Análisis Estadístico Clásico y Robusto de Datos Espaciales

Salvador Huertas Amorós

Tutor: Dr. D. Alfonso García Pérez

Facultad de Ciencias
Universidad Nacional de Educación a Distancia

Trabajo presentado para la obtención del título de
Máster Universitario en Matemáticas Avanzadas de la
UNED.
Especialidad Estadística e Investigación Operativa

Julio 2018

para mi esposa Maridó ....
a quién tanto le debo

Abstract.

El análisis espacial de los datos estadísticos, con la aparición del GPS y su utilización en las aplicaciones informáticas de muy diverso tipo, se ha convertido en uso cotidiano. Aquí nos hemos centrado en el estudio del Variograma y el Kriging Ordinario, analizando las propiedades clásicas y robustas de diversos estimadores del primero y desarrollando su aplicación al problema de las precipitaciones en la Región de Murcia. También se aplica el método de Procesos Puntuales a la Sismicidad de la península Ibérica utilizando el paquete R.

The spatial analysis of statistical data, with the appearance of GPS and its use in computer applications of very different types, has become everyday use. Here we have focused on the study of the Variogram and Ordinary Kriging, analyzing the classic and robust properties of various estimators of the first and developing its application to the problem of rainfall in the Region of Murcia. The method of Point Processes to the Seismicity of the Iberian Peninsula is also applied using the R package.

Índice

1. Introducción.

La estadística de datos espaciales se centra en el estudio de fenómenos aleatorios que ocurren distribuidos en una región del espacio, normalmente de dos o tres dimensiones, con un objetivo primario de modelización probabilística, para poder hacer posteriormente inferencias y predicciones que ayuden a la toma de decisiones en los diferentes campos de aplicación a la realidad.

El Análisis de Datos Espaciales es de gran interés en muchos campos en donde los objetivos pueden ser distintos. En Ecología, por ejemplo, suele ser de interés estimar una distribución espacial que explique las localizaciones acaecidas en una área de estudio o que permita comparar las localizaciones de varias especies.

En Epidemiología el interés suele ser el de poder concluir si las causas de una cierta enfermedad están concentradas en una determinada región. Esto puede conseguirse comparando la distribución espacial de los casos observados con las localizaciones de un conjunto de controles elegidos al azar de la población en estudio.

En Economía, la localización de una nueva empresa es de vital importancia para el incremento de sus beneficios ya que si quiere reducir sus costes logísticos deberá conocer donde se encuentran sus principales clientes.

En otras muchas áreas, incluidas las de carácter militar, está presente la necesidad de asociar los datos estadísticos a un punto espacial y analizar como influye esta información geográfica añadida en el propio dato estadístico.

1.1. Los Sistemas de Información Geográfica.

Como hemos indicado, la información estadística que recopilamos en muchas disciplinas es más valiosa si está *georeferencida*, es decir, si esta información incluye las coordenadas geográficas donde se produce el dato.

Esta necesidad, ha dado lugar a la aparición de los llamados *Sistemas de Información Geográfica* (GIS), que son herramientas informáticas desarrolladas para gestionar esa información que se obtiene en un territorio y, dado que tienen una gran potencia, permiten trabajar con un volumen de datos muy elevado como los que proceden del mundo real. Como características principales de estos sistemas podemos citar:

- Trabajar con una gran cantidad de datos, es decir, permite aplicaciones en lo que hoy en día se llama *Big Data*.
- Capturar datos espaciales, editarlos, almacenarlos, gestionarlos y consultarlos de forma rápida.
- Analizar dichos datos de forma espacial, es decir, utilizando la información proporcionada por sus coordenadas.
- Obtener conclusiones y resultados, tanto desde un punto de vista descriptivo como *inferencial*, lo que permitirá modelizarlos y hacer predicciones.
- Generar resultados y exportarlos: visualizar, crear informes, gráficos, mapas, etc.

En general podemos decir que no son solo herramientas de diseño cartográfico sino que analizan la realidad, la modelizan y la gestionan.

Existen multitud de sistemas *GIS*, tanto de acceso y utilización pública como de carácter privado o de pago, así como repositorios de información de ambas características para estos sistemas. En el estudio presente se ha utilizado el sistema Quantum GIS, también llamado QGIS, que es uso público y que tiene como características principales:

- Es de uso libre. No necesita pago de licencia.
- Existen versiones para diferentes Sistemas Operativos.
- Está muy extendido en la comunidad que usa este tipo software.
- Se integra muy bien con otros sistemas GIS como GRASS, SAGA, ...
- Se integra muy bien con el paquete de análisis estadístico R, también de carácter libre y de amplio uso en la comunidad estadística.

1.2. El modelo Matemático.

El modelo matemático general que se utiliza es, dado un punto del Espacio Euclideo $\mathbf{s} \in R^d$, se supone asociado a dicho punto un dato potencial $Z(\mathbf{s})$

incierto que consideraremos que es una variable aleatoria. Dejemos ahora variar a $\mathbf{s}$ sobre un conjunto índice $D \subset R^d$ de manera que se genera un campo aleatorio multivariado o también llamado proceso aleatorio, donde se sobreentiende que D varia de manera también aleatoria en este modelo general.

$$\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D\}$$

A una realización de la v.a. $\mathbf{Z}(\mathbf{s})$ se denota $\{\mathbf{z}(\mathbf{s}) : \mathbf{s} \in D\}$.

Según se modele la v.a. $\mathbf{Z}(\mathbf{s})$ o el conjunto soporte D, podemos clasificar los modelos resultantes en varias categorías con tratamientos matemáticos muy distintos.

- **Datos Geoestadísticos**. En este modelo, D es un subconjunto fijo de R^d que contiene un rectángulo *d-dimensional* de volumen positivo, y $Z(\mathbf{s})$ es un vector aleatorio en la localización $\mathbf{s} \in D$.
- **Datos en Rejilla**. En este modelo, D es un subconjunto fijo pero numerable (finito o infinito) de puntos de R^d y $Z(\mathbf{s})$ es un vector aleatorio en la localización $\mathbf{s} \in D$.
- **Procesos Puntuales**. En este modelo, D es un un proceso de puntos en R^d o en un subconjunto R^d y $Z(\mathbf{s})$ es un vector aleatorio en la localización $\mathbf{s} \in D$. Aquí D es aleatorio, en el sentido que la variable $Z(\mathbf{s})$ solo se manifiesta en puntos inciertos de R^d.

Cuando $\mathbf{Z}(\mathbf{s})$ es un vector estamos ante un modelo **Multivariante**, y cuando $Z(\mathbf{s})$ es un escalar se trata de un modelo **Univariante**.

2. Geoestadística.

Como hemos indicado en los modelos que estudia la Geoestadística $D \subset R^d$ es fijo y la variable $\mathbf{s}$ de posición se mueve de forma continua en D. Las variables aleatorias $Z(\mathbf{s})$ suelen pertenecer a la misma familia de Funciones de Distribución, aunque en general sus parámetros varían de un punto a otro, dando lugar a que sus medias y varianzas varíen con el valor de $\mathbf{s}$. Esto da lugar a la siguiente clasificación:

- **Modelos Estacionarios**. Cuando la media y la varianza no varían de un punto $\mathbf{s}$ a otro.
- **Modelos Estacionarios en Media**. Cuando la media es constante pero la varianza varía con $\mathbf{s}$.
- **Modelos Estacionarios en Varianza**. Cuando la media varía con $\mathbf{s}$ pero la varianza es constante en el modelo.

Los dos últimos modelos se les llama en general **No Estacionarios**.

Independientemente de las características que hemos mencionado, el aspecto fundamental de todos estos procesos aleatorios está en que las v.a. $Z(\mathbf{s})$ no son independientes unas de las otras, sino que presentan una correlación entre ellas. Esta correlación en general es función de la distancia y suele disminuir cuando la distancia aumenta.

Esta correlación se manifiesta en términos prácticos, en que el valor de $Z(\mathbf{s}_1)$ en un punto $\mathbf{s}_1$ está influenciado por el valor de $Z(\mathbf{s}_2)$ en otro punto $\mathbf{s}_2$ más o menos cercano.

Para el estudio de esta correlación se introduce el concepto de **Variograma** que se define con la expresión:

$$var(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)) = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2), \quad para\ todo\ \ \mathbf{s}_1, \mathbf{s}_2 \in D$$

donde 2γ es el Variograma y γ el Semivariograma.

En general, el Variograma es un función del vector $\mathbf{s}_1 - \mathbf{s}_2$, que lo representamos por el vector $\mathbf{h}$ definido como $\mathbf{h} = \mathbf{s}_1 - \mathbf{s}_2$ y del punto $\mathbf{s}$. En la

ausencia de esta última dependencia se hablará de modelo Estacionarios de Forma Estricta.

La labor del estadístico en estos problemas es, al nivel mas abstracto, determinar los parámetros, la distribución y la correlación entre estas v.a. $Z(\mathbf{s})$, a partir de una muestra de tamaño k que se toma en diferentes puntos de D, $\{Z(\mathbf{s}_1),\ Z(\mathbf{s}_2), \ldots Z(\mathbf{s}_k)\}$.

Por la envergadura del problema propuesto, las soluciones que se obtienen son a nivel numérico y en posiciones determinadas y siempre hablando a lo sumo de valores medios o bajo un intervalo de incertidumbre, puesto que estamos trabajando con variables aleatorias.

Estos problemas también pueden ser tratados de una manera menos rigurosa sin introducir la estadística de probabilidades dando lugar a los llamados **Métodos No Estocásticos**.

2.1. Procesos Estacionarios.

El proceso aleatorio que hemos comentado, se define usualmente con una función de distribución conjunta de n puntos pertenecientes a D. Es decir

$$F_{s_1,s_2,\ldots s_n}(z_1, z_2, \ldots z_n) \equiv P\{Z(\mathbf{s}_1) \leq z_1,\ Z(\mathbf{s}_2) \leq z_2, \ldots Z(\mathbf{s}_n) \leq z_n\} \qquad n \geq 1$$

que debe cumplir la condición de simetría de Kolmogorov's (F permanece constante si a z_j y s_j son sometidas a la misma permutación) y la condición de consistencia,

$$F_{s_1,s_2,\ldots s_{k+l}}(z_1, z_2, \ldots z_k, \infty, \ldots, \infty) = F_{s_1,s_2,\ldots s_k}(z_1, z_2, \ldots z_k)$$

Se define la **Estacionariedad de segundo orden** en un proceso aleatorio espacial, cuando se cumple:

1. Las funciones de distribución en cada punto de D son idénticas, $F_s(z) = P\{Z(\mathbf{s}) \leq z\}$ no depende del punto $\mathbf{s}$. Como consecuencia de ello

- Los valores medios de las variables aleatorias $Z(\mathbf{s})$ son constantes

$$E(Z(\mathbf{s})) = \mu \quad \forall\, \mathbf{s} \in D$$

- Las varianzas también son iguales para todo D.

$$E((Z(\mathbf{s})^2) = cte \quad \forall\, \mathbf{s} \in D$$

2. La covarianza entre las v.a. $Z(\mathbf{s}_1)$ y $Z(\mathbf{s}_2)$ es función del vector diferencia entre $\mathbf{s}_1 - \mathbf{s}_2$.

$$cov(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) = C(\mathbf{s}_1 - \mathbf{s}_2) \quad \forall\, \mathbf{s}_1, \mathbf{s}_2 \in D$$

si además $C(\mathbf{s}_1 - \mathbf{s}_2)$ es función solo de $\parallel \mathbf{s}_1 - \mathbf{s}_2 \parallel$ se dice que el proceso es **isotrópico**.

Se dice que el proceso es **Estacionario de forma estricta**, si aparte de cumplir los puntos anteriores, se verifica que

$$F_{s_{1+h}, s_{2+h}, \ldots s_{n+h}}(z_1, z_2, \ldots z_n) = F_{s_1, s_2, \ldots s_n}(z_1, z_2, \ldots z_n) \quad \forall\, h \in R^d \;\; y\; n \geq 1$$

2.2. El Variograma

Dados dos puntos del espacio D del estudio, se define el **Variograma** entre estos dos puntos $\mathbf{s}_1$ y $\mathbf{s}_1$ por la expresión:

$$var(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)) = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2), \quad para\ todo \;\; \mathbf{s}_1, \mathbf{s}_2 \in D$$

a la función $2\gamma(\mathbf{s}_1 - \mathbf{s}_2)$ se le llama Variograma del proceso aleatorio y se suele escribir, cuando se da el supuesto de estacionariedad, por $2\gamma(\mathbf{h})$, donde $\mathbf{h} = \mathbf{s}_1 - \mathbf{s}_2$.

Por la definición del Variograma, en donde la diferencia de los valores de $Z(\cdot)$ están elevados al cuadrado, se verifica que $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$ y que $\gamma(0) = 0$, sin embargo se tiene que el limite de $\gamma(\mathbf{h}) = c_0$ cuando $\mathbf{h} \to \mathbf{0}$. A este valor de c_0 se le denomina **Efecto Pepita** (Nugget), y como consecuencia el Variograma presenta una discontinuidad en el origen.

Para funciones L_2 continuas esta discontinuidad no puede suceder y la explicación se asocia de errores de medida (c_{MS}) y a variaciones en microescala (c_{ME}) del fenómeno aleatorio.

$$c_0 = c_{MS} + c_{ME}$$

Otras propiedades de c_0 son

1. $2\gamma(\mathbf{h})$ es continua en el origen si solo si $E(Z(\mathbf{s}+\mathbf{h})-Z(\mathbf{s}))^2 \to 0$ cuando $\mathbf{h} \to \mathbf{0}$. Es decir es un proceso L_2-continuo.
2. Si $2\gamma(\mathbf{h})$ no es continuo $Z(\mathbf{s})$ es altamente irregular.
3. Si $2\gamma(\mathbf{h})$ es una constante, entonces $Z(\mathbf{s}_1)$ y $Z(\mathbf{s}_2)$ son incorreladas para $\forall\ \mathbf{s}_1 \neq \mathbf{s}_2$.

Un proceso de media constante en D y que verifica la condición del variograma de depender solo del vector distancia, se le denomina **Intrinsecamente Estacionario**. Si además esta dependencia es función solo del módulo del vector distancia, se le denomina **Isotrópico**.

2.3. Covariograma y Correlograma.

Se llama Covariograma a la función

$$f(\mathbf{s}_1, \mathbf{s}_2) = cov(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) \quad \forall\ \mathbf{s}_1, \mathbf{s}_2 \in D$$

se tiene que

$$var(Z(\mathbf{s}_1) - Z(\mathbf{s}_2)) = var(Z(\mathbf{s}_1)) + var(Z(\mathbf{s}_2)) - 2cov(Z(\mathbf{s}_1)Z(\mathbf{s}_2))$$

de esta expresión se deduce que un proceso Estacionario de Segundo Orden es Intrinsicamente Estacionario y se tiene pues que:

$$cov(Z(\mathbf{s}_1), Z(\mathbf{s}_2)) = C(\mathbf{s}_1 - \mathbf{s}_2) = C(\mathbf{h})$$

Se define el correlograma como la función

$$\rho(\mathbf{h}) = \frac{C(\mathbf{h})}{C(\mathbf{0})}$$

donde $C(\mathbf{0})$ se le llama **Valle** (Sill) del semivariograma y a $C(0) - c_0$ **Valle Parcial** (Parcial Sill).

2.4. Estimación del Variograma.

Por la fórmula de cálculo de la varianza, se tiene

$$E((Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2) = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2) + (\mu(\mathbf{s}_1) - \mu(\mathbf{s}_2))^2$$

que en el caso de medias constantes, se tiene que

$$E((Z(\mathbf{s}_1) - Z(\mathbf{s}_2))^2) = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2)$$

A partir de esta expresión podemos establecer una estimación del variograma tanto clásica como robusta.

2.4.1. Estimador de Matheron. Método de los Momentos.

Bajo el supuesto de constancia de la media, un estimador de $\gamma(\mathbf{h})$, es el llamado estimador de Matheron o también llamado "de los momentos" y que se expresa por la fórmula:

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 \quad \mathbf{h} \in R^d$$

donde

$$N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \ \mathbf{s}_i - \mathbf{s}_j = \mathbf{h};\ i, j = 1, 2, \dots\ n\}$$

y $|N(\mathbf{h})|$ es el número de pares en $N(\mathbf{h})$.

Es importante mencionar que en la fórmula anterior no hay que estimar la media. Una fórmula análoga para el calculo del covariograma sería:

$$\hat{C}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i) - \bar{Z})(Z(\mathbf{s}_j) - \bar{Z})$$

donde

$$\bar{Z} = \sum_{i=1}^{n} Z(\mathbf{s}_i)/n$$

2.5. Ajuste del Modelo del Variograma.

La búsqueda de un variograma que recoja correctamente la variación espacial y la interdependencia del proceso aleatorio, está íntimamente relacionada con la variación espacial de la muestra $(Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots Z(\mathbf{s}_n))'$. El espacio de todas las posibles elecciones de variogramas es muy amplio, y en la práctica se restringe a una familia paramétrica en la que se procura hallar unos parámetros que mejor se acoplen a la muestra. Un ejemplo de familia paramétrica sería:

$$\{2\gamma : \ \gamma(\mathbf{h}) = c_0 + b_l \parallel \mathbf{h} \parallel;\ c_0 \geq 0;\ b_l \geq 0\}$$

En general se pueden expresar estas familias en el formato:

$$P = \{2\gamma : \ 2\gamma(\cdot) = 2\gamma(\cdot, \theta);\ \theta \in \Theta\}$$

Para la búsqueda de parámetros óptimo de θ, existen varios métodos en el que destacan el método de máxima verosimilitud y el de mínimos cuadrados.

2.5.1. Método de la Máxima Verosimilitud.

Aplicado sobretodo a modelos Gausianos y en caso más simple se supone que la muestra $Z(\mathbf{s})$ es multivariada pero independientes es decir $Gau(X\beta, \sigma^2 I)$ donde en este caso el parámetro $\theta = \sigma^2$.

En un contexto más general $Z(\mathbf{s})$ se supone que es de la forma $Gau(X\beta, \Sigma(\theta))$ donde X es una matriz $n \times q$ y de rango $q \leq n$ y que la matriz $n \times n$ $\Sigma(\theta) = (cov(Z(\mathbf{s}_i), Z(\mathbf{s}_j)))$ depende de θ.

El logaritmo de la verosimilitud sería:

$$L(\beta, \theta) = (n/2)log(2\pi) + (1/2)log|\Sigma(\theta)| +$$
$$+(1/2)(Z - X\beta)'\Sigma(\theta)^{-1}(Z - X\beta), \quad \beta \in R^q, \quad \theta \in \Theta$$

Minimizando ésta expresión en los parámetros β y θ, tendríamos estimado la expresión del variograma.

2.5.2. Método de los Mínimos Cuadrados.

Un procedimiento válido para calcular los parámetros sería hacer un gráfico de puntos $\{(h, 2\hat{\gamma}(h\mathbf{e})) : h = h(1), h(2), \dots h(K)\}$ y encontrar una función de variograma que mejor se ajuste a esta gráfica.

Por los diferentes métodos numéricos que hemos descrito para estimar el variograma $(2\hat{\gamma}, 2\tilde{\gamma}, 2\bar{\gamma}, \dots)$ podemos obtener un variograma estimado que llamamos $2\gamma^{\#}(h\mathbf{e})$ y que podemos utilizar para ajustar los parámetros de nuestra familia objetivo. El método de mínimos cuadrados consiste pues en minimizar la función de error cuadrático total que podemos escribir como:

$$\sum_{j=1}^{K} \{2\gamma^{\#}(h(j)\mathbf{e}) - 2\gamma(h(j)\mathbf{e}, \theta)\}^2$$

para alguna dirección **e**. También se puede ampliar esta suma para diferentes direcciones **e** y minimizar de forma conjunta.

2.6. El Variograma en el Programa Estadístico R.

Podemos calcular el Variograma y ajustarlo a un modelo de forma óptima con el programa estadístico R. Para tal fin podemos utilizar diferentes paquetes de R (geoR, gstat, ...) aquí solo vamos a utilizar **geoR**. Lo primero que hay que hacer es instalar el paquete geoR si no lo tenemos y a continuación lo activamos

```
# instalar el paquete geoR
install.packages("geoR")

#activar el paquete
library("geoR", lib.loc="~/R/x86_64-pc-linux-gnu-library/3.4")
```

Este paquete viene con varias muestras de datos que podemos utilizar para nuestro desarrollo, en concreto vamos a utilizar el conjunto de datos llamado "s100".

```
# Listamos los conjunto de datos del paquete
data(package="geoR")

#Cargamos los datos s100
data(s100)

#Hacemos unos graficos descriptivos de los datos
plot(s100)
```

Cuando ejecutamos el último comando, se generan cuatro gráficos que nos aportan información sobre la disposición de los datos (coordenadas x,y) en el primer cuadrante, los valores de los datos según la coordenada x y según la coordenada y y por último un histograma de datos con su densidad de frecuencias. Todo ello lo reflejamos en la figura 1.

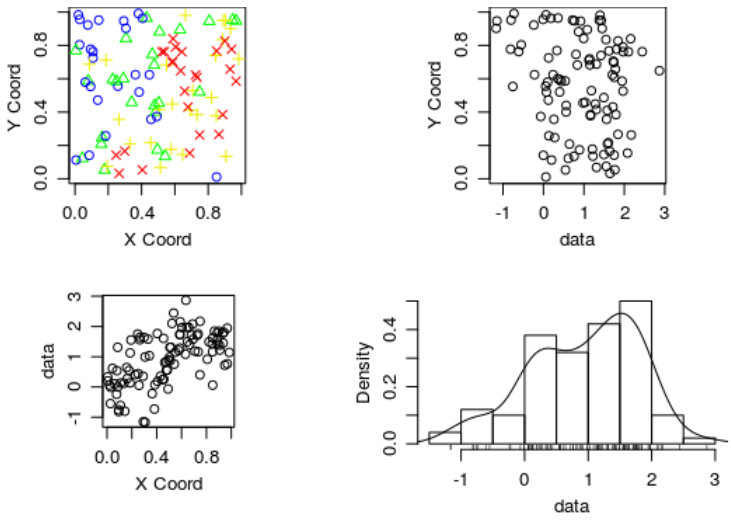

Figura 1: Información Estadística Datos "s100"

2.6.1. Cálculo del Variograma.

El Variograma se calcula con el comando *variog*, y a continuación se puede hacer una gráfica en donde vemos la distancia máxima a la que se ha calculado. El manual dice que es conveniente utilizar como distancia máxima la mitad de la anterior, en nuestro caso $0,6$.

```
#Dividimos la pantalla en dos graficos
oldpar=par(mfrow=c(1,2))

#Calculamos y visualizamos en primer variograma sin acotar distancia
plot(variog(s100))

#Calculamos el variograma acontado distancia
vario=variog(s100,max.dist = 0.6)
plot(vario)

#Dejamos la pantalla como estaba
par(oldpar)
```

Obtenemos como salida la que aparece en la figura 2.

El programa permite calcular el variograma en varios formatos, aparte del conocido (empírico), como son el de nube de puntos, o el variograma

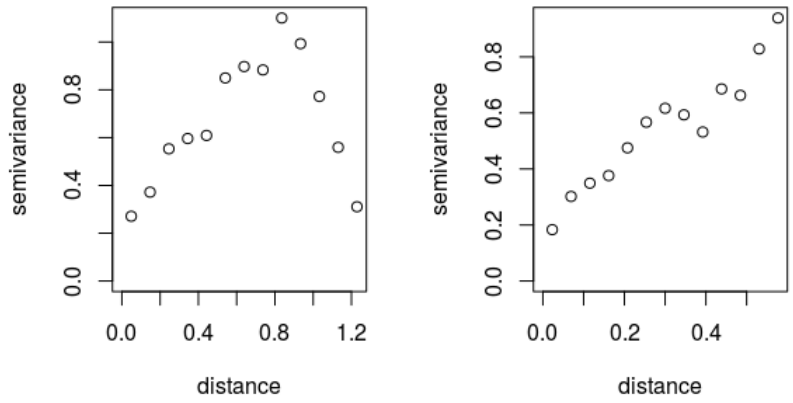

Figura 2: Variograma datos "s100"

suavizado y en forma de cajas. El código que tenemos que introducir es

```
#discretizado
vario=variog(s100,max.dist = 0.6)

#Nube de puntos
vario.c=variog(s100,max.dist = 0.6, op="cloud")

#Suavizado
vario.s=variog(s100,max.dist = 0.6, op="smooth")

#De Cajas
vario.ca=variog(s100,max.dist = 0.6,bin.cloud=TRUE)

#Visualizamos los Variogramas
oldpar=par(mfrow=c(2,2))
plot(vario,main="Variograma Discretizado")
plot(vario.c,main="Variograma de Nube de Puntos")
plot(vario.s,main="Variograma Suavizado");
plot(vario.ca, bin.cloud=TRUE,main="Variograma en Cajas")
par(oldpar)
```

Obtenemos los gráficos de de la figura 3

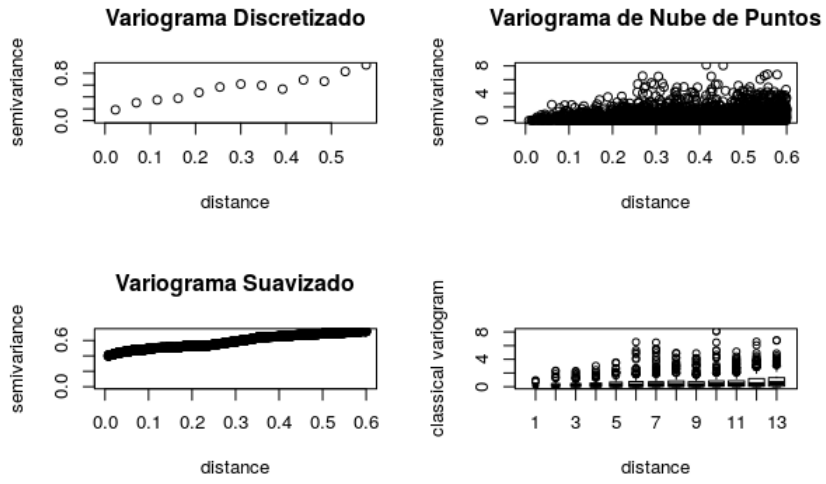

Figura 3: Variograma datos "s100"

2.6.2. Ajuste del Variograma.

El siguiente paso en el estudio del Variograma es ajustarlo una curva, en el que podemos utilizar como criterio de optimización mínimos cuadrados(*variofit*), máxima verosimilitud (*likfit*).

Se puede hacer aparte del criterio anterior de ajuste una variante más que consiste en utilizar pesos relativos (Cressie) o los pesos usuales. La operativa en R se describe a continuación. Esta información será de interés para seleccionar el ajuste más adecuado.

```
#Minimos cuadrados ordinarios para funcion exponencial
vario.ols = variofit(vario, ini=c(1,0.5), weights="equal",cov.model="exponential")

#Minimos cuadrados relativos para funcion exponencial
vario.ols_cre = variofit(vario, ini=c(1,0.5),
    weights="cressie",cov.model="exponential")

#Minimos cuadrados ordinarios para funcion spherica
vario.ols_sph = variofit(vario, ini=c(1,0.5), weights="equal",cov.model="spherical")

#Minimos cuadrados ordinarios para funcion gaussian
vario.ols_gau = variofit(vario, ini=c(1,0.5), weights="equal",cov.model="gaussian")

#Vsualizacion de los parametros del ajuste
```

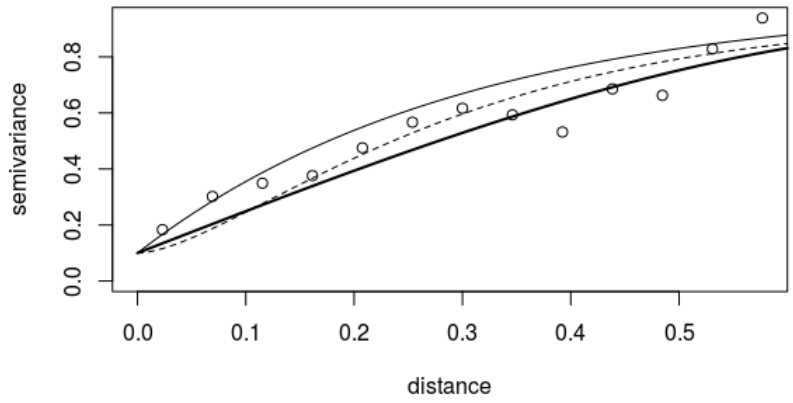

Figura 4: Variograma Ajustes de curvas

```
summary(vario.ols)
summary(vario.ols_cre)
# ...
```

Se puede dibujar un gráfico del variograma y a continuación superponer los diferentes ajustes de funciones mediante el comando *lines.variomodel*.

```
#Visualizamos el Variograma Previo
plot(vario)

#Visualizamos el ajuste exponencial
lines.variomodel(cov.model="exp",cov.pars=c(0.9,0.3), nug=0.1, max.dist=0.6)

#Visualizamos el ajuste matern
lines.variomodel(cov.model="mat",cov.pars=c(0.85,0.2), nug=0.1, max.dist=0.6,
    lty=2, kappa=1)

#Visualizamos el ajuste esferico
lines.variomodel(cov.model="sph",cov.pars=c(0.8,0.8), nug=0.1, max.dist=0.6, lwd=2)
```

El resultado gráfico que se obtiene es el que aparece en la figura 4

A continuación incluimos un ajuste por máxima verosimilitud.

```
#Ajuste por maxima verosimilitud (exponencial)
vario.ml = likfit(s100, ini=c(1,0.5), cov.model="exp")

#Visualizamos el resultado
summary(vario.ml)
```

3. Kriging. Métodos Clásicos.

3.1. Predicción Espacial.

Sea un proceso aleatorio, como los descritos en los apartados anteriores, $\{Z(\mathbf{s}) : \ s \in D \ \subset \ R^d\}$, en el que se han observado n datos $Z(\mathbf{s}_1), \ldots Z(s_n)$. Estos datos serán usados para hacer alguna inferencia sobre el proceso aleatorio o para predecir algún funcional sobre $Z(\mathbf{s})$, es decir $g(\{Z(\mathbf{s}) : \ s \in D\})$. En el más sencillo de los funcionales podría ser $g(Z(\cdot)) = Z(\mathbf{s}_0)$ para algún punto $\mathbf{s}_0 \in D$.

Algunas veces el interés no es $Z(\cdot)$ sino una versión libre de ruido aleatorio de Z, $S(\cdot)$. Es decir:

$$Z(\mathbf{s}) = S(\mathbf{s}) + \epsilon(s) \quad \mathbf{s} \in D$$

donde $\epsilon(\mathbf{s})$ es ruido blanco aleatorio y nuestro interés se centra en predecir el valor del funcional $g(S(\cdot))$.

Por **Predicción Espacial** se entiende estimar el valor de $g(Z(\cdot))$ o el valor de $g(S(\cdot))$ a partir de los valores $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \ldots Z(\mathbf{s}_n)$ que se han medido en las posiciones $\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_n$.

El método **Kriging** es una técnica de predicción espacial basado la minimización de los errores cuadráticos medios y en las propiedades de segundo-orden de los procesos aleatorios $Z(\cdot)$.

El nombre de Kriging se debe al ingeniero de minas D.G. Krige, que desarrolló métodos empíricos para evaluar la distribución del grado del mineral a partir de muestras obtenidas en puntos concretos.

Entre los funcionales a utilizar, aparte del más sencillo o canónico indicado $g(Z(\cdot)) = Z(\cdot)$, se destaca el llamado **Media de Bloques**, que es utilizado ampliamente en aplicaciones prácticas del método, y que se define por las expresión:

$$g(Z(\cdot)) = \int_B Z(s)ds/|B| \quad B \subset D$$

donde $Z(\mathbf{s})$ es un proceso aleatorio y como consiguiente el valor de la integral es una variable aleatoria y $|B|$ es el volumen d-dimensional sobre R^d, resultando pues $g(Z(\cdot))$, una media sobre la región B del proceso aleatorio.

El método o métodos Kriging pueden clasificarse según el criterio estadístico utilizado en su concepción, en:

- Métodos basados en el **Tratamiento Estocástico** de los datos. En este caso, estudio de la predicción espacial estará basada en la consideración de que los datos se comportan según un proceso aleatorio espacial indexado en R^d y por consiguiente tendremos que conocer el origen de la variación aleatoria que hay en el proceso.
- Métodos **Deterministas o No Estocásticos**. Aquí estos problemas se abordan de una forma menos rigurosa, mediante métodos no probabilísticos, que se denominan deterministas y de los que haremos mención al final del estudio.

3.2. Ajuste del Funcional.

Para calcular el funcional o valor de $Z(\cdot)$ en el punto s_0, tendremos que hacerlo siguiendo un criterio de ajuste que debe ser óptimo en algún sentido, y por lo tanto tendremos que introducir consideraciones de Teoria de la Decisión.

Para simplificar supongamos que $Z(\mathbf{s}_0)$ va a ser predicha a partir de unos valores de $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \ldots Z(\mathbf{s}_n)$. El argumento cambia poco si nos referimos a medias de bloques o a estimaciones libres de ruido aleatorio. Usando un formalismo de la teoría de la decisión, $L(Z(\mathbf{s}_0), p(Z, \mathbf{s}_0))$ denota la pérdida incurrida cuando se predice $Z(\mathbf{s}_0)$ con $p(Z, \mathbf{s}_0)$. Un estimador óptimo es aquel que minimiza $E\{L(Z(\mathbf{s}_0), p(Z, \mathbf{s}_0)\}$, donde $E(\cdot)$ denota la esperanza matemática respecto a la distribución conjunta de $Z(\mathbf{s}_0)$ y Z.

Entre las funciones de pérdida a considerar, destaca por sus múltiples

bondades estadísticas, la pérdida **Error Cuadrática**, que se define por la expresión:

$$L(Z(\mathbf{s}_0), p(Z, \mathbf{s}_0)) = (Z(\mathbf{s}_0) - p(Z, \mathbf{s}_0))^2$$

que da lugar a un estimador que minimiza $E\{(Z(\mathbf{s}_0) - p(Z, \mathbf{s}_0))^2 | Z\}$, es decir, la esperanza matemática de $Z(\mathbf{s}_0)$.

$$p^0(Z, \mathbf{s}_0) = E(Z(\mathbf{s}_0)|Z)$$

Esto implica que el error cuadrático medio de la predicción que depende solo de los momentos de primer y segundo orden. Otro aspecto importante es que la región de predicción $100(1-\alpha)\,\%$ es simétrica y la derivación de algunas propiedades son más sencillas.

Otra medida de la pérdida que se utiliza con frecuencia es

$$L(Z(\mathbf{s}_0), p(Z, \mathbf{s}_0)) = |Z(\mathbf{s}_0) - p(Z, \mathbf{s}_0)|^\nu \quad 1 \leq \nu < 2$$

3.3. Estimadores Lineales.

Como se ha visto, en el caso de estimadores para pérdidas error-cuadráticas, el mejor estimador es $E(Z(\mathbf{s}_0)|Z)$ que no siempre es lineal en Z. En vez de buscar el mejor estimador, muchas veces se busca el **mejor estimador lineal**, esto es encontrar $l_1, l_2, \ldots l_n, k$ en

$$p(\mathbf{Z}, \mathbf{s}_0) = \sum_{i=1}^{n} l_i Z(\mathbf{s}_i) + k$$

de manera que $E(Z(\mathbf{s}_0) - p(Z, \mathbf{s}_0))^2$ es minimizado, o equivalentemente minimizar sobre $l_1, l_2, \ldots l_n, k$:

$$E(Z(\mathbf{s}_0) - \sum_{i=1}^{n} l_i Z(\mathbf{s}_i) - k)^2 = var(Z(\mathbf{s}_0) - \sum_{i=1}^{n} l_i Z(\mathbf{s}_i)) +$$

$$+(\mu(\mathbf{s}_0) - \sum_{i=1}^{n} l_i \mu(\mathbf{s}_i) - k)^2$$

donde $\mu(\mathbf{s}_0 = E(Z(\mathbf{s}_0) \quad s \in D$.

La solución de este problema es tomar $k = \mu(\mathbf{s}_0) - \sum_{i=1}^{n} l_i \mu(\mathbf{s}_i)$ y

$$l' = (l_1, \dots \ l_n) = c'\Sigma^{-1}$$

donde $\mathbf{c} = (C(\mathbf{s}_0, \mathbf{s}_1), \dots C(\mathbf{s}_0, \mathbf{s}_n))'$ y Σ es una matriz $n \times n$ cuyo término (i, j) es $C(\mathbf{s}_i, \mathbf{s}_j)$. El estimador óptimo sería pues

$$p^*(Z, \mathbf{s}_0) = c'\Sigma^{-1}(\mathbf{Z} - \boldsymbol{\mu}) + \mu(\mathbf{s}_0)$$

donde $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \dots \mu(\mathbf{s}_n))'$.

El valor del error mínimo cuadrático es

$$\sigma_{sk}^2 = C(\mathbf{s}_0, \mathbf{s}_0) - c'\Sigma^{-1}c$$

3.4. Escala de Variación.

En el estudio de estos modelos se harán, muchas veces, asunciones que serán inverificables. Con objeto de que éstas sean razonables, se deberá prestar atención a la **escala de fluctuación** que el proceso parece mostrar.

La escala tiene dos diferentes significados. Uno es la **escala observacional** de $Z(\mathbf{s})$, esto es, que los instrumentos de medida tienen precisión hasta cierto nivel. El otro es la **escala espacial**, esto es, las observaciones están basadas en cierta agregación y están tomadas a cierta distancia.

No existe un standard que diga que constituye gran escala, pequeña escala o micro-escala, depende del objeto del estudio, la precisión de los datos $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots Z(\mathbf{s}_n))'$, su nivel de agregación, y de las localizaciones espaciales $\{\mathbf{s}_1, \ldots \mathbf{s}_n\}$.

En este sentido, supongamos que los datos $(Z(\mathbf{s}_1), \ldots Z(\mathbf{s}_n))$ representan Z valores en puntos de D y que los modelamos como una realización parcial del proceso

$$\{Z(\mathbf{s}) : \ \mathbf{s} \ \in D \subset R^d\}$$

y que satisface la descomposición

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + W(\mathbf{s}) + \nu(\mathbf{s}) + \epsilon(\mathbf{s})$$

donde

- $\mu(\cdot) \equiv E(Z(\cdot))$ es la media estructural determinista que llamaremos **Variación a Gran Escala**.
- $W(\cdot)$ es de media cero y L_2-*continua* ($E(W(\mathbf{s}+h)-W(\mathbf{s}))^2 \to 0$ cuando $\| \ h \ \| \to 0$), intrínsecamente estacionaria y cuyo variograma es mayor que $min\{\| \ \mathbf{s}_i - \mathbf{s}_j \ \| : \ 1 \leq i < j \leq n\}$. Llamaremos a $W(\cdot)$ **Variación Suave a Pequeña Escala**.
- $\nu(\cdot)$ es de media cero, intrínsecamente estacionaria, independiente de $W(\cdot)$, cuyo variograma es menor que $min\{\| \ \mathbf{s}_i - \mathbf{s}_j \ \| : \ 1 \leq i < j \leq n\}$. Llamaremos a $\nu(\cdot)$ **Variación a Micro-Escala**.
- $\epsilon(\cdot)$ es un ruido blanco de media cero, independiente de $W(\cdot)$ y de $\dot{\nu()}$. Llamaremos a $\epsilon(\cdot)$ **Error de Medida o Ruido** y llamaremos a su varianza $var(\epsilon(s)) = c_{ME}$.

Podemos escribir

$$Z(\mathbf{s}) = S(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \mathbf{s} \in D$$

donde la señal o proceso suave $S(\cdot)$ está dado por $S(\cdot) = \mu(\mathbf{s}) + W(\mathbf{s}) + \nu(\mathbf{s})$. El proceso S se le llama a menudo la versión libre de ruido de Z y en la literatura ingenieril como **Proceso de Estado**.

También podemos escribir

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \qquad \mathbf{s} \in D$$

donde el proceso $\delta(\cdot)$ esta dado por $W(\mathbf{s}) + \nu(\mathbf{s}) + \epsilon(\mathbf{s})$. Cuando la correlación de $\delta(\mathbf{s})$ con $\delta(\mathbf{s}+h)$ puede ser escrita como una función de $h/a(h)$, donde $0 < a(h) < \infty$, entonces $a(h)$ es llamada la **Escala de Correlación Espacial** del proceso en la dirección de h.

La descomposición que hemos indicado en el párrafo anterior no es única y se determina por criterios operacionales. Esto significa que que se puede llegar a diferentes conclusiones según la variación que asignemos a cada una de sus componentes.

3.5. Kriging Ordinario.

Kriging Ordinario hace referencia a una espacial predicción que cumple dos condiciones:

1. Condición del Modelo.

$$Z(\mathbf{s}) = \mu + \delta(\mathbf{s}) \qquad \mathbf{s} \in D,\ \mu \in R,\ \mu \textit{ desconocida}$$

2. Condición del Estimador.

$$p(\mathbf{Z}, B) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i), \qquad \sum_{s=1}^{n} \lambda_i = 1$$

Esta última condición garantiza la existencia de uniforme no sesgo en la media

$$E(p(Z, B)) = \mu = E(Z(B)) \ \ para \ \forall\ \mu \in R$$

Por predicción óptima se entiende en relación a una pérdida error-cuadrática y por tanto hablamos de minimizar la predicción media-error cuadrática, por lo tanto si,

$$g(Z(\cdot)) = Z(B) \equiv \begin{cases} \int_B Z(u)du/|B|, & |B| > 0 \\ ave\{Z(u) : \ u \in B\}, & |B| = 0 \end{cases}$$

El $p(\cdot, B)$ óptimo minimizará el error de predicción medio cuadrático

$$\sigma_e^2 \equiv E(Z(B) - p(Z, B))^2 \tag{1}$$

Sobre la clase de estimadores $\sum_{i=1}^n \lambda_i Z(\mathbf{s}_i)$ que verifican $\sum_{i=1}^n \lambda_i = 1$.

Predicción Óptima. En el Kriging Ordinario la minimización de (1) es realizada sobre $\lambda_1, \dots \lambda_n$ bajo la condición $\sum_{i=1}^n \lambda_i = 1$, donde el modelo asume ajustarse a un variograma

$$2\gamma(\mathbf{h}) = var(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})), \quad \mathbf{h} \in R^d$$

Suponiendo por un momento que $B = \{\mathbf{s}_0\}$, tendremos que minimizar

$$E(Z(\mathbf{s}_0) - \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i))^2 - 2m(\sum_{i=1}^n \lambda_i - 1) \tag{2}$$

con respecto $\lambda_1, \dots \lambda_n$ y m. Siendo m el multiplicador de Lagrange que asegura que $\sum_{i=1}^n \lambda_i = 1$.

Ahora la condición $\sum_{i=1}^n \lambda_i = 1$ implica que

$$(Z(\mathbf{s}_0) - \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i))^2 = -\sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2/2 +$$

$$+2\sum_{i=1}^{n}\lambda_i(Z(\mathbf{s}_0)-Z(\mathbf{s}_i))^2/2$$

que sustituyendo la definición del variograma resulta que (2) se transforma en

$$-\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\mathbf{s}_i-\mathbf{s}_j)+2\sum_{i=1}^{n}\lambda_i\gamma(\mathbf{s}_0-\mathbf{s}_i)-2m(\sum_{i=1}^{n}\lambda_i-1) \qquad (3)$$

diferenciando con respecto $\lambda_1 \ldots \lambda_n, m$, para optimizar, obtenemos un sistema de ecuaciones con $n+1$ incógnitas

$$-\sum_{j=1}^{n}\lambda_i\gamma(\mathbf{s}_i-\mathbf{s}_j)+\gamma(\mathbf{s}_0-\mathbf{s}_i)-m=0 \qquad i=1,2,\ldots n$$

$$\sum_{i=1}^{n}\lambda_i=1$$

Podemos escribir estas ecuaciones en formato matricial

$$\boldsymbol{\lambda}_0=\Gamma_0^{-1}\boldsymbol{\gamma}_0$$

donde

$$\boldsymbol{\lambda}_0=(\lambda_1,\ldots\lambda_n,m)'$$

$$\boldsymbol{\gamma}_0=(\gamma(\mathbf{s}_0-\mathbf{s}_1),\ldots\gamma(\mathbf{s}_0-\mathbf{s}_n),1)'$$

$$\Gamma_0=\begin{cases}\gamma(\mathbf{s}_i-\mathbf{s}_j), & i=1,2\ldots n, \quad j=1,2,\ldots n\\ 1, & i=n+1, \quad j=1,2,\ldots n\\ 0, & i=n+1, \quad j=n+1\end{cases}$$

y Γ_0 es una matriz $n+1\times n+1$ simétrica.

La solución de este sistema de ecuaciones se expresa como

$$\boldsymbol{\lambda}' = (\boldsymbol{\gamma} + \mathbf{1}\frac{(1 - \mathbf{1}'\Gamma^{-1}\boldsymbol{\gamma})}{\mathbf{1}'\Gamma^{-1}\mathbf{1}})'\Gamma^{-1}$$

$$m = -(1 - \mathbf{1}'\Gamma^{-1}\boldsymbol{\gamma})/(\mathbf{1}'\Gamma^{-1}\mathbf{1})$$

donde $\boldsymbol{\lambda} = (\lambda_1, \ldots \lambda_n)'$,$\boldsymbol{\gamma} = (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \ldots \gamma(\mathbf{s}_0 - \mathbf{s}_n))'$ y Γ es una matriz $n \times n$ cuyo elemento (i, j) es $\gamma(\mathbf{s}_i - \mathbf{s}_j)$

El error de predicción mínimo cuadrado es

$$\sigma_k^2(s_0) = \boldsymbol{\lambda_0}'\boldsymbol{\gamma}_0 = \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) + m$$

$$= \boldsymbol{\gamma}'\Gamma^{-1}\boldsymbol{\gamma} - (\mathbf{1}'\Gamma^{-1}\boldsymbol{\gamma} - 1)^2/(\mathbf{1}'\Gamma^{-1}\mathbf{1})$$

o bien

$$\sigma_k^2(s_0) = 2\sum_{i=1}^{n} \lambda_i \gamma(\mathbf{s}_0 - \mathbf{s}_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{s}_i - \mathbf{s}_j)$$

Un intervalo de predicción se puede construir con el 95 % de certeza con la expresión

$$A \equiv (\hat{Z}(\mathbf{s}_0) - 1,96\sigma_k(\mathbf{s}_0), \hat{Z}(\mathbf{s}_0) + 1,96\sigma_k(\mathbf{s}_0))$$

3.6. Kriging en Términos de Covarianzas.

Siguiendo el razonamiento del punto anterior, el error al cuadrado de predicción puede ser escrito directamente como

$$(Z(\mathbf{s}_0) - \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i))^2 = (Z(\mathbf{s}_0) - \mu)^2 +$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j (Z(\mathbf{s}_i) - \mu)(Z(\mathbf{s}_j) - \mu) - 2 \sum_{i=1}^{n} \lambda_i (Z(\mathbf{s}_0) - \mu)(Z(\mathbf{s}_i) - \mu)$$

con la condición $\sum_{i=1}^{n} \lambda_i = 1$.

Supongamos ahora que la condición de modelo se cumple con $\delta(\cdot)$ siendo de media cero y estacionaria de segundo-orden, y que además tiene un covariograma $C(\mathbf{h}),\ \mathbf{h} \in R^d$. La ecuación (2) se transforma en

$$C(\mathbf{0}) + \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(\mathbf{s}_i - \mathbf{s}_j) - 2 \sum_{i=1}^{n} \lambda_i C(\mathbf{s}_0 - \mathbf{s}_i) - 2m(\sum_{i=1}^{n} \lambda_i - 1)$$

que minimizando respecto $\lambda_1, \ldots \lambda_n$ y m resulta la ecuación del kriging

$$\hat{p}(\mathbf{Z}, \mathbf{s}_0) = \boldsymbol{\lambda}' \mathbf{Z}$$

$$\sigma_k^2 = C(\mathbf{0}) - \boldsymbol{\lambda}' \mathbf{c} + m$$

donde

$$\boldsymbol{\lambda}' = (\mathbf{c} + \mathbf{1} \frac{(1 - \mathbf{1}' \Sigma^{-1} \mathbf{c}}{\mathbf{1}' \Sigma^{-1} \mathbf{1}})' \Sigma^{-1}$$

$$m = \frac{(1 - \mathbf{1}' \Sigma^{-1} \mathbf{c})}{\mathbf{1}' \Sigma^{-1} \mathbf{1}}$$

siendo $\mathbf{c} = (C(\mathbf{s}_0 - \mathbf{s}_1), \ldots C(\mathbf{s}_0 - \mathbf{s}_n))'$ y Σ una matriz $n \times n$, cuyo término (i, j) es $C(\mathbf{s}_i - \mathbf{s}_j)$.

3.7. Métodos de Estimación Espacial No Estocásticos.

Estos métodos se caracterizan por no estar apoyados por un modelo probabilístico, siendo procedimientos basados en la práctica para realizar interpolaciones a partir de un conjunto de muestras. Describimos dos métodos (Media Móvil Simple y Media Ponderada de la inversa de la distancia al cuadrado). En general estos modelos no proporcionan una medida del error.

Método Media Móvil Simple. En este modelo se asume que el estimador es "suave" y es elegido en ese contexto. El estimador es:

$$p(\mathbf{z}, \mathbf{s}_0) = \frac{\sum_{i=1}^{n} z(\mathbf{s}_i) I(d_{0,i} \leq r)}{\sum_{i=1}^{n} I(d_{0,i} \leq r)}$$

donde $I(A)$ es la función indicadora del conjunto A. $d_{0,i} \equiv \| \mathbf{s}_0 - \mathbf{s}_i \|$ y r es un parámetro de proximidad que es establece por criterio.

Este modelo se puede especificar de una forma similar y alternativa, ordenando las distancias euclideas de menor a mayor $d_{0,1} \leq d_{0,2} \leq \cdots \leq d_{0,n}$ y definiendo el k-ésimo estimador mas cercano como

$$p(\mathbf{z}, \mathbf{s}_0) = \sum_{i=1}^{n} z(\mathbf{s}_i) I(d_{0,i} \leq d_{0,k})/k$$

Las propiedades de este método que se pueden citar son:

1. Las estimaciones no dependen de parámetros del modelo, aunque si hay que dar valores a r o a k en las fórmulas anteriores.

2. El ordenado de las distancias $\{d_{0,1}, \ \ldots \ d_{0,n}\}$ puede ralentizar el cálculo.

3. Las medias-móviles no son robustas.

4. En general $p(\mathbf{z}, \mathbf{s}_i) \neq z(\mathbf{s}_i)$, esto es, p no es un interpolador exacto. Por el contrario actúa como un "suavizador".

Método de la Media Ponderada de la inversa de la distancia al cuadrado. En este modelo, en vez de permitir a los datos contribuir igualmente a la media, son ponderados de acuerdo a la distancia que existe entre $\mathbf{s}_i$ y $\mathbf{s}_0$. La fórmula del estimador es:

$$p(\mathbf{z}, \mathbf{s}_0) = \sum_{i=1}^{n} d_{0,i}^{-2} z(\mathbf{s}_i) / \sum_{i=1}^{n} d_{0,i}^{-2}$$

Existe una versión robusta del estimador, consistente en usar la mediana-ponderada en los cálculos.

$$p(\mathbf{z}, \mathbf{s}_0) = wt\ med\{\mathbf{z}; d_{0,1}^{-2}, \ldots\ , d_{0,n}^{-2}\}$$

Las propiedades de este método que se pueden citar son:

- El estimador es simple, no requiriendo conocer parámetros del modelo espacial.
- Computacionalmente es muy rápido de calcular, aunque en la versión robusta el número de cálculos es mayor.
- El estimador no es resistente a los atípicos, aunque con la mediana-ponderada si lo es.
- p es interpolador exacto.

3.8. Aplicación del Kriging al Estudio de las Precipitaciones en la Región de Murcia.

Vamos a aplicar los modelos teóricos del Kriging a calcular un mapa de intensidades de precipitaciones de lluvia en la Región de Murcia. Como punto de partida he bajado de la página web **Portal Estadístico de la Región de Murcia**, los datos registrados de volúmenes de agua lluvia en las diferentes estaciones de recogida de información meteorológica http://econet.carm.es/web/crem/inicio/-/crem/sicrem/PU7/sec32.html.

Los datos que he utilizado corresponden a las cifras acumuladas del año 2017. Una muestra aparece en la figura 5.

ESTACIONES_METEOROLOGICAS	VAL2017	ALTITUD	LONGITUD	LATITUD	X	Y
Abanilla	**210,8**	174	102022	381140	**-1,033889**	**38,194444**
Águilas Diputación	**125,7**	26	135132	372502	**-1,586944**	**37,417222**
Águilas Montagro	**162**	65	138302	372426	**-1,641667**	**37,407222**
Alhama Comarza	**180,6**	157	120052	375142	**-1,334722**	**37,861667**
Archena Balneario Automática	**186**	150	118212	380741	**-1,305833**	**38,128056**
Archena H.E.	**204,5**	100	117392	380653	**-1,294167**	**38,114722**
Bullas Depuradora	**219,4**	600	141462	380304	**-1,696111**	**38,051111**
Calasparra	**206,4**	340	142012	381348	**-1,700278**	**38,230000**
Calasparra Agentes Medioambientales	**196,4**	350	142082	381403	**-1,702222**	**38,234167**
Caravaca Archivel	**243,9**	881	200032	380355	**-2,000833**	**38,065278**
Caravaca Fuentes del Marqués	**198,9**	643	152382	380609	**-1,877222**	**38,102500**
Caravaca Los Royos-Aut.	**144,2**	985	203302	375534	**-2,058333**	**37,926111**
Caravaca Polideportivo	**192,9**	623	152042	380609	**-1,867778**	**38,102500**
Cartagena Clause Spain	**152,8**	50	100012	374126	**-1,000278**	**37,690556**

Figura 5: Datos de Precipitaciones (muestra)

En este proceso de búsqueda de los datos me he encontrado con dos problemas:

- Los datos de algunas estaciones no están completos, faltan los valores de algunos meses.
- En la tabla que aparece en internet no figuran las coordenadas de las estaciones meteorológicas.

Para solucionar el segundo punto, que era el más problemático, mandé un correo electrónico a la dirección que aparecía en la web para ver si podían ayudarme. Al día siguiente recibí una hoja Excel con estas coordenadas y procedí a emparejar los datos de precipitaciones con los de coordenadas, con la sorpresa que había entorno al 10 % de coordenadas que no aparecían y también había coordenadas de estaciones que no conseguía encontrar valor de precipitación.

Opté por considerar datos NA a estos casos y me quedé con el resto para hacer el estudio. Quedando un mapa que cubría la región como puede verse en la figura 6.

El siguiente paso fue exportar esta tabla Excel a formato .csv e importarlo

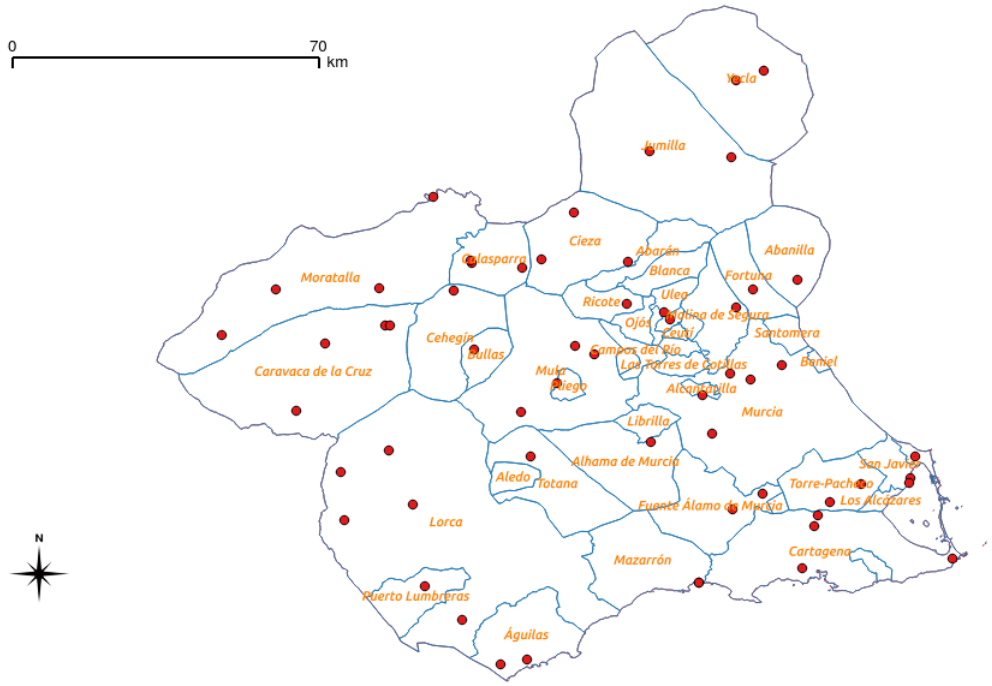

Figura 6: Estaciones Meteorológicas

como una capa de puntos vectorial de formato .cvs, para obtener la capa de puntos que se muestra en la figura 6.

Las diferentes soluciones por Kriging que hemos ido obteniendo, se describen en puntos separados y relacionados mas adelante en la memoria.

3.9. Solución Kriging Ordinario. Mediante Utilidades QGIS.

Se ha aplicado las utilidades que proporciona QGIS al problema de las Precipitaciones en la Región de Murcia, con objeto de obtener un diagrama de intensidades de estas según la zona de la región.

El primer paso ha sido hacer una estimación del Variograma con diferentes formatos de curvas de ajuste y ver cual era la más conveniente en términos de errores cuadráticos. Los resultados que se han obtenido se reflejan en la tabla 1.

Modelos	Valle Parcial	Rango	Pepita	E.Cuadrático
Exponencial (Exp)	3.043,401	41,975	0,000	41.936,32
Esférico (Sph)	2.097,332	51,373	1,365	41.187,38
Gausiano (Gau)	1.756,770	19,490	138,123	52.189,95

Cuadro 1: Comparativa de Ajuste de Variogramas

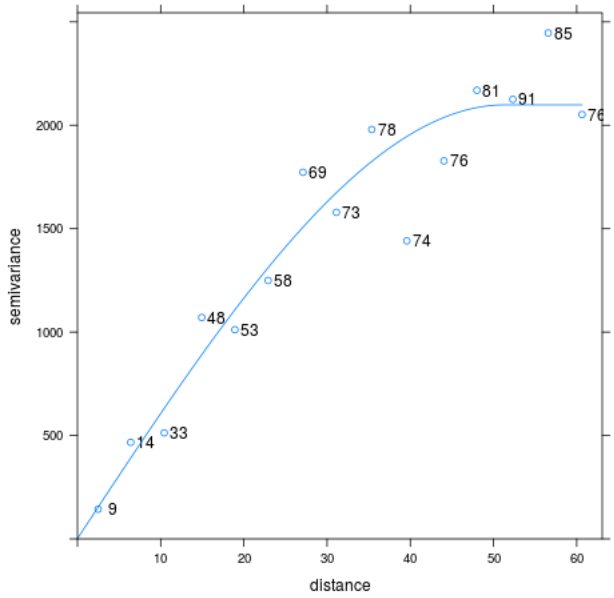

Figura 7: Variograma Precipitaciones Región de Murcia

Como se puede ver el ajuste con menores errores cuadráticos es el esférico y es el que se ha utilizado para realizar el estudio del kriging.

Aplicando el modelo de Kriging Ordinario y ajustando un variograma de forma esférica, obtenemos un ajuste del variograma que podemos visualizar en la figura 7.

La solución que hemos obtenido en formato capa raster la reflejamos en la figura 8 donde se le ha superpuesto un mapa de isovalores para hacerlo mas visual. Adicionalente se incluye un mapa de las región con la capa vectorial de isovalores donde se puede apreciar que las zonas más húmedas son el noroeste que linda con la Sierra del Segura donde nace el río Guadalquivir y

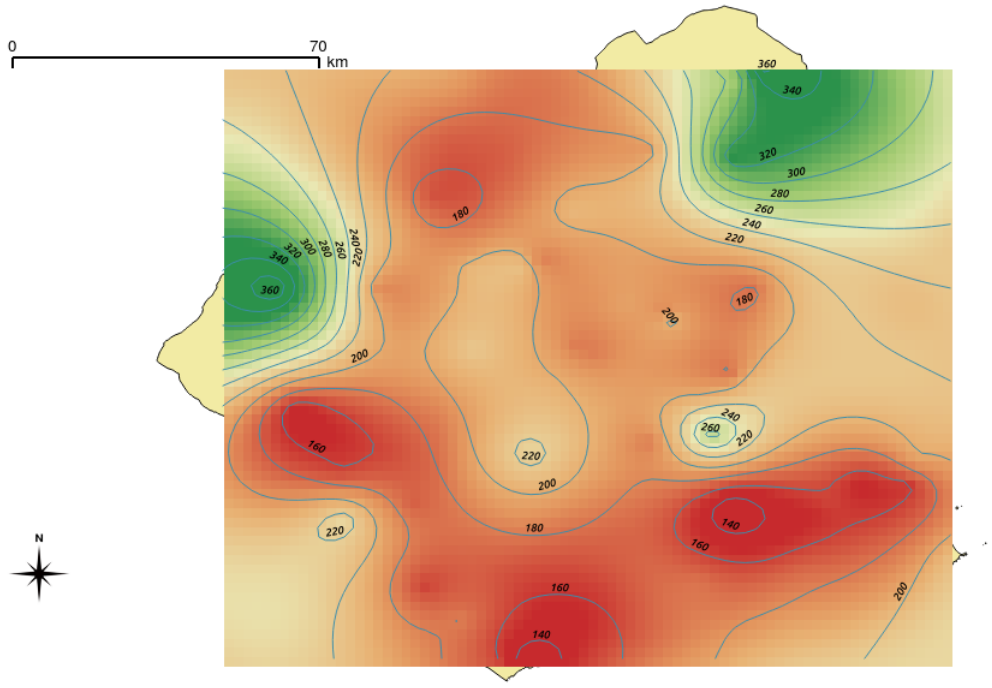

Figura 8: Kriging Precipitaciones Región de Murcia

el río Segura, la zona que linda con el norte de Alicante y también se muestra un pico en la Sierra de Carrascoy. Las menos húmedas son las costeras. Todo ello se muestra en la figura 9

3.10. Solución Problema de las Precipitaciones de Murcia. Método No Estocástico.

Se ha desarrollado un pequeño programa en Python para resolver el problema de las precipitaciones de la Región de Murcia por un método no estocástico. La idea ha sido más didáctica que de otro tipo pues existen utilidades en QGIS que ya lo hacen.

El método aplicado ha sido el de la Media Ponderada de la inversa de la distancia al cuadrado, es decir

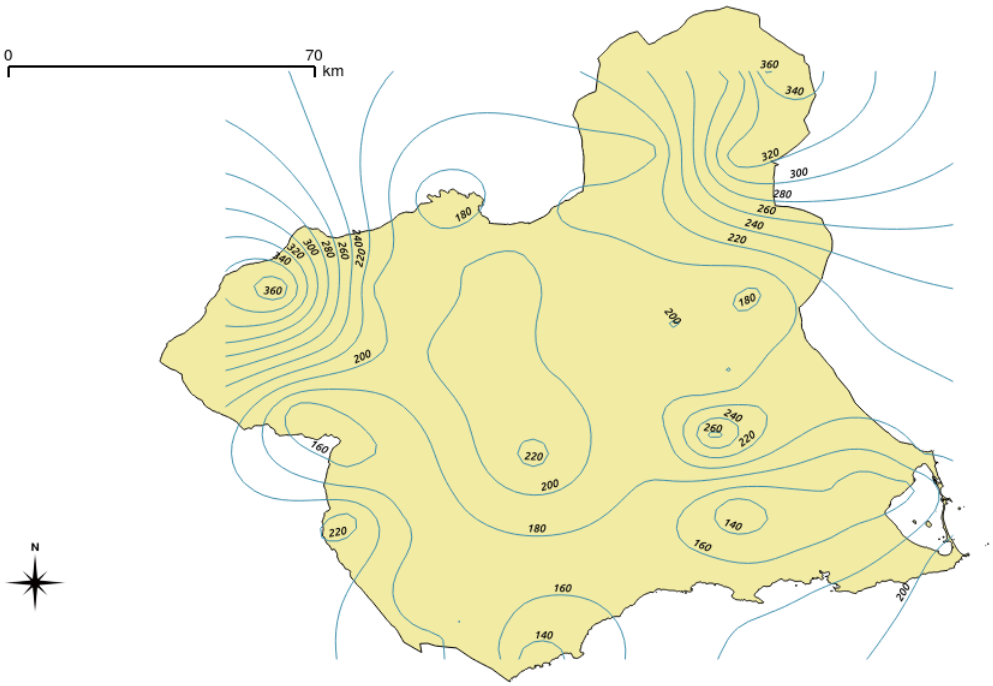

Figura 9: Isovalores Precipitaciones Región de Murcia

$$p(\mathbf{z}, \mathbf{s}_0) = \sum_{i=1}^{n} d_{0,i}^{-2} z(\mathbf{s}_i) / \sum_{i=1}^{n} d_{0,i}^{-2}$$

La construcción de la capa raster se ha realizado sobre una matriz de 250×250, se ha acotado un cuadrado para el estudio comprendido entre la longitud -2,40º a -0,65º y una latitud de 38,83º a 37,30º. La solución obtenida aparece en la figura 10 y en la figura 11. El código del programa desarrollado aparece en el apéndice A.

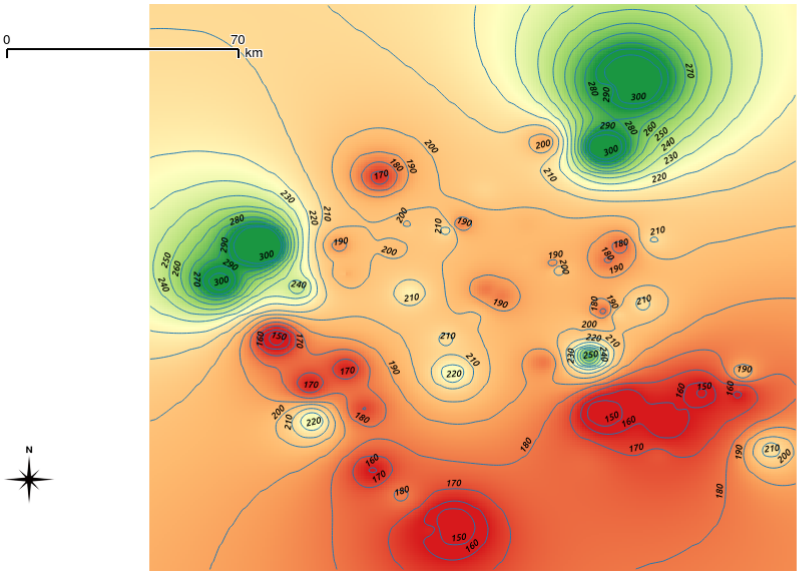

Figura 10: Kriging No Estocástico de Precipitaciones Región de Murcia

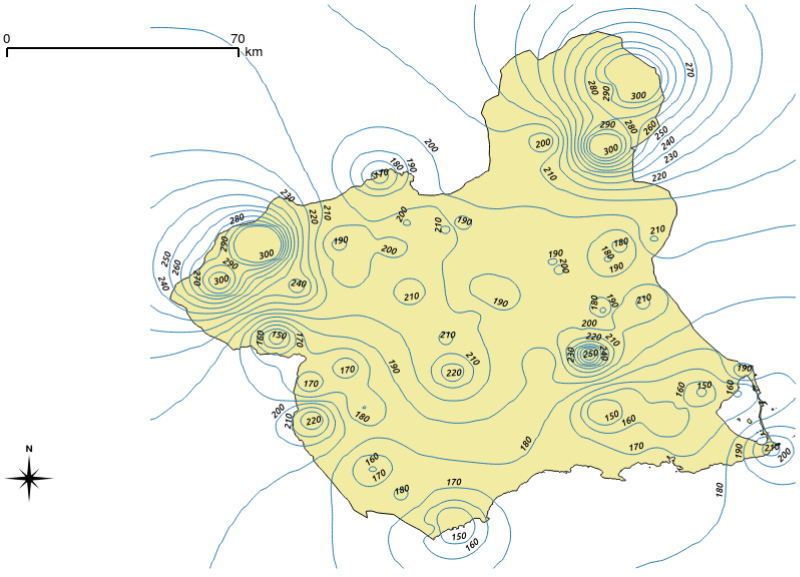

Figura 11: Isovalores Kriging No Estocástico Precipitaciones Región de Murcia

4. Kriging. Métodos Robustos.

4.1. Estimadores Robustos del Variograma.

Quizás el punto más importante para robustizar el Kriging es el cálculo robusto del variograma y se han propuesto varias soluciones que describimos a continuación.

4.1.1. Estimadores de Cressie y Hawkins.

Una primera estimación robusta del variograma en el caso de datos Gaussianos y de media constante fué propuesta por Cressie y Hawkins en 1980, mediante la expresión.

$$2\bar{\gamma}(\mathbf{h}) = \{\frac{1}{|N(\mathbf{h})|}\sum_{N(\mathbf{h})}(|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|)^{1/2}\}^4/(0,457 + 0,494/|N(\mathbf{h})|)$$

y otra análoga para los mismos supuestos

$$2\tilde{\gamma}(\mathbf{h}) = [med\{|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} : (\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})\}]^4/B(\mathbf{h})$$

donde *med* es la mediana del conjunto de puntos $|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2}$ y $B(\mathbf{h})$ es asintóticamente igual a 0,457.

4.1.2. Estimador de Escala. Sustitución del cálculo de la media por la mediana.

Otro estimador robusto llamado de escala se calcula como una variación del estimador de Matheron sustituyendo en el cálculo de este la media por la mediana,

$$med\{(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 : (\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})\}$$

donde *med* es la mediana del conjunto $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$.

4.1.3. Estimadores de Cuantiles.

La estimación robusta del variograma llamada de cuantiles se define por la expresión:

$$[UQ\{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\} - LQ\{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\}]^2$$

donde $UQ\{\}$ y $LQ\{\}$ corresponden a cuantiles superior (75 %) e inferior (25 %) respectivamente de los conjuntos de valores $Z(\mathbf{s}_i) - Z(\mathbf{s}_j) \in N(\mathbf{h})$.

4.1.4. Estimador de Cressie.

En la definición que hemos dado de los estimadores robustos de Cressie y Hawkins para $2\bar{\gamma}$, aparecía una expresión similar en la forma a:

$$\bar{A}(\mathbf{h}) = \sum_{N(\mathbf{h})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} / |N(\mathbf{h})|$$

Bajo la asunción de modelo Gausiano y evaluado $var(\bar{A}(\mathbf{h}))$, Cressie obtuvo una expresión para $cov(\bar{A}(h_s), \bar{A}(h_t))$ y hallando a continuación una expresión para $cov(2\bar{\gamma}(h(i)), 2\bar{\gamma}(h(j))$. A partir de aquí una estimación robusta del parámetro θ se puede hacer minimizando

$$\sum_{j=1}^{K} |N(h(j))| \{ \frac{\bar{\gamma}(h(j))}{\gamma(h(j), \theta)} - 1 \}^2$$

sobre $\theta \in \Theta$.

Un tratamiento más general y moderno del problema de la robustez en la estimación del Variograma pasa por los llamados *M-estimadores de escala* que describimos a continuación.

4.2. Estimadores de Escala.

Dada una Muestra $\{V_1, \ldots, V_n\}$, se llama estimador de escala de dicha muestra a cualquier función positiva $S_n(V_1, \ldots, V_n)$ de dicha muestra, que verifica:

$$S_n(\alpha V_1 + \beta, \ldots, \alpha V_n + \beta) = |\alpha| S_n(V_1, \ldots, V_n) \quad \forall \alpha \in R,\ \forall \beta \in R$$

En el proceso estocástico del Kriging $Z(\cdot)$, la v.a. de diferencias con retardo $\mathbf{h}$ lo podemos expresar como $V(\mathbf{h}) = Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})$ y tiene media cero y varianza $2\gamma(\mathbf{h})$. Entonces si $\{V_1(\mathbf{h}), \ldots, V_{N_h}(\mathbf{h})\}$ es una muestra de $V(\mathbf{h})$ correspondiente a la muestra $\{Z(\mathbf{X}_1), \ldots, Z(\mathbf{X}_n)$ de Z, el variograma clásico de Matheron toma la forma:

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{N_h} \sum_{i=1}^{N_h} V_i(\mathbf{h})^2 \quad \mathbf{h} \in R^d$$

y es simplemente el estimador clásico de la varianza de la muestra de $\{V_1(\mathbf{h})$, $\ldots, V_{N_h}(\mathbf{h})\}$.

Podemos utilizar ahora la teoría de los *M-estimadores de escala* para derivar las propiedades de robustez que vamos buscando. Al final esto nos llevará a calcular un estimador del variograma altamente robusto.

4.3. *M-estimadores* de Escala.

Supongamos que tenemos unas observaciones unidimensionales $V_1(\mathbf{h})$, $\ldots, V_{N_h}(\mathbf{h})$ que están idénticamentes distribuidas y siguiendo una distribución del modelo paramétrico $\{F_\sigma;\ \sigma > 0\}$, donde se verifica que $F_\sigma(\nu) = F(\nu/\sigma)$. Un *M-estimadores de escala* $S_{N_h}(V_1(\mathbf{h}), \ldots, V_{N_h}(\mathbf{h}))$ de σ se define por la ecuación implícita

$$\sum_{i=1}^{N_h} \chi(V_i(\mathbf{h})/S_{N_h}) = 0$$

y corresponde asintóticamente al funcional estadístico S definido por

$$\int \chi(\nu/S(F))dF(\nu) = 0$$

donde χ es una función real, simétrica (par) y suficientemente regular.

La función de influencia de un *M-estimadores de escala* S para una distribución F se calcula con la expresión

$$IF(\nu, S, F) = \frac{\chi(\nu/S(F))S^2(F)}{\int \nu\chi'(\nu/S(F))dF(\nu)}$$

La función de influencia es importante pues su interpretación heurística describe los efectos en el estimador de una contaminación infinitesimal en el punto ν. Como consecuencia un resumen de la función de influencia es el valor γ^* que mide la sensibilidad de S para F y se define como

$$\gamma^* = \sup_{\nu} |IF(\nu, S, F)|$$

Esta cantidad mide la peor influencia que una pequeña contaminación tiene en el estimador. Es deseable que γ^* sea finita, en cuyo caso S es *B-robusto* para F.

Otra importante propiedad de robustez de un estimador de escala es el *punto de ruptura* ε^*, que indica cuantos puntos de datos necesitan ser reemplazados para hacer explotar al estimador (tender a infinito) o tender a cero. En el caso de *M-estimadores de escala* se tiene que

$$\varepsilon^* = min(\frac{-\chi(0)}{\chi(+\infty) - \chi(0)}, \frac{\chi(+\infty)}{\chi(+\infty) - \chi(0)}) \leq \frac{1}{2}$$

En el caso de tomar $\chi(\nu) = |\nu|^q - \int |\nu|^q dF(\nu)$ para $q > 0$, que tienen los llamados L^q *M-estimadores* de escala que se demuestra que nunca son *B-robustos* para todo valor $q > 0$ es decir $\gamma^* = \infty$. y que $\varepsilon^* = 0\,\%$ para cualquier valor de $q > 0$.

4.4. Un Estimador Altamente Robusto del Variograma.

En el contexto de la estimación de escala, se ha propuesto un estimador altamente robusto llamado Q_{N_h} y que se define por la expresión

$$Q_{N_h} = 2,2191\{|V_i(\mathbf{h}) - V_j(\mathbf{h})|; i < j\}_{(k)}$$

donde el factor $2,2191$ es por consistencia con la distribución gausiana

$$k = \binom{[N_h/2] + 1}{2}$$

Esta expresión significa que ordenamos el conjunto de todas las diferencias absolutas $|V_i(\mathbf{h}) - V_j(\mathbf{h})|$ para $i < j$ y calculamos el k-ésimo cuantil, a continuación lo multiplicamos por $2,2191$ y obtenemos Q_{N_h}.

Este estimador Q_{N_h} tiene una $\varepsilon^* = 50\,\%$ como punto de ruptura, el más alto posible y una función de influencia acotada con $\gamma^* = 2,069$ con la distribución gausiana estandard.

Usando estas definiciones de Q_{N_h} se define un estimador altamente robusto del variograma con la expresión

$$2\hat{\gamma}(\mathbf{h}) = (Q_{N_h})^2, \qquad \mathbf{h} \in R^d$$

que por supuesto, este estimador tiene las mismas propiedades de robustez que Q_{N_h}.

4.5. Análisis de los Parámetros de Robustez de los Estimadores.

Vamos a estudiar los diferentes estimadores para determinar sus características en relación a la robustez, así como sus propiedades desde un punto de vista clásico.

4.5.1. El Estimador de Matheron del Variograma. Parámetros de Robustez.

El estimador de Matheron para el Variograma se define por la expresión:

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 \quad \mathbf{h} \in R^d$$

donde

$$N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \ \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}; \ i, j = 1, 2, \dots \ n\}$$

Definimos la Variable aleatoria $Y(\mathbf{h})$ como diferencia de $Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$, y con función de distribución F:

$$Y(\mathbf{h}) = Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$$

De esta manera el Funcional de la estimación será

$$T(F) = \int y^2 dF$$

Podemos ahora aplicarlo a la distribución contaminada resultando pues

$$T((1-\varepsilon)F + \varepsilon\delta_x) = \int y^2 d((1-\varepsilon)F + \varepsilon\delta_x)(y) =$$
$$= (1-\varepsilon) \int y^2 dF + \varepsilon x^2$$

Calculando su derivada respecto a ε en el punto $\varepsilon = 0$ obtenemos la función de influencia,

$$IF(x, T, F) = \frac{d}{d\varepsilon} F_\varepsilon \big|_{\varepsilon=0} = -\int y^2 dF + x^2$$

Tenemos como cálculo de la varianza de y

$$\sigma^2 = \int y^2 dF - \mu^2$$

siendo μ la media de y.

Sustituyendo en la expresión de $IF(x, T, F)$ tenemos

$$IF(x, T, F) = x^2 - (\sigma^2 + \mu^2)$$

Esta expresión nos permite estudiar el comportamiento robusto del estimador de Matheron.

La sensibilidad a grandes errores será

$$\gamma^* = \sup_x |IF(x; T, F)| = \infty$$

La sensibilidad a cambios locales

$$\lambda^* = \sup \frac{|IF(y; T, F) - IF(x; T, F)|}{|y - x|} = \infty$$

pues la pendiente de la parábola tiende a ∞ al crecer x.

El punto de rechazo será

$$\rho^* = \text{ínf}\{r > 0 | IF(x; T, F) = 0 \ \forall |x| > r\} = \infty$$

pues no se anula IF al crecer $|x|$ en cualquier cantidad.

El punto de ruptura será

$$\varepsilon^* = \text{ínf}\{\varepsilon > 0 | T(F_\varepsilon) \textit{ no esta acotado superiormente como } funcion \textit{ de } x\} = 0$$

pues $\forall \varepsilon > 0 \;\; T(F_\varepsilon)$ no está acotada.

Como se evidencia por los valores obtenidos de todos estos parámetros, que el estimador de Matheron se comporta muy mal ante los criterios de robustez. Se han desarrollado modificaciones para su cálculo que intentan mejorar este comportamiento.

4.5.2. Estimador de Cressie y Hawkins.

Como hemos indicado en un apartado anterior, Cressie y Hawkins proponen un estimador del variograma más robusto por medio de la ecuación:

$$2\bar{\gamma}(\mathbf{h}) = \{\frac{1}{|N(\mathbf{h})|}\sum_{N(\mathbf{h})}(|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|)^{1/2}\}^4/(0,457 + 0,494/|N(\mathbf{h})|)$$

El funcional de este estimador es

$$T(F) = \frac{1}{0,457}(\int_{-\infty}^{+\infty} |y|^{1/2} dF(y))^4$$

Aplicando este funcional a la distribución contaminada, se tendrá

$$T((1-\varepsilon)F + \delta_x) = \frac{1}{0,457}((1-\varepsilon)\int_{-\infty}^{+\infty} |y|^{1/2} dF(y) + \varepsilon \int_{-\infty}^{+\infty} |y|^{1/2} d\delta_x(y))^4$$

operando

$$T(F_\varepsilon) = \frac{1}{0,457}((1-\varepsilon)\int_{-\infty}^{+\infty} |y|^{1/2} dF(y) + \varepsilon |x|^{1/2})^4$$

Si llamamos Δ a la integral que nos aparece, se tiene

$$\Delta = \int_{-\infty}^{+\infty} |y|^{1/2} dF(y) = \int_{-\infty}^{0} (-y)^{1/2} dF + \int_{0}^{+\infty} y^{1/2} dF$$

resultará sustituyendo

$$T(F_\varepsilon) = \frac{1}{0,457}((1-\varepsilon)\Delta + \varepsilon|x|^{1/2})^4$$

Derivamos esta expresión de ε y tendremos,

$$\frac{d}{d\varepsilon}T(F_\varepsilon) = \frac{4}{0,457}((1-\varepsilon)\Delta + \varepsilon|x|^{1/2})^3(|x|^{1/2} - \Delta)$$

Tomando limites para ε tendiendo a 0, resulta

$$IF(x; T, F) = \frac{4}{0,457}\Delta^3(|x|^{1/2} - \Delta)$$

Esta ecuación se corresponde a una parábola invertida, que para valores negativos de x toma el mismo valore que su contrario $|x|$. En esta condiciones, los parámetros de robustez serán:

1. Sensibilidad a grandes errores γ^* no esta acotada, por no estarlo la función de influencia.

2. Sensibilidad a cambios locales $\lambda^* = \infty$, por ser de pendiente infinita en el origen $x = 0$.

3. Punto de rechazo $\rho^* = \infty$ por no anularse nunca la función de influencia.

4. Punto de ruptura $\varepsilon^* = 0$, pues para $\forall \varepsilon > 0$ como función de x $T(F_\varepsilon)$ no esta acotada.

4.5.3. Estimador de Escala. Sustitución del cálculo de la media por la mediana.

Una primera mejora de la estimación del Variograma es sustituir la media que aparece en el estimador de Matheron por la mediana, es decir

$$2\hat{\gamma}(\mathbf{h}) = med\{(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 : (\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})\}$$

donde *med* es la mediana del conjunto $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$.

Utilizando la variable $Y(\mathbf{h})$ definida en el apartado anterior, tenemos

$$2\hat{\gamma}(\mathbf{h}) = med\{Y(\mathbf{h})^2\}$$

Podemos definir una nueva variable aleatoria $X = Y^2$, donde se tendrá pues que,

$$2\hat{\gamma}(\mathbf{h}) = med\{X\} \quad y \quad Y = \pm\sqrt{X}$$

Si f es la función de densidad de Y y g la función de densidad de X, se verificará que

$$g(x) = f(\sqrt{x})\frac{1}{2\sqrt{x}} + f(-\sqrt{x})\frac{1}{2\sqrt{x}} \qquad x > 0$$

he integrando podemos obtener la función de distribución

$$G(x) = F(\sqrt{x}) - F(-\sqrt{x}) \qquad x > 0$$
$$G(x) = 0 \qquad x <= 0$$

El funcional correspondiente a la mediana es $T(F) = F^{-1}(1/2)$ que aplicado a la distribución contaminada G_ε resulta

$$T(G_\varepsilon) = G_\varepsilon^{-1}(1/2)$$

Se tiene que

$$G_\varepsilon(G_\varepsilon^{-1}(1/2)) = 1/2$$

Desarrollando la expresión de la distribución contaminada, se tendrá

$$(1-\varepsilon)G(G_\varepsilon^{-1}(1/2)) + \varepsilon\delta_x(G_\varepsilon^{-1}(1/2)) = 1/2$$

Sustituyendo el valor de G en función de la distribución F, se tiene

$$(1-\varepsilon)(F(\sqrt{G_\varepsilon^{-1}(1/2)}) - F(-\sqrt{G_\varepsilon^{-1}(1/2)})) + \varepsilon\delta_x(G_\varepsilon^{-1}(1/2)) = 1/2$$

Derivado respecto a ε se tiene

$$\begin{gathered}
-(F(\sqrt{G_\varepsilon^{-1}(1/2)}) - F(-\sqrt{G_\varepsilon^{-1}(1/2)}))+ \\
+(1-\varepsilon)(f(\sqrt{G_\varepsilon^{-1}(1/2)})\frac{d}{d\varepsilon}G_\varepsilon^{-1}(1/2)\frac{1}{2\sqrt{G_\varepsilon^{-1}(1/2)}}+ \\
+f(-\sqrt{G_\varepsilon^{-1}(1/2)})\frac{d}{d\varepsilon}G_\varepsilon^{-1}(1/2)\frac{1}{2\sqrt{G_\varepsilon^{-1}(1/2)}}) + \delta_x(G_\varepsilon^{-1}(1/2)) + \varepsilon.0 = 0
\end{gathered}$$

Operando se llega a la expresión

$$\frac{d}{d\varepsilon}G_\varepsilon^{-1}(1/2) = \frac{(F(\sqrt{G_\varepsilon^{-1}(1/2)}) - F(-\sqrt{G_\varepsilon^{-1}(1/2)})) - \delta_x(G_\varepsilon^{-1}(1/2))}{(1-\varepsilon)(f(\sqrt{G_\varepsilon^{-1}(1/2)}) + f(-\sqrt{G_\varepsilon^{-1}(1/2)}))}2\sqrt{G_\varepsilon^{-1}(1/2)}$$

Tomando limites para ε tendiendo a cero

$$\frac{d}{d\varepsilon}G_{\varepsilon}^{-1}(1/2) = \frac{(F(\sqrt{G^{-1}(1/2)}) - F(-\sqrt{G^{-1}(1/2)})) - \delta_x(G^{-1}(1/2))}{(f(\sqrt{G^{-1}(1/2)}) + f(-\sqrt{G^{-1}(1/2)}))} 2\sqrt{G^{-1}(1/2)}$$

Obtenemos la función de influencia

$$IF(x;T,G) = \begin{cases} \frac{(F(\sqrt{G^{-1}(1/2)})-F(-\sqrt{G^{-1}(1/2)}))}{(f(\sqrt{G^{-1}(1/2)})+f(-\sqrt{G^{-1}(1/2)}))} 2\sqrt{G^{-1}(1/2)} & \text{si } x > G^{-1}(1/2). \\ \frac{(F(\sqrt{G^{-1}(1/2)})-F(-\sqrt{G^{-1}(1/2)}))-1}{(f(\sqrt{G^{-1}(1/2)})+f(-\sqrt{G^{-1}(1/2)}))} 2\sqrt{G^{-1}(1/2)}, & \text{si } x \leq G^{-1}(1/2) \end{cases} \tag{4}$$

donde $G^{-1}(1/2)$ se calcula resolviendo la ecuación en x:

$$1/2 = F(\sqrt{x}) - F(-\sqrt{x})$$

Se tiene por lo tanto que

$$IF(x;T,G) = \begin{cases} \frac{\sqrt{G^{-1}(1/2)}}{(f(\sqrt{G^{-1}(1/2)})+f(-\sqrt{G^{-1}(1/2)}))} & \text{si } x > G^{-1}(1/2). \\ \frac{-\sqrt{G^{-1}(1/2)}}{(f(\sqrt{G^{-1}(1/2)})+f(-\sqrt{G^{-1}(1/2)}))}, & \text{si } x \leq G^{-1}(1/2) \end{cases} \tag{5}$$

Esta función de influencia es un escalón en el punto $G^{-1}(1/2)$ y constante en el resto de valores de x. En estas condiciones se tiene que

1. Sensibilidad a grandes errores γ^* esta acotada, por estarlo la función de influencia.

2. Sensibilidad a cambios locales $\lambda^* = \infty$, por ser discontinua la curva de influencia.

3. Punto de rechazo $\rho^* = \infty$ por no anularse nunca la función de influencia.

4. Punto de ruptura $\varepsilon^* = 1/2$, pues es una mediana.

4.5.4. Estimador de Cressie y Hawkins - 2.

Otra aproximación en la búsqueda de robustez de la fórmula de Matheron ha sido sustituir esta media de cuadrados por la expresión:

$$2\tilde{\gamma}(\mathbf{h}) = [med\{|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} : (\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})\}]^4/B(\mathbf{h})$$

Definimos como antes $Y(\mathbf{h}) = Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$ de función de distribución F y de densidad f. Se tendrá pues

$$2\tilde{\gamma}(\mathbf{h}) = [med\{|Y(\mathbf{h})|^{1/2} : \mathbf{h} \in N(\mathbf{h})\}]^4/B(\mathbf{h})$$

Definimos una nueva variable aleatoria $X = |Y|^{1/2}$ de tal manera que tenemos

$$X = \begin{cases} Y^{1/2} & Y \geq 0 \\ (-Y)^{1/2} & Y < 0 \end{cases} \tag{6}$$

En estas condiciones si G es la función de distribución de X y g la función de densidad, se tiene

$$g(x) = 2x(f(x^2) + f(-x^2)) \quad para \quad x > 0 \quad y \quad g(x) = 0 \quad x \leq 0$$
$$G(x) = F(x^2) - F(-x^2) \quad para \quad x > 0 \quad y \quad G(x) = 0 \quad x \leq 0$$

La mediana de X será entonces

$$med(X) = G^{-1}(1/2)$$

Y el funcional de variable aleatoria del variograma será

$$T(H) = (G^{-1}(1/2))^4 \frac{1}{0,457}$$

Aplicándolo a la distribución contaminada

$$T(H_\varepsilon) = (G_\varepsilon^{-1}(1/2))^4 \frac{1}{0,457}$$

Podemos calcular la derivada respecto a ε,

$$\frac{dT(H_\varepsilon)}{d\varepsilon} = \frac{4}{0,457}(G_\varepsilon^{-1}(1/2))^3 \frac{dG_\varepsilon^{-1}(1/2)}{d\varepsilon} \tag{7}$$

Calculamos ahora $\frac{dG_\varepsilon^{-1}(1/2)}{d\varepsilon}$, para lo cual se tiene que

$$G_\varepsilon(G_\varepsilon^{-1}(1/2)) = 1/2$$

y sustituyendo la expresión de la distribución contaminada

$$(1-\varepsilon)G(G_\varepsilon^{-1}(1/2)) + \varepsilon\delta_x(G_\varepsilon^{-1}(1/2)) = 1/2$$

Derivando y operando en la expresión tenemos

$$\frac{d}{d\varepsilon}(G_\varepsilon^{-1}(1/2)) = \frac{G(G_\varepsilon^{-1}(1/2)) - \delta_x(G_\varepsilon^{-1}(1/2))}{(1-\varepsilon)g(G_\varepsilon^{-1}(1/2))}$$

Sustituyendo en la ecuación 7, se tiene

$$\frac{dT(H_\varepsilon)}{d\varepsilon} = \frac{4}{0,457}(G_\varepsilon^{-1}(1/2))^3 \frac{G(G_\varepsilon^{-1}(1/2)) - \delta_x(G_\varepsilon^{-1}(1/2))}{(1-\varepsilon)g(G_\varepsilon^{-1}(1/2))}$$

Tomando límites cuando ε tiende a 0, podemos poner

$$\frac{dT(H_\varepsilon)}{d\varepsilon}|_{\varepsilon=0} = \frac{4}{0,457}(G^{-1}(1/2))^3 \frac{1/2 - \delta_x(G^{-1}(1/2))}{g(G^{-1}(1/2))}$$

La función de influencia para el variograma será pues

$$IF(x;T,H) = \begin{cases} \frac{2(G^{-1}(1/2))^3}{0,457g(G^{-1}(1/2))} & x > G^{-1}(1/2) \\ \\ \frac{-2(G^{-1}(1/2))^3}{0,457g(G^{-1}(1/2))} & x \leq G^{-1}(1/2) \end{cases} \tag{8}$$

donde $G^{-1}(1/2)$ es la solución en x a la ecuación

$$1/2 = F(x^2) - F(-x^2)$$

Podemos determinar el comportamiento robusto de esta estimación del variograma

1. Sensibilidad a grandes errores γ^* esta acotada, por estarlo la función de influencia.

2. Sensibilidad a cambios locales $\lambda^* = \infty$, por ser discontinua la curva de influencia.

3. Punto de rechazo $\rho^* = \infty$ por no anularse nunca la función de influencia.

4. Punto de ruptura $\varepsilon^* = 1/2$, pues es una mediana.

4.5.5. Estimador de Cuantiles.

La estimación robusta del variograma llamada de cuantiles se define por la expresión:

$$[UQ\{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\} - LQ\{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\}]^2$$

Utilizando la variable aleatoria $Y(\mathbf{h}) = Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$, nos quedaría,

$$2\tilde{\gamma}(\mathbf{h}) = [UQ\{Y(\mathbf{h})\} - LQ\{Y(\mathbf{h})\}]^2$$

Si la función de distribución de Y es F y g su densidad tenemos que el funcional asociado al variograma en este caso es

$$T(H) = (F^{-1}(1-\alpha) - F^{-1}(\alpha))^2$$

donde en este caso $\alpha = 1/4$ y $1 - \alpha = 3/4$.

Aplicando el funcional a la distribución contaminada se tiene,

$$T(H_\varepsilon) = (F_\varepsilon^{-1}(1-\alpha) - F_\varepsilon^{-1}(\alpha))^2$$

Derivando obtenemos,

$$\frac{dT(H_\varepsilon)}{d\varepsilon} = 2(F_\varepsilon^{-1}(1-\alpha) - F_\varepsilon^{-1}(\alpha))(\frac{d}{d\varepsilon}F_\varepsilon^{-1}(1-\varepsilon) - \frac{d}{d\varepsilon}F_\varepsilon^{-1}(\alpha)) \quad (9)$$

Se tiene $F_\varepsilon(F_\varepsilon^{-1}(\alpha)) = \alpha$, y sustituyendo la distribución contaminada por su expresión, tenemos

$$(1-\varepsilon)F(F_\varepsilon^{-1}(\alpha))\varepsilon + \varepsilon\delta_x(F_\varepsilon^{-1}(\alpha)) = \alpha$$

Derivando se tiene

$$-F(F_\varepsilon^{-1}(\alpha)) + (1-\varepsilon)f(F_\varepsilon^{-1}(\alpha))\frac{dF_\varepsilon^{-1}(\alpha)}{d\varepsilon} + \delta_x(F_\varepsilon^{-1}(\alpha)) = 0$$

operando

$$\frac{dF_{\varepsilon}^{-1}(\alpha)}{d\varepsilon} = \frac{F(F_{\varepsilon}^{-1}(\alpha)) - \delta_x(F_{\varepsilon}^{-1}(\alpha))}{(1-\varepsilon)f(F_{\varepsilon}^{-1}(\alpha))}$$

análogamente para $(1-\alpha)$, se obtiene

$$\frac{dF_{\varepsilon}^{-1}(1-\alpha)}{d\varepsilon} = \frac{F(F_{\varepsilon}^{-1}(1-\alpha)) - \delta_x(F_{\varepsilon}^{-1}(1-\alpha))}{(1-\varepsilon)f(F_{\varepsilon}^{-1}(1-\alpha))}$$

Sustituyendo estas dos expresiones últimas en la ecuación 9, y tomando límites cuando ε tiende a 0, se tiene operando

$$IF(x;T,H) = 2(F^{-1}(1-\alpha)-F^{-1}(\alpha))(\frac{(1-\alpha)-\delta_x(F^{-1}(1-\alpha))}{f(F^{-1}(1-\alpha)} - \frac{(\alpha)-\delta_x(F^{-1}(\alpha))}{f(F^{-1}(\alpha)})$$

evaluando δ_x podemos poner

$$IF(x,T,H) = \begin{cases} 2(F^{-1}(1-\alpha)-F^{-1}(\alpha))(\frac{-\alpha}{f(F^{-1}(1-\alpha)} - \frac{\alpha-1}{f(F^{-1}(\alpha)}) & x < F^{-1}(\alpha) \\ 2(F^{-1}(1-\alpha)-F^{-1}(\alpha))(\frac{-\alpha}{f(F^{-1}(1-\alpha)} - \frac{\alpha}{f(F^{-1}(\alpha)}) & F^{-1}(\alpha) \leq x < F^{-1}(1-\alpha) \\ 2(F^{-1}(1-\alpha)-F^{-1}(\alpha))(\frac{1-\alpha}{f(F^{-1}(1-\alpha)} - \frac{\alpha}{f(F^{-1}(\alpha)}) & F^{-1}(1-\alpha) < x \end{cases}$$

Esta expresión es una función escalonada de dos escalones. La propiedades de robustez que se obtienen de esta función de influencia son

1. Sensibilidad a grandes errores γ^* esta acotada, por estarlo la función de influencia.
2. Sensibilidad a cambios locales $\lambda^* = \infty$, por ser discontinua la curva de influencia.

3. Punto de rechazo $\rho^* = \infty$ por no anularse nunca la función de influencia.

4. Punto de ruptura $\varepsilon^* = \alpha$, pues modificando esa fracción de datos cambiamos las funciones $UQ(Y)$ o $LQ(Y)$.

4.6. Análisis Clásico de los Estimadores.

A continuación procedemos a realizar el estudio clásico de los estimadores robustos.

4.6.1. El Estimador de Matheron del Variograma.

Como hemos indicado en los puntos anteriores el estimador de Matheron, para el cáculo del variograma, se define con la expresión

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 \quad \mathbf{h} \in R^d$$

donde

$$N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \ \mathbf{s}_i - \mathbf{s}_j = \mathbf{h};\ i, j = 1, 2, \dots \ n\}$$

Vamos a calcular las características clásicas de este estimador y que enumeramos a continuación.

1. *Carencia de sesgo.* Se dice que un estimador es insesgado si el valor esperado de su distribución de probabilidad es igual al parámetro.

   En nuestro caso el parámetro es el variograma que se define con la expresión

$$2\gamma(\mathbf{s}_1 - \mathbf{s}_2) = var(Z(\mathbf{s}_1) - Z(\mathbf{s}_2))$$

Si definimos la variable aleatoria $Y(\mathbf{h}) = Z(\mathbf{s}_1) - Z(\mathbf{s}_2)$, la expresión del parámetro es

$$2\gamma(\mathbf{h}) = var(Y(\mathbf{h}))$$

y el estimador sería

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} Y(\mathbf{h})^2$$

Se tiene que

$$E(2\hat{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} E(Y_i^2) = \frac{1}{n} \sum_{i=1}^{n} (\sigma^2 + \mu^2) = \sigma^2 + \mu^2$$

siendo σ^2 la varianza de Y y μ su media.

Si introducimos la condición de procesos estacionarios $\mu = 0$ y se cumple

$$E(2\hat{\gamma}) = \sigma^2 = 2\gamma$$

Por lo tanto el estimador de Matheron es insesgado.

2. *Consistencia.* Un estimador es consistente si, además de carecer de sesgo, se aproxima cada vez más al valor del parámetro a medida que aumenta el tamaño de la muestra. Esto es equivalente a que la varianza del estimador disminuye hacia cero al crecer la muestra.

   En nuestro caso

$$var(\frac{1}{n} \sum_{i=1}^{n} Y_i^2) = \frac{1}{n^2} \sum_{i=1}^{n} var(Y_i^2) = \frac{1}{n} var(Y^2)$$

   pues las v.a. Y_i^2 son independientes por serlo las Y_i. Tomando limites se tiene

$$\lim_{n \to \infty} var(\frac{1}{n} \sum_{i=1}^{n} Y_i^2) = 0$$

   Por lo tanto el estimador de Matheron es Consistente.

3. *Suficiencia.* Un estimador es suficiente cuando en su cálculo se emplea toda la información de la muestra.

   En el caso del estimador de Matheron es suficiente pues los n valores Y_i aparecen en la fórmula de cálculo.

4. *Eficiencia.* Se dice que in estimador $\hat{E}_1$ es más eficiente que otro $\hat{E}_2$ si se cumple que

$$var(\hat{E}_1) < var(\hat{E}_2)$$

   que para el estadístico de Matheron, habría que verificar en cada caso.

4.6.2. Estimador de Cressie y Hawkins.

Cressie y Hawkins proponen un estimador del variograma más robusto por medio de la ecuación:

$$2\bar{\gamma}(\mathbf{h}) = \{\frac{1}{|N(\mathbf{h})|}\sum_{N(\mathbf{h})}(|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|)^{1/2}\}^4/(0,457 + 0,494/|N(\mathbf{h})|)$$

Introduciendo la variable aleatoria Y, se tiene

$$2\bar{\gamma}(\mathbf{h}) = \frac{1}{0,457 + 0,494/n}(\frac{1}{n}\sum_{i=1}^{n}|Y_i|^{1/2})^4$$

Aplicando el operador E a los dos miembros, se tiene

$$E(2\bar{\gamma}(\mathbf{h})) = \frac{1}{0,457 + 0,494/n}\frac{1}{n^4}E((\sum_{i=1}^{n}|Y_i|^{1/2})^4)$$

Con esta información vemos $E(2\bar{\gamma}(\mathbf{h})) \neq 2\gamma$ pues un miembro es una constante y el otro miembro es una función de n. En estas circunstancias tenemos

1. El estimador de Cressie y Hawkins no es insesgado.

2. No es consistente por no ser insesgado.

3. Es suficiente pues utiliza toda la información de la muestra.

Esta afirmación de que el estimador no es insesgado, aunque parece evidente, desde el punto de vista matemático no es riguroso. Intento hacer una demostración por reducción al absurdo en el apéndice, pero obtengo un resultado no del todo satisfactorio. Ver apendice A.

4.6.3. Sustitución de la Media por la Mediana.

En esta aproximación al cálculo del variograma, se sustituye la media de la muestra por la mediana de dicha muestra y que matemáticamente sería

$$2\hat{\gamma}(\mathbf{h}) = med\{(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 : (\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})\}$$

Introduciendo la v.a. Y se tiene que

$$2\hat{\gamma}(\mathbf{h}) = med\{(Y(\mathbf{h}))^2 : \mathbf{h} \in N(\mathbf{h})\}$$

Podemos expresar la función de densidad de Y^2 g en relación con la de Y que llamaremos f. Se puede verificar que se cumple

$$g(x) = f(-\sqrt{x})\frac{1}{2\sqrt{x}} + f(\sqrt{x})\frac{1}{2\sqrt{x}} \quad \forall x > 0$$

y de valor cero para $x \leq 0$.

La función función de distribución será integrando

$$G(x) = F(\sqrt{x}) - F(-\sqrt{x}) \quad \forall x > 0$$

La función de densidad de la mediana es para n impar

$$h(s) = n!g(s)\frac{G(s)^{\frac{n-1}{2}}(1-G(s))^{\frac{n-1}{2}}}{(\frac{n-1}{2})!(\frac{n-1}{2})!}$$

Y la media de la estimación del variograma sería

$$E(2\hat{\gamma}) = \int_0^{+\infty} sh(s)ds$$

Sustituyendo el valor de $h(s)$ y haciendo el cambio de variable $G(s) = u$ se llega a la expresión

$$E(2\hat{\gamma}) = \frac{n!}{(\frac{n-1}{2})!(\frac{n-1}{2})!}\int_0^1 u^{\frac{n-1}{2}}(1-u)^{\frac{n-1}{2}}G^{-1}(u)du$$

y análogamente para n par

$$E(2\hat{\gamma}) = \frac{n!}{(\frac{n}{2}-1)!(\frac{n}{2})!}\int_0^1 u^{\frac{n}{2}-1}(1-u)^{\frac{n}{2}}G^{-1}(u)du$$

En estas expresiones, $E(2\hat{\gamma})$ es función de n que no se cancela en ningún momento. Podemos decir claramente, aunque no de forma rigurosa que

1. Este estimador no es insesgado.
2. No es consistente por no ser insesgado.
3. No es suficiente pues utiliza solo la mitad de información de la muestra a la hora de calcular la mediana.

4.6.4. Estimador de Cuantiles.

En este caso la estimación del variograma, llamada de cuantiles, se define por la expresión:

$$[UQ\{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\} - LQ\{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\}]^2$$

Nuevamente introduciendo la variable Y, se tiene

$$2\hat{\gamma} = [UQ\{Y(\mathbf{h})\} - LQ\{Y(\mathbf{h})\}]^2$$

Haciendo un desarrollo análogo al realizado para la mediana podemos ver que $E(2\hat{\gamma})$ depende del parámetro n y por lo tanto se tiene

1. Este estimador no es insesgado.

2. No es consistente por no ser insesgado.

3. No es suficiente pues utiliza solo la mitad de información de la muestra a la hora de calcular la mediana.

4.7. El Variograma Robusto en R.

Para el cálculo del variograma robusto en R se utilizan la misma función que indicamos en el apartado 2.6 pero incluyendo ahora un parámetro adicional *estimator.type="modulus"*. En R sería

```
#Calculo del Variograma Robusto
varior  = variog(s100,estimator.type="modulus", max.dist=0.6)

#Visualizacion
plot(varior)
```

En las restantes tareas de ajuste de curva se procedería de forma análoga a como lo hicimos en el caso normal, pero ahora utilizando el variograma robusto.

4.8. Kriging Robusto.

El kriging robusto se obtiene como una modificación del Ordinario pero introduciendo consideraciones de estadística robusta. Estas modificaciones son:

1. Estimamos el variograma utilizando uno de los procedimientos robustos que hemos indicado anteriormente y ajustamos un modelo de variograma válido.

2. Usando el variograma anterior calculamos los pesos de kriging para cada posición j, a partir de las restantes $i \neq j$.

$$\hat{Z}_{-j}(\mathbf{s}_j) = \sum_{\substack{i=1 \\ i \neq j}}^{n} \lambda_{ji} Z(\mathbf{s}_i)$$

   calculando la varianza asociada al kriging para cada j $\sigma^2_{-j}(\mathbf{s}_j)$.

3. Con los pesos calculados estimamos una predicción robusta de $Z(\mathbf{s}_j)$ a partir de los nodos vecinos.

$$Z^{@}_{-j}(\mathbf{s}_j) = mediana\ ponderada(\{Z(\mathbf{s}_i) :\ i \neq j\}, \{\lambda_{ji} :\ i \neq j\})$$

4. Editar $Z(\mathbf{s}_j)$ reemplazándolas con las versiones Winsorizadas.

$$Z^{(e)}(\mathbf{s}_j) = \begin{cases} Z^{@}_{-j}(\mathbf{s}_j) + c\sigma_{-j}(\mathbf{s}_j), & si\ Z(\mathbf{s}_j) - Z^{@}_{-j}(\mathbf{s}_j) > c\sigma_{-j}(\mathbf{s}_j) \\ Z(\mathbf{s}_j), & si\ |Z(\mathbf{s}_j) - Z^{@}_{-j}(\mathbf{s}_j)| \leq c\sigma_{-j}(\mathbf{s}_j) \\ Z^{@}_{-j}(\mathbf{s}_j) - c\sigma_{-j}(\mathbf{s}_j), & si\ Z(\mathbf{s}_j) - Z^{@}_{-j}(\mathbf{s}_j) < -c\sigma_{-j}(\mathbf{s}_j) \end{cases}$$

5. Usando el variograma robusto estimamos $Z(B)$. Calculamos para ello los pesos $\{\lambda_{Bi} :\ i = 1, \ldots n\}$ mediante el método del Kriging Ordinario.

$$\hat{p}(\mathbf{Z}, B) = \sum_{i=1}^{n} \lambda_{Bi} Z(\mathbf{s}_i)$$

6. Para hacer la estimación utilizamos los datos editados $\{Z^{(e)}(\mathbf{s}_i) : \ i = 1, \dots n\}$.

$$\hat{Z}(B) = \sum_{i=1}^{n} \lambda_{Bi} Z^{(e)}(\mathbf{s}_i)$$

4.9. Solución Robusta del Problema de las Precipitaciones de Murcia. Método No Estocástico.

Se ha desarrollado un pequeño programa en Python para hacer el Kriging Robusto del problema de las precipitaciones de la Región de Murcia. Análogamente como el método no robusto se ha hecho un cuadriculado de 250×250 de la región de estudio, comprendida entre la longitud -2,40º a -0,65º y una latitud de 38,83º a 37,30º que la cubre completamente.

El algoritmo que se ha utilizado para hacer robusto el kriging ha sido básicamente el de la mediana ponderada, con las siguientes particularidades:

1. Se calcula para cada punto objetivo s_0 de la cuadrícula la distancia d_i a las diferentes estaciones de muestreo.
2. Se calculan los pesos que le corresponden a cada estación con la fórmula:

$$w_i = \frac{1/d_i^2}{\sum_{i=1}^{n} 1/d_i^2} \qquad i = 1, \dots n$$

3. Se ordenan los n valores $w_i * valor_i$ de menor a mayor
4. Se busca el primer valor de m que cumple

$$\sum_{s=1}^{m} w_s > 1/2$$

5. La mediana será el valor medio de $w_{m-1}valor_{m-1}$ y $w_m valor_m$.
6. El valor estimado será:

$$s_0 = \frac{w_{m-1}valor_{m-1} + w_m valor_m}{w_{m-1} + w_m}$$

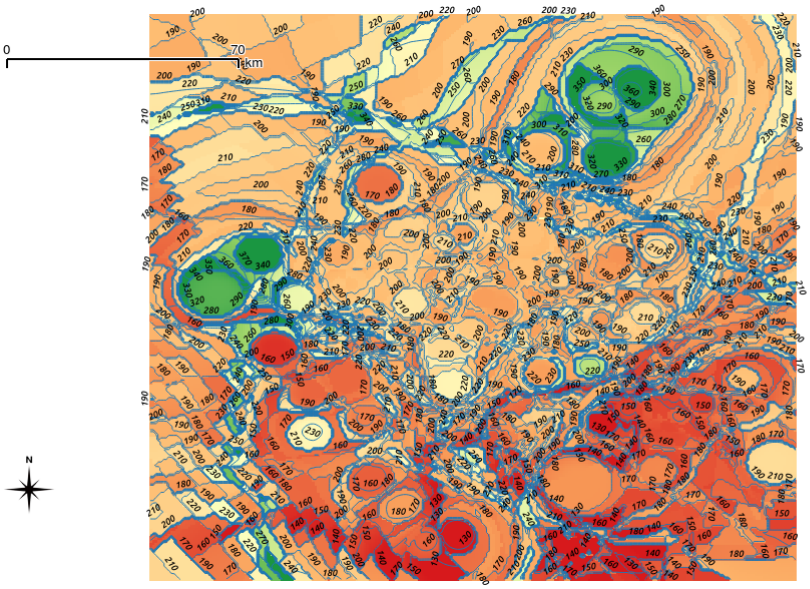

Figura 12: Kriging Robusto Precipitaciones Región de Murcia

Los resultados se muestran en la figura 12 y en la figura 13. Como puede observarse en el resultado siguen apareciendo las mismas zonas húmedas y secas, pero el gráfico se ha vuelto menos continuo. La única explicación que he podido encontrar es que la mediana es un valor más robusto que la media pero tiene por su definición presenta discontinuidades en su variación. No he encontrado explicación para las manchas húmedas que aparecen la parte superior izquierda de la zona de estudio que por lógica no deberían aparecer. El código del programa desarrollado aparece en el apéndice B.

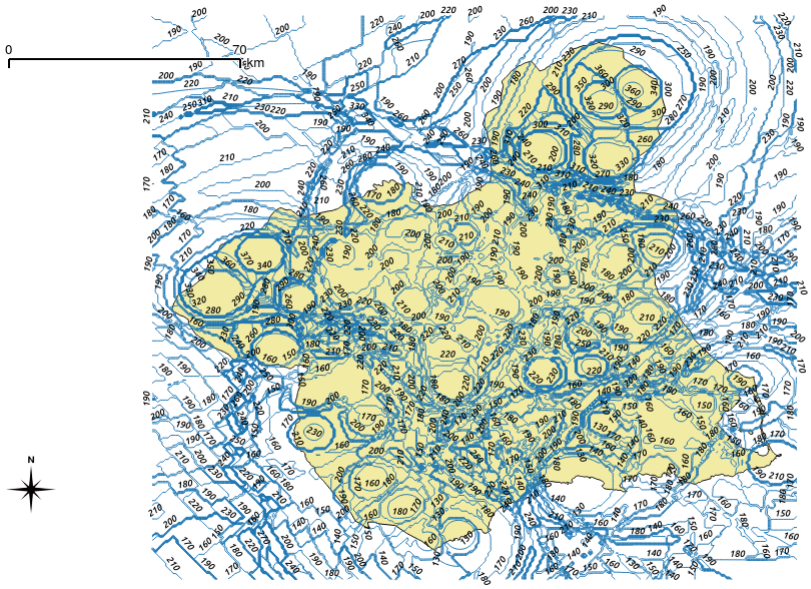

Figura 13: Isovalores Robusto Precipitaciones Región de Murcia

5. Procesos Puntuales.

Los Procesos Puntuales se caracterizan por ser procesos aleatorios espaciales donde el conjunto D en el que se soporta el fenómeno probabilístico es así mismo aleatorio. En el modelo general se tenía que

$$\{\mathbf{Z}(\mathbf{s}) : \ \mathbf{s} \in D\}$$

donde $\mathbf{Z}(\cdot)$ y D son ambos aleatorios.

En el estudio de los procesos puntuales, una cuestión que surge desde el principio es de que manera se encuentran distribuidos los puntos de D, pudiendo darse tres casos.

1. **Aleatoriedad Espacial Completa.** Los puntos de D están distribuidos de forma totalmente aleatoria, lo que es equivalente a que están distribuidos siguiendo un función de distribución de **Poisson homogénea** con intensidad espacial constante λ que no depende de la posición.
2. **Distribución Regular.** Los puntos de D se sitúan de forma regular, como si cada punto se separase de otro una distancia fija. En este caso los puntos tienen una correlación negativa entre ellos lo que les hace separarse al máximo unos de otros.
3. **Distribución por Grupos o Clustering.** Los puntos de D muestran tendencia a agruparse en ciertos lugares. En este caso se puede modelar con una distribución de **Poisson No Homogénea**, siendo λ función de la posición espacial.

En el análisis de problemas reales, es importante averiguar ante cual de los modelos anteriores nos encontramos, para ello se han desarrollado varias técnicas que abordamos a continuación.

5.1. Métodos Quadrat.

El muestreo Quadrat consiste en contar los sucesos en un subconjunto de la región de estudio A. Normalmente estos subconjuntos son rectangulares,

de ahí el nombre, aunque cualquier forma es posible. Los Quadrats pueden situarse aleatoriamente o bien de forma contigua en A.

5.1.1. Método Quadrats Aleatorios.

Básicamente el procedimiento consiste en evaluar cuantos sucesos ocurren en cada quadrat y posteriormente hacemos una clasificación de estos en base al número que hay en cada uno de ellos, dándonos una distribución de frecuencias del número de sucesos por quadrat.

Bajo aleatoriedad espacial completa (*csr*), el número de sucesos en un quadrat A_1, de área $|A_1|$, es una distribución de Poisson de media $\lambda|A_1|$, donde λ es la intensidad de Poisson del proceso. De esta manera, un text para evaluar *csr* es el test de Pearson χ^2 de bondad del ajuste.

Medida de la Regularidad y Agrupamiento. Una vez que se ha determinado que un proceso puntual no es *csr*, el siguiente paso es medir de alguna manera cuanto discrepan de este supuesto y en que sentido (agrupamiento o regularidad). Se han elaborado diferentes índices que incluimos en el cuadro 2.

5.1.2. Método de Aglomerado de Red de Quadrats Contiguos.

El método de los quadrats contiguos consiste básicamente en dividir la región de estudio en cuadrados adyacentes mediante una regilla de lineas que se cruzan y evaluar los sucesos que ocurren en cada uno de los cuadrados.

En el método de Aglomerado, pares de quadrats adyacentes son combinados en bloques de tamaño dos y sucesivamente recombinados en una sucesión de $2, 4, 8, 16, \ldots, K$ quadrats, posteriormente se analizan estos quadrats continuos usando un análisis de la varianza anidado.

Sea A_i y B_i el número de sucesos en el i par de bloques, cada uno conteniendo r quadrats. La suma inter-bloque de cuadrados por bloque de hasta r quadrats es,

Index	Estimador	Regularidad	CSR	Agrupamiento
I	$S^2/\bar{X}$	$I < 1$	$I = 1$	$I > 1$
ICS	$S^2/\bar{X} - 1$	$ICS < 0$	$ICS = 0$	$ICS > 0$
ICF	$\bar{X}^2/(S^2 - \bar{X})$			
$\overset{*}{X}$	$\bar{X} + S^2/\bar{X} - 1$			
IP	$\overset{*}{X}/\bar{X}$			
I_δ	$\frac{n\sum_{i=1}^{n} X_i(X_i-1)}{n\bar{X}(n\bar{X}-1)}$			

X_i es el número de sucesos en en quadrat i.
$\bar{X}$ es la media de sucesos en los quadrats.
S^2 es la varianza de la muestra.

Cuadro 2: Indices Valoración Quadrats

$$SS_r = \frac{1}{r}\sum_{i=1}^{m}(A_i^2 + B_i^2) - \frac{1}{2r}\sum_{i=1}^{m}(A_i + B_i)^2$$
$$= \frac{1}{2r}\sum_{i=1}^{m}(A_i - B_i)^2$$

donde $m = K/2r$.

La media de cuadrados es $MS_r = SS_r/m$. Asumiendo una distribución de Poisson en el conteo de sucesos, se tendrá que

$$E(MS_r) = \lambda|Q|,$$
$$var(MS_r) = [\lambda|Q|\{1 + 4r\lambda|Q|\}]/K$$

donde $|Q|$ es el área del quadrats.

La comparación de sucesivas medias de cuadrados se usa para comprobar aglomerados (clustering) en un tamaño de bloque dado. Para una de aceptación del 95 % y asumiendo normalidad se obtiene región de aceptación:

$$[m^{-1}MS_{r/2}\chi^2_{m,0.025}\ ,\ m^{-1}MS_{r/2}\chi^2_{m,0.975}]$$

5.1.3. Estimadores de la Función de Densidad.

Supongamos que dividimos las región de interés A en una red cuadrada de quadrats de tamaño $a \times a$. Sea $N(s, a)$ el número de sucesos en un cuadrado de $a \times a$ en la localización $\mathbf{s}$ y consideremos

$$\lambda_a(\mathbf{s}) = Pr(N(\mathbf{s}, a) > 0)/a^2$$

Si $\lambda_a(\mathbf{s}) \to \lambda(\mathbf{s})$ cuando $a \to 0$ para todo $s \in A$, al conjunto $\{\lambda(\mathbf{s}) : \mathbf{s} \in A\}$ se le denomina **función intensidad** del proceso puntual.

Para un nodo de la red cuadrada de dimensión $a \times a$, el conteo de sucesos para el quadrat por unidad de área, $N(\mathbf{s}, a)/a^2$, es un estimador insesgado de $\int_0^a \int_0^a \lambda(\mathbf{u}+\mathbf{s})d\mathbf{u}/a^2$. Cuando $\lambda(\cdot)$ no varía mucho sobre el cuadrado $a \times a$, esta integral es aproximadamente igual a $\lambda(\mathbf{s})$. Por lo tanto el conjunto de conteo de sucesos por unidad de área, estima $\{\lambda(\mathbf{s}) : \mathbf{s} \in A\}$.

5.2. Métodos de la Distancia.

Los métodos de la distancia hacen uso de la información precisa de donde han ocurrido los sucesos y tienen la ventaja de no depender de la elección arbitraria del tamaño del quadrat o de su forma.

Métodos del vecino más próximo. En estos métodos las distancias entre los sucesos o bien las distancias entre puntos de muestreo y los sucesos, son calculadas y relacionadas.

Las distancias pueden ser medidas entre sucesos y vecino mas próximo (W), o bien entre punto de muestreo y suceso más próximo (X). Los puntos de muestreo pueden situarse aleatoriamente o bien en una cuadricula.

La distribución teórica para la distancia del vecino más próximo, distancias W y X, bajo el supuesto de aleatoriedad espacial completa (csr) es bien conocida. En R^2, la densidad de la la variable W es

$$g(w) = 2\pi\lambda w e^{-\pi\lambda w^2}, \quad para \;\; w > 0$$

y su función de distribución será

$$G(w) = 1 - e^{-\pi\lambda w^2} \quad para \;\; w > 0$$

Para la variable X de distancia a un punto de muestreo al suceso más próximo, se tiene la misma distribución que W.

Si representamos por d_{ij} la distancia Euclídea entre dos localizaciones i y j, la distancia entre una localización i y la localización vecina más cercana será, lógicamente, $d_i = min_j\{d_{ij}, \; con \; i \neq j\}$, para $i = 1, \ldots n$. Por tanto, fijada una distancia w, el estimador de $G(w)$ será la función de distribución empírica

$$\hat{G}(w) = \frac{numero \; de \; d_i \leq w}{n}$$

Podemos, pues, aplicar tests estadísticos para verificar si la función $\hat{G}(w)$ se ajusta al modelo teórico propuesto y validar de esta manera si estamos ante un caso de csr.

Otros muchos estadísticos han sido propuestos para verificar la aleatoridad espacial completa, basados en una muestra de n puntos. Incluimos varios de ellos en el cuadro 3

5.2.1. Estimadores de la Función de Densidad.

De la misma manera que en los métodos quadrats, una vez que se ha determinado la no csr del proceso y en supuesto de que hemos modelado con una función de Poisson no homogénea, debemos estimar la variación espacial de este proceso.

Variable		Test Estadístico	Asintótico
W	A	$2\lambda^{1/2}\sum W_i/n$	$N(1,(4-\pi)/n\pi)$
W	B	$2\pi\lambda\sum W_i^2$	χ^2_{2n}
X	C	$\pi\lambda\sum X_i^2/n$	$N(1,1/n)$
X	D	$n(\sum X_i^2)/(\sum X_i)^2$	Por Simulación
X	E	$12n(nlog(\sum X_i^2/n) - \sum logX_i^2)/(7n+1)$	χ^2_{n-1}

Cuadro 3: Estadisticos de Vecindad. Bajo supuesto de *csr*

En este caso hay varias posibilidades, una de ellas es utilizar **Métodos Paramétricos**, consistentes en proponer una función cuyos parámetros son estimados por el método de la máxima verosimilitud. Esta vía permite incluir p covariables existentes $Z_j,\ j = 1,\ldots,p$ y utilizar, por ejemplo un modelo log-lineal de la forma

$$log\lambda(\mathbf{s}) = \sum_{j=1}^{p} \beta_j Z_j(\mathbf{s})$$

siendo $Z_j(\mathbf{s})\ j = 1,\ldots,p$ los valores que toman las covariables en la localización $\mathbf{s}$.

La segunda posibilidad en la estimación de la intensidad de un proceso de Poisson no homogéneo son los **Métodos no Paramétricos**, basados en el *Estimador de Núcleo Suavizado* dado por

$$\hat{\lambda}(\mathbf{s}) = \frac{1}{q(\parallel \mathbf{s} \parallel)h^2} \sum_{i=1}^{n} K(\frac{\parallel \mathbf{s} - \mathbf{s}_i \parallel}{h})$$

supuesto que se han observado n sucesos en localizaciones $\mathbf{s}_1,\ldots,\mathbf{s}_n$, siendo K la función núcleo considerada, $q(\mathbf{s})$ una corrección frontera para compensar los valores que se pierden cuando $\mathbf{s}$ está cerca de la frontera de la región de estudio A, y siendo h una medida del nivel de suavizado, también denominado

ancho de banda. Valores pequeños de h darán lugar a estimadores poco suaves y valores grandes a lo contrario.

La función núcleo habitualmente considerada es la denominada función cuártica, también denominada bipondera, y definida para localizaciones $\| \mathbf{s} \| \in (-1, 1)$, como

$$K(\mathbf{s}) = \frac{3}{\pi}(1 - \| \mathbf{s} \|)^2$$

y como 0 para localizaciones $\| \mathbf{s} \| \notin (-1, 1)$.

5.3. Aplicación del Análisis de Procesos Puntuales a la Sismicidad de la Península Ibérica.

Como una aplicación de los modelos teóricos de Procesos Puntuales, vamos a calcular un diagrama raster de los riesgos sísmicos en la península Ibérica. Para ello, accediendo a la web del Instituto Geográfico Nacional de España (http://www.ign.es/web/ign/portal/sis-catalogo-terremotos) me he bajado los datos de seísmos y sus localizaciones de longitud y latitud con objeto de poder hacer el estudio.

He acotado la longitud entre -10º y 5º y la latitud entre 34º y 45º, que incluye toda la península Ibérica y el norte de África, zona ésta muy sísmica y que de alguna manera está influyendo de manera considerable sobre la península. Se han seleccionado los terremotos de magnitud 2,5 o superior en los últimos 50 años.

Una muestra de los datos obtenidos aparecen en la figura 14. Las localizaciones de los seísmos, extendidas en el mapa aparecen en la figura 15.

Para resolver el problema con R he cargado primero las localizaciones de los seísmos en el entorno de R, así como el mapa de la península Ibérica para tener una referencia.

```
#Leer Puntos de Seismos
datos=read.table("/home/Seismos_Espana_2-5.csv",\
    sep="\t",dec=",",encoding="UTF-8",quote="",header=TRUE)
```

Evento	Fecha	Hora	Latitud	Longitud	Prof. (Km)	Inten.	Mag.	Localización
6328	04/01/1968	06:42:14	**37,5983**	**-2,1233**	5		4,0	**SW VÉLEZ-RUBIO.AL**
6329	04/01/1968	17:47:06	**42,6200**	**1,3800**			3,1	**NE ALINS.L**
6331	11/01/1968	10:48:52	**38,4800**	**-8,0583**	5	V	4,0	**SW ÉVORA.POR**
6336	15/01/1968	19:09:20	**42,9300**	**-0,6400**			3,2	**SW LARUNS.FRA**
6339	19/01/1968	20:23:42	**35,9000**	**-3,6500**			2,7	**MAR DE ALBORAN**
6340	20/01/1968	03:08:39	**43,0700**	**-0,6700**			3,4	**SW OLORON STE MARIE.FRA**
6341	22/01/1968	02:39:41	**38,3567**	**0,1833**	40		3,1	**MEDITERRÁNEO-CABO DE PALOS**
6343	22/01/1968	07:19:08	**35,1367**	**-5,8333**	40	V	4,1	**SE TLETA RISSANA.MAC**
6344	22/01/1968	15:30:57	**38,3567**	**0,1833**			3,1	**MAR MEDITERRANEO**
6345	22/01/1968	15:43:03	**36,2800**	**-6,9950**	180		3,8	**GOLFO DE CÁDIZ**
6346	27/01/1968	18:27:57	**35,0000**	**5,0000**		VI	4,2	**RAS EL OUED.ARG**
6349	05/02/1968	05:40:32	**36,0417**	**-5,0400**	18		3,4	**ESTRECHO DE GIBRALTAR**
6350	06/02/1968	11:52:21	**37,6333**	**-4,5833**	5		4,0	**SW ESPEJO.CO**
6352	10/02/1968	21:54:52	**36,8533**	**-1,0550**	5		3,1	**MEDITERRÁNEO-CABO DE PALOS**

Figura 14: Datos de Seísmos Península Ibérica (muestra)

```
#Carga del mapa de Provincias
data.shape<-readOGR(dsn="/home//ll_provinciales_inspire_peninbal_etrs89.shp")

#Visualizar Mapa de Espana
plot(data.shape)
```

Extraemos la matriz de coordenadas y creamos el objeto de referencia de coordenadas. A continuación creamos los objetos de puntos espaciales y podemos hacer una representación el panel gráfico del entorno de R.

```
#Extraemos la Matriz de coordenadas
mat = cbind(datos$Longitud,datos$Latitud)

#Creamos el Sistema de Referencia
llcrs  = CRS("+proj=longlat +ellps=WGS84")

#Creamos el Objeto de Capa de Puntos
sp=SpatialPoints(mat,llcrs)

#Visualizamos los puntos
plot(data.shape,col="red")
plot(sp,add=TRUE,col="blue")
```

Procedemos ahora a calcular el valor óptimo de suavizado, aunque vamos a usar un valor diferente pues el que resulta $h = 0,0355$ es muy bajo y el

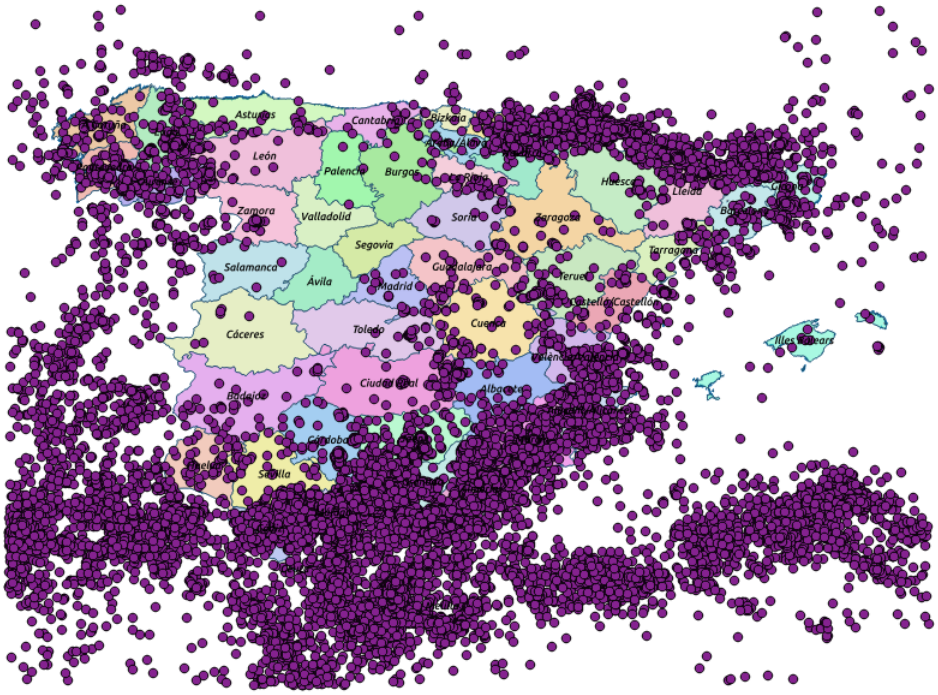

Figura 15: Seísmos Península Ibérica 1968-2017

diagrama sale muy picudo y con los datos de intensidad muy centrados en los puntos de valores máximos. Vamos a usar $h = 0,5000$, pues haciendo pruebas se obtiene un gráfico más razonable.

```
#Creamos el poligono de estudio (longitud y latitud)
poli  = as.points( list (x=c(-10,-10,5,5),y=c(45,34,34,45)))

#Calulamos los errores en funcion de h
suavizados = mse2d(as.points(mat),poli,300,0.15)

#Calculamos el error minimo para el valor h
suavizados$h[which.min(suavizados$mse)]
```

A continuación realizamos el cálculo de la densidad espacial y lo exportamos en una capa raster que acto seguido lo importamos en QGIS y le calculamos una capa vectorial de contornos para visualizar con más detalle las diferencias de intensidad.

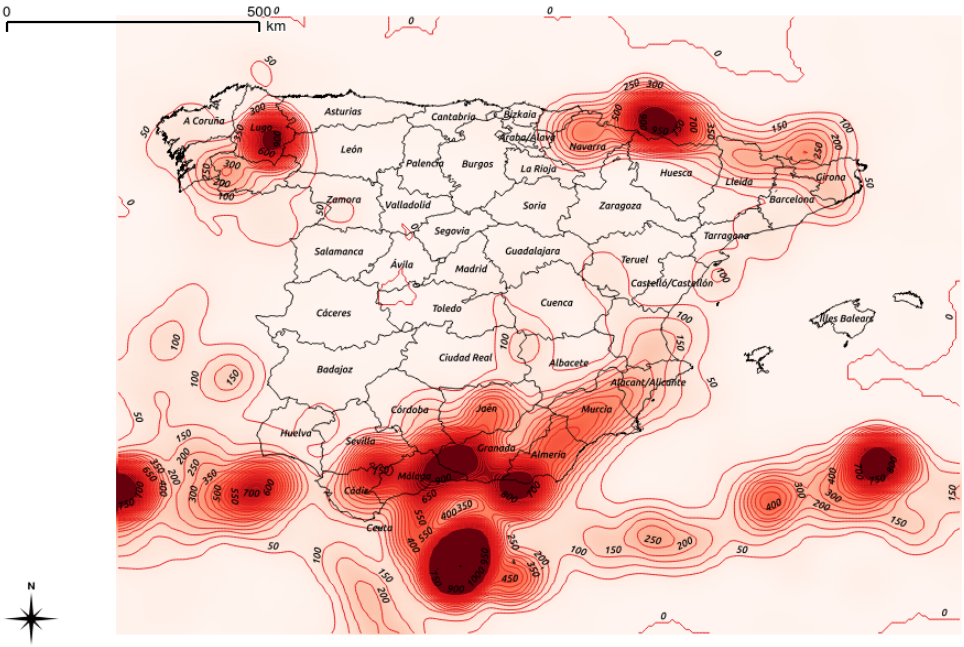

Figura 16: Intensidades Sísmicas Península Ibérica

```
#Calculamos intensidad x,y,z
x1=kernel2d(as.points(mat),poli,h0=0.5,nx=250,ny=250)$x
y1=kernel2d(as.points(mat),poli,h0=0.5,nx=250,ny=250)$y
z1=kernel2d(as.points(mat),poli,h0=0.5,nx=250,ny=250)$z

#Hacemos la Visualizacion de la intensidad
persp(x1,y1,z1)

#Creamos la capa raster
r = raster(t(z1),xmn=-10,xmx=5,ymn=34,ymx=45,crs=llcrs)

#Invertimos la componente "y"
ry=flip(r, direction="y")

#grabamos en el disco
writeRaster(ry, "test_output11", format = "GTiff",overwrite=TRUE)
```

Los gráficos resultantes una vez importados en QGIS aparecen en las figuras 16, 17 y 18

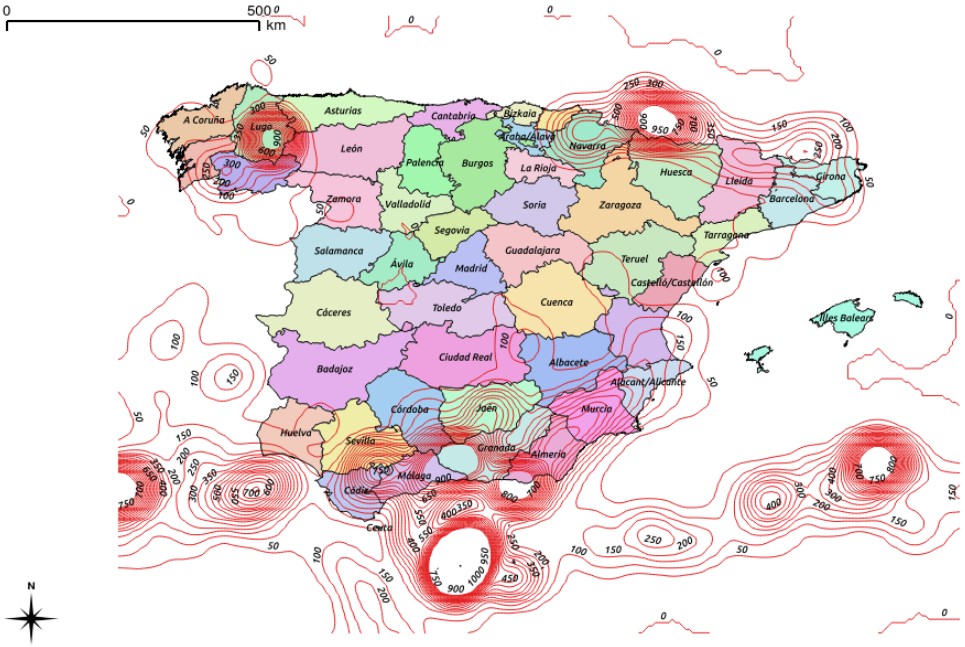

Figura 17: Isovalores Intensidades Sísmicas Península Ibérica

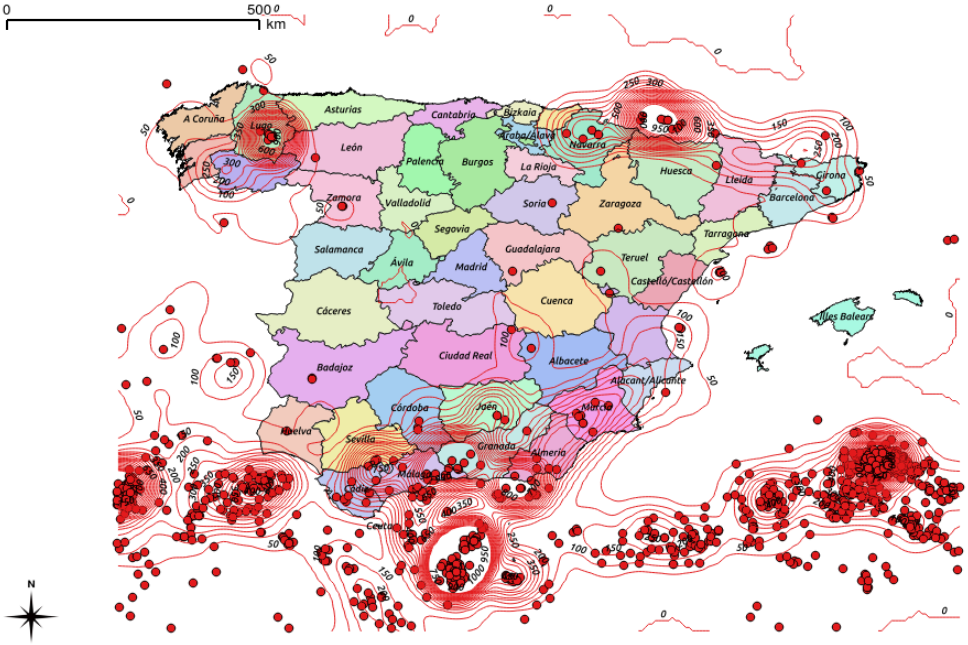

Figura 18: Terremotos Magnitud 4.0 Península Ibérica 1968-2017

A. El Estimador de Cressie y Hawkins No es Insegado.

Hemos llegado a la expresión

$$E(2\bar{\gamma}(\mathbf{h})) = \frac{1}{0,457 + 0,494/n}\frac{1}{n^4}E((\sum_{i=1}^{n}|Y_i|^{1/2})^4)$$

y supongamos que $E(2\bar{\gamma}(\mathbf{h})) = 2\gamma(\mathbf{h})$. Podemos desarrollar la potenciación del sumatorio en cuatro sumatorios,

$$E(2\bar{\gamma}(\mathbf{h})) = \frac{1}{0,457 + 0,494/n}\frac{1}{n^4}E(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}|Y_i|^{1/2}|Y_j|^{1/2}|Y_k|^{1/2}|Y_l|^{1/2})$$

donde se establece que cuando $i = j$ o $j = k$ ... las variables $|Y_i|^{1/2}$ y la variables $|Y_j|^{1/2}$, no son independientes pero a cambio podremos multiplicarlas y consideraremos la esperanza del producto. Podemos aplicar E a la suma y tendremos que,

$$E(2\bar{\gamma}(\mathbf{h})) = \frac{1}{0,457 + 0,494/n}\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}E(|Y_i|^{1/2}|Y_j|^{1/2}|Y_k|^{1/2}|Y_l|^{1/2})$$

Agrupando los sumandos de estos sumatorios, tenemos cinco tipos valores cada uno con una distinta multiplicidad que describimos a continuación

1. Todas la variables Y_i distintas en este caso el sumando vale

$$sum_1 = nV_3^{n-1}E(|Y|^{1/2})^4$$

donde $V_i^n = n!/(n-i)!$.

2. Una variable se repite y el resto no, entonces

$$sum_2 = n\binom{4}{2}V_2^{n-1}E(|Y|)(E(|Y|^{1/2}))^2$$

3. Dos variables se repiten

$$sum_3 = \binom{n}{2}\frac{4!}{2!2!}E(|Y|)^2$$

4. Tres variables se repiten

$$sum_4 = n\binom{4}{3}V_1^{n-1}E(|Y|^{3/2})E(|Y|^{1/2})$$

5. Las cuatro variables se repiten

$$sum_5 = n\binom{4}{4}V_0^{n-1}E(|Y|^2)$$

Se puede comprobar que no sobran ni faltan sumandos pues se cumple

$$n^4 = nV_3^{n-1} + n\binom{4}{2}V_2^{n-1} + n\binom{4}{3}V_1^{n-1} + n\binom{4}{4}V_0^{n-1} + \binom{n}{2}\frac{4!}{2!2!}$$

Podemos separar las partes de los sumandos que depende de n y las que dependen de la variable aleatoria multiplicando por $\frac{1}{0,457+0,494/n}\frac{1}{n^4}$ y tenemos

$$E(2\bar{\gamma}(\mathbf{h})) = \sum_{k=1}^{5} coef_k(n)X_k$$

donde

$$coef_k(n) = \frac{sum_k}{0,457 + 0,494/n}\frac{1}{n^4}$$

y

$$X_1 = E(|Y|^{1/2})^4$$
$$X_2 = E(|Y|)(E(|Y|^{1/2}))^2$$
$$X_3 = E(|Y|)^2$$
$$X_4 = E(|Y|^{3/2})E(|Y|^{1/2})$$
$$X_5 = E(|Y|^2)$$

Si suponemos que el estimador es insesgado se cumplirá

$$\sum_{k=1}^{5} coef_k(n)X_k = 2\gamma$$

y dividiendo por 2γ y haciendo el cambio $X_k/2\gamma = W_k$, tendremos para cada valor de n una ecuación que bajo el supuesto de "no sesgo" se ha de cumplir.

$$\sum_{k=1}^{5} coef_k(n)W_k = 1 \qquad \forall\ n > 4$$

Si ahora damos a n valores $n = 5, 6, 7, 8, 9, 10$ tendremos un sistema de seis ecuaciones con cinco incógnitas que debería ser compatible.

Introducimos estos datos en el ordenador para poder evaluar el sistema, lo que resulta la matriz

```
0.3454480   1.0363440   0.1727240   0.2302987   0.0143937   1.0000000
0.5150391   1.0300783   0.1287598   0.1716797   0.0085840   1.0000000
0.6631410   0.9947115   0.0994711   0.1326282   0.0055262   1.0000000
0.7906627   0.9487952   0.0790663   0.1054217   0.0037651   1.0000000
0.9004012   0.9004012   0.0643144   0.0857525   0.0026798   1.0000000
0.9952607   0.8530806   0.0533175   0.0710900   0.0019747   1.0000000
```

Que tiene un determinante de valor 0. Lo que en principio podría sugerir que el sistema es compatible. Pero he descartado poder hacer esta afirmación pues el sistema tiene un número de condición muy alto $cond = 2.3264e + 17$ y los datos no son fiables a nivel numérico.

B. Código Python para cálculo de Kriging no estocástico.

Se incluye a continuación un código Python que se ejecuta en el entorno de QGIS ,para la realización de un Kriging no estocástico por el método de la distancia ponderada de la inversa de la distancia al cuadrado, aplicándolo al problema de las precipitaciones de la Región de Murcia. El desarrollo de este código, como se ha indicado, tiene más un interés formativo que de valor en sí, pues QGIS da rutinas que lo realizan de forma óptima.

```
# ************************************************************
# Kriging Precipitaciones de la Region de Murcia
# ************************************************************

from qgis.core import *
from qgis.gui import *
from PyQt4.QtCore import *
import numpy as np
import gdal, ogr

name = 'Trabajado_Precipitaciones_Murcia_2017_Definitivo'
layer = QgsMapLayerRegistry.instance().mapLayersByName( name )[0]
qgis.utils.iface.setActiveLayer(layer)

if not layer.isValid():
    print "Layer no Valida"

# ******************************************************
# Lectura de los datos de las estaciones
# ******************************************************

NUM_X_PIXEL = 250
NUM_Y_PIXEL = 250

MIN_LON = -2.40
MAX_LON = -0.65

MAX_LAT = 38.83
MIN_LAT = 37.30

num_estaciones=0
estaciones_x = []
estaciones_y = []
precipitaciones = []
```

```
estaciones_x.append(0)
estaciones_y.append(0)
precipitaciones.append(0)

iter=layer.getFeatures()
for feature in iter:
    num_estaciones=num_estaciones+1
    geom=feature.geometry()
    if geom.type() == QGis.Point:

        x=geom.asPoint()

        estaciones_x.append(x.x())
        estaciones_y.append(x.y())

        idx = layer.fieldNameIndex('VAL2017')
        precipitaciones.append(feature.attributes()[idx])


# ********************************************************
# Cuadriculamos la region para hacer el kriging
# ********************************************************

s0_x = []
s0_y = []
s0_valor = []

for i in range(0,NUM_Y_PIXEL+1):
    s0_x.append([0]*(NUM_X_PIXEL+1))
    s0_y.append([0]*(NUM_X_PIXEL+1))
    s0_valor.append([0]*(NUM_X_PIXEL+1))

for i in range(0,NUM_X_PIXEL+1):
    for j in range(0,NUM_Y_PIXEL+1):
        s0_x[j][i]=MIN_LON+((MAX_LON-MIN_LON)/NUM_X_PIXEL)*(i)
        s0_y[j][i]=MAX_LAT-((MAX_LAT-MIN_LAT)/NUM_Y_PIXEL)*(j)

# ********************************************************
# Aplicamos el Kriging
# ********************************************************

#Creamos un Objeto de Distancia
distance = QgsDistanceArea()
distance.setEllipsoidalMode(True)
distance.setEllipsoid('WGS84')

for i in range(0,NUM_Y_PIXEL+1):
    print i
    for j in range(0,NUM_X_PIXEL+1):
```

```
            suma=0
            sumad=0;
            for t in range(1,num_estaciones+1):

                point1 = QgsPoint(s0_x[i][j],s0_y[i][j])
                point2 = QgsPoint(estaciones_x[t],estaciones_y[t])

                #Medida de la distancia
                m = distance.measureLine(point1, point2)

                suma=suma+precipitaciones[t]/(m**2)
                sumad=sumad+1.0/(m**2)

            s0_valor[i][j]=suma/sumad


# ******************************************************
# Generamos la capa raster
# ******************************************************

geotransform=[MIN_LON, (MAX_LON-MIN_LON)/NUM_X_PIXEL, 0,
    MAX_LAT, 0, -(MAX_LAT-MIN_LAT)/NUM_Y_PIXEL]

Raster = np.array(s0_valor)
driver = gdal.GetDriverByName('GTiff')
outputRaster =
    '/home/salvador/Estudios_Salvador/MaterTFM/raster_prog_precipitaciones.tif'
dst_ds = driver.Create(outputRaster, NUM_X_PIXEL+1, NUM_Y_PIXEL+1, 1,
    gdal.GDT_Float32)
dst_ds.SetGeoTransform(geotransform)
band = dst_ds.GetRasterBand(1)
band.WriteArray( Raster )
band.SetNoDataValue(-9999)

dst_ds = None

print "Calculo Terminado"
```

C. Código Python para cálculo de Kriging Robusto no estocástico.

Se incluye a continuación un código Python que se ejecuta en el entorno de QGIS ,para la realización de un Kriging no estocástico por el método de la Mediana Ponderada.

```
# ************************************************************
# Kriging Robusto Precipitaciones de la Region de Murcia
# ************************************************************

from qgis.core import *
from qgis.gui import *
from PyQt4.QtCore import *
import numpy as np
import gdal, ogr

name = 'Trabajado_Precipitaciones_Murcia_2017_Definitivo'
layer = QgsMapLayerRegistry.instance().mapLayersByName( name )[0]
qgis.utils.iface.setActiveLayer(layer)

if not layer.isValid():
    print "Layer no Valida"

# ******************************************************
# Lectura de los datos de las estaciones
# ******************************************************

NUM_X_PIXEL = 250
NUM_Y_PIXEL = 250

MIN_LON = -2.40
MAX_LON = -0.65

MAX_LAT = 38.83
MIN_LAT = 37.30

num_estaciones=0
estaciones_x = []
estaciones_y = []
precipitaciones = []
ordenacion = []
distancias = []

estaciones_x.append(0)
```

```
estaciones_y.append(0)
precipitaciones.append(0)
ordenacion.append(0)
distancias.append(0)

iter=layer.getFeatures()
for feature in iter:
    num_estaciones=num_estaciones+1
    ordenacion.append(num_estaciones)
    geom=feature.geometry()
    if geom.type() == QGis.Point:

        x=geom.asPoint()

        estaciones_x.append(x.x())
        estaciones_y.append(x.y())

        idx = layer.fieldNameIndex('VAL2017')
        precipitaciones.append(feature.attributes()[idx])

# ********************************************************
# Cuadriculamos la region para hacer el kriging
# ********************************************************

s0_x = []
s0_y = []
s0_valor = []

for i in range(0,NUM_Y_PIXEL+1):
    s0_x.append([0]*(NUM_X_PIXEL+1))
    s0_y.append([0]*(NUM_X_PIXEL+1))
    s0_valor.append([0]*(NUM_X_PIXEL+1))

for i in range(0,NUM_X_PIXEL+1):
    for j in range(0,NUM_Y_PIXEL+1):
        s0_x[j][i]=MIN_LON+((MAX_LON-MIN_LON)/NUM_X_PIXEL)*(i)
        s0_y[j][i]=MAX_LAT-((MAX_LAT-MIN_LAT)/NUM_Y_PIXEL)*(j)

# ************************************************************
# Aplicamos el Kriging
# ************************************************************

#Creamos un Objeto de Distancia
distance = QgsDistanceArea()
distance.setEllipsoidalMode(True)
distance.setEllipsoid('WGS84')

for i in range(0,NUM_Y_PIXEL+1):
    print i
```

```
for j in range(0,NUM_X_PIXEL+1):
    sumad=0;
    distancias = []
    distancias.append(0)

    for t in range(1,num_estaciones+1):

        point1 = QgsPoint(s0_x[i][j],s0_y[i][j])
        point2 = QgsPoint(estaciones_x[t],estaciones_y[t])

        #Medida de la distancia
        m = distance.measureLine(point1, point2)/1000.0

        distancias.append(1.0/(m**2))
        sumad=sumad+1.0/(m**2)

    for t in range(1,num_estaciones+1):
        distancias[t]=distancias[t]/sumad


    # ******************************************
    # Ordenamos las precipitaciones
    # ******************************************

    valores=[]
    valores.append(0)
    for s in range(1,num_estaciones+1):
        valores.append(precipitaciones[s]*distancias[s])

    for s in range(1,num_estaciones+1):
        ordenacion[s]=s

    for s in range(1,num_estaciones+1):
        for k in range(1,num_estaciones):

            if (valores[k]>valores[k+1]):

                orden = ordenacion[k+1]
                valor = valores[k+1]

                ordenacion[k+1]=ordenacion[k]
                valores[k+1] = valores[k]

                ordenacion[k]=orden
                valores[k]=valor

    # ************************************************
    # Calculamos el punto medio
    # ************************************************
```

```
        medio=0;
        l1=0;
        l2=0;
        m1=0;
        m2=0;
        for s in range(1,num_estaciones+1):

            if ((medio+distancias[ordenacion[s]])==0.5):
                l1=s
                l2=s
                m1=medio+distancias[ordenacion[s]]
                m2=medio+distancias[ordenacion[s]];
                break

            if ((medio+distancias[ordenacion[s]])>0.5):
                l1=s-1
                l2=s
                m1=medio
                m2=medio+distancias[ordenacion[s]];
                break

            medio=medio+distancias[ordenacion[s]]


        s0_valor[i][j]=(m1*distancias[ordenacion[l1]]*precipitaciones[ordenacion[l1]]+m2*distancias[ordenacion[l2]]


# ******************************************************
# Generamos la capa raster
# ******************************************************

geotransform=[MIN_LON, (MAX_LON-MIN_LON)/NUM_X_PIXEL, 0,
    MAX_LAT, 0, -(MAX_LAT-MIN_LAT)/NUM_Y_PIXEL]

Raster = np.array(s0_valor)
driver = gdal.GetDriverByName('GTiff')
outputRaster =
    '/home/salvador/Estudios_Salvador/MaterTFM/raster_robusto_prog_precipitaciones.tif'
dst_ds = driver.Create(outputRaster, NUM_X_PIXEL+1, NUM_Y_PIXEL+1, 1,
    gdal.GDT_Float32)
dst_ds.SetGeoTransform(geotransform)
band = dst_ds.GetRasterBand(1)
band.WriteArray( Raster )
band.SetNoDataValue(-9999)

dst_ds = None

print "Calculo Terminado"
```

Referencias

[1] Beazley, David M. "Python, Essential Reference", *Addison Wesley*, Agosto 2015.

[2] Bivand Roger S., Pebesma Edzer , Gómez-Rubio Virgilio *Applied Spatial Data Analysis with R.*, Springer. 2013.

[3] Cabrero Ortega Yolanda y García Pérez Alfonso "Análisis Estadístico de Datos Espaciales con QGIS y R", *Editorial UNED*, Octubre 2015.

[4] Cressie Noel A.C. "Statistics for Spatial Data", *John Wiley and Sons, Inc.* 2015.

[5] Dalgaard Peter "Introductory Statistics with R", *Springer*, 2008.

[6] García Perez Alfonso "M.A.E.A. Métodos Robustos y de Remuestreo", *Editorial UNED*, Septiembre 2014.

[7] Jones Owen, Maillardet Robert and Robinson Andrew , "Introduction to Scientific Programing and Simulation Using R", *CRC Press*, 2009.

[8] Matheron G., "Traité de Geostatistique Appliquée, Tome I. Memoiries du Bureau de Recherches Geologiques et Minieres, No.14", *Editions Technip*. Paris 1962.

[9] Vélez Ibarrola Ricardo y García Pérez Alfonso. "Principios de Inferencia Estadística", *Editorial UNED*, 1993.

[10] Wood Simon N. "Generalized Additive Models. An Introduction with R", *CRC Press*, 2017.