Geoestadística No Paramétrica: Representando información espacial de forma flexible

Ing. Sergio Castillo Páez, PhD.

Universidad de las Fuerzas Armadas ESPE, Ecuador
Universidad de Vigo, España

29 Junio 2018

Contenido

Introducción

- **Geoestadística:** Proporciona modelos y métodos para el análisis de información de procesos continuos asociados a posiciones espaciales.
- **Algunos ejemplos:**
  - Medioambiente: Mapas de Riesgo de Contaminación (Fernandez Casal et al, 2017)
  - Econometría espacial: *Estudios de precios de bienes raíces, crecimiento urbano, selección de localización de negocios, etc.* (G. Arbia y H. Baltagi, 2009)
  - Finanzas: *Optimización de portafolios de inversión mediante kriging* (C. Oliveira et al., 2012)
  - Marketing: *Geomarketing, análisis de políticas de marketing* (E. Bradlow et al., 2005)
  - En estudios sociales: *Mapas de tasas de crimen en una ciudad* (G. Fernández-Avilés Calderón, 2008)

Idea clave: Observaciones más cercanas son similares entre sí, y a medida que la separación entre ellas aumenta, su correlación disminuye (*dependencia espacial*).

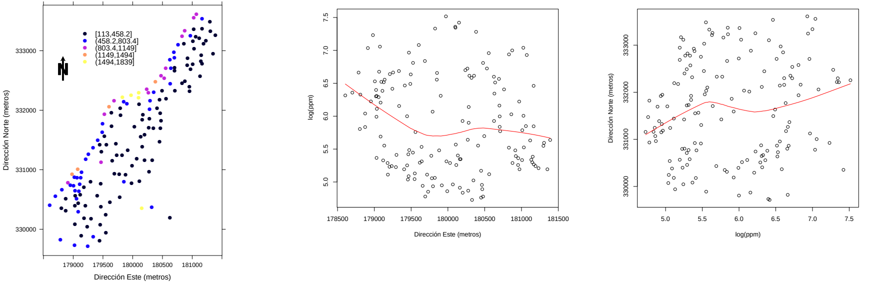


Fig 1: Zinc (ppm) en las riberas del río Meuse. (Fig. Izq.), Dispersión respecto a la dirección Este (Fig. Cent.) y Norte (Fig. Der.) y curvas de regresión no paramétrica.

```r
library (sp); data(meuse)
coordinates(meuse)= ~x+y
spplot(meuse, "zinc")
plot(meuse@coords[,1], meuse@data$zinc, xlab = "Direccion Este", ylab = " log(ppm)")
lines(lowess(meuse@coords[,1], meuse@data$zinc), col = 2)
```

Modelización geoestadística

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

Modelo con tendencia no constante:

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

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

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

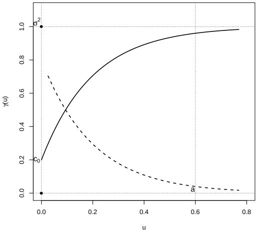
- Usualmente, la dependencia se modela a través del variograma:

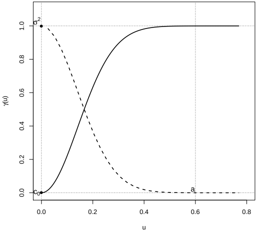
$$\gamma(\mathbf{u}) = \frac{1}{2} Var(\varepsilon\left(\mathbf{x}\right) - \varepsilon\left(\mathbf{x} + \mathbf{u}\right)) \tag{2}$$

- $C(\mathbf{u})$ y $\gamma(\mathbf{u})$ dependen solo del salto u, además:

$$\gamma(\mathbf{u}) = \sigma^2 - C(\mathbf{u}).$$

- **Efecto nugget (c_0):** Comportamiento cerca del origen.
- **Umbral (σ^2):** Comportamiento en el límite del salto.
- **Umbral parcial (c_1):** $\sigma^2 - c_0$.
- **Rango práctico (a):** Distancia mínima a la cual $\gamma(\mathbf{u}) = 0.95\sigma^2$.
- **Anisotropía:** Cuando la forma de $\gamma(\mathbf{u})$ dependiendo de la dirección de $\mathbf{u}$.
  Si $\gamma(\mathbf{u}) = \gamma(||\mathbf{u}||)$ se dice *isotrópico*.





Ejemplos de variogramas y covariogramas.

Estimación Paramétrica

- *Regresión kriging:* Estimación paramétrica de la tendencia y la dependencia basada en residuos.

  1. Obtener una estimación inicial de la tendencia $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\beta}_{mco}$.
  2. Calcular los residuos o errores estimados: $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\boldsymbol{\mu}}$.
  3. Ajustar un variograma válido $\hat{\gamma}(\mathbf{u})$ construido a partir de $\hat{\boldsymbol{\varepsilon}}$.
  4. Se reestima la función $\mu(\cdot)$:

  $$\hat{\boldsymbol{\beta}}_{mcge} = (\mathbf{X}^t\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}$$

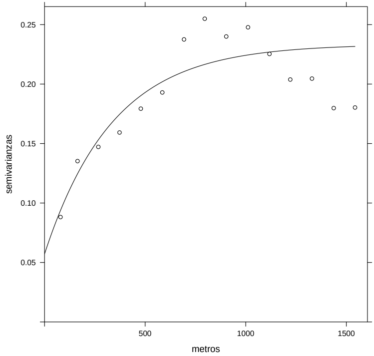
  donde $\hat{\boldsymbol{\Sigma}}$ se obtiene a partir de $\hat{\gamma}(\mathbf{u})$.

- Este proceso se puede repetir de forma iterativa (Algoritmo Neuman - Jacobson)

Estimación Paramétrica
Estimación del variograma

Para estimar $\gamma(\cdot)$ se recurre al *análisis estructural:*

1. Obtener un estimador piloto no paramétrico (p.e. estimador empírico).
2. Selección y ajuste de un modelo válido de variograma (p.e. modelo exponencial)
3. Diagnosis del variograma ajustado (p.e. por Validación cruzada)



Estimación paramétrica del variograma de los residuos a partir de un modelo Exponencial

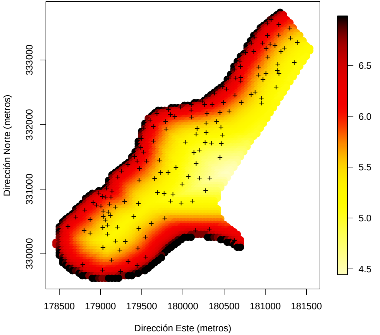
```
library (gstat)
vgm <- variogram(log(zinc) ~ sqrt(dist), meuse)
fit.vgm <- fit.variogram(vgm, model = vgm(1, "Exp", 300, 1))
plot(vgm, fit.vgm, main = " ", xlab = "metros", ylab="semivarianzas", col = 1)
```

Enfoque Paramétrico

- Expuestos a problemas de mala especificación.
- La variabilidad de los residuos subestima la variabilidad del proceso.

Enfoque No Paramétrico

- Obtienen estimaciones más flexibles.
- Facilitan la selección de un modelo.
- Requieren la selección de un parámetro de suavizado (ventana).



Tendencia estimada paramétricamente, a partir de la distancia a la orilla del río (log ppm zinc)

```
model.gstat <- gstat(formula = log(zinc) ~ sqrt(dist), data = meuse, model = fit.vgm)
pred.gstat <- predict(model.gstat, meuse.grid, BLUE = TRUE) #prediccion de tendencia por
      minimos cuadrados generalizados
trend.pred <- pred.gstat$var1.pred
spoints(meuse.grid$x, meuse.grid$y, trend.pred)
```

Se obtiene como la solución para α del siguiente problema:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left\{ Y(\mathbf{x}_i) - \alpha - \beta^T(\mathbf{x}_i - \mathbf{x}) \right\}^2 K_{\mathbf{H}}(\mathbf{x}_i - \mathbf{x}).$$

De forma explícita:

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

donde $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $\mathbf{X}_{\mathbf{x}}$ es la matriz cuya i-ésima fila es igual a $\left(1, (\mathbf{x}_i - \mathbf{x})^t\right)$, $\mathbf{W}_{\mathbf{x}} = diag\left\{K_{\mathbf{H}}(\mathbf{x}_1 - \mathbf{x}), \ldots, K_{\mathbf{H}}(\mathbf{x}_n - \mathbf{x})\right\}$, $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, siendo K una función tipo núcleo d-dimensional, y $\mathbf{H}$ la matriz ventana $d \times d$ simétrica no singular.

Se puede expresar como suavizado lineal de los datos $(\mathbf{x}_i, Y(\mathbf{x}_i))$:

$\hat{\boldsymbol{\mu}} = \mathbf{SY}$, donde $s_{\mathbf{x}}^t$ es la i-ésima fila de la matriz de suavizado $\mathbf{S}$.

Observaciones

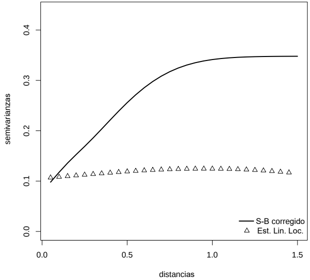
- La ventana **H** se selecciona mediante criterios que deben tomar en cuenta la dependencia espacial
- El estimador lineal local reduce el efecto frontera, y también se puede aplicar para estimar el variograma.
- Se cuenta con un procedimiento NP para corregir el sesgo en la variabilidad de los residuos
- Funciones programadas en el paquete *npsp* de *R*

Simulación del efecto del sesgo en el variograma, debido al uso de residuos

Aplicación a datos: Meuse

Estimación lineal local $\hat{\mu}$ piloto y final

Estimación $\hat{\gamma}$ residual y Modelo S-B ajustado a $\tilde{\gamma}$



- Ejemplo completo de aplicación del paquete **npsp** en **R**:
- Datos de precipitación mensual en EEUU
- `https://rubenfcasal.github.io/npsp/articles/npsp.html`

Referencias

- Cressie, N. (1993) *Statistics for Spatial Data.* Wiley.
- Fernández-Casal R, Francisco-Fernández M (2014) *Nonparametric bias-corrected variogram estimation under non-constant trend*. Stoch Environ Res Risk Assess 28.
- Fernández-Casal, R., Castillo-Páez, S., y Francisco-Fernández, M. (2017) *Nonparametric geostatistical risk mapping*, Stoch Environ Res Risk Assess.
- Pebesma, E.J. (2004) *Multivariable geostatistics in S: the gstat package*. Computers & Geoscience 30: 683-691.

GRACIAS