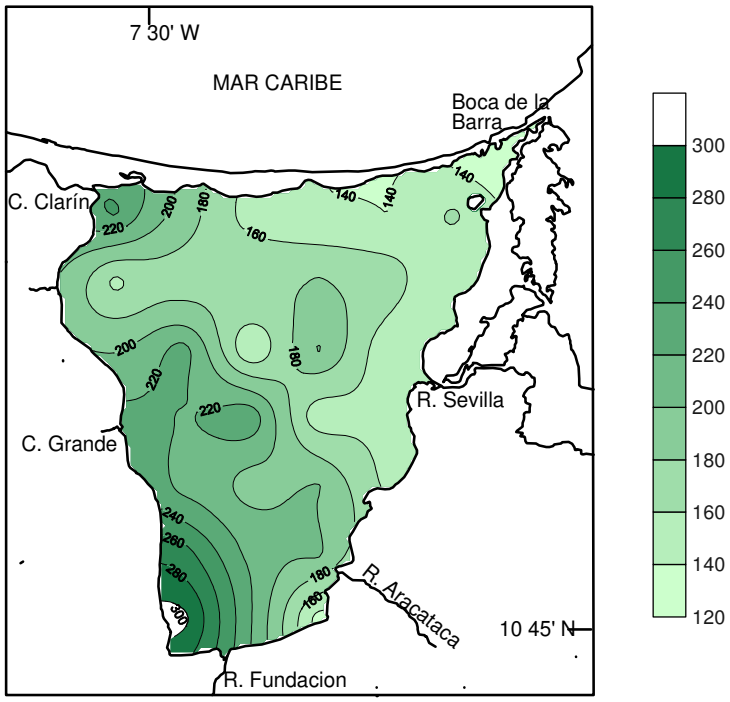
# INTRODUCCION A LA GEOESTADISTICA

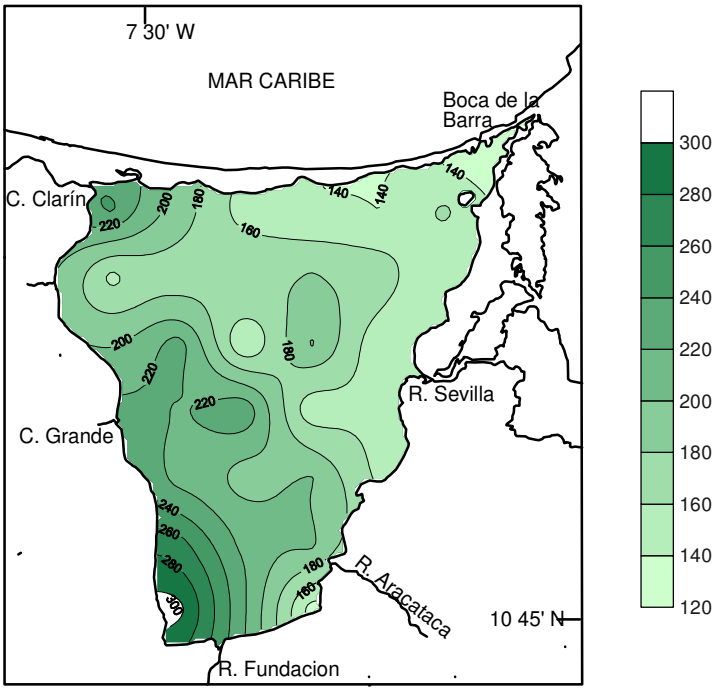



## Teoría y Aplicación

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Sede Bogotá
Facultad de Ciencias
Departamento de Estadística.

# INTRODUCCION A LA GEOESTADISTICA




# Teoría y Aplicación

**Ramón Giraldo Henao.**

*Profesor Asistente*

*Departamento de Estadística*

*Universidad Nacional de Colombia*

*Sede Bogotá*

# *Contenido*

## *4. Predicción Espacial*

4.1. Predicción Espacial Optima
4.2. Definición de Kriging
4.3. Kriging Ordinario
4.4. Otros Métodos Kriging
  4.4.1. Simple
  4.4.2. Bloques
  4.4.3. Universal
  4.4.4. Residual
  4.4.5. Indicador
  4.4.6. LogNormal y Multigaussiano.
4.5. Aplicaciones.

## *5. Temas Especiales*.

5.1. Cokriging Ordinario
5.2. Kriging sobre Ejes Factoriales
5.3. Diseño de Redes de Muestreo.
5.4. Simulación
5.5. Aplicaciones.

## *6. Apéndice*

6.1. Indicador $IGC_i(P)$
6.2. Álgebra de Matrices
6.3. Conceptos de Probabilidad
6.4. Revisión de Algunos Métodos Estadísticos.

## *7. Referencias.*

# *Prefacio*

La necesidad de acudir a herramientas estadísticas para el análisis de datos en todas las áreas del conocimiento, ha hecho que aparezcan con el correr de los años nuevas metodologías que, no obstante se centran en fundamentos probabilísticos comunes, son específicas para cada una de las diversas disciplinas del saber. Algunos ejemplos son, entre otros, la econometría, psicometría o la bioestadística. La gran relevancia que tiene actualmente a nivel mundial el tema ambiental ha hecho que los profesionales en estadística encaminen esfuerzos en el desarrollo de nuevas técnicas apropiadas para el análisis de información enmarcada dentro de este contexto. Como consecuencia de este impulso surgió una nueva rama de la estadística, denominada *environmetrics (estadística ambiental)*. Dentro de esta última, los métodos geoestadísticos juegan un papel preponderante.

El presente documento tiene como propósito servir de consulta a geólogos, biólogos, ecólogos, agrónomos, ingenieros, meteorólogos y todos aquellos profesionales que se encargan del estudio de información ambiental georreferenciada. Se toma como base para las aplicaciones información de variables fisicoquímicas y biológicas medidas en un estuario ubicado en la costa norte de Colombia. La razón fundamental para lo anterior, es que este escrito es uno de los resultados centrales de un proyecto de investigación[1], cuyo objetivo fundamental fue el de evaluar la aplicabilidad de algunos procedimientos estadísticos en el análisis de datos medidos en este tipo de ecosistemas.

El documento tiene un enfoque teórico-práctico. Para el seguimiento completo de la teoría descrita se requiere tener conocimientos básicos de álgebra de matrices y de estadística matemática. Sin embargo aquellas personas que estén poco familiarizadas con estos temas, podrán obviar la lectura de algunas secciones en las que se hacen desarrollos teóricos y centrar su atención en la filosofía de los métodos presentados y en las aplicaciones mostradas en cada uno de los capítulos del documento. Una resumen no exhaustivo de conceptos de álgebra lineal y de estadística es hecho al final en el apéndice.

No obstante en el escrito se cubren diversos temas geoestadísticos y se hacen aplicaciones de métodos recientes, es necesario acudir a la lectura de artículos científicos y textos avanzados para lograr un buen dominio de esta metodología. Un libro formal desde el punto de vista matemático con aplicaciones en diversas disciplinas es Cressie (1993). Otras referencias pueden ser tomadas de la bibliografía.

[1] Proyecto "Análisis y aplicación de técnicas geoestadísticas en la modelación de procesos estocásticos relacionados con variables ecológicas en ambientes estuarinos", cofinanciado por INVEMAR y COLCIENCIAS.

# *Introducción*

El estudio de fenómenos con correlación espacial, por medio de métodos geoestadísticos, surgió a partir de los años sesenta, especialmente con el propósito de predecir valores de las variables en sitios no muestreados. Como antecedentes suelen citarse trabajos de Sichel (1947; 1949) y Krige (1951). El primero observó la naturaleza asimétrica de la distribución del contenido de oro en las minas surafricanas, la equiparó a una distribución de probabilidad lognormal y desarrolló las fórmulas básicas para esta distribución. Ello permitió una primera estimación de las reservas, pero bajo el supuesto de que las mediciones eran independientes, en clara contradicción con la experiencia de que existen "zonas" más ricas que otras. Una primera aproximación a la solución de este problema fue dada por geólogo G. Krige que propuso una variante del método de medias móviles, el cual puede considerarse como el equivalente al krigeado simple que, como se verá más adelante, es uno de los métodos de estimación lineal en el espacio con mayores cualidades teóricas. La formulación rigurosa y la solución al problema de predicción (estimación en muchos textos geoestadísticos) vino de la mano de Matheron (1962) en la escuela de minas de París. En los años sucesivos la teoría se fue depurando, ampliando su campo de validez y reduciendo las hipótesis necesarias (Samper y Carrera, 1990). De la minería las técnicas geoestadísticas, se han "exportado" a muchos otros campos como hidrología, física del suelo, ciencias de la tierra y más recientemente al monitoreo ambiental y al procesamiento de imágenes de satélite.

Aunque la aplicación de la herramienta geoestadística es bastante reciente, son innumerables los ejemplos en los que se ha utilizado esta técnica en estudios ambientales con el ánimo de predecir fenómenos espaciales (Robertson, 1987; Cressie y Majure, 1995; Diggle *et al*., 1995). La columna vertebral del análisis geoestadístico es la determinación de la estructura de autocorrelación entre los datos y su uso en la predicción a través de las técnicas conocidas como kriging y cokriging. Otros temas importantes dentro del estudio de información georreferenciada son el diseño de redes de muestreo (McBratney *et al*., 1981), la geoestadística multivariada (Wackernagel, 1995) y la simulación (Deutsh y Journel, 1992).

La geoestadística es solo una las áreas del análisis de datos espaciales. Es importante reconocer cuando la información georreferenciada es susceptible de ser analizada por medio de dicha metodología. Por ello en el documento se hace inicialmente una definición global de estadística espacial y se describen las características especiales que enmarcan cada una de sus áreas.

En el estudio de información georreferenciada, de forma análoga a como se procede en la aplicación de muchos procedimientos estadísticos, la primera etapa que se debe cumplir es la del análisis exploratorio de datos (AED). Esta busca identificar localización, variabilidad, forma y observaciones extremas. Por ello en el primer capítulo del escrito se hace una revisión de métodos empleados en el AED y se describen algunos particularmente útiles en el contexto del análisis de información georreferenciada. Posteriormente en el segundo capítulo, entrando en materia, se hace definición de conceptos básicos dentro de la teoría geoestadística.

En el tercer capítulo se describen los procedimientos empleados para identificar de manera experimental (con base en datos muestrales) la estructura de autocorrelación espacial, para algunas distancias dadas, de un conjunto de datos de una variable. Se muestra también como generalizar dicha estructura para cualquier distancia entre los sitios de observación. Una vez detectada la autocorrelación espacial, el siguiente paso es la predicción en sitios de la región de estudio donde no se ha hecho medición de la variable de interés. Esto es llevado a cabo por medio de alguno de los procedimientos kriging que son descritos en el capítulo cuatro. Por último, en el capítulo cinco, se hace referencia a temas especiales dentro del análisis geoestadístico como cokriging, componentes principales regionalizados, diseño de redes de muestreo y simulación. En cada sección del documento, después de que han sido expuestos los aspectos teóricos esenciales de cada técnica, se muestran aplicaciones practicas.

# *Capítulo Uno*

## *Datos Espaciales y Análisis Exploratorio*

En las secciones 1.1 y 1.2 se define estadística espacial y se mencionan sus subdivisiones. Lo anterior se hace con el propósito único de que el lector identifique el alcance del tema considerado dentro del escrito. Por ello a partir de la sección 1.3 de este capítulo y en los capítulos siguientes se consideran sólo temas referentes a geoestadística

### 1.1. Estadística Espacial.

Estadística espacial es la reunión de un conjunto de metodologías apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios (puntos del espacio o agregaciones espaciales) de una región. De manera más formal se puede decir que la estadística espacial trata con el análisis de realizaciones de un proceso estocástico $\{Z(s) : s \in D\}$, en el que s$\in R^d$ representa una ubicación en el espacio euclidiano d-dimensional, Z(s) es una variable aleatoria en la ubicación s y s varía sobre un conjunto de índices D$\subset R^d$ .

### 1.2. Areas de la Estadística Espacial.

La estadística espacial se subdivide en tres grandes áreas. La pertinencia de cada una de ellas está asociada a las características del conjunto D de índices del proceso estocástico de interés. A continuación se mencionan dichas áreas y se describen las propiedades de D en cada una de éstas.

**Geoestadística**: Las ubicaciones s provienen de un conjunto D *continuo* y son seleccionadas a juicio del investigador (D *fijo*). Algunos ejemplos de datos que pueden ser tratados con esta metodología son: Niveles de un contaminante en diferentes sitios de una parcela, contenidos auríferos de una mina, valores de precipitación en Colombia medida en las diferentes estaciones meteorológicas en un mes dado o los niveles piezométricos de un acuífero. En los ejemplos anteriores es claro que hay continuidad espacial, puesto que en cualquier sitio de la parcela, de la mina, de Colombia o del acuífero pueden ser medias las correspondientes variables. Es importante resaltar que en geoestadística el propósito esencial es la interpolación y si no hay continuidad espacial pueden hacerse predicciones carentes de sentido. Por ejemplo si la variable medida es producción de café en las fincas cafeteras del departamento del Quindío, hacer interpolación espacial y realizar un mapa de distribución de la producción cafetera puede ser carente de sentido porque podrían hacerse predicciones sobre áreas urbanas o no cultivadas con café. Además de lo anterior las mediciones, no obstante sean georreferenciadas, corresponden a una agregación espacial (finca) más que a un punto del espacio. En la parte de arriba, al comienzo de este párrafo, se mencionó que D debía ser fijo. A este respecto cabe aclarar que el investigador puede hacer selección de puntos del espacio a conveniencia o puede seleccionar los sitios bajo algún esquema de muestreo probabilístico.

- **Lattices (enmallados)**: Las ubicaciones s pertenecen a un conjunto D *discreto* y son seleccionadas por el investigador (D *fijo*). Estas pueden estar regular o irregularmente espaciadas. Algunos ejemplos de datos en lattices son los siguientes**:** Tasa de morbilidad de hepatitis en Colombia medida por departamentos, tasa de accidentalidad en sitios de una ciudad, producción de caña de azúcar en el departamento del Valle del Cauca según municipio, colores de los pixeles en interpretación de imágenes de satélite. En los ejemplos anteriores se observa que el conjunto de ubicaciones de interés es discreto y que estas corresponden a agregaciones espaciales más que a un conjunto de puntos del espacio. Es obvio que la interpolación espacial puede ser carente de sentido con este tipo de datos.

- **Patrones Espaciales**: las ubicaciones pertenecen a un conjunto D que puede ser *discreto o continuo* y su selección no depende del investigador (D *aleatorio*). Ejemplos de datos dentro de esta área son: Localización de nidos de pájaros en una región dada, puntos de imperfectos dentro de una placa metálica, ubicación de los sitios de terremoto en Colombia o cuadrantes de una región con presencia de una especie particular. Debe notarse que en los ejemplos anteriores hay aleatoriedad en la selección de los sitios, puesto que la ubicación de los nidos de los pájaros, de los imperfectos dentro de la placa metálica, de los sitios de terremoto o de los cuadrantes con presencia de la especie, no dependen del criterio del investigador. Una vez se ha hecho la selección de sitios es posible hacer medidas de variables aleatorias en cada uno de ellos. Por ejemplo si en primera instancia se establece la ubicación de árboles de pino dentro de un bosque, es posible que sea de interés medir en cada uno de los árboles el diámetro o la altura. En general el propósito de análisis en estos casos es el de determinar si la distribución de los individuos dentro de la región es aleatoria, agregada o uniforme.

### 1.3. Datos Georrferenciados

Las mediciones de las características de interés en un estudio regionalizado tienen implícitamente asociadas las coordenadas de los sitios en donde estas fueron tomadas. Cuando el área de estudio es considerablemente grande se usa un geoposicionador para establecer dichas coordenadas. En otros casos, por ejemplo en diseños experimentales con parcelas, es suficiente con hacer asignaciones según planos cartesianos. Un esquema general de datos georreferenciados es el siguiente:

| *Sitio* | *Latitud Norte* | *Longitud Este* | $X_1$ | $X_2$ | . | . | . | $X_p$ |
|---|---|---|---|---|---|---|---|---|
| *1* | – | – | $x_{11}$ | $x_{12}$ | . | . | . | $x_{1p}$ |
| *2* | – | – | $x_{21}$ | $x_{22}$ | . | . | . | $x_{2p}$ |
| *3* | – | – | $x_{31}$ | $x_{32}$ | . | . | . | $x_{3p}$ |
| *4* | – | – | $x_{41}$ | $x_{42}$ | . | . | . | $x_{4p}$ |
| . | – | – | . | . | . | . | . | . |
| . | – | – | . | . | . | . | . | . |
| . | – | – | . | . | . | . | . | . |
| *n* | – | – | $x_{n1}$ | $x_{n2}$ | . | . | . | $x_{np}$ |

En la tabla anterior *n* es el número de sitios muestreados y p el de variables medidas en cada uno de ellos. Cada $x_{ij}$ corresponde a la medida de la variable $X_j$ ($j = 1, 2,..., p$) en el sitio $i$ ($i= 1, 2,..., n$), que puede ser cuantitativa o categórica. Algunas de las variables

pueden estar más intensamente muestreadas que las otras ($x_{ij}$ faltantes). Las coordenadas pueden ser planas, geográficas (grados, minutos y segundos) o cartesianas. Sin embargo la posible utilización de unas u otras depende del software empleado para los análisis.

### 1.4. Justificación del Análisis Exploratorio de Datos Espaciales .

En la aplicación de la geoestadística es de suma importancia, al igual que en otros procedimientos estadísticos (por ejemplo los modelos ARIMA dentro de la teoría de series de tiempo), el análisis gráfico. La identificación de valores extremos y su ubicación geográfica, la evaluación de la forma de la distribución y el cálculo de medidas de localización, variabilidad y correlación es muy importante para establecer si algunos supuestos necesarios para la aplicación de la teoría geoestadística son válidos o para definir que procedimiento de predicción es el más conveniente. Por ejemplo, como se verá en el capítulo cuatro, la decisión de usar kriging ordinario o kriging universal se fundamenta en identificar si la media es o no constante en la región. El uso de kriging log-normal se basa en un criterio empírico relacionado con la forma asimétrica de la distribución de los datos muestrales. La decisión de emplear cokriging depende de la detección de asociaciones entre las variables.

### 1.5. Gráficos Exploratorios

Al igual que en un estudio exploratorio clásico, cuando se dispone de información georreferenciada se pueden emplear *histogramas, diagramas de tallos y hojas y de caja y bigotes* (Hoaglin *et al*., 1983) con el propósito de identificar localización, variabilidad, forma y observaciones extremas. Adicionalmente los *gráficos de dispersión* son muy útiles tanto para la detección de relaciones entre las variables como para la identificación de tendencias en el valor promedio de la variable en la región (relación entre la variable medida y las coordenadas geográficas). Un supuesto fundamental en el análisis geoestadístico es que el fenómeno es estacionario, para lo cual, entre otros aspectos, el nivel promedio de la variable debe ser constante en todos los puntos del área de estudio. Una detección de tendencia en el gráfico de dispersión puede ser una muestra de que no se satisface dicho supuesto. El gráfico se construye tomando como eje de las abcisas la variable que representa la coordenada geográfica y en el eje de las ordenadas la variable cuantitativa de estudio. La observación de la nube de puntos resultante, incluso el ajuste de una línea de regresión, permite establecer de manera empírica si existe dicha tendencia. Un gráfico de dispersión entre valores de la variable separados por una distancia espacial dada (*dispersograma rezagado*) es útil en la detección de autocorrelación espacial. Otro gráfico que tradicionalmente se emplea en la descripción de datos espaciales es el de *datos clasificados según puntos de referencia* (media, mediana, cuartíles). Este permite comparar zonas del sistema de estudio respecto a las magnitudes de las variables.

## 1.6 *Aplicación: Estudio exploratorio de la distribución de datos fisicoquímicos y biológicos medidos en el estuario Ciénaga Grande de Santa Marta en Marzo de 1997.*

Con información de las variables salinidad, seston (mg/l), nitritos (µmol/l), silicatos(µmol/l) y clorofila a (µg/l) medidas en una jornada de muestreo realizada en marzo de 1997 en el estuario Ciénaga Grande de Santa Marta (CGSM)(Fig. 1), se realizó un estudio exploratorio de datos. Los resultados encontrados son descritos a continuación: En primera instancia, se evidencia en el diagrama de caja (Fig. 2) y en el gráfico de tallos y hojas (Fig. 3) que, con excepción de la variable nitritos, existe un comportamiento simétrico en las distribuciones de los datos. Se observa también en estas figuras, que en todas las variables se presentan algunos valores "atípicos" o muy alejados del comportamiento general antes mencionado. Lo anterior, antes de ser tomado como un indicador de alta variabilidad o de errores de medición, puede ser considerado como un reflejo del comportamiento espacial de las variables dentro del ecosistema. La simetría de la mayoría de las variables hace pensar que existe una gran zona en donde las condiciones del sistema respecto a la calidad del agua son bastante similares (esto podría ser lo que se conoce como cuerpo de agua de la CGSM) y los valores "alejados" pueden estar representando las condiciones de sitios específicos, particularmente especiales dentro del sistema, como son la zona más estuarina (sitios de muestreo cercanos al sitio Boca de la Barra, Fig.1) y las de desembocaduras de los ríos que bajan de la Sierra Nevada de Santa Marta (costado oriental y sur del sistema, Fig. 1).

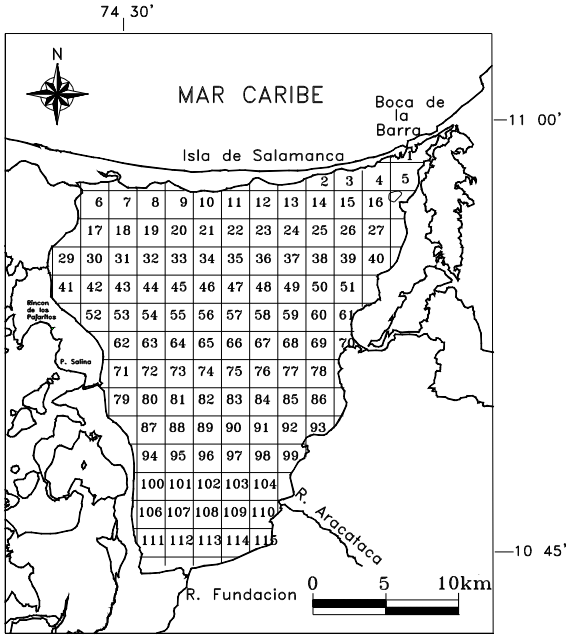



**Figura 1**. Área de estudio y cuadrículas en que fue subdividido el sistema Ciénaga Grande de Santa Marta para realizar la toma de muestras. Cada una de las 115 cuadrículas tiene un área de 4 km2. Los datos fueron tomados en el centro de cada una de ellas.

La afirmación de que no existen problemas de alta variabilidad y que por el contrario los datos medidos son bastante homogéneos, puede confirmarse con los valores de los coeficientes de variación (tabla 1). En su mayoría estos son menores del 30% y por consiguiente indicadores de poca heterogeneidad en la información.

**Tabla 1.** Medidas de localización y variabilidad de algunas variables medidas en la superficie de la columna de agua del estuario Ciénaga Grande de Santa Marta en Marzo de 1997.

| Medida | Salinidad | Seston (mg/l) | Nitritos (umol/l) | Silicatos (umol/l) | Clorofila a (ug/l) |
|---|---|---|---|---|---|
| Media | 17.6 | 218.28 | 0.436 | 244.94 | 132.44 |
| Mediana | 16.9 | 215 | 0.350 | 251.83 | 137.37 |
| Mínimo | 13.02 | 103 | 0.01 | 10.98 | 2.91 |
| Máximo | 34.9 | 318 | 1.61 | 358 | 198.3 |
| Cuartíl Inferior. | 15.97 | 191 | 0.210 | 226.52 | 124.43 |
| Cuartíl Superior. | 18.04 | 248 | 0.6 | 278.43 | 149.29 |
| Desviación Estándar | 2.79 | 41.1 | 0.309 | 61.43 | 31.30 |
| Coeficiente de Variación | 16.1 | 18.8 | 70.8 | 25.07 | 23.7 |

Las medidas de localización (media y mediana, tabla 1) toman valores similares a los reportados en otros estudios para la misma época del año. Una discusión a este respecto se encuentra en Hernández (1986) y Hernández y Gocke (1990).

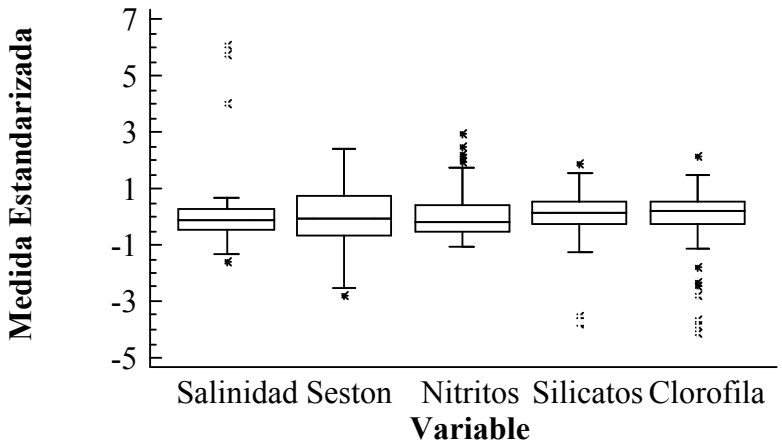



**Figura 2**. Diagramas de caja de algunas variables medidas en la superficie de la columna de agua del estuario Ciénaga Grande de Santa Marta en Marzo de 1997. Las variables fueron estandarizadas antes de construir los diagramas.

El gráfico de dispersión de la variable salinidad (una de las de mayor relevancia en el establecimiento del comportamiento espacial de las variables en el sistema) respecto a las coordenadas latitud y longitud (Fig. 4), permite apreciar una leve tendencia en la magnitud de la variable a lo largo de estas direcciones, lo que hace suponer que, a pesar de la homogeneidad antes mencionada, el valor promedio de la misma no es constante en toda la región. Lo anterior se puede comprobar en el gráfico 5, en donde se aprecia que en una gran parte de la zona centro de la Ciénaga y hacia la desembocadura de los ríos Sevilla y Aracataca la magnitud de la variable es menor a la de los restantes sitios de muestreo. Esta figura revela claramente la influencia que tienen las entradas de agua (tal vez exceptuando

la entrada del río Fundación) en el comportamiento de esta variable. Los valores relativamente altos, respecto a los antes descritos, en la zona occidental pueden ser consecuencia del proceso de lavado de suelos hipersalinos que se da en época de lluvias en el complejo Pajarales (sistema con el que tiene frontera la Ciénaga) y que llegan al sistema a través de los Canales Grande y Clarín (Fig. 5). Los valores "altos" en la zona sur pueden ser de igual forma causados por la influencia del canal Grande y por circulación de las masas de agua dentro del sistema (contrario a las manecillas del reloj)

a). Salinidad

| | | |
|---|---|---|
| | Bajo | 127 |
| 2 | 13º | 5 |
| 4 | 14* | 14 |
| 11 | 14º | 6677889 |
| 18 | 15* | 0011224 |
| 30 | 15º | 566888899999 |
| 46 | 16* | 0011222333333334 |
| (13) | 16º | 5555666777899 |
| 55 | 17* | 011112233344 |
| 43 | 17º | 5556777788889 |
| 30 | 18* | 0000122233344 |
| 17 | 18º | 555556777888 |
| 5 | 19* | 01 |
| | Alto | 283,332,342 |

b). Seston

| | | |
|---|---|---|
| | Bajo | 10 |
| 2 | 1* | 1 |
| 3 | 1T | 2 |
| 5 | 1F | 55 |
| 15 | 1S | 6667777777 |
| 36 | 1o | 888888888999999999999 |
| 20) | 2* | 00000001111111111111 |
| 47 | 2T | 2222222222222333333 |
| 27 | 2F | 44555555 |
| 19 | 2S | 6666777777777 |
| 6 | 2o | 99 |
| 4 | 3* | 0011 |

c). Nitritos

| | | |
|---|---|---|
| 5 | 0 | 16888 |
| 26 | 1 | 000022555777779999999 |
| 43 | 2 | 11113333555558888 |
| (19) | 3 | 0000022244444666699 |
| 50 | 4 | 1133355555777777 |
| 34 | 5 | 02248 |
| 29 | 6 | 000033355599 |
| 17 | 7 | 1668 |
| 13 | 8 | 027 |
| 10 | 9 | 366 |
| 7 | 10 | 2 |
| 6 | 11 | 3 |
| | Alto | 120,126,131,141,161 |

d). Silicatos

| | | |
|---|---|---|
| | Bajo | 1,1,1,1,2 |
| 6 | 1S | 6 |
| 11 | 1o | 88999 |
| 22 | 2* | 00000111111 |
| 43 | 2T | 222222223333333333333 |
| (24 | 2F | 4444444445555555555555 |
| 45 | 2S | 666666666666677777 |
| 27 | 2o | 888888889999 |
| 15 | 3* | 000000111 |
| 6 | 3T | 2223 |
| 2 | 3F | 4 |
| | Alto | 35 |

e). Clorofila a.

| | | |
|---|---|---|
| | Bajo | 2,11,17,43,54,58,75 |
| 8 | 9 | 7 |
| 10 | 10 | 04 |
| 21 | 11 | 01233334579 |
| 41 | 12 | 00033444666677788899 |
| (18) | 13 | 001124444556788889 |
| 47 | 14 | 001111123334444567899 |
| 26 | 15 | 0011333444467788 |
| 10 | 16 | 0266 |
| 6 | 17 | 00138 |
| | Alto | 198 |

**Figura 3.** Diagramas de tallos y hojas de algunas variables medidas en la superficie de la columna de agua del estuario Ciénaga Grande de Santa Marta en Marzo de 1997.

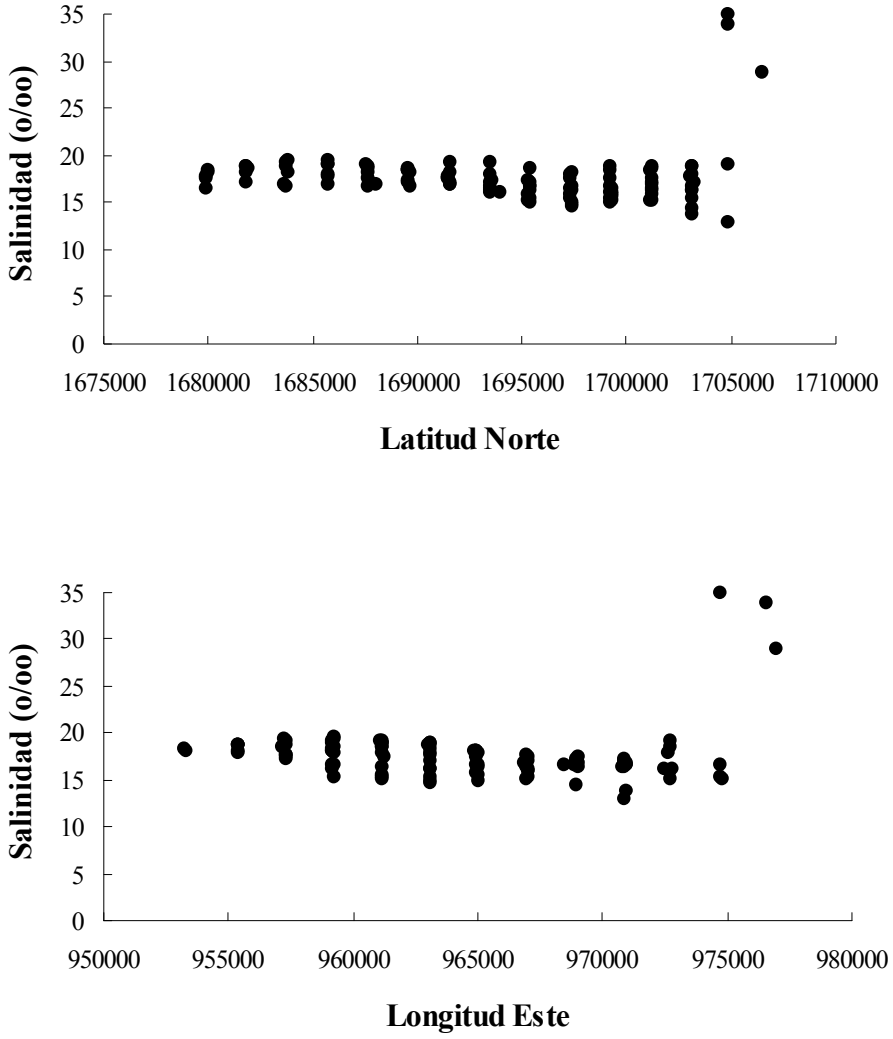



**Figura 4**. Gráficos de dispersión de valores de salinidad respecto a las coordenadas geográficas de medición. Datos tomados en la superficie de la columna de agua del estuario Ciénaga Grande de Santa Marta en Marzo de 1997.

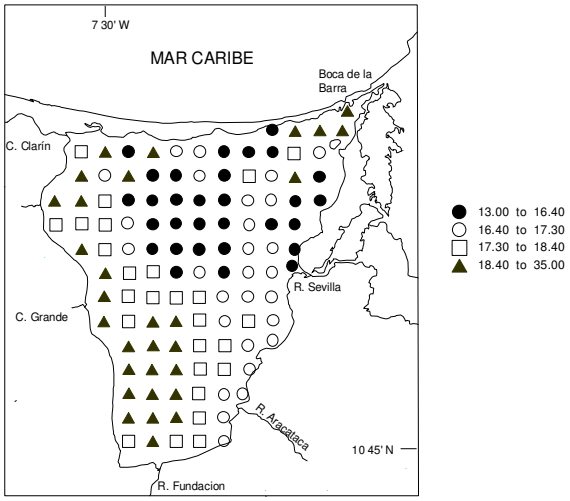



**Figura 5**. Clasificación de observaciones de la variable salinidad en intervalos (según cuartíles) y ubicación de estas dentro del área de estudio. Datos medidos en el estuario Ciénaga Grande de Santa Marta en marzo de 1997.

**Tabla 2**. Matriz de correlación calculada con base en información de algunas variables fisicoquímicas y biológicas medidas en el estuario Ciénaga Grande de Santa Marta en marzo de 1997. Los coeficientes que aparecen en negrita son significativos.

| | Salinidad | Seston | Nitritos | Silicatos | Clorofila a |
|---|---|---|---|---|---|
| Salinidad | 1 | -0.09 | -0.10 | -0.60 | -0.47 |
| Seston | | 1 | -0.33 | 0.06 | 0.45 |
| Nitritos | | | 1 | -0.08 | -0.23 |
| Silicatos | | | | 1 | 0.46 |
| Clorofila a | | | | | 1 |

Por último de la matriz de correlación (tabla 2) es posible afirmar que la abundancia fitoplanctónica, evaluda a través de la concentración de clorofila a, presenta correlación significativa con las variables fisicoquímicas medidas. Este patrón de correlación entre variables bióticas y abióticas en otros trabajos de menor intensidad muestral no ha podido ser detectado. En general los estudios realizados en la Ciénaga Grande de Santa Marta en los que se pretende determinar los patrones de asociación entre las variables biológicas y fisicoquímicas siempre conducen a que la salinidad es la variable de mayor influencia en el régimen de productividad del sistema. Sin embargo estos resultados en primera instancia pueden estar detectando otro tipo de asociaciones. En los capítulos subsiguientes, cuando se realicen los mapas de distribución espacial, se podrán tener más herramientas para discutir respecto a este tema.

# *Capítulo Dos*

## *Definiciones Básicas de Geoestadística*

### 2.1. Definición de Geoestadística

La geoestadística es una rama de la estadística que trata fenómenos espaciales (Journel & Huijbregts, 1978). Su interés primordial es la estimación, predicción y simulación de dichos fenómenos (Myers, 1987). Esta herramienta ofrece una manera de describir la continuidad espacial, que es un rasgo distintivo esencial de muchos fenómenos naturales, y proporciona adaptaciones de las técnicas clásicas de regresión para tomar ventajas de esta continuidad (Isaaks & Srivastava, 1989). Petitgas (1996), la define como una aplicación de la teoría de probabilidades a la estimación estadística de variables espaciales.

La modelación espacial es la adición más reciente a la literatura estadística. Geología, ciencias del suelo, agronomía, ingeniería forestal, astronomía, o cualquier disciplina que trabaja con datos colectados en diferentes locaciones espaciales necesita desarrollar modelos que indiquen cuando hay dependencia entre las medidas de los diferentes sitios. Usualmente dicha modelación concierne con la predicción espacial, pero hay otras áreas importantes como la simulación y el diseño muestral (Cressie, 1989).

Cuando el objetivo es hacer predicción, la geoestadística opera básicamente en dos etapas. La primera es el análisis estructural, en la cual se describe la correlación entre puntos en el espacio. En la segunda fase se hace predicción en sitios de la región no muestreados por medio de la técnica *kriging* (capítulo 4). Este es un proceso que calcula un promedio ponderado de las observaciones muestrales. Los pesos asignados a los valores muestrales son apropiadamente determinados por la estructura espacial de correlación establecida en la primera etapa y por la configuración de muestreo (Petitgas, 1996). Los fundamentos básicos de estas etapas son presentados a continuación.

### 2.2. Variable Regionalizada.

Una variable medida en el espacio de forma que presente una estructura de correlación, se dice que es una variable regionalizada. De manera más formal se puede definir como un proceso estocástico con dominio contenido en un espacio euclidiano d-dimensional $R^d$ , $\{Z(x) : x \in D \subset R^d\}$. Si d = 2, Z (x) puede asociarse a una variable medida en un punto *x* del plano (Díaz-Francés**,** 1993). En términos prácticos *Z(x)* puede verse como una medición de una variable aleatoria (p.ej. concentración de un contaminante) en un punto *x* de una región de estudio.

Recuérdese que un proceso estocástico es una colección de variables aleatorias indexadas; esto es, para cada *x* en el conjunto de índices D, $Z(x)$ es una variable aleatoria. En el caso de que las mediciones sean hechas en una superficie, entonces $Z(x)$ puede interpretarse como la variable aleatoria asociada a ese punto del plano (*x* representa las coordenadas, planas o geográficas, y Z la variable en cada una de ellas). Estas variables

aleatorias pueden representar la magnitud de una variable ambiental medida en un conjunto de coordenadas de la región de estudio.

### 2.3. Momentos de una Variable Regionalizada

Sea $\{Z(x) : x \in D \subset R^d\}$ el proceso estocástico que define la variable regionalizada. Para cualquier *n* puntos $x_1, x_2, ..., x_n$, el vector aleatorio $\vec{Z}(x) = [Z(x_1), Z(x_2), \cdots, Z(x_n)]^T$

está definido por su función de distribución conjunta

$$F[z_1, z_2, \cdots, z_n] = P[Z(x_1) \le z_1, Z(x_2) \le z_2, \cdots, Z(x_n) \le z_n]$$

Conocidas las densidades marginales univariadas y bivariadas se pueden establecer los siguientes valores esperados (momentos univariados y bivariados):

- $E(Z(x_i)) = m(x_i)$
- $V(Z(x_i)) = E[Z(x_i) - m(x_i)]^2 = \sigma_i^2$
- $C(Z(x_i), Z(x_j)) = E[Z(x_i) - m(x_i)][Z(x_j) - m(x_j)]$: Función de autocovarianza
- $\gamma(Z(x_i), Z(x_j)) = \frac{1}{2} E[Z(x_i) - Z(x_j)]^2$ : Función de semivarianza

### 2.4. Estacionariedad

La variable regionalizada es estacionaria si su función de distribución conjunta es invariante respecto a cualquier translación del vector h, o lo que es lo mismo, la función de distribución del vector aleatorio $\vec{Z}(x) = [Z(x_1), Z(x_2), \cdots, Z(x_n)]^T$ es idéntica a la del vector $\vec{Z}(x) = [Z(x_1 + h), Z(x_2 + h), \cdots, Z(x_n + h)]^T$ para cualquier *h*. La teoría geoestadística se basa en los momentos arriba descritos y la hipótesis de estacionariedad puede definirse en términos de estos:

#### 2.4.1 Estacionariedad de Segundo Orden

Sea $\{Z(x): x \in D \subset R^d\}$ una variable regionalizada definida en un dominio D contenido en $R^d$ (generalmente una variable medida en la superficie de una región) se dice que Z(x) es estacionario de *segundo orden* si cumple:

a. $E[Z(x)] = m, \quad k \in R, \quad \forall x \in D \subset R^d$ .
   El valor esperado de la variable aleatoria es finito y constante para todo punto en el dominio.

b. $COV[Z(x), Z(x+h)] = C(h) < \infty$
   Para toda pareja $\{Z(x), Z(x+h)\}$ la covarianza existe y es función única del vector de separación h.

En la figura 6 se muestra el grafico de una variable regionalizada estacionaria. Exceptuando fluctuaciones aleatorias, el valor promedio de la variable no muestra una tendencia definida en alguna dirección.

La existencia de la covarianza implica que la varianza existe, es finita y no depende de *h*, es decir $V(Z(x_i)) = C(0) = \sigma^2$ . Así mismo la estacionariedad de segundo orden implica la siguiente relación entre la función de semivarianza y la de autocovarianza:

$$\begin{aligned}\gamma(Z(x+h), Z(x)) = \gamma(h) &= \frac{1}{2}E[Z(x+h) - m - Z(x) + m]^2 \\ &= \frac{1}{2}\{E(Z(x+h) - m)^2 + E(Z(x) - m)^2 - 2E(Z(x+h) - m)(Z(x) - m)\} \\ &= \frac{1}{2}\sigma^2 + \frac{1}{2}\sigma^2 - E\{(Z(x+h) - m)(Z(x) - m)\} \\ &= \sigma^2 - C(h).\end{aligned}$$

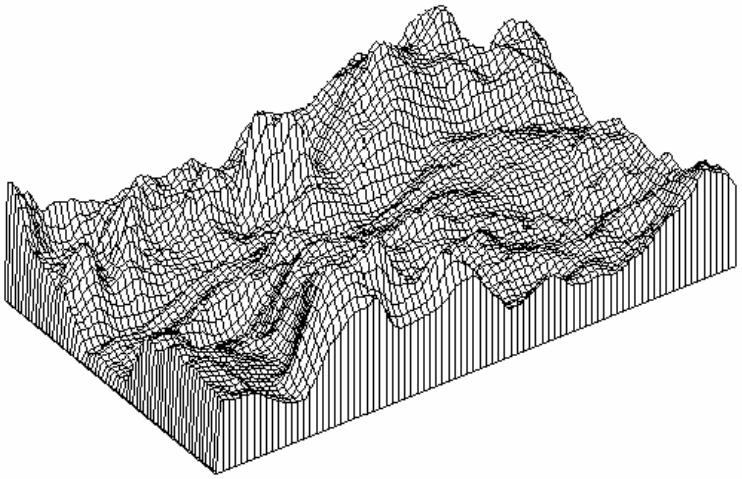


**Figura 6**. Representación de una superficie interpolada para una variable regionalizada estacionaria

### 2.4.2. Estacionariedad Débil o Intrínseca

Existen algunos fenómenos físicos reales en los que la varianza no es finita. En estos casos se trabaja sólo con la hipótesis que pide que los incrementos [Z(x+h) - Z(x)] sean estacionarios, esto es (Clark, 1979):

a. Z(x) tiene esperanza finita y constante para todo punto en el dominio. Lo que implica que la esperanza de los incrementos es cero.

$$E[Z(x+h) - Z(x)] = 0$$

b. Para cualquier vector h, la varianza del incremento está definida y es una función única de la distancia.

$$V\,[\,Z(x+h) - Z(x)] = E[Z(x+h) - Z(x)]^2 = 2\,\gamma\,(h)$$

Es claro que si una variable regionalizada es estacionaria fuerte entonces también será estacionaria débil. El concepto de estacionariedad es muy útil en la modelación de series temporales (Box & Jenkins, 1976). En este contexto es fácil la identificación, puesto que sólo hay una dirección de variación (el tiempo). En el campo espacial existen múltiples direcciones y por lo tanto se debe asumir que en todas el fenómeno es estacionario. Cuando la esperanza de la variable no es la misma en todas las direcciones o cuando la covarianza o correlación dependan del sentido en que se determinan, no habrá estacionariedad. Si la correlación entre los datos no depende de la dirección en la que esta se calcule se dice que el fenómeno es *isotrópico*, en caso contrario se hablará de *anisotropía.* En Isaaks y Srivastava (1989) se definen los posibles tipos de anisotropía y se proponen algunas soluciones. Cressie (1993) discute cual debe ser el tratamiento en caso de que la media no sea constante.

En casos prácticos resulta compleja la identificación de la estacionariedad. Suelen emplearse gráficos de dispersión de la variable respecto a las coordenadas, de medias móviles y de valores clasificados según puntos de referencia, con el propósito de identificar posibles tendencias de la variable en la región de estudio. L a isotropía es estudiada a través del cálculo de funciones de autocovarianza o de semivarianza muestrales (capítulo3) en varias direcciones. Si estas tienen formas considerablemente distintas puede no ser válido el supuesto de isotropía. Finalmente una variable regionalizada será no estacionaria si su esperanza matemática no es constante, esto es si $E[Z(x)] = m(x)$. En la figura 7 se representa una variable regionalizada en la que existe tendencia en el valor promedio de la variable, lo cual es claro indicador de no estacionariedad.

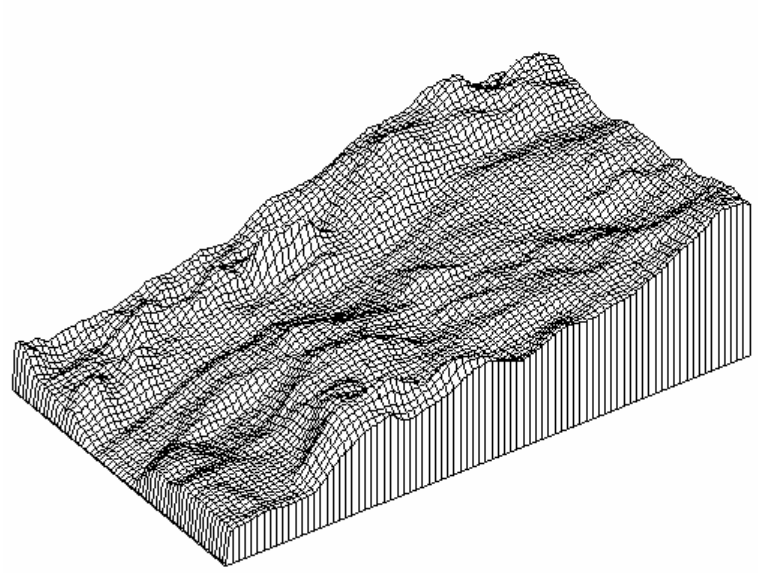


**Figura 7**. Representación de una superficie interpolada para una variable regionalizada no estacionaria

# *Capítulo Tres*

## *Correlación Espacial Muestral y Ajuste de Modelos*

### 3.1. Funciones de Correlación Espacial

La primera etapa en el desarrollo de un análisis geoestadístico es la determinación de la dependencia espacial entre los datos medidos de una variable. Esta fase es también conocida como análisis estructural. Para llevarla a cabo, con base en la información muestral, se usan tres funciones: El semivariograma, el covariograma y el correlograma. A continuación se hace una revisión de los conceptos asociados a cada una de ellas y se describen sus bondades y limitaciones.

### 3.1.1. Variograma y Semivariograma.

Cuando se definió la estacionariedad débil en el capítulo anterior se mencionó que se asumía que la varianza de los incrementos de la variable regionalizada era finita. A esta función denotada por $2\gamma(h)$ se le denomina variograma. Utilizando la definición teórica de la varianza en términos del valor esperado de una variable aleatoria, tenemos:

$$\begin{aligned} 2\gamma(h) &= V\big(Z(x+h)-Z(x)\big) \\ &= E\Big(\big(Z(x+h)-Z(x)\big)^2\Big) - \underbrace{\big(E\big(Z(x+h)-Z(x)\big)\big)^2}_{0} \\ &= E\Big(\big(Z(x+h)-Z(x)\big)^2\Big) \end{aligned}$$

La mitad del variograma $\gamma(h)$, se conoce como la función de semivarianza y caracteriza las propiedades de dependencia espacial del proceso. Dada una realización del fenómeno, la función de semivarianza es estimada, por el método de momentos, a través del semivariograma experimental, que se calcula mediante (Wackernagel, 1995):

$$\bar{\gamma}(h) = \frac{\sum \big(Z(x+h) - Z(x)\big)^2}{2n}$$

donde *Z(x)* es el valor de la variable en un sitio *x*, *Z(x+h)* es otro valor muestral separado del anterior por una distancia *h* y *n* es el número de parejas que se encuentran separadas por dicha distancia. La función de semivarianza se calcula para varias distancia *h*. En la práctica, debido a irregularidad en el muestreo y por ende en las distancias entre los sitios, se toman intervalos de distancia $\{[0, h], (h, 2h], (2h, 3h], \cdots\}$ y el semivariograma experimental corresponde a una distancia promedio entre parejas de sitios dentro de cada intervalo y no a una distancia *h* específica. Obviamente el número de parejas de puntos *n* dentro de los intervalos no es constante.

Para interpretar el semivariograma experimental se parte del criterio de que a menor distancia entre los sitios mayor similitud o correlación espacial entre las observaciones. Por ello en presencia de autocorrelación se espera que para valores de *h* pequeños el

semivariograma experimental tenga magnitudes menores a las que este toma cuando las distancias *h* se incrementan.

### 3.1.2. Covariograma y Correlograma.

La función de covarianza muestral entre parejas de observaciones que se encuentran a una distancia *h* se calcula, empleando la formula clásica de la covarianza muestral, por:

$$C(h) = COV\big(Z(x+h),\ Z(x)\big) = \frac{\sum_{i=1}^{n}(Z(x+h)-m)(Z(x)-m)}{n}$$

$$= \frac{\sum_{i=1}^{n}\big(Z(x+h)\cdot Z(x)\big)}{n} - m^2 = C(h)$$

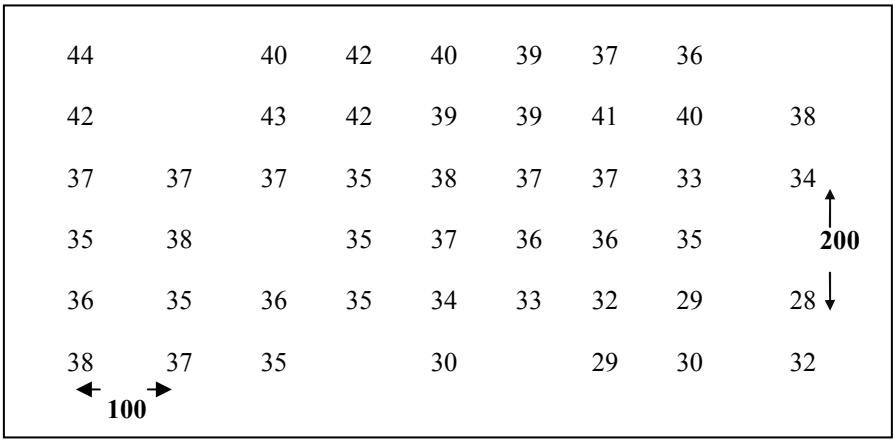
donde *m* representa el valor promedio en todo punto de la región de estudio y *n* es el número de parejas de puntos que se encuentran a una distancia *h*. En este caso es también válida la aclaración respecto a las distancias dadas en el último párrafo de la página anterior.

Asumiendo que el fenómeno es estacionario y estimando la varianza de la variable regionalizada a través de la varianza muestral, se tiene que el correlograma muestral está dado por:

$$r(h) = \frac{COV(Z(x+h), Z(x))}{S_{x+h}\cdot S_x} = \frac{C(h)}{S_x^2} = \frac{C(h)}{C(0)}$$

Bajo el supuesto de estacionariedad cualquiera de las tres funciones de dependencia espacial mencionadas, es decir semivariograma, covariograma o correlograma, puede ser usada en la determinación de la relación espacial entre los datos. Sin embargo como se puede observar en las fórmulas, la única que no requiere hacer estimación de parámetros es la función de semivarianza. Por esta razón, fundamentalmente, en la práctica se emplea el semivariograma y no las otras dos funciones.

A continuación se presenta un ejemplo ilustrativo del cálculo de la función de semivarianza experimental: Suponga que se tienen medidas sobre una variable hipotética cuyos valores están comprendidos entre 28 y 44 unidades y su configuración en una región de estudio es como se presenta en el esquema de la siguiente página. Como se indica en la representación, la distancia entre cada par de puntos contiguos es de 100 unidades. Luego si existe un punto faltante la distancia entre los dos valores ubicados a cada lado de éste será de 200 unidades. Veamos como calcular bajo esta situación el semivariograma experimental. Por simplicidad se calcularán sólo los semivariogramas en sentido (izquierda-derecha) e (inferior-superior), debido a que para obtener un semivariograma experimental en el que sólo se tenga en cuenta la distancia y no la orientación (*semivariograma omnidireccional*), se requeriría calcular la distancia euclidiana para un número considerablemente alto de parejas de puntos.




En primer lugar en sentido izquierda-derecha se encuentran todas las parejas de puntos que están a una distancia de 100 unidades y se calcula el semivariograma como:

$\bar{\gamma}(100) = (38 - 37)^2 + (37 - 35)^2 + (29 - 30)^2 + ... + (37 - 36)^2 /2* 36 = 1.458$

análogamente para la distancia de 200 unidades

$\bar{\gamma}(200) = (40 - 44)^2 + (40 - 40)^2 + (42 - 39)^2 + ... + (29 - 32)^2 /2* 36 = 3.303$

Similarmente se procede para otras distancias y para el sentido inferior-superior. Los valores calculados de el semivariograma se muestran en la siguiente tabla.

**Tabla 3**. Valores de la función de semivarianza experimental en dos direcciones para el conjunto de datos hipotéticos.

| Distancia | Semivarianza Sentido Izquierda-Derecha | Semivarianza Sentido Inferior-Superior |
|---|---|---|
| 100 | 1.45 | 5.34 |
| 200 | 3.30 | 9.87 |
| 300 | 4.31 | 18.88 |
| 400 | 6.69 | 27.53 |

Al graficar los valores de la función de semivarianza experimental dados en la tabla anterior (Fig. 8) se observa que en sentido inferior-superior el semivariograma es mayor que en sentido izquierda-derecha, luego la conclusión más relevante para este conjunto de datos es que la estructura de correlación espacial no sólo depende de la distancia entre los sitios, sino de su orientación. En otras palabras el fenómeno podría ser *anisotrópico.*

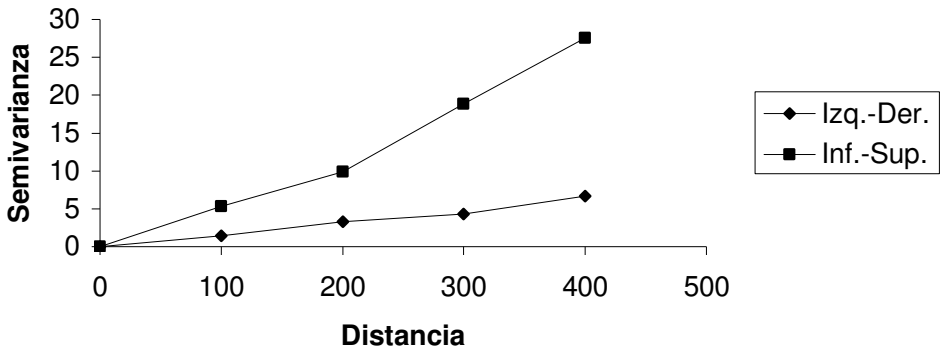



**Figura 8**. Función de semivarianza experimental en dos direcciones para un conjunto de datos hipotéticos.

## 3.2. Modelos Teóricos de Semivarianza.

Como se verá a partir del capítulo cuatro la solución del problema de predicción espacial kriging requiere del conocimiento de la estructura de autocorrelación para cualquier posible distancia entre sitios dentro del área de estudio. En la presentación del semivariograma experimental dada anteriormente se indicó que este es calculado sólo para algunas distancias promedios particulares. Por ello se hace necesario el ajuste de modelos que generalicen lo observado en el semivariograma experimental a cualquier distancia. Existen diversos modelos teóricos de semivarianza que pueden ajustarse al semivariograma experimental. En Samper y Carrera (1990) se presenta una discusión respecto a las características y condiciones que éstos deben cumplir. En general dichos modelos pueden dividirse en no acotados (lineal, logarítmico, potencial) y acotados (esférico, exponencial, gaussiano) (Warrick *et al*., 1986). Los del segundo grupo garantizan que la covarianza de los incrementos es finita, por lo cual son ampliamente usados cuando hay evidencia de que presentan buen ajuste. Todos estos modelos tienen tres parámetros comunes (Fig. 9) que son descritos a continuación:

- **Efecto Pepita**

Se denota por $C_0$ y representa una discontinuidad puntual del semivariograma en el origen (Fig. 9). Puede ser debido a errores de medición en la variable o a la escala de la misma. En algunas ocasiones puede ser indicativo de que parte de la estructura espacial se concentra a distancias inferiores a las observadas.

- **Meseta**

Es la cota superior del semivariograma. También puede definirse como el limite del semivariograma cuando la distancia *h* tiende a infinito. La meseta puede ser o no finita. Los semivariogramas que tienen meseta finita cumplen con la hipótesis de estacionariedad fuerte; mientras que cuando ocurre lo contrario, el semivariograma define un fenómeno natural que cumple sólo con la hipótesis intrínseca. La meseta se denota por $C_1$ o por ($C_0$ + $C_1$) cuando la pepita es diferente de cero. Si se interpreta la pepita como un error en las mediciones, esto explica porque se sugiere que en un modelo que explique bien la realidad, la pepita no debe representar mas del 50% de la meseta. Si el ruido espacial en las

mediciones explica en mayor proporción la variabilidad que la correlación del fenómeno, las predicciones que se obtengan pueden ser muy imprecisas. En la figura 9 se representa este parámetro para el caso de uno de los modelos acotados.

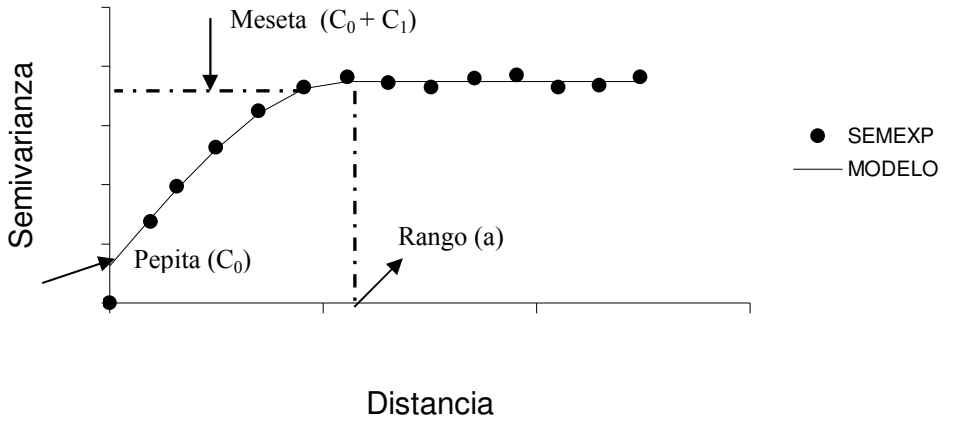



**Figura 9**. Comportamiento típico de un semivariograma acotado con una representación de los parámetros básicos. SEMEXP corresponde al semivariograma experimental y MODELO al ajuste de un modelo teórico.

- **Rango**

En términos prácticos corresponde a la distancia a partir de la cual dos observaciones son independientes. El rango se interpreta como la zona de influencia. Existen algunos modelos de semivariograma en los que no existe una distancia finita para la cual dos observaciones sean independientes; por ello se llama rango efectivo a la distancia para la cual el semivariograma alcanza el 95% de la meseta. Entre más pequeño sea el rango, más cerca se esta del modelo de independencia espacial. El rango no siempre aparece de manera explícita en la fórmula del semivariograma. En el caso del modelo esférico (3.2.1), el rango coincide con el parámetro **a**, que se utilizará en las ecuaciones más adelante. Sin embargo, en el modelo exponencial (3.2.2), el rango efectivo es a/3 y en el modelo gaussiano (3.2.3) es a/√3.

### 3.2.1. Modelo Esférico

Tiene un crecimiento rápido cerca al origen (Fig. 10), pero los incrementos marginales van decreciendo para distancias grandes, hasta que para distancias superiores al rango los incrementos son nulos. Su expresión matemática es la siguiente:

$$\gamma(h)=\begin{cases} C_0 + C_1\left(\frac{3}{2}\left(\frac{h}{a}\right)-\frac{1}{2}\left(\frac{h}{a}\right)^3\right) & h \le a \\ C_0 + C_1 & h > a \end{cases}$$

En donde $C_1$ representa la meseta, $a$ el rango y $h$ la distancia.

### 3.2.2. Modelo Exponencial

Este modelo se aplica cuando la dependencia espacial tiene un crecimiento exponencial respecto a la distancia entre las observaciones. El valor del rango es igual a la distancia para la cual el semivariograma toma un valor igual al 95% de la meseta (Fig. 10). Este modelo es ampliamente usado. Su expresión matemática es la siguiente:

$$\gamma(h) = C_0 + C_1\left(1 - \exp\left(\frac{-3h}{a}\right)\right)$$

### 3.2.3. Modelo Gaussiano

Al igual que en el modelo exponencial, la dependencia espacial se desvanece solo en una distancia que tiende a infinito. El principal distintivo de este modelo es su forma parabólica cerca al origen (Fig.10). Su expresión matemática es:

$$\gamma(h) = C_0 + C_1\left(1 - \exp\left(\frac{-h^2}{a^2}\right)\right)$$

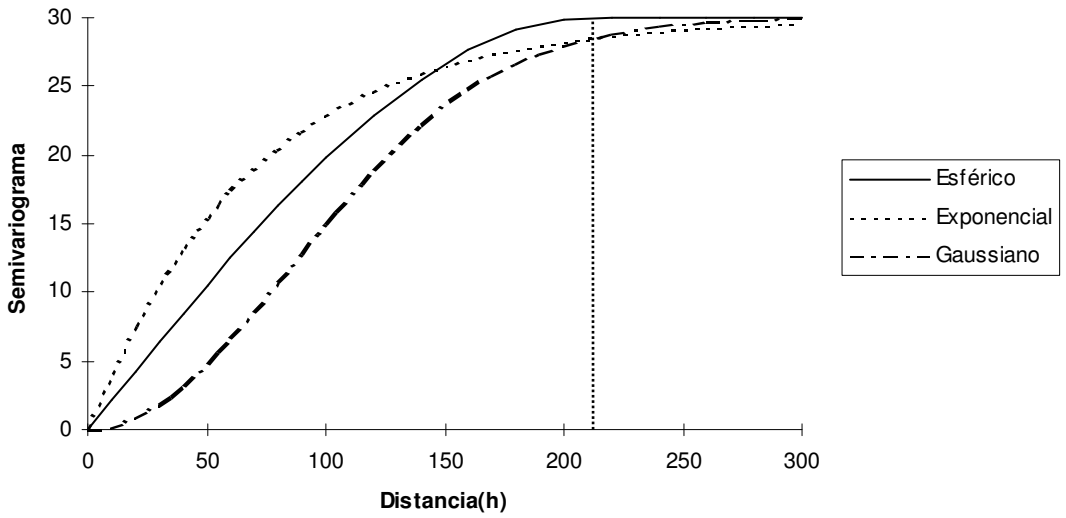



**Figura 10**. Comparación de los modelos exponencial, esférico y Gaussiano. La línea punteada vertical representa el rango en el caso del modelo esférico y el rango efectivo en el de los modelos exponencial y gaussiano. Este tiene un valor de 210, respecto a una escala simulada entre 0 y 300. El valor de la meseta es 30 y el de la pepita 0. El 95% de la meseta es igual a 28.5.

### 3.2.4. Modelo Monómicos.

Corresponden a los modelos que no alcanzan la meseta (Fig. 11). Su uso puede ser delicado debido a que en algunos casos indican la presencia de no estacionariedad en alguna dirección.

Su fórmula matemática es la siguiente:

$$\gamma(h) = kh^{\theta} \qquad 0 < \theta < 2$$

Obviamente cuando el parámetro θ es igual a uno el modelo es lineal y k representa la pendiente de la ecuación de regresión con intercepto cero. Gráficamente se pueden representar así:

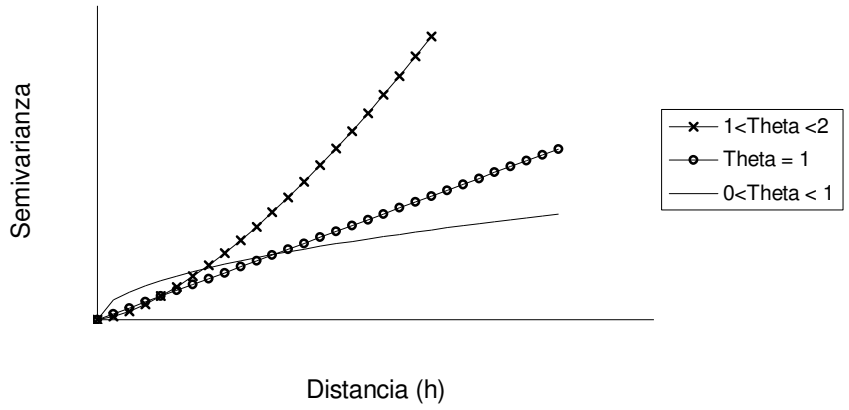



**Figura 11**. Comportamiento típico de los modelos de semivarianza monómicos.

### 3.2.5. Modelo de Independencia (Pepita Puro).

Es indicativo de carencia de correlación espacial entre las observaciones de una variable (Fig. 12). Es común sumar este modelo a otro modelo teórico de semivarianza, para obtener lo que se conoce como semivariograma anidado. Lo anterior se sustenta en una propiedad de los semivariogramas que dice que cualquier combinación lineal de semivariogramas con coeficientes positivos es un semivariograma. Su expresión matemática es:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 & h > 0 \end{cases}, \text{ donde } C_0 > 0$$

Su representación gráfica es la siguiente:

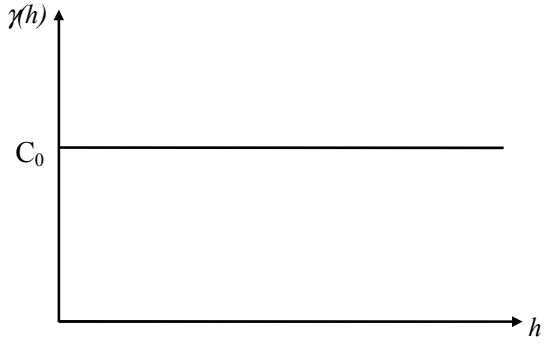



**Figura 12**. Modelo de semivarianza teórico para variables sin correlación espacial.

La estimación de los parámetros de los modelos teóricos descritos puede ser llevada a cabo, entre otros métodos, por máxima verosimilitud (Cressie, 1993) o regresión no lineal (Gotway, 1991). Algunos paquetes de computo geoestadísticos como GS+ (Gamma Design Software, 1999) traen incorporados procedimientos iterativos como el de Gauss-Newton para llevar a cabo la estimación. Otros como GeoEAS (Englund y Sparks, 1988) sólo permiten el ajuste a sentimiento por el método de ensayo y error.

Como se mencionó en la sección 4.2. cuando la autocorrelación no es igual en todas las direcciones entonces se dice que hay anisotropía. Esa puede ser geométrica o zonal. La primera se presenta cuando los semivariogramas calculados en varias direcciones tienen igual meseta pero varían en el rango. En el segundo caso todos los semivariogramas direccionales tiene igual rango pero diferente meseta. Algunas transformaciones apropiadas para solucionar la anistropía y hacer válida la construcción de un semivariograma omnidireccional se pueden encontrar en Isaaks y Srivastava (1989), Samper y Carrera (1990) y Cressie (1993) .

### *3.3. Aplicación: Estimación de Modelos de Semivarianza para algunas variables fisicoquímicas y biológicas medidas en el estuario Ciénaga Grande de Santa Marta.*

En esta sección se hace una interpretación práctica de resultados encontrados al hacer estimación de modelos teóricos de semivarianza para un conjunto de variables medidas en el estuario Ciénaga Grande de Santa Marta (IGAC, 1973). Se consideran para el análisis datos tomados en dos niveles de la columna de agua (superficie y fondo), de las variables salinidad, oxígeno disuelto (mg/l), sólidos en suspensión (mg/l), nitritos (µmol/l) y clorofila "a"(µg/l), Además se estudian valores de profundidad (m) y transparencia (m). Para cada variable se obtuvieron 115 observaciones tomadas a lo largo de todo el sistema mediante muestreo sistemático de cuadrículas (cuadrículas de 4 $km^2$)(Fig. 1). La ubicación en cada punto de muestreo fue realizada mediante un geoposicionador GPS 100 SRVY II (Garmin, 1993). Los semivariogramas experimentales y los correspondientes modelos estimados se hallaron por medio del software GS+ (Gamma Design Software, 1999).

- **Resultados y Discusión.**

Los semivariogramas experimentales encontrados (Figs. 13 y 14) indican que las variables presentan estructuras de dependencia espacial, puesto que en ningún caso la semivarianza es constante en función de la distancia. Los rangos encontrados en los modelos teóricos ajustados a los semivariogramas (tabla 4), superan los 11 km y en algunos casos éste parámetro alcanza los 25 km, lo cual resulta relativamente alto, teniendo en cuenta que la distancia entre los extremos sur y norte del sistema (la más amplia) no supera los 30 km. Lo anterior es un indicador de fuerte dependencia espacial para el caso considerado. Esto es sin duda conveniente puesto que desde un punto de vista teórico es conocido que un alto valor en el rango permite obtener curvas de predicción más suavizadas reduciendo las magnitudes en varianzas de predicción (Díaz-Francés, 1993) . Es importante resaltar respecto a los otros dos parámetros, que en ningún caso el valor de la pepita supera el 50% del valor de la meseta (tabla 4), lo cual, según Díaz-Francés (1993), es recomendable para que el modelo de correlación espacial describa bien la realidad. Si el ruido espacial en las mediciones explica en mayor proporción la variabilidad que la correlación del fenómeno, las predicciones pueden ser muy imprecisas.

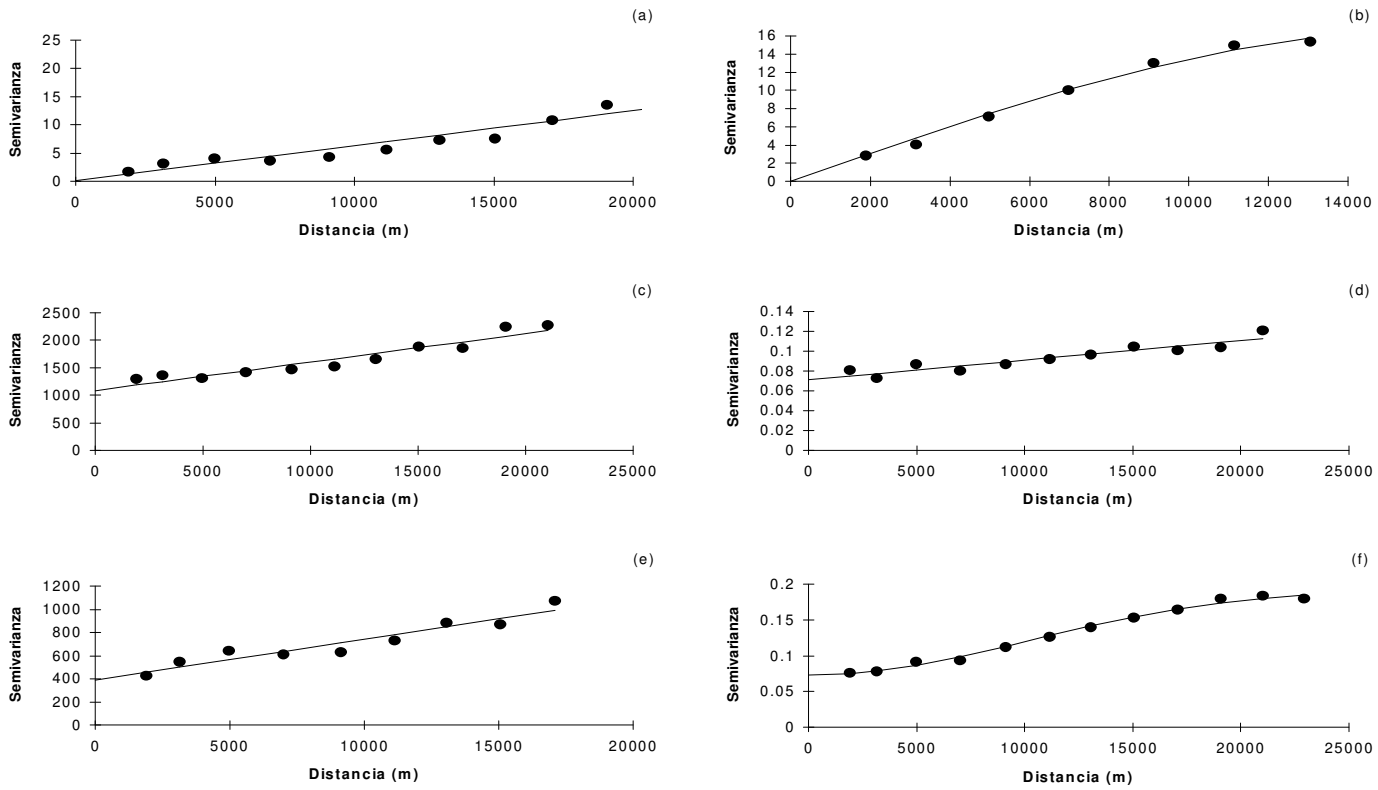



Figura 13. semivariogramas experimentales (calculados con los datos muestrales) y ajustes de modelos teóricos para las variables medidas en la superficie de la columna de agua de la Ciénaga Grande de Santa Marta en marzo de 1997. a) salinidad; b)oxígeno; c) sólidos en suspensión; d) nitritos; e) clorofila a; f)profundidad.

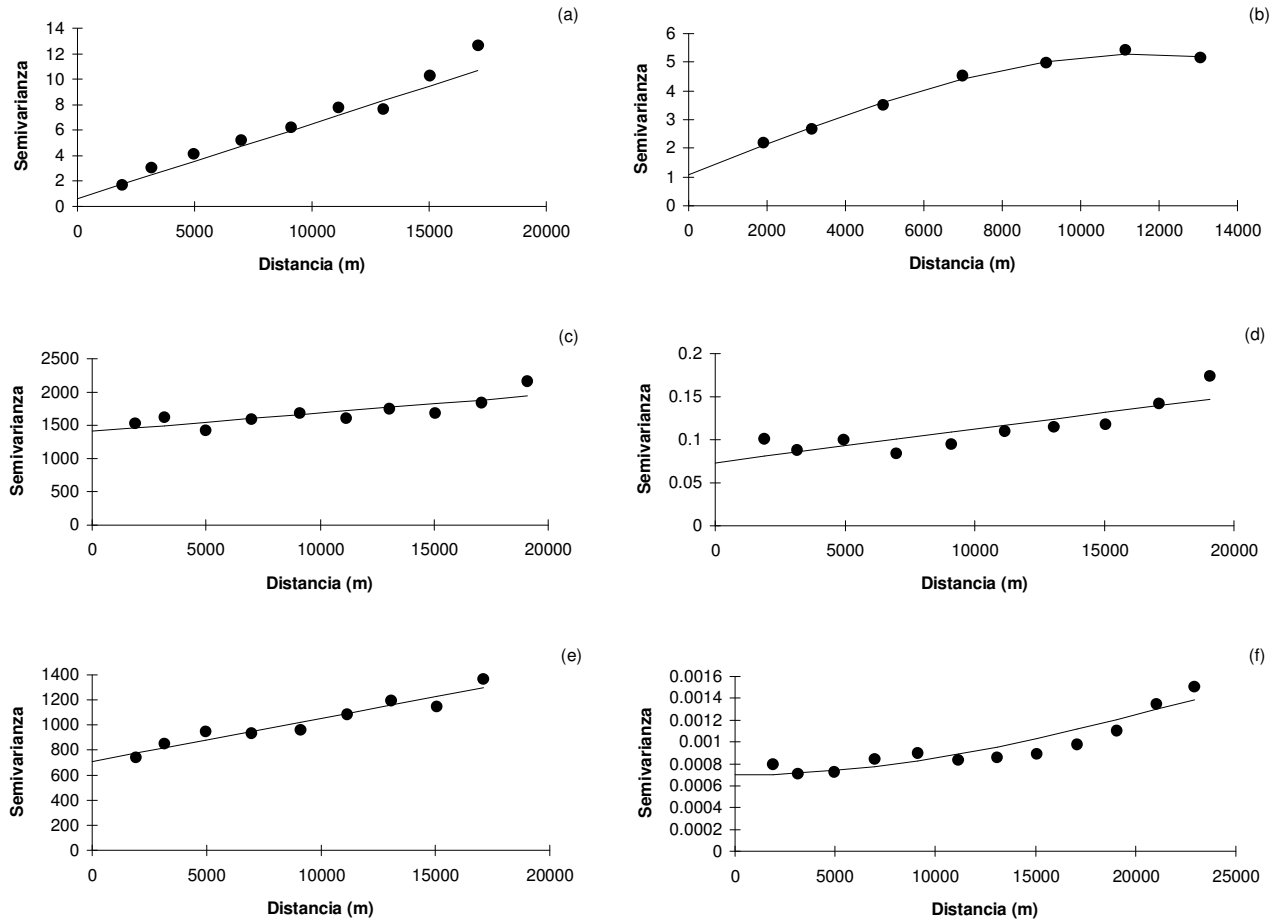



Figura 14. Semivariogramas experimentales (calculados con los datos muestrales) y ajustes de modelos teóricos para las variables medidas en el fondo de la columna de agua de la Ciénaga Grande de Santa Marta en marzo de 1997. a) salinidad; b)oxígeno; c) sólidos en suspensión; d) nitritos; e) clorofila a; f) transparencia.

**Tabla 4**. Modelos teóricos ajustados a semivariogramas experimentales de variables fisicoquímicas y biológicas medidas en dos niveles de la columna de agua de la Ciénaga Grande de Santa Marta, durante una jornada de muestreo realizada en marzo de 1997.

| Variable | Nivel | Modelo | Pepita | Meseta | Rango (m) | $r^2$ |
|---|---|---|---|---|---|---|
| Salinidad | Superficie | Lineal | 0.179 | 12.309 | 20000 | 0.89 |
| | Fondo | Lineal | 0.627 | 11.752 | 20000 | 0.99 |
| Oxígeno | Superficie | Gaussiano | 1.830 | 14.320 | 12940 | 0.99 |
| | Fondo | Esférico | 1.080 | 4.211 | 11650 | 0.99 |
| Sólidos en suspensión | Superficie | Lineal | 1087 | 1138 | 22000 | 0.90 |
| | Fondo | Lineal | 1408 | 557 | 20000 | 0.67 |
| Nitritos | Superficie | Lineal | 0.071 | 0.043 | 22000 | 0.87 |
| | Fondo | Lineal | 0.073 | 0.077 | 20000 | 0.70 |
| Clorofila a | Superficie | Lineal | 389.2 | 623.2 | 18000 | 0.91 |
| | Fondo | Lineal | 710 | 616.4 | 18000 | 0.91 |
| Profundidad | | Gaussiano | 0.073 | 0.121 | 24850 | 0.99 |
| Transparencia | | Gaussiano | 0.0069 | 0.0019 | 25000 | 0.85 |

Se puede afirmar que las variables oxígeno disuelto, profundidad y transparencia cumplen con la hipótesis de estacionariedad fuerte, dado que sus modelos son acotados (Biau *et al*., 1997; Samper y Carrera, 1990). De otro lado salinidad, sólidos en suspensión, nitritos y clorofila "a", sólo cumplen la hipótesis intrínseca (estacionariedad débil) puesto que sus modelos son lineales (Evangelos y Flatman, 1988; Samper y Carrera, 1990).

Debido a que los resultados arriba descritos respecto a los semivariogramas experimentales y al ajuste de modelos teóricos, confirman la hipótesis de autocorrelación espacial en las características medidas en el estuario de estudio, es posible afirmar que los métodos geoestadísticos pueden ser una herramienta de gran utilidad en la modelación e interpretación de fenómenos observados en este tipo de ecosistemas. Cuando se utilicen métodos estadísticos tradicionales (regresión, análisis de varianza, técnicas multivariadas, muestreo) para el análisis de este tipo de información, debe involucrarse en los correspondientes modelos la estructura de correlación espacial implícita en los datos.

# *Capitulo Cuatro*

## *Predicción Espacial*

### 4.1. Predicción Espacial Optima.

De la teoría de la decisión se conoce que si $Z_0$ es una cantidad aleatoria y $Z^*_0$ es su predictor [2], entonces $L(Z_0; Z_0^*)$ representa la pérdida en que se incurre cuando se predice $Z_0$ con $Z_0^*$ y el mejor predictor será el que minimice $E\{L(Z_0; Z_0^*)/Z\}$ con $Z = \{Z_1, Z_2, \cdots, Z_n\}'$, es decir el predictor óptimo es el que minimice la *esperanza condicional* de la función de pérdida. Si $L(Z_0; Z_0^*) = [Z_0 - Z_0^*]^2 \Rightarrow Z_0^* = E(Z_0 / Z)$. La expresión anterior indica que para encontrar el predictor óptimo se requiere conocer la distribución conjunta de la *n+1* variables aleatorias.

### 4.2. Definición de Kriging.

La palabra kriging[3] (expresión anglosajona) procede del nombre del geólogo sudafricano D. G. Krige, cuyos trabajos en la predicción de reservas de oro, realizados en la década del cincuenta, suelen considerarse como pioneros en los métodos de interpolación espacial. Kriging encierra un conjunto de métodos de predicción espacial que se fundamentan en la minimización del error cuadrático medio de predicción. En la tabla 5 se mencionan los tipos de kriging y algunas de sus propiedades. En la secciones 4.3 y 4.4, se hace una presentación detallada de ellos.

**Tabla 5**. Tipos de predictores kriging y sus propiedades.

| TIPO DE PREDICTOR | NOMBRE | PROPIEDADES |
|---|---|---|
| LINEAL | • Simple<br>• Ordinario<br>• Universal | • Son óptimos si hay normalidad multivariada.<br>• Independiente de la distribución son los mejores predictores linealmente insesgados. |
| NO LINEAL | • Indicador<br>• Probabilístico<br>• Log Normal, Trans-Gaussiano<br>• Disyuntivo | • Son predictores óptimos. |

[2] La palabra *estimación* es utilizada exclusivamente para inferir sobre parámetros fijos pero desconocidos; *predicción* es reservada para inferencia sobre cantidades aleatorias.
[3] Algunos textos indican que en español la palabra adecuada sería krigeado.

Los métodos kriging se aplican con frecuencia con el propósito de predicción, sin embargo estas metodologías tienen diversas aplicaciones, dentro de las cuales se destacan la simulación y el diseño de redes óptimas de muestreo (capítulo 5).

### 4.3. Kriging Ordinario

Suponga que se hacen mediciones de la variable de interés Z en los puntos $x_i$, $i= 1, 2,..., n$, de la región de estudio, es decir se tienen realizaciones de las variables $Z(x_1), \ldots, Z(x_n)$, y se desea predecir $Z(x_o)$, en el punto $x_0$ donde no hubo medición. En esta circunstancia, el método kriging ordinario propone que el valor de la variable puede predecirse como una combinación lineal de las $n$ variables aleatorias así:

$$Z^*(x_0) = \lambda_1 Z(x_1) + \lambda_2 Z(x_2) + \lambda_3 Z(x_3) + \lambda_4 Z(x_4) + \lambda_5 Z(x_5) + \ldots + \lambda_n Z(x_n)$$

$$= \sum_{i=1}^{n} \lambda_i Z(x_i)$$

en donde los $\lambda_i$ representan los pesos o ponderaciones de los valores originales. Dichos pesos se calculan en función de la distancia entre los puntos muestreados y el punto donde se va a hacer la correspondiente predicción. La suma de los pesos debe ser igual a uno para que la esperanza del predictor sea igual a la esperanza de la variable. Esto último se conoce como el requisito de insesgamiento.

Estadísticamente la propiedad de insesgamiento se expresa a través de:

$$E\left(Z^*(x_0)\right) = E\left(Z(x_0)\right)$$

Asumiendo que el proceso es estacionario de media $m$ (desconocida) y utilizando las propiedades del valor esperado, se demuestra que la suma de las ponderaciones debe ser igual a uno:

$$E\left(\sum_{i=1}^{n} \lambda_i Z(x_i)\right) = m$$

$$\sum_{i=1}^{n} \lambda_i E\left(Z(x_i)\right) = m$$

$$\sum_{i=1}^{n} \lambda_i m = m$$

$$m\sum_{i=1}^{n} \lambda_i = m \Rightarrow \sum_{i=1}^{n} \lambda_i = 1$$

Se dice que $Z^*(x_0)$ es el mejor predictor, lineal en este caso, porque los pesos se obtienen de tal manera que minimicen la varianza del error de predicción, es decir que minimicen la expresión:

$$V\left(Z^*(x_0) - Z(x_0)\right)$$

Esta última es la característica distintiva de los métodos kriging, ya que existen otros métodos de interpolación como el de distancias inversas o el poligonal, que no garantizan varianza mínima de predicción (Samper y Carrera, 1990). La estimación de los pesos se obtiene minimizando $V[Z^*(x_0)-Z(x_0)]$ sujeto a $\sum_{i=1}^{n}\lambda_i = 1$.

Se tiene que $V[Z^*(x_0) - Z(x_0)] = V[Z^*(x_0)] - 2COV[Z^*(x_0),Z(x_0)] + V[Z(x_0)]$

Desagregando las componentes de la ecuación anterior se obtiene los siguiente:

$$V[Z^*(x_0)]=V\left[\sum_{i=1}^{n}\lambda_i Z(x_i)\right]=\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j COV[Z(x_i),Z(x_j)]$$

En adelante se usará la siguiente notación: $COV[Z(x_i),Z(x_j)] = C_{ij}$ y $V[Z(x_0)]=\sigma^2$

De lo anterior $COV[Z^*(x_0),Z(x_0)]=COV\left[\sum_{i=1}^{n}\lambda_i Z(x_i),Z(x_0)\right]$

$$=\sum_{i=1}^{n}\lambda_i COV[Z(x_i),Z(x_0)] = \sum_{i=1}^{n}\lambda_i C_{i0}$$

Entonces reemplazando, se tiene que:

$$V[Z^*(x_0)-Z(x_0)]=\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C_{ij}-2\sum_{i=1}^{n}\lambda_i C_{i0}+\sigma^2 \quad (0)$$

Luego se debe minimizar la función anterior sujeta a la restricción $\sum_{i=1}^{n}\lambda_i=1$. Este problema de minimización con restricciones se resuelve mediante el método de multiplicadores de Lagrange.

$$\sigma_k^2=\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C_{ij}-2\sum_{i=1}^{n}\lambda_i C_{i0}+\sigma^2+\underbrace{2\mu}_{\substack{Multiplicador\\ de\,Lagrange}}\underbrace{\left(\sum_{i=1}^{n}\lambda_i-1\right)}_{0}$$

Siguiendo el procedimiento acostumbrado para obtener valores extremos de una función, se deriva e iguala a cero, en este caso con respecto a $\lambda_i$ y $\mu$:

$$\frac{\partial(\sigma_k^2)}{\partial\lambda_1}=\frac{\partial\left[(\lambda_1^2 C_{11}+2\lambda_1\sum_{j=2}^{n}\lambda_j C_{1j}+\sum_{i=2}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C_{ij})-2\sum_{i=1}^{n}\lambda_i C_{i0}+\sigma^2+2\mu\left(\sum_{i=1}^{n}\lambda_i-1\right)\right]}{\partial\lambda_1}$$

$$=\underbrace{\left(2\lambda_1 C_{11}+2\sum_{j=2}^{n}\lambda_j C_{1j}\right)}_{\Downarrow}-2C_{10}+2\mu$$

$$=2\sum_{j=1}^{n}\lambda_j C_{1j}-2C_{10}+2\mu=0\Rightarrow\sum_{j=1}^{n}\lambda_j C_{1j}+\mu=C_{10}\ (1)$$

De manera análoga se determinan las derivadas con respecto a $\lambda_2$ , ..., $\lambda_n$ :

$$\frac{\partial(\sigma_k^2)}{\partial\lambda_2} = 2\sum_{j=1}^{n}\lambda_j C_{2j} - 2C_{20} + 2\mu = 0 \Rightarrow \sum_{j=1}^{n}\lambda_j C_{2j} + \mu = C_{20} \quad (2)$$

$$\vdots$$

$$\frac{\partial(\sigma_k^2)}{\partial\lambda_n} = 2\sum_{j=1}^{n}\lambda_j C_{nj} - 2C_{n0} + 2\mu = 0 \Rightarrow \sum_{j=1}^{n}\lambda_j C_{nj} + \mu = C_{n0} \quad (3)$$

por último derivamos con respecto a $\mu$:

$$\frac{\partial(\sigma_k^2)}{\partial\mu} = 2\sum_{i=1}^{n}\lambda_i - 2 = 0 \Rightarrow \sum_{i=1}^{n}\lambda_1 = 1 \quad (4)$$

De (1), (2), (3), (4) resulta un sistema de *(n + 1)* ecuaciones con *(n + 1)* incógnitas, que matricialmente puede ser escrito como:

$$\begin{pmatrix} C_{11} & . & . & . & C_{1n} & 1 \\ . & . & & & . & . \\ . & & . & & . & . \\ . & & & . & . & . \\ C_{n1} & . & . & . & C_{nn} & 1 \\ 1 & . & . & . & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ . \\ . \\ . \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} C_{10} \\ . \\ . \\ . \\ C_{n0} \\ 1 \end{pmatrix}$$

$$C_{ij} \quad \bullet \quad \lambda \quad = \quad C_{i0}$$

por lo cual los pesos que minimizan el error de predicción se determinan mediante la función de covariograma a través de:

$$\lambda \quad = \quad C_{ij}^{-1} \quad \bullet C_{i0}.$$

Encontrando los pesos se calcula la predicción en el punto $x_o$. De forma análoga se procede para cada punto donde se quiera hacer predicción.

- **Varianza de Predicción del Kriging Ordinario**

Multiplicando (1), (2) y (3) por $\lambda_i$ se obtiene:

$$\lambda_i\left(\sum_{j=1}^{n}\lambda_i C_{ij} + \mu\right) = \lambda_i C_{i0} \qquad \forall i, i = 1, 2, \cdots, n.$$

Sumando las *n* ecuaciones

$$\sum_{i=1}^{n}\lambda_i\sum_{j=1}^{n}\lambda_i C_{ij} + \sum_{i=1}^{n}\lambda_i\mu = \sum_{i=1}^{n}\lambda_i C_{io}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_i C_{ij} = \sum_{i=1}^{n}\lambda_i C_{io} - \sum_{i=1}^{n}\lambda_i\mu$$

Sustituyendo la expresión anterior en (0)

$$\sigma_k^2 = \sigma^2 + \sum_{i=1}^{n} \lambda_i C_{i0} - \sum_{i=1}^{n} \lambda_i \mu - 2\sum_{i=1}^{n} \lambda_i C_{i0}$$

$$\sigma_k^2 = \sigma^2 - \sum_{i=1}^{n} \lambda_i C_{i0} - \mu \quad (5)$$

- **Estimación de Ponderaciones por medio de la Función de Semivarianza**

Los pesos $\lambda$ pueden ser estimados a través de la función de semivarianza, para lo cual se requiere conocer la relación entre las funciones de covariograma y de semivarianza. Antes de esto conveniente tener en cuenta la siguiente notación:
$\sigma^2 = V(Z(x))$, $\gamma_{ij} = \gamma(h)$, donde h es la distancia entre los puntos i y j y análogamente
$C_{ij} = C(h)$.
La relación entre las dos funciones en cuestión es la siguiente:

$$\gamma_{ij} = \frac{1}{2} E\left[(Z(x_j) - Z(x_i))^2\right]$$

$$= \frac{1}{2} E\left[(Z(x_j))^2 - 2(Z(x_j)Z(x_i) + (Z(x_i))^2\right]$$

$$= \frac{1}{2} E\left[(Z(x_j)^2\right] - E\left[Z(x_j)Z(x_i)\right] + \frac{1}{2} E\left[(Z(x_i))^2\right]$$

$$= \frac{1}{2}\left[E\,(Z(x_j))^2 - k^2\right] + \frac{1}{2}\left[E\,(Z(x_j))^2 - k^2\right] - \left[E\left(Z(x_j)Z(x_i)\right) - k^2\right]$$

$$= \frac{1}{2}\left[V(Z(x))\right] + \frac{1}{2}\left[V(Z(x))\right] - COV\left[Z(x_j)Z(x_i)\right]$$

$$= V\left[Z(x)\right] - COV\left[Z(x_j)Z(x_i)\right]$$

$$= \sigma^2 - C_{ij} \Rightarrow C_{ij} = \sigma^2 - \gamma_{ij} \qquad (6)$$

Reemplazando (6) en (1), (2) y (3) se determinan los pesos óptimos λ en términos de la función de semivarianza:

$$\frac{\partial(\sigma_k^2)}{\partial\lambda_1} = \sum_{j=1}^{n} \lambda_j C_{1j} + \mu - C_{10} = \sum_{j=1}^{n} \lambda_j (\sigma^2 - \gamma_{1j}) + \mu - (\sigma^2 - \gamma_{10})$$

$$= \sigma^2 \sum_{j=1}^{1} \lambda_j - \sum_{j=1}^{n} \lambda_j \gamma_{1j} + \mu - \sigma^2 + \gamma_{10}$$

$$= \sigma^2 - \sum_{j=1}^{n} \lambda_j \gamma_{1j} + \mu - \sigma^2 + \gamma_{10} \Rightarrow \sum_{j=1}^{n} \lambda_j \gamma_{1j} - \mu = \gamma_{10}$$

Similarmente,

$$\frac{\partial(\sigma_k^2)}{\partial\lambda_2} = \sum_{j=1}^{n} \lambda_j \gamma_{2j} - \mu = \gamma_{20}$$

$$\vdots$$

$$\frac{\partial(\sigma_k^2)}{\partial\lambda_n} = \sum_{j=1}^{n} \lambda_j \gamma_{nj} - \mu = \gamma_{n0}$$

El sistema de ecuaciones se completa con (4). De acuerdo con lo anterior los pesos se obtienen en términos del semivariograma a través del sistema de ecuaciones:

$$\begin{pmatrix} \gamma_{11} & \cdot & \cdot & \cdot & \gamma_{1n} & 1 \\ \cdot & \cdot & & & \cdot & \cdot \\ \cdot & & \cdot & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot & \cdot \\ \gamma_{n1} & \cdot & \cdot & \cdot & \gamma_{nn} & 1 \\ 1 & \cdot & \cdot & \cdot & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \cdot \\ \cdot \\ \cdot \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \gamma_{10} \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{n0} \\ 1 \end{pmatrix}$$

Para establecer la expresión de la correspondiente varianza del error de predicción en términos de la función de semivarianza se reemplaza (6) en (5), de donde:

$$\sigma_k^2 = \sigma^2 - \left[\sum_{i=1}^{n} \lambda_i \left(\sigma^2 - \gamma_{ij}\right)\right] + \mu$$

$$\sigma_k^2 = \sigma^2 - \sigma^2 \sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \lambda_i \gamma_{ij} + \mu$$

$$\sigma_k^2 = \sum_{i=1}^{n} \lambda_i \gamma_{io} + \mu$$

Los pesos de kriging ordinario también pueden ser estimados mediante el uso del correlograma aplicando la siguiente relación: $\rho_{ij} = C_{ij} / \sigma^2$ , caso en el que la correspondiente varianza de predicción estaría dada por (Isaaks y Srivastava, 1989):

$$\sigma_k^2 = \sigma^2 \left(1 - \sum \lambda_i \gamma_{io} + \mu\right)$$

- **Validación del kriging.**

Existen diferentes métodos para evaluar la bondad de ajuste del modelo de semivariograma elegido con respecto a los datos muestrales y por ende de las predicciones hechas con kriging. El más empleado es el de validación cruzada, que consiste en excluir la observación de uno de los *n* puntos muestrales y con los *n-1* valores restantes y el modelo

de semivariograma escogido, predecir vía kriging el valor de la variable en estudio en la ubicación del punto que se excluyó. Se piensa que si el modelo de semivarianza elegido describe bien la estructura de autocorrelación espacial, entonces la diferencia entre el valor observado y el valor predicho debe ser pequeña. Este procedimiento se realiza en forma secuencial con cada uno de los puntos muestrales y así se obtiene un conjunto de *n* "errores de predicción". Lo usual es calcular medidas que involucren a estos errores de predicción para diferentes modelos de semivarianza y seleccionar aquel que optimice algún criterio como por ejemplo el del mínimo error cuadrático medio (MECM). Este procedimiento es similar a la conocida técnica de remuestreo Jacknife (Efron, 1982) empleada en diversos contextos estadísticos para calcular varianzas de estimación, entre otros aspectos. Una forma descriptiva de hacer la validación cruzada es mediante un gráfico de dispersión de los valores observados contra los valores predichos. En la medida en que la nube de puntos se ajuste más a una línea recta que pase por el origen, mejor será el modelo de semivariograma utilizado para realizar el kriging.

- **Representación de las predicciones**

Una vez se ha hecho la predicción en un conjunto de puntos diferentes de los muestrales vía kriging, se debe elaborar un mapa que dé una representación global del comportamiento de la variable de interés en la zona estudiada. Los más empleados son los mapas de contornos, los mapas de residuos y los gráficos tridimensionales. En el caso de los mapas de contornos, en primer lugar se divide el área de estudio en un enmallado y se hace la predicción en cada uno de los nodos de éste mismo. Posteriormente se unen los valores predichos con igual valor, generando así las líneas de contorno (isolíneas de distribución). Este gráfico permite identificar la magnitud de la variable en toda el área de estudio. Es conveniente acompañar el mapa de interpolaciones de la variable con los correspondientes mapas de isolíneas de los errores y de las varianzas de predicción (posiblemente estimados a través de métodos matemáticos), con el propósito de identificar zonas de mayor incertidumbre respecto a las predicciones.

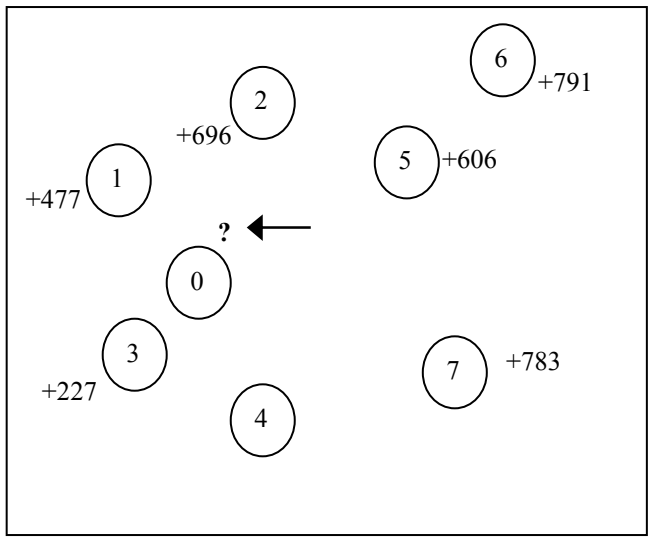
- **Intervalos de Confianza.**

Asumiendo que los errores de predicción siguen una distribución normal estándar y que son independientes, un intervalo de confianza del 100(1-$\alpha$)%, 0<$\alpha$ <1, para *Z(x)* es:

$\left[ z^*(x) - z_{1-\alpha/2}\sigma_k \, , z^*(x) + z_{1-\alpha/2}\sigma_k \right]$ con $z^*(x)$ el valor calculado de la predicción y $z_{1-\alpha/2}$ el percentil de una normal estándar.

- **Ilustración**

Suponga que se tiene una configuración de datos como la que se presenta en el esquema de abajo. Con base en siete datos observados (valores al lado del signo + por fuera de los círculos numerados de 1 a 7) se quiere predecir un valor de la variable en el punto donde se encuentra el signo de interrogación, por fuera del circulo con el número cero.




La matriz de distancia euclidianas entre los sitios es la siguiente:

| Sitio | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.00 | 4.47 | 3.61 | 8.06 | 4.49 | 6.71 | 8.94 | 13.45 |
| 1 | | 0.00 | 2.24 | 10.44 | 13.04 | 10.05 | 12.17 | 17.80 |
| 2 | | | 0.00 | 11.05 | 13.00 | 8.00 | 10.05 | 16.97 |
| 3 | | | | 0.00 | 4.12 | 13.04 | 15.00 | 11.05 |
| 4 | | | | | 0.00 | 12.37 | 13.93 | 7.00 |
| 5 | | | | | | 0.00 | 2.24 | 12.65 |
| 6 | | | | | | | 0.00 | 13.15 |
| 7 | | | | | | | | 0.00 |

Asumiendo que la estructura de correlación espacial entre los datos es estimada por un modelo exponencial $\gamma(h) = 10\,(1 - \exp(-3h/10))$ (pepita cero, meseta 10 y rango 10) o en términos de la función de autocovarianza por $C(h) = 10\,(\exp(-3h/10))$, se encuentran las siguientes matrices que permiten encontrar los pesos para la predicción:

$$C_{ij} = \begin{pmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} & C_{17} & 1 \\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} & C_{27} & 1 \\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36} & C_{37} & 1 \\ C_{41} & C_{42} & C_{43} & C_{44} & C_{45} & C_{46} & C_{47} & 1 \\ C_{51} & C_{52} & C_{53} & C_{54} & C_{55} & C_{56} & C_{57} & 1 \\ C_{61} & C_{62} & C_{63} & C_{64} & C_{65} & C_{66} & C_{67} & 1 \\ C_{71} & C_{72} & C_{73} & C_{74} & C_{75} & C_{76} & C_{77} & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 10 & 5.11 & 0.44 & 0.20 & 0.49 & 0.26 & 0.05 & 1 \\ 5.11 & 10 & 0.36 & 0.20 & 0.91 & 0.49 & 0.06 & 1 \\ 0.44 & 0.36 & 10 & 2.90 & 0.20 & 0.11 & 0.36 & 1 \\ 0.20 & 0.20 & 2.90 & 10 & 0.24 & 0.15 & 1.22 & 1 \\ 0.49 & 0.91 & 0.20 & 0.24 & 10 & 5.11 & 0.22 & 1 \\ 0.26 & 0.49 & 0.11 & 0.15 & 5.11 & 10 & 0.19 & 1 \\ 0.05 & 0.06 & 0.36 & 1.22 & 0.22 & 0.19 & 10 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

$$C_{ij}^{-1} = \begin{pmatrix} 0.127 & -0.077 & -0.013 & -0.009 & -0.008 & -0.009 & -0.012 & 0.136 \\ -0.077 & 0.129 & -0.010 & -0.008 & -0.015 & -0.008 & -0.011 & 0.121 \\ -0.013 & -0.010 & 0.098 & -0.042 & -0.010 & -0.010 & -0.014 & 0.156 \\ 0.009 & -0.008 & -0.042 & 0.102 & -0.009 & -0.009 & -0.024 & 0.139 \\ -0.008 & -0.015 & -0.010 & -0.009 & 0.130 & -0.077 & -0.012 & 0.118 \\ -0.009 & -0.008 & -0.010 & -0.009 & -0.077 & 0.126 & -0.013 & 0.141 \\ -0.012 & -0.011 & -0.014 & -0.024 & -0.012 & -0.013 & 0.085 & 0.188 \\ 0.136 & 0.121 & 0.156 & 0.139 & 0.118 & 0.141 & 0.188 & -2.180 \end{pmatrix}$$

$$C_{io} = \begin{pmatrix} C_{10} \\ C_{10} \\ C_{10} \\ C_{10} \\ C_{10} \\ C_{10} \\ C_{10} \\ 1 \end{pmatrix} = \begin{pmatrix} 2.61 \\ 3.39 \\ 0.89 \\ 0.58 \\ 1.34 \\ 0.68 \\ 0.18 \\ 1 \end{pmatrix}, \text{ de donde } \lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_{41} \\ \lambda_5 \\ \lambda_6 \\ \lambda_7 \\ \mu \end{pmatrix} = C_{ij}^{-1} C_{io} = \begin{pmatrix} 0.173 \\ 0.318 \\ 0.129 \\ 0.086 \\ 0.151 \\ 0.057 \\ 0.086 \\ 0.907 \end{pmatrix}.$$

con base en el vector estimado de parámetros se encuentra que

$$Z_0^* = \sum_{i=1}^{7} \lambda_i Z_i = (0.173)(477) + (0.318)(696) + ... + (0.086)(0.18) = 592.$$

$$\text{con } \sigma_k^2 = \sigma^2 - \sum_{i=1}^{7} \lambda_i C_{io} - \mu = 10 - [(0.173)(2.61) + \cdots + (0.086)(0.18)] - 0.907$$

### 4.4. Otros Métodos Kriging

A continuación se mencionan algunos aspectos generales de otros métodos de predicción espacial. Un estudio riguroso de ellos puede hacerse en Cressie (1993), Deutsch y Journel (1998) y Samper y Carrera (1990).

#### 4.4.1. Kriging Simple

Suponga que hay una variable regionalizada estacionaria con media ($m$) y covarianza conocidas. De manera análoga a como se define en modelos lineales (por ejemplo en diseño de experimentos) el modelo establecido en este caso es igual a la media más un error aleatorio con media cero. La diferencia es que en este caso los errores no son independientes.

Sea *Z(x)* la variable de interés medida en el sitio $x$.

$E[Z(x)] = m$

$Z(x) = m + \varepsilon(x)$, con $E[\varepsilon(x)] = 0$.

El predictor de la variable de interés en un sitio $x_0$ donde no se tiene información se define como:

$Z^*(x_0) = m + \varepsilon^*(x_0)$,

con $\varepsilon^*(x_0)$ que corresponde a la predicción del error aleatorio en el sitio $x_0$. Despejando de la ecuación anterior $\varepsilon^*(x_0) = Z^*(x_0) - m$.

El predictor del error aleatorio se define por:

$$\varepsilon^*(x_0) = \sum_{i=1}^{n} \lambda_i \varepsilon(x_i) = \sum_{i=1}^{n} \lambda_i (Z(x_i) - m).$$

de donde el predictor de la variable de estudio es:

$$Z^*(x_0) = m + \left[ \sum_{i=1}^{n} \lambda_i (Z(x_i) - m) \right] = m + \sum_{i=1}^{n} \lambda_i \varepsilon(x_i)$$

El predictor es insesgado si:

$E(Z^*(x_0)) = E(Z(x_0)) = m$. Luego el predictor será insesgado cuando $E(\varepsilon^*(x_0)) = 0$.

$E(\varepsilon^*(x_0)) = \sum_{i=1}^{n} \lambda_i \varepsilon(x_i) = \sum_{i=1}^{n} \lambda_i (0) = 0$. Por consiguiente, a diferencia del kriging ordinario, en este caso no existen restricciones para las ponderaciones tendientes al cumplimiento de la condición de insesgamiento.

La estimación de los pesos del método kriging ordinario se obtiene de tal forma que se minimice $V\left(\varepsilon^*(x_0)-\varepsilon(x_0)\right)$.

$$
\begin{aligned}
V\left(\varepsilon^*(x_0)-\varepsilon(x_0)\right) &= E\left(\varepsilon^*(x_0)-\varepsilon(x_0)\right)^2 \\
&= E\left(\left(\sum_{i=1}^{n}\lambda_i\varepsilon(x_i)\right)-\varepsilon(x_0)\right)^2 \\
&= E\left[\left(\sum_{i=1}^{n}\lambda_i\varepsilon(x_i)\right)\left(\sum_{j=1}^{n}\lambda_j\varepsilon(x_j)\right)-2E\left(\left(\sum_{i=1}^{n}\lambda_i\varepsilon(x_i)\right)(\varepsilon(x_0))\right)+E(\varepsilon(x_0))^2\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j E\left(\varepsilon(x_i)\varepsilon(x_j)\right)-2\sum_{i=1}^{n}\lambda_i E(\varepsilon(x_i)\varepsilon(x_0))+E(\varepsilon(x_0))^2
\end{aligned}
$$

usando:

i. $E[\varepsilon(x_0)]=0$

ii. $E\left(\varepsilon(x_i)\varepsilon(x_j)\right)=COV\left(\varepsilon(x_i),\varepsilon(x_j)\right)=C_{ij}$

iii. $E(\varepsilon(x_o))^2=\sigma^2$

$$
V\left(\varepsilon^*(x_0)-\varepsilon(x_0)\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C_{ij}-2\sum_{i=1}^{n}\lambda_i C_{i0}+\sigma^2 \qquad (7)
$$

derivando respecto a $\lambda_1$ se tiene:

$$
\frac{\partial V\left(\varepsilon^*(x_0)-\varepsilon(x_0)\right)}{\partial\lambda_1}=\frac{\partial}{\partial\lambda_1}\left(\lambda_1^2C_{11}+2\lambda_1\sum_{j=2}^{n}\lambda_jC_{1j}+\sum_{i=2}^{n}\sum_{j=2}^{n}\lambda_i\lambda_jC_{ij}-2\lambda_1C_{10}-2\sum_{i=2}^{n}\lambda_iC_{i0}+\sigma^2\right)
$$

$$
\begin{aligned}
&= 2\lambda_1C_{11}+2\sum_{j=2}^{n}\lambda_jC_{1j}-2C_{10} \\
&= 2\sum_{i=1}^{n}\lambda_iC_{1i}-2C_{10}
\end{aligned}
$$

igualando a cero

$$
\sum_{i=1}^{n}\lambda_1C_{1i}=C_{10}\ .
$$

En general para cualquier *i, i= 1, 2, ..., n,* se obtiene:

$$
\frac{\partial}{\partial\lambda_i}=\sum_{j=1}^{n}\lambda_jC_{ij}=C_{i0} \qquad (8)
$$

Con las *n* ecuaciones resultantes se construye el siguiente sistema de ecuaciones:

$$\begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} C_{10} \\ C_{20} \\ \vdots \\ C_{n0} \end{pmatrix}$$

- **Varianza de Predicción Kriging Simple.**

Se tiene de (7) que:

$$V\left(\varepsilon^*(x_0) - \varepsilon(x_0)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C_{ij} - 2\sum_{i=1}^{n}\lambda_i C_{i0} + \sigma^2$$

$$\sigma_k^2 = \sum_{i=1}^{n}\lambda_i\sum_{j=1}^{n}\lambda_j C_{ij} - 2\sum_{i=1}^{n}\lambda_i C_{i0} + \sigma^2$$

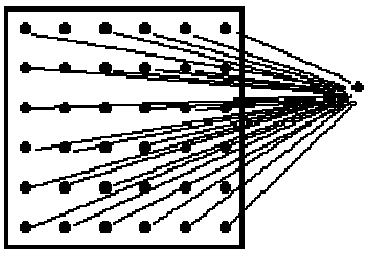
reemplazando (8) en (7)

$$\sigma_k^2 = \sum_{i=1}^{n}\lambda_i C_{i0} - 2\sum_{i=1}^{n}\lambda_i C_{i0} + \sigma^2$$

$$\sigma_k^2 = \sigma^2 - \sum_{i=1}^{n}\lambda_i C_{i0}$$

### 4.4.2. Kriging en Bloques.

En los dos métodos kriging hasta ahora descritos el objetivo ha estado centrado en la predicción puntual. A menudo, sin embargo, se requiere estimar un bloque, o más precisamente, estimar el valor promedio de la variable dentro de un área local.



El valor promedio dentro del bloque es estimado por :

$$\overline{Z}(A) = \sum_{i=1}^{n}\lambda_i Z(x_i)$$

Del sistema de ecuaciones para el kriging ordinario se tiene:

$$\begin{pmatrix} C_{11} & . & . & . & C_{1n} & 1 \\ . & . & & & . & . \\ . & & . & & . & . \\ . & & & . & . & . \\ C_{n1} & . & . & . & C_{nn} & 1 \\ 1 & . & . & . & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ . \\ . \\ . \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} C_{10} \\ . \\ . \\ . \\ C_{n0} \\ 1 \end{pmatrix}$$

Consecuentemente el vector del lado derecho de la igualdad en el sistema de arriba debe modificarse para incluir las covarianzas respecto al bloque. La covarianza de un punto al bloque corresponde a la covarianza promedio entre el punto muestreado *i* y todos los puntos dentro del bloque (en la práctica un enmallado regular de puntos dentro del bloque es usado como se muestra en la figura de la página anterior). El sistema de ecuaciones del kriging en bloques está dado por:

$$\begin{pmatrix} C_{11} & . & . & . & C_{1n} & 1 \\ . & . & & & . & . \\ . & & . & & . & . \\ . & & & . & . & . \\ C_{n1} & . & . & . & C_{nn} & 1 \\ 1 & . & . & . & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ . \\ . \\ . \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} \overline{C}_{1A} \\ . \\ . \\ . \\ \overline{C}_{nA} \\ 1 \end{pmatrix}$$

donde el vector de covarianzas al lado derecho de la igualdad en el sistema anterior es contiene las covarianzas entre las variables $Z(x_1), Z(x_2), \cdots, Z(x_n)$ y el bloque A donde se quiere hacer la estimación.

$$\overline{C}_{iA} = \frac{1}{|A|} \sum_{j\,/\,j \in A} C_{iA} \ .$$

La varianza del error de predicción del kriging en bloques está dada por:

$\sigma^2_{kB} = \overline{C}_{AA} - \left( \sum_{i=1}^{n} \lambda_i \overline{C}_{iA} + \mu \right)$, con $\overline{C}_{AA} = \frac{1}{|A|^2} \sum_{i\,/\,i \in A} \sum_{j\,/\,j \in A} C_{ij}$ igual a la covarianza entre pares de puntos dentro del bloque.

Isaaks y Srivastava (1989) muestran a través de ejemplos que el kriging en bloques coincide con el promedio de predicciones hechas por kriging ordinario sobre cada uno de los puntos del enmallado dentro del bloque. Así mismo indican que en la práctica es suficiente con un enmallado cuadrado (6x6) para obtener estimaciones estables en los bloques.

### 4.4.3. Kriging Universal.

En los supuestos hechos hasta ahora respecto a los métodos kriging se ha asumido que la variable regionalizada es estacionaria (al menos se cumple con la hipótesis intrínseca). En muchos casos, la variable no satisface estas condiciones y se caracteriza por exhibir una tendencia. Por ejemplo en hidrología los niveles piezométricos[4] de una acuífero pueden mostrar una pendiente global en la dirección del flujo (Samper y Carrera, 1990). Para tratar este tipo de variables es frecuente descomponer la variable *Z(x)* como la suma de la tendencia, tratada como una función determinística, más una componente estocástica estacionaria de media cero. Asuma que:

$$Z(x) = m(x) + \varepsilon(x)$$

con $E(\varepsilon(x)) = 0$, $V(\varepsilon(x)) = \sigma^2$ y por consiguiente $E(Z(x)) = m(x)$.

La tendencia puede expresarse mediante:

$$m(x) = \sum_{l=1}^{P} a_l f_l(x)$$

donde las funciones $f_l(x)$ son conocidas y p es el número de términos empleados para ajustar *m(x)*.

El predictor kriging universal se define como:

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)$$

este será insesgado si:

$$E(Z^*(x_0)) = m(x_0)$$

$$E\left(\sum_{i=1}^{n} \lambda_i Z(x_i)\right) = m(x_0)$$

$$\left(\sum_{i=1}^{n} \lambda_i m(x_i)\right) = m(x_0)$$

$$\sum_{i=1}^{n} \lambda_i \left(\sum_{l=1}^{P} a_l f_l(x_i)\right) = \sum_{l=1}^{p} a_l f_l(x_0)$$

$$\sum_{l=1}^{p} a_l \left(\sum_{i=1}^{n} \lambda_i f_l(x_i)\right) = \sum_{l=1}^{p} a_l f_l(x_0) \quad \Rightarrow \quad \sum_{i=1}^{n} \lambda_i f_l(x_i) = \sum_{l=1}^{p} f_l(x_0)$$

[4] Piezómetro: Instrumento utilizado para medir coeficientes de compresibilidad de sólidos, líquidos y gases

La obtención de los pesos en el kriging universal, análogo a los otros métodos kriging, se hace de tal forma que la varianza del error de predicción sea mínima.

$$
\begin{aligned}
V\left(Z^*(x_0)-Z(x_0)\right) &= E\left(Z^*(x_0)-Z(x_0)\right)^2 \\
&= E\left(\left(\sum_{i=1}^{n}\lambda_i\left(m(x_i)\ -\varepsilon(x_i)\right)\right)-\left(m(x_0)-\varepsilon(x_0)\right)\right)^2 \\
&= E\left[\left(\sum_{i=1}^{n}\lambda_i m(x_i)-m(x_0)\right)+\left(\sum_{i=1}^{n}\lambda_i\varepsilon(x_i)-\varepsilon(x_0)\right)\right]^2 \\
&= E\left[\left(\sum_{i=1}^{n}\lambda_i\varepsilon(x_i)-\varepsilon(x_0)\right)^2\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j E\left(\varepsilon(x_i)\varepsilon(x_j)\right)-2\sum_{i=1}^{n}\lambda_i E\left(\varepsilon(x_i)\varepsilon(x_0)\right)+E\left(\varepsilon(x_0)\right)^2
\end{aligned}
$$

Usando

$C_{ij} = COV\left(\varepsilon(x_i),\varepsilon(x_j)\right)$

$\sigma^2 = E\left(\varepsilon(x_0)\right)^2$

se tiene

$$V\left(Z^*(x_0)-Z(x_0)\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C_{ij}\ -2\sum_{i=1}^{n}\lambda_i C_{io}+\sigma^2\ .$$

Luego incluyendo la restricción dada por la condición de insesgamiento, se debe minimizar:

$$\sigma_{ku}^2=\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C_{ij}\ -2\sum_{i=1}^{n}\lambda_i C_{io}+\sigma^2+\sum_{l=1}^{p}\mu_l\left[\sum_{i=1}^{n}\lambda_i f_l(x_i)-f_l(x_0)\ \right]$$

o en términos de la función de semivarianza

$$\sigma_{ku}^2=-\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma_{ij}+2\sum_{i=1}^{n}\lambda_i\gamma_{io}+\sum_{l=1}^{p}\mu_l\left[\sum_{i=1}^{n}\lambda_i f_l(x_i)-f_l(x_0)\ \right]$$

derivando la expresión anterior respecto a $\lambda_1, \lambda_2, \cdots, \lambda_n,\ \mu_1, \mu_2, \cdots, \mu_p$ e igualando a cero las correspondientes derivadas se obtienen las siguientes ecuaciones:

$$\sum_{j=1}^{n}\lambda_j\gamma_{ij}+\sum_{l=1}^{P}\mu_l f_l(x_i)=\gamma_{i0}\quad i = 1, 2, ..., n$$

$$\sum_{j=1}^{n}\lambda_j f_l\left(x_j\right)=f_l(x_0)\qquad\qquad j = 1, 2, ..., p$$

en términos matriciales

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & f_{11} & \cdots & f_{p1} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & f_{12} & \cdots & f_{p2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{n2} & f_{1n} & \cdots & f_{pn} \\ f_{11} & f_{12} & \cdots & f_{1n} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ f_{p1} & f_{p2} & \cdots & f_{pn} & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \gamma_{10} \\ \gamma_{20} \\ \vdots \\ \gamma_{n0} \\ f_{10} \\ \vdots \\ f_{p0} \end{pmatrix}$$

donde $f_{lj} = f_l(x_j)$ es la l-ésima función en el punto j-ésimo.

La varianza de predicción del kriging universal está dada por (Samper y Carrera, 1990):

$$\sigma^2_{ku} = \sum_{i=1}^{n} \lambda_i \gamma_{i0} + \sum_{l=1}^{p} \mu_l f_l(x_0).$$

Nótese que si $p = 1$ y $f_l(x) = 1$, el sistema de ecuaciones del kriging universal y la varianza de predicción coinciden con las del kriging ordinario. En este orden de ideas puede decirse que el kriging ordinario es un caso particular del kriging universal.

### 4.4.4. Kriging Residual.

La técnica kriging residuales empleada bajo las mismas circunstancias del kriging universal, es decir en aquellos casos en que la variable regionalizada no es estacionaria debido a la presencia de tendencia espacial en el valor promedio de la variable. La hipótesis central del kriging residual consiste en suponer conocida la tendencia *m(x)*. A partir de ella se calculan los residuos con base en los cuales se aplica kriging ordinario. La estimación de la tendencia es generalmente llevada a cabo por medio de mínimos cuadrados. La predicción en un sitio no muestreado es igual a la tendencia estimada más la predicción del error, es decir:

$$Z^*(x_0) = \hat{m}(x_0) + e^*(x_0)$$

$$e^*(x_0) = \sum_{i=1}^{n} \lambda_i e(x_i)$$

los pesos o ponderaciones son estimados por kriging ordinario como se muestra en la sección 4.2. La varianza de predicción de la variable de interés coincide con la varianza de predicción de los errores. En la figura 15 se muestra un esquema con el procedimiento kriging residual en el caso de una tendencia lineal.

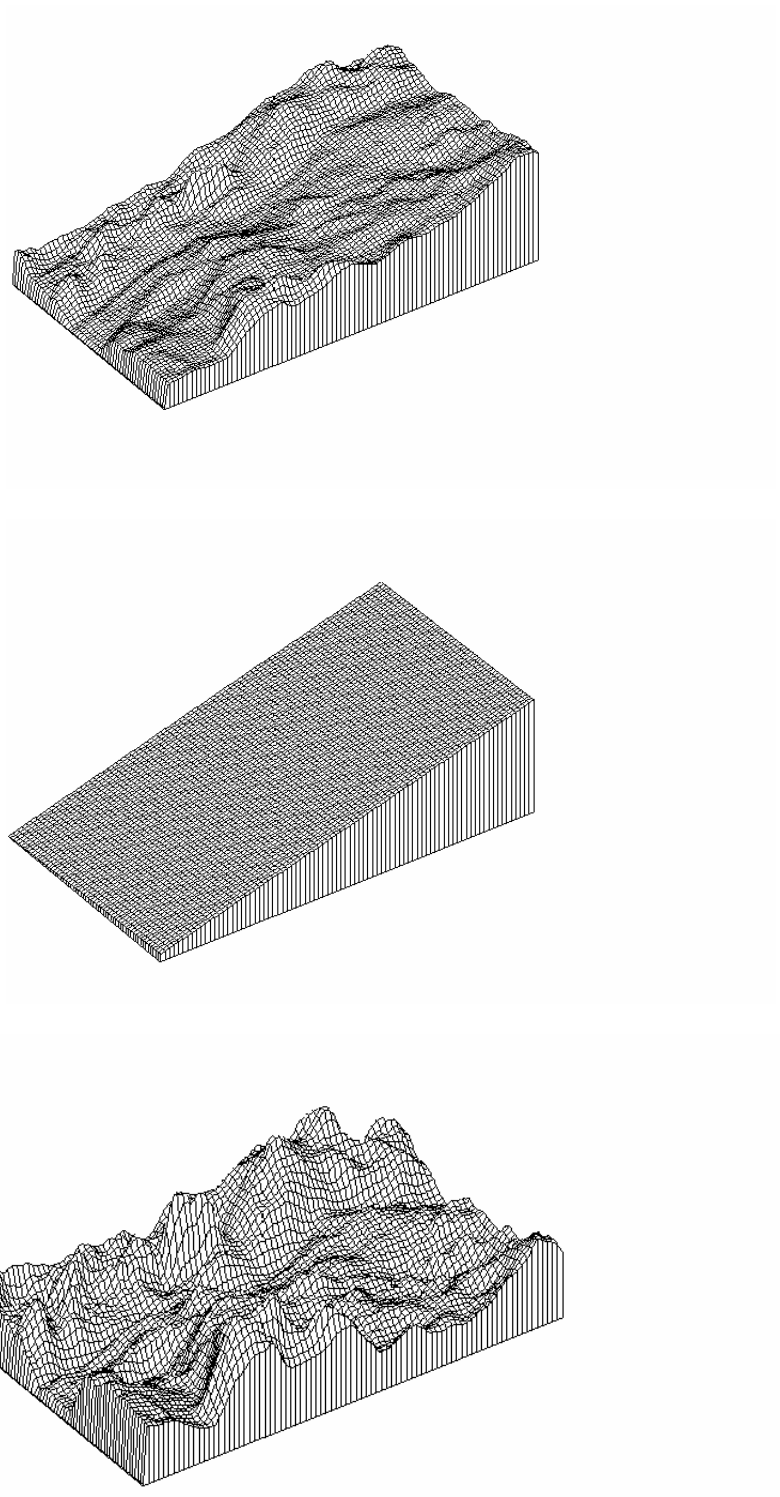


**Figura 15**. Representación del procedimiento kriging residual. La superficie interpolada (arriba) es igual a la suma de la tendencia lineal (centro) más la predicción de los errores (abajo).

### 4.4.5. Kriging Indicador

Suponga que se tiene una variable regionalizada $\{Z(x): x \in D \subset R^d\}$ estacionaria. Se define la siguiente transformación:

$$I(x_i, z_l) = \begin{cases} 1 & Si\ Z(x_i) \le z_l \\ 0 & Otro\ caso \end{cases}$$

Algunas propiedades son las siguientes:

*i.* $\Pr(I(x, z_l) = 1) = \Pr(Z(x) \le z_l) = F(z_l)$

*ii.* $E(I(x, z_l)) = 1\Pr(I(x, z_l) = 1) + 0\Pr(I(x, z_l) = 0)$

$= 1\Pr(I(x, z_l) = 1) = F(z_l)$

El predictor kriging indicador es igual a:

$$I^*(x_0, z_l) = \sum_{i=1}^{n} \lambda_i(z_l) I(x_i, z_l)$$

es decir que la predicción de la función indicadora en el sitio $x_0$ es igual a una combinación lineal de las *n* funciones indicadoras evaluadas en los sitios de medición. Samper y Carrera (1990) muestran que el kriging indicador es un estimador de la probabilidad acumulada hasta el límite *z* definido en la función indicadora. El predictor kriging indicador (dado que predice probabilidades acumuladas) tiene las siguientes restricciones:

*i.* $0 \le I^*(x, z_l) \le 1$

*ii.* $I^*(x, z_l) \le I^*(x, z_l')$ si $z_l \le z_l'$

Una condición suficiente para que estas restricciones se cumplan es que
$\lambda_i(z_l) = \lambda_i$ con $0 \le \lambda_i \le 1,\ \forall i, \forall z_l$.
Sin embargo en la práctica las ponderaciones se estiman de tal forma que el predictor sea insesgado de varianza mínima.
Para la condición de insesgamiento:

$$E(I^*(x_0, z_l)) = E(I(x_0, z_l)) = F(z_l)$$

$$\sum_{i=1}^{n} \lambda_i(z_l) E(I(x_i, z_l)) = F(z_l)$$

$$\sum_{i=1}^{n} \lambda_i(z_l) F(z_l) = F(z_l) \ \Rightarrow \ \sum_{i=1}^{n} \lambda_i(z_l) = 1$$

Después de llevar a cabo el proceso de derivación sobre la expresión de la varianza del error de predicción (obtenida de forma análoga a como se hizo en kriging ordinario), se obtiene el siguiente sistema de ecuaciones:

$$\sum_{i=1}^{n} \lambda_i(z_l)\gamma_{ij} + \mu = \gamma_{i0} \quad i=1, 2, ..., n$$

$$\sum_{i=1}^{n} \lambda_i = 1.$$

donde $\gamma_{ij} = \gamma(h)$, la función de semivarianza evaluad para la distancia entre los sitios *i, j*.
La varianza del error de predicción se encuentra de la misma forma a como se mencionó en la sección 4.2.

### 4.4.6. Kriging Log-Normal y Multi-Gaussiano

En estos dos procedimientos se hacen transformaciones de la variable regionalizada con el propósito de normalizar en cada sitio de la región de estudio.
El primero de estos consiste en aplicar kriging ordinario a la transformación logarítmica de los datos. Veamos:

Sea $\{Z(x): x \in D\}$ una variable regionalizada log-normal. Es decir que $Y(x) = Log(Z(x))$ tiene distribución normal. Algunas veces se requiere adicionar una constante positiva de tal forma que $Y(x)$ esté definida.
El predictor kriging log-normal es:

$$Y^*(x_0) = \sum_{i=1}^{n} \lambda_i Y(x_i).$$

Los pesos se obtienen de manera análoga al kriging ordinario. El semivariograma usado es el de los valores transformados. La complicación práctica puede darse al hacer la retransformación a la escala original, puesto que $Z^*(x_0) = \exp(Y^*(x_0))$ es un predictor sesgado.
Se puede demostrar que un predictor insesgado es (Cressie, 1993):

$Z^*(x_0) = \exp\left( Y^*(x_0) + \frac{\sigma_{ko}^2}{2} - \mu \right)$, donde $\sigma_{ko}^2$ es la varianza de predicción obtenida en el sitio $x_0$ por medio de kriging ordinario sobre los valores transformados y $\mu$ es el multiplicador de Lagrange empleado para la condición de insesgamiento sobre la escala de valores transformados.

Respecto al kriging multi-gaussiano, suponga que se tiene una variable regionalizada $\{Z(x): x \in D\}$ estacionaria. Este procedimiento consiste en hacer una transformación de *Z(x)* tal que los valores transformados sigan una distribución normal estándar. En ese sentido es una generalización del kriging log-normal.

Los pasos del método kroging multi-gaussiano son los siguientes:

*i.* Se encuentra la función de probabilidad acumulada empírica $F_n(Z(x))$.

*ii.* Se calculan con base en $F_n(Z(x))$ los "scores" normales estándar (Fig. 16), es decir los valores de una distribución de probabilidad normal estándar para los cuales la

probabilidad acumulada corresponde a $F_n(Z(x))$. En otras palabras se encuentra $U(x) = \Phi^{-1}(F_n(Z(x)))$.

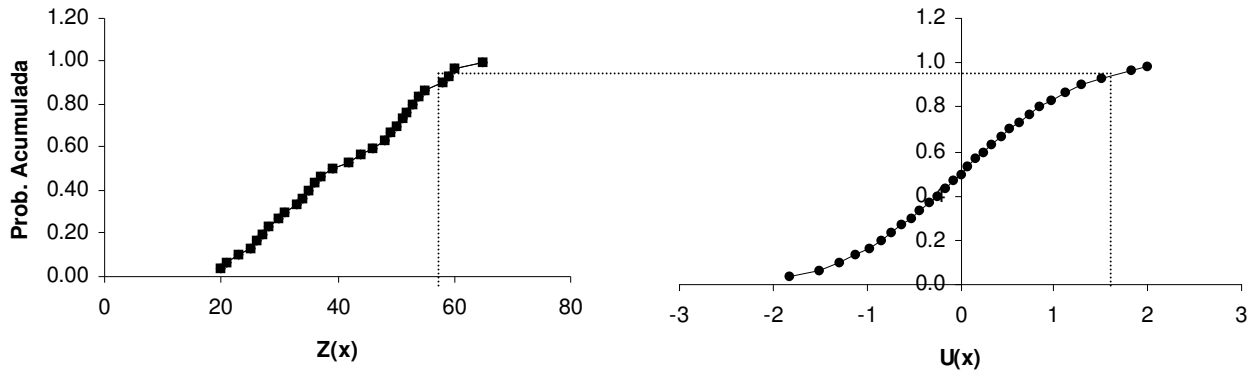



Figura 16. Representación de la transformación a scores normales

*iii.* Se realiza kriging simple sobre los valores transformados.

### 4.5. *Aplicación*: *Estudio de la distribución espacial de variables fisicoquímicas y biológicas medidas en el estuario Ciénaga Grande de Santa Marta*

Con base en la información descrita en la aplicación de la sección 3.3. y empleando los semivariogramas ajustados a las variables allí mencionadas, se generaron mapas de isolíneas de cada una de ellas (figuras 17 a 22) y se realizó interpretación de los mismos en un contexto ecológico. Como apoyo en la descripción se emplean medidas descriptivas de las variables (tabla 6).

**Tabla 6**. Medidas descriptivas de variables fisicoquímicas y biológicas medidas durante un muestreo realizado en marzo de 1997 en la Ciénaga Grande de Santa Marta.

| Variable | Nivel de la columna | Promedio | Mínimo | Máximo | Coeficiente de Variación (%) |
|---|---|---|---|---|---|
| Salinidad | Superficie | 17.6 | 13.02 | 34.9 | 16.1 |
| | Fondo | 18.04 | 13.94 | 33.9 | 15.5 |
| Oxígeno | Superficie | 8.80 | 3.03 | 16.2 | 36.9 |
| | Fondo | 5.68 | 2.64 | 13.4 | 36.8 |
| Sólidos en suspensión | Superficie | 218.2 | 103 | 318 | 18.8 |
| | Fondo | 215.35 | 86 | 310 | 19.6 |
| Nitritos | Superficie | 0.43 | 0.01 | 1.61 | 70.8 |
| | Fondo | 0.42 | 0.01 | 2.39 | 81.7 |
| Clorofila a | Superficie | 132.44 | 2.91 | 198.35 | 23.8 |
| | Fondo | 136.19 | 2.91 | 194.75 | 26.4 |
| Profundidad | | 1.47 | 0.25 | 2.50 | 24.1 |
| Transparencia | | 0.27 | 0.20 | 0.35 | 10.8 |

### 4.5.1. Mapas de Distribución

**Salinidad**.

Los valores medidos oscilaron entre 13.02 y 34.9 en el nivel superficial de la columna de agua y entre 13.94 y 33.9 en el fondo de la misma, con valores promedios de 17.6 y 18.04, respectivamente (tabla 6). De los mapas de isolíneas de distribución de dicha variable (Fig. 17), es posible concluir que existe gran homogeneidad en todo el cuerpo de agua, con excepción de la zona nororiental, puesto que los valores máximos y mínimos predichos varían sólo alrededor de 14 y 19 unidades.

En los dos mapas (superficie y fondo) de la figura 17, se evidencia la influencia que tienen las entradas de agua dulce y marina sobre la magnitud de la variable dentro del sistema. Las salinidades máximas se encuentran en la zona nororiental (zona estuarina) donde hay entrada de agua marina a través del sitio denominado Boca de la Barra (Fig. 1), encontrándose allí valores superiores a 30 unidades. Hacia la zona centro del cuerpo de agua se presentan las menores magnitudes de la variable, valores entre 15 y 16 unidades, lo que parece ser consecuencia del aporte de agua dulce que se da en la desembocadura de uno de los tres ríos (Río Sevilla) que baja de la Sierra Nevada de Santa Marta (SNSM). Así mismo se puede observar que en el sector occidental del sistema se presentan valores intermedios a los de las zonas antes mencionadas (alrededor de 19 unidades). Lo anterior puede deberse al efecto de intercambio de aguas, por medio de los canales Grande y Clarín, con el ecosistema Complejo Pajarales (Fig. 1), en donde se da un proceso de lavado de suelos hipersalinos en época de lluvias o cuando hay inundaciones. Dada la similitud en magnitud y forma de distribución que se observa en los mapas de superficie y fondo (Fig. 17), se podría pensar que para la época seca del año, no existe estratificación de la columna de agua respecto a esta variable.

Los valores de salinidad observados y predichos a través de la técnica kriging, resultan bajos respecto a los registrados para esta misma época en otros estudios (Giraldo *et al*., 1995). Lo anterior podría deberse a un posible aumento en los caudales de los ríos que desembocan en la CGSM, durante los meses de lluvia precedentes al muestreo, como consecuencia del efecto del fenómeno del niño en la región a finales del año 1996. No obstante lo anterior, puede pensarse, dada la gran homogeneidad en la distribución, que para la fecha del muestreo no se estaban presentando aportes considerables de agua dulce, por parte de los ríos que desembocan en la CGSM, lo que significa un período de relativa calma para los organismos que dependen de la salinidad para sus funciones y distribución (Reid y Wood, 1976).

**Oxígeno Disuelto**.

Los valores medidos de esta variable oscilaron entre 3.03 (mg/l) y 16.2 (mg/l) en la superficie de la columna de agua y entre 2.09 (mg/l) y 13.4 (mg/l) en el fondo de la misma, con valores promedios de 8.8 (mg/l) y 5.68 (mg/l), respectivamente (tabla 6).
Las correspondientes isolíneas (Fig. 18), indican que en el fondo de la columna de agua se presenta mayor homogeneidad en la distribución, puesto que los valores predichos varían entre 4.5 mg/l y 6.5 mg/l, con excepción de una pequeña zona en el sector nororiental frente a la desembocadura del río Sevilla (valores entre 6.5 y 9.5 mg/l), mientras que en superficie

existe considerable diferencia entre los valores ajustados en el centro del sistema (entre 9 mg/l y 13 mg/l) y los estimados para la zona sur y noroccidental del mismo (magnitudes alrededor de 4 mg/l). Lo anterior sugiere la ocurrencia de procesos de estratificación en el sistema hacia la zona central del espejo de agua, donde la productividad se concentra aportando grandes volúmenes de oxígeno al agua durante el día (Reid y Wood, 1976; Welch, 1992; Mancera y Vidal, 1994). Las isolíneas, para ambos niveles de la columna de agua, muestran que hacia las fronteras del sistema los valores del gas disminuyen. Este comportamiento podría ser explicado al considerarse que en estas zonas existe intercambio de flujos entre el sistema y otros cuerpos de agua, además de aportes de hojarasca y material orgánico, provenientes del manglar.

**Sólidos en suspensión**.

Los valores para la variable, presentan algunas diferencias entre los dos planos de muestreo. Los mínimos y máximos fueron de 103 mg/l y 318 mg/l en la superficie de la columna de agua y de 86 mg/l y 310 mg/l en el fondo de la misma, con promedios de 218.2 mg/l y 215.3 mg/l, respectivamente (tabla 6).

El mapa de distribución superficial (Fig. 19), revela la influencia que tienen los aportes de agua sobre la magnitud de esta variable en el sistema. Se observa que las mayores concentraciones se presentan en las zonas de las desembocaduras de los ríos Fundación y Aracataca además de la del caño Clarín (por medio del cual se da el aporte de agua del río Magdalena) y que las menores magnitudes se dan en el sector de intercambio de agua dulce y marina (desde la zona centro y nororiental hacia el sitio denominado Boca de la Barra). Una excepción a este comportamiento se da en la desembocadura del río Sevilla y del caño Grande en donde las concentraciones de los sólidos en suspensión son muy similares a las observadas en el resto del cuerpo de agua.

El patrón de comportamiento de la variable en el fondo de la columna de agua es muy similar al descrito en el párrafo de arriba respecto a los valores superficiales; es decir mayores concentraciones hacia las desembocaduras de los ríos y caños (zonas sur y nor-occidental) y menores magnitudes en la zona nororiental. Sin embargo, la diferencia entre los valores predichos en estas fronteras y los del resto del sistema (valores entre 220 mg/l y 210 mg/l), no resultan significativos como en el caso de la distribución superficial (valores entre 245 mg/l y 175 mg/l).

Una posible explicación a la diferencia en magnitud de los valores de superficie y fondo es que los flujos de agua dulce son menos densos y presentan mayores concentraciones de sólidos en suspensión, por lo cual al ingresar al sistema y encontrarse con las aguas salobres del mismo (más pesadas), tienden a permanecer en la superficie (lo cual puede causar estratificación. (Wheaton, 1977; Welch, 1992 ; Jay *et al*, 1997; Nixon, 1997).

**Nitritos.**

El ión nitrito presentó valores entre 0.01 µmol/l y 1.61 µmol/l para la superficie de la columna de agua y entre 0.01 µmol/l y 2.39 µmol/l en el fondo de la misma. Los valores promedios fueron de 0.43 µmol/l y 0.42 µmol/l, respectivamente (tabla 6).

Las mediciones superiores a 1 µmol/l se dieron en dos estaciones de muestreo, ubicadas en la zona norte del sistema (cuadrículas 14 y 15, Fig. 1). Dado que lo anterior no fue el patrón generalizado, las isolíneas de distribución en superficie y fondo (Fig. 20) presentan sólo valores alrededor de los promedios arriba mencionados. Este resultado es esperable, puesto que los nitritos generalmente se dan en bajas concentraciones (Mancera, 1990; Hernández y Gocke, 1990). Day *et al.* (1989), indican que esto puede ser debido al consumo continuo de las comunidades fitoplanctónicas y a la precipitación en los sedimentos como consecuencia de lo cambios en las condiciones del agua estuarina.

En ambos casos (superficie y fondo) los valores interpolados (entre 0.2 µmol/l y 0.7 µmol/l) para esta variable, revelan la presencia de un gradiente sur-norte, dándose las mayores concentraciones en el sector más estuarino. Lo anterior podría ser consecuencia de aportes de materia orgánica por parte de las poblaciones cercanas a esta zona (Welch, 1992). Los mapas de distribución espacial no revelan estratificación de la columna de agua para esta variable, dada la similitud en los valores predichos en superficie y fondo (Fig. 20).

**Clorofila "a".**

Los valores medidos de clorofila "a" oscilaron entre 2.91 µg/l y 198.35 µg/l en la superficie de la columna de agua y entre 2.91 µg/l y 194.75 µg/l en el fondo de la misma. Los valores promedios fueron 132.44 µg/l y 136 µg/l, respectivamente. Los bajos coeficientes de variación (menores del 30%), en ambos casos (superficie y fondo), indican relativa homogeneidad en las mediciones de esta variable (tabla 6). Las isolíneas de distribución calculadas con los datos predichos (Fig. 21) presentan algunos aspectos comunes. En ambos mapas (superficie y fondo) se observa que los valores máximos (alrededor de 160 µg/l) se dan en el sur del sistema hacia la desembocadura del río Fundación y las menores concentraciones se presentan en el sector más nororiental (valores menores de 50 µg/l). La diferencia radica en el comportamiento en la zona centro del espejo de agua. Mientras que en la superficie se presenta alta variabilidad (valores entre 90 y 160 µg/l), en el fondo de la columna de agua los valores son muy homogéneos (entre 130 µg/l y 150 µg/l) y se ajustan claramente a una tendencia creciente en sentido sur - nororiente. El comportamiento distribucional representado en los mapas de isolíneas puede estar de acuerdo con las condiciones climáticas de la época. Bula-Meyer (1989) y Sánchez (1996), indican que en la época más seca del año (época en la que se realizó el muestreo) predominan los vientos Alisios y que la circulación de las masas de agua en el sistema obedece a la fuerza del viento y a los cambios de marea en el Mar Caribe. Esto hace pensar que los flujos de agua son más lentos en la zona centro del sistema y por consiguiente, al no presentarse un recambio de agua muy fuerte, se favorece el desarrollo de las comunidades fitoplanctónicas, puesto que estas consumen los nutrientes que se liberan desde el sedimento por acción de los vientos (Welch, 1992).

Giraldo (1996), encontró un comportamiento similar en la distribución de esta variable con datos promedios de la época, calculados con información de varios años anteriores a 1995. Sin embargo en la zona de la desembocadura de los ríos, específicamente en la del Fundación, los valores reportados por dicho autor resultaron considerablemente más bajos a los encontrados en el presente estudio. Lo anterior puede estar indicando, como se mencionó en la interpretación de los resultados obtenidos con la salinidad, una disminución

de la entrada de agua dulce en esta zona, causándose así el mismo efecto de baja circulación y alta asimilación de nutrientes, comentado en el párrafo de arriba para la zona centro del sistema.

**Profundidad (Batimetría) y Transparencia (Secchi).**

Las medidas resúmenes de la variable profundidad (tabla 6) y el mapa de distribución calculado con los datos predichos (Fig. 22) confirman lo reportado en estudios anteriores (Wiedemann, 1973), en los cuales se afirma que la CGSM es un sistema somero, con una profundidad promedio alrededor de 1.5 m. El mencionado mapa revela la presencia de gradientes positivos en sentido oriente occidente y sur nor-occidente, respectivamente. Este comportamiento puede estar relacionado con los procesos de sedimentación que se han venido presentando en los últimos años cerca al sitio Boca de la Barra (cuadrícula 1, Fig. 1) como consecuencia de la disminución de los flujos de agua que entran al sistema a través de ríos que bajan de la SNSM y de los canales que comunican con el río Magdalena.

De otro lado respecto a la variable transparencia, medida por medio de la profundidad del disco de Secchi (Reid y Wood, 1976) se puede concluir, de acuerdo con los valores encontrados (tabla 6) y predichos (Fig. 22), que esta es una característica muy homogénea en el sistema de estudio. Las isolíneas de distribución indican que en general los valores esperados en este ecosistema no son superiores a 30 cm. Este bajo nivel de transparencia está de acuerdo con el resultado encontrado para la variable clorofila "a" (altas concentraciones en gran parte del sistema, como consecuencia de un aumento en las entradas de nutrientes), puesto que como lo muestra Welch (1992), la relación entre estas dos variables es de tipo inverso. Según resultados reportados por este autor se espera que para niveles de visibilidad del disco secchi, inferiores a 1 m se den concentraciones de clorofila "a" superiores a 80 µg/l.

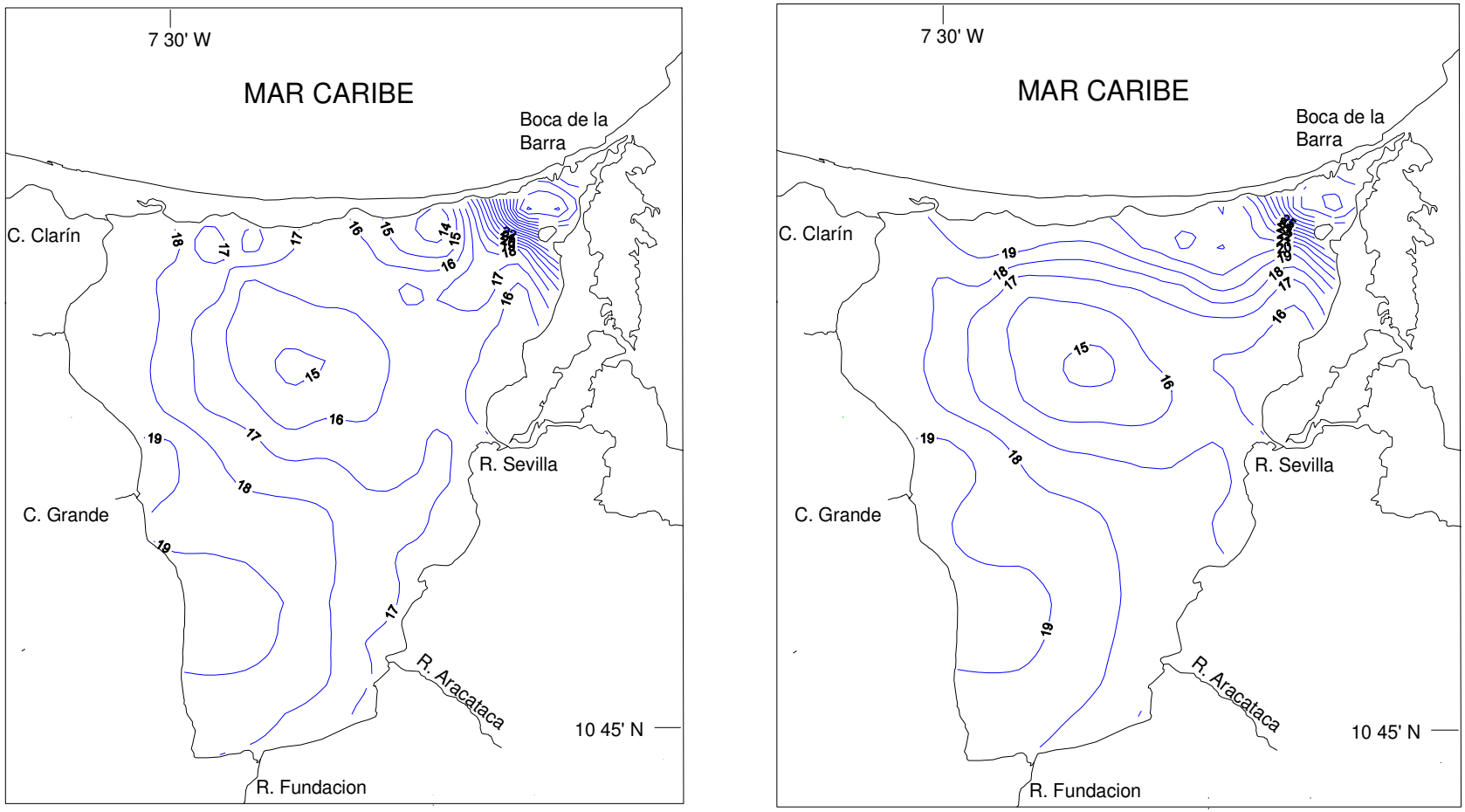



Figura 17. Distribución espacial de la salinidad del agua en la Ciénaga Grande de Santa Marta durante una jornada de muestreo realizada en marzo de 1997. El mapa de la izquierda corresponde a los valores en la superficie de la columna de agua y el de la derecha a los niveles en el fondo de la misma.

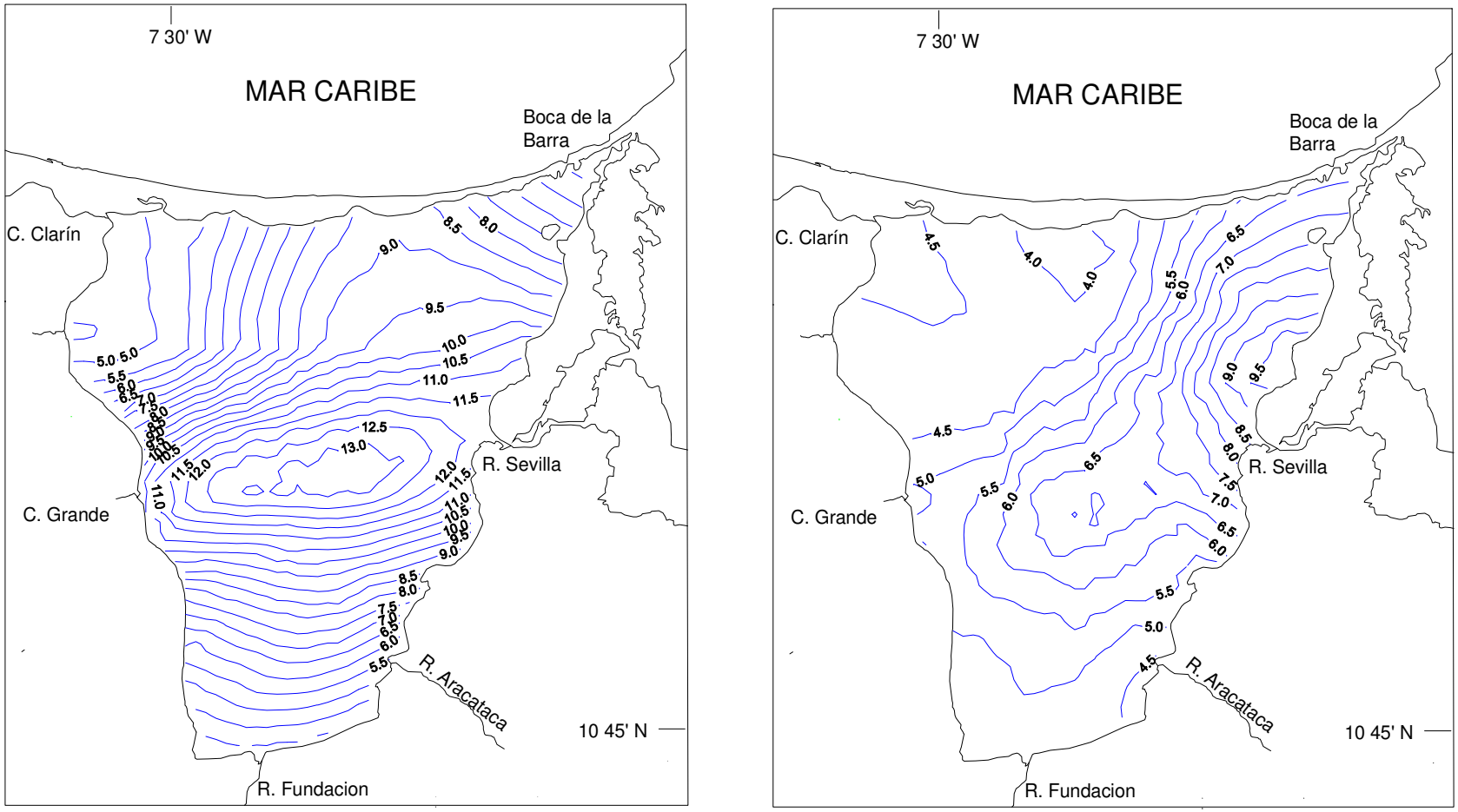



Figura 18. Distribución espacial del oxígeno disuelto (mg/l) en la Ciénaga Grande de Santa Marta durante una jornada de muestreo realizada en marzo de 1997. El mapa de la izquierda corresponde a los valores en la superficie de la columna de agua y el de la derecha a los niveles en el fondo de la misma.

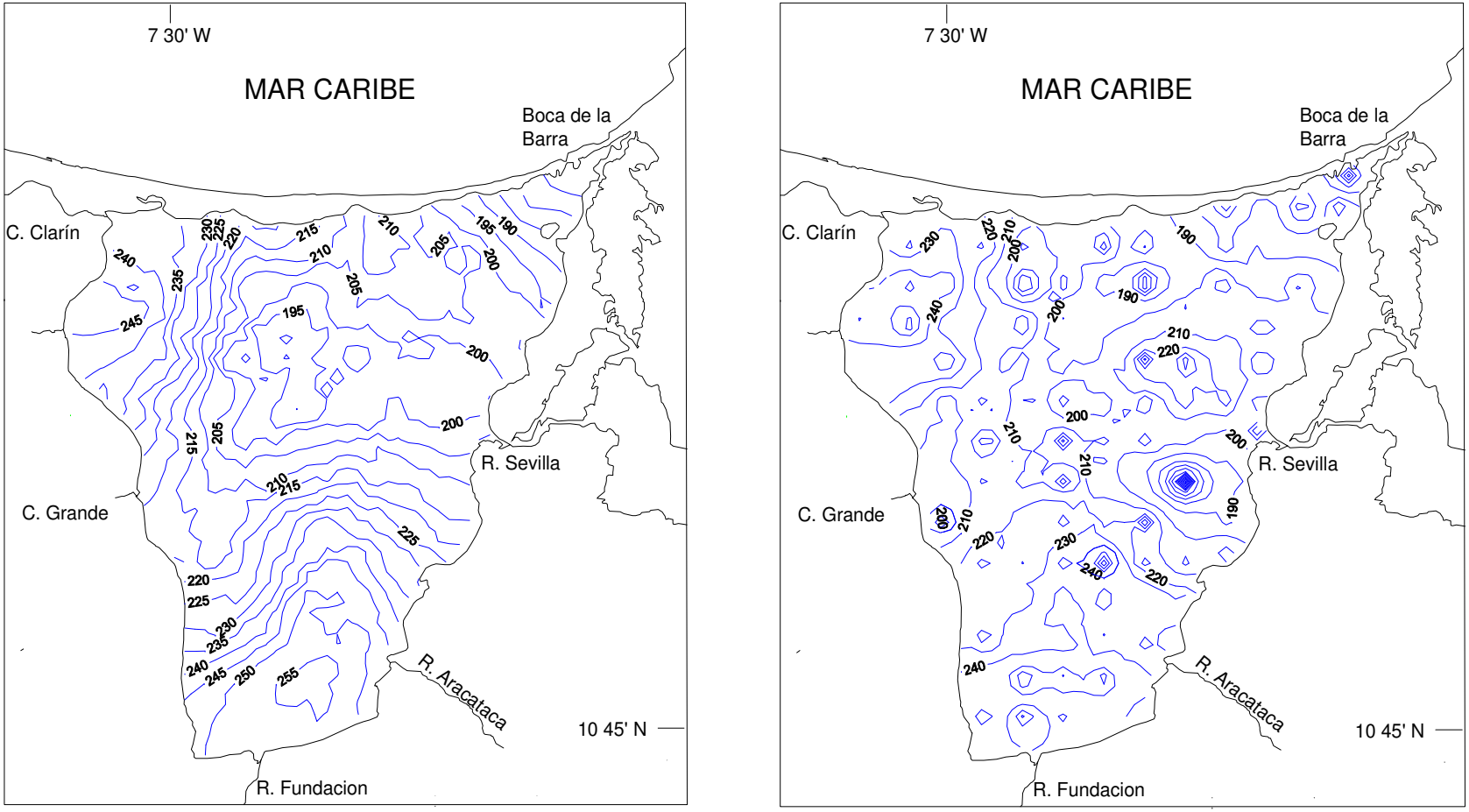



Figura 19. Distribución espacial de los sólidos en suspensión (mg/l) en la Ciénaga Grande de Santa Marta durante una jornada de muestreo realizada en marzo de 1997. El mapa de la izquierda corresponde a los valores en la superficie de la columna de agua y el de la derecha a los niveles en el fondo de la misma.

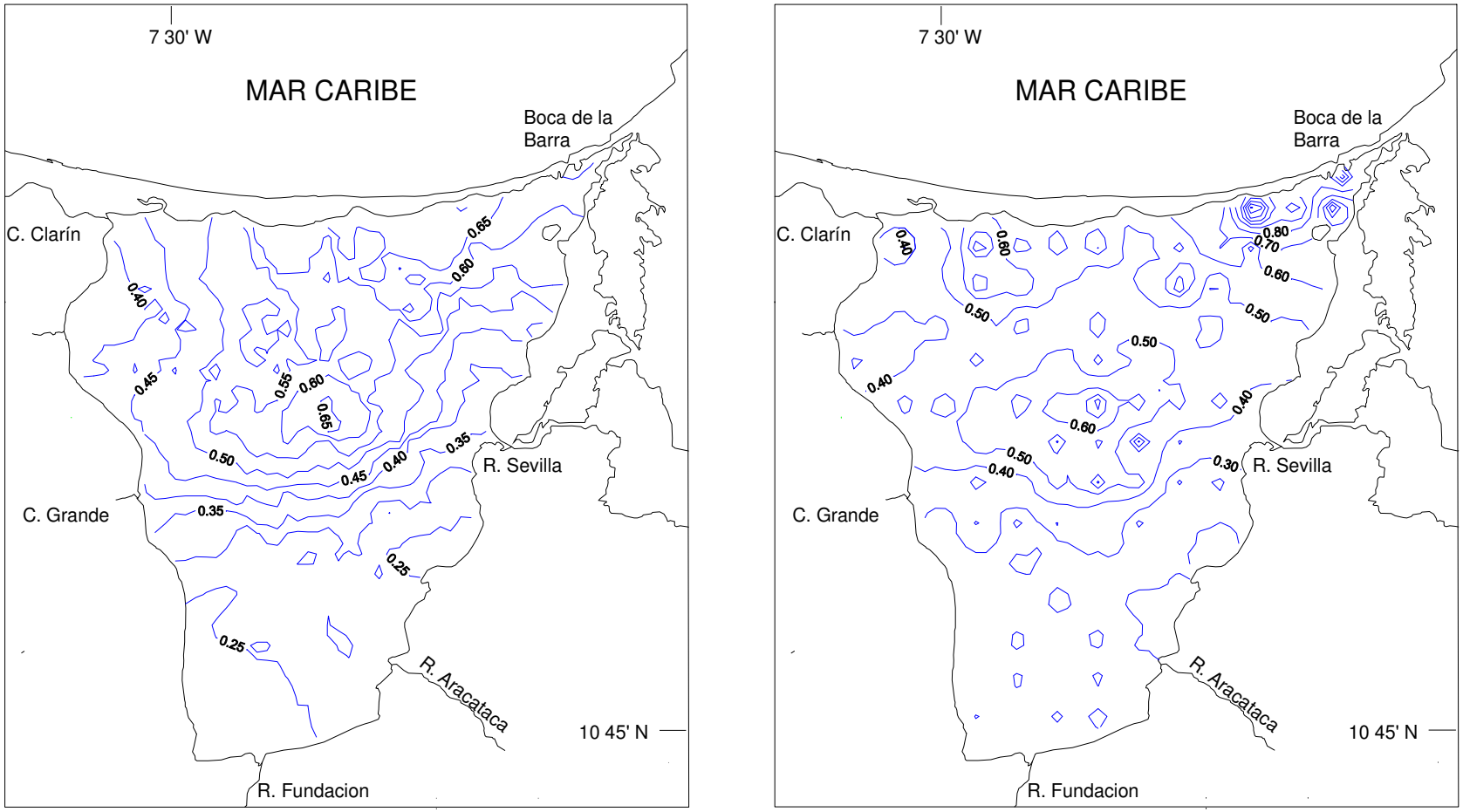



Figura 20. Distribución espacial de nitritos (µmol/l) en la Ciénaga Grande de Santa Marta durante una jornada de muestreo realizada en marzo de 1997. El mapa de la izquierda corresponde a los valores en la superficie de la columna de agua y el de la derecha a los niveles en el fondo de la misma.

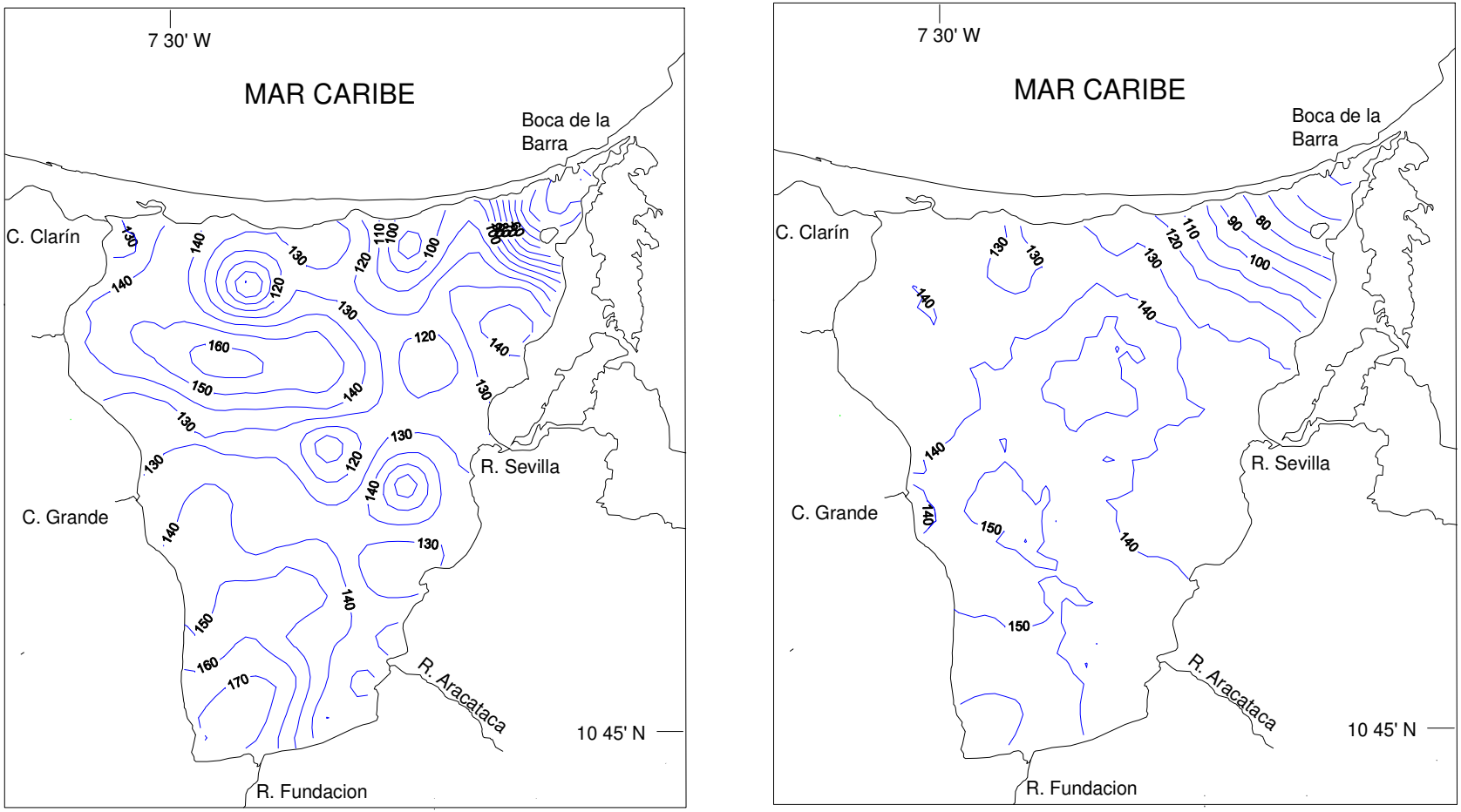



Figura 21. Distribución espacial de clorofila a (µg/l) en la Ciénaga Grande de Santa Marta durante una jornada de muestreo realizada en marzo de 1997. El mapa de la izquierda corresponde a los valores en la superficie de la columna de agua y el de la derecha a los niveles en el fondo de la misma.

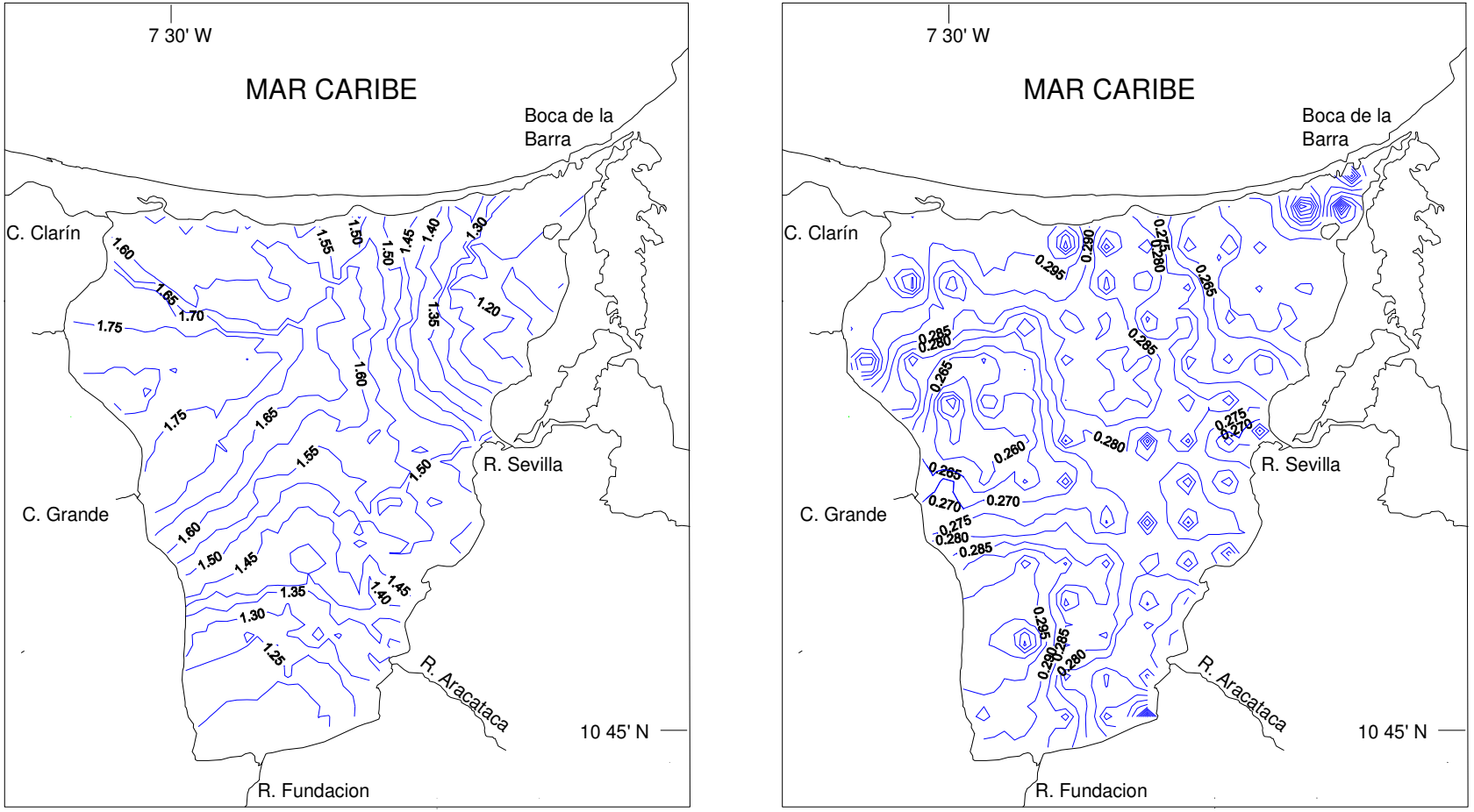



Figura 22. Distribución espacial de la profundidad (m) (izquierda) y transparencia (m) (derecha) en la Ciénaga Grande de Santa Marta durante una jornada de muestreo realizada en marzo de 1997.

# *Capitulo Cinco*

## *Temas Especiales*

En este capítulo se presentan algunos temas que no aparecen explícitamente en libros clásicos de geoestadística, tales como el diseño de redes muestrales o el análisis de componentes principales sobre variables regionalizadas. Así mismo se revisa la teoría del análisis cokriging y de simulación de fenómenos espaciales, bajo el supuesto de normalidad multivariada.

### 5.1. Cokriging Ordinario

Si se tienen dos variables regionalizadas $Z_{v1}(x)$ y $Z_{v2}(x)$ tomadas en sitios de la región de estudio, no necesariamente iguales, entonces el semivariograma cruzado entre ellas, se estima por:

$$\gamma_{v_1v_2}(h)=\frac{1}{2n_h}\sum^{n_h}\left\{Z_{v_1}(x+h)-Z_{v_1}(x)\right\}\left\{Z_{v_2}(x+h)-Z_{v_2}(x)\right\} \tag{9}$$

Donde $n_h$ es el número de parejas de datos que se encuentran a una distancia h (Bogaert *et al*., 1995).

- **Modelo Lineal de Corregionalización (MLC)**

El MLC asume que todos los semivariogramas simples y cruzados pueden expresarse como una suma de modelos básicos (exponencial, esférico, gaussiano, etc.) idénticos. Para el caso de dos variables:

$$\begin{aligned}\gamma_{v_1}(h)&=\alpha_0\gamma_0(h)+\ldots+\alpha_m\gamma_m(h)\\ \gamma_{v_2}(h)&=\beta_0\gamma_0(h)+\ldots+\beta_m\gamma_m(h)\\ \gamma_{v_1v_2}(h)&=\delta_0\gamma_0(h)+\ldots+\delta_m\gamma_m(h)\end{aligned} \tag{10}$$

donde $\gamma_{v1}(h)$ y $\gamma_{v2}(h)$ son los semivariogramas simples, $\gamma_{v1v2}(h)$ es el semivariograma cruzado. $\gamma_0(h), \gamma_1(h), \ldots, \gamma_m(h)$ son los modelos básicos de semivariograma y $\alpha$, $\beta$ y $\delta$ son constantes.

Matricialmente:

$$\Gamma(h)=\begin{pmatrix}\gamma_{v_1}(h) & \gamma_{v_1v_2(h)}\\ \gamma_{v_1v_2(h)} & \gamma_{v_2}(h)\end{pmatrix}=\sum_{s=0}^{m}B_s\gamma_s(h),\ \text{donde}$$

$$B_s=\begin{pmatrix}\alpha_s & \delta_s\\ \delta_s & \beta_s\end{pmatrix}\quad \gamma_s(h)=\begin{pmatrix}\gamma_s(h) & 0\\ 0 & \gamma_s(h)\end{pmatrix} \tag{11}$$

*Γ(h)* se le conoce como matriz de corregionalización.

- **Predictor Cokriging**

El método de predicción espacial *cokriging* consiste en hacer predicción espacial de una variable con base en su información y en la de algunas variables auxiliares que este correlacionadas espacialmente con ella. El predictor cokriging tiene la siguiente expresión en el caso en el que se considera una sola variable auxiliar:

$$\hat{Z}^*_{v_1}(x_o)=\sum_{i=1}^{n_1} a_i Z_{v_1}(x_i)+\sum_{j=1}^{n_2} b_j Z_{v_2}(x_j) \qquad (12)$$

el lado izquierdo de la igualdad en la ecuación anterior representa la predicción de la variable de interés en el punto $x_0$ no muestreado. $Z_{v_1}(x_i)$ con $i=1, 2, \ldots, n_1$, representa la variable primaria. Así mismo, $Z_{v_2}(x_j)$ con $j=1, 2, \ldots, n_2$, representa la variable auxiliar. $a_i$ y $b_j$, con $i=1, 2, \ldots, n_1$ y $j=1, 2, \ldots, n_2$ respectivamente, representan los pesos o ponderaciones de las observaciones de las variables primaria y auxiliar y se estiman con base en el MLC ajustado a los semivariogramas simples y cruzados. Los pesos $a_i$ y $b_j$ se estiman de manera análoga al proceso descrito para el método kriging ordinario, es decir estos serán los que minimizan la varianza del error de predicción sujeta a la restricción de que el predictor sea insesgado. La estimación de los parámetros se obtiene resolviendo el siguiente sistema de ecuaciones (Isaaks y Srivastava, 1989):

$$\begin{pmatrix}
\gamma_{v1}(1,1) & \cdots & \gamma_{v1}(n,1) & \gamma_{v1v2}(1,1) & \cdots & \gamma_{v1v2}(m,1) & 1 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\gamma_{v1}(1,n) & \cdots & \gamma_{v1}(n,n) & \gamma_{v1v2}(1,n) & \cdots & \gamma_{v1v2}(m,n) & 1 & 0 \\
\gamma_{v1v2}(1,1) & \cdots & \gamma_{v1v2}(n,1) & \gamma_{v2}(1,1) & \cdots & \gamma_{v2}(m,1) & 0 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\gamma_{v1v2}(1,m) & \cdots & \gamma_{v1v2}(n,m) & \gamma_{v2}(1,m) & \cdots & \gamma_{v2}(m,m) & 0 & 1 \\
1 & \cdots & 1 & 0 & \cdots & 0 & 0 & 0 \\
0 & \cdots & 0 & 1 & \cdots & 1 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
a_1 \\ \vdots \\ a_n \\ b_1 \\ \vdots \\ b_m \\ \mu_1 \\ \mu_2
\end{pmatrix}=
\begin{pmatrix}
\gamma_{v1}(0,1) \\ \vdots \\ \gamma_{v1}(0,n) \\ \gamma_{v1v2}(0,1) \\ \vdots \\ \gamma_{v1v2}(0,m) \\ 1 \\ 0
\end{pmatrix} \qquad (13)$$

La matriz del lado izquierdo contiene los valores de las funciones de semivarianza y de semivarianza cruzada calculadas para todas las distancias entre las parejas de puntos consideradas. Las dos ultimas filas de dicha matriz son las correspondientes a la restricción de insesgamiento del predictor. $a_i$ y $b_j$ con $i = 1, 2, ..., n$ y $j = 1, 2, ...., m$, son los parámetros a estimar, $\mu_1$ y $\mu_2$ son los multiplicadores de Lagrange empleados para la restricción de insesgamiento y el vector del lado derecho contiene los valores de la funciones de semivarianza y semivarianza cruzada evaluados para las distancia entre los sitios de muestreo (de ambas variables) y el sitio donde se desea hacer la predicción. Las dos últimas filas del vector están asociadas a la condición de insesgamiento. La correspondiente varianza de predicción del método cokriging se calcula como (Bogaert *et al*, 1995):

$$\sigma_k^2 = Cov(Z_{v1}(x_0), Z_{v1}(x_0)) + \mu_1 + \sum_{i=1}^{n} a_i Cov(Z_{v1}(x_i), Z_{v1}(x_0)) + \sum_{j=1}^{m} b_j Cov(Z_{v2}(x_j), Z_{v2}(x_0)) \qquad (11)$$

donde $\mu_1$ es el multiplicador de Lagrange empleado para la restricción dado por la condición de insesgamiento $\left(\sum_{i=1}^{n} a_i = 1\right)$.

$Cov(Z_{vi}(x_k), Z_{vi}(x_l)) = \sigma_{vi}^2 - \gamma_{vivi}(k,l)$ es la función de covarianza espacial de la variable *i, i=1,2,* evaluada para la distancia entre los sitios de muestreo *k, l.*

La ventaja del método *cokriging* frente al *kriging* radica en el hecho de que cuando la variable auxiliar está ampliamente correlacionada con la variable de interés se puede obtener un disminución en la varianza de predicción, no obstante dicha variable tenga menor densidad de muestreo. En situaciones en las que la variable objetivo tiene costos altos de muestreo se recomienda la aplicación de esta metodología (Bogaert *et al*., 1995).

- **Kriging Probabilístico**

Es un predictor basado en cokriging que utiliza como variables predictoras una variable indicadora y una variable generada a través de la *transformación uniforme.*

Sea $Z(x_i)$ la variable observada, $i = 1,2, \ldots, n$, entonces se definen las siguientes transformaciones:

- $I(x_i, z) = \begin{cases} 1 & Si\ Z(x_i) \le z \\ 0 & Otro\ caso \end{cases}$
- $U(x_i) = \frac{R(Z(x_i)}{n}$ para todo *i, i = 1,2,. . . , n*.

con $R(Z(x_i))$ igual al rango (posición que ocupa dentro de los datos ordenados de menor a mayor) la i-ésima observación muestral. La predicción de probabilidad de éxito en el sitios de interés está dada por:

$$I^*(x_0) = \sum_{i=1}^{n} \lambda_i I(x_i, z) + \sum_{i=1}^{n} v_i U(x_i)$$

Los pesos $\lambda_i$ y $v_i$ se estiman mediante el sistema de ecuaciones del método cokriging.

### 5.2. Análisis de Componentes Principales Regionalizado (ACPR)

El ACPR se fundamenta en la realización de análisis de componentes principales (ACP) (apéndice, sección 6.4.2) con base en varias matrices de corregionalización (sección 5.1). El caso más simple de ACPR es cuando se aplica ACP con base en la matriz de corregionalización a distancia cero (matriz de correlación tradicional). En este caso la técnica consiste en generar los ejes principales de la forma tradicional (Manly, 1994), posteriormente realizar la correspondiente interpretación de estos en términos de la variabilidad explicada por cada componente respecto a cada variable original y finalmente llevar a cabo un análisis geoestadístico a través de la estimación de la función de semivarianza y de la aplicación de algún procedimiento kriging con base en los datos de los ejes generados. La interpretación del mapa de predicciones obtenida sobre los componentes permite obtener una visión integral del comportamiento conjunto de las variables consideradas dentro del sistema de estudio.

El procedimiento se puede realizar con matrices de correlación obtenidas para diferentes distancias (fijadas de antemano) y calculadas a través de la función de semivarianza, covarianza cruzada o de correlación cruzada (si las variables tienen diferentes escalas de medida se recomienda emplear la función de correlación cruzada). Lo anterior implica que se deben calcular $\binom{n}{2}$ funciones de correlación espacial cruzada, siendo *n* el número total de variables involucradas en el estudio. Esto puede ser una limitante computacional del método, cuando se incremente el número de variables. En la práctica se acostumbra a seleccionar grupos de pocas variables (alrededor de 5, consideradas como las más relevantes) que estén muy relacionadas espacialmente y con base en la información de estas hacer el análisis para dos o tres matrices de corregionalización (incluyendo la de distancia cero).

- **Correlación Intrínseca.**

Se dice que un conjunto de variables regionalizadas tiene correlación intrínseca cuando la estructura de correlación de las variables es independiente de la distancia espacial (puede haber correlación para distancia cero), es decir cuando las funciones de semivarianza cruzada, covarianza cruzada o correlación cruzada, entre las variables, son constantes en función de la distancia. La detección de correlación intrínseca puede hacerse a través de las variables originales o por medio de los componentes principales generados. En la sección 6.4.2., se menciona que los ejes principales deben ser independientes, luego se espera que para cualquier distancia la función de semivarianza cruzada entre cualquier para de componentes principales esté alrededor de cero. En caso contrario habrá relación espacial entre las variables. Existen dos formas de llevar a cabo el ACPR dependiendo de si las variables tienen o no correlación intrínseca.

- **ACPR en Presencia de Correlación Intrínseca.**

El algoritmo en este caso se inicia con el cálculo de la matriz de corregionalización para distancia cero (matriz de correlación clásica) (en las otras distancias no hay correlación entre las variables). Posteriormente se aplica el ACP clásico se generan ejes factoriales que explican, se espera que en un alto porcentaje, la variabilidad contenida en el conjunto total de variables (idealmente dos o tres componentes deberían explicar más del 90% de la varianza total). Con base en la magnitud y le signo de los vectores propios se identifica el peso de cada variable original en los correspondientes ejes. Finalmente se obtiene un mapa de distribución espacial, cumpliendo con las etapas básicas del análisis geoestadístico, que permita dar una interpretación simultánea del comportamiento de las variables involucradas en el análisis.

- **ACPR sin Correlación Intrínseca.**

En este caso se debe establecer en primer lugar un modelo lineal de corregionalización entre las variables originales. Con base en este se calculan varias matrices de corregionalización (una para cada distancia *h* fijada) y con cada una de ellas se realiza un ACP clásico. Los resultados obtenidos en cada análisis permiten establecer relaciones entre las variables que no son observados en análisis clásicos de correlación .

## 5.3. Diseño de Redes de Muestreo

- **Selección de Variables**

Cuando se va a iniciar el estudio de un ecosistema natural, deben establecerse aspectos referentes a su ubicación geográfica, a las condiciones climáticas, hídricas y geológicas del mismo. La revisión debe incluir los antecedentes históricos de las variables ecológicas e incluso los factores económicos, socio-culturales y demográficos que enmarcan a la región de estudio. Todos estos aspectos permiten planear, desde un punto de vista logístico la realización del muestreo.

Podría decirse que existen dos tipos de variables que deben tenerse en cuenta cuando se realiza un estudio ambiental. Aquellas que están directamente relacionadas con el fenómenos ecológico de estudio (contaminación, calidad del agua, manglar, bentos, plancton, etc.) y aquellas que permiten decidir respecto a aspectos relacionados con las frecuencia del muestreo (variables climáticas) y la ubicación estratégica de sitios de muestreo (por ejemplo variables hídricas como dirección y velocidad de las masas de agua, fuentes de entrada y salida, etc)

Así mismo la selección de las variables es consecuencia del tipo de estudio que se quiere llevar a cabo. Estos pueden ser de línea base, monitoreo y evaluación. En el primero de estos no hay antecedentes históricos respecto al fenómeno de interés, se asume total desconocimiento respecto a la relación, comportamiento y distribución de las variables en el ecosistema y por consiguiente se debe evaluar un número grande de variables, con amplia intensidad de muestreo en la que se cubra toda la región de estudio, de forma tal que se pueda caracterizar de manera general el sistema. Si existe conocimiento de la región de estudio, hay estudios preliminares que posibilitan el planteamiento de estructuras de correlación espacial y temporal de las variables y se quieren establecer los cambios que se están dando en el ecosistema, por ejemplo por actividades antrópicas, es entonces un estudio de monitoreo. En este caso debe establecerse con base en la información disponible tanto la frecuencia como la ubicación de los puntos de muestreo. Por último cuando hay conocimiento del ecosistema respecto al fenómeno de interés y se quieren observar posibles variaciones muy puntuales respecto al patrón temporal o espacial tradicionalmente observado, el estudio se denomina de evaluación. En este último caso el objetivo puede ser el de conservar o mitigar posibles daños más que el de hacer diagnóstico como en el caso del monitoreo.

- **Selección de la Red Optima de Muestreo**

Como en cualquier procedimiento estadístico en el que se hace inferencia, en geoestadística cuando se hace predicción en sitios o puntos de la región de estudio no observados, a través de cualquiera de las técnicas kriging, es necesario evaluar la precisión de tal predicción. Lo anterior se realiza, como se estableció en el capitulo 4 y en la sección 5.1, por medio del cálculo de la varianza del error de predicción. De la sección 4.2, para el caso del kriging ordinario, la varianza de predicción se calcula por:

$$\sigma^2 = \sum_{i=1}^{n} \lambda_{ii}\gamma_{io} + \mu$$

Es evidente, de la ecuación anterior, que la varianza de predicción no es constante como en el caso clásico y que además no depende de los valores medidos de la variable sino de su estructura de correlación, evaluada a través de la función de semivarianza.

McBratney *et al* (1981) muestran que, para cualquier densidad muestral, la distancia máxima entre un punto de observación y un punto a interpolar es mínima cuando la configuración de los puntos es hecha en un enmallado triangular, por lo cual bajo esta distribución de puntos se obtendrán las menores varianzas de predicción. Sin embargo este mismo autor y Warrick *et al* (1986) indican que por razones logísticas referentes a la ubicación de los sitios en el campo y minimización de recorridos, el enmallado cuadrado es preferible.

De acuerdo con lo anterior el problema del diseño muestral se limita a establecer para varias redes de muestreo, de diferentes densidades, con enmallado triangular equilátero o cuadrado, la relación entre las varianzas de predicción máximas (las obtenidas en el centro de cada triángulo o cuadrado) y los costos asociados a ellas. De esta forma se deduce el costo necesario para alcanzar cierto grado de certidumbre o, a la inversa, la varianza de predicción si se prefija el costo.

### 5.4. Simulación

A continuación se describe el método de simulación de variables regionalizadas con densidad conjunta normal multivariada (gaussiana) (Cressie, 1993) que se fundamenta en la técnica de descomposición de Cholesky (Anderson, 1984).

Suponga que se desea simular el vector $\vec{Z}(x) = \begin{pmatrix} Z(x_1) \\ Z(x_2) \\ \vdots \\ Z(x_n) \end{pmatrix}$ correspondiente a $n$ variables aleatorias en $n$ sitios de muestreo de interés $x_1, x_2, ..., x_n$. Asuma que el proceso estocástico tiene vector de medias y matriz de varianzas y covarianzas

$$E\left(\vec{Z}(x)\right) = \vec{\mu}(x) = \begin{pmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_n) \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & C_{12} & \cdots & C_{1n} \\ C_{21} & \sigma_2^2 & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{pmatrix}, \text{ con } C_{ij} = COV\left(Z(x_i), Z(x_j)\right).$$

La matriz de varianzas y covarianzas es descompuesta por el método de Cholesky.

$\Sigma = LL^T$, con L una matriz triangular inferior. Entonces el vector simulado se define como:

$\vec{Z}(x) = \vec{\mu}(x) + L\vec{\varepsilon}$, donde $\vec{\varepsilon} \sim N_n\left(\vec{0}, I\right)$.

Usando teoremas referentes a la distribución de combinaciones lineales de vectores con distribución normal multivariada (Anderson, 1984) se comprueba que el vector simulado tiene vector de medias $\vec{\mu}(x)$ y matriz de covarianzas $\Sigma$.

## 5.5. Aplicaciones

### 5.5.1. *Comparación de los métodos Kriging y Cokriging con base en resultados de análisis espaciales de Variables Fisicoquímicas y Biológicas Medidas en el Estuario Ciénaga Grande de Santa Marta.*

Se emplea la información de las variables profundidad (m), secchi (cm), salinidad, oxígeno (mg/l) y clorofila "a" (µg/l) descrita en la aplicación 3.3. para mostrar las ventajas del método cokriging respecto al kriging.

Se mencionó en la sección 4.1. que la técnica cokriging es preferible al kriging cuando hay información de variables auxiliares relacionadas espacialmente con la variable de interés y ésta última tiene altos costos de muestreo. En este caso se usan como variables auxiliares profundidad, secchi, salinidad y oxígeno y como variable objetivo clorofila "a". Las cuatro primeras tienen bajos costos de muestreo (son medidas *in-situ*) mientras que los altos costos de los insumos de laboratorio necesarios para la obtención de las medidas de clorofila "a" pueden ser limitantes del uso de una red de muestreo densa para dicha variable. Aunque no se presentan en el trabajo los semivariogramas cruzados entre las variables consideradas como auxiliares y clorofila "a", estos muestran la presencia de fuertes estructuras de dependencia espacial entre ellas.

Debido a que la aplicación del método cokriging resulta dispendiosa cuando se consideran dos o más variables auxiliares (es complejo el ajuste del modelo lineal de corregionalización), se decidió tomar la información de las variables auxiliares de forma "condensada" a través del indicador $IGC_i(4)$ (Giraldo, 2002; apéndice, sección 6.1), calculado con base en la información dicotomizada de dichas variables.. El criterio para dicotomizar cada variable fue el de comparación de cada valor observado con su correspondiente mediana. Se asignó el valor 1 para valores mayores o iguales que la mediana y 0 en caso contrario. Se aplicó el método cokriging para realizar predicciones de la variable clorofila "a" con base en sus observaciones e información auxiliar del indicador $IGC_i(4)$. Para detectar la eficiencia del método se redujo a aproximadamente la mitad (54 datos) la información original de clorofila "a" (en adelante se denomina a ésta como CLORO 2) y se dejó la información completa (114 datos) del $IGC_i(4)$.El ajuste del modelo lineal de corregionalización y el cálculo de las varianzas de predicción fue realizado en el software The Spatial Interpolation of Agroclimatic Data (Bogaert *et al.*, 1995).

- **Resultados y Discusión**

En primera instancia se calcularon semivariogramas simples de las variables CLORO 2 e $IGC_i(4)$ y los semivariogramas cruzados entre estas tres variables. Sólo se muestra el semivariograma experimental cruzado (Fig. 23). Este indica que las dos variables consideradas presentan correlación espacial inversa, es decir que valores altos de productividad biológica (alta clorofila "a") están asociados con valores bajos del indicador $IGC_i(4)$ en zonas circundantes (incluso mayores a 10 km). Los valores bajos del indicador $IGC_i(4)$ están asociados a magnitudes por debajo de la mediana en las variables profundidad, secchi, salinidad y oxígeno disuelto (ver tabla de interpretación del $IGC_i(4)$ en el apéndice). De lo anterior se concluye que zonas con alta biomasa fitoplanctónica están asociadas a baja profundidad, alta turbidez y a masas de agua con baja salinidad y bajo nivel de oxígeno (tal vez como consecuencia del consumo de éste durante las horas del día).

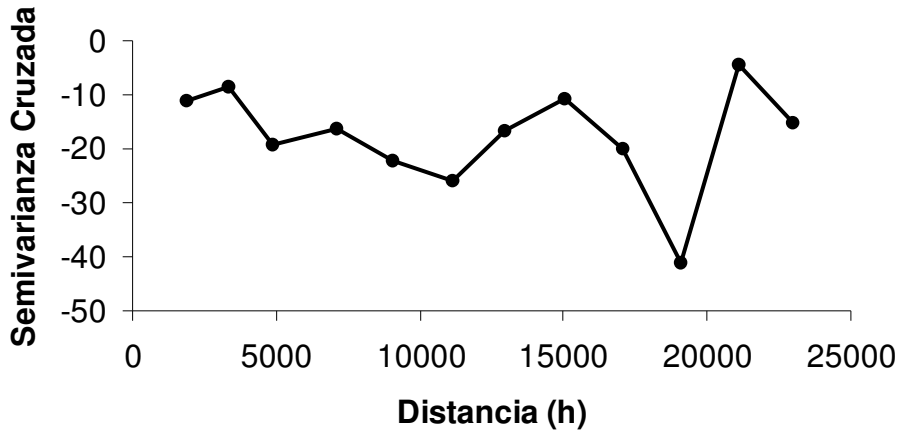



**Figura 23**. Semivariograma experimental cruzado entre las variables clorofila "a" (54 datos) e $IGC_i(4)$. Información tomada en marzo de 1997 en la Ciénaga Grande de Santa Marta.

Una vez calculados los semivariogramas experimentales se ajustó el modelo lineal de corregionalización entre las variables CLORO 2 e $IGC_i(4)$ (tabla 7), el cual incluye efecto pepita puro y un modelo esférico. Con base en el modelo lineal de corregionalización se realizaron las predicciones, de la variable CLORO 2, a través del método cokriging en 53 sitios de muestreo (aquellos en los que fue eliminada inicialmente la información de clorofila "a") y se calculó la varianza de predicción máxima, mínima y promedio (tabla 8). Utilizando la información de los 54 datos de clorofila "a", se llevó a cabo predicción en los 53 sitios restantes a través del método kriging y se calcularon nuevamente las varianzas de predicción máxima, mínima y promedio (tabla 8). Los resultados obtenidos por estos dos métodos fueron comparados con la varianza máxima de predicción obtenida por Giraldo *et al*. (2001) para la variable clorofila "a" utilizando la información completa (tabla 8).

**Tabla 7**. Modelo de corregionalización ajustado a los semivariogramas experimentales (simples y cruzado) de las variables clorofila "a" (V1, 54 datos) e IGC(4) (V2, 114 datos). La información original fue medida en una muestreo realizado en marzo de 1997 en la Ciénaga Grande de Santa Marta.

| | Modelo Ajustado |
|---|---|
| $\gamma_{v1}(h)$ | 131.82 + 535.4 Esférico (8000) |
| $\gamma_{v2}(h)$ | 3.89 + 9.59 Esférico (8000) |
| $\gamma_{v1v2}(h)$ | -1.18 - 18.70 Esférico (8000) |

**Tabla 8**. Varianzas de predicción mínima, máxima y promedio ((μg)$^2$ /l) para la variable clorofila "a", usando los métodos kriging y cokriging (con base en información auxiliar de la variable IGC(4)). Entre paréntesis se encuentran la ganancia en precisión respecto al método kriging con información incompleta.

| MÉTODO | VARIANZA DE PREDICCIÓN MÁXIMA | VARIANZA DE PREDICCIÓN MÍNIMA | VARIANZA DE PREDICCIÓN PROMEDIO |
|---|---|---|---|
| Kriging con datos en 107 sitios de muestreo (información completa) | 379 (25%) | 0 | |
| Kriging con datos en 54 sitios de muestreo (información reducida) | 506 (0%) | 0 | 194.147(0%) |
| Cokriging con datos en 54 sitios de muestreo para la variable clorofila "a" y 114 datos para la variable $IGC_i(4)$ | 488 (4%) | 0 | 190.06 (2.1%) |

De los resultados de la tabla anterior se concluye, como era de esperarse, que la reducción en el número de observaciones (reducción del número de sitios de muestreo) a cerca de la mitad, ocasiona un aumento en la varianza de predicción de la variable considerada. No obstante dicho aumento es menor cuando se aplica el método cokriging utilizando como variable auxiliar la variable $IGC_i(4)$. Al hacer la predicción de la variable clorofila "a" con menos información empleando el método cokriging se gana un 4% en términos de la varianza máxima y un 2.1 % , en términos de la varianza promedio (promedio de 53 varianzas estimadas), de precisión respecto al método kriging con datos incompletos.

Teniendo en cuenta lo anterior y que la variable clorofila "a" presenta altos costos de muestreo (pasar de 107 sitios a 54 disminuye en más del 200% (cerca de 2 millones de pesos) los costos (Giraldo *et al.,* 2001) se recomienda en este caso la aplicación del método de cokriging. Con la conclusión anterior no se indica que el número óptimo de puntos de muestreo para la variable clorofila "a" en el ecosistema de estudio debe ser 54, sólo se muestra que en los casos en que se tiene información de variables auxiliares como las aquí utilizadas es preferible el uso del predictor cokriging. Un estudio del número óptimo de sitios de muestreo requiere del diseño de una red de muestreo. Para lo anterior se puede consultar Giraldo *et al* (2001).

### 5.5.2. *Estudio Multivariado del Comportamiento Espacial de Variables Fisicoquímicas y Biológicas Medidas en el Estuario Ciénaga Grande de Santa Marta.*

Se emplea la información de las variables salinidad, sólidos en suspensión (mg/l), nitritos (μmol/l), silicatos (μmol/l) y clorofila "a"(μg/l), tomada de la base de datos descrita en las aplicaciones anteriores, para mostrar el uso de los componentes principales en la descripción de la distribución conjunta de estas.

- **Resultados y Discusión**

Con base en la matriz de correlación de Pearson clásica (en distancia cero) y aplicando la metodología descrita en la sección 6.4.2. se generaron los ejes factoriales. Los dos primeros componentes explican el 71% de la variabilidad contenida en las 5 variables consideradas (tabla 9). Se consideran para el análisis sólo estos dos componentes. La función de semivarianza cruzada entre los dos primeros componentes principales (Fig. 24 revela que para distancias diferentes de cero hay asociación entre los ejes principales y por consiguiente se puede concluir que no existe correlación intrínseca entre las variables originales. Lo anterior indica que un estudio exhaustivo de la correlación entre las variables

requeriría del cálculo de matrices de corregionalización para diferentes distancias. En el documento se presenta sólo el caso en que la distancia es cero, puesto que el propósito es identificar la distribución espacial de las 5 variables de forma simultánea .

Las variables de mayor importancia en la construcción del primer eje principal son salinidad (en sentido directo), silicatos y clorofila "a" (en sentido inverso). Análogamente las variables de más peso en la construcción del componente dos son sólidos en suspensión y clorofila "a" en sentido directo y nitritos en sentido inverso (tabla 10).

**Tabla 9**. Porcentajes de varianza explicados por los componentes principales generados con información de cinco variables fisicoquímicas y biológicas medidas en el estuario Ciénaga Grande de Santa Marta en marzo de 1997.

| COMPONENTE | VALOR PROPIO | % DE VARIANZA | % ACUMULADO |
|---|---|---|---|
| 1 | 2.23 | 44.610 | 44.610 |
| 2 | 1.34 | 26.962 | 71.562 |
| 3 | 0.70 | 14.124 | 85.746 |
| 4 | 0.37 | 7.426 | 93.172 |
| 5 | 0.34 | 6.828 | 100.000 |

Con la información georreferenciada de los dos componentes principales se realizaron análisis geoestadísticos y se obtuvieron mapas de distribución espacial (Figs. 25 y 26). El mapa de distribución del componente uno (Fig. 25) indica que en gran parte del sistema el nivel de salinidad está por debajo del promedio y que sería esperable encontrar allí altas concentraciones de silicatos y clorofila "a". Hacia la zona norte del sistema se dan por el contrario magnitudes de salinidad por encima de su promedio y bajas concentraciones de silicatos y clorofila "a". En resumen podría pensarse, respecto a la información aportada por el primer eje principal, que la productividad biológica puede verse favorecida por altas concentraciones de nutrientes y bajos niveles de salinidad en sitios aledaños.

La información aportada por el mapa de distribución espacial del componente dos (Fig. 26) confirma en gran medida lo descrito respecto a la distribución espacial de la clorofila "a" dentro del sistema, es decir concentraciones altas en gran parte del cuerpo de agua y bajos niveles hacia la zona norte. En este caso se puede concluir que las magnitudes altas de esta variable están asociadas a niveles altos sólidos en suspensión y , contrario a lo esperado, a bajos niveles en el ion nitrito.

Los resultados descritos de forma conjunta a través de la interpretación de las figuras 25 y 26 son completamente coherentes con los reportados de forma univariada para las variables originales en la sección 4.5.1.

**Tabla 10**. Pesos de las variables en la construcción de los dos primeros componentes principales. Información original medida en marzo de 1997 en la Ciénaga Grande de Santa Marta**.**

| VARIABLE | COMPONENTE 1 | COMPONENTE 2 |
|---|---|---|
| Salinidad | 0.5060 | 0.4150 |
| Sólidos en Suspensión | -0.3468 | 0.5668 |
| Nitritos | 0.2084 | -0.6219 |
| Silicatos | -0.5049 | -0.3334 |
| Clorofila "a" | -0.5703 | 0.9140 |

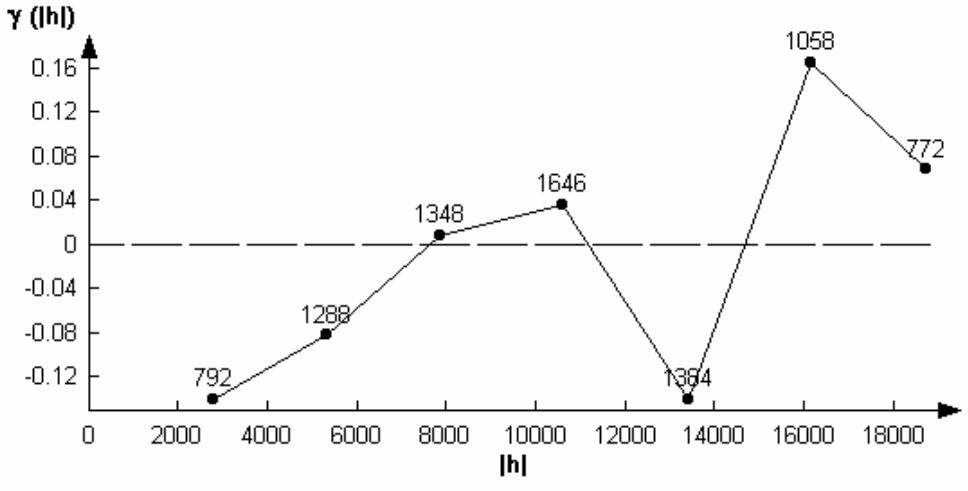



**Figura 24**. Función de semivarianza cruzada entre los dos primeros componentes principales generados con información de algunas variables fisicoquímicas y biológicas medidas en marzo de 1997 en el estuario Ciénaga Grande de Santa Marta.

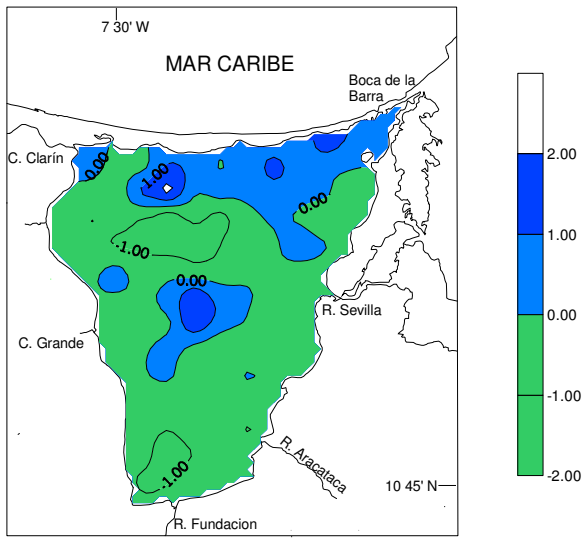



**Figura 25**. Distribución espacial del primer componente principal generado con información de variables fisicoquímicas y biológicas medidas en el estuario Ciénaga Grande de Santa Marta en marzo de 1997.

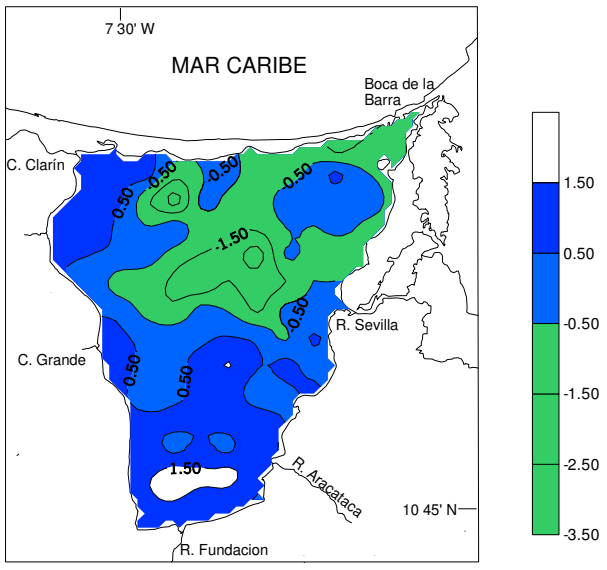



**Figura 26**. Distribución espacial del segundo componente principal generado con información de variables fisicoquímicas y biológicas medidas en el estuario Ciénaga Grande de Santa Marta en marzo de 1997.

### 5.5.3. *Diseño de una Red de Muestreo para el Estuario Ciénaga Grande de Santa Marta.*

En esta sección se hace una aplicación de la metodología propuesta por McBratney *et al* (1981) con el objetivo de diseñar una red óptima de muestreo para la CGSM (ampliamente descrita en secciones anteriores). En las dos últimas décadas la CGSM ha venido dando muestras de deterioro y por ello se han implementado algunas obras civiles que buscan su recuperación. Para el monitoreo de los cambios que se están dando en el ecosistema se hace necesario contar con un conjunto de sitios de muestreo que haga posible lograr una visión integrada del comportamiento de las principales variables que rigen sus procesos de productividad.

Se analizaron datos en la superficie de la columna de agua de las variables temperatura, salinidad, seston total, profundidad, silicatos, clorofila, oxígeno disuelto, nitritos y clorofilas "a" y "c" tomados en los mismos puntos de muestreo descritos en la sección 3.3.

Se simularon redes de muestreo con cuadrículas de 4, 9, 16, 25 y 36 $km^2$, respectivamente (Fig. 27) y se estimaron las correspondientes varianzas de predicción de cada variable en cada época tomando como base los modelos de correlación espacial estimados en la sección 3.3.

La comparación del error estándar de predicción y de los costos asociados al muestreo de cada variable en cada red, posibilitó el establecimiento de un conjunto de sitios de muestreo óptimo bajo estos dos criterios.

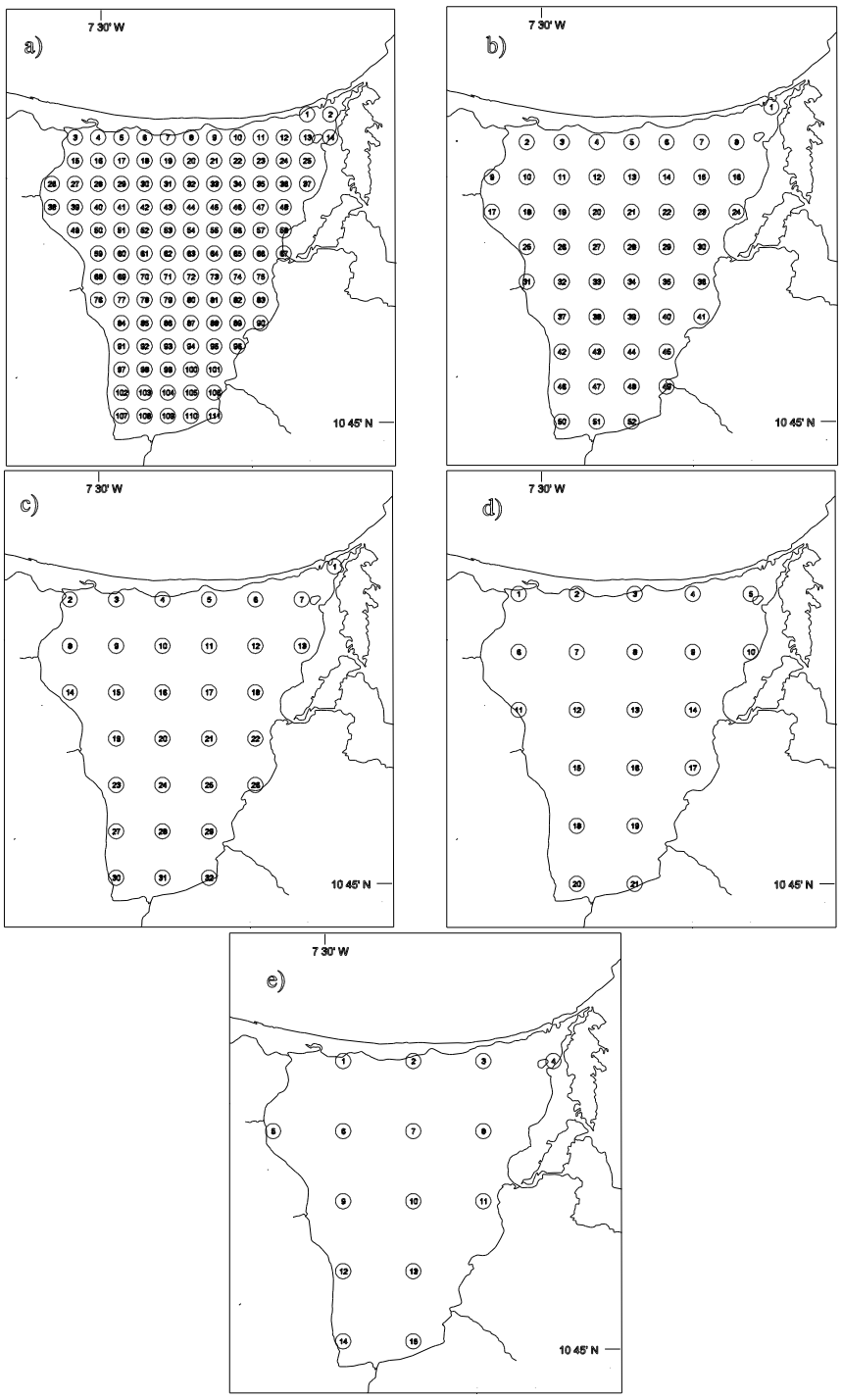



**Figura 27**. Redes de muestreo bajo las cuales se hicieron las estimaciones de las varianzas de predicción de cada una de las variables consideradas, asumiendo los modelos de semivarianza estimados. Distancias entre puntos de muestreo: a) 2000 m; b) 3000 m; c) 4000 m; d) 5000 m y e) 6000 m.

- **Resultados y Discusión**

Si bien es posible que en la fecha del muestreo se estuviese dando un fenómeno de intervención debido a los cambios climatológicos dados en el año inmediatamente anterior a este, para los propósitos del trabajo esto no resulta ser un inconveniente puesto que de hecho se asume que el establecimiento del conjunto óptimo de puntos de muestreo no depende de

la magnitud de la variables sino de la estructura de dependencia espacial presente en la región de estudio para cada una de ellas.

Como era de esperarse los errores estándar de predicción estimados son crecientes en función de la distancia que existe entre los puntos de muestreo (tabla 11). La variable en la cual se consigue mayor ganancia relativa en precisión al pasar de la red menos densa (Fig. 27 e), a la más densa (Fig. 27.a) es la salinidad. En dicha variable se consigue aumentar en un 35 % la precisión (tabla 12). Otras variables como temperatura, oxígeno disuelto, silicatos y clorofila "a" tienen aumentos en precisión que oscilan entre 15.9 y 23.8 % (tabla 12). Por último en las variables profundidad, nitritos, seston total y clorofila "c", sólo se consigue aumentar la precisión en porcentajes que están entre el 5,7 y el 10.1 % (tabla 12). Obviamente si se comparan las redes intermedias, redes con distancias entre puntos de muestreo de 3000, 4000 y 5000 m (Fig. 27.b, 27.c y 27.d), con la red de 6000 m (Fig. 27.e), resultan mucho menores los incrementos relativos en precisión (tabla 12).

**Tabla 11**. Error estándar (raíz cuadrada de la varianza) de predicción máxima de cada variable considerada para redes de muestreo con cuadrículas de 4, 9, 16, 25 y 36 $km^2$.

| Variables | TAMAÑO DE LA RED (Distancia entre los puntos de muestreo en metros) | | | | |
|---|---|---|---|---|---|
| | **2000** | **3000** | **4000** | **5000** | **6000** |
| Profundidad (m) | 0.2825 | 0.2874 | 0.2930 | 0.3002 | 0.307 |
| Temperatura (°C) | 0.6380 | 0.6690 | 0.7046 | 0.7632 | 0.8373 |
| Salinidad | 0.9096 | 1.0511 | 1.1676 | 1.2965 | 1.4075 |
| Oxígeno disuelto (mg/L) | 1.5145 | 1.5917 | 1.6752 | 1.7977 | 1.9431 |
| Seston total (mg/L) | 35.6363 | 36.4021 | 37.0459 | 37.8076 | 38.5197 |
| Nitritos (umol/L) | 0.2832 | 0.2875 | 0.2913 | 0.2958 | 0.3003 |
| Silicatos (umol/L) | 47.6524 | 50.207 | 52.3806 | 54.6797 | 56.6932 |
| Clorofila a (ug/L) | 19.4634 | 21.2041 | 22.5233 | 23.5582 | 24.2163 |
| Clorofila c (ug/L) | 6.1071 | 6.2977 | 6.4536 | 6.6336 | 6.7967 |

**Tabla 12**. Ganancia en precisión en porcentaje (cociente entre los respectivos errores estándar de predicción) de cada una de las redes de muestreo (observada y simuladas) en cada variable respecto a la red de 6000 metros (la menos densa).

| Variables | TAMAÑO DE LA RED (Distancia entre los puntos de muestreo en metros) | | | | |
|---|---|---|---|---|---|
| | **2000** | **3000** | **4000** | **5000** | **6000** |
| Profundidad (m) | 8.0 | 6.4 | 4.6 | 2.2 | 0 |
| Temperatura (°C) | 23.8 | 20.1 | 15.8 | 8.8 | 0 |
| Salinidad | 35.4 | 25.3 | 17.0 | 7.9 | 0 |
| Oxígeno disuelto (mg/L) | 22.1 | 18.1 | 13.8 | 7.5 | 0 |
| Seston total (mg/L) | 7.5 | 5.5 | 3.8 | 1.8 | 0 |
| Nitritos (umol/L) | 5.7 | 4.3 | 3.0 | 1.5 | 0 |
| Silicatos (umol/L) | 15.9 | 11.4 | 7.6 | 3.6 | 0 |
| Clorofila a (ug/L) | 19.6 | 12.4 | 7.0 | 2.7 | 0 |
| Clorofila c (ug/L) | 10.1 | 7.3 | 5.0 | 2.4 | 0 |

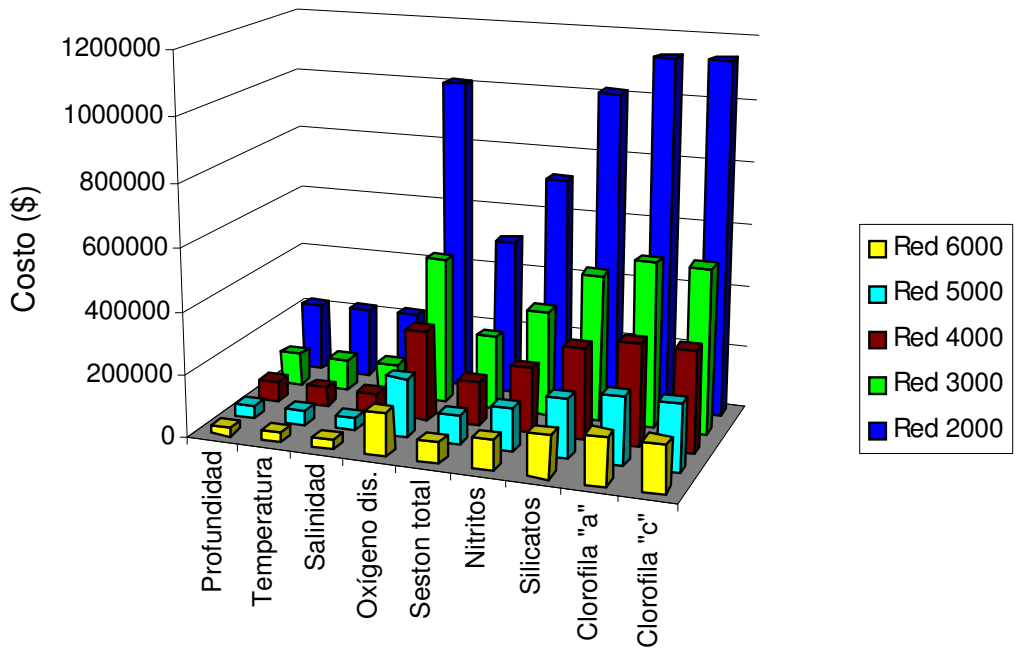



**Figura 28**. Costos de muestreo de variables fisicoquímicas y biológicas en la Ciénaga Grande de Santa Marta, según diferentes espaciamientos entre sitios de muestreo (se asumen muestreos sistemáticos de cuadrículas).

De otro lado si se estudian los costos de muestreo asociados a cada variable bajo cada una de las densidades muestrales (Fig. 28), se observa que existe considerable diferencia, con excepción de las variables temperatura, profundidad y salinidad, entre la red de 2000 m y las restantes respecto a dichos costos. Para algunas de las variables (oxígeno disuelto, silicatos y clorofilas) pasar de la red de 3000 m a la 2000 m, implica incrementar el costo de muestreo de cada una de ellas en más de $300000

En conclusión para las variables temperatura y salinidad sería mucho más conveniente hacer un muestreo intensivo (red más densa) dado que se consigue, comparando con la red menos densa, aumentar la eficiencia en porcentajes considerables (23 y 35%, respectivamente, tabla 12), con costos netos que se incrementan sólo alrededor de $100000 (Fig. 28). En la variable profundidad, si bien los costos de muestreo no se incrementan significativamente (Fig. 28), es más aconsejable muestrear en la red menos densa dado que la eficiencia se incrementa en máximo un 7% al compararla con las restantes redes (tabla 12). En las variables nitritos, seston total y clorofilas "a" y "c" hay poco aumento de la eficiencia al pasar de la red de 6000 m a otras con mayor número de puntos (tabla 12) y por el contrario los costos, especialmente en la red de 2000 m, tienen aumentos considerables, lo que hace que se planteen las redes menos densas (5000 m y 6000 m entre puntos de muestreo) como las más adecuadas para el seguimiento de estas variables. En las restantes variables (oxígeno disuelto, silicatos y clorofila “a”) es un poco más compleja la decisión dado que se obtienen aumentos considerables en los costos (Fig. 28), pero también incrementos de eficiencia (tabla 12). De todas formas es claro que se debe descartar en este caso la red con distancias entre puntos de muestreo de 2000 m dado que entre esta y la red de 3000 m, la eficiencia relativa aumenta en un máximo del 8 % (tabla 12) con costos que se duplican o triplican para algunas variables (Fig. 28).

# *Apéndice*

## 6.1. Indicador $IGC_i(P)$.

Suponga que se tiene información dicotómica sobre P variables en n sitios de muestreo. La estructura de esta información se presenta a continuación:

| *Sitio* | $X$ | $Y$ | $V_1$ | $V_2$ | $\cdots$ | $V_P$ |
|---|---|---|---|---|---|---|
| *1* | $x_1$ | $y_1$ | *0 ó 1* | *0 ó 1* | $\cdots$ | *0 ó 1* |
| *2* | $x_2$ | $y_2$ | *0 ó 1* | *0 ó 1* | $\cdots$ | *0 ó 1* |
| *3* | $x_3$ | $y_3$ | *0 ó 1* | *0 ó 1* | $\cdots$ | *0 ó 1* |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| *n* | $x_n$ | $y_n$ | *0 ó 1* | *0 ó 1* | $\cdots$ | *0 ó 1* |

donde $X$ y $Y$ representan las correspondientes coordenadas de ubicación geográfica (grados, planas o cartesianas), $V_1, \ldots, V_p$, son P variables que indican la presencia (1) o ausencia (0) de la característica. Si las variables observadas son cuantitativas se pueden categorizar baja diversos criterios. Uno de los más empleados es el de comparación de cada valor con su correspondiente mediana muestral.

La información referente a tales variables puede presentarse en una matriz de la siguiente estructura:

$$\begin{pmatrix} \eta_{11} & \eta_{12} & \cdots & \eta_{1P} \\ \eta_{21} & \eta_{22} & \cdots & \eta_{2P} \\ \vdots & \vdots & \vdots & \vdots \\ \eta_{n1} & \eta_{n2} & \cdots & \eta_{nP} \end{pmatrix}, \qquad \eta_{ij} = \begin{cases} 1 & \text{Si la variable } j \text{ en el sitio } i \text{ cumple} \\ & \text{el criterio de dicotomización} \\ 0 & \text{en otro caso} \end{cases}$$

Se define el indicador del número de "éxitos" por sitio como:

$$\sum_{j=1}^{P} \eta_{ij} = \eta_{i\bullet}, \quad \forall i, i = 1, \cdots, n$$

y el indicador del número de "exitos" par la variable $j$ como:

$$\sum_{j=1}^{n} \eta_{ij} = \eta_{\bullet j}, \ \forall\, j, j = 1,2,\cdots,P$$

Además sea:

$$K_i = \sum_{j=1}^{P} \delta_{ij}, \quad donde \;\; \delta_{ij} = \begin{cases} j & Si\, \eta_{ij} = 1 \\ \eta_{i\bullet} - P & Si\, \eta_{ij} = 0 \end{cases}$$

El indicador $IGC_i(P)$ en el sitio *i*-ésimo se calcula como:

$$IGC_i(P) = P(P-2) + \eta_{i\bullet} + K_i - h(P, \eta_{i\bullet})$$

donde:

$$h(P, \eta_{i\bullet}) = \begin{cases} -2P & Si\ \eta_{i\bullet} = 0 \\ 0 & Si\ \eta_{i\bullet} = 1 \\ (a_m - b_m P) & Si\ m = \eta_{i\bullet} - 2 \;\; y \;\; \eta_{i\bullet} = 2, 3, \cdots, P \end{cases}$$

con :

$$a_0 = 0$$
$$a_m = a_{m-1} + (m^2 + m), \qquad m = 1, 2, \cdots$$

y

$$b_0 = -1$$
$$b_m = b_{m-1} + (m-1), \qquad m = 1, 2, \cdots$$

En la tabla 13 se presentan los valores los valores de $a_m$, $b_m$ y $h(p, \eta_{i\bullet})$ necesarios para el cálculo de la variable $IGC_i(P)$ cuando el número de presencias de especies está entre 2 y 15.

**Tabla 13**. Valores de $a_m$ y $b_m$ en la ecuación 6 para $\eta_{i\bullet}$ = 2, 3, . . ., 15. P es el número de variables de tipo presencia - ausencia consideradas.

| $\eta_{i\bullet}$ | $m$ | $a_m$ | $b_m$ | $(a_m - b_m P)$ |
|---|---|---|---|---|
| 2 | 0 | 0 | -1 | $p$ |
| 3 | 1 | 2 | -1 | $(2+p)$ |
| 4 | 2 | 8 | 0 | $8$ |
| 5 | 3 | 20 | 2 | $(20-2p)$ |
| 6 | 4 | 40 | 5 | $(40-5p)$ |
| 7 | 5 | 70 | 9 | $(70-9p)$ |
| 8 | 6 | 112 | 14 | $(112-14p)$ |
| 9 | 7 | 168 | 20 | $(168-20p)$ |
| 10 | 8 | 240 | 27 | $(240-27p)$ |
| 11 | 9 | 330 | 35 | $(330-35p)$ |
| 12 | 10 | 440 | 44 | $(440-44p)$ |
| 13 | 11 | 572 | 54 | $(572-54p)$ |
| 14 | 12 | 728 | 65 | $(728-65p)$ |
| 15 | 13 | 910 | 77 | $(910-77p)$ |

Para ilustrar el cálculo de la variable IGC(P), se presenta en la tabla 14 el caso particular de P = 4. Se utilizan todos los posibles arreglos de ceros y unos que se pueden formar con 4 datos.

**Tabla 14**. Cálculo de los valores de la variable $IGC_i(4)$ con todos los posibles arreglos formados con 4 símbolos de dos tipos (ceros y unos).

| ARREGLO | $\eta_{i1}$ | $\eta_{i2}$ | $\eta_{i3}$ | $\eta_{i4}$ | $\eta_{i\bullet}$ | $\delta_{i1}$ | $\delta_{i2}$ | $\delta_{i3}$ | $\delta_{i4}$ | $K_i$ | $P(P-2)$ | $IGC(P)_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | -4 | -4 | -4 | -4 | -16 | 8 | 0 |
| 2 | 1 | 0 | 0 | 0 | 1 | 1 | -3 | -3 | -3 | -8 | 8 | 1 |
| 3 | 0 | 1 | 0 | 0 | 1 | -3 | 2 | -3 | -3 | -7 | 8 | 2 |
| 4 | 0 | 0 | 1 | 0 | 1 | -3 | -3 | 3 | -3 | -6 | 8 | 3 |
| 5 | 0 | 0 | 0 | 1 | 1 | -3 | -3 | -3 | 4 | -5 | 8 | 4 |
| 6 | 1 | 1 | 0 | 0 | 2 | 1 | 2 | -2 | -2 | -1 | 8 | 5 |
| 7 | 1 | 0 | 1 | 0 | 2 | 1 | -2 | 3 | -2 | 0 | 8 | 6 |
| 8 | 0 | 1 | 1 | 0 | 2 | -2 | 2 | 3 | -2 | 1 | 8 | 7 |
| 9 | 1 | 0 | 0 | 1 | 2 | 1 | -2 | -2 | 4 | 1 | 8 | 7 |
| 10 | 0 | 1 | 0 | 1 | 2 | -2 | 2 | -2 | 4 | 2 | 8 | 8 |
| 11 | 0 | 0 | 1 | 1 | 2 | -2 | -2 | 3 | 4 | 3 | 8 | 9 |
| 12 | 1 | 1 | 1 | 0 | 3 | 1 | 2 | 3 | -1 | 5 | 8 | 10 |
| 13 | 1 | 1 | 0 | 1 | 3 | 1 | 2 | -1 | 4 | 6 | 8 | 11 |
| 14 | 1 | 0 | 1 | 1 | 3 | 1 | -1 | 3 | 4 | 7 | 8 | 12 |
| 15 | 0 | 1 | 1 | 1 | 3 | -1 | 2 | 3 | 4 | 8 | 8 | 13 |
| 16 | 1 | 1 | 1 | 1 | 4 | 1 | 2 | 3 | 4 | 10 | 8 | 14 |

Se observa en la tabla anterior que el $IGC_i(4)$ es una variable discreta monótona creciente con valores entre 0 y 14. Para este caso los valores del indicador definen puntualmente, excepto cuando el $IGC_i(4)$ igual a 7, lo ocurrido respecto al número de unos y a la posición de los mismos dentro de las sucesiones. Valores del $IGC_i(4)$ entre 1 y 4 indican que hubo un solo uno en la sucesión y cada número revela la posición que éste ocupa en la misma. Valores entre 5 y 9 corresponden a sucesiones en las que hubo dos unos, con $IGC_i(4)$ igual a 5 cuando los dos unos están en las primeras dos posiciones de las sucesión ordenada y a 9 cuando están en las dos últimas. Los valores 6, 7 y 8 reflejan la transición de las dos primeras a las dos últimas posiciones. Valores entre 10 y 13 indican que hubo tres unos y cada uno de estos valores corresponde a una única sucesión. El valor del $IGC_i(4)$ será 10 cuando los unos estén en las tres primeras posiciones e igual a 13 cuando estén en las tres ultimas. Los valores 0 y 14 se obtendrán cuando en la sucesión no haya unos o todos los valores sean iguales a uno, respectivamente.

## 6.2. Álgebra de Matrices.

La gran mayoría de métodos estadísticos, incluyendo la geoestadística, pueden ser tratados de forma mucho más sencilla a través del uso del álgebra de matrices. Por ésta razón es útil, si no esencial, tener un cierto conocimiento mínimo de ésta área de las matemáticas. Lo anterior es cierto siempre y cuando el interés sea usar los métodos como una herramienta. La notación del álgebra matricial algunas veces puede resultar desanimante. Sin embargo, no es difícil entender sus principios básicos.

### 6.2.1. Matriz

Una matriz A de tamaño (mxn) es un arreglo rectangular de m filas con n columnas.

$$A = \begin{pmatrix} a_{11} & a_{12} & . & . & . & a_{1n} \\ a_{21} & a_{22} & . & . & . & a_{2n} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ a_{m1} & a_{m2} & . & . & . & a_{mn} \end{pmatrix}$$

### 6.2.2. Suma y Producto de Matrices

El procesos aritmético de adición, sustracción, multiplicación y división tiene sus contraparte con matrices. Si A y D son dos matrices de orden 3x2, entonces su suma se define como:

$$A + D = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} + \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{32} & d_{23} \end{pmatrix} = \begin{pmatrix} a_{11}+d_{11} & a_{12}+d_{12} \\ a_{21}+d_{21} & a_{22}+d_{22} \\ a_{31}+d_{31} & a_{32}+d_{32} \end{pmatrix}$$

En el caso de la multiplicación se debe cumplir que el número de columnas de la primera matriz sea igual ala número de filas de la segunda.

$$A \bullet B = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \bullet \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix} = \begin{pmatrix} \sum a_{1i}b_{i1} & \sum a_{1i}b_{i2} & \sum a_{1i}b_{i3} \\ \sum a_{2i}b_{i1} & \sum a_{2i}b_{i2} & \sum a_{2i}b_{i3} \\ \sum a_{3i}b_{i1} & \sum a_{3i}b_{i2} & \sum a_{3i}b_{i32} \end{pmatrix}$$

### 6.2.3. Inversa y Determinante de una Matriz.

Si *k* es un número, es cierto que *k* x $k^{-1}$ = 1. De forma similar si A es una matriz cuadrada (número de filas igual al número de columnas) su inversa es $A^{-1}$, donde $AA^{-1}$ = $A^{1}A$ = I, con I igual a la matriz idéntica (matriz de unos en la diagonal y cero por fuera de ella). Un ejemplo de matriz inversa es:

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix}$$

Esto puede comprobarse observando que:

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \bullet \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

la inversa de una matriz 2x2, si existe, puede determinarse fácilmente por medio del siguiente cálculo:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = \begin{pmatrix} a_{22}/\Delta & -a_{12}/\Delta \\ -a_{21}/\Delta & a_{11}/\Delta \end{pmatrix}$$

Donde $\Delta = a_{11}a_{22} - a_{12}a_{21}$. La cantidad $\Delta$ es llamada el ***determinante*** de la matriz. Claramente la inversa no está definida si el determinante es igual a cero. Con matrices grandes el cálculo de la inversa es tedioso y se debe usar un programa de computo para realizarlo.

### 6.2.4. Valores y Vectores Propios.

Dada una matriz **A** de orden (n x n)**,** si existe un vector ***x*** (n x 1) y un número $\lambda$ tal que

$$Ax = \lambda x. \quad \text{ó} \quad (A - \lambda I)x = 0$$

donde I es la matriz idéntica de orden (n x n) y 0 es un vector (n x 1), entonces se llama a $\lambda$ y ***x***, respectivamente, *valor y vector propio* de la matriz A. Pueden encontrarse hasta n valores propios y hay tantos vectores propios como valores propios se encuentren. Los valores de $\lambda$ deben satisfacer que el determinante de $A - \lambda I = 0$**.** Los vectores propios se calculan después de reemplazar los valores propios encontrados en la expresión $Ax = \lambda x$. Al igual que con la inversa, para matrices grandes se debe emplear un software especializado para su obtención. A continuación, a manera de ilustración, se realiza el cálculo de los vectores y valores propios de una matriz de orden 2 x 2.

Sea $A = \begin{pmatrix} 6 & 3 \\ 3 & 4 \end{pmatrix}$, entonces

$$|A - \lambda I| = 0 \Rightarrow \left| \begin{pmatrix} 6 & 3 \\ 3 & 4 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0$$

$$\left| \begin{pmatrix} 6 & 3 \\ 3 & 4 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right| = 0$$

$$\begin{vmatrix} (6-\lambda) & 3 \\ 3 & (4-\lambda) \end{vmatrix} = 0$$

$$(6-\lambda)(4-\lambda) - 9 = 0$$

$$\lambda^2 - 10\lambda + 15 = 0$$

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$\lambda = \frac{-(-10) \pm \sqrt{100 - 4(15)}}{2} = \frac{10 \pm \sqrt{40}}{2}$$

$$\lambda = 8.1623, \quad \lambda = 1.8377$$

Para cada valor propio existe un vector propio, el cual se obtiene reemplazando el valor propio correspondiente en la primera expresión de la página anterior y usando la condición de que los respectivos vectores propios estén normalizados.

Un vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ se dice que está normalizado si satisface que $\sqrt{x_1^2 + x_2^2} = 1$.

Teniendo en cuenta lo anterior se calculan los vectores propios de la siguiente forma:

$(A - \lambda I)x = 0$

$$\begin{pmatrix} (6-\lambda) & 3 \\ 3 & (4-\lambda) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$(6-\lambda)x_1 + 3x_2 = 0$$
$$3x_1 + (4-\lambda)x_2 = 0$$

Restando las dos ecuaciones anteriores y factorizando, obtenemos:

$$x_1(6-\lambda-3) + x_2(3-4+\lambda) = 0$$
$$x_1(3-\lambda) + x_2(-1+\lambda) = 0$$

$$\boxed{x_1 = \frac{(1-\lambda)x_2}{(3-\lambda)}}$$

Entonces para $\lambda = 8.1623$ y $\lambda = 1.8377$ se tiene respectivamente:

$x_1 = 1.3847x_2$ y $x_1 = -0.7207x_2$ . Ahora utilizando la restricción de que los vectores estén normalizados se obtiene:

$$x_1^2 = (1.3847)^2\left(1 - x_1^2\right)$$
$$x_1^2 + (1.3847)^2 x_1^2 = (1.3847)^2$$
$$x_1^2\left(1 + 1.3847^2\right) = (1.3847)^2$$
$$x_1^2 = \frac{(1.3847)^2}{\left(1 + 1.3847^2\right)} \quad \Rightarrow \quad x_1 = \frac{1.3847}{\sqrt{\left(1 + 1.3847^2\right)}} = 0.8107$$

Reemplazando el valor de $x_1$, obtenemos que $x_2 = \frac{x_1}{1.3847} = \frac{0.8107}{1.3847} = 0.5855$.

Luego el vector propio asociado al valor propio $\lambda = 8.1623$ es $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.8107 \\ 0.5855 \end{pmatrix}$

Efectuando un procedimiento similar se puede comprobar que el vector propio asociado al valor propio $\lambda = 1.8377$ es $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -0.5847 \\ 0.8113 \end{pmatrix}$

En resumen dada la matriz del ejemplo entonces se puede comprobar que:

$$\left[\begin{pmatrix} 6 & 3 \\ 3 & 4 \end{pmatrix} - \begin{pmatrix} 8.1623 & 0 \\ 0 & 8.1623 \end{pmatrix}\right]\begin{pmatrix} 0.8107 \\ 0.5855 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

y, con el segundo valor y vector propio, que

$$\left[\begin{pmatrix} 6 & 3 \\ 3 & 4 \end{pmatrix} - \begin{pmatrix} 1.8377 & 0 \\ 0 & 1.8377 \end{pmatrix}\right]\begin{pmatrix} -0.5847 \\ 0.8113 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

### 6.3. Conceptos de Probabilidad

A continuación se presenta una revisión no exhaustiva y a manera introductoria de conceptos básicos de la teoría de probabilidades. Un estudio profundo y formal de estos se puede hacer en Mood *et al* (1963) .

#### 6.3.1. Variable Aleatoria

Si X es una función que le asigna a cada uno de los resultados de un experimento aleatorio (aquel cuya respuesta no puede ser establecida de antemano) un número real, entonces X se llama una *Variable Aleatoria*. Estas pueden ser discretas o continuas.

#### 6.3.2. Función de Probabilidad

Si X es una variable aleatoria discreta. Se llamará a f(x) = P (X = x) función de probabilidad de la variable aleatoria X, si satisface las siguientes propiedades:

i. $f(x) \geq 0 \quad \forall x \in R_X$

ii. $\sum_x f(x) = 1$.

Si existe una función f(x) tal que:

i. $f(x) \geq 0\ ,\ -\infty < x < \infty$

ii. $\int_{-\infty}^{\infty} f(x)dx = 1$

iii. $P(a < X\ b) = \int_a^b f(x)dx$ para cualquier a y b, entonces f(x) es la función de densidad de probabilidad de la variable aleatoria continua X.

La función de probabilidad acumulada, notada como F(x), es igual a $P(X \le x)$ y se evalúa a través de una sumatoria o de una integral dependiendo de si X es discreta o continua.

### 6.3.2.1. Valor Esperado y Varianza

Si X es una variable aleatoria, el valor esperado de una función de la variable aleatoria X, $g(X)$ está dado por:

$$E(g(X)) = \begin{cases} \sum_x g(x)f(x) & X \text{ discreta} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & X \text{ continua} \end{cases}$$

como caso particular,

$$E(X) = \mu = \begin{cases} \sum_x xf(x) & X \text{ discreta} \\ \int_{-\infty}^{\infty} xf(x)dx & X \text{ continua} \end{cases}$$

La varianza de la variable aleatoria X está definida como:

$$V(X) = \sigma^2 = E(X-\mu)^2 = \begin{cases} \sum_x (x-\mu)^2 f(x) & X \text{ discreta} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & X \text{ continua} \end{cases}$$

La raíz cuadrada de la varianza se denomina desviación estándar y se denota por $\sigma$.

Se cumple que:

1. $E(aX) = aE(X)$, con *a* constante
2. $E(aX+b) = aE(X)+b$, con *a* y *b* constantes
3. $V(aX) = a^2V(X)$ y *a* constante
4. $V(X) = E(X^2) - [E(X)]^2$

### 6.3.2.2. Función de Probabilidad Binomial y Normal.

**Modelo Binomial**

Suponga que hay un experimento que consiste en examinar *n individuos* y evaluar o medir en cada uno de ellos si tienen o no una característica dada (sólo hay dos posibles resultados).Sea *p* la probabilidad de ¨éxito¨ y *q = 1-p* la de ¨fracaso¨ en cada uno de los n ensayos. Se asume que esta probabilidad es constante en cada uno de ellos.

Sea X = Número de éxitos en los n ensayos, entonces asumiendo conocido p entonces es posible establecer las probabilidades de ocurrencia de cada evento mediante la siguiente ecuación, denominada *modelo de probabilidad binomial*:

$$P(X=x)=\binom{n}{x}p^{x}\,(1-p)^{n-x} \qquad x=0,1,2,\ldots,\mathrm{n}$$

En este modelo:

$$\mu=E(X)=np$$
$$\sigma^{2}=V(X)=np(1-p)$$

**Modelo Normal**

El modelo de probabilidad normal (Gaussiano) es útil para encontrar las probabilidades asociadas a eventos de variables aleatorias cuyas distribuciones de frecuencias son simétricas alrededor del valor promedio. Algunos ejemplos de este tipo de variables aleatorias son los siguientes:

Sea $\mu$ el valor promedio de la variable (E(X)) y $\sigma^2$ su correspondiente varianza (V(X)), entonces las probabilidades de ocurrencia de eventos asociados a los posibles resultados de la variable estudiada pueden ser encontrados usando la siguiente expresión, llamada *modelo de probabilidad normal*:

$$P(a\le X\le b)=\int_{a}^{b}\frac{1}{\sqrt{2\pi}\sigma}e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^{2}}dx.$$

Obviamente resultaría muy dispendioso tener que calcular estas integrales para cada valor de a, b, $\mu$ y $\sigma$ . Por esta razón se acude a un procedimiento llamado estandarización, el cuál consiste en hacer la transformación $Z=\frac{X-\mu}{\sigma}$. La variable anterior tendrá (si la distribución de frecuencias de X se ajusta a un modelo de probabilidad normal con media $\mu$ y varianza $\sigma^2$) una distribución de frecuencias que se ajusta a un modelo de probabilidad normal con media cero y varianza uno, es decir que:

$$P(a\le X\le b)=\left(\left(\frac{a-\mu}{\sigma}\right)\le Z\le\left(\frac{b-\mu}{\sigma}\right)\right)=(z_1<Z<z_2)=\int_{z_1}^{z_2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^{2}}$$

La ecuación anterior también puede resultar difícil de evaluar, sin embargo para cualquier valor de a, b, $\mu$ y $\sigma$ las correspondientes probabilidades pueden hallarse, sin necesidad de resolver la integral, empleando la *tabla de distribución acumulada normal estándar* que aparece en los textos de estadística.

### 6.3.3. Función de Probabilidad Bivariada.

Si X y Y son dos variables aleatorias discretas. La probabilidad de X = x y Y = y está determinada por la función de probabilidad bivariada $f(x,y)=P[X=x,Y=y]$ donde :

i. $f(x,y)\ge 0,\forall\, x,y\in R_X,R_Y$

ii. $\sum_{x}\sum_{y}f(x,y)=1$

Si existe una función $f(x,y)$ tal que la probabilidad conjunta:

$$P[a < X < b, c < Y < d] = \int_a^b \int_c^d f(x,y)dydx$$

para cualquier valor de a, b, c y d en donde $f(x,y) \geq 0$, $-\infty < x, y < \infty$ y $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dydx = 1$, entonces $f(x,y)$ es la función de probabilidad bivariada de X y Y.

La función de probabilidad acumulada $F(x,y)$ es igual a $P[X \leq x, Y \leq y]$ y se evalúa a través de una doble sumatoria o de una doble integral dependiendo de si las variables aleatorias son discretas o continuas, respectivamente.

### 6.3.3.1. Función de Probabilidad Marginal

Si X y Y son dos variables aleatorias con función de probabilidad conjunta $f(x,y)$. Las funciones de probabilidad marginales de Y y Y están dadas por

$$f(x) = \sum_y f(x,y)$$
$$f(y) = \sum_x f(x,y)$$

si X y Y son variables aleatorias discretas

ó por

$$f(x) = \int_{-\infty}^{\infty} f(x,y)dy$$
$$f(y) = \int_{-\infty}^{\infty} f(x,ydx)$$

si X y Y son variables aleatorias continuas

### 6.3.3.2. Función de Probabilidad Condicional

Sean X y Y dos variables aleatorias con función de densidad conjunta $f(x,y)$. La función de probabilidad condicional de la variable aleatoria X, denotada por $f(x/y)$, para un valor fijo y de Y, está definida por:

$f(x/y) = \frac{f(x,y)}{f(y)}$, donde $f(y)$ es la función de probabilidad marginal de Y de manera tal que $f(y) > 0$ .

De manera análoga, la función de probabilidad condicional de Y para un valor fijo x de X se define como:

$f(y/x)=\frac{f(x,y)}{f(x)}$, donde $f(x)$ es la función de probabilidad marginal de X de manera tal que $f(x)>0$ .

### 6.3.3.3. Independencia Estadística.

Sean X y Y dos variables aleatorias con función de densidad conjunta $f(x,y)$. X y Y son independientes si y sólo si:

$$f(x,y)=f(x)f(y)$$

donde $f(x)$ y $f(y)$ son las funciones de probabilidad marginales.

### 6.3.3.4. Valor Esperado, Varianza y Covarianza

Sean X y Y dos variables aleatorias que se distribuyen conjuntamente. El valor esperado de una función de X y Y, $g(x,y)$, se define como:

$$E(g(X,Y))=\begin{cases}\sum_x\sum_y g(x,y)f(x,y) & si\ X\ y\ Y\ son\ discretas\\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dydx & si\ X\ y\ Y\ son\ continuas\end{cases}$$

La covarianza entre X y Y, denotada por Cov (X, Y), se define como:

$$E[(X-\mu_X)(Y-\mu_Y)]=E(XY-X\mu_Y-Y\mu_X+\mu_X\mu_Y)=E(XY)-E(X)E(Y)$$

donde $\mu_X$ y $\mu_Y$ representan los valores esperados de X y Y respectivamente.

Si la covarianza de X y Y se divide por el producto de las desviaciones estándar de X y Y, el resultado es una cantidad sin dimensiones que recibe el nombre de coeficiente de correlación y se denota por $\rho(X,Y)$.

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

#### 6.3.3.4.1. Propiedades del Valor Esperado y la Varianza.

Si X y Y son dos variables aleatorias con densidad conjunta, entonces se cumple que:

1. $E(X+Y)=E(X)+E(Y)$
2. $V(X\pm Y)=V(X)+V(Y)\pm 2Cov(X,Y)$

3. $V\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(X_i, X_j)$.

Observación: $Cov(X_i, X_j) = Cov(X_j, X_i)$ y $Cov(X_i, X_i) = V(X_i)$

Como caso particular:

$$V(a_1 X_1 \pm a_2 X_2) = a_1^2 V(X_1) + a_2^2 V(X_2) \pm 2Cov(X_1, X_2)$$

3. Si $E(X) = E(Y)$, entonces $\frac{1}{2}E\left[(X-Y)^2\right] = \frac{1}{2}V(X) + \frac{1}{2}V(Y) - Cov(X,Y)$.

## 6.4. Algunos Métodos Estadísticos.

### 6.4.1. Regresión Simple

En el modelo de regresión simple se establece una relación lineal entre la esperanza condicional de una variable aleatoria Y dados unos valores fijos de una variable X.

**Modelo Poblacional**

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$E(Y / X_i) = \hat{Y}_i = \beta_0 + \beta_1 x_i$$

$Y_i$ : i-ésimo valor de la variable respuesta o dependiente en la población
$x_i$ :i-ésimo valor de la variable predictora o independiente en la población
$\beta_0$ y $\beta_1$ son parámetros poblacionales que representan el intercepto y la pendiente, respectivamente
$\varepsilon_i$ : i-ésimo error aleatorio en la población.

**Supuestos del Modelo.**

1. $E(\varepsilon_i) = 0$
2. $V(\varepsilon_i) = \sigma^2$
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$
4. $\varepsilon_i \sim N(0, \sigma^2)$

**Modelo Muestral**

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$
$$y_i = \hat{y}_i + e_i$$

$y_i$ : i-ésimo valor de la variable respuesta en la muestra
$x_i$ : i-ésimo valor de la variable predictora .

$\hat{\beta}_0$ *y* $\hat{\beta}_1$ son las estimaciones de los parámetros con base en la información muestral.

$e_i$ : i-ésimo erro muestral.

**Estimación de** $\beta_0$ *y* $\beta_1$

Uno de los métodos de estimación de los parámetros es el de mínimos cuadrados, que consiste en encontrar los estimadores que hacen mínima la suma de cuadrados de los errores, es decir aquellos valores que hacen más pequeña $\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ .

$\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2$ . Derivando e igualando a cero se obtiene:

$$\frac{\partial\sum_{i=1}^{n}\varepsilon_i^2}{\partial\beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) = 0 \quad \text{y} \quad \frac{\partial\sum_{i=1}^{n}\varepsilon_i^2}{\partial\beta_1} = -2\sum_{i=1}^{n}X_i(Y_i - \beta_0 - \beta_1 x_i) = 0.$$

Al simplificar las dos ecuaciones anteriores y distribuir las sumas se tiene:

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i Y_i = \beta_0\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} x_i^2$$

Las dos ecuaciones anteriores se conocen como ecuaciones normales. Dadas las realizaciones $y_1, y_2, \ldots, y_n$ las ecuaciones pueden resolverse para encontrar los estimados de los parámetros:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n} x_i$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\sum_{i=1}^{n} x_i y_i = \left(\bar{y} - \hat{\beta}_1\bar{x}\right)\sum_{i=1}^{n} x_i + \hat{\beta}_1\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = \left(\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\right)\sum_{i=1}^{n} x_i + \hat{\beta}_1\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = \frac{\sum_{i=1}^{n} y_i * \sum_{i=1}^{n} x_i}{n} - \hat{\beta}_1 \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} y_i * \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Se puede demostrar que los errores estándar estimados de los estimadores de los parámetros corresponden a:

$$s(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \quad \text{y} \quad s(\hat{\beta}_0) = s\left(\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} (x_i - \bar{x})^2}}\right), \text{ con } \quad s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

### 6.4.2. Análisis de Componentes Principales.

El análisis de componentes principales es uno de los métodos multivariados más difundidos, que permite la estructuración de un conjunto de datos de múltiples variables de una población, cuya distribución de probabilidades no necesita ser conocida (Lebart *et al.*, 1995).

Se trata de una técnica matemática que no requiere un modelo estadístico para explicar la estructura probabilística de los errores. Sin embargo, si es posible suponer que la población muestreada tiene una distribución conjunta normal multivariada, podrá estudiarse la significación estadística de los componentes y será posible utilizar la muestra efectivamente observada para efectuar pruebas de hipótesis, que contribuyan a conocer la estructura de la población original, con un cierto grado de confiabilidad, fijado *a priori* o *a posteriori* (Pla, 1986).

Los objetivos más importantes del análisis de componentes principales son:

i. Generar nuevas variables que puedan expresar la información contenida en el grupo original de datos.

ii. Reducir la dimensionalidad del problema que se está estudiando, como paso previo para futuros análisis.

iii. Eliminar, cuando sea posible, algunas variables originales, en el caso de que aporten poca información.

Este análisis se basa en una transformación lineal de las observaciones originales. Esta transformación es conocida en el campo del álgebra vectorial como generación de vectores y valores propios. Las nuevas variables generadas se llaman *componentes principales* y poseen algunas características estadísticas deseables, tales como la independencia (cuando se asume la multinormalidad) y en todos los casos la no correlación. Esto significa que si las variables originales no están correlacionadas, los componentes principales no ofrecen ventaja alguna.

Generación de los Componentes Principales

Se ha dicho que los componentes principales tienen ciertas características que son deseables:

*a)* Los componentes principales *no están correlacionados.*

b) Cada componente principal sintetiza la máxima variabilidad residual contenida en los datos. Es decir, el primer componente sintetiza la máxima variabilidad posible en el conjunto de datos originales; el segundo componente sintetiza la máxima variabilidad restante, sujeta a la condición de no correlación con el primer componente, y así hasta el p-ésimo componentes.

c) Cada componente contiene información de todas las variables pero en diferentes proporciones.

Matricialmente se expresa la generación de los componentes a través de:

$$Y_{(nxp)} = X_{(nxp)} L_{(pxp)} D^{-1}_{(pxp)}$$

donde :

Y: Matriz cuyas columnas representan las nuevas variables (componentes principales). Estas tienen la propiedad de ser no correlacionadas.

X: Matriz de datos originales

L: Matriz de vectores propios de: a) $X^TX$, si X es la matriz de datos originales; b) S (matriz de varianzas y covarianzas) si X es centrada; c) R (matriz de correlación) si X está estandarizada.

D: Matriz diagonal con valores en la diagonal iguales a la raiz cuadrada de los valores propios de $X^TX$ , S o R.

La transformación lineal para generar los componentes principales (matriz Y) se fundamenta en el proceso de diagonalización de una matriz, $X^TX$ , S o R., según el caso, a través del teorema de descomposición del valor singular

# Referencias


Anderson, T. W. 1984. An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, New York.

Biau, G., E. Zorita, H. von Storch & H. Wackernagel. 1997. Estimation of precipitation by kriging in EOF space. GKSS, 97, E45.

Box, G. E. P. y G.M. Jenkins. (1976). Time Series Analysis Forecasting and Control. Holden -Day, San Francisco.

Bogaert, P., P. Mahau & F. Beckers. 1995. The Spatial Interpolation of Agroclimatic Data. Cokriging Software and Source Code. FAO, Rome.

Bula-Meyer, G. 1985. Un nuevo núcleo de surgencia en el Caribe colombiano detectado en correlación con las macroalgas. Bol. Ectrópica 12:3-25.

Carr, J. , D. Myers y Ch. Glass. 1985. Cokriging - A Computer Program. Computers & Geosciences. 11(2), 111-127.

Clark, I. 1979. Practical Geostatistics. Elsevier Publishing, New York.

Cressie, N. 1989. Geostatistics. The American Statistician. 43(4): 611(23).

Cressie, N. 1993. Statistical for Spatial Data. John Wiley & Sons, New York.

Cressie, N. & M. M. Majure. 1995. Non-Point Source Pollution of Surface Waters over a Watershed. Programme Abstracts of the third SPRUCE International Conference. Merida, Mexico.

Day, J., C. Hall, M. Kemp, & A. Yánez-Arancibia. 1989. Estuarine Ecology. John Wiley & sons, New York.

Deutsch , C. V. & A. G. Journel. 1992. GSLIB: Geostatistical Software Library and User´s Guide. Oxford University Press, New York.

Díaz- Francés, E. (1993). Introducción a Conceptos Básicos de Geoestadística. Memorias Seminario Estadística y Medio Ambiente. Centro de Investigación en Matemáticas, CIMAT. Guanajuato, México.

Diggle, P., L. Harper y S. Simon. (1995). Geoestatistical Analysis of Residual Contamination from Nuclear Weapons Testing. Programme Abstracts of the third SPRUCE International Conference. Merida, Mexico.

Englund, E. & A. Sparks. 1988. GeoEAS, User´s Guide. EPA, Las Vegas.

Evangelos A. & G. T. Flatman. 1988. On Sampling Nonstationary Spatial Autocorrelated Data. Computers and Geosciences. 14(5): 667(86).

Gamma Design. 1995. GS+. Geostatistical software for the Agronomic and Biological Science, version 2.3. Plainwell, Michigan.

Garmin International, Inc. 1993. Garmin Communication and Navigation. GPS 100 SRVY II personal surveyor. Owner´s manual. Lenexa..

Giraldo, R., D. Ospína & N. Méndez. 2001. Design of a Sampling Network for an Estuary in the Colombian Caribbean. Rev. Acad. Col. Cienc. 25(97):509-518

Giraldo, R. 2002. Construcción de un Indicador para el Estudio Conjunto de la Distribución Espacial de Múltiples Variables Binarias. Tesis de Maestría en

Estadística. Departamento de Estadística. Universidad Nacional de Colombia, Bogotá.

Giraldo, R. 1996. Geoestadística Aplicada a Datos Multivariados Provenientes del Monitoreo de las Aguas de la Ciénaga Grande de Santa Marta y el Complejo Pajarales, Caribe Colombiano. tesis de grado Especialización en Estadística. Universidad Nacional de Colombia..

Giraldo, R., J. Martínez, L. H. Hurtado, S. Zea & R. Madera. 1995. Análisis de Clasificación de Series Temporales: El Caso de la Salinidad en la Ciénaga Grande de Santa Marta, Colombia. An. Inst. Invest. Mar. Punta Betín 24: 123-134.

Hernández, C.A. & K. Gocke. 1990. Productividad Primaria en la Ciénaga Grande de Santa Marta, Colombia. An. Inst. Invest. Mar. Punta Betín 19 - 20 : 101 - 119

Hoaglin, D. F., F. Mosteller & J. Tukey. 1983. Understanding Robust and Exploraory Data Análisis. John Willey & Sons, New York.

IGAC, 1973. Monografía del Departamento del Magdalena. Inst. Geogr. Agustín Codazzi, Bogotá.

Isaaks, E. & R. M. Srivastava. 1989. Applied Geostatistics. Oxford University Press, New York.

Jay, D.A., R.J. Uncles, J. Largier, W.R. Geyer, J.Vallino & W.R. Boynton. 1997. A Review of Recent Developments in Estuarine Scalar Flux Estimation. Estuaries. 20(2): 262-280.

Journel, A.G. y Ch. J. Huijbregts. 1978. Mining Geostatistics, Academics Press, New York.

Krige, D. G. 1951. A Statistical Approoach to Some Basic Mine Valuation Problems on the Witwatersrand. Journal of the Chemical, Metallurgical and Mining Society of South Africa, 52: 119-139.

Lebart, L. A., A. Morineau & M. Piron. 1995. Statistique Exploratoire Multidimenssionnelle. Dunod, Paris

Mancera, J. E. 1990. Caracterización Ecológica de la Salina Artificial Pozos Colorados, Caribe colombiano. An. Inst. Invest. Mar. Punta Betín 19-20 : 121 - 138.

Mancera, J. E. & L. A. Vidal. 1994. Florecimiento de microalgas relacionado con mortandad masiva de peces en el complejo lagunar Ciénaga Grande de Santa Marta, Caribe colombiano. An. Inst. Invest. Mar. Punta Betín 23: 103 - 117.

Matheron, G. 1962. Traite de Geostatistique Apliquee, Tome I. Memoires bureau de Recherches Geologiques et Minieres, N 24. Editions Bureau de Recherche et Minieres, Paris.

McBratney, A. B., Webster, R. and Burgess, T. M. 1981. 'The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalized Variables I', Computers and Geosciences. 7(4): 331-334

Mood, A., F. A. Graybill & D. C. Boes. 1974. Introduction to the Theory of Statistics. McGraw-Hill, New York.

Myers, D. E. 1987. Optimization of Sampling Locations for Variogram Calculations. Water Resources Research. 23(3): 283(93).

Nixon, S.W. 1997. Prehistoric Nutrient Inputs and Productivity in Narrangansett Bay. Estuaries. 20(2): 253-261

Petitgas, P. 1996. Geostatistics and Their Applications to Fisheries Survey Data 5: 114-142. In: B. A. Megrey & E. Mosknes, (E). Computers and Fisheries Research. Chapman-Hall, Londres.

Pla, L.1986. Análisis Multivariado: Método de Componentes Principales. Monografía No 27. Serie de matemática. Secretaría General de la OEA.

Reid, G. K. & R. D. Wood. 1976. Ecology of Inland Waters and Estuaries. Second edition. D. Van Nostrand, New York

Roldán, G. 1992. Fundamentos de Limnología Neotropical. Editorial Universidad de Antioquia, Medellín.

Robertson, G. P. 1987. Geostatistics in Ecology: Interpolating with Know Variance. Ecology 68(3) : 744-748.

Samper, F.J. & J. Carrera 1990. Geoestadística. Aplicaciones a la Hidrogeología Subterránea. Centro Internacional de Métodos Numéricos en Ingeniería. Universitat Politécnica de Catalunya. Barcelona.

Sánchez, C. 1996. Variación Espacial y Temporal de la Ictiofauna de la Ciénaga Grande de Santa Marta, Complejo de Pajarales y Ciénagas delCostado Occidental de la Isla de Salamanca, Caribe colombiano. Tesis Biología. Fac. Cienc. Depto. Biol. Univ. Nacional de Colombia, Santafé de Bogotá.

Vidal, L.A. 1995. Estudio del Fitoplancton en el Sistema Lagunar Estuarino Tropical Ciénaga Grande de Santa Marta, Colombia, durante el año 1987. tesis de grado M. Sc. Universidad Nacional de Colombia, 270 pp.

Wackernagel. H. 1995. Multivariate Geostatistics. An Introduction with Applications. Springer-Verlag, Berlín.

Warrick, A. W., D. E. Myers & D. R. Nielsen. 1986. Geostatistical Methods Applied to Soil Science. Methods of Soil Analysis. Part 1. Physical and Mineralogical Methods- Agronomy Monograph 9: 53 - 81.

Wheaton, F.W. 1977. Aquacultural Engineering. Krieger Publishing Company. Malabar, Florida.

Welch, E. B. 1992. Ecological Effects of Wastewater. Applied Limnology and Pollutant Effects. (second edition). Chapman & Hall, Londres.

Wiedemann, H. V. 1973. Reconnaissance of the C.G.S.M., Colombia: Physical Parameters and Geologid history. Milt. Inst. Colombo-Aleman Inv. Cientif. 7: 85-119.