UNIVERSITA' DEGLI STUDI DI MILANO
BICOCCA

# SCUOLA DI ECONOMIA E STATISTICA

**STATISTICA SPAZIALE**

**Prova di Laboratorio**

**Appello del 28/06/2017**

| | |
|---|---|
| **Cognome** | **Nome** |
| **Numero di Matricola** | |

**Esercizio 1** (p. 19)

Nel file dat.csv sono presenti le seguenti variabili:

`x: ascissa del punto di misura`

`y: ordinata del punto di misura`

`valori: concentrazione in mg/L di una sostanza`

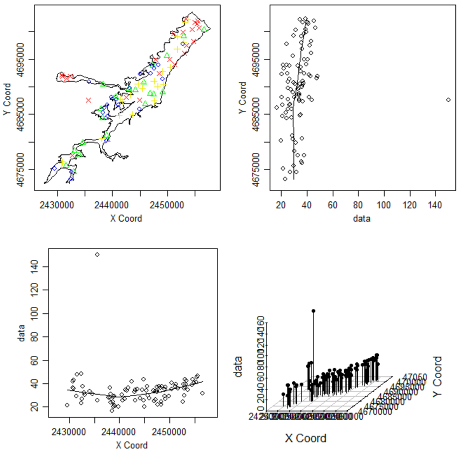
I file `PE.shp`, `PE.shx`, `PE.dbf` riportano i confini geografici della regione di studio a cui i dati sono riferiti che rappresenta un bacino collocato in una regione italiana.

1. Si esegua una sintetica analisi esplorativa dei dati, evidenziandone le caratteristiche principali. Si dica quale potrebbe essere il sistema di riferimento geografico utilizzato e il significato delle coordinate dei punti di misura. Si riportino i risultati nello spazio predisposto [2]
2. Si calcoli il variogramma empirico della variabile "valori" usando 16 distanze (bin) per la sua costruzione. Se ne produca il grafico e si trascrivano, negli spazi sotto riportati, i valori del variogramma empirico e i valori delle corrispondenti distanze, approssimati al primo decimale. Si discuta sinteticamente come sono trattate eventuali inconsistenze nei dati. Si commenti il variogramma ottenuto. (4)
3. Si esegua un'opportuna diagnostica grafica di tipo Monte Carlo per evidenziare se, nel dataset considerato, la dipendenza spaziale risulta significativamente rilevante oppure no. Si riporti e si commenti il grafico nello spazio predisposto. Ai fini dell'analisi Montecarlo si inizializzi la generazione casuale con il valore 17628 e si eseguano 80 simulazioni indipendenti per implementare la diagnostica. Si riportino, nello spazio preposto, **tutte e sole** le istruzioni R necessarie a questa analisi. (3)
4. Si stimi il variogramma tramite un modello sferico usando i minimi quadrati ordinari, fissando la massima distanza a cui calcolare il variogramma empirico a 35000 e inizializzando la procedura con il valore 25 per la soglia parziale, 10 per il nugget e 5000 per il range. Si riportino i valori stimati dei parametri: range, soglia e nugget. Si calcoli il nugget relativo in termini percentuali interpretando sinteticamente il valore ottenuto. Si scriva l'equazione del modello di semivariogramma stimato. Si riportino tutti i valori approssimati al primo decimale. (4)
5. Si riporti il grafico del modello stimato e del variogramma empirico. Si riporti le istruzioni R utilizzate per la stima e il grafico del variogramma. (2)
6. Usando le stime ottenute al punto 4, si riporti nello spazio di seguito predisposto la mappa ottenuta tramite kriging ordinario corredata da un'opportuna legenda usando una griglia regolare 50×50 per discretizzare la superficie. Si rappresenti solamente la superficie delle concentrazioni all'interno dello shape file della regione di studio. Si riportino sulla mappa i punti di misura del campione. Nel caso non si sia in grado di stimare il variogramma tramite OLS si utilizzino per la previsione i valori iniziali indicati al punto 4. Si riporti nello spazio indicato il codice utilizzato per questa analisi. (4)

**Svolgimento** (aumentare lo spazio dedicato alla risposta dove necessario)

## Punto 1

### Riportare di seguito i risultati dell'analisi esplorativa



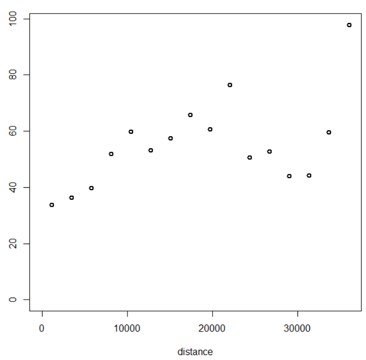

| x | y | valori |
|---|---|---|
| Min. :2429574 | Min. :4673244 | Min. : 16.64 |
| 1st Qu.:2438615 | 1st Qu.:4686067 | 1st Qu.: 26.85 |
| Median :2444961 | Median :4690243 | Median : 33.85 |
| Mean :2444007 | Mean :4689813 | Mean : 33.93 |
| 3rd Qu.:2449732 | 3rd Qu.:4694112 | 3rd Qu.: 37.73 |
| Max. :2456584 | Max. :4702377 | Max. :150.00 |

Il dataset presenta un evidente outlier e un punto esterno alla regione di studio. Tale punto esterno coincide proprio con l'outlier. Data la doppia inconsistenza di questo dato è opportuno rimuoverlo dalle osservazioni usate per le analisi successive. Esclusa tale osservazione non si evidenziano altre criticità.

Dall'ordine di grandezza delle coordinate geografiche si intuisce che si tratta di coordinate cartografiche di tipo Gauss-Boaga che esprimono in metri la distanza dall'Equatore (ordinata) e dal parallelo di riferimento (ascissa) a meno di una traslazione (false easting).

## Punto 2

### Riportare di seguito il grafico del variogramma empirico




### Riportare di seguito il commento sul grafico precedente

Nel grafico è riportato il variogramma empirico dopo aver eliminato l'outlier dal dataset. Il valore del nugget è intorno a 30 della soglia intorno a 60. Non ci sono particolari evidenze di outlier o trend. L'andamento più altalenante dei bin finali è potenzialmente attribuibile al basso numero di coppie campionarie presenti (nell'ultimo bin ce ne sono 37)

Riportare di seguito il valore del variogramma empirico stimato e le relative distanza

Riportare di seguito i valori di $2\hat{\gamma}(h)$

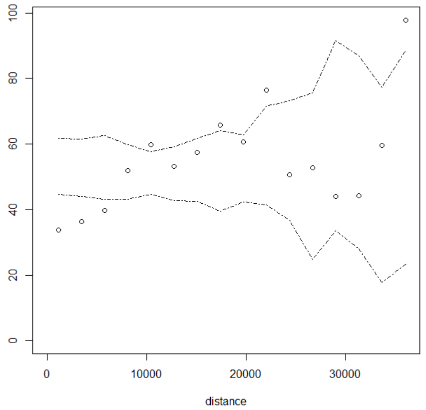
```
33.8 36.3 39.8 51.9 59.7 53.3 57.5 65.7 60.7 76.5 50.6 52.7 44.1 44.2 59.7 97.9
```

Riportare di seguito i valori di *h*

```
1162.5  3487.4  5812.4  8137.3 10462.3 12787.2 15112.2 17437.1 19762.1 22087.0 24412.0
26736.9 29061.9 31386.8 33711.8 36036.7
```

## Punto 3

Riportare di seguito il grafico richiesto




Riportare di seguito l'interpretazione del grafico

… …

Riportare di seguito il codice richiesto

```
dat=read.table("dat.csv", header=T, sep=";")
dat=dat[dat$valori!=max(dat$valori),]
dgeo <- as.geodata(dat, coords.col=1:2, data.col=3)
set.seed(17628)
dat.om.env <- variog.mc.env(dgeo, obj.var = vario,nsim=80)
plot(vario, envelope = dat.om.env)
```

## Punto 4

Riportare di seguito le stime dei parametri del modello esponenziale

| nugget | soglia parziale | range | nuggget relativo |
|---|---|---|---|
| 27.8 | 28.4 | 13643.4 | 49.4% |

Equazione del modello (semivariogramma)

$$\gamma(h) = 27.5 + 28.4\left(1.5\,\frac{h}{13643.4} - 0.5\left(\frac{h}{13643.4}\right)^3\right) \quad \text{se } 0 < h < 13643.4$$

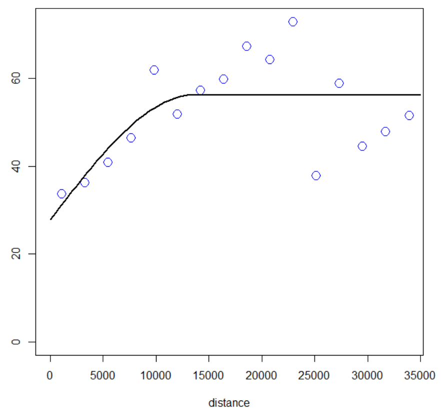
$$\gamma(h) = 56.2 \quad \text{altrimenti}$$

Riportare di seguito le interpretazioni per le stime richieste (nugget relativo)

… … …

## Punto 5

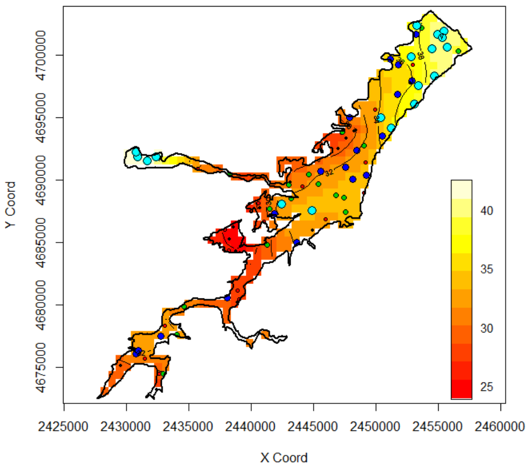
Riportare di seguito il grafico del variogramma stimato




```
varioTr   <-   variog(dgeo,   estimator.type="classical",uvec=16,max.dist=35000)
plot(varioTr, main="",col="blue",cex=1.8)
variosf<-variofit(varioTr, ini.cov.pars=c(25,5000), weights="equal",
                  cov.model="spherical", fix.nugget=FALSE, nugget=10)
lines(variosf,lwd=2)
```

## Punto 6

Riportare di seguito la mappa di previsione



## Riportare di seguito il codice richiesto

```
poly<-readShapePoly("PE.shp",verbose=TRUE)
X=bbox(poly)[1,]
Y=bbox(poly)[2,]
size=50
YY<-round(seq(ceiling(min(Y)),floor(max(Y)), length=size),2)
XX<-round(seq(ceiling(min(X)),floor(max(X)), length=size),2)
griglia<-expand.grid(X=XX,Y=YY)
krg.or <- krige.conv(geodata=dgeo, loc=griglia,
                     krige=krige.control(cov.pars=varioexp$cov.pars,
                                         cov.model="spherical",
                                         nugget=varioexp$nugget),
                     borders=p)
image(krg.or)
legend.krige(x.leg=c(2456000, X[2]), y.leg=c(Y[1],4690000), val=krg.or$predict,
             vert=TRUE, off=0.7, cex=0.9, scale.vals=pretty(krg.or$predict)
       )
contour(krg.or,add=T)
points.geodata(dgeo,add=T,pt.divide="quintiles", col=1:5)
```

**Esercizio 2** (p.12)

Il file `datpp.csv` riporta dati relativi ad un point pattern realizzatosi in una regione quadrata di lato 10 e vertice basso a sinistra sull'origine degli assi cartesiani. Il file contiene due colonne:

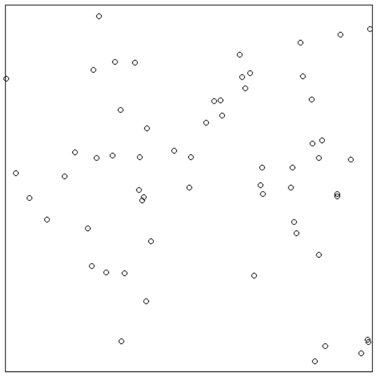
```
x:      ascissa dell'evento
y:      ordinata dell'evento
```

1. Si rappresenti su un grafico il point pattern e la relativa finestra e lo si riporti nello spazio indicato. (2)
2. Si riporti il numero di eventi del point pattern e si stimi l'intensità del processo di punto supponendo che il processo sia di Poisson omogeneo. Si interpreti il valore ottenuto. (2)
3. Si esegua un test per la presenza di CSR contro un'ipotesi alternativa che il processo abbia una struttura clusterizzata. SI utilizzi il metodo dei Quadrat Count usando un insieme di 25 celle ottenute suddividendo il range delle ascisse e quello delle ordinate in 5 classi. Si riporti il risultato ottenuto negli spazi predisposti e lo si interpreti. Si riporti l'istruzione R utilizzata. (3)
4. Si riporti il grafico dell'intensità stimata tramite il metodo kernel ponendo pari a 5 la finestra di lisciamento sia per i valori di ascissa che per quelli di ordinata. Si aggiungano al grafico gli eventi osservati in colore bianco. Si riporti la mappa nell'area predisposta del presente file. Si riporti l'istruzione R utilizzata. (3)
5. Tra i due metodi usati per stimare l'intensità (quello al punto 2. e quello al punto 4.) quale ritenete essere il più adeguato e perché? Riportare la risposta nello spazio preposto (2)

**Svolgimento** (aumentare lo spazio dedicato alla risposta se necessario)

### Punto 1

Riportare di seguito la mappa del point pattern



### Punto 2

Riportare di seguito il valore dell'intensità stimata e l'interpretazione di tale valore

Numero eventi= 61

Intensità stimata = 0.61 punti per unità di superficie

Riportare di seguito l'interpretazione dell'intensità stimata

… … …

### Punto 3

Riportare di seguito il test per la CSR

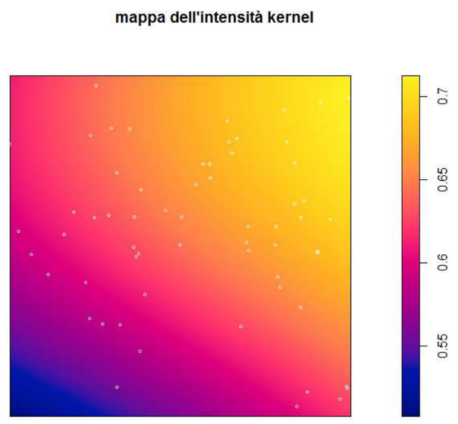
```
Statistica test osservata = 44.238
df = 24
p-value = 0.006987

te0<-quadrat.test(pp,5, alternative= "clustered"); te0
```

La struttura CSR risulta poco compatibile con i dati osservati e quindi la si respinge in favore di una struttura a cluster.

### Punto 4

Riportare di seguito il grafico della stima kernel dell'intensità




```
Z <- density.ppp(pp, 5)
plot(Z,main="mappa dell'intensità kernel");
plot(pp,add=T,cex=0.6,col="white")
```

## Punto 5

Riportare di seguito il commento richiesto

… …