

O3 Pollution

Di Prospero, Iezzi, Valentini

Introduzione

In questo elaborato verrà affrontato un problema di analisi di **dati spazialmente continui o geostatistici**, con l'obiettivo finale di pervenire alla ricostruzione del fenomeno di interesse nell'intero dominio oggetto di studio, ossia prevedere il valore assunto dal fenomeno stesso in corrispondenza di siti non campionati sfruttando le informazioni presenti all'interno di un campione di siti in cui il dato è stato invece campionato. Nel caso che verrà affrontato, il fenomeno d'interesse è rappresentato dal livello di **OZONO** rilevato nel 2014 negli **USA** (dominio oggetto di studio). Per pervenire alla previsione del livello di ozono in corrispondenza di tutta la regione oggetto di studio verrà affrontato un problema di **interpolazione** dei dati disponibili per stimarne i valori dove non si hanno misurazioni. Il metodo di interpolazione più usato nell'ambito dei dati spazialmente continui o geostatistici è il **kriging**. Nel kriging, questa interpolazione spaziale si basa sull'autocorrelazione, cioè l'assunto che il fenomeno di interesse vari nello spazio con continuità. Pertanto, si considera la **legge di Tobler**, in base al quale “osservazioni prese da siti vicini tendono ad essere più simili di osservazioni prese a siti distanti”. Sulla base di tale ipotesi ci si attende quindi che la dipendenza spaziale si indebolisca all'aumentare della distanza considerata.

Descrizione del dataset pollution e selezione del dato d'interesse

Il set di dati “**Pollution**” contiene 50428 osservazioni, ciascuna delle quali si riferisce al valore rilevato di un determinato inquinante in corrispondenza di un certo sito e relativamente ad un determinato anno di riferimento. Per ciascuna osservazione vengono rilevate 6 variabili.

```
rm(list=ls())
library(readr)
library(knitr)
library(dplyr)
library(geoR)
library(ggplot2)
library(maps)
library(viridis)
library(akima)

Polluttion <- read_csv("C:/Users/Admin/OneDrive/Desktop/Polluttion.csv")
```

Per meglio descrivere la composizione del Dataset mostriamo le prime sei righe che lo caratterizzano:

Year	Site	Longitude	Latitude	pollutant	mean.obs.value
2005	100010002	-75.5568	38.9867	O3	0.0494486
2005	100010002	-75.5556	38.9847	PM25	13.1458333
2005	100010003	-75.5181	39.1550	EC	0.5897667
2005	100010003	-75.5181	39.1550	NH4	1.9013000
2005	100010003	-75.5181	39.1550	NO3	1.9445500
2005	100010003	-75.5181	39.1550	OC	1.6482184

Nella prima colonna (Year) vengono riportati gli anni di rilevazione. Nella seconda colonna (Site) sono riportate le labels che identificano i siti in cui gli inquinanti sono stati rilevati. Nella terza e nella quarta colonna sono riportate la longitudine (Longitude) e la latitudine (Latitude), ossia le coordinate spaziali dei siti in cui vengono rilevati i valori degli inquinanti. Nella quinta colonna (pollutant) sono riportate le labels relative al tipo di inquinante considerato. Infine, la sesta colonna (mean.obs.value) contiene il valore rilevato del tipo di inquinante considerato, e pertanto rappresenta il dato d'interesse Z. In questo elaborato il fenomeno di interesse che si intende analizzare è rappresentato dal **livello di ozono** rilevato nel **2014** nei diversi stati degli **USA**. Pertanto, mediante i comandi che seguono si procede ad una selezione delle righe del Dataset **“Pollutant”** relative all'anno 2014 e al livello di ozono, che presenta lable “O3”, per poi creare un DataFrame contenente tali dati (che verrà denominato **“dat”**) mediante l'utilizzo della funzione filter() che prende in input: - il Dataset da cui selezionare i dati di interesse (“Pollution”); - le condizioni di selezione delle righe (Year == 2014 , pollutant == “O3”). Si procede inoltre col moltiplicare per 100 il dato di interesse, in modo tale da migliore la sua visualizzazione, e col creare una matrice denominata **“s”** contenente le coordinate dei siti spaziali.

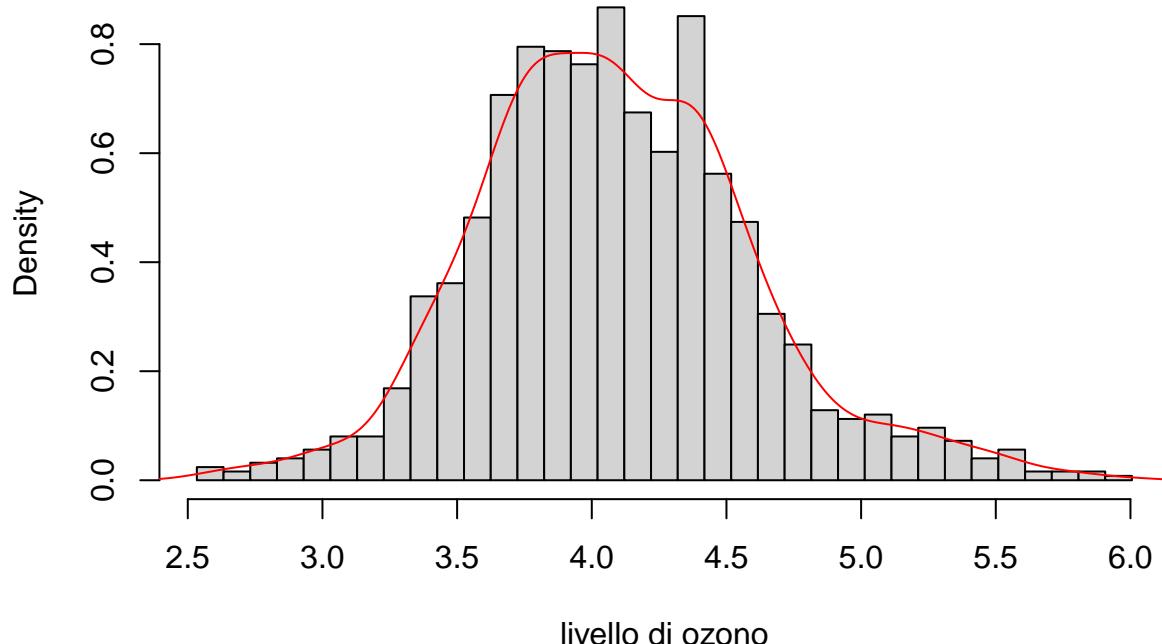
```
dat <- filter(Pollutution, Year == 2014, pollutant == "O3")
Y <- 100*dat[,6]
Y <- as.numeric(unlist(Y))
s <- as.matrix(dat[,3:4])
```

Analisi esplorativa dei dati

Il primo passo dell'analisi esplorativa dei dati selezionati consiste nello studio della distribuzione della variabile contenente i valori rilevati di ozono mediante una rappresentazione grafica del relativo istogramma rappresentante le densità di frequenza, su cui viene sovrapposta anche una curva rappresentante la funzione di densità della variabile stessa:

```
nclassi <- round(sqrt(length(Y)))                                # numero di classi
classi <- seq(min(Y), max(Y), length=nclassi+1) # vettore delle classi
hist(Y,freq=FALSE, breaks=classi ,                               # istogramma
     main = "Istogramma e curva di densità relativa al livello di ozono" ,
     xlab = "livello di ozono")
lines(density(Y) , col= "red")                                    # f. di densità
```

Istogramma e curva di densità relativa al livello di ozono



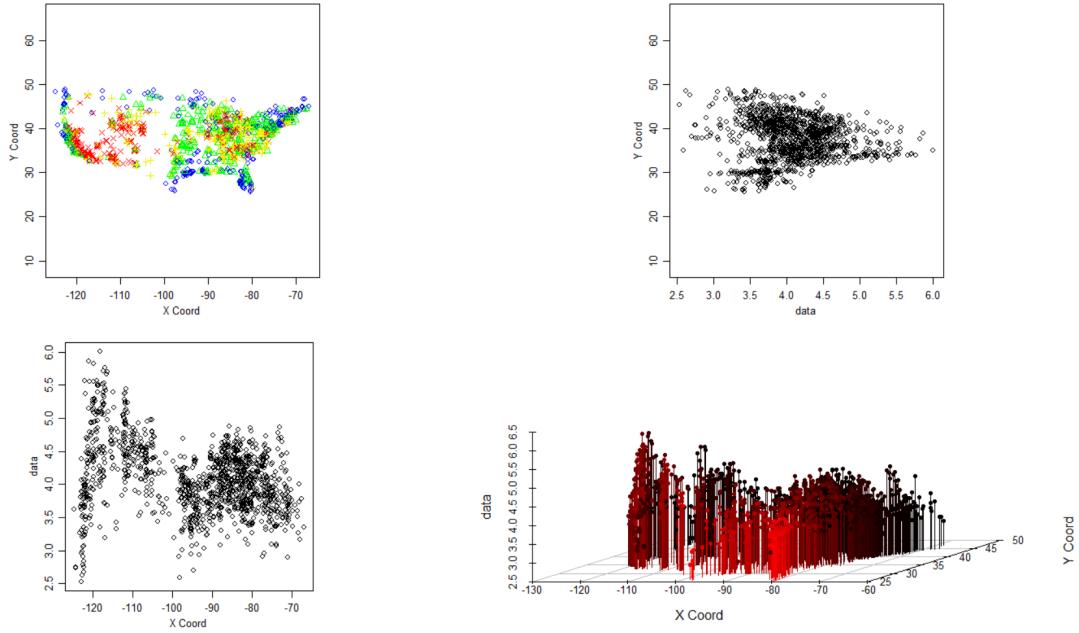
Possiamo notare che la rappresentazione dell’istogramma relativo al livello di ozono risulta essere approssimativamente simmetrica, ed in particolare osservando la parte centrale della distribuzione possiamo notare la presenza di 2 picchi. L’istogramma nel suo complesso presenta quindi una forma a campana quasi simmetrica, con 2 picchi nella parte centrale della distribuzione. La curva di densità si presenta anch’essa quasi simmetrica e quindi la distribuzione può approssimativamente essere considerata di tipo Gaussiana (Normale Gaussiana).

Procediamo con l’analisi esplorativa dei dati creando l’oggetto “data” di tipo Data-Frame contenente le coordinate spaziali dei siti ed il dato d’interesse **Z**, vale a dire il livello di ozono:

```
data <- data.frame(coord = s[,1:2], Y=Y)
data <- as.geodata(data)
```

La creazione di tale oggetto ci permette di pervenire alla definizione di una variabile di tipo **geodata** mediante la funzione `as.geodata()`, la quale richiede una sequenza precisa delle colonne che dovranno essere così ordinate: coordinata x, coordinata y e dato z. Combinando tale funzione con la funzione `Plot()` otteniamo una rappresentazione grafica sommaria della variabile di tipo **geodata**, che ci permette di osservare la distribuzione dei siti spaziali sulla regione di interesse e la distribuzione relativa al valore di ozono rilevato in corrispondenza di tali siti spaziali:

```
par(mfrow=c(1,1))
plot(data, scatter3d=TRUE, highlight=TRUE)
```



L'output ottenuto consiste in un panel di 4 Subplot. Nel primo Subplot (in alto a sinistra) possiamo osservare la disposizione dei siti spaziali rispetto alle coordinate x e y e il relativi valori, suddivisi in classi di appartenenza definite da differenti colori, della presenza di ozono. Nel secondo e nel terzo Sublot (in alto a destra ed in basso a sinistra), i valori rilevati di ozono sono posti in funzione delle coordinate spaziali x ed y; osservando tali subplot non si evidenziano trend. L'ultimo subplot consiste in uno Scatterplot in 3D in cui sul piano vengono rappresentate le coordinate spaziali X ed y caratterizzanti ciascun sito, mentre sulla terza dimensione viene riportato il valore di ozono rilevato su ciascun sito campionario. Osservando il grafico si nota la presenza di un andamento non costante che caratterizza la distribuzione spaziale del livello di ozono. Pertanto il processo non è stazionario in media.

Nel seguente grafico effettuiamo una rappresentazione grafica del variogramma campionario andando a suddividere la nuvola di punti in classi (bin). All'interno di ogni classe si calcola la media delle differenze in valore al quadrato tra due siti, per ottenere in tal modo un valore rappresentativo.

```

pollutant.cloud1 <- variog(data , option = "cloud")

## variog: computing omnidirectional variogram

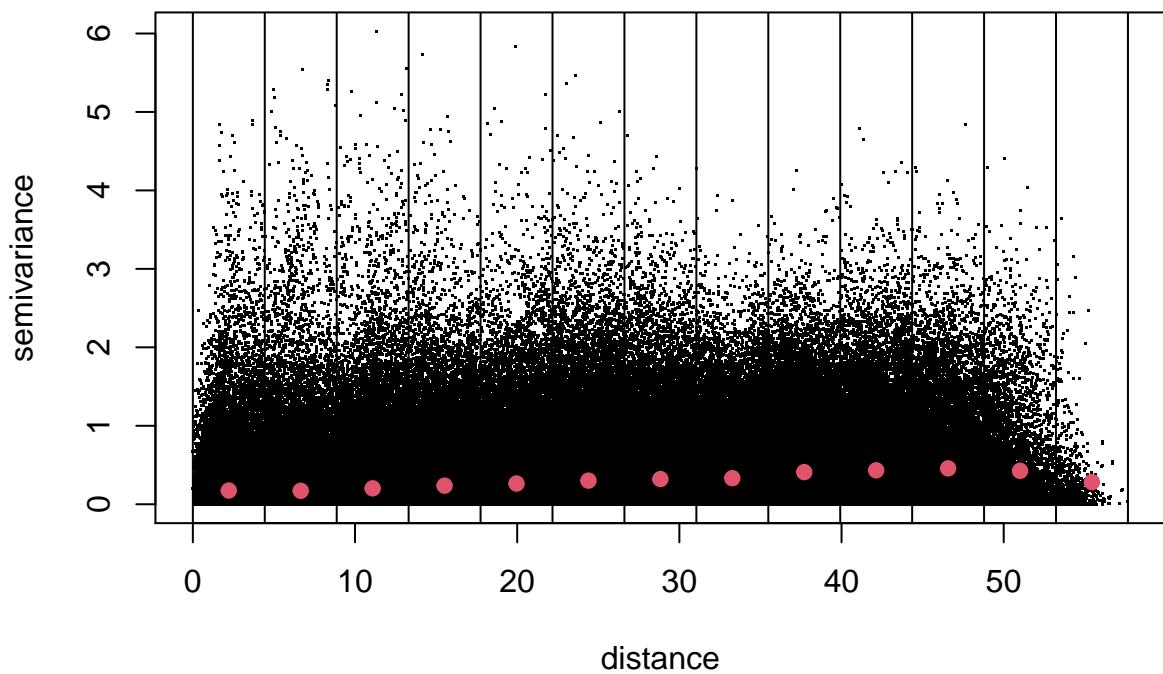
plot( pollutant.cloud1 , main='Variogramma campionario diviso in classi di distanza',
      pch='.',cex=0.8) # variogramma empirico
pollutant.bin<-variog(data,option="bin")

## variog: computing omnidirectional variogram

points( pollutant.bin$u , pollutant.bin$v ,pch=19 , col=2) # calcola la media in ogni classe
abline(v= pollutant.bin$bins.lim) # abline serve per rappresentare le classi

```

Variogramma campionario diviso in classi di distanza

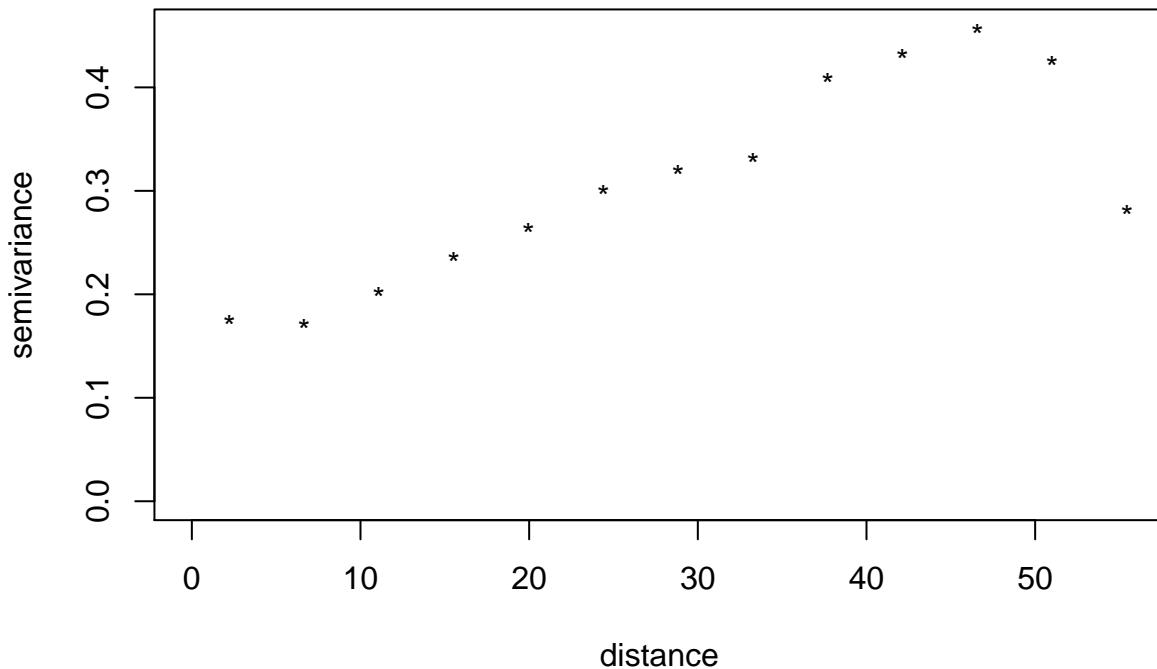


Nella rappresentazione grafica ottenuta i puntini rossi rappresentano le dissimilarità medie per ogni classe. Osservando l'andamento di tali punti si nota che è presente una bassa dissimilarità.

L'analisi esplorativa prosegue con lo studio della struttura di correlazione spaziale dei nostri dati, che viene stimata ricorrendo al variogramma empirico. In prima approssimazione ipotizziamo che il processo sia **isotropico** e calcoliamo il variogramma empirico mediante la funzione variog():

```
pollutant.cloud2 <- variog(data) # variogramma campionario  
## variog: computing omnidirectional variogram  
plot(pollutant.cloud2 , main = "variogramma empirico" , pch = "*") # plottiamo il variogramma campionario
```

variogramma empirico



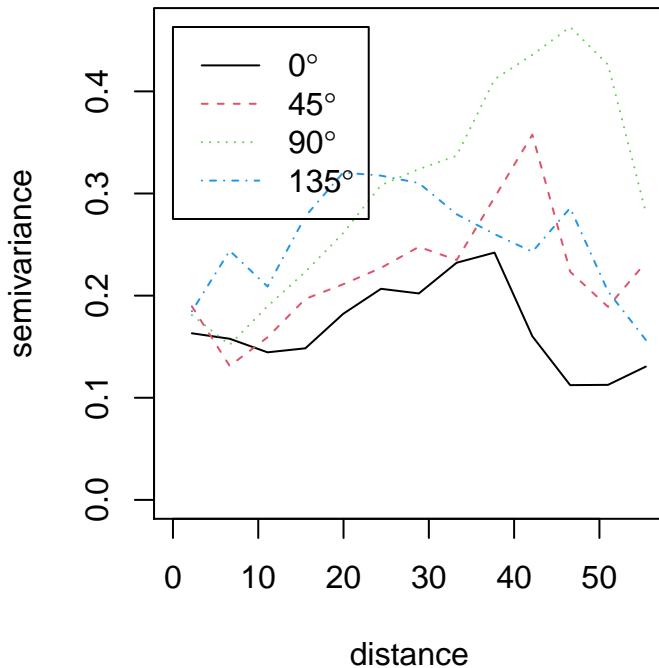
Da un punto di vista teorico il variogramma è una funzione che parte dall'origine per poi crescere fino ad un livello massimo (detto **sella**) in corrispondenza del quale si assesta. Per cui dopo una certa distanza h (detta **range**) il variogramma tende ad essere parallelo all'asse delle ascisse. Il range rappresenta quindi la distanza oltre la quale non dovrebbe essere presente correlazione spaziale. Pertanto, tutti i siti che si troveranno a distanze maggiori del punto di range avranno un contributo minimo o nullo nella previsione. Da un punto di vista campionario possiamo osservare che il comportamento del variogramma empirico è diverso da quanto ipotizziamo di aspettarci: esso infatti tende a seguire un andamento crescente (fino ad una distanza pari a circa 50) e a non stabilizzarsi. Ciò è dovuto al fatto che, come è stato già osservato, esiste un trend nei dati (variabilità di larga scala). Inoltre, il variogramma non parte dall'origine ma da un valore di semivarianza pari a circa 0.2; ciò può essere dovuto alla presenza di un errore legato allo strumento di misura (cioè un Nugget). Ne consegue quindi che il **processo non è stazionario in media**. Analizziamo ora il variogramma nelle quattro direzioni principali (N-S, NE-SW, E-W e SE-NW) per verificare se il processo è isotropico o meno:

```
par(mfrow=c(1,1),pty='s')
pollutant.bin4<-variog4(data)

## variog: computing variogram for direction = 0 degrees (0 radians)
## tolerance angle = 22.5 degrees (0.393 radians)
## variog: computing variogram for direction = 45 degrees (0.785 radians)
## tolerance angle = 22.5 degrees (0.393 radians)
## variog: computing variogram for direction = 90 degrees (1.571 radians)
## tolerance angle = 22.5 degrees (0.393 radians)
## variog: computing variogram for direction = 135 degrees (2.356 radians)
## tolerance angle = 22.5 degrees (0.393 radians)
## variog: computing omnidirectional variogram
```

```
plot(pollutant.bin4)
title(main='Variogramma campionario nelle quattro direzioni :\n N-S, NE-SW, E-W e SE-NW')
```

Variogramma campionario nelle quattro direzioni : N-S, NE-SW, E-W e SE-NW



Osservando l'output ottenuto, si potrebbe ipotizzare che il processo sia **anisotropico**: infatti la struttura del variogramma varia al variare della direzione considerata. La rappresentazione grafica suggerisce che bisognerebbe tener conto della direzionalità per calcolare la correlazione tra i siti.

Modellazione della variazione di primo ordine (o di larga scala) ed analisi di variogramma sui residui spaziali

Una delle condizioni necessarie per poter sfuggire il predittore di tipo lineare del Kriging consiste nel rispettare le ipotesi di stazionarietà del secondo ordine (stazionarietà in media, stazionarietà in covarianza). Dallo scatterplot in 3D osservato nel corso dell'analisi esplorativa possiamo affermare che IL PROCESSO NON RISULTA ESSERE STAZIONARIO IN MEDIA. in quanto abbiamo osservato che la media non si mantiene costante al variare dei siti (e quindi lungo il dominio di interesse). In altre parole, la distribuzione spaziale del livello di ozono si caratterizza per presentare un trend che pertanto dovrà essere rimosso. Si procede quindi alla determinazione di un modello deterministico per poter rappresentare la variabilità di larga scala (trend del processo). Successivamente, dopo essere perventi alla definizione di un modello deterministico (e quindi un iperpiano di regressione) andiamo a sottrarre il trend ai dati originali identificando in tal modo i residui del modello. Una volta ricavati i residui del modello occorrerà verificare se sono correlati spazialmente oppure rispettano le ipotesi di base del teorema di Gauss-Markov. Se i residui del modello risultano essere correlati spazialmente sarà necessario cogliere l'informazione inerente la struttura di correlazione spaziale, per poterla poi sfruttare durante la fase di previsione mediante Kriging. Predisponiamo quindi un analisi di regressione. Con le seguenti righe di codice definiamo i seguenti oggetti: - X1 ed Y1 i quali rappresentano

le coordinate spaziali dei siti (latitudine e longitudine) - z che rappresenta il dato di interesse, ossia il livello di ozono rilevato in corrispondenza di ciascun sito osservato.

```
x1 <- s[,1]
x2 <- s[,2]      # coordinate spaziali
z <- Y # dato di interesse (livello di ozono)
```

Proseguiamo l'analisi con il Fit di un **trend del primo ordine (o lineare nelle coordinate)** e di un **trend del secondo ordine (o quadratico)**, i quali risultati vengono riportati in appendice, per poi scegliere il modello che meglio si adatti ai dati:

```
aq.fit.1st <- lm(z~x1+x2)                      # Fit trend del primo ordine (lineare)
aq.fit.2nd<- lm(z~x1+x2+I(x1^2)+I(x2^2)+x1*x2) # Fit trend del secondo ordine (quadratico)
```

Per pervenire alla scelta del modello (e quindi del trend), si procede con l'utilizzo dei criteri **AIC** e **BIC**:

```
AIC(aq.fit.1st , aq.fit.2nd)
```

```
##           df      AIC
## aq.fit.1st  4 1768.557
## aq.fit.2nd  7 1518.260
```

```
BIC(aq.fit.1st ,aq.fit.2nd)
```

```
##           df      BIC
## aq.fit.1st  4 1789.096
## aq.fit.2nd  7 1554.204
```

Entrambi i criteri suggeriscono la scelta di un modello **quadratico** rappresentante un **trend del secondo ordine**. Definiamo una griglia di dimensioni 20x20, la quale avrà come estremi i valori minimi e massimi delle coordinate x1 e x2, ed incrociamo tutte le possibili coppie di valori rilevabili in “x1grid” ed “y1grid” mediante la funzione expand.grid():

```
x1grid<-seq(min(x1),max(x1),length=20)    # griglia di valori di x1 (longitudine)
x2grid<-seq(min(x2),max(x2),length=20)    # griglia di valori di x2 (latitudine)
aq.grid<- expand.grid(x1=x1grid,x2=x2grid)
```

Effettuiamo poi la previsione del trend sulle osservazioni contenute nella griglia utilizzando i coefficienti del modello di regressione quadratico precedentemente stimato e la funzione predict(), la quale prende in input:
- i parametri del modello stimato (“**aq.fit.2nd**”)
- la griglia relativa ai siti sui quali prevedere il dato (“**aq.grid**”)

```
aq.surf2<-predict(aq.fit.2nd , newdata=aq.grid) # previsione con trend del secondo ordine
```

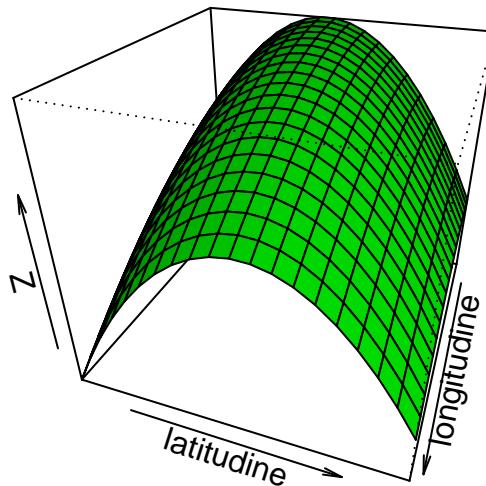
L'utilizzo della funzione predict() permette in questo caso di definire un iperpiano di regressione sulle nuove coordinate contenute in “aq.grid”. Con i seguenti comandi si produce una rappresentazione grafica in 3D dell'iperpiano di regressione definito utilizzando una funzione di tipo quadratico:

```

persp(x1grid,x2grid,matrix(aq.surf2,20,20),xlab="longitudine",ylab="latitudine",zlab= "Z",theta=110,
      phi=30,expand=0.9, col = "green",ltheta = 120, shade = 0.75, ticktype = "simple",
      main="Trend del secondo ordine (quadratico)",box=TRUE)

```

Trend del secondo ordine (quadratico)



Mediante i comandi che seguono ricaviamo i residui relativi al modello di regressione stimato “**aq.fit.2nd**”, rappresentante un **trend del secondo ordine**, e realizziamo una rappresentazione grafica della decomposizione del processo mediante un’operazione di **interpolazione**. Tale operazione viene realizzata mediante la funzione `interp()`, la quale applica il concetto delle **splines** sulle coordinate **x** ed **y** per poter andar ad interpolare il dato di interesse. Eseguiamo quindi un’operazione di interpolazione sul dato originale e sui residui del modello di regressione “**aq.fit.2nd**”:

```

# residui del modello QUADRATICO (stazionari in media):

res_trend=aq.fit.2nd$res

# operazione di interpolazione sul livello di ozono (Z):

pollutant.interp<-interp(x1,x2,z)

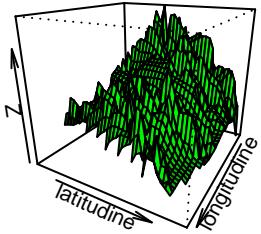
# operazione di interpolazione sui residui del modello QUADRATICO Fittato:

res.interp <- interp(x1,x2, res_trend)

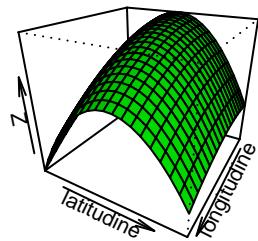
```

e rappresentiamo graficamente la decomposizione del processo mediante una rappresentazione in 3D:

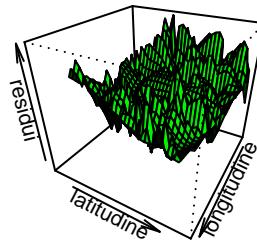
Perspective plot



Trend quadratico



Residui

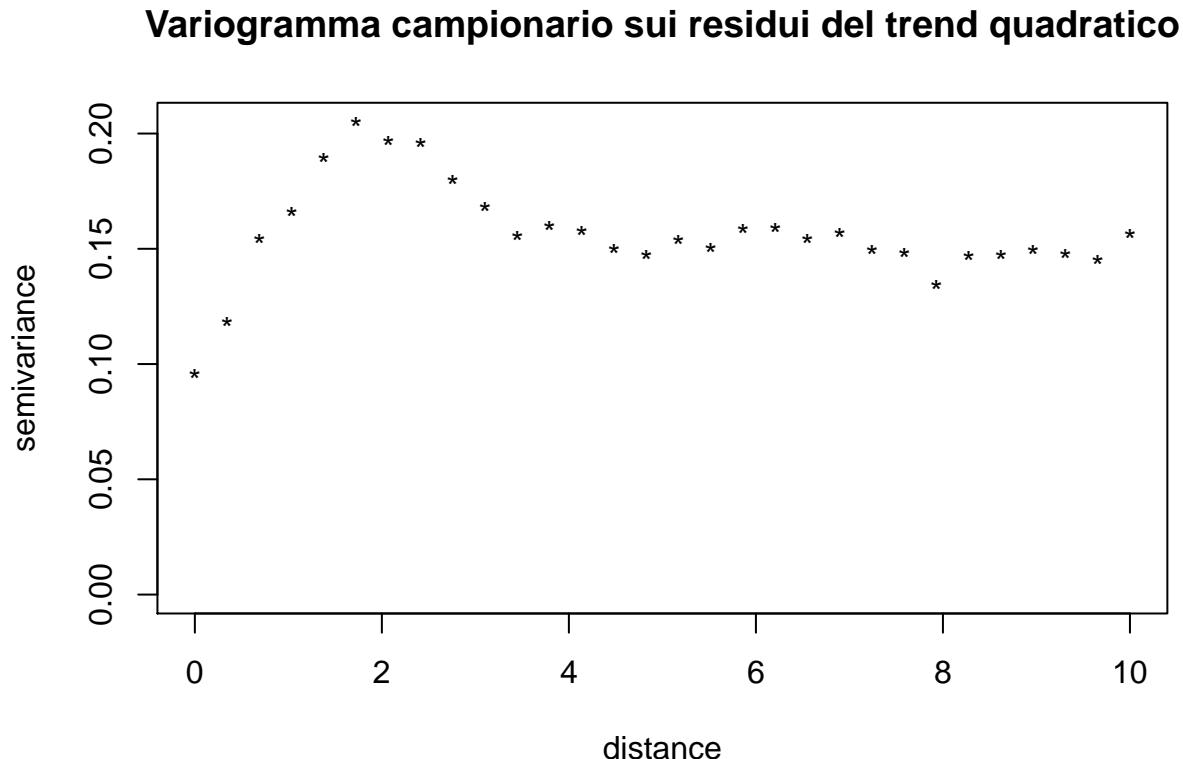


Il primo grafico a sx fornisce una rappresentazione del **processo spaziale** ottenuta mediante un'operazione di interpolazione. Il grafico al centro rappresenta invece il **Trend di fondo del processo (variabilità di larga scala)**, il quale viene modellato mediante un modello di regressione **quadratico**. Fittato il modello di regressione quadratico si procede col ricavare i residui di tale modello (“`res_trend`”), i quali ci permettono di pervenire all'ultima rappresentazione grafica a destra. Osservando le rappresentazioni grafiche ottenute possiamo notare che, sebbene mediante un modello di regressione quadratico si riesce a cogliere il trend di fondo del processo (**variabilità di larga scala**), la conoscenza di tale informazione non risulta essere sufficiente per pervenire alla definizione di un predittore in grado di fornire previsioni accurate circa il livello di ozono, in corrispondenza di siti in cui esso non viene campionato. Infatti, osservando il grafico a destra si può notare che **i residui del modello di regressione risultano essere spazialmente correlati**. Pertanto, i residui del modello di regressione rappresentante il Trend di fondo del processo non rispettano le ipotesi di base del teorema di Gauss-Markow. Occorrerà quindi analizzare la struttura di correlazione spaziale presente nei residui, andando a modellare tale correlazione spaziale (e quindi la **variabilità di piccola scala**).

L'analisi prosegue quindi con lo studio della dipendenza spaziale che caratterizza la **variazione di secondo ordine o di piccola scala**. In particolare si procede a stimare nuovamente la struttura di covarianza sui residui “`res_trend`” partendo dal calcolo del variogramma empirico sugli stessi. Per costruire il variogramma empirico viene prima generato il vettore di distanze “`dist`”, impostando un range di distanze in cui la distanza minima è pari a 0 mentre la distanza massima è pari a 10; il numero di osservazioni presenti tra il minimo è il massimo è pari a 30 :

```
dist <- seq(0,10,length=30) # Vettore delle distanze  
  
vg <- variog(coord = s[,1:2], data = res_trend, uvec = dist)  
  
## variog: computing omnidirectional variogram
```

Eseguiamo quindi una rappresentazione grafica del variogramma empirico sui residui spaziali:

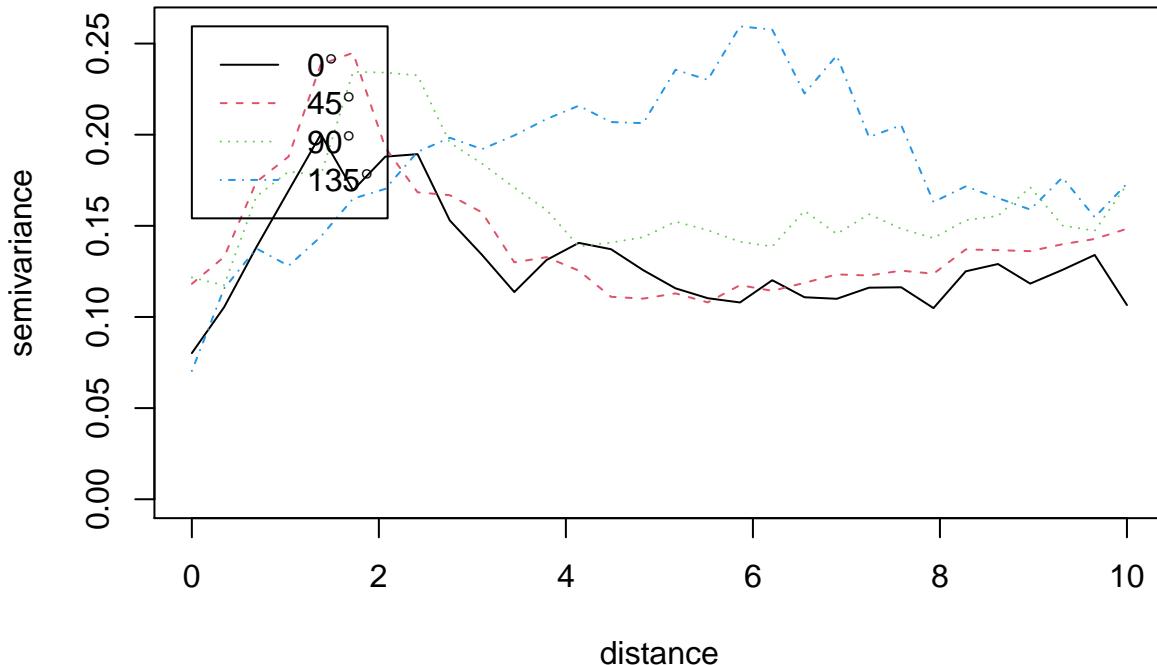


Osservando il grafico ottenuto si nota che il variogramma presenta una struttura diversa da quella osservata precedentemente considerando il dato originale ; rimosso il trend del processo otteniamo che il variogramma campionario calcolato sui residui del modello di regressione quadratico cresce fino ad un'altezza pari all'incirca a 0.20 , per poi stabilizzarsi. Tuttavia, possiamo osservare che oltre una distanza pari all'incirca a 2 il variogramma campionario decresce, per poi tendere a stabilizzarsi dopo una distanza pari a circa 3 e ad un'altezza pari all'incirca a 0.15.

Analizziamo il variogramma dei residui nelle quattro direzioni principali (N-S, NE-SW, E-W e SE-NW) per verificare se effettivamente il processo dev'essere o meno considerato anisotropico:

```
## variog: computing variogram for direction = 0 degrees (0 radians)
##      tolerance angle = 22.5 degrees (0.393 radians)
## variog: computing variogram for direction = 45 degrees (0.785 radians)
##      tolerance angle = 22.5 degrees (0.393 radians)
## variog: computing variogram for direction = 90 degrees (1.571 radians)
##      tolerance angle = 22.5 degrees (0.393 radians)
## variog: computing variogram for direction = 135 degrees (2.356 radians)
##      tolerance angle = 22.5 degrees (0.393 radians)
## variog: computing omnidirectional variogram
```

Variogramma campionario dei residui nelle quattro direzioni: N-S, NE-SW, E-W e SE-NW



L'output ottenuto conferma che il processo dovrebbe essere considerato anisotropico: la struttura di variogramma continua infatti a variare al variare della direzione considerata, suggerendo quindi che occorrerebbe tener conto della direzionalità per calcolare la correlazione tra i siti. Inoltre possiamo notare che, a differenza del variogramma calcolato precedentemente, la struttura di continuità spaziale del fenomeno risulta essere mutata. In particolare, il variogramma calcolato nella direzione Sud Est-Nord Ovest presenta una maggiore pendenza rispetto ai variogrammi calcolati nelle restanti direzioni considerate. La rappresentazione grafica ottenuta suggerisce una maggior continuità spaziale nella direzione Nord-Sud. Per la nostra analisi assumiamo però che il processo sia ISOTROPICO.

Stima della struttura di covarianza dei residui spaziali

Per poter sfruttare le informazioni contenute nel variogramma empirico calcolato sui residui è necessario che quest'ultimo venga interpolato da un modello matematico. La modellazione del variogramma rappresenta un passo molto delicato, in quanto occorre scegliere il modello teorico (e quindi la curva) che meglio approssima la nuvola dei punti del variogramma empirico. Esistono vari modelli teorici di variogramma per l'approssimazione dei variogrammi empirici; fra questi i più utilizzati sono: - il modello esponenziale, - il modello sferico, - il modello Gaussiano, - il modello wave o hole-effect.

Tali modelli teorici si caratterizzano per essere parametrizzati, e pertanto dipendono da i seguenti parametri che dovranno essere stimati:

- la sella parziale, che rappresenta la varianza del segnale vero (latente) non influenzata dall'errore di misura,
- il Nugget, che rappresenta la varianza dell'errore di misura,
- il range, che rappresenta la distanza massima oltre la quale non si osserva più correlazione spaziale (il variogramma si assesta).

Per pervenire alla definizione del modello teorico di variogramma da impiegare per interpolare il variogramma empirico proviamo a stimare i valori di sella parziale, Nugget e Range osservando il grafico del variogramma empirico, per poi eseguire un confronto grafico volto ad osservare in che modo il modello teorico così ottenuto si adatta alla nuvola dei punti del variogramma empirico. Sulla base dell'output relativo alla rappresentazione grafica del variogramma empirico vengono fissati i seguenti valori di sella parziale, Nugget e Range:

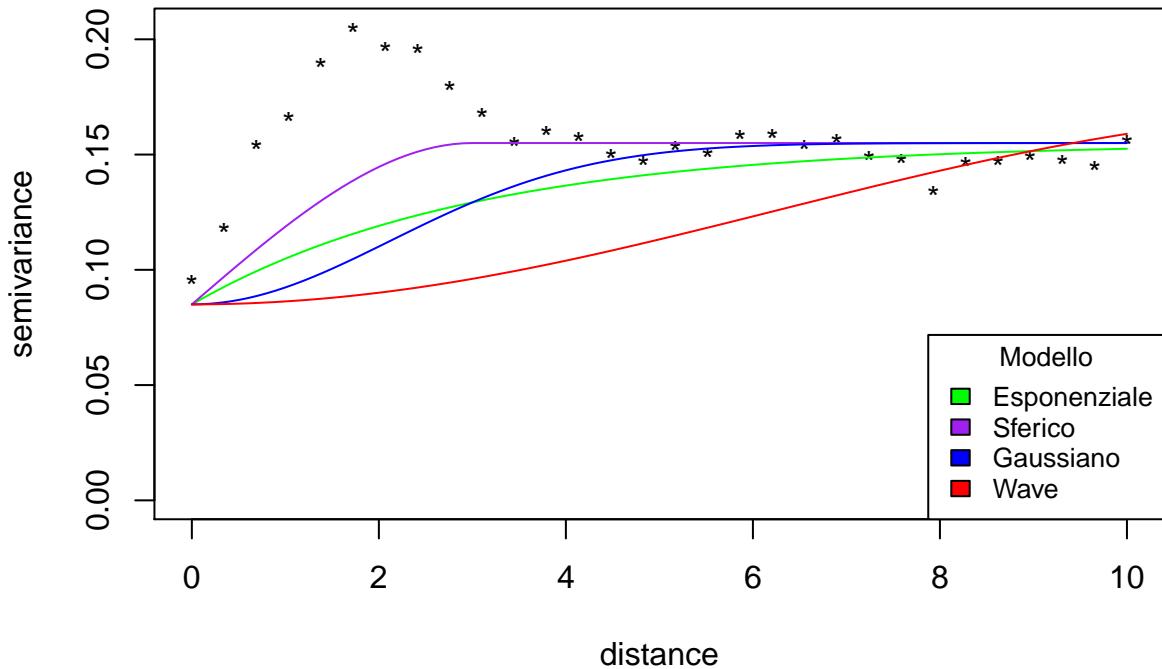
```
# parametri fissati ad occhio:

tau2=0.085 # Nugget
sig2=0.070 # sella parziale (gli ho detratto il nugget)
phi= 3 # Range
```

Sulla base di tali valori si procede col Fitting dei quattro modelli teorici di variogramma maggiormente utilizzati (ossia, il modello esponenziale, il modello sferico, il modello gaussiano ed il modello wave):

Infine, si esegue un confronto grafico fra modelli teorici Fittati ed il variogramma empirico:

variogramma campionario con i vari modelli



Dal confronto grafico dei modelli teorici di variogramma fittati ed il variogramma empirico calcolato sui residui spaziali emerge che, dati i parametri di sella parziale, Nugget e Range ipotizzati, il modello esponenziale sembra adattarsi meglio alla nuvola dei punti del variogramma empirico. Si procede ora con l'applicazione del metodo della massima verosimiglianza per stimare i tre parametri incogniti caratterizzanti la struttura di correlazione spaziale del fenomeno (sella parziale, Nugget, Range). In particolare, consideriamo i modelli teorici di variogramma precedentemente Fittati e, per ciascuno di essi, applichiamo il metodo della massima verosimiglianza per stimare i parametri incogniti. Successivamente, si procede con la scelta del modello da adottare sulla base del valore della verosimiglianza e sulla base dei criteri BIC ed AIC. Con le seguenti righe di codice perveniamo alla stima dei parametri incogniti caratterizzanti i modelli teorici di variogramma considerati:

```

# Stima di massima verosimiglianza utilizzando il modello ESPONENZIALE

mle.nc.exp<-likfit(data,ini.cov.pars=c(0.1,2), cov.model="exponential",trend= "2nd" ,
nugget = 0.1 , fix.nugget=FALSE)

# Stima di massima verosimiglianza utilizzando il modello SFERICO

mle.nc.sferic<-likfit(data,ini.cov.pars=c(0.1,2), cov.model="spherical",trend= "2nd" ,
nugget = 0.1 , fix.nugget=FALSE)

# Stima di massima verosimiglianza utilizzando il modello GAUSSIANO

mle.nc.gaus<-likfit(data,ini.cov.pars=c(0.1,2), cov.model= "gaussian" ,trend= "2nd",

# Stima di massima verosimiglianza utilizzando il modello WAVE

mle.nc.wawe<-likfit(data,ini.cov.pars=c(0.1,2), cov.model="wave",trend= "2nd" ,
nugget = 0.1 , fix.nugget=FALSE)

```

Successivamente, mettiamo a confronto i valori di verosimiglianza ottenuti, i valori di AIC ed i valori di BIC:

```

kable(cbind(rbind("esponenziale", "sferico", "gaussiano","wave"),rbind(mle.nc.exp$loglik , mle.nc.sferic
caption = "Valori di verosimiglianza")

```

Table 2: Valori di verosimiglianza

esponenziale	-434.558037306561
sferico	-436.210720652392
gaussiano	-454.701507489114
wave	-480.411518391649

```

kable(cbind(rbind("esponenziale", "sferico", "gaussiano","wave"),rbind(mle.nc.exp.BIC , mle.nc.sferic.B
caption = "Valori di BIC")

```

Table 3: Valori di BIC

mle.nc.exp.BIC	esponenziale	933.330092277215
mle.nc.sferic.BIC	sferico	936.635458968878
mle.nc.gaus.BIC	gaussiano	973.617032642321
mle.nc.wawe.BIC	wave	1025.03705444739

```

kable(cbind(rbind("esponenziale", "sferico", "gaussiano","wave"),rbind(mle.nc.exp.AIC , mle.nc.sferic.A
caption = "Valori di AIC")

```

Table 4: Valori di AIC

mle.nc.exp.AIC	esponenziale	887.116074613122
mle.nc.sferic.AIC	sferico	890.421441304785
mle.nc.gaus.AIC	gaussiano	927.403014978228

Table 4: Valori di AIC

mle.nc.wawe.AIC	wave	978.823036783298
-----------------	------	------------------

Dai risultati ottenuti emerge che tutti gli approcci adottati per pervenire alla scelta del modello suggeriscono l'utilizzo di un modello esponenziale; infatti: - confrontando i valori di verosimilanza ottenuti l'evidenza empirica suggerisce di considerare un modello teorico esponenziale (“mle.nc.exp\$loglik”), dato che in corrispondenza di tale modello otteniamo il più alto valore di verosimiglianza, - confrontando invece i valori di AIC e BIC ottenuti l'evidenza empirica suggerisce ancora una volta di considerare un modello teorico esponenziale (“mle.nc.exp.BIC”, “mle.nc.exp.AIC”), dato che in corrispondenza di tale modello otteniamo i valori più bassi degli indici.

Pertanto si può concludere che l'evidenza empirica suggerisce la scelta di un modello esponenziale. Osserviamo ora l'output generato dalla stima di massima verosimiglianza del modello esponenziale:

```
mle.nc.exp # modello ESPONENZIALE
```

```
## likfit: estimated model parameters:
##      beta0      beta1      beta2      beta3      beta4      beta5      tausq    sigmasq
## "-9.7554" "-0.1087" " 0.4774" "-0.0008" "-0.0077" "-0.0012" " 0.0795" " 0.1680"
##      phi
## " 4.5978"
## Practical Range with cor=0.05 for asymptotic range: 13.77379
##
## likfit: maximised log-likelihood = -434.6
```

L'output mostra le stime dei seguenti parametri: - beta0, beta1, ..., beta5 ossia i parametri relativi al trend (rappresentante la variabilità di larga scala); in questo caso vengono stimati cinque parametri in quanto è stato preso in considerazione un trend del secondo ordine(quadratico),

- “tausq”: la stima del parametro di Nugget, il quale rappresenta la varianza dell'errore di misura,
- “sigmasq”: la stima del parametro di sella, la quale rappresenta la varianza del segnale latente non disturbata dall'errore di misura,
- “phi”: la stima del parametro di Range,
- “Practical Range”: la stima del valore di Range effettivo, vale a dire il valore di range oltre il quale la correlazione si assesta, e risulta essere più piccola di 0.05,
- “likfit” che rappresenta invece il valore della verosimiglianza.

In particolare, il modello di variogramma esponenziale stimato partirà da un'altezza pari a 0.079 circa (“tausq”) per poi raggiungere asintoticamente il valore di sella ad un'altezza pari a 0.168 (“sigmasq”). La distanza oltre la quale il variogramma comincia ad assestarsi è pari invece pari a 4.6 circa (“phi”), mentre la distanza oltre la quale la correlazione risulta essere più piccola di 0.05 è pari a 13.77 circa (“Practical Range”) Il valore di verosimiglianza (il più alto fra i valori ottenuti) è invece pari a -434.6 (“likfit”).

Previsione del livello di ozono mediante l'algoritmo Kriging

In questa sezione si procede con la previsione del livello di ozono in corrispondenza dei siti non campionati sfruttando le informazioni presenti all'interno del campione di dati fin qui analizzato. Il kriging è un metodo di regressione che permette di pervenire alla definizione di un previsore costruito come combinazione lineare dei

dati osservati, con l'obiettivo di ottenere una stima del valore assunto dal fenomeno d'interesse (nel nostro caso il livello di ozono) in corrispondenza di siti in cui il dato non è stato campionato. Conoscendo il valore di ozono in corrispondenza dei siti contenuti nel campione a nostra disposizione siamo in grado determinare il valore di ozono in corrispondenza di siti per i quali non esistono misure mediante una combinazione lineare dei dati a nostra disposizione.

Nelle righe di codice che seguono viene generata una griglia su cui verrà effettuata la previsione: tale griglia avrà dimensioni 701x251 e quindi conterrà 175951 siti.

```
X0 <- seq(-135,-65,0.1)
Y0 <- seq(25,50,0.1)
s0 <- as.matrix(expand.grid(X0,Y0))
```

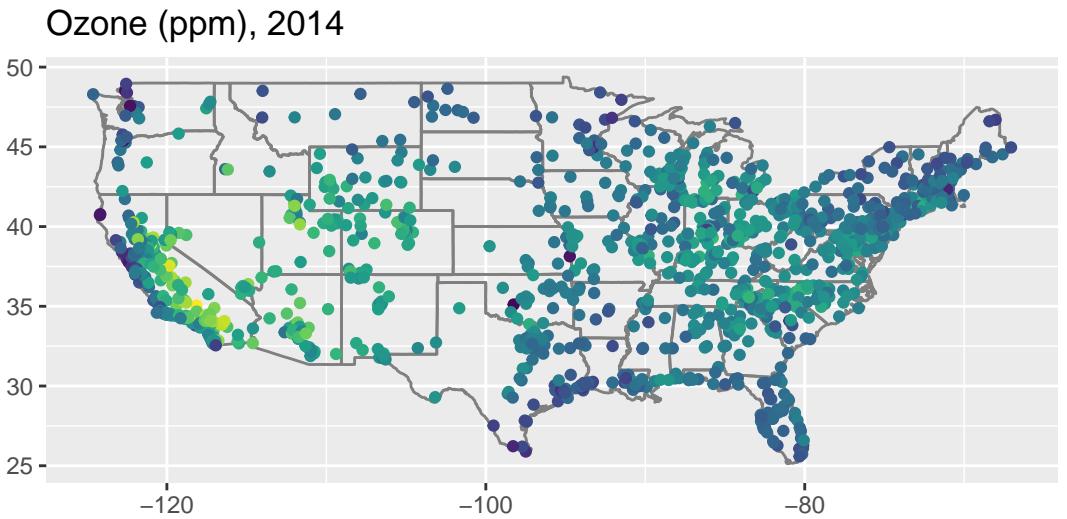
Passiamo ora all'applicazione del kriging. Il kriging è un metodo che permette di interpolare i dati geospatiali, correlati tra di loro, in siti non osservati inizialmente; questo avviene utilizzando i valori osservati come predittori che contribuiscono all'ottenimento del valore dell'ozono in nuovi punti. Imponiamo un kriging di tipo universale a cui passiamo come parametri i valori di nugget, sella e range stimati tramite la massima verosimiglianza, il modello di covarianza esponenziale e un trend di tipo quadratico.

```
krige.par.exp<-krige.control(type.krige='ok',cov.pars=mle.nc.exp$cov.pars,
                                cov.model= "exponential",trend.d= "2nd" ,
                                trend.l= "2nd")
kriging.exp<-krige.conv(data,locations= s0, krige=krige.par.exp)
```

Rappresentazione grafica dei risultati ottenuti

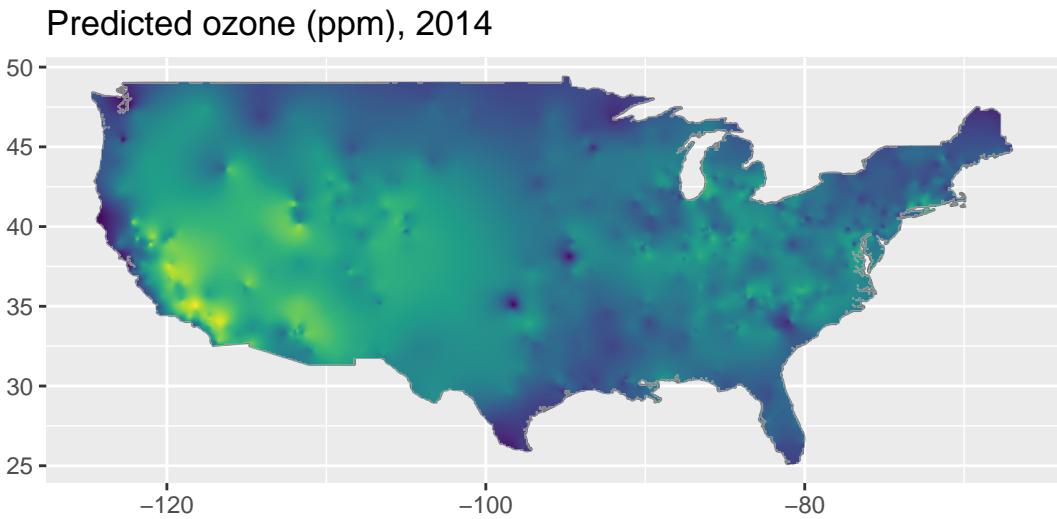
Nel grafico seguente rappresentiamo i livelli di ozono rilevati nei siti osservati:

```
df <- data.frame(long=s[,1],lat=s[,2],Y=Y)
ggplot(df, aes(long, lat)) +
  borders("state") +
  geom_point(aes(colour = Y)) +
  scale_colour_gradientn(colours = viridis(10)) +
  xlab("")+ylab("")+labs(title="Ozone (ppm), 2014")+
  coord_fixed()
```



Mentre nel grafico successivo andiamo a rappresentare la previsione dei livelli di ozono effettuate tramite il kriging sulla griglia di valori creata in precedenza:

```
df.qd.exp <- data.frame(long=s0[,1],lat=s0[,2],Y=kriging.exp$pred)
ggplot(df.qd.exp, aes(long, lat)) +
  borders("state") +
  geom_raster(aes(fill = Y)) +
  scale_fill_gradientn(colours = viridis(10))+
  xlab("")+ylab("")+labs(title="Predicted ozone (ppm), 2014")+
  coord_fixed()
```



Il grafico previsionale mostra come i livelli di Ozono più elevati sono presenti nella parte sud-occidentale degli Stati Uniti e in particolar modo nello stato della California.

Rappresentazione grafica su mappa dell'errore di previsione

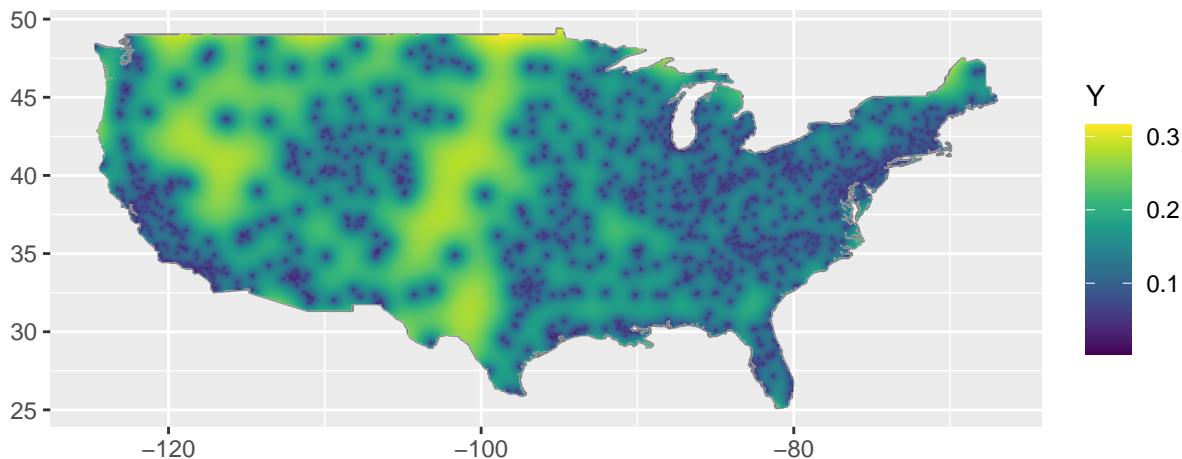
Plot dell'errore quadratico medio (ossia, delle varianze degli errori di previsione del kriging):

```
se.exp <- sqrt(kriging.exp$krige.var)
df.var.exp <- data.frame(long=s0[,1],lat=s0[,2],Y=se.exp)

# Rappresentazione grafica:

ggplot(df.var.exp, aes(long, lat)) +
  borders("state") +
  geom_raster(aes(fill = Y)) +
  scale_fill_gradientn(colours = viridis(10))+
  xlab("")+ylab("")+labs(title="Errore quadratico medio di previsione: trend lineare e modello esponenziale")+
  coord_fixed()
```

Errore quadratico medio di previsione: trend lineare e modello esponenziale



Nel grafico soprastante i puntini in blu scuro sono le osservazioni. Nell'area dell'america centrale l'errore di previsione è elevato (giallo/verde chiaro) perchè c'è molta distanza tra i siti di osservazione, quindi c'è poca influenza da aree vicine causando un aumento di incertezza previsiva.

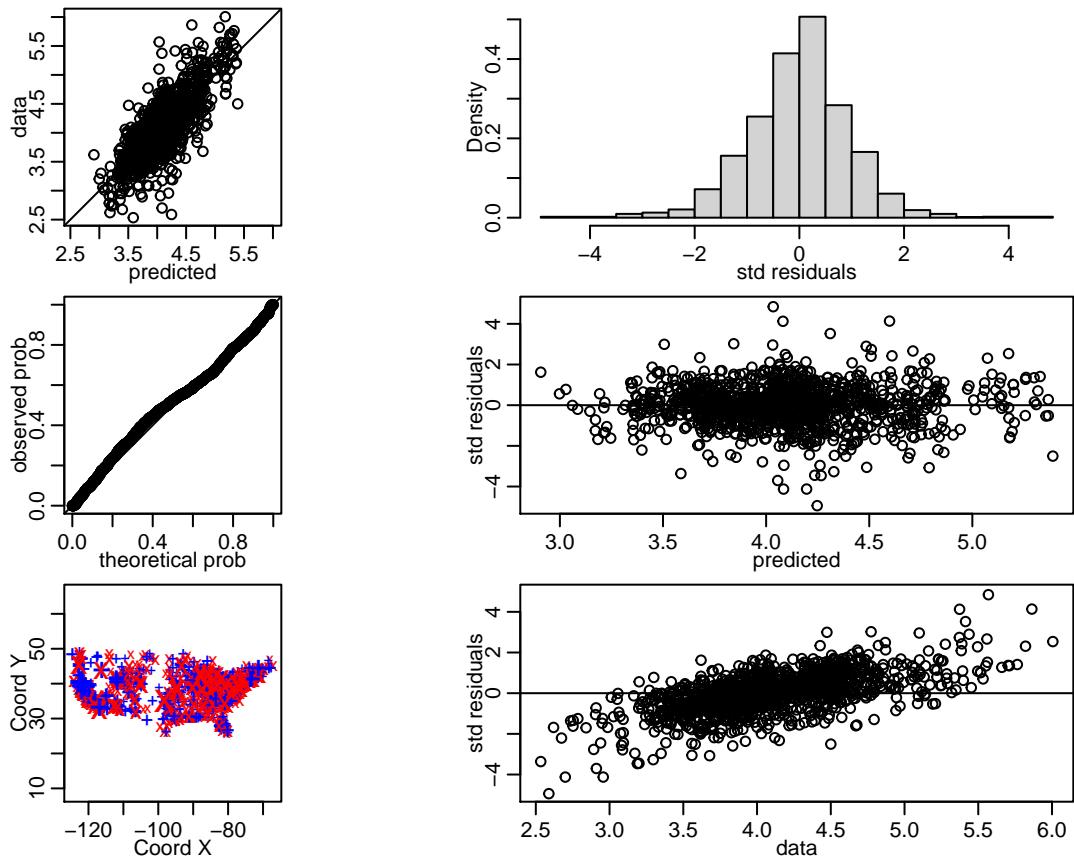
Utilizziamo la tecnica di Leave-One-Out CV per verificare la bontà di previsione del modello stimato:

```
cv.exp<-xvalid(data,model=mle.nc.exp)
```

```
## xvalid: number of data locations      = 1255
## xvalid: number of validation locations = 1255
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## xvalid: end of cross-validation
```

Le seguenti 2 righe di codice servono per fare il setting per la parametrizzazione della rappresentazione grafica: in particolare le funzioni utilizzate servono per mettere insieme tutti i panel che otteniamo quando andiamo a plottare i risultati ottenuti:

```
par.ori <- par(no.readonly = TRUE)
par(mfcol=c(3,2), mar=c(2.3,2.3,.5,.5), mgp=c(1.3, .6, 0))
plot(cv.exp , error = FALSE, ask = FALSE)
```



Il grafico in alto a sinistra confronta le previsioni con i dati: vediamo quindi la coerenza che esiste fra il dato previsto e il dato vero. Il secondo grafico a sinistra è un Q-Q Plot: i quantili dei valori previsti rispetto a quelli della normale per vedere di quanto ci allontaniamo dall'ipotesi di normalità. Il grafico in alto a dx riporta la rappresentazione dell'istogramma dei residui, che risulta più o meno simmetrica. I restanti 2 grafici di destra sono dei classici grafici utilizzati nell'ambito della regressione che confrontano i valori previsti rispetto ai residui e i dati rispetto ai residui: si può osservare che esiste un pattern abbastanza parallelo all'asse delle ascisse e non vi sono forme ad imbuto che si allargano e che portano ad ipotizzare la presenza di eteroschedasticità.

Appendice

```
summary(aq.fit.1st)

##
## Call:
## lm(formula = z ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82917 -0.30317  0.00038  0.32787  1.66145
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.2340428  0.1454982 22.227 <2e-16 ***
## x1           0.1454982  0.1454982  1.000  0.3125
## x2           0.3278700  0.1454982  2.227  0.0291 *
```

```

## x1      -0.0102085  0.0008776 -11.633  <2e-16 ***
## x2      -0.0027388  0.0029378 -0.932    0.351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4885 on 1252 degrees of freedom
## Multiple R-squared:  0.1005, Adjusted R-squared:  0.09906
## F-statistic: 69.94 on 2 and 1252 DF,  p-value: < 2.2e-16

```

```
summary(aq.fit.2nd)
```

```

##
## Call:
## lm(formula = z ~ x1 + x2 + I(x1^2) + I(x2^2) + x1 * x2)
##
## Residuals:
##       Min     1Q   Median     3Q    Max 
## -1.72810 -0.27079  0.01522  0.26948  1.54257
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.026e+01  1.090e+00 -9.414   < 2e-16 ***
## x1          -6.799e-02  1.581e-02 -4.299 1.85e-05 ***
## x2          5.819e-01  3.490e-02 16.676   < 2e-16 ***
## I(x1^2)     -1.908e-04  6.269e-05 -3.044  0.00239 **  
## I(x2^2)     -7.137e-03  5.036e-04 -14.171  < 2e-16 ***
## x1:x2      5.493e-04  2.087e-04   2.632  0.00859 **  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4417 on 1249 degrees of freedom
## Multiple R-squared:  0.2667, Adjusted R-squared:  0.2637
## F-statistic: 90.83 on 5 and 1249 DF,  p-value: < 2.2e-16

```

```
summary(mle.nc.exp)
```

```

## Summary of the parameter estimation
## -----
## Estimation method: maximum likelihood
##
## Parameters of the mean component (trend):
##   beta0   beta1   beta2   beta3   beta4   beta5
## -9.7554 -0.1087  0.4774 -0.0008 -0.0077 -0.0012
##
## Parameters of the spatial component:
##   correlation function: exponential
##   (estimated) variance parameter sigmasq (partial sill) =  0.168
##   (estimated) cor. fct. parameter phi (range parameter) =  4.598
##   anisotropy parameters:
##   (fixed) anisotropy angle = 0 ( 0 degrees )
##   (fixed) anisotropy ratio = 1
##
## Parameter of the error component:

```

```

##      (estimated) nugget =  0.0795
##
## Transformation parameter:
##      (fixed) Box-Cox parameter = 1 (no transformation)
##
## Practical Range with cor=0.05 for asymptotic range: 13.77379
##
## Maximised Likelihood:
##      log.L n.params      AIC      BIC
## "-434.6"      "9"    "887.1"    "933.3"
##
## non spatial model:
##      log.L n.params      AIC      BIC
## "-752.1"      "7"    "1518"    "1554"
##
## Call:
## likfit(geodata = data, trend = "2nd", ini.cov.pars = c(0.1, 2),
##        fix.nugget = FALSE, nugget = 0.1, cov.model = "exponential")

summary(mle.nc.sferic)

## Summary of the parameter estimation
## -----
## Estimation method: maximum likelihood
##
## Parameters of the mean component (trend):
##   beta0   beta1   beta2   beta3   beta4   beta5
## -9.6865 -0.1213  0.4485 -0.0010 -0.0080 -0.0017
##
## Parameters of the spatial component:
##   correlation function: spherical
##      (estimated) variance parameter sigmasq (partial sill) =  0.2535
##      (estimated) cor. fct. parameter phi (range parameter) =  12.24
##   anisotropy parameters:
##      (fixed) anisotropy angle = 0 ( 0 degrees )
##      (fixed) anisotropy ratio = 1
##
## Parameter of the error component:
##      (estimated) nugget =  0.0812
##
## Transformation parameter:
##      (fixed) Box-Cox parameter = 1 (no transformation)
##
## Practical Range with cor=0.05 for asymptotic range: 12.2429
##
## Maximised Likelihood:
##      log.L n.params      AIC      BIC
## "-436.2"      "9"    "890.4"    "936.6"
##
## non spatial model:
##      log.L n.params      AIC      BIC
## "-752.1"      "7"    "1518"    "1554"
##
## Call:

```

```

## likfit(geodata = data, trend = "2nd", ini.cov.pars = c(0.1, 2),
##        fix.nugget = FALSE, nugget = 0.1, cov.model = "spherical")

summary(mle.nc.gaus)

## Summary of the parameter estimation
## -----
## Estimation method: maximum likelihood
##
## Parameters of the mean component (trend):
##   beta0   beta1   beta2   beta3   beta4   beta5
## -8.9623 -0.0551  0.5421 -0.0003 -0.0074 -0.0001
##
## Parameters of the spatial component:
##   correlation function: gaussian
##   (estimated) variance parameter sigmasq (partial sill) = 0.0948
##   (estimated) cor. fct. parameter phi (range parameter) = 1.762
##   anisotropy parameters:
##   (fixed) anisotropy angle = 0 (0 degrees)
##   (fixed) anisotropy ratio = 1
##
## Parameter of the error component:
##   (estimated) nugget = 0.0915
##
## Transformation parameter:
##   (fixed) Box-Cox parameter = 1 (no transformation)
##
## Practical Range with cor=0.05 for asymptotic range: 3.048752
##
## Maximised Likelihood:
##   log.L n.params      AIC      BIC
## "-454.7"      "9"    "927.4"    "973.6"
##
## non spatial model:
##   log.L n.params      AIC      BIC
## "-752.1"      "7"    "1518"     "1554"
##
## Call:
## likfit(geodata = data, trend = "2nd", ini.cov.pars = c(0.1, 2),
##        fix.nugget = FALSE, nugget = 0.1, cov.model = "gaussian")

```

```
summary(mle.nc.wawe)
```

```

## Summary of the parameter estimation
## -----
## Estimation method: maximum likelihood
##
## Parameters of the mean component (trend):
##   beta0   beta1   beta2   beta3   beta4   beta5
## -9.6979 -0.0789  0.5187 -0.0004 -0.0071 -0.0002
##
## Parameters of the spatial component:
##   correlation function: wave

```

```

##      (estimated) variance parameter sigmasq (partial sill) = 0.1537
##      (estimated) cor. fct. parameter phi (range parameter) = 1.028
##      anisotropy parameters:
##          (fixed) anisotropy angle = 0 ( 0 degrees )
##          (fixed) anisotropy ratio = 1
##
## Parameter of the error component:
##      (estimated) nugget = 0.1038
##
## Transformation parameter:
##      (fixed) Box-Cox parameter = 1 (no transformation)
##
## Practical Range with cor=0.05 for asymptotic range: 3.074805
##
## Maximised Likelihood:
##      log.L n.params      AIC      BIC
##      "-480.4"      "9"    "978.8"    "1025"
##
## non spatial model:
##      log.L n.params      AIC      BIC
##      "-752.1"      "7"    "1518"    "1554"
##
## Call:
## likfit(geodata = data, trend = "2nd", ini.cov.pars = c(0.1, 2),
##       fix.nugget = FALSE, nugget = 0.1, cov.model = "wave")

```