Universidad Adolfo Ibáñez
Facultad de Ciencias

**Prueba N°2: Análisis Estadístico de datos espaciales**

1. Considere las 500 coordenadas espaciales en el archivo *coords1.txt*. En estas coordenadas se observa una variable disponible en el archivo *data1.txt*. Se considera como modelo un campo aleatorio Gausiano
$$Y(\mathbf{s}) = \mu + \sigma Z(\mathbf{s})$$
donde $Z$ es un campo aleatorio Gausiano de media 0 y varianza 1, $\mu$ es el parámetro de media y $\sigma^2$ el parámetro de varianza.

   Para el campo aleatorio $Z$, se eligen dos modelos del correlación que son dos casos especiales del modelo de Matérn
$$\mathcal{M}_{\nu,\alpha}(\mathbf{h}) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{||\mathbf{h}||}{\alpha}\right)^{\nu} \mathcal{K}_\nu \left(\frac{||\mathbf{h}|}{\alpha}\right)$$
es decir, el modelo 1 dado por
$$\mathcal{M}_{\frac{1}{2},\alpha}(\mathbf{h}) = e^{-\frac{||\mathbf{h}||}{\alpha}}$$
y el modelo 2 dado por
$$\mathcal{M}_{\frac{3}{2},\alpha}(\mathbf{h}) = e^{-\frac{||\mathbf{h}||}{\alpha}} \left(1 + \frac{||\mathbf{h}||}{\alpha}\right)$$
Utilizando el paquete *GeoModels* del software estadístico R:

   - Representar gráficamente y comentar la estimación del semivariograma asumiendo isotropía.
   - Representar gráficamente la distribución de los datos y comentar si el modelo Gausiano es razonable
   - Estimar con el método de máxima verosimilitud los parámetros de media, varianza y parámetros dependencia espacial asociados a los modelos de correlación 1 y 2.
   - Elegir el mejor modelo de correlación utilizando el criterio de información de Akaike (AIC).
   - Elegir el mejor modelo de correlación según su capacidad predictiva, en particular utilizando el criterio del mínimo (root mean squared error), mediante validación cruzada (ocupe 100 iteraciones y el 10 % de los datos como validación en cada iteración).
   - Con el modelo elegido y sus respectivas estimaciones, realizar una representación gráfica de la predicción y error cuadrático medio asociado en las localizaciones espaciales contenidas en al archivo *loc_to_pred.txt*.

2. El archivo *data_temp.txt* contiene 1500 coordenadas espaciales (proyectadas) con la asociada temperatura media del mes de abril 2015 y alturas en una específica región de España.

   Se quiere aplicar es siguiente modelo de regresión espacial lineal
$$Y(\mathbf{s}) = \beta_0 + \beta_1 A(\mathbf{s}) + \sigma Z(\mathbf{s})$$
donde $Z$ es un campo aleatorio Gausiano de media 0 y varianza 1 y $A$ son las alturas. Se asume que el modelo de correlación de $Z$ es un caso especial del modelo Wendland generalizado
$$\mathcal{GW}_{0,4,\alpha} = \begin{cases} \left(1 - \frac{||\mathbf{h}||}{\alpha}\right)^4 & 0 \leq ||\mathbf{h}|| \leq \alpha \\ 0 & ||\mathbf{h}|| > \alpha, \end{cases}$$

El objetivo es estimar $\beta_0$ y $\beta_1$ (que expresa el impacto de la altura en la temperatura media), el parámetro de varianza $\sigma^2$ y el parámetro de dependencia espacial $\alpha$.

Utilizando el paquete *GeoModels* del software estadístico R:

- Estime los parámetros $(\beta_0, \beta_1, \sigma^2, \alpha)$ con el método de máxima verosimilitud compuesta basado en parejas condicionales utilizando una función peso con vecindad de orden 3. ¿Cuál es el efecto de la altura sobre la temperatura media?.
- Estime los errores estándar asociados ocupando el método bootstrap.
- Compare el semivariograma empírico de los residuos con el semivariograma estimado. ¿Es el modelo de dependencia de Wendland un modelo razonable?.

3. Utilizando el paquete *GeoModels* del software estadístico R:

   - En la coordenadas espaciales *coords1.txt* y tiempos $(1, 2, 3, 4)$, simule con el método de descomposición de Cholesky un campo aleatorio Gausiano espacio-tiempo del tipo

     $$Y(\mathbf{s}, t) = \mu + \sigma Z(\mathbf{s}, t)$$

     siendo $Z$ un campo aleatorio Gausiano de media cero y varianza 1, $\mu = 1$, $\sigma^2 = 2$ y la función del correlación de $Z$

     $$\rho(\mathbf{h}, u) = e^{-\frac{||\mathbf{h}||}{\alpha_s} - \frac{|u|}{\alpha_t}}.$$

     con $\alpha_s$=0.2/3 y $\alpha_t = 2/3$.
   - Estime los parámetros $\mu$, $\sigma^2$, $\alpha_s$ y $\alpha_t$ con el método de máxima verosimilitud compuesta basado en parejas condicionales con función peso dada por una vecindad espacial de orden 3 y temporal de orden 1.
   - Ocupando la estimaciones obtenidas al punto anterior, calculen la predicción optima en las coordenadas espaciales (0.3,0.2) y (0.5,0.5) y tiempo 2.

4. Considere un campo aleatorio de tipo Gausiano $Z = \{Z(\mathbf{s}), \quad \mathbf{s} \in A \subset R^2\}$ de media constante $E(Z(\mathbf{s})) = \mu$, varianza unitaria $Var(Z(\mathbf{s})) = 1$ y función de correlación conocida $\rho(\mathbf{h})$.

   Se desea estimar $\mu$ con el método de máxima verosimilitud compuesta basado en parejas marginales, es decir, encontrar el estimador

   $$\widehat{\mu} = \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \left( \sum_{i=1}^{n} \sum_{j \neq i}^{n} l_{ij}(\mu) w_{ij} \right)$$

   donde $l_{ij}(\mu)$ es la log-verosimilitud asociada al vector aleatorio bivariado Gausiano $(Z(\mathbf{s}_i), Z(\mathbf{s}_j))^T$ y $w_{ij}$ son pesos conocidos.

   - Encuentre una forma explícita del estimador $\widehat{\mu}$.
   - Demuestre que el estimador es insesgado.

5. Considere un campo aleatorio estacionario $Z = \{Z(\mathbf{s}), \quad \mathbf{s} \in A \subset R^d\}$ con $E(Z(\mathbf{s})) = \mu$, $V(Z(\mathbf{s})) = \sigma^2$ y función de correlación $\rho(\mathbf{h})$. Además sea $Z = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))$.

   - Sea $g(Z)$ un predictor de $Z(\mathbf{s}_0)$ donde $g$ es una función arbitraria. Mostrar que el error cuadratico medio

     $$EQM(g(Z)) = E((Z(\mathbf{s}_0) - g(Z))^2)$$

     es minimo cuando $g(Z) = E(Z(\mathbf{s}_0)|Z)$ y que $EQM(E(Z(\mathbf{s}_0)|Z)) = V(Z(\mathbf{s}_0)|Z)$.

- Sea $g(Z) = \lambda_0 + \lambda^T Z$ un predictor lineal de $Z(\mathbf{s}_0)$. Mostar que, si el predictor es insesgado es decir $E(g(Z)) = E(Z(\mathbf{s}_0)) = \mu$, entonces el error cuadratico medio

$$EQM(g(Z)) = E((Z(\mathbf{s}_0) - g(Z))^2)$$

es minimo cuando $g(Z) = \mu + r^T R^{-1}(Z - 1\mu)$ y que $EQM(\mu + r^T R^{-1}(Z - 1\mu)) = \sigma^2(1 - r^T R^{-1} r)$.
($r = [\rho(\mathbf{s}_0 - \mathbf{s}_i))]_{i=1}^n, \quad R = [\rho(\mathbf{s}_i - \mathbf{s}_j))]_{i,j=1}^n$ y 1 es un vector de unos).

6. Considere un campo aleatorio estacionario de tipo Bernoulli $B = \{B(\mathbf{s}), \quad \mathbf{s} \in A \subset R^2\}$ definido como

$$B(\mathbf{s}) = \begin{cases} 1 & Z(\mathbf{s}) < \mu(\mathbf{s}) \\ 0 & e.o.c. \end{cases}.$$

donde $Z$ es un campo aleatorio Gausiano de media 0 y varianza 1 y función de correlación $\rho(\mathbf{h})$.

- Encuentre la distribución marginal del campo aleatorio $B$, es decir,

$$Pr(B(\mathbf{s}) = m), \quad m = 0, 1$$

- Encuentre la distribución bivariada del campo aleatorio $B$, es decir,

$$Pr(B(\mathbf{s}_i) = m, B(\mathbf{s}_j) = n), \quad m, n = 0, 1$$

- Encuentre la función de covarianza del campo aleatorio $B$