ALMA MATER STUDIORUM – UNIVERSITA' DI BOLOGNA

DIPARTIMENTO DI SCIENZE STATISTICHE

"PAOLO FORTUNATI"

Corso di Laurea in Scienze Statistiche

ANALISI DI DATI GEOSTATISTICI SULLA CONCENTRAZIONE GIORNALIERA DI PM_{10} IN ITALIA

(Epidemiologia Ambientale)

Presentata da:

Susanna Casarini

matricola: 0000938926

Relatore:

Prof Massimo Ventrucci

APPELLO III

ANNO ACCADEMICO 2021 /2022

Indice

Introduzione……..4

Obiettivi

La concentrazione di PM_{10} in Italia

1 Modelli geostatistici spaziali……...……5

1.1 Dipendenza spaziale dei dati geostatistici

1.2 Processi stocastici spaziali

1.3 Correlazione spaziale nei processi Gaussiani isotropici

1.4 Variogramma teorico e funzione di covarianza spaziale

1.5 Variogramma empirico e sample variogram

1.6 Modellazione spaziale di un processo non stazionario in media

1.7 Famiglia Matérn

1.8 Stima dei parametri

1.9 Previsione spaziale

2 Analisi esplorative per dati geostatistici spazio-temporali……14

2.1 Processi stocastici Gaussiani spazio-temporali

2.2 Struttura di covarianza e variogramma spazio-temporale

3 Applicazione dei metodi: analisi della concentrazione giornaliera di PM10 in Italia...16

3.1 Introduzione

3.2 Materiali e Metodi

3.3 Analisi esplorative spaziali

3.4 Modellazione spaziale

3.5 Previsione spaziale

3.6 Correlazione spazio-temporale

Conclusioni……..27

Bibliografia……..29

Introduzione

Obiettivi

La geostatistica è una branca della statistica spaziale che studia i fenomeni spaziali continui (Diggle e Ribeiro, 2007). Il seguente lavoro di tesi ha l'obiettivo di descrivere sia dal punto di vista teorico, che dal punto di vista pratico, la modellazione spazio-temporale e la previsione spaziale di dati geostatistici. Per la descrizione dei metodi di modellazione spaziale è stato fatto riferimento principalmente al libro "Model based geostatistics" di Diggle e Ribeiro (2007), nel quale vengono descritti i metodi di stima di un modello spaziale con un approccio model-based, mentre per la parte di analisi spazio-temporale è stato fatto riferimento al lavoro di Padoan e Bevilacqua (2015). Nel primo capitolo del lavoro di tesi sono spiegati i processi stocastici spaziali, la modellazione di un processo spaziale Gaussiano, i metodi di stima del modello e la previsione spaziale, mentre nel secondo capitolo sono descritti i metodi di studio della dipendenza spazio-temporale di dati geostatisitici. Nel terzo capitolo viene illustrata l'applicazione dei metodi descritti nei primi capitoli a dati sulla concentrazione di PM_{10}.

La concentrazione di PM_{10} in Italia

Nel terzo capitolo del seguente lavoro di tesi, i metodi geostatistici sono stati applicati a dati sulla concentrazione giornaliera di particolato atmosferico (PM_{10}) nel territorio italiano nell'anno 2015. Il particolato atmosferico è un inquinante costituito da particelle solide o liquide in sospensione, che provengono sia da processi naturali che da processi antropogenici; in particolare la sigla "PM_{10}" indica le particelle con diametro minore di 10μ. Lo studio dell'esposizione al PM_{10} è di grande interesse nella scelta delle politiche di mitigazione dell'inquinamento e di riduzione degli effetti sulla salute. Infatti, secondo gli studi del progetto ESCAPE (2013), l'esposizione a questo inquinante causa l'aumento del rischio di tumore ai polmoni, di eventi coronarici acuti, di mortalità per cause non accidentali e di mortalità per cause cardiovascolari. Inoltre un'esposizione eccessiva aumenta l'incidenza di demenza, morbo di Parkinson e sclerosi multipla (The Lancet, gennaio 2017). Nell'ottobre 2013 l'Agenzia Internazionale per la Ricerca sul Cancro ha classificato l'inquinamento atmosferico esterno come cancerogeno di gruppo 1 per l'uomo, per il tumore al polmone.

1 Modelli geostatistici spaziali

In questo paragrafo sono descritti i processi stocastici spaziali, specificando in modo particolare i processi spaziali stazionari e la correlazione spaziale del processo. Nei paragrafi successivi viene descritta la specificazione del modello spaziale stocastico Gaussiano.

Dipendenza spaziale dei dati geostatistici

Per studiare un fenomeno spaziale è possibile analizzare i dati geostatistici ottenuti da misurazioni del fenomeno stesso. I dati geostatistici sono solitamente definiti come coppie (x_i, y_i) per i=1,...,n in cui y_i corrisponde alla misurazione del fenomeno ed è la realizzazione di una variabile aleatoria Y_i associata ad una serie di locazioni x_i situate in una regione spaziale continua. La distribuzione della variabile risposta Y_i dipende da un processo spaziale continuo S(x) non direttamente osservabile e specificato in ogni sito x_i, S(x) è chiamato anche segnale, termine coniato da Diggle e Ribeiro (2007). In generale $Y_i = S(x_i) + Z_i$, dove Z_i è chiamato rumore ed è l'errore associato alla risposta; è indipendente dal processo, incorrelato spazialmente e si distribuisce normalmente con media nulla e varianza τ^2.

Alla base dell'analisi geostatistica vi è il concetto di "dipendenza spaziale" del fenomeno, che può essere spiegata dalla prima legge della geografia di Tobler (1970) "Tutto è correlato a tutto il resto, ma le cose vicine sono più correlate di quelle lontane", infatti le misure Y_i del fenomeno, osservate in diverse locazioni, non sono indipendenti tra loro.

Processi stocastici spaziali

Per studiare un fenomeno Y_i è necessario soffermarsi sul segnale S(x), che è un processo stocastico spaziale continuo definito sul dominio spaziale $D \in R^2$, dove D corrisponde alla regione spaziale studiata. Per conoscere questo processo è necessario conoscere la distribuzione congiunta del vettore casuale spaziale $(S(x_1), \ldots, S(x_n))^T$, che solitamente nella modellazione spaziale viene assunto come processo stazionario Gaussiano, quindi ha una distribuzione Gaussiana multivariata con media $\mu(x)$ e matrice di covarianza Σ. I processi Gaussiani vengono spesso utilizzati per la modellazione spaziale poiché con pochi parametri è possibile modellare diversi fenomeni. Infatti il vettore spaziale è spiegato interamente dalla media $\mu(x)$ e dalla matrice di covarianza Σ, che insieme alla funzione di correlazione, sono definite come segue: $\mu(x) = E[S(x)]$, $c(x_i,x_j) = Cov[S(x_i), S(x_j)]$, $\sigma^2(x) = Var[S(x)] = c(x,x)$, $\rho(x_i,x_j) = \frac{c(x_i,x_j)}{\sigma(x_i)\sigma(x_j)}$.

Vi sono alcune proprietà dei processi spaziali, come la stazionarietà e l'isotropia, che possono semplificare la modellazione e la previsione del fenomeno. Un processo stocastico si definisce

stazionario se è invariante per traslazione, quindi se la media spaziale $\mu(x)$ è costante e se la covarianza tra $S(x_i)$ e $S(x_j)$ dipende solo dal vettore differenza x_i-x_j, di conseguenza in un processo isotropico la varianza è costante, infatti $c(x,x)=c(0)= \sigma^2$. Un processo stazionario S(x) si definisce isotropico se è invariante per rotazione, quindi se la covarianza $c(x_i,x_j)$ dipende solo dalla distanza Euclidea $u=||x_i - x_j||$; ne segue che più due locazioni x_i e x_j sono lontane, più la correlazione tra $S(x_i)$ e $S(x_j)$ diminuisce.

Correlazione spaziale nei processi Gaussiani isotropici

Nei prossimi capitoli ci si soffermerà sulla descrizione della correlazione spaziale nei processi Gaussiani isotropici. Nel caso di isotropia il processo S(x) si distribuisce normalmente con media μ costante e matrice di covarianza Σ, in cui le covarianze $c(||x_i-x_j||)$ sono solo funzione della distanza euclidea e, nella diagonale principale, la varianza σ^2 è costante. Di conseguenza Σ può essere scritta come prodotto tra la varianza σ^2 e la matrice di correlazione R; la funzione di covarianza è quindi definita dalla seguente formula: $c(x_i, x_j)=\sigma^2\rho(u)$. Anche la correlazione assume una forma semplice: $\rho(u) = \frac{c(u)}{\sigma 2}$, è funzione solo della distanza u tra i siti ed è simmetrica, infatti $||x_i - x_j|| = ||x_i - x_j||$.

Variogramma teorico e funzione di covarianza spaziale

Per studiare la struttura di covarianza di un processo spaziale viene utilizzata la funzione $\gamma(x_i,x_j)$, chiamata variogramma. Nell'approccio model-based (Diggle e Ribeiro, 2007) il variogramma ha una funzione prettamente esplorativa, che verrà spiegata nel dettaglio nei paragrafi sulla stima dei parametri del modello. La funzione $\gamma(x_i,x_j)$ è chiamata anche semivariogramma ed è definita dalla seguente relazione $\gamma(x_i,x_j) = ½ \text{ Var}(S(x_i) - S(x_j)) = ½ [\sigma^2(x_i) + \sigma^2(x_j) – 2c(x_i, x_j)]$. In un processo isotropico il variogramma dipende solo dalla distanza Euclidea tra i siti e si semplifica come segue: $\gamma(x_i, x_j) = ½ [2\sigma^2 – 2c(u)] = \sigma^2(1 - \rho(u))$. Nel seguente paragrafo si farà riferimento al Variogramma nel caso di isotropia, rappresentato dalla seguente figura.

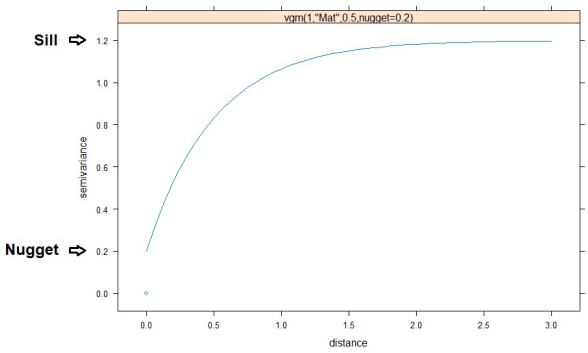

Figura 1.1 - Variogramma teorico di un processo isotropico

L'osservazione del grafico del Variogramma è fondamentale per l'analisi esplorativa della correlazione spaziale. Come si può osservare dalla Figura 1, il valore di $\gamma(x_i,x_j)$ aumenta all'aumentare della distanza e si annulla quando i siti coincidono, perchè la correlazione tra le variabili è massima, quindi uguale all'unità. All'origine degli assi vi può essere una discontinuità, come si può vedere dalla Figura 1.1; in questo caso si parla di nugget effect, termine introdotto da Georges Matheron nel 1962, che indica un effetto dovuto a variazioni spaziali minori della distanza più piccola campionaria studiata. Il valore del variogramma in funzione di una distanza nulla per un processo affetto da errore è chiamato nugget ed è uguale alla varianza τ^2 del rumore. Se viene raggiunta la distanza u=ϕ, chiamata range, per cui la covarianza si annulla, il valore del variogramma viene chiamato sill ed è specificato dalla formula c(0)=$\tau^2+\sigma^2$. Spesso da un punto di vista grafico il range non può essere definito, poiché la correlazione non si annulla mai e la sill non viene raggiunta se non asintoticamente, ma può essere definito il practical range, che corrisponde alla distanza in cui la funzione di correlazione assume il valore 0.05. La sill e il range sono strumeni utili per testare l'isotropia del processo; infatti, se la varianza del processo non è costante, il variogramma non mostra una convergenza verso un valore definito, quindi il processo non è né stazionario, né isotropico.

Un altro strumento che permette di studiare la dipendenza spaziale del fenomeno è la funzione di covarianza spaziale c(u)=$\sigma^2\rho$(u), che per un processo isotropico dipende solo dalla distanza Euclidea tra i siti. Quando la distanza è nulla, la covarianza coincide con la varianza, mentre all'aumentare della distanza la covarianza diminuisce e tende a zero.

Sia il variogramma che la funzione di covarianza dipendono dalla distanza Euclidea, ma hanno ruoli diversi nel processo di modellazione spaziale. Nell'approccio model-based il Variogramma ha uno scopo esplorativo e viene utilizzato per fornire i valori iniziali dei parametri della funzione di covarianza agli algoritmi impiegati nella stima del modello (Diggle

e Ribeiro, 2007), che saranno descritti successivamente nella sezione sulla stima dei parametri. La funzione di covarianza spaziale è invece utilizzata nella fase di modellazione del processo, che sarà spiegata nei paragrafi sulla modellazione spaziale.

Variogramma empirico e sample variogram

Per studiare la funzione di covarianza spaziale, date le $y_1 \ldots y_n$ osservazioni associate alle $x_1 \ldots x_n$ locazioni, viene utilizzato il variogramma empirico, dove i valori in ascissa corrispondono alle distanze Euclidee u_{ij} tra i siti xi: i=1,...,n e i valori in ordinata sono le semivariances $\gamma_{ij} = ½ (y_i - y_j)^2$. Ad ogni punto del variogramma empirico (u_{ij}, γ_{ij}) corrisponde dunque una coppia di siti (x_i, x_j). Il variogramma empirico è però di difficile interpretazione, infatti ogni misura y è affetta da errore e la varianza lungo le ascisse u_{ij} non è costante, quindi vengono applicate tecniche di smoothing, che permettono di ottenere una versione lisciata del variogramma empirico, chiamata sample variogram. Per definire il sample variogram, le distanze vengono suddivise in k intervalli regolari e per ogni intervallo di distanza viene individuato il punto medio $\tilde{u}_l$, per l=1,…,k, in corrispondenza del quale vengono calcolati i valori $\hat{\gamma}(\tilde{u}_l) = \frac{1}{2|N(\tilde{u}_l)|}\sum_{(x_i,x_j)\in N(\tilde{u}_l)}(Y_i - Y_j)^2$, dove $N|(\tilde{u}_l)|$ è l'insieme delle coppie di punti in cui la distanza u_{ij} è compresa nell'intervallo l_k.

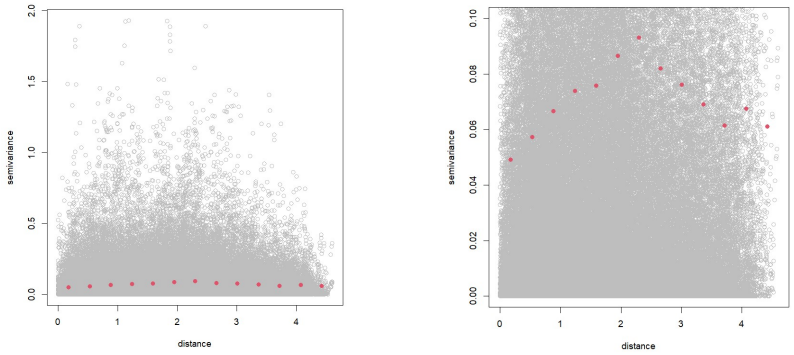

Figura 1.2 - A sinistra: variogramma empirico e corrispondente sample variogram (in rosso). A destra: stesso grafico rappresentato a sinistra, con una riduzione di scala nell'asse delle ordinate, applicata per ottenere una migliore interpretazione.

L'osservazione del sample variogram è fondamentale per studiare la struttura di covarianza del fenomeno e fare assunzioni sul segnale. Infatti se dall'osservazione del sample variogram non è possibile individuare una sill, si assume che il processo sia non stazionario; la media del processo è funzione di x e viene chiamata trend. In questo caso la correlazione spaziale non si esaurisce all'interno dell'area di studio ed è necessario modellare un trend su larga scala e

analizzare la correlazione residua rispetto al trend specificato; questa modellazione verrà trattata nei prossimi capitoli.

Modellazione spaziale di un processo non stazionario in media

Una volta effettuate le analisi esplorative sul fenomeno, è possibile definire un modello e stimarne i parametri. Per la stima del modello geostatistico vi sono due approcci: l'approccio classico e l'approccio model-based. Nel primo, chiamato anche curve fitting, la stima dei parametri del trend spaziale avviene separatamente dalla stima della funzione di covarianza, che è calcolata direttamente dal sample variogram. Un approccio alternativo è invece quello model-based, termine coniato da Diggle, Town e Moyeed (1998), dove, una volta assunto un modello esplicito per il processo spaziale, vengono stimati contemporaneamente i parametri della funzione di covarianza e quelli del trend di larga scala, utilizzando il metodo della Massima Verosimiglianza. Nell'approccio model-based il sample variogram ha solo uno scopo esplorativo e non viene utilizzato nella stima del modello, se non per definire i valori iniziali dei parametri della funzione di correlazione. In seguito verrà descritta la modellazione seguendo un approccio model-based.

Come è stato descritto nei paragrafi precedenti, il processo stocastico Y_i è definito dalla seguente relazione $Y_i = S(x_i) + Z_i \quad \forall\, i = 1, \dots, n$. Se il processo è non stazionario, il segnale $S(x_i)$ è a sua volta definito dalla somma del trend spaziale $\mu(x_i)$ e del processo spaziale residuo $\varepsilon(x_i)$: $S(x_i) = \mu(x_i) + \varepsilon(x_i) \quad \forall\, i = 1, \dots, n$. Il trend spaziale è la parte deterministica del processo e cattura le variazioni del fenomeno su larga scala, mentre il processo spaziale residuo descrive le variazioni di piccola scala rispetto alla struttura catturata dal trend e corrisponde alla parte casuale del modello. Nel caso di un processo spaziale non stazionario, $\mu(x)$ non è costante e può essere stimata in funzione delle coordinate, utilizzando metodi di regressione in cui latitudine e longitudine sono le variabili esplicative. Viene così definito il trend di larga scala su un piano in R^2. La media $\mu(x)$ in generale può essere definita da termini di primo ordine o di ordine superiore; l'aumento di termini migliora l'adattamento del modello ai dati, ma allo stesso tempo rende il modello più complesso e meno parsimonioso.

Una volta specificato il trend $\mu(x)$, si studia la componente residua del processo $\varepsilon(x)=S(x)-\mu(x)$. Se un processo $S(x)$ è tale per cui $S(x)-\mu(x)$ è stazionario, allora $S(x)$ viene definito "stazionario per covarianza" (Diggle e Ribeiro, 2007). Negli studi geostatistici la componente residua viene spesso formalizzata tramite modelli Gaussiani isotropici, specificando un modello teorico per descrivere la funzione di covarianza spaziale. Per definire un modello che catturi la struttura della covarianza spaziale nel trend di piccola scala, viene studiato il sample variogram

dei residui rispetto al trend di larga scala. Se dal sample variogram risulta una struttura di correlazione nei residui, viene ipotizzato un modello teorico di correlazione spaziale e vengono definiti i valori iniziali dei parametri del modello, che sono utilizzati negli algoritmi iterativi per la stima dei parametri. Come spiegato nei paragrafi precedenti, nell'approccio model-based il variogramma ha solo uno scopo esplorativo. Se lo scopo dell'analisi è di specificare un modello a fini previsivi, è necessario definire un modello teorico di correlazione spaziale che si adatti ai dati, che rispetti eventuali assunzioni e che dipenda da pochi parametri interpretabili. Una delle assunzioni necessarie per un modello di correlazione di un processo spaziale stazionario è che la funzione di correlazione sia definita positiva; infatti, la correlazione tra S(x) e S(x-u) decresce all'aumentare della distanza tra le due locazioni. È quindi utile considerare famiglie di funzioni di covarianza spaziale per le quali le assunzioni di positività siano già verificate.

Famiglia Matérn

Una delle famiglie di funzioni di correlazione più utilizzate è la famiglia Matérn, che prende il nome dello statistico Bertil Matérn (1960). In questa famiglia, la correlazione $\rho(u)$ è funzione di un parametro k>0 che definisce la forma della funzione, chiamata smoothness del processo, e di un parametro ϕ>0 che definisce la scala della funzione e determina il practical range.

$$\rho(u) = \frac{1}{2^{k-1}\Gamma(k)}\left(\frac{u}{\phi}\right)^{k} K_k\left(\frac{u}{\phi}\right) \quad \phi, k > 0$$

Dove K_k è la funzione di Bessel di ordine k. È necessario specificare che k e ϕ non sono parametri indipendenti, infatti il practical range è funzione di k e di conseguenza non è possibile comparare funzioni Matérn con un parametro di scala fissato e parametri di forma differenti. Alcuni dei rapporti tra k e ϕ descritti da Diggle e Ribeiro (2007) sono riportati di seguito:

- se k=0.5, practical range = 3ϕ,
- se k=1.5, practical range=4.75ϕ
- se k=2.5, practical range=5.92ϕ

Una volta fissato k è possibile definire la funzione Matérn rispetto a un solo parametro ϕ. Nel caso in cui k venga fissato a 0.5, la funzione Matérn diventa una funzione esponenziale, che all'aumentare di ϕ ha un tasso di decadimento maggiore: $\rho(u) = \exp\left(-\frac{u}{\phi}\right)$, ϕ>0. La famiglia di funzioni Matérn è la più utilizzata nelle applicazioni geostatistiche perchè è molto flessibile; infatti la presenza di due parametri permette di specificare molti modelli di covarianza diversi tra loro.

Solitamente nella modellazione dei residui, un processo spaziale con dati geostatistici viene ipotizzato Gaussiano isotropico, per cui $\varepsilon \sim N(0, \sigma^2 R)$, dove R è la matrice di correlazione spaziale i cui elementi dipendono dalla distanza Euclidea e dai parametri della funzione di correlazione k e ϕ. Se si assume questa forma per i residui, ne segue che il vettore spaziale $(Y_1 — Y_n)^T$ ha la seguente distribuzione normale: $(Y_1,...,Y_n)^T \sim N(D\beta, \sigma^2 R+\tau^2 I)$. La media μ è un vettore n-dimensionale corrispondente al prodotto tra la matrice D delle covariate e il vettore β dei coefficienti. La varianza corrisponde alla somma delle varianze di $S(x_i)$ e Z_i poiché sono processi indipendenti tra loro.

Stima dei parametri

Nel seguente capitolo vengono descritti i metodi di stima di Massima Verosimiglianza dei parametri incogniti necessari per la specificazione del modello geostatistico.

In seguto alla formulazione del modello $Y \sim N(D\beta, \sigma^2 R + \tau^2 I)$ è necessario stimare i parametri incogniti: i parametri del trend β, la varianza dell'errore di misura τ^2 e i parametri σ^2, k e ϕ dai quali dipende la matrice di covarianza Σ. Solitamente si procede stimando più modelli con diversi valori di k fissati. I modelli ottenuti vengono poi confrontati e viene scelto il modello "migliore" utilizzando alcuni criteri di selezione. In seguito ci si soffermerà sul metodo di stima di Massima Verosimiglianza, che solitamente viene scelto per la stima dei parametri, poiché generalmente le stime di Massima Verosimiglianza risultano essere consistenti, efficienti e asintoticamente normali. La stima dei parametri incogniti avviene tramite metodi iterativi di massimizzazione della funzione di log-verosimiglianza, che utilizzano parametri iniziali dedotti dal sample variogram. La funzione di log-verosimiglianza è specificata come segue: $l(\beta, \sigma^2, \phi, \tau^2) = \log\left(\prod_{i=1}^{n} f(y_i)\right) = -\frac{1}{2} n \log(2\pi) \frac{1}{2} \log\left\{|(\sigma^2 R(\phi) + \tau^2 I)|\right\} = -\frac{1}{2}(y - D\beta)^T (\sigma^2 R(\phi) + \tau^2 I)^{-1} (y - D\beta)$.

Per massimizzare la funzione di log-verosimiglianza, si fissa un valore per k e la matrice di correlazione viene riparametrizzata da un parametro V: $\Sigma = \sigma^2 V$. Il parametro V è specificato dalla matrice di correlazione R e dal parametro $v^2 = \frac{\tau^2}{\sigma^2}$, che rapporta l'errore di misurazione con la variabilità del processo. Il parametro V può quindi essere definito dalla seguente formula: $V = R + v^2 I$. Per ottenere le stime $\hat{\beta}(V)$ e $\hat{\sigma}^2(V)$ si fissa la matrice V, assegnando valori iniziali al parametro $v^2 = \frac{\tau^2}{\sigma^2}$ e al nugget ϕ, ottenuti dall'osservazione del sample variogram. Le stime $\hat{\beta}(V)$ e $\hat{\sigma}^2(V)$ vengono ottenute massimizzando la funzione di log-verosimiglianza:

$\hat{\beta}(V) = (D^T V^{-1} D)^{-1} D^T V^{-1} y$; $\hat{\sigma}^2(V) = n^{-1}\left[y - D\hat{\beta}(V)\right]^T V^{-1}\left[y - D\hat{\beta}(V)\right]$.

Sostituendo le stime $\hat{\beta}(V)$ e $\hat{\sigma}^2(V)$ nella funzione di log-verosimiglianza, si ottiene la funzione di "verosimiglianza profilo", che dipende da V, quindi dal range ϕ e dal parametro v^2:
$L_0(v^2, \phi) = \frac{1}{2}\{n\log(2\pi) + n\log\hat{\sigma}^2(V) + \log|V| + n\}$. La verosimiglianza profilo viene poi ottimizzata numericamente rispetto a ϕ e v^2 e le stime ottenute sono sostituite nuovamente nelle formule di stima di $\hat{\beta}(V)$ e $\hat{\sigma}^2(V)$.

Previsione spaziale

Nel seguente paragrafo verranno illustrati i principali metodi di kriging, termine coniato da G. Matheron (1963), che indica la previsione spaziale in ambito geostatistico. L'obiettivo del kriging è di prevedere il valore del processo S(x) in un sito x_o in cui non è stato osservato il processo. Per prevedere il processo nel punto x_o non osservato è necessario specificare un previsore puntuale, chiamato anche point predictor $\hat{T}$= t(Y), che è funzione del processo Y e restituisce il valore previsto del fenomeno in un punto x_0 non osservato. Per scegliere un previsore ottimale per fare inferenza sul processo, si cerca il $\hat{T}$ che minimizzi il mean square error MSE($\hat{T}$) = E[($\hat{T}$-T)^2]. Si dimostra che il MSE($\hat{T}$) è minimo quando $\hat{T}$= E(T|Y), per la dimostrazione si faccia riferimento a Diggle e Ribeiro (2007). L'errore della previsione $\hat{T}$ è chiamato varianza di previsione Var(T|Y) e corrisponde alla Varianza il cui valore atteso è il MSE.

Per specificare il previsore $\hat{T}$ e il MSE($\hat{T}$) nel caso di un modello Gaussiano, è necessario definire il valore atteso E(T|Y) e la varianza Var(T|Y). Come descritto da Diggle e Ribeiro (2007), data una variabile casuale multivariata X = (X_1, X_2) con distribuzione congiunta Gaussiana con media μ = (μ_1, μ_2) e covarianza $\boldsymbol{\Sigma}$=$\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$, la condizionata $X_1|X_2$ si distribuisce come una Normale con media $\mu_{1|2}$ e covarianza $\boldsymbol{\Sigma}_{1|2}$ con $\mu_{1|2} = \mu_1+\Sigma_{12}{\Sigma_{22}}^{-1}(X_2 - \mu_2)$ e $\boldsymbol{\Sigma}_{1|2} = \Sigma_{11} - \Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{21}$. Ne segue che, dato il processo spaziale Y ~ N(Dβ ,σ^2R+τ^2I), il valore del previsore è $\hat{T}$= E(T|Y)= μ+$r^T V^{-1}$(Y-μ), dove r è un vettore n-dimensionale e corrisponde alla funzione di correlazione: $r = [\hat{\rho}(||x - x_1||), \ldots, \hat{\rho}(||x - x_n||)]^T$. Inoltre nel modello Gaussiano, la varianza previsiva Var(T|Y) = σ^2(1-$r^T V^{-1}$r) e non dipende da Y, infatti Var(T|Y)= MSE($\hat{T}$), quindi la varianza previsiva coincide con il Mean Square Error di $\hat{T}$. La previsione S(x) del processo può essere scritta come combinazione lineare delle osservazioni Y_i con pesi a_i(x), chiamati pesi di kriging, assegnati ad ogni osservazione i=1,..,n. Questi pesi dipendono da $r^T V^{-1}$, quindi dalle stime dei parametri della funzione di covarianza.

La previsione S(x) è dunque definita dalla seguente combinazione lineare: $S(x) = \mu + \sum_{i=1}^{n} a_i(x)(Y_i - \mu)$. Il fattore ($Y_i$-$\mu$) specifica che la previsione nel punto x_i dipende sia dall'informazione ottenuta dai dati osservati, che dalla media del processo. Infatti $S(x_0)$ varia sia al variare del punto x_0 in cui si vuole stimare il fenomeno, che al variare delle osservazioni $x_1 \ldots x_n$ in cui si è stimato il fenomeno. Se la correlazione tra x_0 e i punti osservati è vicina allo zero, viene data più importanza alla media; se al contrario la correlazione ha valori alti, viene data più rilevanza ai dati osservati.

Le previsioni spaziali dipendono anche dai parametri di correlazione k, ϕ e τ^2. In particolare, al variare del parametro di forma k, varia la smoothness del processo: per un k maggiore la forma del processo tra i punti osservati è più "lisciata" rispetto alla forma di un processo stimato con k minore. Al variare del parametro di scala ϕ varia il range: all'aumentare di ϕ aumenta la distanza in corrispondenza della quale la correlazione si annulla, si invece il range è molto piccolo la correlazione si annullerà anche per siti molto vicini. Al variare del nugget τ^2 si ha un impatto sulla forma del previsore in corrispondenza dei punti osservati: all'aumentare del nugget, la previsione in un punto x_i osservato, si sposta verso la media generale più che verso la misurazione del fenomeno nel punto osservato. Per un processo Gaussiano con τ^2=0, in cui μ è assunta costante e nota, la previsione prende il nome di Simple Kriging e lo stimatore viene specificato dalla seguente formula: $\hat{T} = \mu r^T V^{-1}(y-\mu)$. Questo stimatore è un interpolatore definito "esatto" se la stima $\hat{S}(x_i)$ del processo in un punto x_i campionato coincide con la misura y_i, oppure se la varianza della previsione è nulla in ogni punto x, quindi se τ^2 è nulla.

2 Analisi esplorative per dati geostatistici spazio-temporali

Nel seguente capitolo sono descritti i metodi di esplorazione di dati geostatistici definiti nello spazio e nel tempo. Nei prossimi paragrafi sono specificati i processi stocastici Gaussiani spazio-temporali per dati geostatistici e i metodi di studio della struttura di covarianza spazio-temporale. Per la stesura di questo capitolo e per l'analisi spazio-temporale sui dati di PM_{10} è stato fatto riferimento principalmente all'articolo "Analysis of Random Fields Using CompRandFld", di S. A. Padoan e M. Bevilacqua (2015).

Processi stocastici Gaussiani stazionari spazio-temporali

Per studiare i fenomeni spazio-temporali vengono analizzati dati geostatistici definiti sia nello spazio che nel tempo. Questi dati sono specificati da coordinate spaziali x_i e da un istante temporale t_p per i=1,...,n e p=1,...,m. Ogni misura y_{ip} del fenomeno, associata alla coppia (x_i,t_p), è la realizzazione di un processo stocastico spazio-temporale Y(x,t) definito sul dominio I=SxT, dove S è il dominio spaziale $\subseteq R^d$ e T è il dominio temporale $\subseteq R$. In generale un fenomeno spazio-temporale è specificato dalla seguente formula: $Y(x_i,t_p)=S(x_i,t_p)+Z_{ip}$. Analogamente al caso spaziale, $Z_{ip}(x_i,t_p)\sim N(0,\tau^2)$ è un processo white noise indipendente dal fenomeno e S(x,t) è un processo stocastico spazio-temporale, che nel seguente lavoro di tesi viene sempre assunto Gaussiano stazionario ed è quindi definito esclusivamente dalla media e dalla struttura di covarianza del processo stesso. Per la proprietà di stazionarietà, la media e varianza del processo sono costanti, mentre la covarianza è invariante per traslazione, quindi dipende solo dalla distanza spaziale $u=x_i-x_j$ e dalla distanza temporale $h=t_p-t_q$.

Per il processo Y(x,t) valgono le seguenti formule: $E[Y(x,t)]=E([S(x,t)]=\mu$, $Var[S(x,t)]=\sigma^2=c(0,0)$, $Var[Y(x,t)]=\omega^2=\tau^2+\sigma^2$, $Cov[Y(x_i,t_p),Y(x_j,t_q)]=c(u,h)$, $Corr[Y(x_i,t_p),Y(x_j,t_q)]=\rho(u,h)=c(u,h)/c(0,0)$.

A differenza della proprietà di stazionarietà, solitamente per un processo spazio-temporale l'assunzione di isotropia descritta nel caso spaziale non è adeguata, perché a differenza dello spazio, il tempo ha una direzione definita (Gneiting e Guttorp, 2010). Un processo spazio-temporale viene quindi definito "spazialmente isotropico" se la componente spaziale è invariante per rotazione per un istante temporale fissato. Sotto questa assunzione, la covarianza $c(u,h_0)$ dipende solamente dalla distanza Euclidea $||u||$ per ogni istante temporale h_0 fissato.

Struttura di covarianza e variogramma spazio-temporale

Analogamente al caso spaziale, per studiare il fenomeno spazio-temporale è fondamentale analizzare la dipendenza spaziale e temporale tra le diverse osservazioni, che può essere specificata dalla funzione di covarianza spazio-temporale c(u,h) e dal variogramma spazio-temporale γ(u,h). La funzione di covarianza viene definita dalla seguente formula: γ(u,h)=c(0,0)-c(u,h)=σ^2-c(u,h)= σ^2[1- ρ(u,h)], mentre il variogramma empirico spazio temporale viene calcolato con la formula $\hat{\gamma}(u_{ij}, h_{pq}) = \frac{1}{2|N(u_{ij},h_{pq})|}\sum_{(x_{ip},x_{jq})\in N(u_{ij},h_{pq})}(Y_{ip} - Y_{jq})^2$, dove $N|(u_{ij}, h_{pq})|$ è l'insieme delle coppie di osservazioni con distanza spaziale $u_{ij} = ||x_i - x_j||$ e distanza temporale $h_{pq} = ||t_p - t_q||$. Per interpretare al meglio il variogramma viene poi calcolato il relativo sample variogram: le distanze spaziali vengono suddivise in intervalli con valori centrali $\tilde{u}_l$ per l=1,…,k e ad ogni coppia ($\tilde{u}_l$, h_{pq}) è associata alla media dei valori del variogramma empirico contenuti nell'intervallo delle distanze spaziali u_{ij} alla distanza temporale h_{pq}.

Per un'analisi più accurata della componente temporale del fenomeno è anche possibile calcolare il variogramma empirico temporale utilizzando gli stessi metodi descritti per il variogramma spaziale nei paragrafi precedenti. A differenza del caso spaziale non vengono però calcolati gli intervalli di distanza e il sample variogram è funzione delle stesse distanze temporali h_{pq}: $\hat{\gamma}(h_{pq}) = \frac{1}{2|N(h_{pq})|}\sum_{(x_p,x_q)\in N(h_{pq})}(Y_p - Y_q)^2$, dove $N|(h_{pq})|$ è l'insieme delle coppie di punti con distanza temporale h_{pq}.

Lo studio di questi sample variogram risulta utile per stimare due modelli, uno spaziale e uno temporale, oppure un unico modello spazio-temporale, con l'obiettivo di ottenere una previsione spazio-temporale dei dati. In questo lavoro di tesi non viene approfondita la stima dei modelli di correlazione spazio-temporali, né la previsione spazio-temporale; per un approfondimento del tema si consultino i lavori di Porcu, Mateu e Christakos (2009) o di Cressie and Wikle (2011).

3 Applicazione dei metodi: analisi della concentrazione giornaliera di PM_{10} in Italia

Nel seguente capitolo le principali tecniche di analisi spazio-temporale e di previsione spaziale illustrate nei capitoli precedenti verranno applicate a dati geostatistici di PM_{10}.

Introduzione

Lo scopo dell'analisi è di studiare la variabilità spaziale della concentrazione giornaliera di particolato atmosferico (PM_{10}) nell'anno 2015, misurata in 500 centraline collocate in Italia, con l'obiettivo di specificare un modello geostatistico adeguato e stimare la previsione spaziale dell'inquinante su tutto il territorio italiano. L'analisi esplorativa spaziale è stata effettuata su quattro giornate, con l'obiettivo di analizzare anche la differenza della distribuzione dell'inquinante in diversi istanti temporali. Successivamente è stata effettuata un'analisi esplorativa spazio-temporale sull'intero anno di studio. Lo studio della distribuzione spaziale e spazio-temporale del PM_{10} è fondamentale per prendere adeguate decisioni sulle politiche di mitigazione dell'inquinamento e per valutare l'esposizione della popolazione a questo inquinante. La seguente analisi ha però prettamente lo scopo di illustrare i modelli geostatistici, infatti il modello previsivo utilizzato non è il più adeguato per effettuare la previsione su questi dati. Inoltre, le covariate ambientali non sono state introdotte nella fase di previsione spaziale, né nell'analisi spazio-temporale.

Materiali e Metodi

Il dataset utilizzato (pm10) contiene i valori delle concentrazioni medie giornaliere di PM_{10} ($\mu g/m^3$) nel 2015, misurate dall'Agenzia Regionale per le Politiche Ambientali (ARPA) e raccolte dall'Istituto Italiano per la Protezione e la Ricerca (ISPRA). Il dataset, contenente misurazioni dell'inquinante su oltre 500 siti e i valori dei alcune covariate ambientali, è stato ottenuto dal materiale dello studio "Spatio-temporal modelling of PM10 daily concentrations in Italy using the SPDE appraoch" (Fioravanti et al., 2020) presente nella pagina Github di G. Fioravanti (https://github.com/guidofioravanti/spde_spatio_temporal_pm10_modelling_italy). Nella seguente analisi spaziale è stata studiata la variabilità spaziale della concentrazione di PM_{10} nelle giornate del 20 marzo, del 21 giugno, del 23 settembre e del 22 dicembre 2015 e dal dataset originale sono state selezionate le seguenti variabili:

1- 'pm10': concentrazione media giornaliera di pm10 ($\mu g/m^3$)

2- 'x': coordinate spaziali di longitudine del sito osservazionale

3- 'y': coordinate spaziali di latitudine del sito osservazionale

4- 'q_dem': altitudine (metri) del sito di monitoraggio

5- ‘d_a1’: distanza in linea d’aria dalla strada più vicina (metri)
6- ‘sp’: pressione atmosferica (hPa)
7- ‘tp’: precipitazioni totali (mm)
8- ‘t2m’: temperatura media a 2 metri (°C)
9-‘pbl00’: strato limite planetario alle 00:00 (metri)
10-‘pbl12’: strato limite planetario alle 12:00 (metri)
11-‘aod550’: profondità ottica a 550 nm (nm)
12-‘i_surface’: impeabilizzazione dei suoli (%)

Per la rappresentazione grafica della concentrazione di inquinante è stata utilizzata la mappa delle regioni italiane dell’Istituto Nazionale di Statistica (Istat). L’analisi esplorativa dei dati, la modellazione e la previsione sono state effettuate utilizzando il software R. I principali pacchetti utilizzati sono: geoR, devtools, FMCC, rgdal, lubridate, raster, CompRandFld. Lo studio spaziale della concentrazione del PM_{10} per ogni giornata è suddiviso in due sezioni: una prima parte di analisi esplorativa spaziale dei dati e di stima di un modello geostatistico e una seconda parte di previsione spaziale. Nell’analisi esplorativa spaziale è stato studiato l’andamento dei dati originali rispetto alle coordinate spaziali ed è stata ricercata la presenza di un trend spaziale di larga scala. In seguito, è stata studiata la correlazione spaziale su piccola scala tramite lo studio dei sample variogram per specificare un modello geostatistico. Utilizzando un approccio model-based, descritto precedentemente nei paragrafi sulla stima dei parmetri, è stata fatta inferenza sui parametri della funzione teorica di covarianza ipotizzata e sui parametri del trend di larga scala, con il Metodo della Massima Verosimiglianza. Successivamente, utilizzando un modello specificato dalle sole coordinate spaziali, omettendo le covariate ambientali, è stata effettuata la previsione spaziale sul territorio italiano con il metodo del simple kriging. In seguito alle analisi spaziali, la concentrazione di inquinante è stata studiata da un punto di vista spazio-temporale tramite un’analisi esplorativa dei dati giornalieri medi per ogni mese dell’anno.

Analisi esplorative spaziali

Per studiare la concentrazione dell’inquinante sia da un punto di vista sia spaziale, che spazio-temporale, le analisi esplorative sono state eseguite sui dati originali di quattro giornate: gli equinozi e i solstizi nell’anno 2015. Di seguito sono riportati i valori sintetici di PM_{10} nelle diverse giornate.

	Minimo	1° quartile	Mediana	Media	3° quartile	Massimo
20 marzo	4.90	28.65	39.40	42.10	53.00	104.00
21 giugno	1.00	8.00	10.29	11.92	13.99	54.80
23 settembre	1.00	11.00	14.60	15.15	18.00	44.30
22 dicembre	2.00	30.77	50.00	53.45	73.44	182.90

Tabella 3.1 - Statistiche spaziali sintetiche della concentrazione di PM_{10}.

Confrontando le statistiche sintetiche dei dati nei quattro giorni, si nota che il range dei valori di concentrazione di PM_{10} misurati al 22 dicembre è maggiore del range dei valori misurati negli altri tre giorni e la media è maggiore nel giorno di dicembre. Nelle giornate del 21 giugno e del 23 settembre i valori sono più bassi rispetto a marzo e dicembre; infatti, sia la media che la mediana del PM_{10} al 22 dicembre sono superiori al valore massimo raggiunto dal PM_{10} il 23 settembre.

Modellazione spaziale

Per ogni giornata è stata fatta un'analisi esplorativa dei dati originali, dai quali sono risultate distribuzioni positivamente asimmetriche. Quindi sono stati studiati i residui rispetto a diversi trend spaziali, osservando il grafico dei quartili, l'istogramma della distribuzione e i grafici della correlazione tra il PM_{10} e ognuna delle coordinate spaziali. Dalle analisi ottenute, si può ipotizzare che i residui siano incorrelati rispetto a entrambe le coordinate spaziali per ogni giornata, tranne nella giornata di giugno, dove si può osservare un trend decrescente rispetto alla latitudine. Gli istogrammi sono incentrati sullo zero e le forme sono simili a una Normale. La distribuzione spaziale dei quartili di concentrazione sembra essere leggermente differente nelle quattro giornate, con qualche similarità in giugno e dicembre.

Per un'analisi più accurata del trend di larga scala, sono stati stimati più modelli di primo e secondo ordine, che differiscono per l'inclusione o l'esclusione di alcune covariate ambientali. Per le analisi successive sono stati scelti i modelli con un indice R^2 corretto maggiore. Nei modelli selezionati, l'indice R^2 di marzo e giugno è superiore a 0.6, mentre a settembre e a dicembre è inferiore a 0.5, quindi il modello non spiega esaustivamente la variabilità del PM_{10}. In particolare il modello per il mese di settembre ha un indice R^2 di 0.2823, dunque il trend ipotizzato non è adeguato a catturare la variabilità del fenomeno su larga scala. Nei modelli stimati, le coordinate spaziali non hanno lo stesso effetto sulla concentrazione di PM_{10}. La longitudine ha una relazione positiva con la concentrazione di PM_{10} sia nella giornata di marzo che in quella di giugno, mentre nella giornata di settembre vi è una relazione negativa. La

latitidune ha invece una relazione positiva con la concentrazione di PM_{10} nelle giornate di marzo e dicembre, mentre nella giornata di giugno ha una relazione negativa.

Dopo la modellazione del trend di larga scala, è stata studiata la correlazione spaziale dei residui rispetto al trend specificato, con l'obiettivo di definire un modello teorico per la correlazione spaziale su piccola scala. Di seguito sono riportati i grafici dei variogrammi empirici e dei sample variogram per ognuna delle giornate analizzate, specificati utilizzando i dati standardizzati. Nello studio dei sample variogram è opportuno fare riferimento soprattutto all'andamento dei variogrammi per distanze non troppo elevate, infatti si può notare dalla nuvola di punti del variogramma empirico che in prossimità di grandi distanze vi sono meno coppie di osservazioni, il che rende la stima del variogramma instabile.

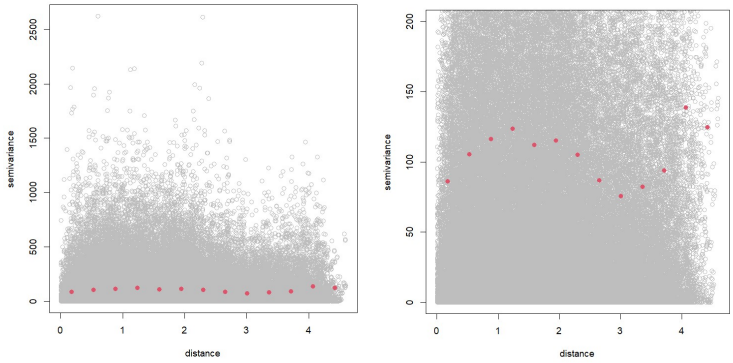

Figura 3.1 A sinistra: variogramma empirico e sample variogram (in rosso) dei residui del un trend di second'ordine specificato dalle coordinate spaziali e dalle covariate ambientali per la giornata del 20 marzo. A destra: stesso grafico rappresentato a sinistra, con una riduzione di scala nell'asse delle ordinate, applicata per ottenere una migliore interpretazione.

Dall'osservazione del sample variogram della giornata di marzo, si può assumere che i residui del trend specificato si distribuiscano come un processo Gaussiano isotropico. Infatti è possibile definire una sill, che viene raggiunta ad una distanza di poco superiore all'unità. Il variogramma non si assesta sul valore della sill, ma l'andamento del variogramma per distanze elevate è meno rilevante per le ragioni specificate precedentemente.

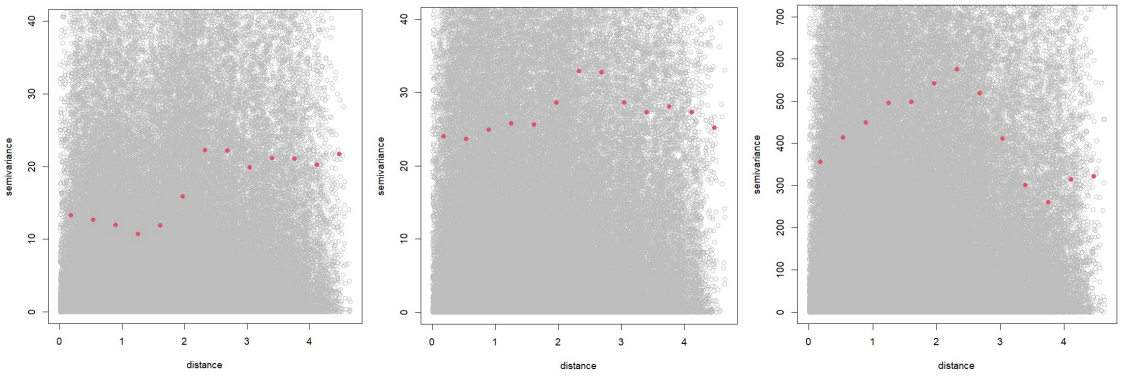

Figura 3.2 Variogrammi empirici e sample variograms dei residui (in rosso) per il 21 giugno (a sinistra), per il 23 settembre (al centro) e per il 22 dicembre (a destra).

Nei sample variograms delle giornate di giugno e settembre non è possibile individuare una sill, dunque i residui non rispettano l'assunzione di stazionarietà del processo. Nel sample variogram della giornata di dicembre è possibile individuare una sill ed è dunque stato assunto un processo Gaussiano isotropico per i residui. In seguito all'analisi dei sample variogram sono stati specificati i modelli per le giornate del 20 marzo e del 22 dicembre, dove è possibile assumere un processo Gaussiano isotropico per i residui. Il trend di larga scala per le giornate di giugno e settembre non è stato specificato correttamente; per uno studio più approfondito di queste osservazioni sarebbe necessario inserire nel modello informazioni aggiuntive.

Sia per la giornata di marzo che per quella di dicembre, la funzione teorica di correlazione ipotizzata è una funzione Matérn. Dai sample variogram sono stati dedotti i valori per la varianza d'errore τ^2, per la varianza σ^2 e per il range ϕ, con il quale è stato calcolato il practical range, come descritto nel paragrafo sulla famiglia di correlazione Matérn. Questi valori sono stati utilizzati come valori iniziali negli algoritmi iterativi per la stima dei parametri, effettuata con il Metodo della Verosimiglianza, come descritto nel paragrafo sulla stima dei parametri. Seguendo l'approccio model-based, sono stati infatti stimati contemporaneamente i parametri del trend di larga scala e della funzione teorica di correlazione per ogni giornata, specificando tre modelli con parametri k differenti. Di seguito sono riportati i coefficienti stimati per i modelli con parametro k fissato a 0.5.

Intercetta	**x**	**y**	$\mathbf{x^2}$	$\mathbf{y^2}$	$\mathbf{x * y}$	**d_a1**	**t2m**
-134.1383	90.4631	-12.8825	-0.7916	0.9314	-5.4733	0.0000	-2.6014
tp	**q_dem**	**sp**	**pbl00**	**pbl12**	**i_surface**	**aod550**	
0.8545	-0.0095	0.1531	0.0214	0.0037	0.0548	30.7201	
$\boldsymbol{\tau^2}$	$\boldsymbol{\sigma^2}$	$\boldsymbol{\phi}$	**Practical range**				
36.1216	77.8601	0.2331	0.6984473				

Tabella 3.2 – Stima dei parametri del modello del 20 marzo.

Intercetta	**x**	**y**	$\mathbf{x^2}$	$\mathbf{y^2}$	$\mathbf{x * y}$	**d_a1**	**t2m**
-871.5610	27.4926	72.2194	3.3708	-1.7531	-2.6234	0.0000	-5.4755
tp	**q_dem**	**sp**	**pbl00**	**pbl12**	**i_surface**	**aod550**	
-1.7486	-0.0378	0.2697	0.0387	-0.0038	0.1622	33.4156	
$\boldsymbol{\tau^2}$	$\boldsymbol{\sigma^2}$	$\boldsymbol{\phi}$	**Practical range**				
96.6657	346.0573	0.1196	0.358239				

Tabella 3.3 – Stima dei parametri del modello del 22 dicembre.

La rappresentazione dell'andamento dei valori di PM_{10} rispetto ai corrispettivi valori stimati è illustrata nella figura 3.3.

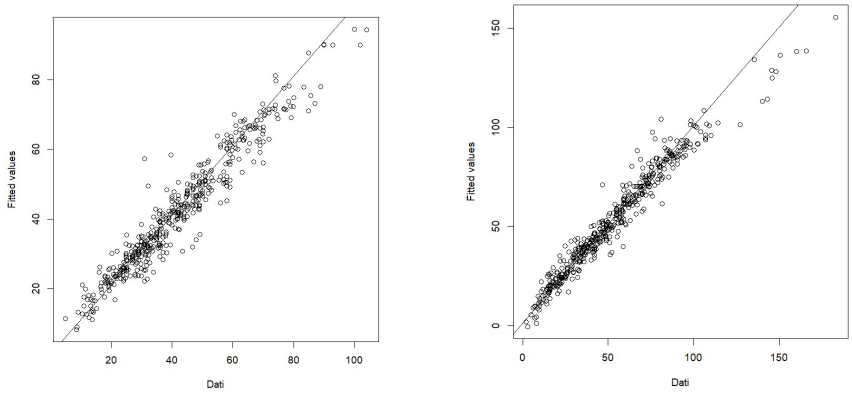

Figura 3.3 - Confronto tra valori osservati e stimati al 20 marzo (sinistra) e 22 dicembre (destra).

In entrambi i grafici (Figura 3.3) i punti sono posizionati vicino alla bisettrice, quindi le stime del modello si avvicinano ai valori dei dati osservati. Nel grafico relativo al 22 dicembre i punti sembrano essere più vicini alla retta rispetto ai valori del 20 marzo. In ambedue i giorni, per valori alti di PM_{10} i punti tendono a discostarsi maggiormente dalla bisettrice; infatti, i valori stimati tendono ad essere più bassi di quelli misurati. I modelli specificati non sono stati

utilizzati per la previsione spaziale per mancanza di dati relativi alle covariate nello spazio di previsione.

Previsione spaziale

In seguto alle analisi di esplorazione dei dati e di modellazione geostatistica, sono stati specificati due modelli per la previsione spaziale nelle giornate del 20 marzo e del 22 dicembre. Come anticipato nei paragrafi precedenti, per la previsione spaziale sono stati formalizzati due modelli in cui la variabilità di larga scala è stata definita da un trend lineare di second'ordine senza le covariate ambientali, per mancanza dei dati delle covariate sulle locazioni di previsione. La previsione spaziale è stata effettuata su una griglia di 251001 locazioni di previsione. Per ogni giornata è stato stimato un trend spaziale di second'ordine, funzione delle sole coordinate, con una struttura di correlazione definita da una funzione Matérn con k=0.5, ipotizzando come distribuzione residua un processo Gaussiano isotropico. Di seguito sono riportate le stime dei parametri del modello e il valore di Massima Verosimiglianza ottenuti tramite i metodi di stima descritti nel primo capitolo.

Intercetta	**x**	**y**	$\mathbf{x^2}$	$\mathbf{y^2}$	$\mathbf{x * y}$	$\boldsymbol{\tau^2}$	$\boldsymbol{\sigma^2}$	$\boldsymbol{\phi}$
-951.928	14.222	81.258	-1.329	-1.390	-7.956	37.081	106.430	0.214
Practical range		Massima verosimiglianza						
0.6410058		-1663						

Tabella 3.4 - Stime dei parametri del modello per la previsione spaziale del 20 marzo.

Intercetta	**x**	**y**	$\mathbf{x^2}$	$\mathbf{y^2}$	$\mathbf{x * y}$	$\boldsymbol{\tau^2}$	$\boldsymbol{\sigma^2}$	$\boldsymbol{\phi}$
-1093.444	49.725	113.096	-4.373	-2.932	-0.980	95.015	623.588	0.131
Practical range		Massima verosimiglianza						
0.3923255		-2067						

Tabella 3.5 - Stime dei parametri del modello per la previsione spaziale del 22 dicembre.

Successivamente sono state calcolate le previsioni per le locazioni non osservate e le varianze di previsione, che nel caso della distribuzione Gaussiana coincidono con il Mean Square Error di previsione, come descritto nei capitoli precedenti. Dal simple kriging sono state ottenute le mappe di previsione per entrambi i giorni e le mappe relative all'errore di previsione.

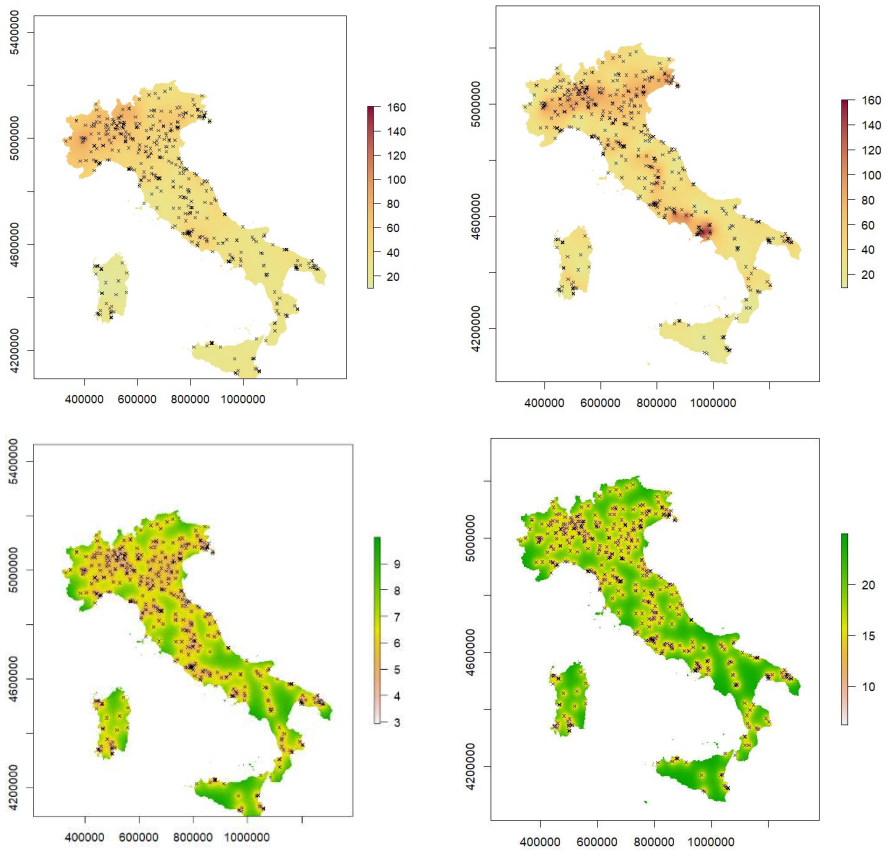

Figura 3.4 - Mappe di previsione spaziale del 20 marzo (a sinistra in alto) e del 22 dicembre (a destra in alto). Mappe dell'incertezza per le previsioni del 20 marzo (a sinistra in basso) e del 22 dicembre (a destra in basso).

Confrontando le mappe di previsione, si osserva che la distribuzione della concentrazione di PM_{10} nel territorio italiano differisce visibilmente nelle due giornate analizzate. Al 20 marzo la più alta concentrazione si trova nell'area Nord-Ovest, mentre nel resto del territorio italiano la concentrazione è piuttosto bassa e uniforme. Al 22 dicembre le zone con maggiore concentrazione si vedono nell'area di Roma e di Napoli, mentre nella zona della Pianura Padana i valori sono più alti che nel resto dell'Italia, ma non raggiungono valori eccessivamente elevati. Inoltre la scala di valori per la giornata di marzo indica con il colore rosso le concentrazioni di inquinante maggiori, che si trovano tra i 60 e i 100 $\mu g/m^3$, mentre nella giornata di Dicembre, le concentrazioni maggiori corrispondono a valori tra i 100 e i 160 $\mu g/m^3$. Le mappe degli errori indicano la distribuzione della dell'errore di previsione. Si nota che nei punti di previsione lontani dalle centraline l'incertezza è molto alta. In particolare, nella mappa relativa al 22 dicembre la deviazione standard è tra i valori 15 e 25 $\mu g/m^3$ su quasi tutto il territorio, mentre nelle analisi relative al 22 marzo l'incertezza è minore e raggiunge al massimo

un valore di 10 μg/m^3. Un'incertezza di queste dimensioni associata a valori di PM_{10} tra i 20 e i 60 μg/m^3 rende la previsione poco esaustiva.

Correlazione spazio-temporale

Il confronto tra i dati nelle giornate di marzo, giugno, settembre e dicembre ha mostrato una netta differenza nei livelli di concentrazione dell'inquinante e nella distribuzione spaziale tra le diverse giornate dell'anno. Ma nell'ottica di un'analisi sapzio-temporale, lo studio di sole quattro giornate risulta essere poco soddisfacente. Per approfondire lo studio sulla distribuzione spaziale di inquinante in funzione del tempo, il dataset originale è stato aggregato ottenendo le medie mensili di PM_{10} in 470 siti spaziali, quindi su 12 istanti temporali. Di seguito sono riportate le statistiche descrittive mensili e i box-plot dei dati aggregati, dalle quali è possibile effettuare una prima analisi temporale.

	Minimo	**1° quartile**	**Mediana**	**Media**	**3° quartile**	**Massimo**
Gennaio	2,00	22,69	32,66	34,66	45,40	105,13
Febbraio	3,185	20,727	28,039	30,791	39,625	74,143
Marzo	5,444	22,258	27,794	28,957	35,852	56,287
Aprile	5,70	17.43	21.35	21,45	25,24	41,36
Maggio	5,323	16,492	19,414	19,713	22,944	43,350
Giugno	6,633	16,623	19,535	19,824	22,993	41,137
Luglio	6,718	20,931	24,214	24,723	28,308	47,616
Agosto	9,129	17,996	20,869	21,326	24,279	46,800
Settembre	5,40	15,88	19,17	19,76	23,32	50,58
Ottobre	3,194	15,337	19,855	21,434	27,110	47,291
Novembre	2,036	22,899	32,341	33,909	43,845	77,033
Dicembre	3.032	30,046	46,673	46,525	59,528	158,276

Tabella 3.6 - Statistiche temporali sintetiche della concentrazione di PM_{10} per ogni mese dell'anno 2015.

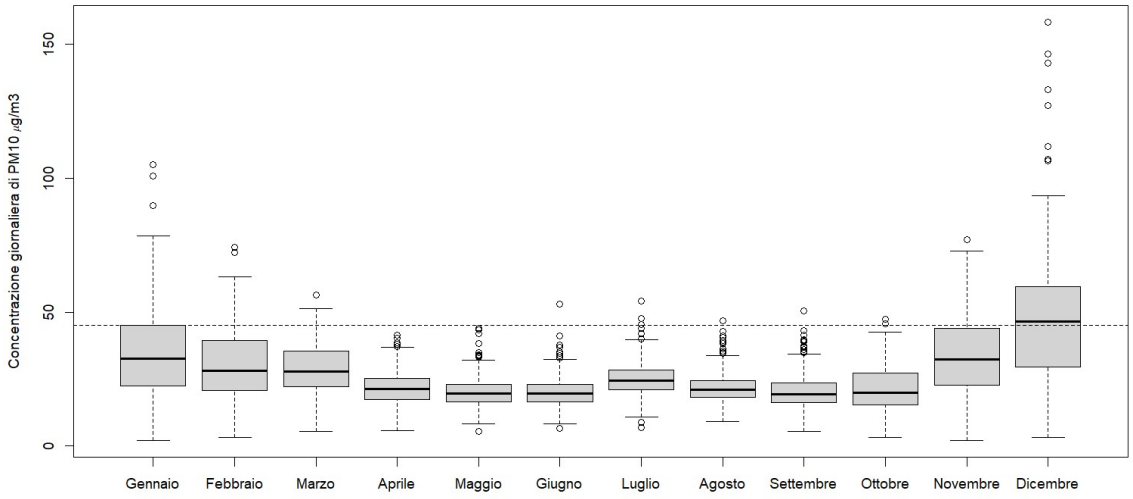

Figura 3.5 – Box-plot delle medie mensili della concentrazione di PM_{10} nell'anno 2015. La linea tratteggiata indica l'esposizione giornaliera al PM_{10} oltre la quale vi sono effetti sulla salute (45 $\mu g/m^3$) (OMS Ufficio Regionale per l'Europa, 2022).

Dalle statistiche descrittive e dall'osservazione del box-plot è ben visibile un andamento decrescente della concentrazione di inquinante dal mese di gennaio al mese di aprile, oltre il quale vi è un andamento piuttosto costante fino ad ottobre. Nel mese di novembre e dicembre vi è invece un aumento di inquinante, che supera i livelli di gennaio. La linea tratteggiata presente nel grafico corrisponde all'esposizione giornaliera al PM_{10} oltre la quale vi sono effetti sulla salute (45 $\mu g/m^3$), secondo i riferimenti dell'Organizzazione Mondiale della Sanità (OMS Ufficio Regionale per l'Europa, 2022). Per un'analisi più approfondita della correlazione spazio-temporale sono stati costruiti i sample variogram spaziali, temporali e spazio-temporali ottenuti tramite i metodi descritti nel secondo capitolo, senza introdurre nell'analisi le covariate ambientali. I grafici sono stati calcolati utilizzando il pacchetto R CompRandFld, assegnando ad ogni mese un valore da 1 a 12 in ordine crescente partendo da gennaio. Le distanze temporali sono state quindi calcolate tramite la differenza tra questi valori.

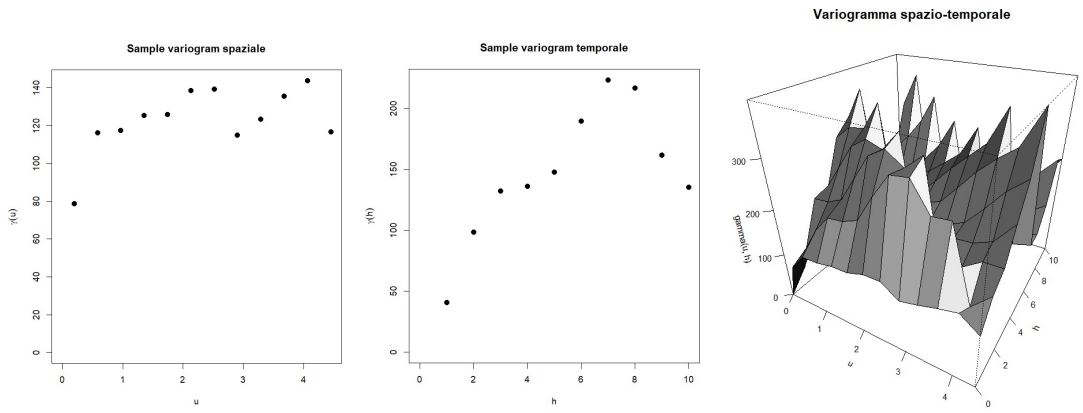

Figura 3.6 A sinistra: sample variogram spaziale. Al centro: sample variogram temporale. A destra: Sample variogram spazio-temporale.

Il sample variogram spaziale ha una forma riconducibile alla famiglia di funzioni di correlazione Matérn ed è possibile assumere isotropia spaziale per il processo, infatti cresce fino alla distanza 2 per poi stabilizzarsi tra i valori 120 e 140, considerando che all'aumentare della distanza vi sono meno coppie di osservazioni coinvolte quindi la stima del variogramma è meno precisa. Il sample variogram temporale invece ha un andamento monotono crescente fino alla distanza 7, in corrispondenza della quale si ottiene il valore massimo, per poi discendere nuovamente. L'andamento del variogramma temporale è in linea con le analisi esplorative, nelle quali si è osservato che la concentrazione di PM_{10} nei mesi invernali è più alta rispetto a quella nei mesi estivi, infatti alla distanza temporale 7 corrispondono coppie di osservazioni misurate a poco più di mezzo anno di distanza, quindi in stagioni diverse, mentre le coppie misurate a una distanza temporale minore o maggiore di 7 si trovano in stagioni dell'anno più vicine e simili.

Come descritto nel capitolo sui metodi di analisi spazio-temporale, i grafici ottenuti permettono di fare ipotesi sulla struttura spazio-temporale più adatta ai dati presi in esame, sono quindi utili per effettuare un'analisi esplorativa preliminare alla stima di un modello spazio-temporale. La stima del modello non sarà sviluppata in questo lavoro di tesi, ma per un'approfondimento sulla previsione spazio-temporale effettuata su questi dati è possibile consultare il lavoro di Fioravanti G. citato in precedenza.

Conclusioni

Nella parte applicativa del lavoro di tesi è stata studiata la concentrazione giornaliera di PM_{10} in Italia. Dall'analisi preliminare dei dati in quattro giornate rappresentative dell'anno 2015 si è potuta osservare una differenza nella concentrazione spaziale di PM_{10} tra le giornate analizzate, sia per la distribuzione geografica, che per i livelli di concentrazione raggiunti: le giornate di marzo e dicembre presentano concentrazioni di inquinante mediamente maggiori rispetto alle giornate di giugno e settembre. Nella giornata di marzo vi sono alte concentrazioni della zona nord-ovest della penisola, mentre nelle giornate di giugno e settembre vengono raggiunte alte concentrazioni anche nelle zone meridionali del paese e nelle isole; il 22 dicembre invece le concentrazioni maggiori si osservano nell'area della Pianura Padana e in alcune zone del centro-Italia. È stata rilevata la presenza di un trend di larga scala per le diverse giornate, determinato in modo significativo dalle covariate ambientali inserite nell'analisi. Nello studiare la correlazione spaziale della componente di piccola scala è stato assunto un processo Gaussiano isotropico per i residui rispetto al trend di larga scala solo nelle giornate di marzo e dicembre, per cui è stato deciso di proseguire con la modellazione e la previsione di queste due giornate. L'analisi del variogramma effettuata in diversi istanti temporali ha mostrato risultati molto diversi nelle due giornate; questo suggerisce che il processo generatore dei dati è non stazionario nel tempo. Per un'analisi spaziale più accurata sarebbe necessario inserire nel modello altre covariate ambientali e utilizzare un complesso modello spazio-temporale sfruttando le informazioni di tutto il dataset. Con un approccio model-based sono stati stimati i parametri dei trend di larga scala dei due modelli e delle funzioni di correlazione Matérn per la componente di piccola scala. Infine, la previsione spaziale è stata effettuata utilizzando un modello in cui la concentrazione dell'inquinante è stata specificata solo in funzione delle coordinate spaziali, infatti le covariate ambientali non sono state inserite nel modello previsivo. Per un'analisi previsiva più accurata sarebbe necessario reperire le informazioni ambientali per ogni punto di previsione, in modo da poter inserire nel modello previsivo anche le covariate. Le mappe previsive ottenute mostrano una distribuzione della concentrazione del PM_{10} in Italia molto differente nelle due giornate: nella giornata di marzo la concentrazione risulta essere molto alta nell'area nord-ovest della penisola, mentre nel mese di dicembre i valori maggiori si hanno vicino alle città di Napoli e Roma e nella zona della Pianura Padana. I risultati previsivi non sono però soddisfacenti, in quanto l'incertezza associata alla previsione raggiunge valori molto alti. Una possibile soluzione potrebbe essere quella di inserire per ogni punto di previsione le misure delle covariate ambientali risultate più significative dalla modellazione del fenomeno. Nell'ultima analisi esplorativa spazio-temporale effettuata è stato analizzato

l'andamento della concentrazione di PM_{10} durante tutto l'anno. Dai risultati ottenuti si osserva che la concentrazione di inquinante è generalmente maggiore nei mesi invernali, mentre diminuisce nettamente tra il mese di marzo e il mese di ottobre, infatti le osservazioni meno correlate sono misurate a 6 o 7 mesi di distanza. I variogrammi spazio-temporali calcolati potrebbero essere utilizzati per specificare un modello con cui effettuare una previsione nello spazio e nel tempo.

In conclusione, le analisi effettuate mostrano che la dipendenza spaziale della concentrazione giornaliera di PM_{10} in Italia varia sia in funzione di fattori ambientali, che in funzione della stagione dell'anno. Utilizzando metodi di previsione spazio-temporale più complessi e introducendo nello studio più informazioni sui fenomeni ambientali correlati alla concentrazione di inquinante, sarebbe possibile ottenere delle mappe di previsione spazio-temporale e utilizzare queste informazioni per determinare adeguate strategie di mitigazione del PM_{10}.

Bibliografia

Chen H. (2017). *Living near major roads and the incidence of dementia, Parkinson's disease, and multiple sclerosis: a population-based cohort study*. The Lancet.

Cressie N. A. e Wikle C. K. (2011). *Statistics for Spatial-Temporal Data*. Probability and Statistics. John Wiley & Sons.

Diggle P. J. e Riberio Jr P. J (2007). *Model-based Geostatistics*. Springer

Diggle P J., Town J. e Moyeed R. A. (1998). *Model-based geostatistics*. Applied Statistics, 47, 3, 299-350.

Fioravanti G., Martino S., Cameletti M., Cattani G. (2020). *Spatio-temporal modelling of PM10 daily concentrations in Italy using the SPDE approach*. Atmospheric Environment 248. Elsevier.

Gneiting T. e Guttorp P. (2010). *Continuous Parameter Spatio-Temporal Processes.* Handbook of Spatial Statistics. Chapman e Hall/CRC

IARC (2015). *Outdoor Air Pollution.* IARC Monographs on the Evaluation of Carcinogenic Risks to Humans Volume 109.

Matheron G. (1963). *Traité de géostatistique appliquée.* Edizioni Technip, Francia.

OMS Ufficio Regionale per l'Europa (2022). *Linee guida globali OMS sulla qualità dell'aria. Particolato ($PM_{2,5}$ e PM_{10}), ozono, biossido di azoto, anidride solforosa e monossido di carbonio. Sintesi*.

Padoan S. A., Bevilacqua M. (2015). *Analysis of Random Fields using CompRandFld*. Journal of Statistical Software Volume 63.

Porcu E, Mateu J, Christakos G (2009). *Quasi-Arithmetic Means of Covariance Functions with Potential Applications to Space-Time Data*. Journal of Multivariate Analysis, 8(100), 1830-1844.

Raaschou-Nielsen O (2013). *Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE)*. Lancet Oncol.

Tobler W. (1970). *A computer movie simulating urban growt in the Detroit region*. Economic Geography, 46: 234-240.

Warsono, K.P.S., Bartolucci, A.A., Bae, S. (2001). *Mathematical modeling of environmental data.* Math. Comput. Model. 33, 793-800.