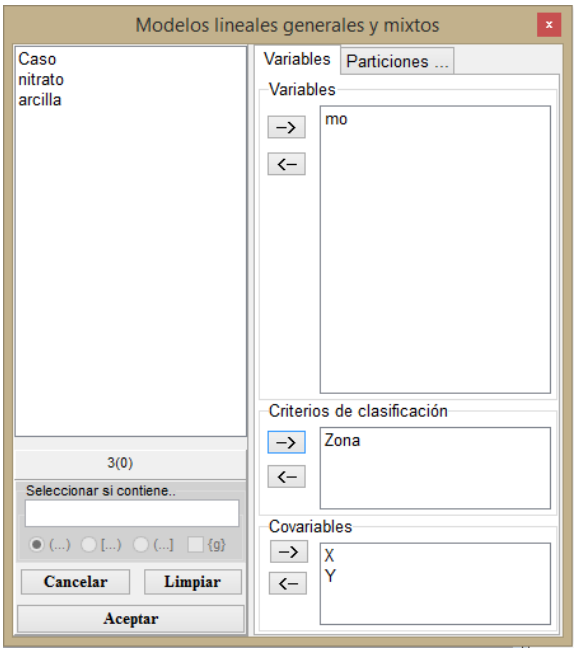


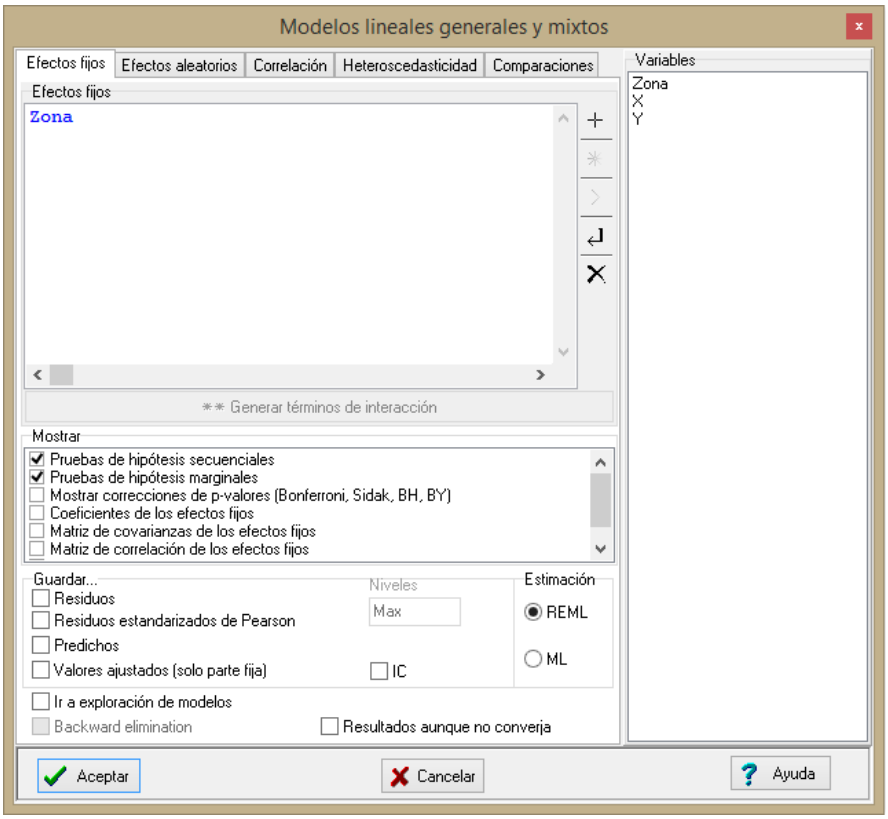
Caso	X	Y	Zona	mo	nitrato	arcilla
1	312594.63	5801201.56	1.00	4.43	8.41	33.91
2	312595.39	5801053.24	1.00	4.38	9.51	33.79
3	312668.63	5800984.74	1.00	4.27	9.06	32.78
4	312486.52	5801081.42	1.00	4.53	10.07	34.48
5	312523.40	5800973.23	1.00	4.42	9.96	32.37
6	312743.65	5800767.24	2.00	4.19	9.22	30.92
7	312598.22	5800755.66	2.00	4.23	9.18	30.80
8	312342.28	5800921.11	1.00	4.34	10.89	31.14
9	312526.11	5800675.51	2.00	3.96	9.26	31.04
10	312599.43	5800606.95	2.00	4.28	9.35	33.02
11	312416.43	5800704.44	2.00	3.82	9.49	30.38
12	312380.35	5800664.34	2.00	3.89	10.19	30.83
13	312417.95	5800555.31	2.00	4.11	11.08	32.88
14	312234.93	5800652.61	2.00	4.29	12.07	32.14
15	312089.92	5800640.62	1.00	4.85	14.07	32.88
16	312236.05	5800503.97	1.00	4.67	14.10	33.62

Para realizar la comparación de medias utilizando modelos con correlación espacial se debe seleccionar “Estadística” submenú “Modelos Lineales Mixtos”.

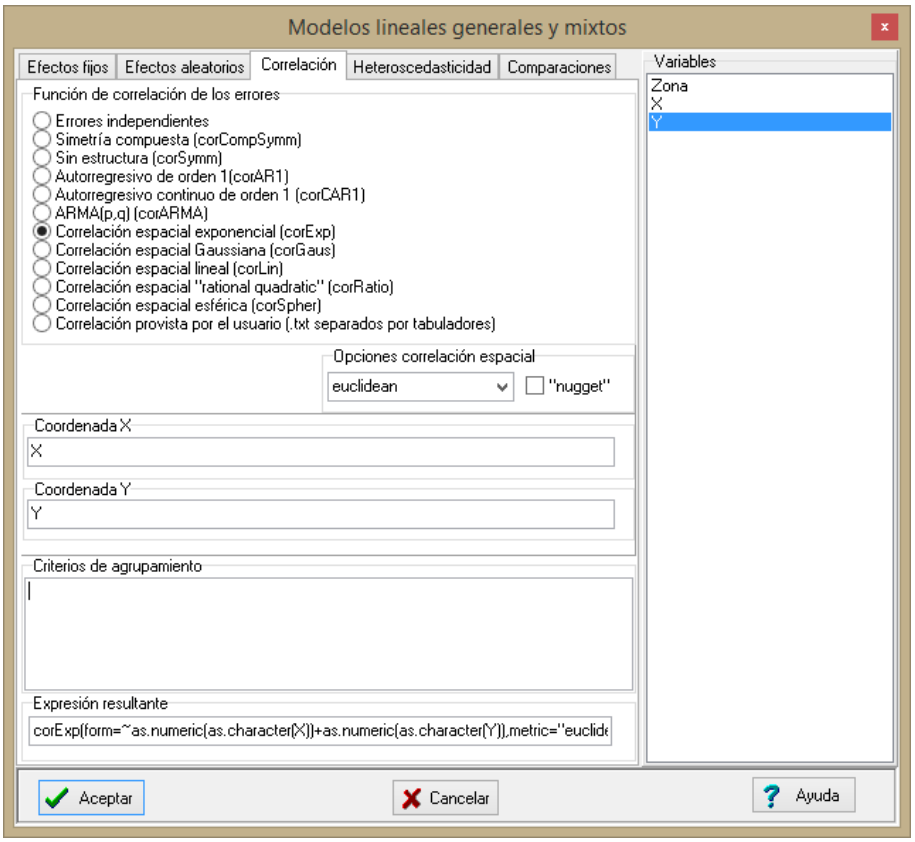


Los modelos se especifican a partir de datos de materia orgánica clasificados por zona, cargando las coordenadas como covariables. En el paso siguiente, la variable zona se coloca en la solapa “Efecto fijo”.

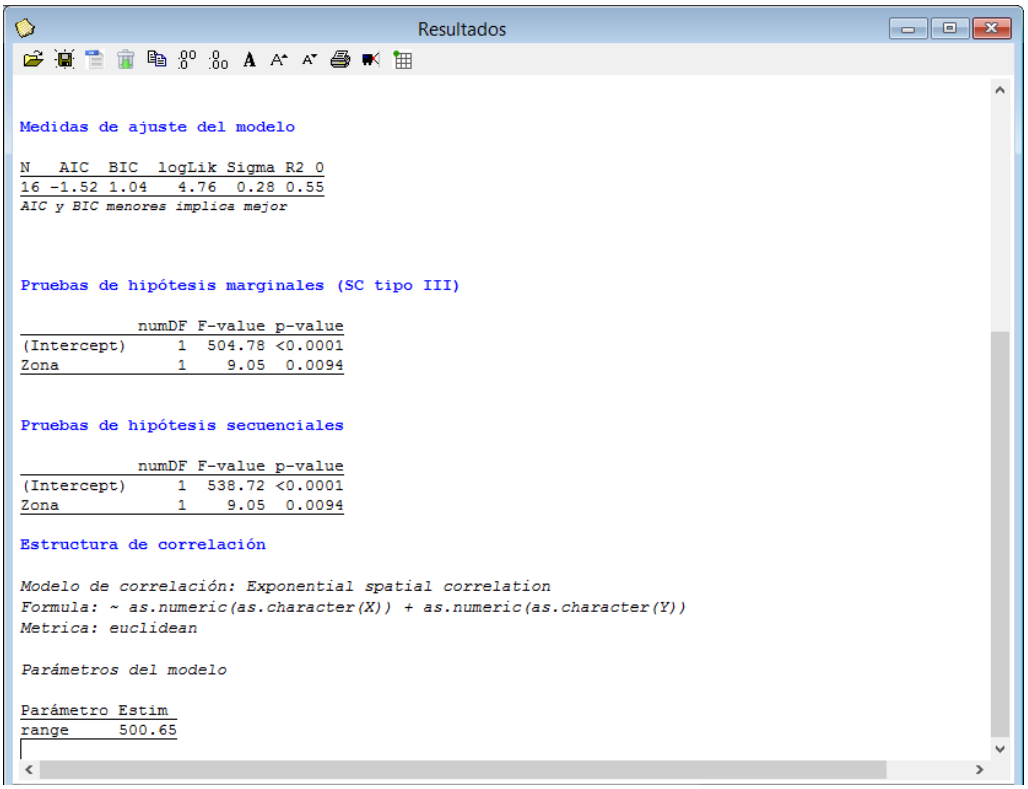




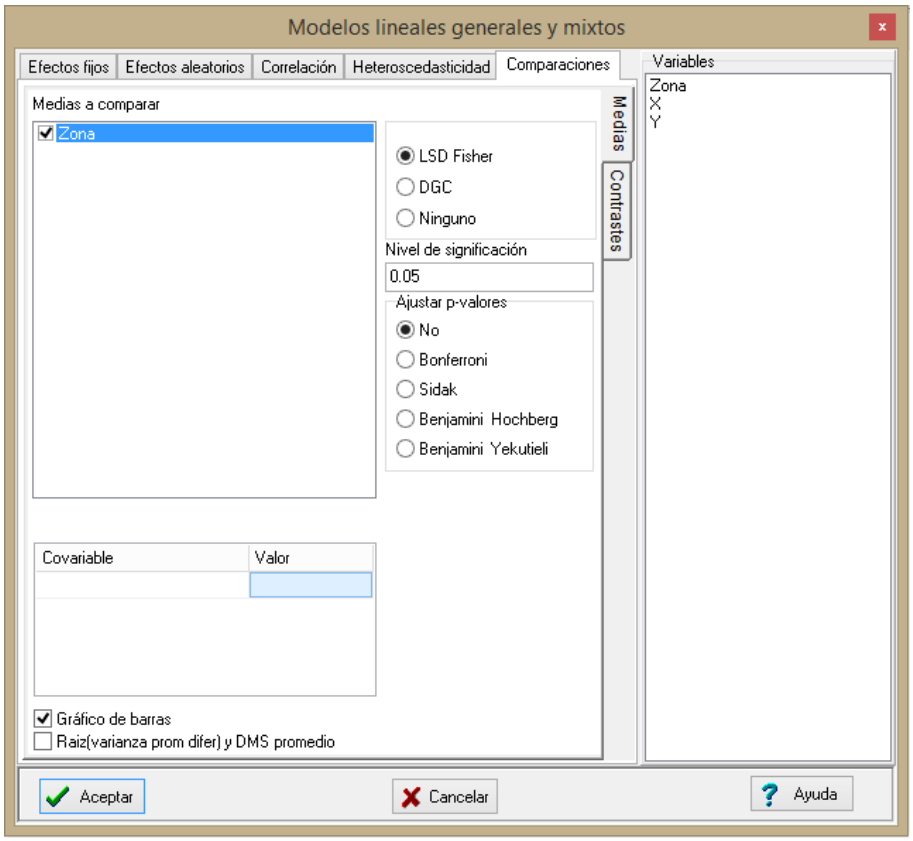
La función de correlación se declara en la solapa “Correlación”. En esta pantalla también puede escogerse la medida de distancia, la presencia o ausencia de efecto nugget y debe declararse las variables con las coordenadas en el eje X e Y.



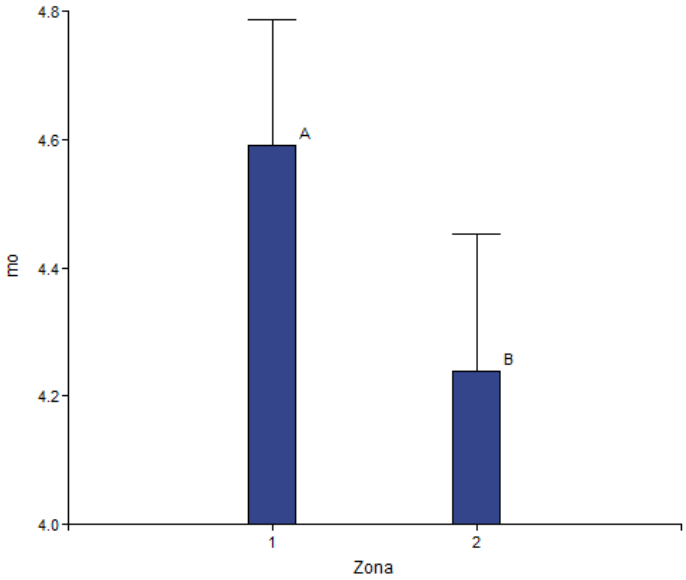
Los resultados desplegados muestran, entre otras cosas, el Criterio de Información de Akaike, que permite la comparación de modelos con diferentes estructuras de correlación.



Paso 7. *Comparación entre zonas*. El gráfico de barras de las diferencias en materia orgánica para las diferentes zonas de manejo puede pedirse desde la ventana de modelos mixtos, en la solapa "Comparaciones". Para realizar la comparación de medias debe tildarse "zona" en "Medias a comparar" y "Gráfico de barras" en la parte inferior izquierda de la ventana.



En la ventana de resultados se adicionan detalles de la comparación entre zonas y el gráfico de barras se despliega en la ventana de gráficos. Desde las herramientas gráficas puede editarse el color y la escala.


4.8
4.6
4.4
4.2
4.0
mo
A
B
1
2
Zona

REFERENCIAS BIBLIOGRÁFICAS

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.

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.

Arce G. R., (2005). Nonlinear signal processing: a statistical approach. John Wiley and Son, New Jersey.

Barbieri P., Echeverría H., 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.

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

Bivand R., (2014). spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.5-75/r559. http://R-Forge.R-project.org/projects/spdep/

Bivand R., Keitt T. and Rowlingson B. (2014). rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.8-16. http://CRAN.R-project.org/package=rgdal

Bland J. M. Altman D. G., (1995). Multiple significance tests: the Bonferroni method. BMJ. 310 (6973), 170.

Bremner J., Keeney D., (1966). Determination and isotope-ratio analysis of different forms of nitrogen in soil: 3. Exchangeable ammonium, nitrate and nitrite by extraction-distillation methods. Soil Sci. Soc. Am. Proc. 30, 577‑582.

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

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

Córdoba M., Bruno, C., Costa, J. L., Balzarini, M., (2013). Subfield management class delineation using cluster analysis from spatial principal components of soil variables. Comp. Electron. Agric. 97: 6-14.

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

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

Di Rienzo J. A., Casanoves F., Balzarini M. G., Gonzalez L., Tablada M., Robledo C. W. InfoStat versión 2014. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar

Draper N. R. and Smith H., (1988). Applied Regression Analysis. John Wiley and Sons, New York.

Dray S. Legendre P. Peres-Neto P. R., (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecol. Model. 196(3), 483–493.

Farahani H. J. and Flynn R. L., (2007). Map quality and zone delineation as affected by width of parallel swaths of mobile agricultural sensors. Biosyst. Eng. 96, 151–159.

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.

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

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.

Gallardo A. y Maestre F. T., (2008). Métodos geoestadísticos para el análisis de datos ecológicos espacialmente explícitos. En: Maestre FT, Escudero A, Bonet A (eds) Introducción al Análisis Espacial de Datos en Ecología y Ciencias Ambientales: Métodos y Aplicaciones. Universidad Rey Juan Carlos, pp 215–272

Isaaks E. H. and Srivastava R. M., (1989). An Introduction to Applied Geostatistics. Oxford Univ. Press, New York, 561 pp.

Journel A. G. and Huijbregts C. J., (1978). Mining geostatistics. Academic Press, Inc., London, UK.

Khosla R., Westfall D. G., Reich R. M., Mahal J. S., Gangloff W. J., (2010). Spatial variation and site-specific management zones. En: Margaret, O. (Ed.), Geostatistical applications in precision agriculture. Springer, New York, pp. 195–219.

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.

Lee J. and Wong D. W. S., (2001). Statistical Analysis with Arcview GIS. John Wile and Son, New York, 192 pp.

Matheron G., (1971). The theory of regionalized variables and its applications. Cahiers du Centre de Morphologie Mathrmatique de Fontainebleau, No. 5, Paris, 211 pp.

Meyer D., Dimitriadou E., Hornik K., Weingessel A. and Leisch F., (2014). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-4. http://CRAN.R-project.org/package=e1071.

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.

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., (2013). Precision agriculture and geostatistics: How to manage agriculture more exactly. Significance, 10(2): 17–22.

Patterson H. D. and Thompson R., (1971). Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554.

Pebesma E. J., (2004). Multivariable geostatistics in S: the gstat package. Comp. Geosci. 30: 683-691.

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., Sarkar D. and R Core Team. (2014). _nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-118, <URL: http://CRAN.R-project.org/package=nlme>.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ribeiro Junior P. J. and Diggle P. J., (2001). geoR: a package from geostatistical analysis. RNEWS 1(2):15–18.

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. Bs As.

Satorre E. H. and Slafer G. A., (1999). Wheat production systems of the Pampas. In: Satorre, E. M. and Slafer, G. A. (Eds.), Wheat Ecology and Physiology of Yield Determination. The Haworth Press Inc., New York, pp. 333–348.

Schabenberger O. and Gotway C. A., (2004). Statistical Methods for Spatial Data Analysis. Taylor and Francis. Chapman and Hall/CRC, 488pp.

Schwarz G., (1978). Estimating the dimension of a model. Ann. Stat. 6:461–464.

Searle S.R., Casella G. and McCullogh C.E., (1992). Variance Components. Wiley, New York.

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.

Taylor J. A., McBratney A. B. and Whelan B. M., (2007). Establishing Management Classes for Broadacre Agricultural Production. Agron. J. 99: 1366–1376.

Veris Technologies, (2001). Frequently asked questions about soil electrical conductivity. Veris Technologies, Salina. KS. http://www.veristech.com (acceso 9/02/14).

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.

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.

Windham M. P., (1981). Cluster validity for fuzzy clustering algorithms. 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.

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.

ANEXO I

DESCRIPCIÓN DE LA BASE DE DATOS DE ILUSTRACIÓN

Los datos utilizado para la ilustración fueron recolectados de un lote en producción agrícola (84 ha) bajo agricultura continua con rotaciones de cultivos anuales, ubicado en la región pampeana Argentina. La región 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).

El grupo de Calidad y Manejo de Suelo y Agua del INTA Balcarce, coordinado por el Dr. José Luis Costa, registró mediciones georreferenciadas de conductividad eléctrica aparente (CE) en dos profundidades 0-30 cm (CE30) y 0-90 cm (CE90), elevación, profundidad de suelo y rendimiento de trigo (Tg), entre otras variables. 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.

Luego de zonificar el lote, se tomaron 8 puntos de muestreo de suelos georreferenciadas dentro de cada zona 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 de materia orgánica (MO) 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, 1966). De la base de datos original se generaron tres bases de datos usados en la ilustración del manual de buenas prácticas para el análisis de los datos espaciales. La primera, denominada *CE30.txt*, contiene los datos de mediciones de CE30 y las coordenadas en grados decimales de los

sitios de observaciones del lote. La base *datos2.txt* contiene también datos de mediciones de CE30 pero con coordenadas cartesianas. El archivo *Pred.txt* contiene la predicción espacial para el re-escalado de las variables CE, elevación y profundidad del suelo, usando una grilla común a todas ellas de 10 × 10 m. Finalmente la base *Muestreo.txt* contiene datos del muestreo de suelo de MO, nitratos y arcilla. Las bases de datos y los script de R para realizar los análisis presentados en este manual, se encuentran disponibles en: https://drive.google.com/folderview?id=0B_8UVonay55CfkZlYnQ5d3lBcjFER3NYTlkwZFdISnlseUFNSWZJVmdoakdnLWpfM1FrNVU&usp=sharing

UNC Universidad Nacional de Córdoba

FCA Facultad de Ciencias Agropecuarias

Secretaría de CIENCIA y TECNOLOGIA

Ministerio de INDUSTRIA, COMERCIO, MINERÍA Y DESARROLLO CIENTÍFICO TECNOLÓGICO

GOBIERNO DE LA PROVINCIA DE CÓRDOBA

Programa de Transferencia de Resultados de Investigación y Comunicación Pública de la Ciencia (PROTRI)

El Programa PROTRI de la Secretaría de Ciencia y Tecnología del Gobierno de la Provincia de Córdoba, procura identificar los resultados, experiencias o saberes transferibles generados por los grupos de investigación de las universidades, empresas o centros de ciencia y tecnología cordobeses, para promover el intercambio fructífero con otras áreas del sector social y productivo provincial, potencialmente usuarios de nuevos conocimientos y mejores prácticas, persiguiendo una mejora en la calidad de vida y un aumento de las oportunidades territoriales.
El Programa financia: ciclos de capacitación o asesoramiento, documentos de divulgación científica, guías/manuales de buenas prácticas, infografías impresas, cuadernos de experimentos, infografías digitales y videos cortos. Para postular a un subsidio, cada equipo de investigación formula su proyecto a partir de una demanda, de un compromiso específico previamente acordado con algún sector social, científico, educativo o productivo, que será finalmente el receptor de la transferencia.

Dirección de Promoción de Actividades Científicas
Subsecretaría de Promoción Científica