PROJECTE FINAL DE CARRERA

DIPLOMATURA D'ESTADÍSTICA

Comparació de mètodes d'aplicació de processos espaials amb distribució lognormal

Alumnes: Clara Foz Altarriba
Bibiana Prat Pubill

Directora: Vera Pawlowsky Glahn
Departament: Matemàtica Aplicada III
Data d'entrega: 4 de juliol del 2.000

Facultat de Matemàtiques
i Estadística
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Índex

Capítol 1: Introducció

El projecte està estructurat de manera que en un principi s'introdueixen una sèrie d'explicacions de les eines geoestadístiques i, en concret, es remarquen les que posteriorment s'usaran en el desenvolupament de l'estudi.
A continuació i d'una forma esquemàtica s'explica el procediment de desenvolupament i treball del projecte que es detalla al capítol 5.
Finalment, en els capítols 6 i 7 s'analitzen els resultats i s'extreuen les conclusions.

L'objectiu del projecte és comparar dos mètodes d'estimació de dades amb dependència espaial en el cas de dades amb distribució experimental asimètrica. El primer mètode s'anomena 'krigeat lognormal', el segon 'krigeat indicador'.

El 'krigeat lognormal' parteix de la hipòtesis que les dades són realitzacions d'un procés espaial lognormal i consisteix en aplicar la transformació logarítmica a les

dades, procedir a una estimació habitual sota la hipòtesis de procés gaussià i tornar després a l'espai mostral original mitjançant la transformació adequada, que serà $\exp(\mu+1/2\sigma^2)$. També és possible obtenir intervals de predicció seguint el mateix esquema.

El 'krigeat indicador' es basa en una transformació no lineal, utilitzant la funció indicatriu I(x,t), de les dades de manera que fixant un valor t:

$$I(x,t)=\begin{cases}0,\text{si } Z(x)>t;\\ 1,\text{si } Z(x)\le t;\end{cases}$$

on Z(x) és la funció aleatòria. L'estimació per diferents valors de t permet aproximar la funció de densitat de la variable aleatòria Z(x) en cada punt de l'espai físic.

El programa de domini públic GSLIB (Geostatistical Software Library) permet simular dades amb densitat normal (0,1) i una dependència espaial donada emprant el mètode anomenat simulació gaussiana. Aplicant la transformació exponencial s'obté una realització d'una funció aleatòria lognormal (0,1) i amb aquesta realització es pot fer l'estudi de comparació dels dos mètodes per un cas particular. A més, amb aquest programa ha estat amb el que s'han realitzat les estimacions.

Altres programes usats han estat els següents:

- MINITAB, per la seva facilitat a l'hora de manipular dades i la seva sortida gràfica.
- Geo-Eas per obtenir el semivariograma experimental
- Excel per a modelar el semivariograma
- Compilador de Fortran per a l'execució del GSLIB i pel càlcul de l'stress.

Totes aquestes eines es complementen i ens han permès treballar d'una forma més eficient.

Capítol 2: Nocions de geoestadística

S'ha cregut necessari fer un incís a la geoestadística per tal d'introduir al lector en aquest camp.

A més, la geoestadística té unes metodologies de treball, suposicions estadístiques, vocabulari, etc. molt concrets i que, en certs aspectes, difereixen dels de l'estadística clàssica que és amb la que estem acostumats a treballar.

És per això que s'ha volgut aclarir tot això abans de començar a explicar el que s'ha fet en aquest projecte.

2.1 El semisemivariograma

El semivariograma és un estadístic que, al contrari de la covariància, evalua com decreix en promig la semblança entre dues variables aleatòries a mesura que la distància entre elles augmenta.

En estadística clàssica definim la covariància teòrica entre dues variables Y1, Y2 com:

$$\mathrm{COV}(Y1, Y2) = \mathrm{E}[(Y1 - \mu 1)(Y2 - \mu 2)]$$

on Y1 i Y2 no depenen de la seva ubicació espaial.

En estadística espaial redefinim les variables Y1 i Y2. Usem Z(x) en lloc de Y1 i Z(x+h) en lloc de Y2. On Z(x) és la variable regionalitzada i h és el vector distància.

Obtenim llavors:

$$\mathrm{COV}[Z(x), Z(x+h)] = \mathrm{E}[(Z(x) - \mathrm{E}[Z(x)])(Z(x+h) - \mathrm{E}[Z(x+h)])]$$

que és la funció d'autocovariància.

Ara ja podem definir el semivariograma teòric i ho fem com la semivariància dels increments:

$$\gamma[Z(x), Z(x+h)] = \frac{1}{2}\mathrm{Var}[Z(x) - Z(x+h)] = \frac{1}{2}\mathrm{E}[((Z(x) - Z(x+h)) - \mathrm{E}[(Z(x) - Z(x+h))])^2]$$

Per tal de poder ajustar models a aquesta funció teòrica hem d'assumir que Z(x) és estacionària de segon ordre:

1.1 El valor esperat de Z(x) existeix i no depèn de x:

$$\mathrm{E}[Z(x)] = \mu \quad \forall x$$

1.2 La covariància existeix i només depén de la distància h:

$$\mathrm{E}[(Z(x) - \mu)(Z(x+h) - \mu)] = C(h)$$

En geoestadística, aquesta condició es debilita introduint la hipòtesi intrínseca que és menys restrictiva:

1.1 El valor esperat de Z(x) existeix i no depèn de x:

$$E[Z(x)] = \mu \quad \forall x$$

1.2 Per cada vector h, l'increment (Z(x)-Z(x+h)) té variància finita i no depèn d'x:

$$\frac{1}{2}\mathrm{Var}[Z(x) - Z(x+h)] = \frac{1}{2}E[(Z(x) - Z(x+h))^2] = \gamma(h) \quad \forall x$$

La forma més senzilla de poder mostrar un semivariograma és fent un gràfic que mostri la distància entre el parell de mostres (h) en l'eix horitzontal i el valor del semivariograma en l'eix vertical. Per definició h comença en el valor 0, encara que òbviament és impossible agafar dues mostres diferents separades per distància 0.

Per fer-nos una idea del semivariograma a nivell pràctic, considerem el cas en que h=0; agafem dues mostres situades exactament a la mateixa posició i mesurem els seus valors. Assumirem que la diferència és zero per definició, encara que si les dues mostres estan separades per una distància h que tendeix a zero, però que exactament no ho és, podem trobar-hi petites diferències, això causaria una discontinuïtat a l'origen del semivariograma. El salt que hi pot haver en l'eix vertical des del valor 0 fins a l'origen del valor del semivariograma en distancies de separació extremament petites s'anomena l'efecte pepita i s'haurà de tenir en compte a l'hora d'ajustar un model apropiat de semivariograma.

Ara suposem que agafem dues mostres una mica separades entre sí; si tornem a comparar els dos valors veurem que hi ha una petita diferència i, per tant, el semivariograma prendrà un valor positiu petit. A mesura que les mostres s'agafen a més distància entre si, cada cop seran més diferents i, per tant, cada cop el semivariograma prendrà valors més elevats. En el cas teòric ideal, tindríem que quan la distància és suficientment gran les mostres són independents entre si i, per tant, el valor del semisemivariograma a partir d'aquest punt serà més o menys constant.

La distància a partir de la qual les mostres esdevenen independents les unes de les altres es denota per *a* i s'anomena el rang d'influència de la mostra o abast. El valor de γ al que arriba el gràfic és anomenat la meseta.

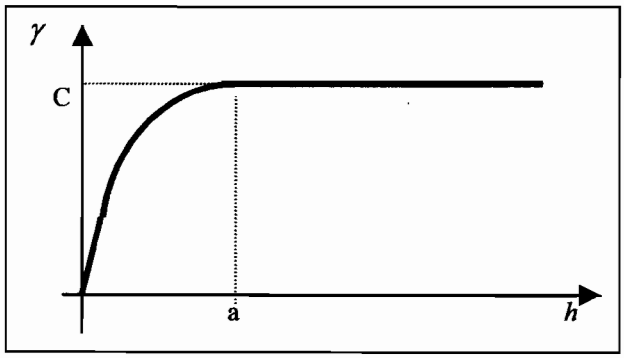


Figura 2.1.1 La forma ideal per un semivariograma

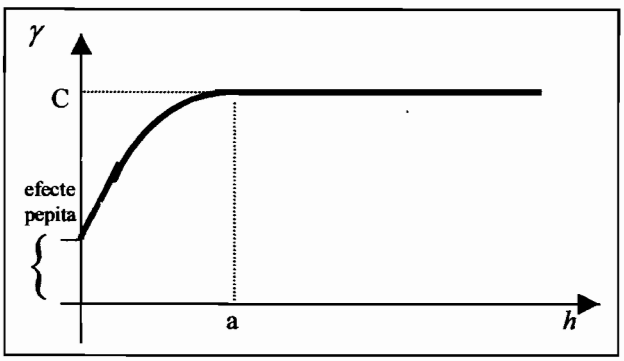


Figura 2.1.2 La forma ideal per un semivariograma amb efecte pepita

El semivariograma descriu la relació de dependència espaial mitjançant la variància dels increments de la mateixa manera que les densitats descriuen la dispersió de la probabilitat sobre el domini de definició d'una variable aleatòria. Igual que aquestes últimes, han de satisfer una sèrie de situacions, motiu per el qual s'han estudiat models concrets que les compleixin.

Hi ha molts models de semivariograma, però només uns quants són usats habitualment. Els més comuns són el model esfèric i l'exponencial.

De la mateixa forma que les funcions de densitat, hi ha dues vies de modelització:

- Raonament teòric i ajustament de paràmetres a partir de la mostra.
- Ajustament a un semivariograma experimental mitjançant tècniques més o menys automàtiques.

A la pràctica s'usa més aquest segon mètode amb un ajustament 'a sentiment'.

Com que en general a priori no podem conèixer el semivariograma que seguiran les dades, ens haurem de basar en un estimador d'aquest que anomenarem semivariograma experimental ($\gamma^*(h)$) i es defineix a partir del semivariograma teòric anteriorment citat de la següent manera:

$$\gamma^*(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left(z(x_i) - z(x_i + h)\right)^2$$

on N(h) és el nombre de parells de punts mostrals separats una distància h.

Definim el semivariograma experimental sota la hipòtesi intrínseca, i això implica que la mitjana ha de ser constant. La variació d'aquesta mitjana és el que s'anomena deriva. Conseqüentment, un semivariograma experimental apropiat requereix l'eliminació d'aquesta deriva.

Amb el semivariograma experimental descrivim l'estructura espaial d'una variable. Aquesta estructura ens farà falta a l'hora de fer les estimacions a partir del krigeat, per tant haurem d'ajustar un model.
Els models de semivariograma venen definits per dos paràmetres abast (a) i meseta (C) (Figura 2.1.1) a vegades també s'ha de tenir en compte l'efecte pepita explicat anteriorment (Figura 2.1.2).

L'abast és el valor d'h en que el semivariograma assoleix un valor constant anomenat meseta. Així s'assumeix que dos punts separats per una distància més gran que l'abast no s'influeixen mútuament.

Els models de semivariograma que usarem són el següents.

1. Model esfèric : Creix linealment fins a trobar, a una distància 'a', la meseta C ($\approx\sigma^2$).

$$\gamma(h)=\begin{cases} C\left[1.5\frac{|h|}{a}-0.5\left(\frac{|h|}{a}\right)^3\right] & \text{si } |h|\le a \\ C & \text{si } |h|>a \end{cases}$$

Un cas particular d'aquest model es produeix quan el rang és igual a zero. S'anomena efecte pepita i la notació és C_0.

2. Model exponencial:

$$\gamma(h)=C\left[1-\exp\left(-\frac{|h|}{a}\right)\right]$$

2.2 El krigeat, introducció al mètode

El krigeat és una tècnica d'estimació espaial que permet conèixer el valor d'una determinada variable regionalitzada en punts de la zona no mostrejats a partir dels punts coneguts.

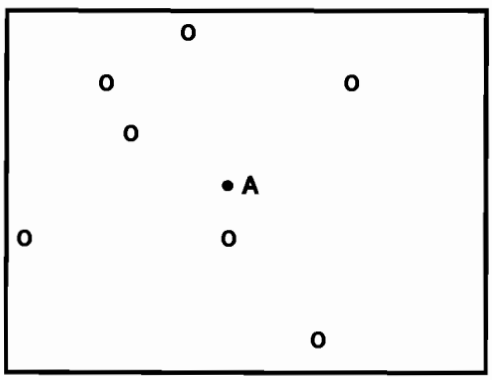


Figura 2.2.1: Mostreig hipotètic i punt a estimar (A)

Que l'estimació que realitzem sia més o menys bona depèn de diferents factors:

- El primer és la grandària de la mostra escollida. Com més gran sigui aquesta, més serà la informació que tinguem de la variable en estudi i, conseqüentment, la variància disminueix.
- Que la qualitat de les dades no sigui la mateixa en una mostra que en una altra. Per tant, no tindrem la mateixa informació en tots els punts mostrejats. Això és solventable donant pesos diferents a les mostres en el moment de fer l'estimació.
- Que les mostres no hagin estat recollides d'una forma regular dins de la regió i per tant, hi haurà zones de les quals no tindrem informació i d'altres que en tindrem massa. Amb les tècniques anomenades de *disgregació* podem procedir a un buidatge automàtic de certes zones.
- S'ha de tenir també en compte que els punts propers entre sí s'influencien, mentre que no esperarem que ho facin els punts que estan allunyats.
- No és el mateix estimar una variable regionalitzada que té poca variabilitat espaial, que una que sigui totalment canviant i amb salts bruscos.

Tots aquests factors, que pot ser que es donin individualment, o en el pitjor dels casos, tots alhora, els té en conte el mètode del krigeat.

Els mètodes clàssics d'estadística es basen en la independència entre mostres per fer l'estimació d'un punt. Aquesta suposició implica que els pesos a donar a cada punt de la mostra seran els mateixos, és a dir, tenim equiprobabilitat de succés.
En aquest cas l'estimació de la variància es realitza fent la suma quadràtica de les diferències entre el valor estimat en un punt determinat (A), és a dir, la mitjana, i el valor real en cada un dels punts mostrejats

$$s^2 = \frac{\sum_{i=1}^{n} (\overline{X} - x_i)^2}{n}$$

Com hem dit prèviament, un punt és influenciat per altres punts propers mentre que difícilment ho serà per un que està allunyat. Aquest fet ens fa plantejar ja d'entrada una estimació basada en un sistema de pesos:

$$\hat{v} = \sum_{i=1}^{n} w_i v_i$$

on $v_1,\ldots, v_n$ són els valors de les variables i w_i els pesos assignats a cada punt.

Per tant, per estimar la variància ho farem a partir d'una combinació lineal ponderada:

$$Var\left\{\sum_{i=1}^{n} w_i V_i\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j Cov\{V_i V_j\}$$

El que es pretén ara és minimitzar l'estimació de la variància respecte els pesos (w_i), és a dir, trobar els pesos que minimitzen l'error quadràtic mig.
Això s'aconsegueix igualant a zero la primera derivada de l'error respecte cada un dels pesos desconeguts:

$$\frac{\partial \sigma_\varepsilon^2}{\partial w_i} = 0,$$
$$i = 1,2,3,..n$$

El producte resultant és un sistema d'n equacions, la solució del qual ens proporciona els pesos requerits per dur a terme el krigeat en localitzacions puntuals. El krigeat requereix la solució de tants sistemes d'equacions com localitzacions a estimar.

Hi ha moltes formes de realitzar el krigeat, totes elles inicialment formulades per la estimació d'un atribut espaial continu en punts geogràfics no mostrejats, preferiblement dins de l'espai definit per la localització de les dades.

Passarem a descriure a continuació els tipus de krigeat que utilitzarem en el desenvolupament del projecte. Cal tenir en compte que el krigeat (tant el simple com l'ordinari) han estat construïts per obtenir el millor estimador lineal no esbiaixat (B.L.U.E.) per valors no mostrejats.

2.2.1 El krigeat lognormal

El krigeat lognormal és equivalent al krigeat simple o ordinari (que més endavant detallem) dels logaritmes de les dades.
Considerem la logtransformació y(u)=ln(z(u)), on z(u) és la variable original estrictament positiva. El krigeat simple o ordinari de les dades logarítmiques dóna una estimació y*(u) per ln(z(u)). Malauradament, quan volem tornar a les nostres dades originals, si apliquem la transformació exp(y*(u)), el que aconseguim és un estimador esbiaixat de Z(u). La transformació no esbiaixada és:

$$z^*(u) = \exp [\, y^*(u) + \sigma^2(u) / 2]$$

on $\sigma^2(u)$ és la variància lognormal del mètode de krigeat usat.

Per tant, usem una transformació basada en la correcció per la desviació tipus.
A continuació passem a detallar els mètodes de krigeat més usats quan es vol fer una estimació per krigeat lognormal.

2.2.1.1 El krigeat simple

La formulació i l'aplicació del krigeat simple basat en un model de funció aleatòria, requereix alguns requisits:

Requisit 1: s'assumeix que la mostra és una realització parcial d'una funció aleatòria Z(x) on x denota la localització espaial.

Requisit 2: La funció aleatòria ha de ser estacionària de segon ordre, cosa que suposa que els moments implicats siguin insensibles a qualsevol translació en l'espai i depenen només de la distància Euclidiana en l'espai físic de mostreig.

$$E[Z(x)] = m$$

$$E[(Z(x)-m)(Z(x+h)-m)] = E[Z(x)Z(x+h)] - m^2 = Cov(x, x+h) = Cov(h)$$

on E(.) denota el valor esperat

m és un escalar constant que representa la mitjana

h és la distància vectorial en l'espai mostrejat.

Cov(.) és la covariància de la funció aleatòria.

Requisit 3: El krigeat simple requereix que la mitjana de la variable regionalitzada en estudi sigui coneguda.

Un cop s'han tingut en compte aquests requisits, podem passar a la formulació de l'estimador.

Sent Z una funció aleatòria estacionària de segon ordre amb mitjana m, l'estimació $Z^*_{sk}(x_0)$ en la posició x_0 ve donada per la següent combinació lineal de variables aleatòries a les posicions x_i considerades en la mostra:

$$Z^*_{sk}(x_0) = m + \sum_{i=1}^{k} w_i(Z(x_i) - m)$$

El propòsit principal del krigeat simple és trobar un conjunt de pesos per a la estimació en l'equació anterior de forma que es minimitzi l'error quadràtic mig.

La variància d'aquest estimador ve donada per:

$$\sigma^2_{SK}(u) = C(0) - \sum_{i=1}^{n} w_i(u)C(u - u_0) \geq 0$$

2.2.1.2 El krigeat ordinari

El krigeat ordinari va sovint acompanyat del qualificatiu de B.L.U.E. És "linear" (lineal) perquè les seves estimacions són combinacions lineals ponderades de les dades existents. És "unbiased" (no esbiaixat) perquè intenta que la mitjana residual (m_r) sia igual a zero. I és "best" (millor) perquè fa la variància dels errors (σ^2_R) mínima.

L'avantatge que presenta el krigeat ordinari davant d'altres mètodes d'estimació és precisament que fa mínima la variància residual, ja que n'hi ha que també són lineals i no esbiaixats.

Com hem dit anteriorment, el krigeat simple requereix el coneixement de la mitjana per resoldre el problema de trobar els pesos que minimitzin la variància de l'error d'estimació. Si la mitjana no és coneguda serà convenient utilitzar el krigeat ordinari, l'estimació del qual és independent de la mitjana, ja que s'afegeix la restricció de que els pesos han de sumar 1, d'aquesta manera, els pesos de la mitjana a l'equació (2.2.2.1) sumen 0 i l'estimació depèn només de la mostra.

$$Z^*(u) = \sum_{i=1}^{n} w_i(u) Z(u_i) + \left(1 - \sum_{i=1}^{n} w_i(u)\right) m \qquad (2.2.2.1)$$

Així doncs, podem afirmar que el krigeat ordinari no és res més que un krigeat simple que compleix la condició que els pesos tenen una suma unitària.

Considerem una funció aleatòria Z estacionària de segon ordre. L'estimador $Z^*_{ok}(x_0)$ en la posició x_0 ve donada per les següents combinacions lineals de les variables aleatòries a les posicions x_i considerades en la mostra:

$$Z^*_{ok}(x_0) = \sum_{i=1}^{k} w_i Z(x_i)$$

$$\text{subjecte a} : \sum_{i=1}^{k} w_i = 1$$

Les principals propietats del krigeat ordinari són:

- Minimitza l'error quadràtic mig.
- Interpolació exacte amb variància de krigeat 0.
- Independència de la translació dels nodes de referència.
- Dependència del patró de la mostra.
- Independència de la variància de krigeat per les observacions individuals.

2.2.3 El krigeat indicador

El krigeat indicador consisteix en fer una transformació no lineal de les dades per aproximar numèricament la distribució acumulada.
El krigeat indicador té sentit usar-lo quan les dades són qualitatives o bé, quan es vol fer l'estimació de la funció de distribució local d'una variable regionalitzada; aquest és el nostre cas i per tant en el que ens centrem,

Donada una funció aleatòria Z(x), l'indicador I(x,t) és la transformació binària

$$I(x,t) = \begin{cases} 0, & \text{si } Z(x) > t \\ 1, & \text{si } Z(x) \leq t \end{cases}$$

on t és el punt de tall o cut-off

Els punts de tall els definim nosaltres i seran, per tant, aquells que ens interessin. Aquesta transformació té la propietat que el valor esperat és igual a la probabilitat acumulada de la variable.

Per qualsevol variable de la funció aleatòria Z(x), la distribució acumulada és igual al valor esperat de l'indicador de la variable

$$F(t)_x = E[I(x,t)]$$

Llavors, el mapejat de les probabilitats acumulades de cada punt de tall es redueix a mapejar el valor esperat dels indicadors.

El procediment a seguir per aproximar la distribució acumulada d'una funció aleatòria per krigeat indicador és la següent:

1. Escollir els punts de tall i generar tants conjunts de dades com punts de tall tinguem. Si fem molts punts de tall llavors haurem de repetir moltes vegades el mateix procediment i si n'agafem pocs no obtindrem resultats fiables. Per tant s'ha de ser molt precís a l'hora de triar-ne el nombre.

2. Modelar el semivariograma per cada punt de tall.

3. Fer el krigeat ordinari de cada conjunt de dades indicador separadament, el que genera tantes xarxes com cut-off.

4. Aproximació numèrica de la funció de distribució acumulada.

Un dels problemes que presenta el krigeat indicador en el moment d'usar-lo en la realitat, és que per produir el mateix tipus de resultats, l'esforç demanat pel krigeat indicador és més gran que el que cal per fer un aproximament basat en la normalitat dels errors, quan aquest darrer és aplicable; ja que hem de fer un model de semivariograma i una estimació per krigeat per cada punt de tall. Computacionalment, el GSLIB té solventat aquest darrer inconvenient ja que permet fer el krigeat introduint tots els models de semivariograma alhora. Però els semivariogrames s'han de fer un a un.

Un dels avantatges que té el krigeat indicador és que permet l'extracció d'informació en dades poc precises. Per exemple, suposem que en un lloc determinat el valor de la variable en estudi no és exactament conegut, però tenim motius per pensar que és menor de 10 unitats. Aquest tipus d'informació, que amb el krigeat ordinari no es pot treballar del tot, de vegades resulta que donant valors indicadors i per tant usant l'estimació per krigeat indicador podem adquirir coneixements d'aquesta variable.

2.3 Simulació seqüencial gaussiana.

La simulació seqüencial gaussiana, es basa en la simulació de dades a partir d'una variable continua de distribució normal que, a més, ha de ser estandarditzada. En geoestadística s'usa la simulació per tal de poder conèixer el valors de la variable en estudi en punts no mostrejats. En el nostre cas l'hem usat per tal de poder tenir dades a partir de les quals fer l'estudi. Per tant només detallarem la seva mecànica:

1. Determinar la funció de distribució $F_Z(z)$, que representa tota l'àrea d'estudi i no només els punts mostrejats.
2. Transformar $F_Z(z)$ en una distribució Normal (0,1).
3. Comprovar que realment la nova variable (Y) es distribueix normalment. En el cas que no s'hagi pogut conservar el model gaussià, considerar models alternatius com ara una barreja de poblacions gaussianes o bé plantejar-se fer un altre tipus de simulació estocàstica.
4. A partir d'aquí ja podem usar mètodes computacionals que actuen de la següent manera:
   4.1 Es crea una 'ruta' que consisteix en la visita una sola vegada de cada node (u) de la xarxa. De cada node visitat es reté un nombre màxim de nodes veïns per tal de poder fer simulació condicionada.
   4.2 Escollir un tipus de krigeat i un tipus de semivariograma per tal de poder determinar la mitjana i la variància de la funció de distribució condicionada Y(u) en la localització u.
   4.3 Treure un valor simulat $y^{(l)}(u)$ de la funció de distribució condicionada.
   4.4 Afegir el valor simulat $y^{(l)}(u)$ al conjunt de dades.
5. Continuar amb el següent node fins que es simuli el darrer.

Capítol 3: Mètodes de comparació

De la mateixa manera que s'ha cregut oportú fer un incís en la geoestadística d'una forma general, s'ha optat per fer una explicació de les tècniques estadístiques que s'usaran al llarg d'aquest estudi.

De fet només s'explicarà l'stress, ja que les altres tècniques són prou conegudes i tant sols s'anomenaran.

Cal tenir en compte a l'hora d'aplicar mètodes d'estadística clàssica que la majoria d'ells assumeixen que les dades són independents entre sí, hipòtesis que al treballar amb dades distribuïdes en l'espai deixa de ser certa [ref. 2] és per això que utilitzarem l'ATREC com a estadístic determinant, ja que en la seva computació prescindeix d'aquesta hipòtesis.

3.1 STRESS

Per tal de poder dur a terme la comparació entre els dos mètodes d'estimació i les dades originals (simulades), i per tal de poder determinar quina de les dues estimacions s'aproxima més a la realitat, ens basarem en un estadístic anomenat STRESS que mesura el grau d'allunyament entre parells de variables. Passem a detallar-ho a continuació .

Suposem que tenim un conjunt X d'n dades (X=1, ..., n) ; per cada parell de dades (r,s) obtenim una mesura de no similaritat o diferència (δ_{rs}).
Si representem aquests punts en l'espai podem calcular les distàncies entre cada parell (d_{rs}).

L'objectiu d'aquest mètode és trobar una configuració en l'espai d'n punts de manera que les dues mesures anteriors (δ_{rs} i d_{rs}) siguin el més semblants possibles per a tot parell de punts (r, s).

Aquesta explicació la reduirem a conjunts de dades que pertanyen a l'espai d'R^2 i a distàncies euclidianes ja que és el cas que ens ocupa.

- Suposem que el punt r-èssim d'X té les coordenades $x_r = (x_{r1}, x_{r2})^t$
- Suposem que la distància entre els punts r i s és euclidiana

$$d_{rs} = \left[\sum_i (x_{ri} - x_{si})^2 \right]^{1/2}$$

- Definim les diferències $\{\hat{d}_{rs}\}$ com una funció de les distàncies de $\{d_{rs}\}$:

$$\hat{d}_{rs} = f(d_{rs})$$

on f és una funció monòtona de manera que $\hat{d}_{rs} \leq \hat{d}_{tu}$ sempre que $\delta_{rs} \leq \delta_{tu}$

Ara podem definir la funció de pèrdua L

$$L = \left\{ \frac{\sum_{r,s} (d_{r,s} - \hat{d}_{r,s})^2}{\sum_{r,s} d_{r,s}^2} \right\}^{1/2}$$

Aquesta funció de pèrdua és l'usada més habitualment. L'objectiu és trobar una configuració que provoqui una pèrdua mínima.

Podem redefinir la funció de pèrdua L de la següent manera:

$$S = \sqrt{\frac{S^*}{T^*}} \in [0,1]$$

Si definim $\{\hat{d}_{rs}\}$ com la regressió quadràtica de $\{d_{rs}\}$ a $\{\delta_{rs}\}$, llavors S s'anomena l'STRESS de la configuració; S* l'STRESS per fila. El numerador T de la fórmula, en el cas de l'STRESS, s'usa com a factor normalitzador, i és per això que l'STRESS no té dimensió.

Aquesta mesura d'STRESS és la que hem aplicat amb la finalitat de veure fins a quin punt les dues estimacions que s'han realitzat s'assemblen a la realitat. Obtindrem com a resultat un nombre comprés entre 0 i 1; com més proper sigui a zero, menys significativa serà la diferència entre les dues variables comparades i viceversa. Així , en la comparació de les dades obtingudes a través d'una estimació amb les dades inicialment simulades (dades reals), com més petit sigui l'STRESS, més bona serà la simulació obtinguda.

3.2 Altres mètodes

Tal i com s'ha explicat a l'inici del present capítol, aquests mètodes només s'anomenen pel fet de que són prou coneguts.

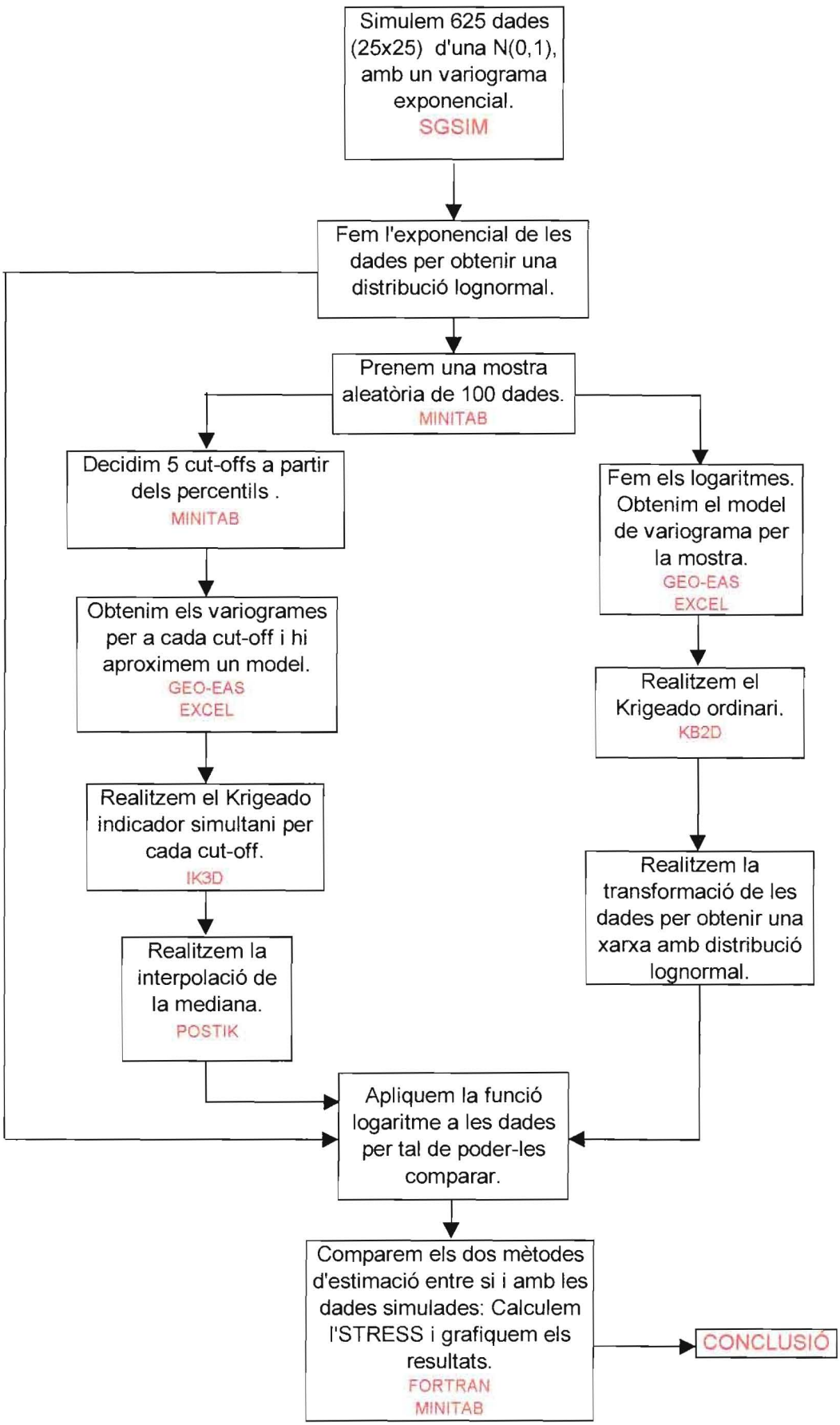
- Descriptiva univariant
- Regressió lineal
- Intervals de confiança
- Gràfics univariants, bivariants, de superfície

Capítol 4: Diagrama de fluxe

La metodología de treball usada en l'elaboració d'aquest projecte, queda reflectida en aquest capítol d'una forma resumida.

Per acompanyar l'explicació s'adjunta el diagrama de fluxe.

FLUXOGRAMA DEL DESENVOLUPAMENT DEL PROJECTE



Creació de dades i recollida de la mostra

Simulació de les 625 dades:

En aquest projecte treballarem sobre una xarxa de 25 x 25 nodes on cada un dels punts contindrà una dada provinent d'una distribució lognormal. S'ha escollit aquest model perquè l'objectiu és estudiar el comportament de dos estimadors de krigeat amb variables lognormals i podrem fer la transformació a una N(0,1) fàcilment.
Per a poder simular l'esmentada xarxa, generarem 625 dades d'una N(0,1) amb l'aplicació SGSIM del programa GSLIB, i tot seguit aplicarem la transformació exponencial per obtenir una distribució lognormal, d'aquesta manera tindrem una base de realitat simulada sobre la que podrem començar a treballar.
Per tal de poder fer aquesta simulació hem decidit utilitzar un semivariograma exponencial
basant-nos en la lectura de diferents articles publicats respecte aquest tipus d'estudis [ref. 2], en els que, en la majoria de casos, el semivariograma de les dades recollides era també exponencial. D'aquesta manera pretenem que les nostres dades segueixin un model estàndard de realitat.
Aquestes 625 dades representen la nostra xarxa real, és a dir, el resultat al qual ens volem aproximar a partir de les simulacions que farem a continuació.

Mostra de 100 dades:

De les 625 dades que conformen el total de la xarxa, pendrem una mostra aleatòria de 100, que representarà la mostra recollida a partir de la qual realitzarem les estimacions.

Krigeat indicador.

Decidir cut-off i modelar semivariogrames:

Per tal de poder realitzar el krigeat indicador serà necessari escollir prèviament els cut-off o punts de tall sobre els quals ens basarem per tal de poder-lo implementar, aquests seran determinats a partir de l'observació dels percentils, d'aquesta manera podrem dividir les dades en blocs de grandària similar. Cal tenir en compte que el nombre de cut-off ha de ser suficientment gran perquè el krigeat tingui la precisió necessària.

Un cop haguem decidit els percentils caldrà observar els semivariogrames corresponents a aquests i determinar el model que segueixen per tal de poder dur a terme el krigeat indicador.

Realització del Krigeat indicador:

Utilitzarem l'aplicació IK3D del GSLIB.

El fitxer resultant contindrà n columnes (una per cada cut-off) amb 625 valors compresos entre 0 i 1 segons correspongui.

Interpolació de la mediana:

Un cop haguem realitzat el krigeat indicador serà necessari interpolar la mediana a través de l'aplicació POSTIK, aquesta, a més de calcular el valor del quantil p corresponent a la funció de distribució condicional, fa correccions en les relacions d'ordre sobre el fitxer obtingut del IK3D.

En el nostre cas el quantil és p=0.5 ja que el que pretenem és estimar la mediana.

D'aquesta manera obtindrem una columna amb les 625 dades resultants de l'estimació de la mediana realitzades a partir del primer mètode: el krigeat indicador.

Apliquem els logaritmes

Per tal de poder dur a terme les comparacions amb l'altre mètode d'estimació i amb la mostra haurem d'aplicar la transformació logarítmica. D'aquesta manera la comparació entre dues variables és molt més simple i els gràfics seran més fàcils d'interpretar.

Krigeat Lognormal

Fer logaritmes i modelar el semivariograma

En primer lloc transformarem la nostra variable en una distribució normal aplicant el logaritme a les dades.

A continuació ajustarem un model de semivariograma per la mostra de dades transformades a N(0,1). Aquest model serà utilitzat en la realització del krigeat.

Realització del krigeat ordinari

Utilitzem l'aplicació KB2D per a poder realitzar el krigeat ordinari sobre la mostra de dades lognormals, transformada.

El resultat és una estimació de les 625 dades que conformen la nostra xarxa, conservant intactes els valors mostrals observats.

Comparació dels dos mètodes

Un cop realitzats els dos mètodes d'estimació donarem pas al darrer apartat del projecte que consistirà en la comparació del krigeat indicador amb el krigeat lognormal, i a continuació la comparació de cada un dels dos mètodes amb les dades "reals". Es tracta de veure principalment si els resultats obtinguts son similars entre ells i si s'aproximen a la realitat (en el nostre cas, coneguda). Per a poder-ho fer ens basarem en mètodes estadístics de comparació d'un conjunt de dades.

També es pretén determinar quin dels dos mètodes és millor, és a dir quin dels dos mètodes de krigeat estima més fidedignament la realitat. Això ho aconseguirem amb la comparació entre les dades que configuraven la nostra xarxa de partida i les obtingudes a partir dels dos mètodes.

En cas de que els resultats siguin similars, es recomanaria l'utilització en un futur, del krigeat lognormal, ja que és un mètode molt més simple i ràpid que el krigeat indicador.

Capítol 5: Desenvolupament del projecte

Tot estudi geoestadístic segueix un guió de treball:

- Determinació i anàlisi de l'estructura espaial
- Modelització de la variabilitat espaial
- Estimació de punts no mostrejats
- Simulació
- Conclusions i descripció de la zona estudiada.

En el nostre cas, malgrat no tenir dades reals i que el nostre objectiu no sia el coneixement de la variable regionalitzada sinó el comparar dos mètodes d'estimació, el procés serà el mateix. L'única diferència és que nosaltres fem una simulació per tal d'obtenir una base de realitat a partir de la qual començar a treballar i no per tenir realitzacions de punts no mostrejats. Per tant, en el nostre cas, la simulació esdevé el primer pas a realitzar.

5.1 La simulació

Per tal de tenir una regió en la que poder treballar, construïm una xarxa de 25x25x1, és a dir, treballem a R^2.

Volem obtenir una realització en cada un dels nodes de la xarxa, per tant realitzem una simulació de 625 dades. Figura 5.1.1

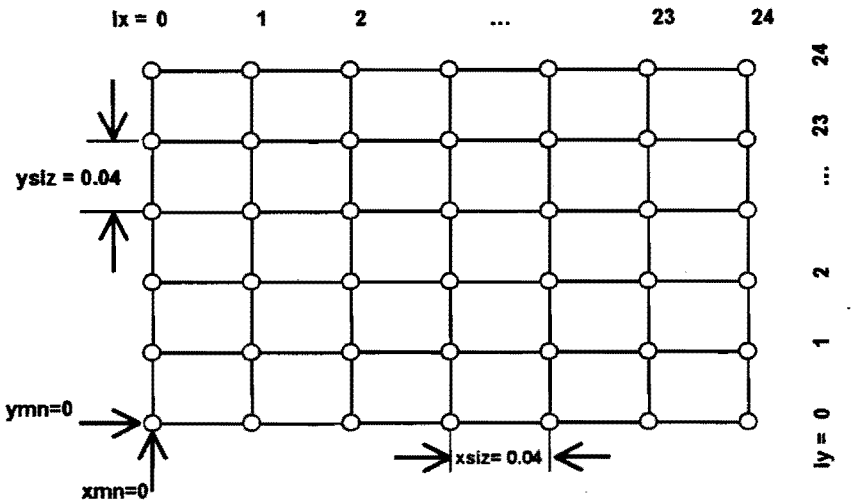


Figura 5.1.1: Xarxa sobre la que es treballa i punts mostrejats.

En el camp de la geología és molt difícil obtenir gaires mostres d'una regió, per això els mètodes de simulació i d'estimació són tant importants en geoestadística. Com que a nosaltres ens interesa tenir una xarxa complerta (perquè el que ens interessa és veure quin dels dos mètodes d'estimació és millor i no el saber com és la regió), el que fem és simular realitzacions d'una N(0,1), amb semivariograma exponencial de meseta unitat, en tots els nodes de la nostre xarxa. S'ha escollit aquest tipus de semivariograma per ser el més habitual en el camp de la geología. [ref 2]

En aquestes dades els hi apliquem exponencials i per tant la nostra realitat segueix una distribució lognormal de mitjana unitat.

El fet de treballar sobre dades lognormals ve justificat perque hi ha molts fenomens naturals que segueixen aquesta distribució [ref 11] (cerca de determinat mineral, bosses de petroli, quantitat de sediments, exploracions geoquimiques,...)

Per tant, la nostre realitat és aquesta, és a dir, la distribució lognormal de la variable regionalitzada. La N(0,1) tant sols la fem servir per ser fàcil de maniobrar computacionalment. [ref 16]

Per tal d'assegurant-se que la nostre realitat és mes o menys exacta mirem la següent taula:

Variable	N	Mean	Median	StDev	SE Mean	Min	Max	Q1	Q3
exp(sim)	625	1.0546	0.7742	1.0075	0.0403	0.0250	9.5965	0.3934	1.3721

Taula 5.1.1 Descriptiva de la variable regionalitzada exp(sim)

Amb els següents gràfics comprovem que realment les nostres dades segueixen una distribució lognormal.

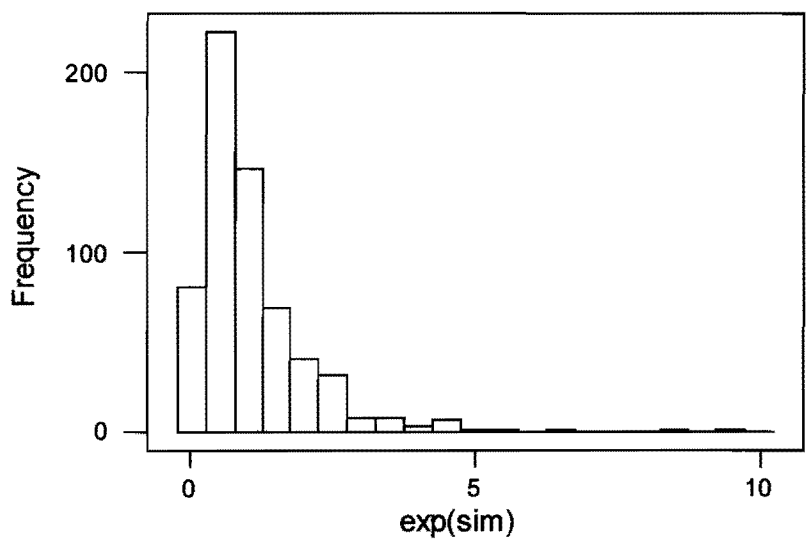


Figura 5.1.2 Histograma de la realitat.

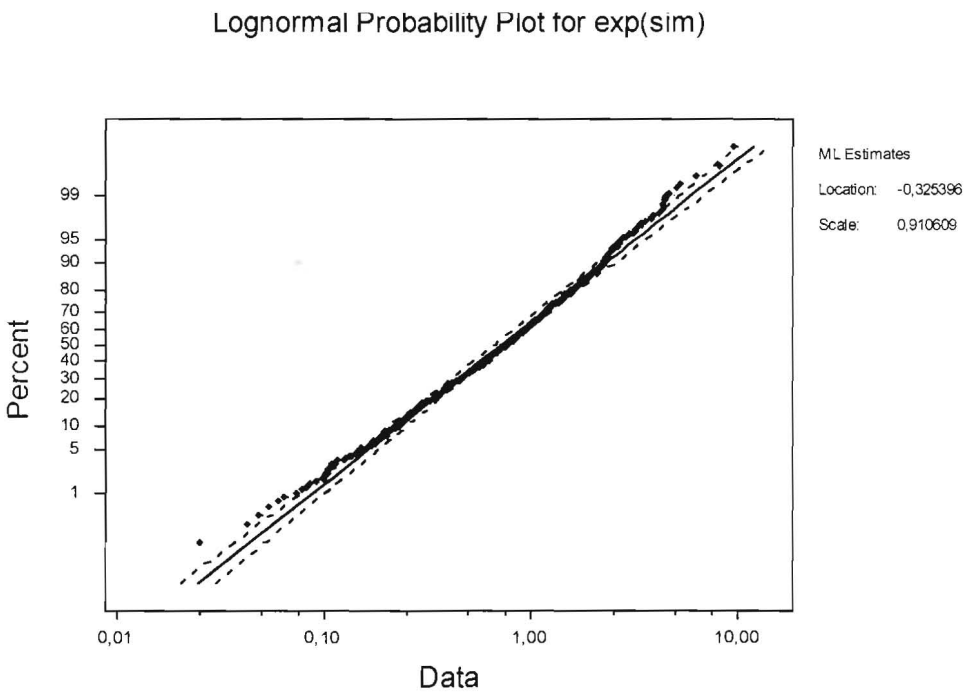


Figura 5.1.3: Lognormal probability plot de la variable regionalitzada exp(sim)

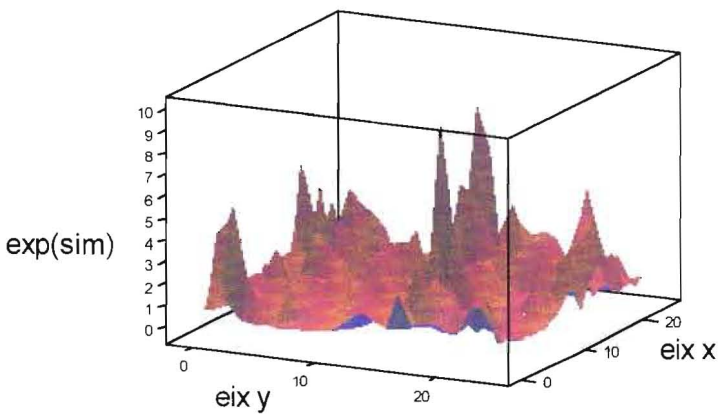


Figura 5.1.4: Surface plot de la variable exp(sim).

El surface plot és una eina molt útil, ja que ens dóna una visió de com es la nostra realitat a l'espai. A més un cop haguem fet les dues estimacions, també podrem fer el surface plot i amb un cop d'ull podrem treure una primera impressió de si les estimacions han estat bones o no.

5.2 La mostra

A partir de les dades que hem obtingut amb la simulació prenem una mostra aleatòria de 100 dades a partir de la qual realitzarem les dues estimacions de la realitat. Cal tenir en compte que a partir d'aquest moment es treballarà sobre la mostra i no sobre les dades simulades que en l'experimentació real són desconegudes.

La mostra s'ha pres de forma que constitueixi una bona representació del total de la xarxa. La figura 5.2.1 mostra la localització de cada una de les dades que formen part de la mostra sobre la xarxa inicial de 25x25 nodes.

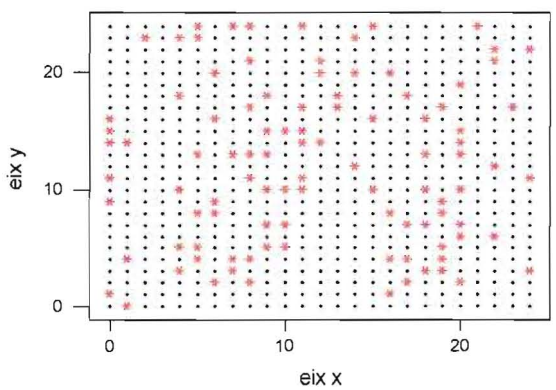


Figura 5.2.1 Distribució de la mostra dins la xarxa real.

La mostra obtinguda conserva les propietats de les dades provinents de la simulació, d'aquesta manera podem afirmar que és una bona extrapolació de la realitat. Vegem a continuació la descriptiva (Taula 5.2.1) i l'histograma de la mostra (Figura 5.2.2) amb el gràfic de probabilitats (Figura 5.2.3). Podem observar que segueix, efectivament una distribució lognormal.

Variable	N	Mean	Median	TrMean	StDev	SE Mean
mostra	100	1,0552	0,8330	0,9458	0,9566	0,0957

Variable	Minimum	Maximum	Q1	Q3
mostra	0,0250	4,4185	0,3281	1,3714

Taula 5.2.1: Descriptiva de la mostra

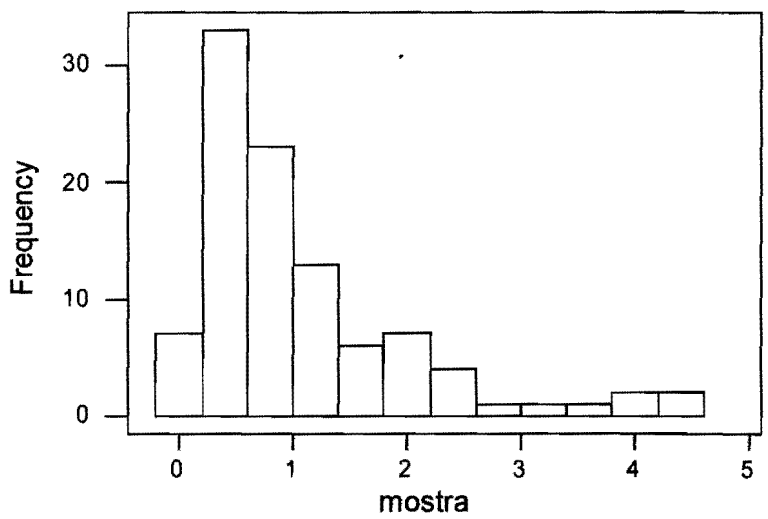


Figura 5.2.2 Histograma de la mostra

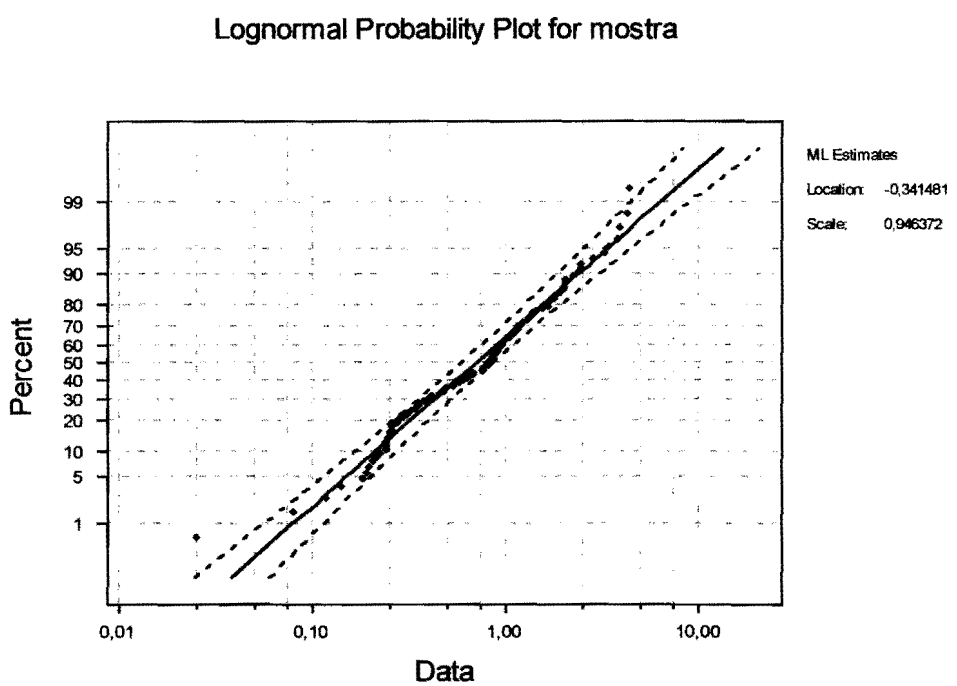


Figura 5.2.3 Plot de probabilitat lognormal.

La superfície descrita per la mostra podem representar-la també gràficament. (Figura 5.2.4)

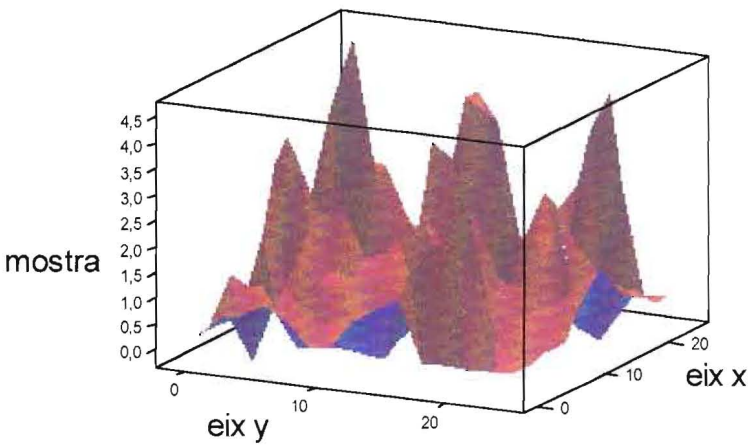


Figura 5.2.4 Superfície definida per la mostra

5.3 L'estimació per krigeat indicador

Per tal de realitzar el krigeat indicador definim 5 punts de tall o *cut-off*. Aquests punts de tall els escollim de forma que guardin simetria respecte el percentil del 50%, tal i com mostra la taula 5.3.1

Categoria	Percentil	Percentatge
1	9,0	0,09952
2	30,0	0,37637
3	50,0	0,73142
4	70,0	1,27046
5	91,0	2,54092

Taula 5.3.1 Els cinc punts de tall escollits.

Un cop decidides les categories o punts de tall, necessitem saber el semivariograma que segueixen cada una d'elles per a poder dur a terme el krigeat indicador, ja que l'estimació realitzada a partir d'aquest mètode requereix una modelització dels cinc semivariogrames experimentals.
El procediment és el següent:

Partim de la mostra de 100 dades que segueix una distribució lognormal.

Per cada una de les categories creem una columna de 0 i 1 de manera que si aquell punt concret és per sota del punt de tall el valor que pren la variable dicotòmica és 1 i si és per sobre pren el valor 0.

Els semivariogrames els creem amb aquestes cinc columnes de zeros i uns.

Per cada semivariograma provem d'ajustar un model esfèric i un d'exponencial. Aquests dos tipus de models els anirem ajustant a ull movent el valor del rang i de l'abast. Partim d'aquests models per ser els més habituals en la realitat, i ens quedarem amb el que ajusti millor dels dos [ref. 2].

A continuació presentem els resultats d'aquests ajustaments per a cada punt de tall.

Categoria 1 $\gamma(h)$

Model Exponencial: Abast = 8
Meseta = 0.03

Model Esfèric: Abast = 7.5
Meseta = 0.03

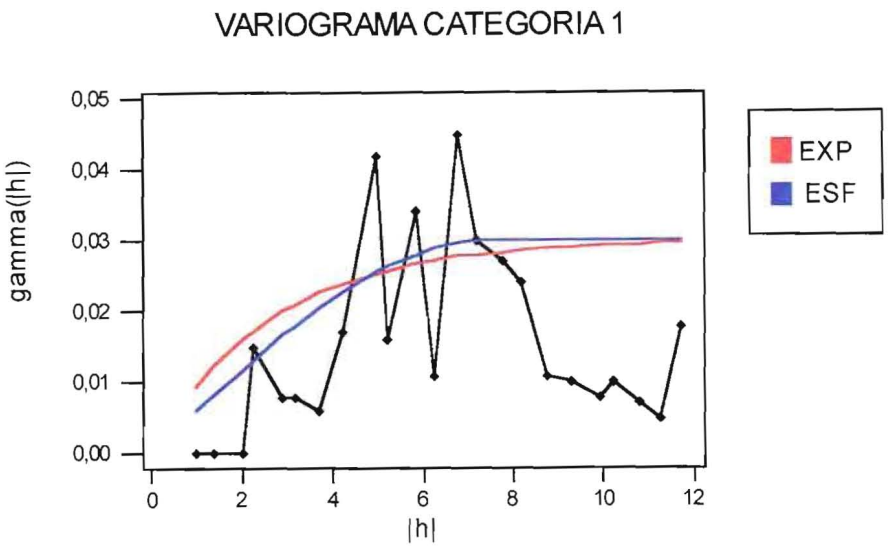


Figura 5.3.1: Model de semivariograma esfèric i exponencial per la categoria 1.

Categoria 2

Model Exponencial: Abast = 4
Meseta = 0.22

Model Esfèric: Abast = 4
Meseta = 0.22

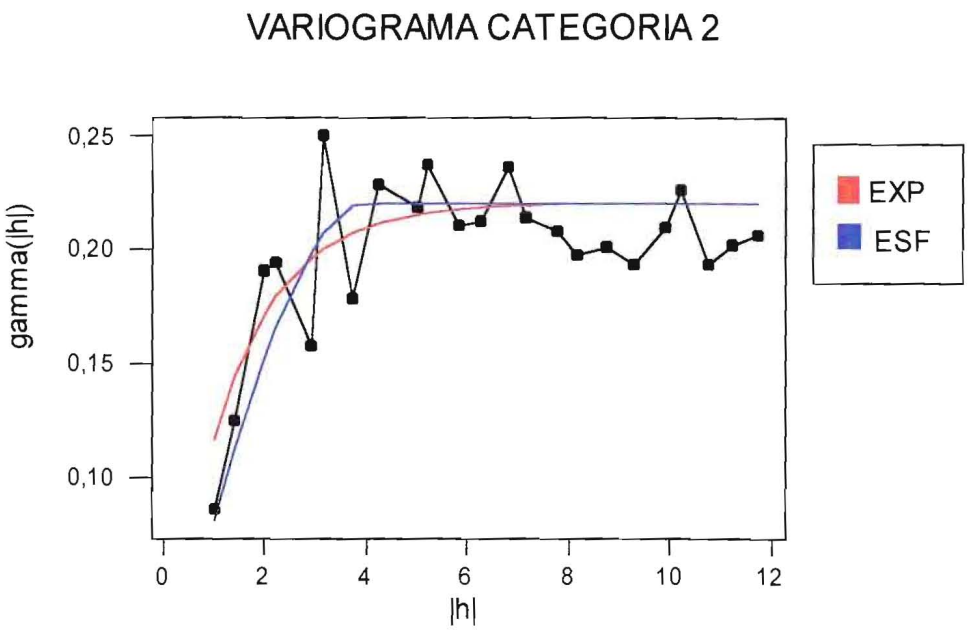


Figura 5.3.2: Model de semivariograma esfèric i exponencial per la categoria 2.

Categoria 3

Model Exponencial: Abast = 7
Meseta = 0.27

Model Esfèric: Abast = 6
Meseta = 0.265

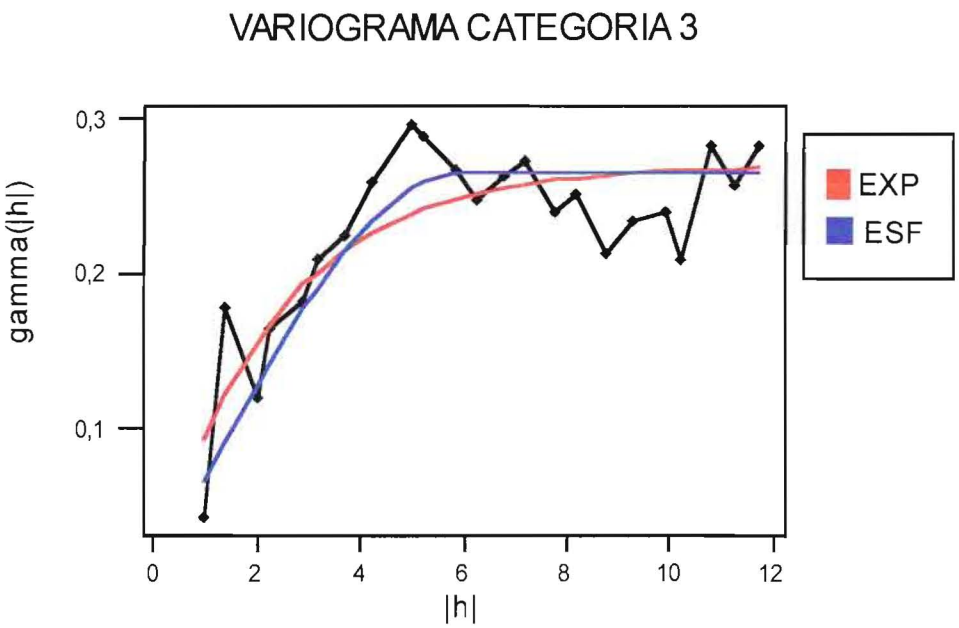


Figura 5.3.3: Model de semivariograma esfèric i exponencial per la categoria 3.

Categoria 4

Model Exponencial: Abast = 5
Meseta = 0.2

Model Esfèric: Abast = 4
Meseta = 0.2

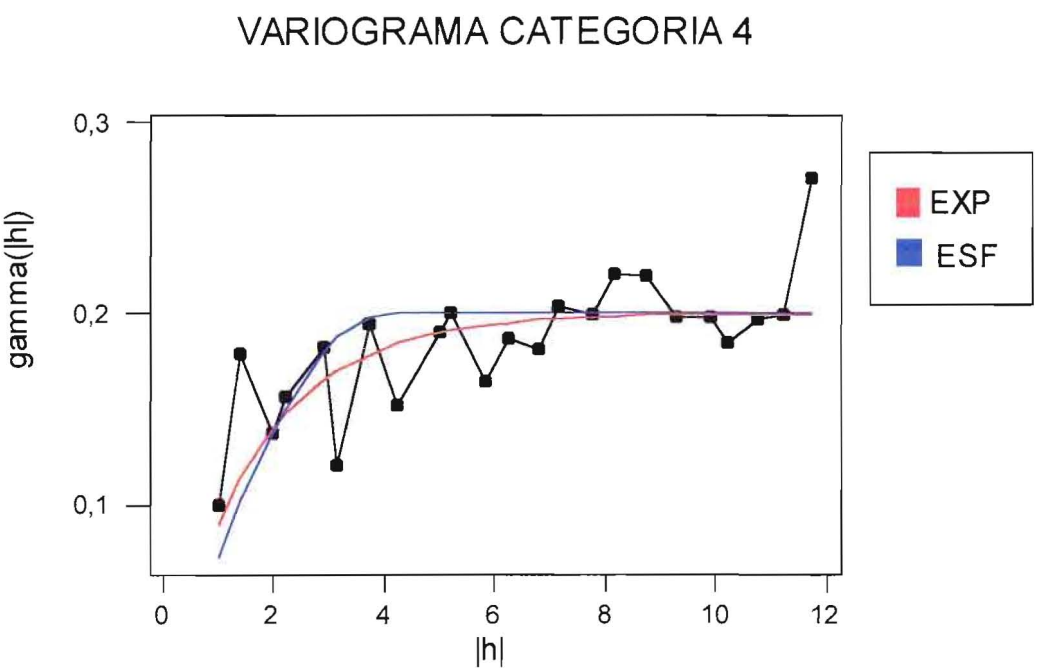


Figura 5.3.4: Model de semivariograma esfèric i exponencial per la categoria 4.

Categoria 5

Model Exponencial: Abast = 4
Meseta = 0.06

Model Esfèric: Abast = 3
Meseta = 0.06

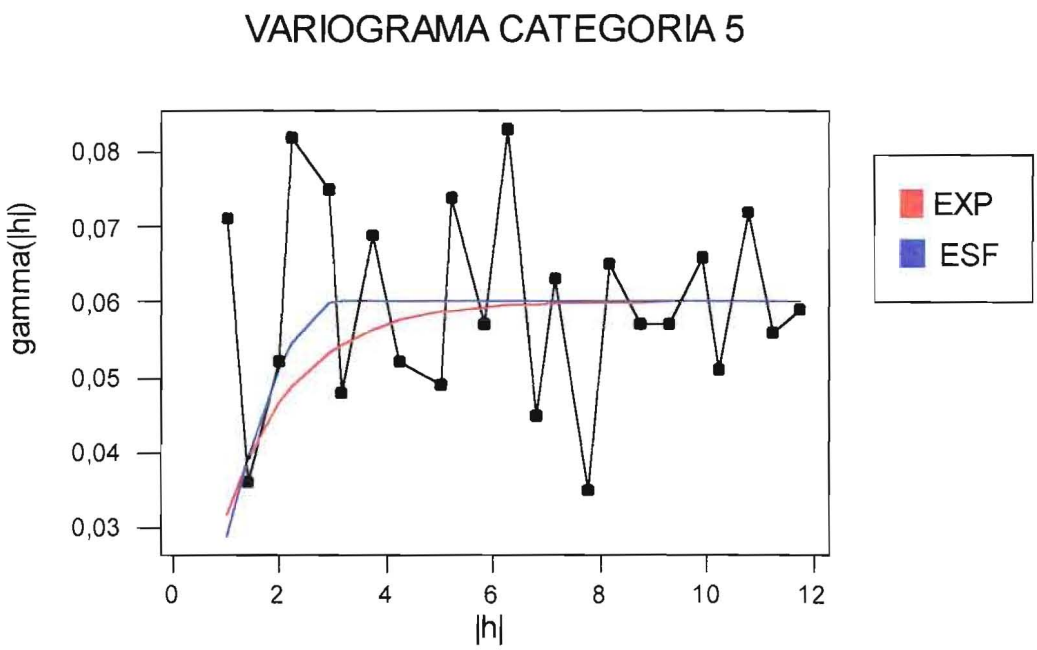


Figura 5.3.5: Model de semivariograma esfèric i exponencial per la categoria 5.

Mirant els cinc gràfics, veiem que el model que ajusta millor és l'esfèric en tots els casos, encara que no hi ha diferències massa significatives respecte el model exponencial. A més observem una certa simetria pel que fa a la meseta i abast de cada un dels models esfèrics de cada categoria. Taula 5.3.2

	Cat 1	**Cat 2**	**Cat 3**	**Cat 4**	**Cat 5**
Meseta	0.03	0.22	0.265	0.2	0.06
Abast	7.5	4	6	4	3

Taula 5.3.2: Meseta i abast per cada categoria.

Aquesta simetria s'observa sobretot per les categories 2, 3 i 4
Els models corresponents a les categories 1 i 5 no ajusten del tot bé degut a que són les categories dels extrems i, per tant la gran majoria de mostres si no totes, estaran per sobre o per sota respectivament.

Ara ja podem fer l'estimació per krigeat indicador. Al programa ik3d, li introduïm els cinc models de semivariogrames per tal que ens calculi l'estimació.

El resultat que obtenim són cinc columnes de 625 dades on els valors van de 0 a 1. Concretament ens calcula la probabilitat de que el valor atorgat a cada punt sigui inferior al d'aquell cut-off.
El que es pretén ara és, a partir d'aquest conjunt de probabilitats, interpolar la mediana per a cada un dels punts de la xarxa, per tant el que fem és executar el programa postik en l'opció 3, de manera que ens tregui l'estimació de la mediana.

A continuació passem a detallar en que consisteix el postik. Per tal de poder-ho expressar millor ens centrarem en un punt concret de la xarxa, el (3,0).

Els tercer valor de cada una de les 5 columnes obtingudes amb el krigeat indicador, és a dir, els valors corresponents a la mostra situada en el node (3,0), representen la probabilitat de que el valor de la variable en aquest punt sigui major que cada un dels cut-off respectivament, així podem afirmar que:

$$P(x_i < c1) = P(x_{(3,0)} < 0{,}0995) = 0.0000$$
$$P(x_i < c2) = P(x_{(3,0)} < 0{,}3764) = 0.0397$$
$$P(x_i < c3) = P(x_{(3,0)} < 0{,}7314) = 0.0409$$
$$P(x_i < c4) = P(x_{(3,0)} < 1{,}2705) = 0.2079$$
$$P(x_i < c5) = P(x_{(3,0)} < 2{,}5402) = 1.0000$$

Si ara grafiquem aquests 5 valors obtenim una estimació de la probabilitat que té el punt estudiat (3,0) de ser més petit que qualsevol valor, és a dir, l'estimació de la funció de probabilitat en aquest punt. Com que el que es pretén estimar en el nostre estudi és la mediana, mirarem quin valor aproxima per la probabilitat del 0.5, és a dir:

$$m \text{ tal que } P(x_{(3,0)} < m) = 0.5$$

Per això tracem una línia a l'alçada 0.5 i d'aquesta manera obtenim el valor estimat de la mediana. Aquest valor correspon al valor que obtenim amb el POSTIK, que és, per aquesta primera observació, 0.6983

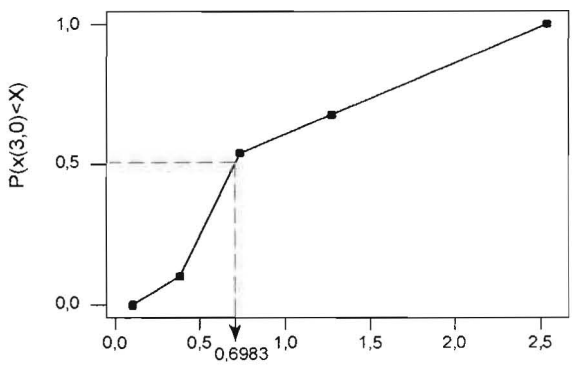


Figura 5.3.6: Els 5 cut-off aplicats al punt (3,0)

L'aplicació del POSTIK, realitza aquesta estimació per a cada un dels punts obtinguts (625) i, d'aquesta manera podem obtenir una interpolació de la mediana per cada un dels valors a estimar.

5.4 L'estimació per krigeat lognormal

Per tal de dur a terme el krigeat lognormal utilitzem l'aplicació KB2D del programa GSLIB.

A més de les dades mostrals, per realitzar una estimació dels punts no mostrejats es necessita una estimació del semivariograma, és a dir, de la variància espaial. Això és una aproximació del rang i de la meseta. Alguns autors utilitzen la variància mostral com a estimació de la meseta del semivariograma, però si el semivariograma experimental ens representa una meseta clarament definida es pot prendre aquesta com a estimació de la variància poblacional. [ref. 1]

També cal remarcar que per trobar una estimació raonable de la meseta es necessita una dimensió de dades superior a tres cops el rang del semivariograma (abast). [ref.1] Primerament ens caldrà veure amb quin model de semivariograma volem estimar les dades. Per fer-ho realitzem el semivariograma a partir de la mostra lognormal transformada a una N(0,1) i ajustem el model més adient.

Per tal d'escollir el model més apropiat pel semivariograma experimental ajustem els dos models que son més freqüents en aquest tipus de dades [ref. 2] : l'exponencial i l'esfèric.

MODEL EXPONENCIAL: Aquest model es formula de la següent manera:

$$\gamma(h) = c * \left[1 - \exp\left(\frac{-3h}{a}\right)\right]$$

on a és l'abast i c la meseta

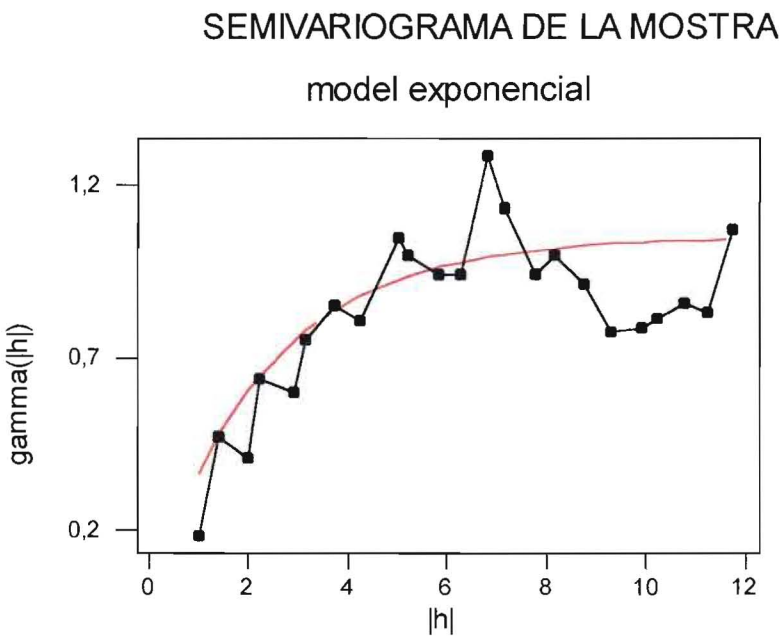


Figura 5.4.1 : Semivariograma experimental amb un model exponencial

Aquest model obtingut "a ull" té els següents paràmetres: abast=7

meseta=1.05

Aplicant aquest valors al model general obtenim la formulació per a les nostres dades:

$$\gamma(h) = 1.05 * \left[1 - \exp\left(\frac{-3h}{7}\right)\right]$$

MODEL ESFÈRIC: La formulació general és:

$$\gamma(h) = \begin{cases} c * \left(1{,}5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3 \right), & si\, h \le a \\ c, & si\, h > a \end{cases}$$

on a és l'abast i c la meseta

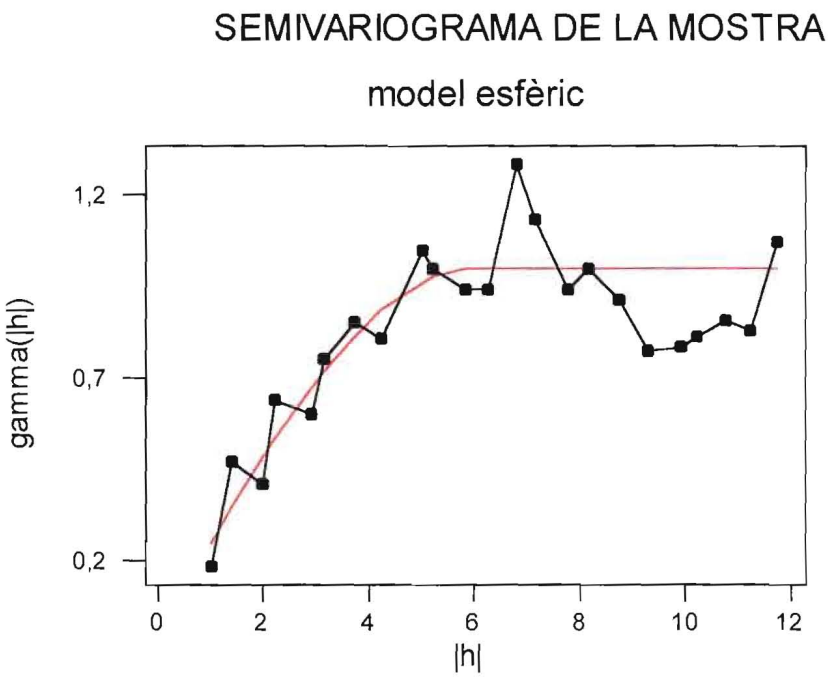


Figura 5.4.2: Semivariograma experimental amb un model esfèric.

Aquest model l'hem obtingut de la mateixa manera que l'anterior: a ull i els seus paràmetres, que també aplicarem a la forma general són: abast=6 i meseta=1

El nostre model esfèric és per tant:

$$\gamma(h)=\begin{cases}1*\left(1{,}5\frac{h}{6}-0.5\left(\frac{h}{6}\right)^3\right), & si\ h\le 6\\ 1, & si\ h>6\end{cases}$$

Al ser els dos models molt semblants en quan a precisió, escollim el model exponencial per ser el més usat en els estudis realitzats fins ara [ref. 2], encara que utilitzar el model esfèric tampoc seria una determinació errada i els resultats no serien massa diferents dels obtinguts.

Un cop escollit per tant el model de variograma que usarem per a realitzar l'estimació lognormal passarem a executar el KB2D. Per fer-ho i tal com s'indica en estudis realitzats [ref. 1], hem de partir d'una transformació dels valors de la mostra per tal de poder treballar amb dades normals. Aquesta transformació és:

$$Y(u)=\ln X(u)$$

on $X\sim \text{lognormal}$ i, en conseqüència $Y\sim N(0,1)$

A partir de les dades de la mostra transformades i conservant la seva localització espaial, ja podem realitzar l'estimació a partir del krigeat ordinari.

El resultat és una estimació de les 625 posicions conservant, en els punts pertanyents a la mostra, les dades mostrals transformades.

Per obtenir el resultat final del krigeat lognormal l'únic que haurem de fer és una segona transformació, de la forma següent:

$$Z(u)=\exp\left\{Y(u)+\frac{\sigma_{KO}^2}{2}\right\}$$

Així obtenim l'estimació de les 625 dades en l'espai que segueixen una distribució lognormal no esbiaixada.

Si mirem el plot de probabilitats (Figura 5.4.3) i l'histograma (Figura 5.4.4) corresponent a la estimació podem comprovar que efectivament les dades segueixen una distribució lognormal.

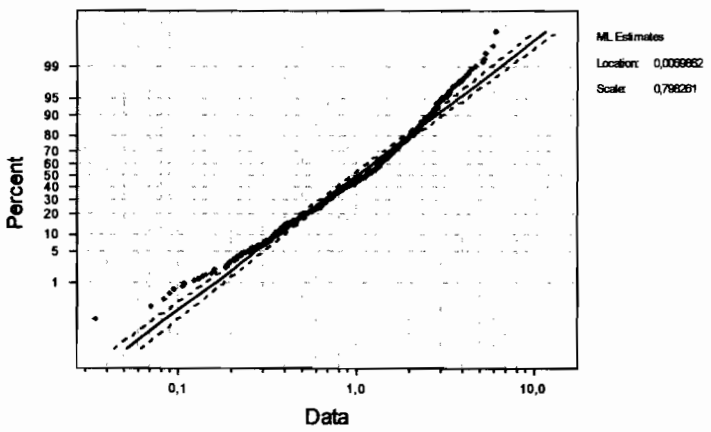


Figura 5.4.3: Lognormal probability plot

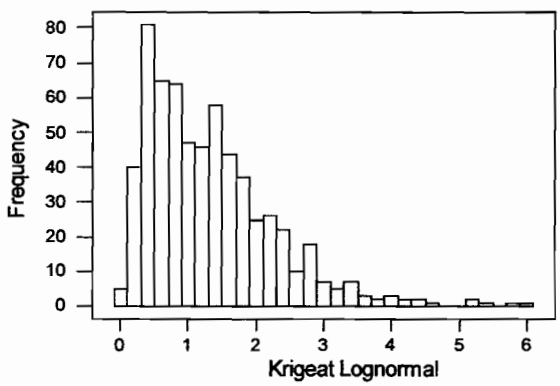


Figura 5.4.4: Histograma de l'estimació per krigeat lognormal

Finalment podem representar la superfície que hem estimat a partir del krigeat lognormal utilitzant un gràfic de superfície (Figura 5.4.5)

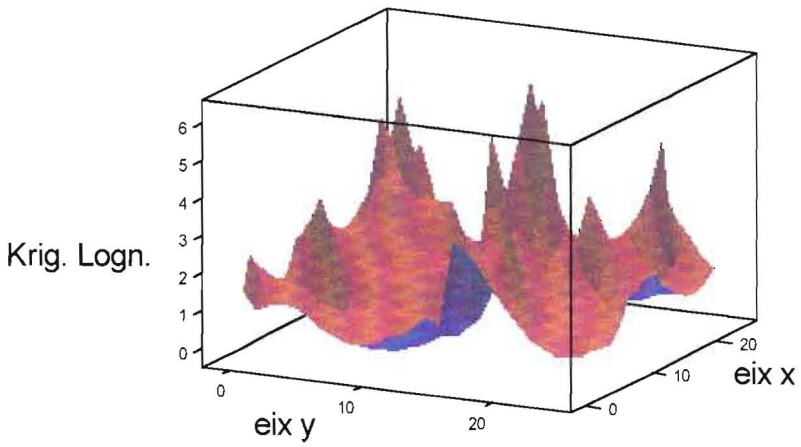


Figura 5.4.5: Gràfic de superfície pel krigeat lognormal

Capítol 6: Comparació dels resultats obtinguts

En aquest capítol analitzem els resultats obtinguts en l'estimació per krigeat lognormal i krigeat indicador i els comparem amb la simulació original. També comparem els dos mètodes d'estimació entre sí.

Ens basem en la regressió lineal i amb l'STRESS (estadístic descrit a l'apartat 3.1). Aquestes comparacions són les que, posteriorment, ens ajudaran a extreure les conclusions.

6.1 Regressió lineal

Per tal de dur a terme la comparació entre els dos mètodes d'estimació i les dades que simulen la realitat, hem expressat els resultats obtinguts gràficament mitjançant els gràfics de punts ajustant en cada cas la recta de regressió i dibuixant la bisectriu. Cal tenir en compte que per augmentar la claredat i per a simplificar la interpretació dels resultat hem transformat les dades a una N(0,1) encara que el nostre estudi es basi en un model lognormal.

Passem a continuació a analitzar els gràfics.

En primer lloc (Figura 6.1.1) tenim el gràfic de punts entre les dades obtingudes amb el primer mètode d'estimació, el krigeat lognormal, i les dades provinents de la simulació de la realitat:

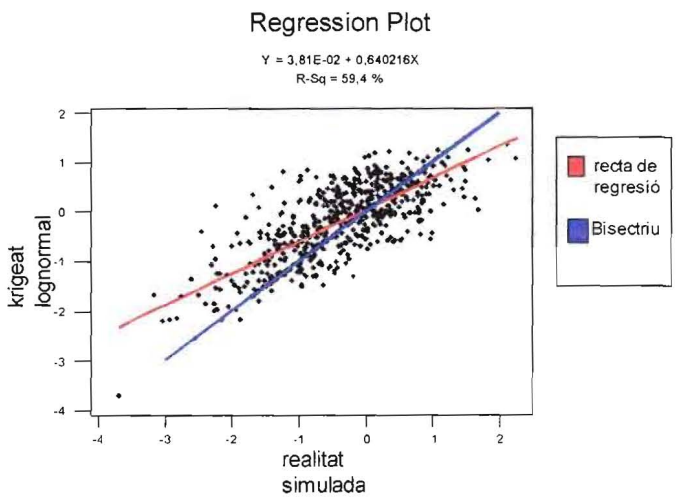


Figura 6.1.1 Recta de regressió entre el krigeat lognormal i la simulació de la realitat

Si mirem aquest gràfic podem observar que el valor de l'estimació realitzada sobreestima els valors petits de la simulació i, en canvi en els valors grans, subestima.

El coeficient de correlació es del 59.4%, és a dir que el mètode del krigeat lognormal explica un 59.4% de la realitat.

Passem ara a realitzar el gràfic entre el segon mètode d'estimació, el krigeat indicador, i la realitat simulada (Figura 6.1.2):

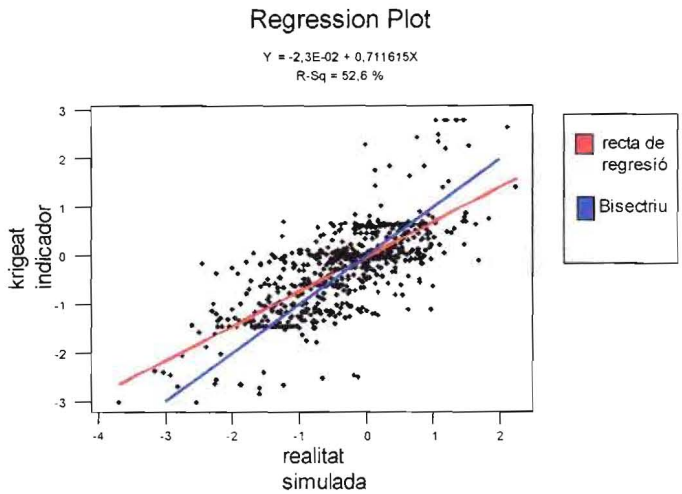


Figura 6.1.2: Recta de regressió entre el krigeat indicador i la simulació de la realitat

Podem observar que el gràfic obtingut és molt semblant a l'anterior i que ens trobem amb el mateix problema de sobreestimació dels valors petits i subestimació dels valors grans de la simulació.

El valor del coeficient de correlació també és força semblant: 52.6% Així doncs la realitat és explicada en un 52.6% pel krigeat indicador.

Finalment i per tenir una idea de la relació que hi ha entre els dos mètodes emprats en el projecte, realitzarem un últim gràfic comparatiu (Figura 6.1.3) que representa els dos mètodes d'estimació: el krigeat lognormal i el krigeat indicador:

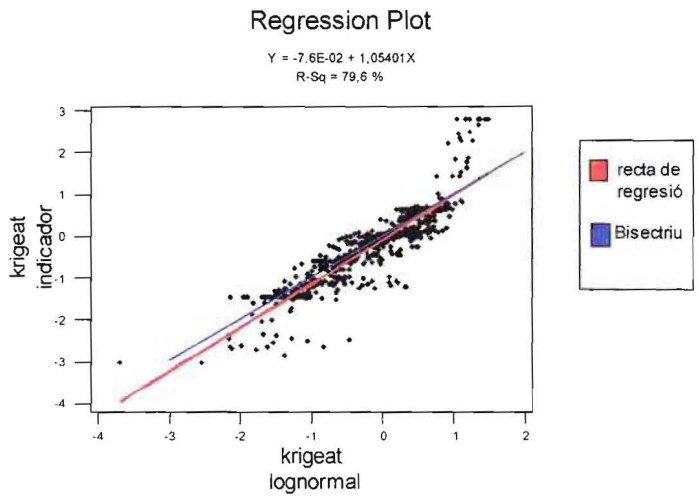


Figura 6.1.3 Recta de regressió entre el krigeat lognormal i el krigeat indicador

En aquest cas podem observar que la recta de regressió coincideix gairebé a la perfecció amb la bisectriu, cosa que vol dir que els dos mètodes són molt semblants entre sí i que, per tant, és normal que l'aproximació de cada un d'ells a la realitat sigui semblant.

El valor de la regressió és més gran que en els altres dos casos: 79.6%, és a dir que els dos mètodes no donen resultats significativament diferents entre si.

Podem observar, un conjunt de dades que prenen valors especialment elevats en l'estimació per krigeat indicador. Si analitzem aquests punts podem comprovar que són punts que tenen una probabilitat de ser més grans que el cinquè cut-off (percentil 0,97) igual a 1. Per tant, aquest tipus de punts no s'han de tenir en compte en el moment d'extreure conclusions. De fet, això ja es sospitava en el gràfic anterior, però en que en aquell hi ha una elevada dispersió i per tant l'efecte queda dissimulat.

La taula 6.1.1 resumeix els paràmetres més importants tant per la realitat simulada com per les dues estimacions per krigeat.

Variable	N	Mean	Median	TrMean	StDev	SE Mean
K.log	625	-0,3121	-0,1840	-0,2825	0,7989	0,0320
K.ind	625	-0,2550	-0,0865	-0,2651	0,8940	0,0358
realitat	625	-0,3254	-0,2559	-0,3067	0,9113	0,0365

Variable	Minimum	Maximum	Q1	Q3	95,0 % CI
K.log	-3,6890	1,4860	-0,7940	0,2595	(-0,3749; -0,2494)
K.ind	-2,9957	2,7893	-0,8517	0,1866	(-0,3252; -0,1848)
realitat	-3,6890	2,2614	-0,9330	0,3164	(-0,3970; -0,2538)

Taula 6.1.1: Descriptiva univariant per la realitat i per les estimacions.

6.2 STRESS

Per tal de calcular l'STRESS apliquem les formules conegudes (veure apartat 3.1) a les nostres dades i obtenim així una mesura per a poder comparar les estimacions realitzades. Passem a veure a continuació la seva aplicació i els resultats obtinguts:

Comparació entre la realitat simulada i els resultats del krigeat lognormal:

$$STRESS = \sqrt{\frac{\sum_{I<J}\left(\delta_{IJ} - \delta^{*}_{IJ}\right)^2}{\sum_{I<J}\left(\delta^{*}_{IJ}\right)^2}} = 0.6308$$

on:

$\delta_{ij} = |x(i) - x(j)| \Rightarrow$ distància euclidiana entre els parells de punts provinents de la simulació transformada a N(0,1)

$\delta^{*}_{ij} = |y(i) - y(j)| \Rightarrow$ distància euclidiana entre els parells de punts provinents de les estimacions per krigeat lognormal

x : vector que conté les dades provinents de la simulació

y : vector que conté els punts obtinguts amb el krigeat lognormal transformats.

Comparació entre la realitat simulada i els resultats del krigeat indicador:

$$STRESS = \sqrt{\frac{\sum_{i<j}\left(\delta_{ij} - \delta_{ij}^{*}\right)^2}{\sum_{i<j}\left(\hat{\delta}_{ij}\right)^2}} = 0.8092$$

on:

$\delta_{ij} = |x(i) - x(j)| \Rightarrow$ distància euclidiana entre els parells de punts provinents de la transformació de la simulació a N(0,1)

$\hat{\delta}_{ij} = |z(i) - z(j)| \Rightarrow$ distància euclidiana entre els parells de punts provinents de les estimacions per krigeat indicador.

x : vector que conté les dades provinents de la simulació

z : vector que conté els punts obtinguts amb el krigeat indicador. transformats.

Cal tenir en compte que per tal de poder aplicar aquest estadístic, hem hagut de transformar les dades per tal de que seguissin una N(0,1) en els tres casos i simplificar, d'aquesta manera la comparació.

Si ens fixem en els dos resultats obtinguts podem concloure que el valor de l'STRESS utilitzant el krigeat lognormal és significativament inferior al valor obtingut utilitzant el krigeat indicador. Així doncs la informació que ens aporta l'STRESS es que el krigeat lognormal té més precisió d'estimació que el krigeat indicador.

Finalment hem calculat també l'STRESS entre els dos mètodes de krigeat per tal de veure fins a quin punt les dues estimacions obtenen resultats similars:

$$STRESS = \sqrt{\frac{\sum_{i<j}\left(\delta_{ij}^{*} - \hat{\delta}_{ij}\right)}{\sum_{I<J}\left(\hat{\delta}_{ij}\right)^2}} = 0.5325$$

En aquest cas l'STRESS és més proper a 0 que en els altres dos casos; això vol dir que les dues estimacions obtenen resultats més similars entre si que a la realitat, cosa que era d'esperar si tenim en compte que s'ha utilitzat la mateixa mostra per obtenir els resultats en els dos mètodes.

Capítol 7: Conclusions

L'objectiu d'aquest estudi és la comparació de dos mètodes d'estimació espaial, (krigeats), per veure si els dos tenen la mateixa potència estimativa, o bé si algun d'ells s'aproxima més a la realitat que l'altre.

El punt de partida ha estat la simulació d'una realitat constituïda per 625 dades localitzades sobre una xarxa de 25x25 nodes que segueixen una distribució lognormal. Del total de dades simulades hem extret una mostra aleatòria de 100, a partir de la qual hem realitzat els dos mètodes de krigeat a comparar: el lognormal i l'indicador. Els dos krigeats realitzen una estimació del tramat original obtinguda a partir de la mostra.

Arribats a aquest punt podem fer l'analisis comparatiu dels dos tipus d'estimació.

Gràficament, mitjançant els plots (Figures 6.1.1 i 6.1.2) entre els krigeats i la realitat de la simulació, hem observat que malgrat ser la bisectriu semblant a la regressió, aquesta darrera no és tan bona com s'esperava, encara que s'ha de tenir en compte que en tot estudi de simulació de realitzacions de dades i en general quan intentem reproduir una realitat fent veure que no la coneixem a partir d'una mostra la variabilitat resultant és elevada.
En ambdós gràfics s'observa una elevada dispersió, sobretot en l'estimació per krigeat indicador. Per tant, a primer cop d'ull, podem començar a pensar que les estimacions no han estat gaire precises, malgrat que el nivell de dispersió de les dades és manté en les tres realitzacions.
Un altre gràfic que analitzem, és el del krigeat indicador amb el krigeat lognormal (Figura 6.1.3); en aquest cas veiem que la bisectriu i la regressió coincideixen força i que la dispersió ha disminuït si la comparem amb les gràfiques anteriors. Això ens porta a pensar que l'estimació pels dos mètodes es bastant similar, encara que, com ja hem vist, la obtinguda a través del krigeat lognormal és lleugerament millor.

A partir del resultat obtingut amb l'aplicació de les fórmules de l'STRESS, arribem a les mateixes conclusions que gràficament, és a dir, que el krigeat lognormal ens proporciona una estimació més bona de la realitat que el krigeat indicador , encara que cap dels dos mètodes sigui òptim:

STRESS(k.lognormal, simulació)=0.63075
STRESS (k.indicador, simulació)=0.80924

Tant el krigeat indicador com el lognormal han estimat uns valors que no concorden amb la realitat si mirem l'stress.
L'objectiu era respondre a una pregunta: Quin dels dos mètodes de krigeat és millor, el lognormal o l'indicador? La resposta és que tots dos fan una estimació similar, però el krigeat lognormal ens proporciona més exactitud que el krigeat indicador.

Finalment caldria remarcar que aquestes conclusions s'han extret a partir de la realització d'una sola aplicació del procés i, per tant, per arribar a una conclusió final sobre el problema que ens ocupa, al no basant-nos en proves teòriques sinó en

simulacions pràctiques, s'hauria de realitzar tot el procediment per varies mostres i canviant els factors modificables que poden afectar als resultats finals del procés, com és ara l'elecció dels cut-off en el krigeat indicador, així com la quantitat de punts de tall, la distància entre lags per cercar els parells de punts; estudiant la sensibilitat envers ells.

Aquest exercici quedaria obert per a futures investigacions sobre la comparació de mètodes de krigeat en l'àrea de la geoestadística.

Bibliografia

[ref 1] Barnes, Randal J.. *The variogram sill and the sample variances*: Mathematical Geology, Vol. 23, No4, 1991.

[ref 2] Bogaert, P. *On the optimal estimation of the cumulative distribution function in presence of spatial dependence:* Mathematical Geology, Vol. 3, No2, 1999.

[ref 3] Clark, I. *Practical Geostatistical.* Applied science publishers, London, 1979

[ref 4] Cox T.F. and Cox M.A.A., *Multidimensional Scaling.* Monographs on statistics and applied probability 59, 1994

[ref 5] Deutsch, C.V. and Journel, A.G. *GSLIB: Geostatistical software library and user's guide.* New York, Oxford University Press, 1998

[ref 6] Dowd, P.A. *Lognormal Kriging-The general Case.* Mathematical Geology, vol. 14, No 5, 1982

[ref 7] Isaaks, E.H. and Srivastava, R.M. *(An introduction to) Applied geostatistics.* New York, Oxford, Oxford University Press, 1989

[ref 8] Journel, A.G. and C. J. Huijbregts, *Mining Geoestatistics*: Academic Press, London, England, 1978

[ref 9] Journel, A.G. *Fundamentals of Geoestatistics in Five Lessons* American geophysical Union, Washington, 1989

[ref 10] Journel, A.G. *The lognormal aproach to predicting local distributions of selective mining unit grades:* Mathematical Geology, Vol. 12, No4, 1980

[ref 11] Link, Richard F. and Koch, Jr., George S. *Some consequences of Applying Lognormal Theory to Pseudolognormal Distributions*: Mathematical Geology, Vol. 7, No2, 1975.

[ref 12] Merayo, F.G. *Fortran 90.* Paraninfo, Madrid, 1999

[ref 13] Olea, R.A. and Pawlowsky, V. *Compensating for estimation smoothing in kriging:* Mathematical Geology, Vol. 28, No 4, 1996

[ref 14] Olea, R.A. *Geoestatistics for engineers and earth scientists.*kansas Geolical Survey, 1999

[ref 15] Rendu, Jean-Michel M. *Kriging, Logarithmic kriging and conditional expectation: Comparison of theory with actual results.*1979

[ref 16] Rivoard, J. *A review of lognormal estimators for In Situ reserves*: Mathematical Geology, vol. 22, No 2, 1990

A. GSLIB

El programa usat per la realització d'aquest projecte ha estat la versió 98 del GSLIB (Geostatistical Software Library) que és el programa més difús i més complert pel que es refereix a geoestadística.
El GSLIB consisteix en un conjunt de llibreries que s'han d'executar en FORTRAN. L'usuari només ha de modificar el fitxer de paràmetres segons les seves necessitats.
Les llibreries usades en aquest projecte són les que a continuació detallem.

A.1 Sequential Gaussian Simulation (SGSIM)

Aquest programa, en principi, està pensat per simular dades a partir d'una variable regionalitzada, és a dir, es té una mostra en una zona i es vol estendre el comportament de la variable en estudi a tota la zona. El fitxer de paràmetres que hem usat és el següent:

```
START OF PARAMETERS:
                                  \file with data
1  2  0  3  5  0                  \  columns for X,Y,Z,vr,wt,sec.var.
-1.0       1.0e21                 \  trimming limits
0                                 \transform the data (0=no, 1=yes)
sgsim.trn                         \  file for output trans table
0                                 \  consider ref. dist (0=no, 1=yes)
histsmth.out                      \  file with ref. dist distribution
1  2                              \  columns for vr and wt
0.0    15.0                       \  zmin,zmax(tail extrapolation)
1       0.0                       \  lower tail option, parameter
1      15.0                       \  upper tail option, parameter
1                                 \debugging level: 0,1,2,3
sgsim.dbg                         \file for debugging output
sgsim.out                         \file for simulation output
1                                 \number of realizations to generate
50    1    1.0                    \nx,xmn,xsiz
50    1    1.0                    \ny,ymn,ysiz
1     1    1.0                    \nz,zmn,zsiz
69069                             \random number seed
4     8                           \min and max original data for sim
12                                \number of simulated nodes to use
1                                 \assign data to nodes (0=no, 1=yes)
0     3                           \multiple grid search (0=no, 1=yes),num
0                                 \maximum data per octant (0=not used)
10.0  10.0  10.0                  \maximum search radii (hmax,hmin,vert)
 0.0   0.0   0.0                  \angles for search ellipsoid
1     0.60   1.0                  \ktype: 0=SK,1=OK,2=LVM,3=EXDR,4=COLC
../data/ydata.dat                 \  file with LVM, EXDR, or COLC variable
4                                 \  column for secondary variable
1    0                            \nst, nugget effect
2    1 0.0   0.0   0.0            \it,cc,ang1,ang2,ang3
         10.0  10.0  10.0         \a_hmax, a_hmin, a_vert
```

- **Fitxer amb dades:** En cas de que no es disposi de dades inicials a partir de les quals fer la simulació, es deixa en blanc i el programa fa una simulació incondicionada d'una N(0,1).

  En cas de tenir el fitxer amb dades inicials definim també:
  - **Posició de les variables:** eix x, eix y, eix z, variable a simular, pesos, variable secundària.
  - **Límits:** Només s'usaran els valors del fitxer que estiguin dins d'aquest interval.
- **Transformació:** Si les dades es generen sense transformar s'assumeix que aquestes segueixen una distribució Normal (0,1). En cas contrari haurem de determinar els següents camps:
  - **Fitxer de sortida de les dades transformades.**
  - **Distribució de referència:** Si aquest camp pren el valor 0, la transformació es basarà en la distribució real de les dades inicials. Si pren el valor 1, la distribució serà l'especificada a continuació.
    - `Fitxer.out`
      - Variable a simular i pesos.
  - **Límits:** Mínim i màxim valor permès.
- **Nivell de debucatge:** Segons el nivell de debucatge que hi posem tindrem més o menys sortida al fitxer sgsim.dbg i l'execució serà més o menys lenta.
- **Fitxer amb el resultat de la simulació:** `sgsim.out`
- **`nsim`:** Nombre de simulacions a realitzar.
- **`nx, xmn, xsiz`:** Definició de la xarxa (eix x)

  `ny, ymn, ysiz`: Definició de la xarxa (eix y)

  `nz, zmn, zsiz`: Definició de la xarxa (eix z)
- **`Num`:** Nombre aleatori a partir del qual es realitza la simulació.
- **`ndmin, ndmax`:** Nombre mínim i màxim de punts de les dades inicials utilitzats a l'hora de simular un node de la xarxa.
- **`ncnode`:** Nombre màxim de nodes simulats a usar per la simulació d'un altre node.
- **Mètode de busqueda:**

  0: les dades son superbloc i els nodes prèviament simulats en espiral.

  1: les dades es recoloquen als nodes de la xarxa i es fa la busqueda en espiral.
- **`multgrid`:** Si pren el valor 0 la simulació es fa en espiral, en cas contrari es fa simulació de xarxa múltiple i s'especifica:
  - Nombre de refinaments.

- **noct:** Si pren el valor 0, no es fa busqueda per octant, en cas contrari especifica el nombre de dades originals a usar per octant.
- **Radis:** Radis de busqueda en la màxima i mínima direcció horitzontal i en la direcció vertical.
- **Angles:** Especifiquen la anisotropia.
- **ktype:** Tipus de krigeat a usar.
- **Nst, c0:** Nombre d'estructures del semivariograma i efecte pepita.
- **Definició detallada del semivariograma.**

A.2 2D Kriging Program (kb2d)

El krigeat lognormal el fem amb aquest programa.

```
START OF PARAMETERS:
mostra.dat                      \file with data
1   2   3                       \   columns for X, Y, and variable
-1.0e21   1.0e21                \   trimming limits
2                               \debugging level: 0,1,2,3
kb2d.dbg                        \file for debugging output
kb2d.out                        \file for kriged output
25    0  1.0                    \nx,xmn,xsiz
25    0  1.0                    \ny,ymn,ysiz
1    1                          \x and y block discretization
4    8                          \min and max data for kriging
10.0                            \maximum search radius
1    2.302                      \0=SK, 1=OK,   (mean if SK)
1   0                           \nst, nugget effect
1   1.05  0  7   7              \it, c, azm, a_max, a_min
```

- **Fitxer amb les dades**
  - **Posició de les variables:** eix x, eix y, variable a estimar.
  - **Límits:** Mínim i màxim valor permès.
- **Nivell de debucatge:** Segons el nivell de debucatge que hi posem tindrem més o menys sortida al fitxer `kb2d.dbg` i l'execució serà més o menys lenta.
- **`nx, xmn, xsiz`:** Definició de la xarxa (eix x)

  `ny, ymn, ysiz`: Definició de la xarxa (eix y)
- **`nxdis, nydis`:** Si el valor és 1 es fa un krigeat per punt. Altrament es fa per bloc i especifiquem el nombre de punts a usar per l'estimació de cada bloc.
- Nombre mínim i màxim de dades que s'usen per krigear un bloc.
- **Radis de busqueda**
- **Tipus de krigeat**
- **`Nst, c0`:** Nombre d'estructures del semivariograma i efecte pepita.
- **Definició detallada del semivariograma.**

A.3 Indicator Kriging Program (IK3D)

Amb l'ik3d el krigeat que es realitza pot ser ordinari o simple i tant de funcions continues com categòriques.

El fitxer de paràmetres usat és el següent:

```
                  Parameters for IK3D
                  *******************

START OF PARAMETERS:
1                                 \1=continuous(cdf),0=categorical(pdf)
0                                 \option: 0=grid, 1=cross, 2=jackknife
jack.dat                          \file with jackknife data
1   2   0    3                    \   columns for X,Y,Z,vr
5                                 \number thresholds/categories
0.1 0.38 0.73 1.27 2.54           \   thresholds / categories
0.09 0.3 0.5 0.7 0.91             \   global cdf / pdf
mostra.dat                        \file with data
1   2   0    3                    \   columns for X,Y,Z and variable
direct.ik                         \file with soft indicator input
1   2   0    3  4  5  6           \   columns for X,Y,Z and indicators
-1.0e21   1.0e21                  \trimming limits
2                                 \debugging level: 0,1,2,3
ik3d.dbg                          \file for debugging output
ik3d.out                          \file for kriging output
25   0    1.0                     \nx,xmn,xsiz
25   0    1.0                     \ny,ymn,ysiz
1    0.0    1.0                   \nz,zmn,zsiz
4    8                            \min, max data for kriging
10.0  10.0  10.0                  \maximum search radii
 0.0   0.0   0.0                  \angles for search ellipsoid
0                                 \max per octant (0-> not used)
0   2.5                           \0=full IK,1=Median IK(threshold num)
1                                 \0=SK, 1=OK
1    0                            \One   nst, nugget effect
1    0.03  0.0  0.0  0.0          \      it,cc,ang1,ang2,ang3
           7.0  7.0  7.0          \      a_hmax, a_hmin, a_vert
1    0                            \Two   nst, nugget effect
1    0.22  0.0   0.0   0.0        \      it,cc,ang1,ang2,ang3
           4.0   4.0   4.0        \      a_hmax, a_hmin, a_vert
1    0                            \Three nst, nugget effect
1    0.265 0.0  0.0  0.0          \      it,cc,ang1,ang2,ang3
           6.0  6.0  6.0          \      a_hmax, a_hmin, a_vert
1    0                            \Four  nst, nugget effect
1    0.2   0.0   0.0   0.0        \      it,cc,ang1,ang2,ang3
           4.0   4.0   4.0        \      a_hmax, a_hmin, a_vert
1    0                            \Five  nst, nugget effect
1    0.06  0.0   0.0   0.0        \      it,cc,ang1,ang2,ang3
           3.0  3.0  3.0          \      a_hmax, a_hmin, a_vert
```

- **Tipus de variable:** Si la variable de la qual hem de fer el krigeat indicador és continua o categòrica.
- **Nombre de categories**
- **Categories:** Hi ha d'haver `ncat` valors

- **Valors:** De la funció de distribució acumulada o de la funció de densitat.
- **Fitxer amb les dades**
  - **Posició de les variables:** eix x, eix y, eix z, variable a estimar.
- **`Fitxer.ik`:** En el cas que tinguem les categories fetes en un altre fitxer.
  - **Posició de les variables:** eix x, eix y, eix z, variables indicadores.
  - **Límits:** Mínim i màxim valor permès.
- **Nivell de debucatge:** Segons el nivell de debucatge que hi posem tindrem més o menys sortida al fitxer `ik3d.dbg` i l'execució serà més o menys lenta.
- **Fitxer amb el resultat de la simulació:** `ik3d.out` Si un node no s'estima apareix el valor –9.9999.
- **nx, xmn, xsiz:** Definició de la xarxa (eix x)
  ny, ymn, ysiz: Definició de la xarxa (eix y)
  nz, zmn, zsiz: Definició de la xarxa (eix z)
- Nombre mínim i màxim de dades que s'usen per krigear un bloc.
- **Radis:** Radis de busqueda en la màxima i mínima direcció horitzontal i en la direcció vertical.
- **Angles:** Especifiquen la anisotropia.
- **noct:** Si pren el valor 0, no es fa busqueda per octant, en cas contrari especifica el nombre de dades originals a usar per octant.
- Si `mik=0` llavors es fa un krigeat indicador complert. Si `mik=1` llavors es fa una aproximació per la mitjana del krigeat indicador on els pesos del krigeat de la categoria o els punts de tall més propers a `mikcut` s'usen per tots els punts de tall.
- Si `ktype=0` llavors es fa el krigeat simple. Si `ktype=1` llavors es fa el krigeat ordinari.
- Per cada `ncat` fa falta un semivariograma.

A.1.4 Postprocessing of IK Results (postik)

El fitxer de sortida de ik3d requereix un procés addicional abans de ser usat.

```
                  Parameters for POSTIK
                  *********************

START OF PARAMETERS:
ik3d.out                          \file with IK3D output (continuous)
postik.out                        \file for output
3   0.5                           \output option, output parameter
5                                 \number of thresholds
0.1 0.38 0.73 1.27 2.54           \the thresholds
0   1      0.75                   \volume support?, type, varred
mostra.dat                        \file with global distribution
3   0    -1.0    1.0e21           \   ivr,  iwt,  tmin,  tmax
0.0    30.0                       \minimum and maximum Z value
1   1.0                           \lower tail: option, parameter
1   1.0                           \middle    : option, parameter
1   2.0                           \upper tail: option, parameter
100                               \maximum discretization


option 1 = E-type
       2 = probability and mean above threshold(par)
       3 = Z percentile corresponding to (par)
       4 = conditional variance
```

- **Fitxer amb les dades:** És el fitxer de sortida de ik3d
- **Fitxer de sortida:** ~.out
- **iout i outpar:** Segons el valor iout obtindrem un tipus d'output o un altre.
  - iout = 1 : Mitjana de la distribució i la variància condicionada
  - iout = 2 : Probabilitat i mitjana sobre i sota el valor outpar
  - iout = 3 : Valor Z (p quantil) corresponent al valor de la funció de distribució acumulada outpar = p
  - iout = 4 : Calcula la variància condicionada.
- **Nombre de categories**
- **Categories** : Les mateixes que a ik3d
- **ivol:** Factor de correcció.
- **tmin, tmax:** Els valors que surten fora d'aquest interval s'ignoren
- **zmin, zmax:** Els valors mínim i màxim permesos.
- **ltail:** Especifica l'extrapolació a la cua inferior de la distribució.
  - **ltail=1:** Interpolació lineal
  - **ltail=2:** Interpolació per model exponencial amb w=ltpar

  - **ltail=3:** Interpolació lineal entre quantils.
- **middle:** Especifica l'extrapolació al mig de la distribució.
  - **middle=1:** Interpolació lineal
  - **middle=2:** Interpolació per model exponencial amb w=`midpar`
  - **middle=3:** Interpolació lineal entre quantils.
- **utail:** Especifica l'extrapolació a la cua superior de la distribució.
  - **utail=1:** Interpolació lineal
  - **utail=2:** Interpolació per model exponencial amb w=`utpar`
  - **utail=3:** Interpolació lineal entre quantils.
- **maxdis:** Paràmetre de discretització màxima. El valor de 50 és el més tipic.

B. GEO-EAS (Version 1.2.1)

El GEO-EAS (Geoestatistical Environmental Assessment Software) és un programa de software interactiu que s'utilitza per tal de realitzar anàlisis geoestadistics de dades espaials.

Nosaltres utilitzarem aquest programa per tal de poder calcular els semivariogrames que necessitarem per a poder realitzar els diferents mètodes de krigeat.

Les dues aplicacions del GEO-EAS que es fan servir en el projecte són el PREVAR i el VARIOGRAMA , que passem a detallar a continuació:

PREVAR: L'aplicació PREVAR ens proporciona un fitxer amb extensió .PCF (Pair Comparison File) que conté la comparació dels parells de dades de la mostra a utilitzar. Aquest fitxer serà utilitzat en l'aplicació del VARIO per calcular els semivariogrames.

- Les distàncies i direccions son calculades per a cada parell de punts mostrejats del fitxer d'entrada i son escrites en el fitxer PCF.

- Els límits dels valors coordinats i les distàncies a tenir en compte a l'hora de buscar els parells de punts son imposades per l'usuari per tal de limitar la grandària del fitxer PCF.

VARIO: Aquesta aplicació calcula i gràfica els semivariogrames. Les dades son llegides del fitxer PCF produït pel PREVAR.

- El fitxer PCF conté les distàncies, direccions i parell de punters per tots els parells de punts de la mostra 2D en un fitxer en format Geo-Eas.
- S'han de decidir les toleràncies per els parells de direccions i la distància entre els intervals de lags (que en el nostre cas serà de 0.5).
- Es podem realitzar gràfics de punts, de barres i boxplots.
- Podem ser llistats els resultats individuals per a una determinada classe (i.e. distància entre punts)
- Podem realitzar l'aproximació a traves de 4 models de semivariograma (exponencial, esfèric, Gaussià i lineal)
- Els resultats intermedis podem ser guardats en un fitxer amb format Geo-Eas.

El fitxer d'entrada que ens demana el Geo-Eas per a realitzar els semivariogrames és un fitxer que conté les dades i les corresponents localitzacions en l'espai, en format Geo-Eas.

Decisions preses:

Aquest programa, com hem pogut veure també pot aproximar models als semivariogrames obtinguts però hem preferit aplicar les formules directament per als diferents models en un programa a part per tal de que clarificar d'on provenia cada un dels valors utilitzats en la realització de la modelització del semivariograma experimental.

Escollim la distància entre lags igual a 0.5. Cada lag és la distància a la que es buscaran parells de punts, així tindrem les següents distàncies de busqueda:
1, 1.5,2,...
És a dir, en primer lloc ens buscarà tots els parells de punts separats a una distància de 1, després els parells de punts separats per 1.5 unitats i així successivament.

C. EXCEL

Un cop hem obtingut els valors per a realitzar els semivariogrames experimentals a partir del Geo-Eas, utilitzarem el full de càlcul EXCEL per tal de poder fer un ajust del model de semivariograma més apropiat.

Per fer-ho aplicarem les dues formules de semivariograma més freqüents en aquest tipus d'estudi, el semivariograma exponencial i el semivariograma esfèric (veure capítol 2.2) i anirem utilitzant diferents valors de meseta i d'abast i graficant-los fins a obtenir un model que ajusti les dades.
Ho farem de la següent manera:
Creem dues columnes amb les dades obtingudes amb el Geo-Eas: Una contindrà els valors de les h (lags) , és a dir, l'eix de les x's, i l'altra el valor del semivariograma per cada lag.
A continuació generarem dues columnes: la columna corresponent a l'abast (a) i la corresponent a la meseta (c) amb el mateix valor per totes les files. Finalment crearem les dues últimes columnes que seran l'aplicació de les fórmules pels dos models:

Model esfèric

$$\gamma(h)=\begin{cases} C\left[1.5\frac{|h|}{a}-0.5\left(\frac{|h|}{a}\right)^3\right] & \text{si } |h|\le a \\ C & \text{si } |h|>a \end{cases}$$

Model exponencial:

$$\gamma(h)=C\left[1-\exp\left(-\frac{|h|}{a}\right)\right]$$

C: meseta que anirem variant per trobar l'ajust òptim.
a: abast que també anirà variant fins a ajustar-se a les dades.
h: Diferents Lags o distàncies de separació entre els parells de variables. Utilitzarem la columna generada pel Geo-Eas amb els seus valors.

Un cop realitzades les operacions de forma automàtica, graficarem els resultats i anirem modificant els valors per tal de trobar l'ajust correcte.

Aquesta operació s'ha dut a terme 6 cops: 5 per els semivariogrames experimentals dels cut-off per a poder realitzar el krigeat indicador i un darrer pel semivariograma experimental de la mostra per a poder realitzar el krigeat lognormal.

D. FORTRAN

El compilador de FORTRAN l'hem utilitzar per poder executar el GSLIB i també per calcular l'stress.

El programa que hem usat per l'stress és el següent:

```
PROGRAM STRESS

        real dimension X(625), Y(625), Z(625)
        real i,j,k1,k2,k3,s1,s2,s3,s12,s13,s23

        open (1, file='C:\4000\stress\dades.DAT')
         DO i=1,625,1
            read (1,*) X(i),Y(i),Z(i)

            do j=i+1,625,1
                     k1=|X(i)-X(j)|
                     k2=|Y(i)-Y(j)|
                     k3=|Z(i)-Z(j)|
                     s1=s1+k1**2
                     s2=s2+k2**2
                     s3=s3+k3**2
                     s12=s12+(k1-k2)**2
                     s13=s13+(k1-k3)**2
                     s23=s23+(k2-k3)**2
            end do

        end do
        close(1)
        write(*,*) 'k1=',k1,'k2=',k2,'k3=',k3
        write(*,*) 's1=',s1,'s2=',s2,'s3=',s3
        write(*,*) 's12=',s12,'s13=',s13,'s23=',s23
        pause
end
```

El fitxer d'entrada consta de tres columnes: En la primera hi ha les dades corresponents a la realització del krigeat lognormal transformades a N(0,1), en la segona hi podem trobar les realitzacions provinents del krigeat indicador també transformades a N(0,1). En la tercera i última columna situarem les dades que simulen la realitat també transformades.

Els resultats d'aquest programa son els sumatoris que s'han d'utilitzar per tal de calcular l'STRESS:

$$s1 = \sum_{i<j} \left(\delta^*_{ij}\right)^2$$

$$s2 = \sum_{i<j} \left(\hat{\delta}_{ij}\right)^2$$

$$s3 = \sum_{i<j} \left(\delta_{ij}\right)^2$$

$$s12 = \sum_{i<j} (\delta^*_{ij} - \hat{\delta}_{ij})^2$$

$$s13 = \sum_{i<j} (\delta_{ij} - \delta^*_{ij})^2$$

$$s12 = \sum_{i<j} (\delta_{ij} - \hat{\delta}_{ij})^2$$

i s'apliquen de la següent manera.

$$STRESS = \sqrt{\frac{\sum_{I<J} \left(\delta_{IJ} - \hat{\delta}^*_{IJ}\right)^2}{\sum_{I<J} \left(\hat{\delta}^*_{IJ}\right)^2}}$$