MÓDULO DE GEOESTADÍSTICA PARA EXCEL: APLICACIÓN A PESCA

Diana María González Troncoso
Centro Oceanográfico de Vigo. Cabo Estai, Vigo

RESUMEN

La Geoestadística es una ciencia cada vez más usada en distintos ámbitos científicos. Por ello, parece necesario que existan herramientas sencillas que permitan usarla. En el presente trabajo se expone una herramienta en Excel fácilmente utilizable y accesible a todos los interesados, y se aplica a los datos obtenidos para el fletán negro por medio de campañas científicas llevadas a cabo por el Centro Oceanográfico de Vigo en Svalbard.

INTRODUCCIÓN

La *Geoestadística* nació en la década de los 50 como una ciencia minera, con el fin de evaluar las reservas minerales útiles. Por lo tanto, es una ciencia joven, en pleno desarrollo. El término fue concebido por Georges Matheron (Matheron, 1962, 1963a, 1963b) a partir de trabajos previos de H. Sichel, D. G. Krige y B. Matern.

La Geoestadística es una ciencia aplicada que estudia las variables distribuidas espacialmente, partiendo de una muestra representativa del fenómeno en estudio. Se basa en el hecho de que los datos van a estar correlacionados espacialmente, esto es, que un dato va a estar relacionado con datos cercanos, y esta dependencia va perdiendo fuerza a medida que nos alejamos del dato.

El objetivo de la Geoestadística (como de toda ciencia estadística) es la predicción. Para ello se suelen aplicar las técnicas *Kriging*, que básicamente proporcionan una predicción de Z(s), es decir, del valor del proceso en una posición espacial s, a partir de una muestra dada $\{Z(s_1),...,Z(s_n)\}$. Para conseguir dicha estimación, primero ha de caracterizarse la relación de dependencia entre las distintas localizaciones espaciales; en este sentido, puede recurrirse a la estimación del covariograma o del variograma.

Numerosas son las ciencias que utilizan los métodos que desarrolla la Geoestadística. Entre ellas se encuentra la Oceanografía, la cual usaremos en el presente trabajo. Por lo tanto, cada vez se hace más necesaria la existencia de herramientas para poder aplicarla. En este trabajo se expone una herramienta programada en Visual Basic para Excel, pues hoy en día casi todos los ordenadores vienen dotados de esta aplicación.

CONCEPTOS DE GEOESTADÍSTICA

ESTACIONARIEDAD

Lo primero que se estudia de los datos es la estacionariedad. Modelándola, denotaremos la variable espacial en la posición s como Z(s), con $s \in D \subset IR^k$. Un conjunto de la forma {Z(s), $s \in D$} recibe el nombre de *proceso estocástico* o *proceso aleatorio*. En este caso, designaremos la distribución del proceso como sigue:

$$F_{s_1,\ldots,s_n}(z_1,\ldots,z_n) = P(Z(s_1) \le z_1,\ldots,P(Z(s_n) \le z_n)$$

cumpliendo ciertas hipótesis de regularidad. Además, se le suelen imponer hipótesis relativas al valor de la media y de la varianza, definiéndose de ese modo los distintos tipos de estacionariedad, que son las siguientes:

1. $E[Z(s)] = \mu, \forall s \in D$

2a. $Cov[Z(s),Z(t)] = c(s-t), \forall s,t \in D$ Estacionariedad de segundo orden

2b. $Var[Z(s)-Z(t)] = 2\gamma(s-t), \forall s,t \in D$ Estacionariedad intrínseca

La función c(h) es conocida como *covariograma*, y la función $\gamma(h)$ como *semivariograma*. Nótese que estas funciones dependen no sólo de la distancia entre dos posiciones, sino también de la dirección que las separa. Si el covariograma y el variograma no dependen de la dirección, se dice que el proceso es *isotrópico*. En caso contrario, recibe el nombre de *anisotrópico*.

Suponiendo la existencia de varianza finita, se puede demostrar que:

Estac. de segundo orden => Estac. Intrínseca => $\gamma(s-t) = c(0) - c(s-t) \quad \forall s, t \in D$

Si bien para procesos estacionarios de segundo orden es lo mismo estudiar el variograma que el covariograma, en la práctica se suele recurrir a la estimación del variograma ya que es menos restrictiva con respecto al proceso, por ser válida para procesos intrínsecos donde no es necesario exigir la existencia de varianza finita.

FOCOS DE POSIBLE NO ESTACIONARIEDAD

Para comprobar focos de posible no estacionariedad, se suelen estudiar las distribuciones de los valores de los siguientes variogramas:

Variograma nube

Propuesto por Chauvet (1982), el *variograma nube* en una dirección unitaria dada e es la representación grafica de los valores $\{(Z(s_i + he) - Z(s_i))^2\}$ para distintos valores de h.

Variograma nube de las raíces de las diferencias

Descrito en Cressie (1993), el *variograma nube de las raíces de las diferencias* en una dirección unitaria dada e es la representación grafica de los valores $\{|Z(s_i + he) - Z(s_i)|^{1/2}\}$ para distintos valores de h.

En ambos casos se puede representar un Box-Plot para cada h, de modo que en aquellos casos en los que la diferencia entre la media y la mediana sea muy grande se podría analizar la existencia de valores atípicos. El variograma nube de las raíces de las diferencias es más robusto que el otro.

ESTIMACIÓN DEL VARIOGRAMA

Para estimar esta función existen numerosos estimadores. Los más usados son los siguientes:

Estimador muestral o empírico

Propuesto por Matheron (1962, 1963b), este estimador está basado en el método de los momentos, y es del siguiente modo:

$$2\hat{g}(h) = \frac{\sum_{N(h)} (Z(s_i) - Z(s_j))^2}{|N(h)|} \qquad \text{con } h \in IR^k \text{ y } N(h)=\{(s_i,s_j) / s_i - s_j = h, i,j = 1,...,n\}$$

Este estimador, a pesar de ser insesgado, es poco robusto, muy sensible a la presencia de valores atípicos.

Estimador Robusto

Propuesto por Cressie y Hawkins (1980), este estimador es del siguiente modo:

$$2\tilde{g}(h) = \frac{\left[\sum_{(s_i,s_j) \in N(h)} \left| Z(s_i) - Z(s_j) \right|^{1/2} \right]^4}{|N(h)|^4 \left(0.457 + \frac{0.494}{|N(h)|} \right)}$$

Como indica su nombre, este estimador es más robusto, es decir, es menos sensible a la presencia de valores atípicos que el anterior.

Ambos estimadores se calculan para un vector h dado de modo que, fijada la dirección unitaria y en dicha dirección, se consideran los pares de datos que distan ||h|| unidades en esa dirección. En caso de isotropía, se evaluaría el estimador sobre una distancia $h \in IR$ y se considerarían todos lospares de datos distanciados h unidades, con independencia de la dirección del vector que les separa.

Siguiendo las recomendaciones de Journel y Huijbregts (1978), se debe tener un mínimo de 30 pares en N(h) para obtener la estimación. Para conseguir esto, a veces se recurre a considerar una región de tolerancia en torno a h que garantice la existencia de suficientes pares de datos; esta región de tolerancia debe ser pequeña y puede variar para cada h.

Métodos paramétricos

Entre las propiedades que ha de cumplir un variograma, hay que citar que ha de ser *condicionalmente definido negativo*, esto es, para cualquier colección de posiciones espaciales $\{s_1,...,s_n\}$ y para cualquier colección de números reales $\{a_1,...,a_n\}$ tales que $\sum_{i=1}^{n} a_i = 0$, se cumple que:

$$\sum_i \sum_j a_i a_j g(s_i - s_j) \leq 0$$

Sin esta propiedad, se podrían dar estimaciones negativas del error cuadrático medio de la predicción, lo cual sería un absurdo. Aunque generalmente los variogramas empírico y robusto ajustan bien los datos, el problema es que ninguno de ellos es condicionalmente definido negativo. Por ello lo que se suele hacer es, una vez estudiado el variograma empírico o el robusto para hacerse una idea de cómo es la forma del variograma teórico, ajustar éste a través de una familia paramétrica de curvas del modo:

$$P=\{2\gamma / 2\gamma(.) = 2\gamma(.,\theta), \theta \in \Theta\}$$

donde $2\gamma(.,\theta)$ es una función de θ condicionalmente definida negativa. Las funciones que usualmente se utilizan, en caso de isotropía, son las dadas por los modelos lineal, esférico, exponencial, cuadrático-racional, oscilatorio y potencial.

Estos modelos son válidos en presencia de isotropía. Si tenemos anisotropía, lo más usual es ajustar un variograma distinto en cada dirección.

Generalmente se estima el variograma teórico por el método de mínimos cuadrados ponderados, que se basa en construir previamente algún estimador no paramétrico de manera que se selecciona el parámetro θ de la familia de variogramas paramétricos válidos P que mejor se aproxime al estimador empírico. Este método consiste en minimizar:

$$\underset{q}{Min} \sum_{j=1}^{n} \frac{|N(h_j)|}{g_q(h_j)^2} \left(\breve{g}(h_j) - g_q(h_j)\right)^2 = \underset{q}{Min} \sum_{j=1}^{n} |N(h_j)| \left(\frac{\breve{g}(h_j)}{g_q(h_j)} - 1\right)^2$$

donde $\breve{g}(.)$ es el estimador no paramétrico escogido (generalmente, el empírico o el robusto) y g_q es el variograma paramétrico (función de q) a estimar, perteneciente a nuestra familia paramétrica. De este modo obtenemos un estimador del variograma que da más peso a aquellos h_j a los que corresponde mayor número de pares con esa diferencia.

KRIGING

La predicción espacial o Kriging, nombre puesto en honor a Krige, es el objetivo último de la Geoestadística. Nuestro objetivo será hallar una estimación del valor de la variable Z en un cierto

punto de nuestro entorno s_0, $\hat{Z}(s_0)$, a partir de la muestra $\{Z(s_1),...,Z(s_n)\}$. Para hacer esto, antes de nada hemos de caracterizar la dependencia espacial existente entre los datos, y esto lo hemos hecho en apartados anteriores gracias al variograma.

Aunque existen varios métodos de Kriging, el más usado es el *Kriging Ordinario* (KO), que consiste en estimar $Z(s_0)$ mediante el mejor estimador lineal insesgado, BLUE en inglés (Best Linear Unbiased Estimator). Se puede ver que conseguir esto es lo mismo que resolver el siguiente problema de optimización:

$$MinE[(\hat{Z}(s_o) - Z(s_0))^2] + 2m(\sum_{i=1}^{n} l_i - 1)$$

PROGRAMACIÓN Y RESULTADOS

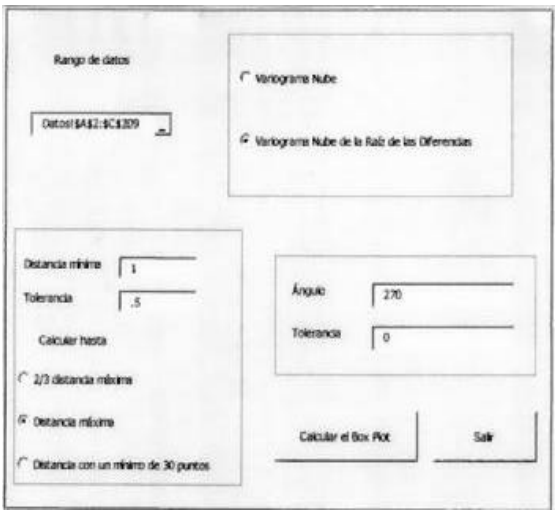
El objetivo principal del trabajo era aplicar las técnicas Kriging, pero para ello había que realizar antes una serie de cálculos. Por ello, he creado herramientas para calcular los Variogramas Nube, los Variogramas Muestrales, los Variogramas Paramétricos, y también para realizar un estudio exploratorio sobre la Isotropía. Finalmente pude programar el Kriging, realizando gráficas en 3 dimensiones, aunque en el presente trabajo no se muestran.

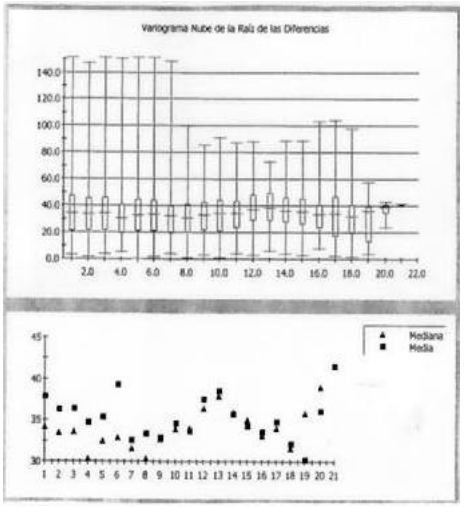
DATOS USADOS

Los datos usados para el análisis de la población del fletán negro (*Reinhardtius hippoglossoides*) proceden de las campañas científicas *Fletán Ártico* llevadas a cabo por el Equipo de Pesquerías Lejanas del Centro Oceanográfico de Vigo en aguas del área de Protección del Archipiélago Svalbard, División IIb del CIEM, amablemente cedidos por dicho equipo. Para este estudio se usaron los datos de las campañas realizadas desde el año 1997 al año 2001, todas ellas en Octubre. En estas campañas la especie predominante es sin duda el fletán negro, que, por ejemplo, en el año 2000 fue 68 veces la captura de la segunda especie más abundante, el lirio o bacaladilla.

ESTUDIO DE LA ESTACIONARIEDAD

En el presente trabajo se presentan la interfaz de entrada de la estacionariedad y la salida del programa para el año 2000 del estudio de la estacionariedad mediante los variogramas nube. Es el siguiente:

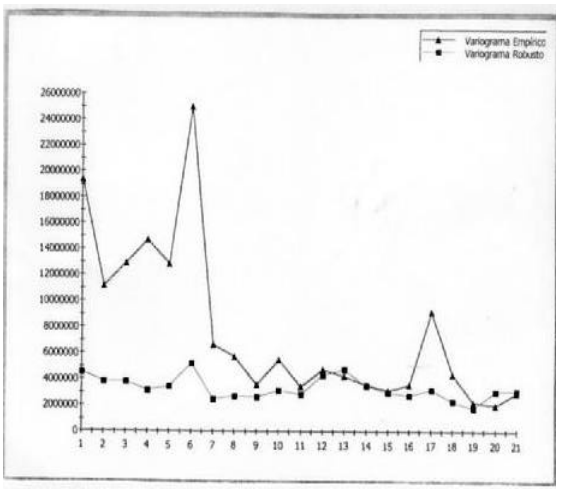




Si estudiamos el rango intercuartílico en los Box-Plot (amplitud de las cajas), vemos que la dispersión en torno a la mediana se mantiene bastante constante, excepto tal vez en las últimas distancias, seguramente debido a que en esos casos tenemos pocos pares de puntos. Este hecho hace pensar en la estacionariedad de los datos. Habría también que fijarse en la diferencia entre la media y la mediana, pues en aquellos casos en los que difieren mucho habría que analizar posibles focos atípicos. Esto ocurre bastante en este año.

VARIOGRAMA MUESTRAL

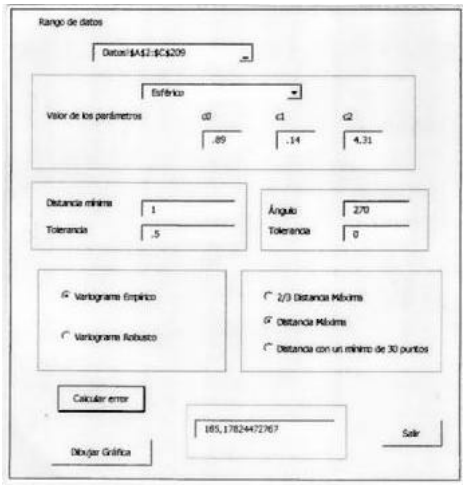
El estudio del variograma muestral se muestra también para el año 2000. Se representaron los dos variogramas muestrales a la vez, el Variograma Empírico y el Variograma Robusto. En ellos se puede seguir la evolución de la varianza a medida que se consideran datos más distanciados:

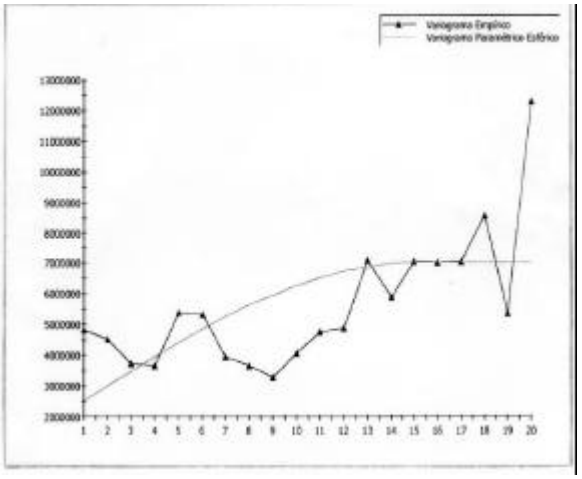


En esta gráficas se aprecia el carácter poco robusto del variograma empírico con respecto al robusto, pues aquel se ve claramente influenciado por los valores atípicos localizados mediante el variograma nube.

VARIOGRAMA PARAMÉTRICO

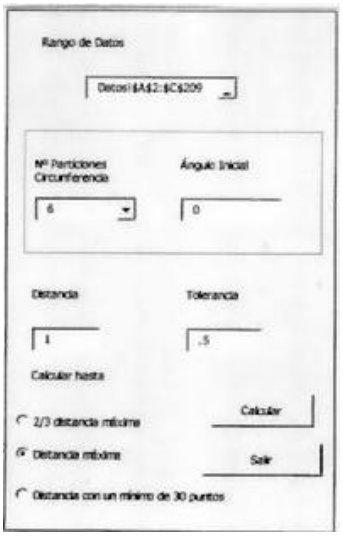
El estudio del variograma paramétrico resulta más arduo que los anteriores, como consecuencia de tener que encontrar los parámetros que minimicen la función. Por lo tanto, este apartado se estudió sólo para el año 97 y para el variograma empírico. A la vista de los datos, parece que los modelos que mejor ajustan ese variograma son el esférico y el exponencial. Opté por el modelo esférico a partir de la forma de los variogramas muestrales, si bien este procedimiento subjetivo podría mejorarse implementando en un futuro el mecanismo de selección propuesto por Gorsich and Genton (1999) basado en considerar una aproximación a la primera derivada del variograma.

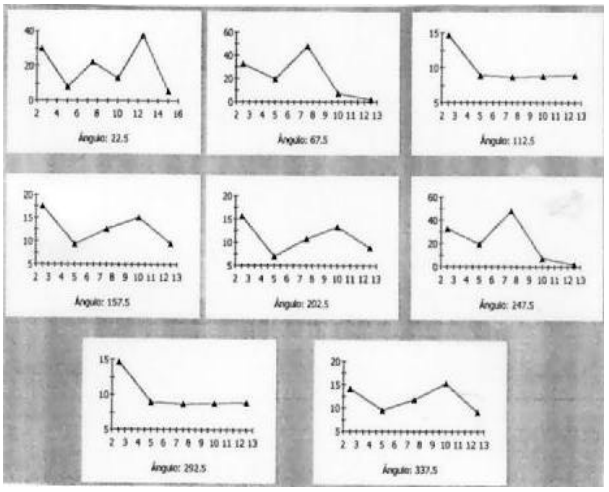




ISOTROPÍA

De manera general y por conocimientos prácticos, se puede decir que el proceso según el cual se distribuye el fletán negro es anisotrópico, pues se suele distribuir en dirección Norte-Sur. Analizando, por ejemplo, el gráfico resultante para el año 2000, se ve la anisotropía de esta especie, pues dependiendo del ángulo usado, tenemos resultados diferentes:





AGRADECIMIENTOS

Mi más sincero agradecimiento al Equipo de Pesquerías Lejanas del Centro Oceanográfico de Vigo, y en particular a Xabier Paz, por la cesión de los datos del fletán negro usados en el presente trabajo.

Mi agradecimiento, también, a Pilar García Soidán y a Manuel Febrero Bande por su ayuda en la realización de esta ponencia.

BIBLIOGRAFÍA

Chauvet, P. (1982): "The Variogram Cloud". *Internal Note* N-725. Centre de Geoestatistique Fontaineblau

Cressie, N. (1993): *Statistics for Spatial Data*. Ed. Wiley, Nueva York

Cressie, N. and Hawkins, D. M. (1980): "Robust Estimation of the Variogram". *Journal of the International Association for Mathematical Geology* 12, n.2: 115-125

Gorsich, D. and Genton, M. (1999): "Variogram Model Selection via Nonparametric Derivate Estimation". *Journal of the International Association for Mathematical Geology* 32, n.3: 249-270

Halvorson, M. (1999): *Aprenda Visual Basic 6.0 Ya*. Ed. MacGrawHill

Journel, A. G. and Huijbregts, C. J. (1978): *Minig Geoestatistic*. Academic Press, London

González Troncoso, D. (2002): *Módulo de Estadística Espacial en Visual Basic para Excel*. Universidad de Vigo.

Matheron, G. (1962): *Traite de Geoestatistique Appliquee. Tome 1*. Memoires du Bureau de Recherches Geologiques et Minieres. Ed. Technip, París

Matheron, G. (1963a): *Traite de Geoestatistique Appliquee. Tome 2: Le Krigeage*. Memoires du Bureau de Recherches Geologiques et Minieres. Ed. Technip, París

Matheron, G. (1963b): "Principles of Geostatistics". *Economic Geology*, 58: 1246-1266

Paz, X. and Román, E. (2000): "Spanish Bottom Trawl Survey "Fletán Ártico 97-99" in the slope of Svalbard Area, ICES Division IIb: 1997-99". *ICES CM 2000- Artic Fisheries Working Group. Working Document* 39, 34 pp.

Paz, X, Casas, J. M. and González Troncoso, D. (2002): "Spanish Bottom Trawl Survey "Fletán Ártico 2001" in the slope of Svalbard Area, ICES Division Iib". *ICES CM 2002- Artic Fisheries Working Group. Working Document* 31, 17 pp.

Román, E. and Paz, X. (2001): "Spanish Bottom Trawl Survey "Fletán Ártico 2000" in the slope of Svalbard Area, ICES Division IIb". *ICES CM 2001- Artic Fisheries Working Group. Working Document* 32, 15 pp.