Minicurso: Estadística para datos Espacio-temporales

Dra. Carolina Euán

https://www.cimat.mx/~caroe/

Cómo son los Datos Espacio-temporales

- Características: Datos con dependencia temporal y espacial

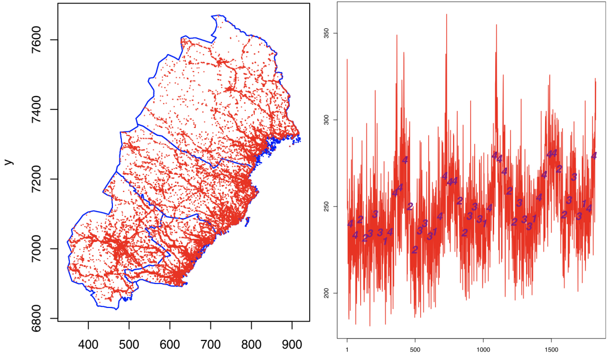
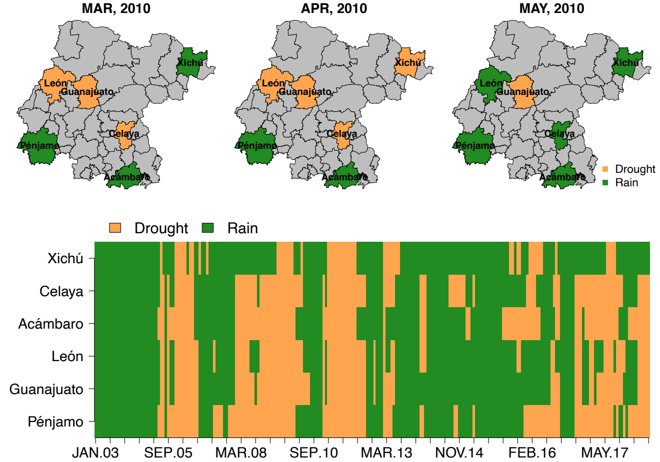


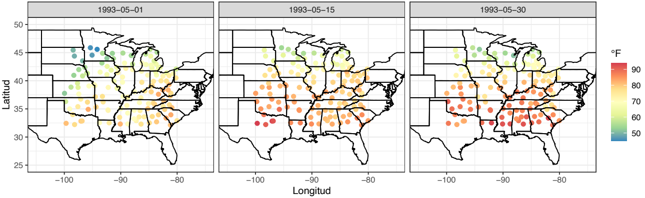
Fig. 1. The spatial and temporal components of the spatio-temporal emergency alarm call data. The left panel shows the spatial locations of the calls while the right panel presents the call counts over the study region as a function of time. The scales of measurements for the spatial locations and the temporal count data are meter (scaled by 1000) and day, respectively. The numbers in the right panel plot indicate the smoothed seasons-of-a-year, which are Fall (1), Spring (2), Summer (3), and Winter (4), of the emergency alarm calls; the first time point corresponds to January 1, 2014.

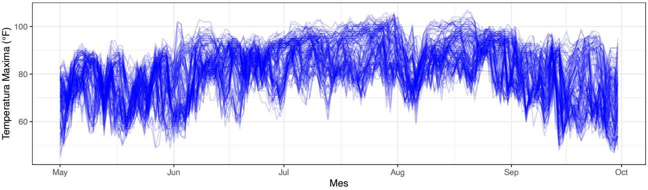
Eventos de Sequía en Guanjauato



Temperatura en EU

Temperatura máxima diaria en el este de EU durante Mayo-Septiembre 1993





Contenido

Notacion Inicial

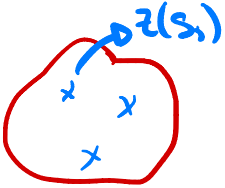
Referencia: Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019), Spatio-Temporal Statistics with R, Boca Raton, FL: Chapman & Hall/CRC

s - ubic esp

- $Z(\mathbf{s})$ denota un proceso espacial.
- $Z(t, \mathbf{s})$ denota un proceso espacio-temporal.

$\rightarrow$ Comenzaremos estudiando datos espaciales y en la ultima sesión lo expandiremos a datos espacio temporales.

Cómo son nuestros Datos?



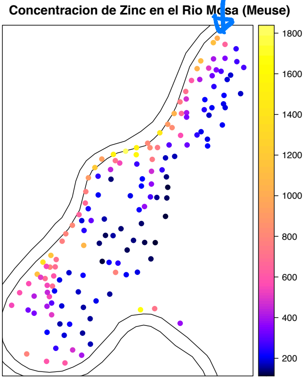
- m ubicaiones, $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m$.
- T tiempos, $t_1, \ldots, t_T$.
- Muestra (datos espaciales): $\{Z(\mathbf{s}_1), Z(\mathbf{s}_2), \ldots, Z(\mathbf{s}_m)\}$

Datos del Rio Mosa

El Río Mosa es un importante río Europeo

```
> library(sp)
> data(meuse)
> data(meuse.riv)
```

En este caso: $m = 155$ y $Z(\mathbf{s}_i)$ es la concentración de Zinc.



11

s_i

$Z(s_i)$

Contenido

Midiendo Dependencia Espacial

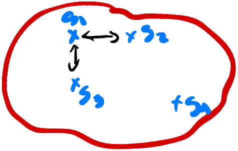
Sea $Z(\mathbf{s})$ un proceso espacial definimos su función de covarianza como

$$\mathrm{Cov}(\mathbf{s}_1, \mathbf{s}_2) = \mathrm{Cov}(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) = \mathbb{E}[\{Z(\mathbf{s}_1) - \mu\}\{Z(\mathbf{s}_2) - \mu\}]$$

donde $\mu = \mathbb{E}(Z(\mathbf{s}))$.

Lo más común es tener una muestra de tamaño 1. Entonces necesitamos ciertos supuestos:

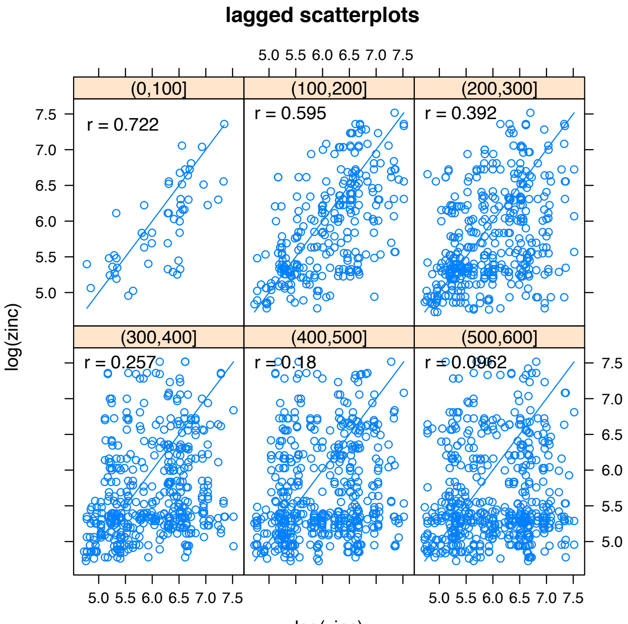
1. Isotropia: $\mathrm{Cov}(\mathbf{s}_1, \mathbf{s}_2) = \mathrm{Cov}(|\mathbf{s}_1 - \mathbf{s}_2|)$, $|\cdot|$ denota norma euclideana.



$$\mathrm{Cov}(Z(s_1), Z(s_2))$$
$$= \mathrm{Cov}(Z(s_1), Z(s_3))$$
$$\text{si} \quad |s_1 - s_2| = |s_1 - s_3|$$

Datos del Rio Mosa

```
> hscat(log(zinc) ~ 1, meuse, (0:6) * 100)
```



$(z(s_i), z(s_j))$ si $|s_i - s_j| = h$

Variograma

- La estrategia es transformar los datos

$V(X) = E[X^2] - (E[X])^2.$

- Definimos $h = |\mathbf{s}_1 - \mathbf{s}_2|$, luego

$$\gamma_Z(h) = 1/2\ \mathbb{E}(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2$$

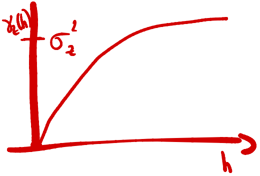
$$\mathbb{E}(Z(s_1) - Z(s_2))^2 = Var(Z(s_1) - Z(s_2))$$

$$= Var(Z(s_1)) + Var(Z(s_2)) - 2Cov(Z(s_1), Z(s_2))$$

$$= 2\sigma_Z^2 - 2Cov(Z(s_1), Z(s_2))$$

$Var(Z(s_1)) = Var(Z(s_2)) = \sigma_Z^2$

$$\Rightarrow \gamma_Z(h) = \sigma_Z^2 - Cov(s_1, s_2)$$



Variograma

- Entonces, $\gamma_Z(h) = \sigma^2 - \text{Cov}(\mathbf{s}_1, \mathbf{s}_2)$, donde $\sigma^2 = Var(Z(\mathbf{s}))$

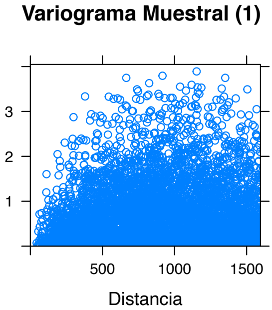
- Ahora tenemos mas valores muestra de $Z(\mathbf{s}_1) - Z(\mathbf{s}_2)$.

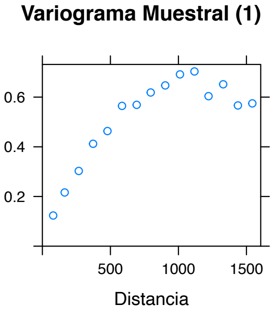
- Asi el estimador muestral del variograma es

$$\hat{\gamma_Z}(h) = \frac{1}{2 * N_h} \sum_{i=1}^{N_h} (Z(\mathbf{s}_i) - Z(\mathbf{s}_i + h))^2$$

Datos del Rio Mosa

```
> plot(variogram(log(zinc) ~ 1, meuse,cloud = TRUE),
+       xlab="Distancia",ylab="",
+       main="Variograma Muestral (1)")
> plot(variogram(log(zinc) ~ 1, meuse),
+       xlab="Distancia",ylab="",
+       main="Variograma Muestral")
```





Variograma

- Entonces, $\gamma_Z(h) = \sigma^2 \cancel{+} \mathrm{Cov}(\mathbf{s}_1, \mathbf{s}_2)$, donde $\sigma^2 = Var(Z(\mathbf{s}))$

- Asi el estimador muestral del variograma es

$$\hat{\gamma_Z}(h) = \frac{1}{2 * N_h} \sum_{i=1}^{N_h} (Z(\mathbf{s}_i) - Z(\mathbf{s}_i + h))^2$$

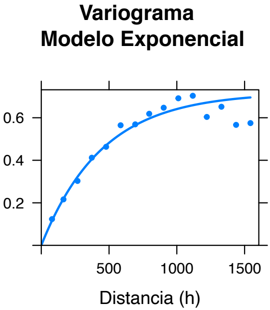
- NOTA - Necesitamos que la funcion de covarianza estimada sea no-negativa definida. Para ello, una version muestral del variograma no es suficiente!

$a_1, a_2 \in \mathbb{R}^p$, M

$a_1 M a_2 \geq 0$.

Datos del Rio Mosa

```
> v<-variogram(log(zinc) ~ 1, meuse)
> model<-fit.variogram(v, vgm(1, "Exp", 500, 0))
> plot(v, model=model,pch=20,lwd=2,xlab="Distancia (h)",
+       ylab="",main="Variograma - Modelo Exponencial")
```



$\gamma_Z(h) = \sigma_Z^2 + Cov(h)$

$a\exp\{-|\mathbf{s}_1 - \mathbf{s}_2|/b\} = Cov(\mathbf{s}_1, \mathbf{s}_2; a, b)$

$\rightarrow \underbrace{\sum_h \left(\hat{\gamma}_Z(h) - \gamma_Z(h; a, b)\right)^2}_{\hat{a},\ \hat{b}\quad LS.}$

Contenido

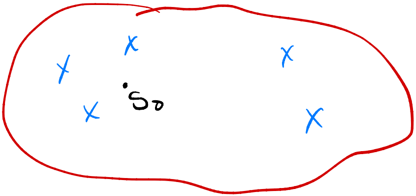
Predicción

- En el estudio de datos espaciales, un objetivo es poder predecir en ubicaciones donde no observamos datos.

- El objetivo es construir un predictor lineal que use la información observada.

$$Z(\mathbf{s}_0) = \lambda_0 + \sum_{i=1}^{m} \lambda_i Z(\mathbf{s}_i)$$

- Se necesita conocer la covarianza entre $Z(\mathbf{s}_0)$ y cada $Z(\mathbf{s}_i)$.



Mejor predictor lineal

$\to$ Menor error $\mathbb{E}\left(Z(s_0)-\hat{Z}(s_0)\right)^2 \to$ mínimo.

$$Z=\begin{pmatrix} Z(s_1) \\ \vdots \\ Z(s_m) \end{pmatrix}$$

$$\mathbb{E}\left(Z(s_0)-\lambda_0-\lambda^T Z\right)^2$$

$$\lambda=\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix}$$

$$= Cov\left(Z(s_0)-\lambda_0-\lambda^T Z,\ Z(s_0)-\lambda_0-\lambda^T Z\right)$$

$$+\left[\mathbb{E}\left(Z(s_0)-\lambda_0-\lambda^T Z\right)\right]^2 \quad , \quad \mathbb{E}(Z(s)) = \mu(s)$$

$$=\left(\mu(s_0)-\lambda_0-\lambda^T\mu\right)^2 + k_0 - 2\lambda^T k + \lambda^T K \lambda$$

$$\to K^{-1} \checkmark \qquad \Rightarrow \qquad \left.\begin{aligned} \lambda &= K^{-1}k \\ \lambda_0 &= (s_0) - k^T K^{-1}\mu \end{aligned}\right\} \checkmark$$

$$K = Cov(Z, Z) \quad , \quad k = Cov(Z(s_0), Z)$$

Mejor predictor lineal -Kriging

$$Z(s) = \underbrace{X(s)^T \beta}_{\mu(s)} + \varepsilon(s)$$

media "0"

$$\beta_0 = \mu$$

$$\rightarrow \mathbb{E}(\lambda_0 + \lambda^T Z) = \mathbb{E}(Z(s_0))$$

$$\Leftrightarrow \lambda_0 + \lambda^T \underline{\overline{X}} \beta = X(s_0)^T \beta$$

Se encuentra λ - $\mathcal{J}$: es BLUP

$$\lambda^T Z = k^T K^{-1} (Z - \underline{\overline{X}} \hat{\beta}) + X(s_0)^T \hat{\beta}$$

Mejor predictor lineal -Kriging

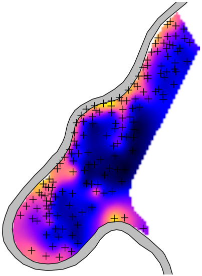
$$\hat{\beta} \quad GLS \;, \quad (M^T K^{-1} M)^{-1} M^T K^{-1} Z$$

$$M = \underline{\overline{X}} .$$

Kriging Ordinario - Rio Mosa

El Río Mosa es un importante río Europeo

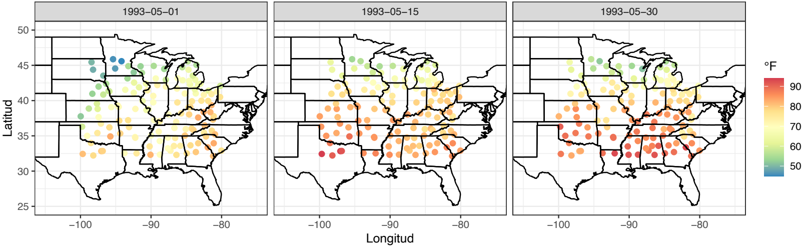
```
> krige(log(zinc) ~ 1,
+ meuse, meuse.grid, model)
```



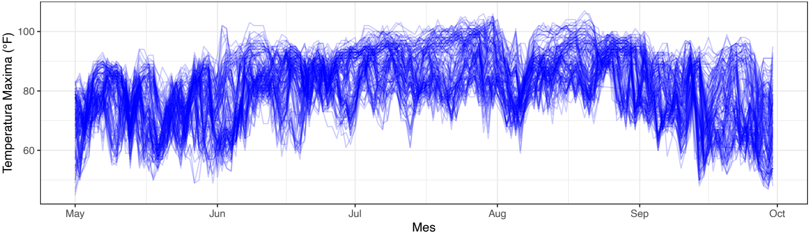
Contenido

Temperatura en EU - Espacial

```
> data("NOAA_df_1990", package = "STRbook")
```



Temperatura en EU - Temporal



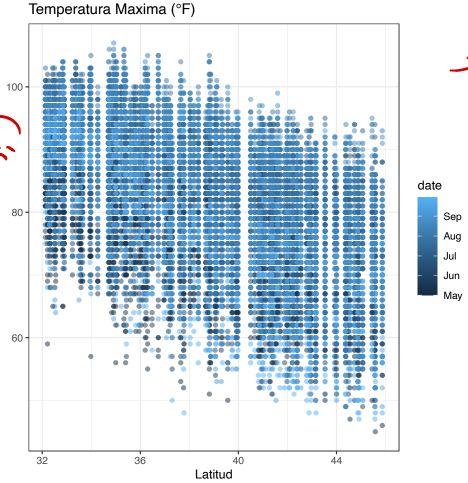
Temperatura en EU - Efecto de Latitud

$z(t, s_i)$

$\rightarrow S_i = (s_{1i}, s_{2i})$

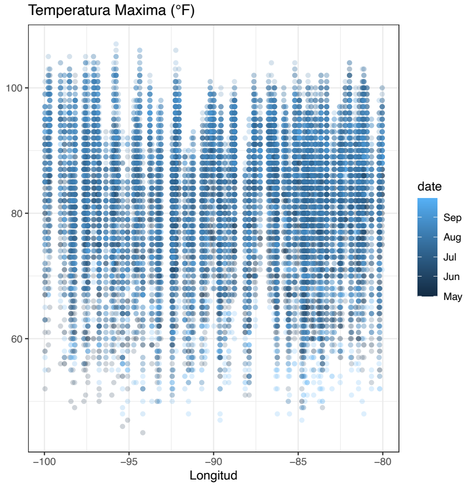
S

$z(t, s_i)$



$\rightarrow s_{1i}$

Temperatura en EU - Efecto de Longitud

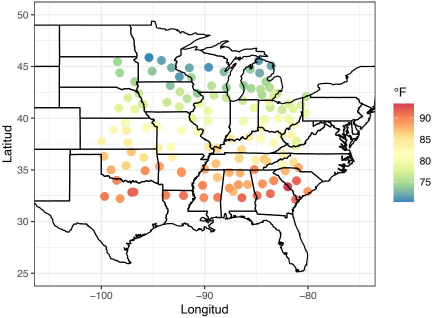


Contenido

Media Espacial

La media espacial muestral se calcula promediando en el tiempo

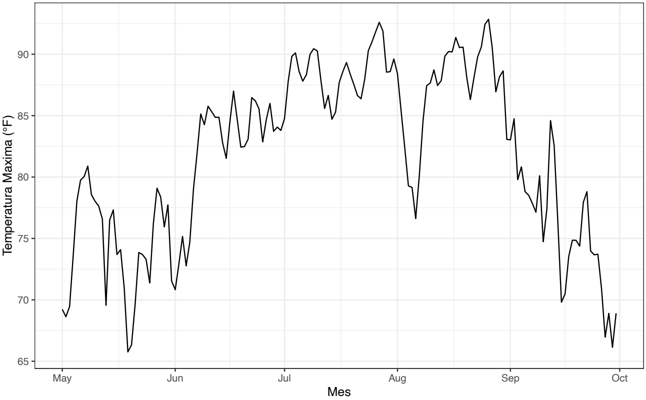
$$\hat{\mu}_z(\mathbf{s}_i) = \frac{1}{T}\sum_{j=1}^{T} Z(\mathbf{s}_i, t_j)$$




```
> library("dplyr")
> TmaxMean <- Tmax %>% group_by(id,lat,lon) %>%
+   summarise(z = mean(z))
```

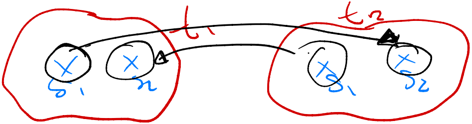
Serie de Tiempo de Promedios

$$\hat{\mu}_z(t_j) = \frac{1}{m}\sum_{i=1}^{m} Z(\mathbf{s}_i, t_j)$$




```
> TmaxMeanTS <- Tmax %>% group_by(date) %>%
+   summarise(z = mean(z))
```

Covarianza



La covarianza resulta,

$$\mathrm{Cov}_Z(Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2)) = \mathbb{E}[\{Z(\mathbf{s}_1, t_1) - \mu(\mathbf{s}_1, t_1)\}\{Z(\mathbf{s}_2, t_2) - \mu(\mathbf{s}_2, t_2)\}]$$

y el variograma,

$$\gamma_Z(\mathbf{s}_1, \mathbf{s}_2, t_1, t_2) = \frac{1}{2} var(Z(\mathbf{s}_1, t_1) - Z(\mathbf{s}_2, t_2))$$

Si asumimos isotropia y estacionariedad:

$$\gamma_Z(h, \tau) = \mathrm{Cov}_Z(0, 0) - \mathrm{Cov}_Z(h, \tau).$$

Media Espacial

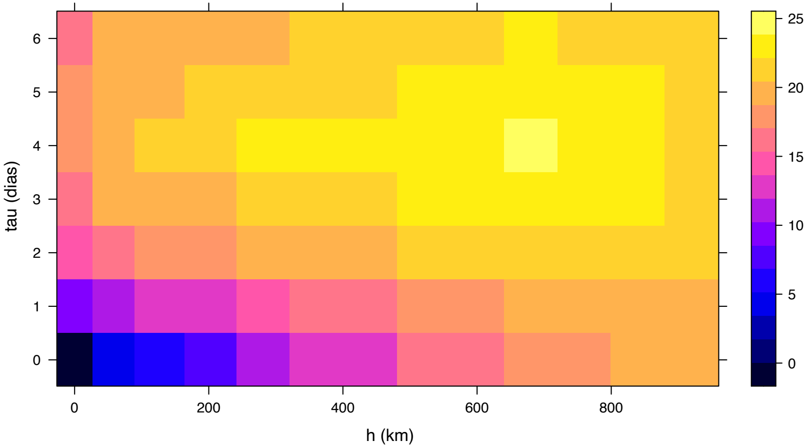
El estimador muestral resulta ser

$$\hat{\gamma}_Z(h,\tau) = \frac{1}{\#N_h}\frac{1}{\#N_\tau}\sum_{\mathbf{s}_i,\mathbf{s}_k\in N_h}\sum_{t_i,t_l\in N_\tau}(Z(\mathbf{s}_i,t_j) - Z(\mathbf{s}_k,t_l))^2$$

En R...

```
> data("STObj3", package = "STRbook")
> STObj4 <- STObj3[, "1993-07-01::1993-07-31"]
> VTmax<-variogram(object = z~1 +lat,
+             data = STObj4, # Julio 93
+             width = 80, # spatial bin (80 km)
+             cutoff = 1000, # pts < 1000 km
+             tlags = 0.01:6.01)
```

Variograma para Temperaturas Max



Contenido

Modelos Estadísticos para datos Espacio-temporales

- Es este caso tendremos tres posibles obejetivos:
- 1- Predicción de un valor desconocido en una nueva ubicación pero en el mismo periodo de observación.
- 2- Inferencia del efecto de Covariables
- 3- Prediction de un valor futuro.

Predición lineal determinística (Interpolación)

- Muestra $\{Z(\mathbf{s}_1, t_1), \ldots, Z(\mathbf{s}_m, t_1), Z(\mathbf{s}_1, t_2), \ldots, Z(\mathbf{s}_m, t_2), \ldots, Z(\mathbf{s}_1, t_T), \ldots, Z(\mathbf{s}_m, t_T)\}$

- Estimador pesado con kernel

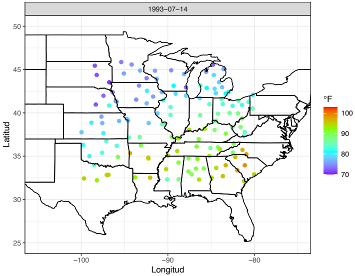
$$\hat{Z}(\mathbf{s}_0, t_0) = \sum_{j=1}^{T} \sum_{i=1}^{m} \omega_{ij}(\mathbf{s}_0, t_0) Z(\mathbf{s}_i, t_j),$$

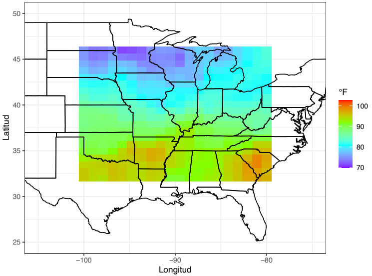
donde los pesos $\omega_{ij}(\mathbf{s}_0, t_0)$ es usualmente determinado por un kernel. Por ejemplo, el kernel Gaussiano radial,

$$\kappa_{ij}(\mathbf{s}_0, t_0; \theta) = \exp\left(-\frac{1}{\theta}|(\mathbf{s}_i, t_j) - (\mathbf{s}_0, t_0)|^2\right),$$

con $\omega_{ij}(\mathbf{s}_0, t_0) = \frac{\kappa_{ij}(\mathbf{s}_0, t_0; \theta)}{\sum_{j=1}^{T}\sum_{i=1}^{m} \kappa_{ij}(\mathbf{s}_0, t_0; \theta)}$

Predición lineal determinística (Interpolación)





Predición usando Regresión

$Y = X\beta + \varepsilon$

Consideramos un modelo lineal con p covariables:

$$Z(\mathbf{s}_i, t_j) = \beta_0 + \beta_1 X_1(\mathbf{s}_i, t_j) + \ldots + \beta_p X_p(\mathbf{s}_i, t_j) + \varepsilon(\mathbf{s}_i, t_j).$$

En un primer análisis asumimos que $\varepsilon(\mathbf{s}_i, t_j)$ son errores indepedientes con varianza σ_ε^2.

Predición usando Regresión

Consideramos un modelo lineal con las siguientes covariables:

- Media constante en $\mathbf{s}$ and t.
- Covariables en longitud ($\mathbf{s}_{1,i}$), latitud ($\mathbf{s}_{2,i}$).
- Tendencia lineal en tiempo, t_j.
- Y posibles interacciones tiempo-espacio

El modelo resultante es de la forma:

$$Z(\mathbf{s}_i, t_j) = \beta_0 + \beta_1 \mathbf{s}_{1,i} + \beta_2 \mathbf{s}_{2,i} + \beta_3 t_j + \beta_4 \mathbf{s}_{1,i} \mathbf{s}_{2,i} + \beta_5 \mathbf{s}_{1,i} t_j + + \beta_6 \mathbf{s}_{2,i} t_j + \varepsilon(\mathbf{s}_i, t_j).$$

Predición usando Regresión

En R...

```
> ModLin<- lm(z ~ (lon + lat + day)^2,     # model
+                data = select(TmaxJ, -id))
```

Nota: En realidad es un poco más complicado que un simple modelo de regresión.

Predición usando Kriging y Procesos Gaussianos

Aquí tambien consideraremos que podemos tener error de medicón.

- $Z(\mathbf{s}_i, t_j) = Y(\mathbf{s}_i, t_j) + \epsilon(\mathbf{s}_i, t_j)$, donde ϵ son iid.
- $Y(\mathbf{s}_i, t_j) = \mu(\mathbf{s}_i, t_j) + \eta(\mathbf{s}_i, t_j)$, μ representa la media que puede depender del tiempo y espacio y $\eta(\mathbf{s}_i, t_j)$ es un proceso Gaussiano espacio temporal con media zero.

S-T Predictor usando Kriging

Usando Notación Matricial

$\rightarrow Z = Y + \varepsilon$ $\Delta \sim \varepsilon$ error de medición en este caso $N(0, \sigma_\varepsilon^2)$ $\Longrightarrow$ Si asumimos $Y(s,t)$ es un proceso Gaussiano

$\rightarrow Y = \mu + \eta$

$$\begin{pmatrix} Y(s_0, t_0) \\ Z \end{pmatrix} \sim N\left(\underbrace{\begin{pmatrix} x(s_0, t_0)' \\ X \end{pmatrix}}_{1+m^T} \beta, \begin{bmatrix} C_{00} & C_0^T \\ C_0 & C_Z \end{bmatrix} \right)$$

Donde

$\rightarrow C_0 = \text{cov}(Y(s_0; t_0), Z)$ $\rightarrow C_{00} = \text{Var}(Y(s_0, t_0))$

$\rightarrow C_Z = \text{Cov}(Z) = C_y + C_\varepsilon$

Bajo el supuesto distribucional el mejor predictor lineal resulta ser $\mathbb{E}(Y(s_0, t_0) | Z)$

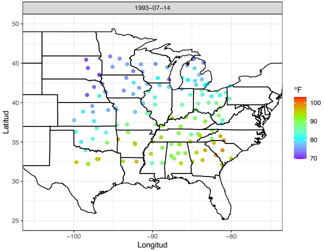
S-T Predictor usando Kriging

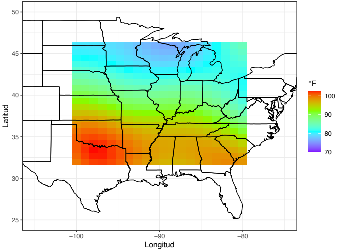
Entonces,

$$\hat{Y}(s_0, t_0) = x(s_0, t_0)'\beta + c_0' C_z^{-1}(Z - X\beta)$$

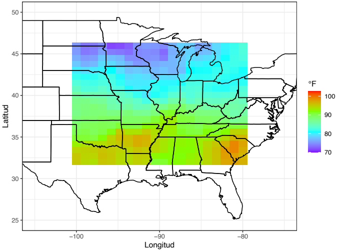
y $\sigma^2_{\tilde{Y},} = C_{00} - c_0^T C_z^{-1} c_0$

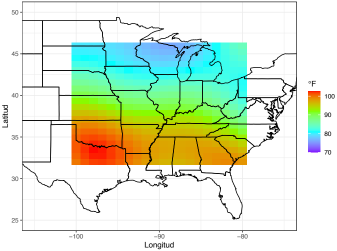
S-T Predictor usando Kriging





Comparación





Referencia

- Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019), Spatio-Temporal Statistics with R, Boca Raton, FL: Chapman & Hall/CRC
- Stein, M. (1999) Interpolation of Spatial Data - Some Theory for Kriging. Springer.

Carolina Euán
caroe@cimat.mx
https://www.cimat.mx/~caroe/

Home

TIES – GRASPA 2021

TIES-GRASPA 2021 CONFERENCE

https://sites.google.com/view/tiesgraspa2021/home

Unirse a TIES: https://www.isi-web.org/component/rsform/form/70-ties-membership-form-2020

WORKSHOP 1

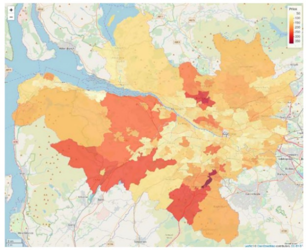
01-02 July 2021, 12-4pm

Bayesian spatial and spatio-temporal modeling with R-NIMBLE

by Tullia Padellini and Garyfallos Konstantinoudis, Imperial College London.

Registrations will open soon

WORKSHOP 2



13-14 September 2021, 12-4pm

Spatio-temporal modeling and visualisation of small-area data in R

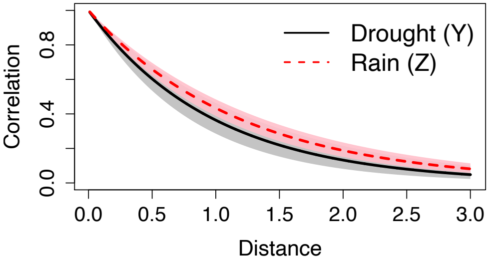
by Duncan Lee, University of Glasgow.

See here for the description of the workshop.

Registrations will open soon

Drought Data Analysis

- We fit the Ber VAR model with the spatial model with $\exp(\lambda||s_i - s_j||)$ as the correlation function.
- Under the Ber VAR model, $X_{i,t} = Y_{i,t}X_{i,t-1} + Z_{i,t}(1 - X_{i,t-1})$. Then, Y_t models the dynamics of the dry events and Z_t models the dynamics of the rain.
- Correlation function of the latent processes:



- Confident bands are computed by the asymptotic variance of the MLE estimators and the $MSE = 0.125$.

Drought Data Analysis

Cross correlation plots of X_t.