Contribuciones a la Estadística Espacial No Paramétrica

Sergio Castillo Páez (UVIGO, ESPE)

II CONFERENCIA DE MATEMÁTICOS ECUATORIANOS

París, Abril 2016

Indice

Un ejemplo introductorio

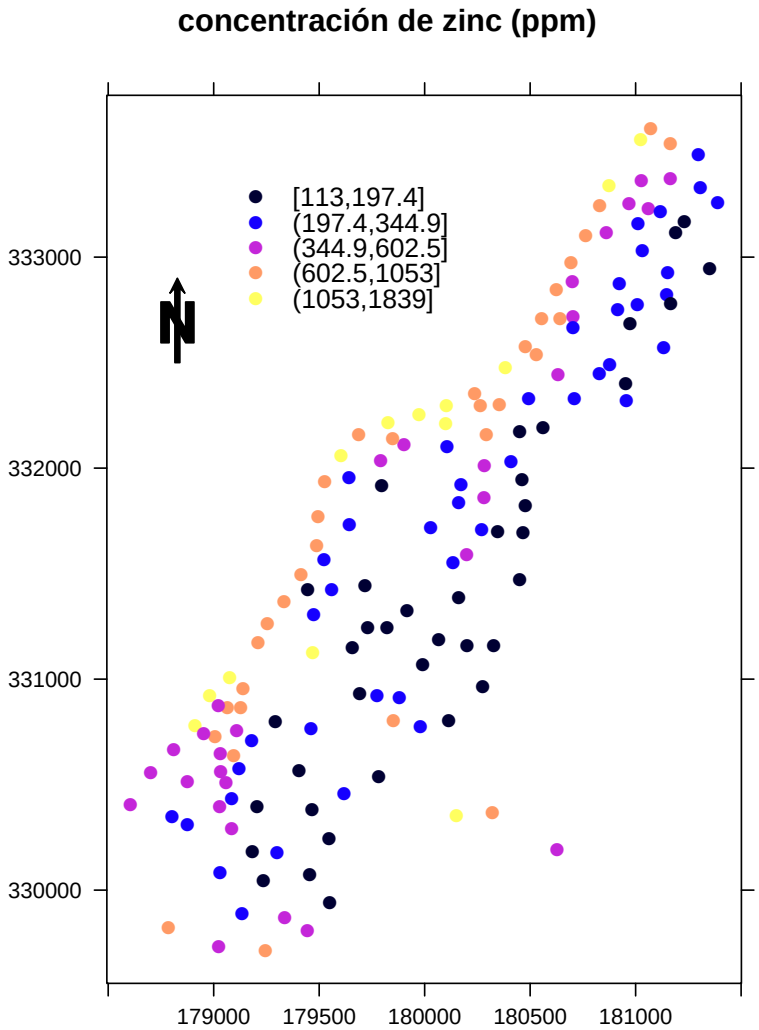

Figura 1. Concentración de zinc medida sobre la superficie de las riberas del río Meuse (Pebesma, 2004)

Modelo geoestadístico

- Proceso espacial: $\left\{ Y(\mathbf{x}), \mathbf{x} \in D \subset \mathbb{R}^d \right\}$, con dominio D continuo.
- Modelo:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{1}$$

  - $\mu(\cdot)$ función tendencia (determinística).
  - $\varepsilon(\cdot)$ proceso de error estacionario de segundo orden, de media cero y covariograma:

$$C(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$$

- Usualmente, la dependencia se modela a través del variograma:

$$\gamma(\mathbf{u}) = \frac{1}{2} Var(\varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x}+\mathbf{u})) \tag{2}$$

Objetivos principales

- A partir de n valores observados $\mathbf{Y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))^t$, puede interesar:
  - Estimar la tendencia del proceso: $\hat{\mu}(\cdot)$
  - Obtener la dependencia estimada: $\hat{\gamma}(\cdot)$
  - Realizar inferencias sobre el proceso espacial:
    - Predicciones en regiones no observadas: $\hat{Y}(\mathbf{x}_0)$.
    - Intervalos de confianza para $\mu(\cdot)$ y $\gamma(\cdot)$.
    - Mapas de riesgos: $P(Y(\mathbf{x}_0) \geq c)$.

Enfoque paramétrico y no paramétrico

- Enfoque paramétrico tradicional:

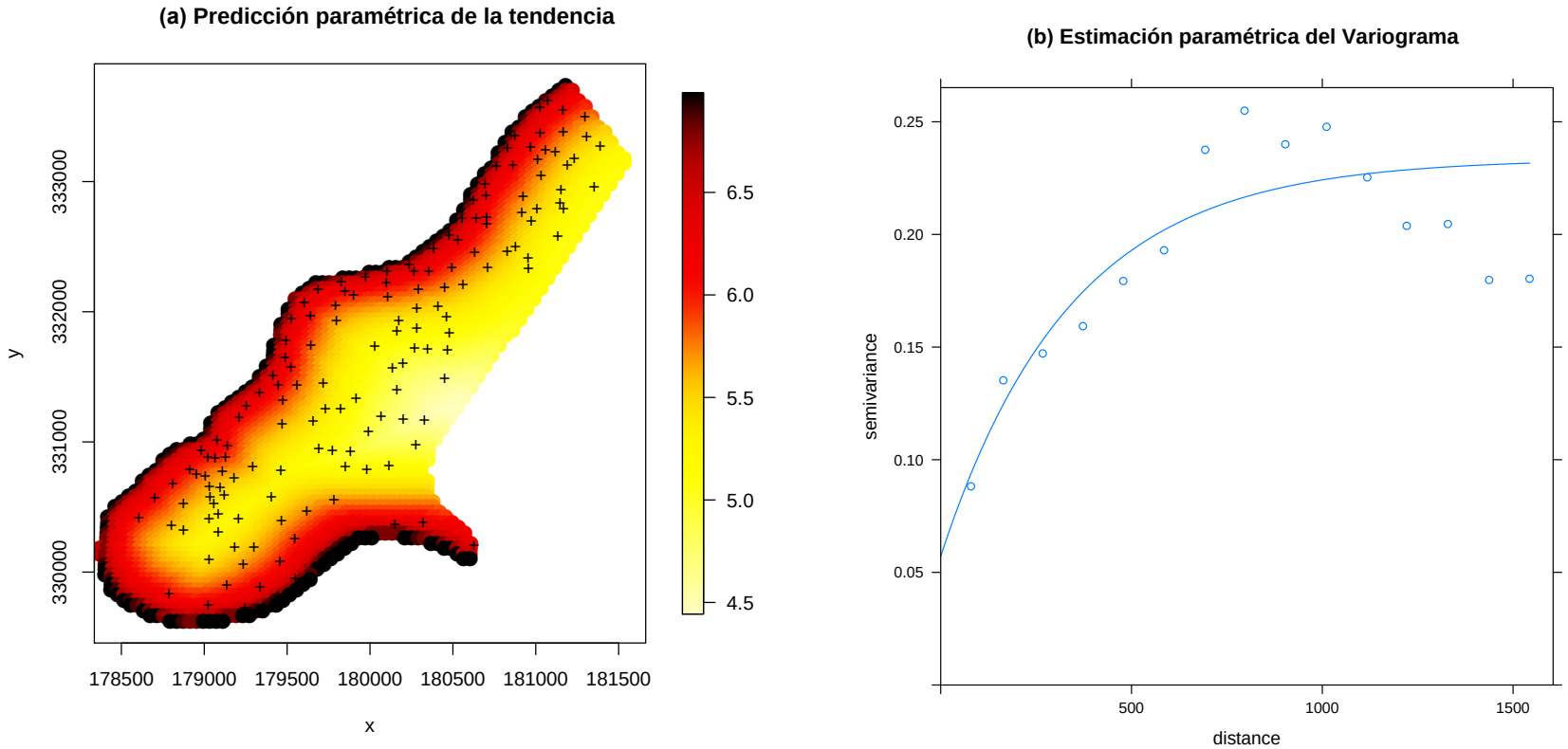

Figura 2. (a) Predicción paramétrica de la tendencia de Log(Zinc) tomando como variable explicativa la raíz cuadrada de la distancia al río, y (b) Estimación paramétrica del variograma de los residuos a partir de un modelo Exponencial

Enfoque paramétrico y no paramétrico

- Enfoque no paramétrico:
  - No están expuestos a problemas de mala especificación .
  - Obtienen estimaciones más flexibles.
  - De utilidad en inferencia paramétrica y facilitan la selección de un modelo.
  - Requieren la selección de un parámetro de suavizado (ventana).
  - Se propone utilizar el estimador lineal local por sus propiedades teóricas (reducción efecto frontera) y son más fáciles de implementar (paquete `npsp` de `R`).

Estimación NP de la tendencia

Estimador lineal local multivariante:

- Se obtiene por suavizado lineal de los datos $(\mathbf{x}_i, Y(\mathbf{x}_i))$, tal que:

$$\hat{\mu}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^t \left(\mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x}\right)^{-1} \mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{Y} = s_\mathbf{x}^t \mathbf{Y}, \tag{3}$$

  - $\mathbf{e}_1 = (1, 0, .., 0)^t$.
  - $\mathbf{X}_\mathbf{x}$ matriz cuya i-ésima fila es $(1, (\mathbf{x}_i - \mathbf{x})^t)$.
  - $\mathbf{W}_\mathbf{x} = diag\left\{K_\mathbf{H}(\mathbf{x}_1 - \mathbf{x}), \ldots, K_\mathbf{H}(\mathbf{x}_n - \mathbf{x})\right\}$,
  - $K_\mathbf{H}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, donde K es una función tipo núcleo d-dimensional.
  - $\mathbf{H}$ es una matriz definida positiva de orden d, y representa el parámetro de suavizado o ventana.

- Matriz de suavizado $\mathbf{S}$: matriz $n \times n$ con $s_{\mathbf{x}_i}^t$ en la fila i tal que: $\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$.

Criterios para selección de la ventana

- Validación cruzada tradicional (VC), suponiendo independencia:

$$CV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}(Y(\mathbf{x}_i) - \hat{m}_{-i}(\mathbf{x}_i))^2$$

siendo $\hat{m}_{-i}(\mathbf{x}_i)$ la estimación obtenida eliminando el dato i.

- VC generalizada con corrección de sesgo para dependencia (Francisco-Fernández y Opsomer, 2005):

$$CGCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)}{1 - \frac{1}{n}tr\,(\mathbf{SR})}\right)^2$$

siendo $\mathbf{R}$ la matriz de correlaciones (estimada).

Influencia de la ventana en la estimación NP

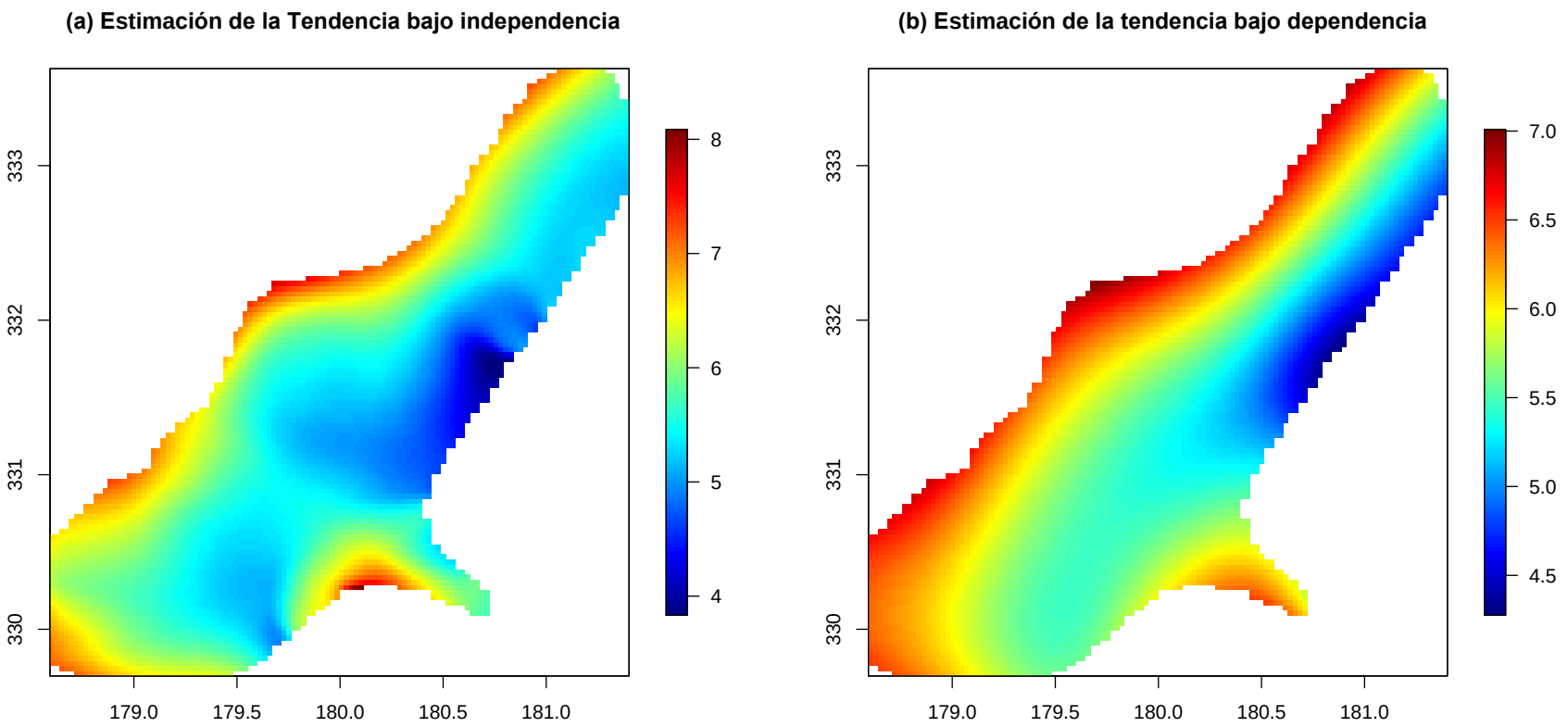

Figura 3. (a) Estimación NP de la tendencia de Log(Zinc) con ventana $\mathbf{H} = diag(0{,}5329, 0{,}5683)$ *bajo independencia y (b) Estimación NP de la tendencia de Log(Zinc) con ventana* $\mathbf{H} = diag(1{,}0945, 1{,}5631)$ *bajo dependencia de los datos.*

- La matriz ventana **H** controla el grado de suavizado de la estimación lineal local.

Estimación NP del variograma

- Se realiza a partir de los residuos: $r(\mathbf{x}_i) = Y(\mathbf{x}_i) - \hat{\mu}(\mathbf{x}_i)$.
- Si la media se supone constante: $r(\mathbf{x}_i) = Y(\mathbf{x}_i)$
- Puede verse como un caso particular de regresión:

$$\gamma(\mathbf{u}) = \frac{1}{2}\mathbb{E}(\varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x}+\mathbf{u}))^2$$

- La estimación se puede obtener por suavizado lineal utilizando (3) sobre los datos $(||\mathbf{u}||, \frac{1}{2}(r(\mathbf{x}) - r(\mathbf{x}+\mathbf{u}))^2)$, usando una ventana seleccionada por VC.
- Estos estimadores no son condicionalmente definido-negativos (no es factible realizar predicciones kriging), por lo que se debe ajustar un modelo de variograma válido.
- Modelo "no paramétricos"de Shapiro - Botha (paquete *npsp*).

Corrección del sesgo de variograma

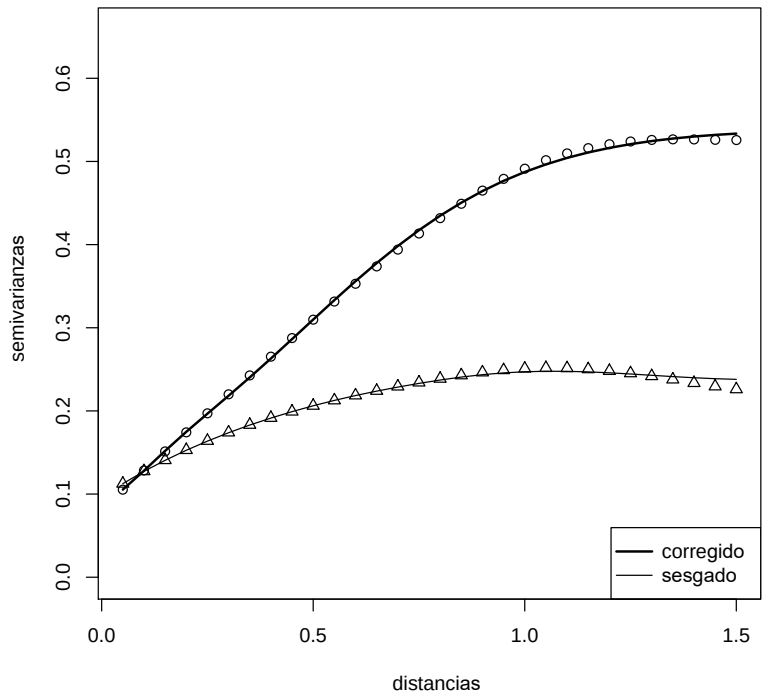

Figura 4. Variograma sesgado y corregido de los residuos estimados luego de eliminar la tendencia del Log(Zinc).

- La corrección del sesgo se realiza mediante un proceso iterativo basado en la relación: $\mathbf{\Sigma_r} = \mathbf{\Sigma} + \mathbf{S\Sigma S}^t - \mathbf{\Sigma S}^t - \mathbf{S\Sigma}$, siendo $\mathbf{\Sigma_r}$ la matriz de covarianza de los residuos.

Inferencias sobre el proceso espacial

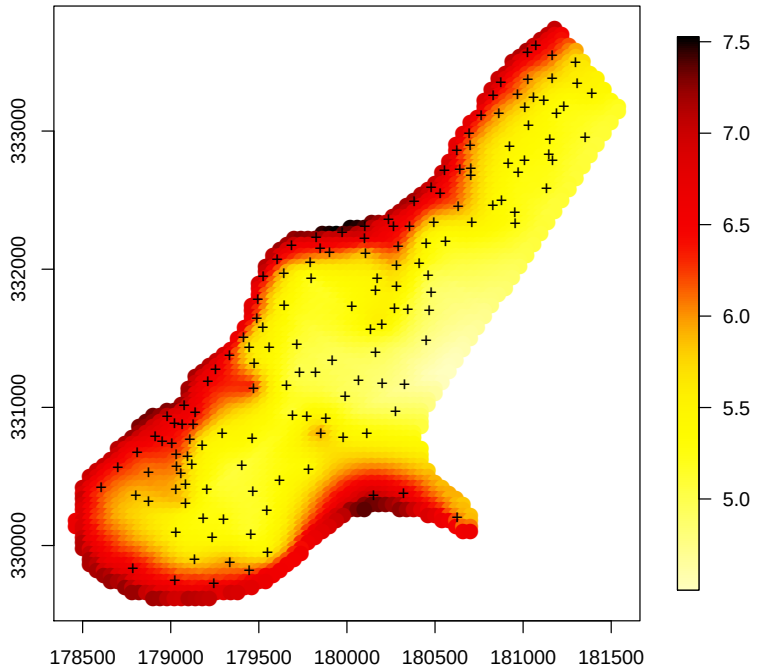

Figura 5. Predicción de Log(Zinc) mediante kriging no paramétrico.

- Se pueden construir predicciones, intervalos de confianza, mapas de riesgo, etc.
- Aplicaciones en minería, monitoreo ambiental, procesamiento de imágenes satelitales, meteorología, etc.

Nuevos criterios para la selección de ventana

- Propuestas para seleccionar $\mathbf{H}$ del estimador lineal local de la tendencia de un proceso espacial (Fernández-Casal, Castillo-Páez y García-Soidán, 2016):

$$CCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{m}_{-i}(\mathbf{x}_i)\right)^2 + \frac{2}{n}tr\left(\mathbf{S}_{-N_1}\boldsymbol{\Sigma}\right)$$

  o más generalmente:

$$CMCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{m}_{-N(i)}(\mathbf{x_i})\right)^2 + \frac{2}{n}tr\left(\mathbf{S}_{-N}\boldsymbol{\Sigma}\right)$$

  para algún vecindario N y siendo $\boldsymbol{\Sigma}$ la matriz de covarianzas de $\mathbf{Y}$.

Método Bootstrap No Paramétrico (NPB)

- Realizar inferencias sobre la variabilidad del estimador lineal local del variograma de un proceso espacial con y sin tendencia.
- Proponer un método bootstrap, basado en la descomposición de Cholesky de la matriz de covarianzas $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^t$.

1. A partir de una ventana $\mathbf{H}$, obtener el estimador lineal local de la tendencia, tal que: $\hat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$.
2. Calcular los residuos $\mathbf{r} = \mathbf{Y} - \mathbf{S}\mathbf{Y}$ y estimar el variograma sesgado $\hat{\gamma}(\mathbf{u})$
3. Obtener $\hat{\mathbf{\Sigma}}_r$ y $\mathbf{L}_r$ ajustando un Modelo Shapiro - Botha a $\hat{\gamma}(\mathbf{u})$.
4. Corregir el sesgo de $\hat{\mathbf{\Sigma}}_r$ para así obtener $\hat{\mathbf{\Sigma}}$ y $\mathbf{L}$.
5. Generar $\mathbf{e}^*$ por remuestreo independiente de $\mathbf{e} = \mathbf{L}_\mathbf{r}^{-1}\mathbf{r}$.
6. Construir la muestras bootstrap $\mathbf{Y}^* = \mathbf{S}\mathbf{Y} + \mathbf{r}^*$, donde $\mathbf{r}^* = \mathbf{L}\mathbf{e}^*$.

Método Bootstrap No Paramétrico (NPB)

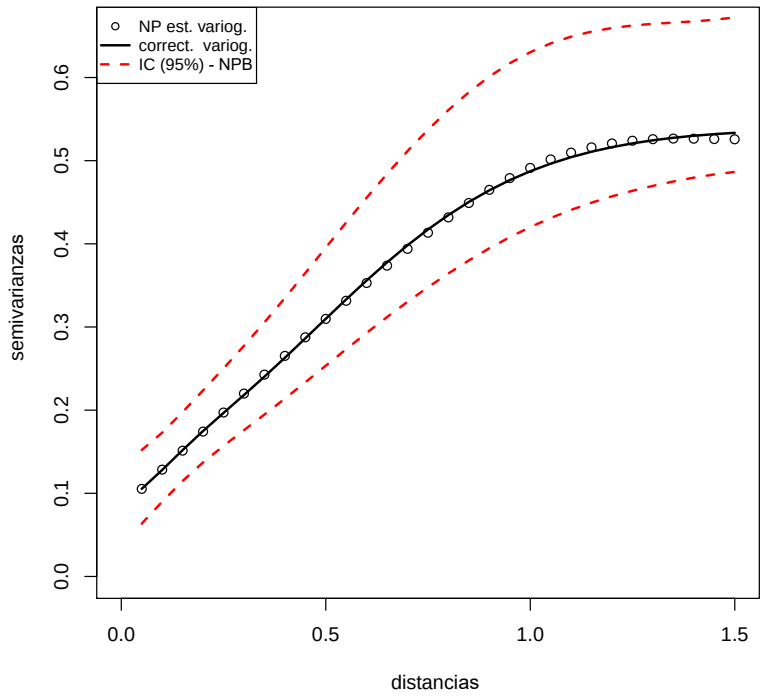

Figura 6. Intervalo de confianza al 95 % para el estimador lineal local del variograma de los residuos de datos "meuse", utilizando el NPB.

- Estudios numéricos muestran que el NPB conduce a mejores resultados que otros métodos bootstrap como el MBB o SPB (Castillo-Páez, S., Fernández-Casal, R., y García-Soidán, P., 2016).

Mapas de riesgo geoestadístico no paramétrico

- A partir del NPB es factible contruir mapas de riesgo basados en la predicción kriging.
- Se estima a partir de la probabilidad de que la variable Y exceda un valor crítico c en un ubicación específica $\mathbf{x}_0$:

$$r_c(\mathbf{x}_0) = P(Y(\mathbf{x}_0) \geq c).$$

- El proceso propuesto por Fernández-Casal, R., Castillo-Páez, S., y Francisco-Fernández, M. (2016), implica:
  1. Aplicar el método NPB para construir B réplicas bootstrap $Y^*(\mathbf{x}_i), i = 1, \ldots, n$ del proceso espacial original.
  2. Obtener la predicción kriging $\hat{Y}^*(\mathbf{x}_0)$ en cada localización no muestreada $\mathbf{x}_0$ a partir de cada muestra bootstrap.
  3. El mapa para $r_c(\mathbf{x}_0)$ se construye mediante las frecuencias observadas en las que $\hat{Y}^*(\mathbf{x}_0) \geq c$.

Mapas de riesgo geoestadístico no paramétrico

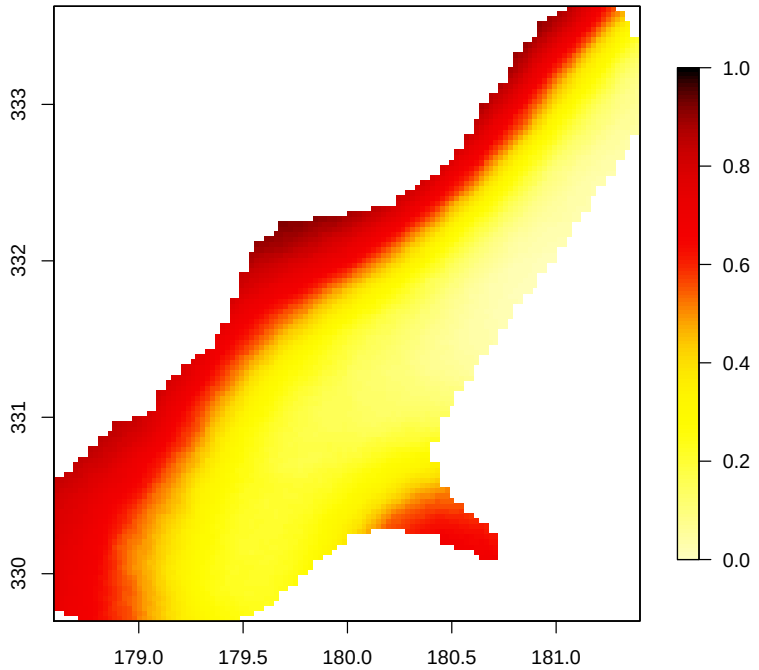

Figura 7. Mapa de probabilidad estimada de observar una concentración de log(zinc) mayor o igual a un valor crítico de $c = 6{,}0$ ppm.

Estimación NP en procesos espaciales heterocedásticos

- Se considera el modelo:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x})\varepsilon(\mathbf{x}),$$

  - $\mu(\cdot)$ función tendencia, $\sigma(\cdot)$ función varianza (determinísticas).
  - $\varepsilon(\cdot)$ proceso de error estacionario de segundo orden, de media cero, varianza unitaria y correlograma:

$$\rho(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$$

  - En este caso: $\gamma(\mathbf{u}) = 1 - \rho(\mathbf{u})$.

- Objetivo: Estimar no paramétricamente las características del proceso, i.e., $\hat{\mu}(\mathbf{x})$, $\hat{\sigma}(\mathbf{x})$ y $\hat{\gamma}(\mathbf{u})$.
- Se proponen nuevos estimadores para $\hat{\sigma}(\mathbf{x})$ y una modificación del proceso iterativo para la corrección del sesgo del variograma bajo heterocedasticidad.

MUCHAS GRACIAS

Referencias

- Fernández-Casal R, Francisco-Fernández M (2014) Nonparametric bias-corrected variogram estimation under non-constant trend. Stoch Environ Res Risk Assess 28.
- Francisco-Fernández M, Opsomer JD (2005) Smoothing paremeter selection methods for nonparametric regression with spatially correlated errors. Canadian J Stat 33:279-295.
- García-Soidán, P., González-Manteiga, W., Febrero-Bande, M. (2003) Local linear regression estimation of the variogram. Stat Prob Lett 64.
- Pebesma, E.J. (2004) Multivariable geostatistics in S: the gstat package. Computers & Geoscience 30: 683-691.

Nuevas contribuciones pendientes de publicación.

- Fernández-Casal, R., Castillo-Páez, S. y García-Soidá, P. (2016) Bandwidth selection for local linear trend estimation, presentado en el Congreso Internacional METMA 2104, Turín, Italia.
- Castillo-Páez, S., Fernández-Casal, R., y García-Soidán, P. (2016) Bootstrap methods for inference on the variogram, presentado en el Congreso SEIO 2015, Pamplona, España.
- Fernández-Casal, R., Castillo-Páez, S., y Francisco-Fernández, M. (2016) Nonparametric geostatistical risk mapping, presentado en el Congreso Internacional METMA 2104, Turín, Italia.