



UNIVERSITA' DEGLI STUDI DI BERGAMO
DIPARTIMENTO DI INGEGNERIA GESTIONALE E DELL'INFORMAZIONE[°]
QUADERNI DEL DIPARTIMENTO[†]

Department of Management and Information Technology

Working Paper

Series “*Mathematics and Statistics*”

n. 13/MS – 2006

***Dati artificiali e misure in campo nei modelli
dinamici per la previsione spaziale delle polveri
fini nella Regione Piemonte.***

by

Alessandro Fassò and Michela Cameletti

[°] Viale Marconi. 5, I – 24044 Dalmine (BG), ITALY, Tel. +39-035-2052339; Fax. +39-035-562779

[†] Il Dipartimento ottempera agli obblighi previsti dall'art. 1 del D.L.L. 31.8.1945, n. 660 e successive modificazioni.

COMITATO DI REDAZIONE[§]

Series Economics and Management (EM): Stefano Paleari, Andrea Salanti

Series Information Technology (IT): Stefano Paraboschi

Series Mathematics and Statistics (MS): Luca Brandolini, Alessandro Fassò

[§] L'accesso alle *Series* è approvato dal Comitato di Redazione. I *Working Papers* ed i *Technical Reports* della Collana dei Quaderni del Dipartimento di Ingegneria Gestionale e dell'Informazione costituiscono un servizio atto a fornire la tempestiva divulgazione dei risultati dell'attività di ricerca, siano essi in forma provvisoria o definitiva.

Dati artificiali e misure in campo nei modelli dinamici per la previsione spaziale delle polveri fini nella Regione Piemonte.

Risultati preliminari.

Alessandro Fassò

Michela Cameletti

Dipartimento di Ingegneria Gestionale e dell'Informazione
Università degli Studi di Bergamo *

26 ottobre 2006

Sommario

In riferimento ad un tema di grande attualità come l'inquinamento atmosferico e la qualità dell'aria, la statistica si propone come strumento di analisi indispensabile in un approccio modellistico integrato. In questo working paper, in particolare, viene presentato un modello gerarchico spazio-temporale con una procedura di stima in due passi. L'analisi di un caso reale, relativo alla Regione Piemonte nell'anno 2004, permette di testare la robustezza e l'efficienza del modello e, allo stesso tempo, di integrare dati misurati dalla rete di monitoraggio con dati provenienti da una catena di modelli deterministici. I risultati ottenuti sono comunque da considerarsi provvisori poiché la metodologia utilizzata è in fase di completamento e miglioramento.

*Viale Marconi, 5, 24044 Dalmine (BG). E-mail: alessandro.fasso@unibg.it

1 Introduzione

Il fenomeno dell'inquinamento atmosferico ha assunto negli ultimi anni un ruolo centrale nelle discussioni che coinvolgono le istituzioni pubbliche e i cittadini. Questi ultimi sono alla ricerca di informazioni dettagliate e tempestive riguardanti la qualità dell'aria del proprio territorio, anche in relazione alle recenti evidenze scientifiche che mostrano come gli inquinanti dispersi nell'atmosfera abbiano un effetto nocivo sulla salute umana. Le istituzioni, dal canto loro, sono chiamate a rispettare norme e provvedimenti vigenti a livello europeo, nazionale e regionale in termini di monitoraggio, di valutazione della qualità dell'aria e di azioni attuative nel caso di superamento delle soglie di allarme. Purtroppo la complessità del fenomeno dell'inquinamento atmosferico, soggetto come è all'influenza di numerose variabili (meteorologiche, antropiche, chimiche, ecc.), ostacola la possibilità di un'analisi univoca e completa del fenomeno. In letteratura esistono soprattutto approcci puramente deterministici basati su equazioni di diffusione e su modelli di dispersione (Finzi e Brusasca (1991); Finzi *e altri* (2000)). A tale proposito il DM n. 261 del 2002, insieme al DL n. 351 del 1999, fornisce le direttive tecniche per la costruzione di modelli di dispersione da utilizzare come strumenti di valutazione della qualità dell'aria. Nell'allegato 1 del DM, in particolare, viene espressamente affermato che i risultati delle simulazioni modellistiche sono connotati da un certo grado di incertezza che deriva dall'incertezza intrinseca del modello che rappresenta la realtà con un certo grado di approssimazione e di inaccuratezza, e da quella associata ai dati di input. La soluzione proposta per questo problema è tutt'altro che rigorosa: viene suggerito di stimare orientativamente in un fattore due l'incertezza di un valore medio annuo in un generico punto. Eppure la statistica potrebbe essere un utile strumento per risolvere il problema; essa, infatti, può adottare modelli stocastici in grado di cogliere le dinamiche casuali e residuali del fenomeno, considerando così le fonti di incertezza e ciò che non può essere esplicitamente definito. E' pur vero che i modelli statistici risultano essere spesso troppo poveri perché non includono al loro interno i processi chimico-fisici di trasporto, diffusione e trasformazione degli inquinanti. Un approccio integrato del tipo deterministico-statistico potrebbe, quindi, essere la soluzione ideale per determinare, con la dovuta incertezza e completezza, la distribuzione spaziale e temporale delle concentrazioni degli inquinanti atmosferici. Ciò permetterebbe di raggiungere gli obiettivi del DM (previsione anche in zone non attrezzate con centraline, determinazione dei contributi relativi delle varie sorgenti emissive, valutazione dell'impatto di inquinanti non misurati dalla rete di monitoraggio, ecc.) in maniera rigorosa e con una evidenza scientifica, così da poter informare correttamente i cittadini e orientare le

amministrazioni pubbliche verso gli opportuni interventi di risanamento. Questo lavoro si pone in quest'ottica di approccio integrato allo studio del fenomeno, utilizzando i dati provenienti da più modelli deterministici in un modello statistico studiato per la mappatura spaziale della concentrazione di polveri sottili (*PM10*). In particolare, il lavoro è così strutturato: nel paragrafo 2 viene introdotto un modello statistico gerarchico di tipo state-space caratterizzato da una variabile latente che interpreta il vero livello di inquinamento atmosferico da polveri sottili non direttamente osservabile. Successivamente, nel paragrafo 3 viene presentata la metodologia utilizzata per la stima dei parametri ignoti del modello; la procedura, in particolare, è suddivisa in due passi e utilizza, rispettivamente, il metodo dei minimi quadrati e l'algoritmo EM. Nella seconda parte del lavoro, viene analizzato un caso reale relativo alla Regione Piemonte nell'anno 2004. In particolare, dopo aver definito la struttura della rete di monitoraggio e della catena modellistica della Regione, vengono presentate le stime dei parametri e una mappa di previsione spaziale. Infine, il paragrafo finale delle conclusioni viene dedicato agli sviluppi in corso conseguenti a questo lavoro, con l'auspicio che una collaborazione tra scienziati provenienti da diversi settori sia presto realizzabile.

2 Un modello generale gerarchico di tipo state-space

Si consideri una rete di monitoraggio composta da n centraline localizzate nei punti $(s_1, \dots, s_n)$ appartenenti alla regione di interesse D . Sia, inoltre, Z un processo spazio-temporale n -dimensionale che interpreta la concentrazione di polveri sottili e sia Y un processo temporale latente univariato che può essere definito come il vero livello di inquinamento atmosferico comune a tutta la regione D . Indicando con $z_t = (z(s_1, t), \dots, z(s_n, t))$ il vettore delle realizzazioni del processo Z all'istante t nelle n centraline ($t = 1, 2, \dots, T$), è possibile introdurre una prima equazione di misura che esprime il processo osservato incorporando l'errore di misura ε :

$$\begin{matrix} z_t & = & u_t & + & \varepsilon_t \\ (n \times 1) & & (n \times 1) & & (n \times 1) \end{matrix} \quad (1)$$

Il termine u_t , che rappresenta la concentrazione di polveri al netto dell'errore di misura, viene definito nella seconda equazione di misura in funzione di p variabili esogene, opportunamente ponderate con il vettore dei coefficienti β

costante nel tempo, del processo latente y_t e dell'errore di modello ω :

$$\begin{matrix} u_t & = & X_t\beta & + & k_ty_t & + & \omega_t \\ (n \times 1) & & (n \times p)(p \times 1) & & (n \times 1)(1 \times 1) & & (n \times 1) \end{matrix} \quad (2)$$

Infine, nell'equazione di stato viene definita la dinamica temporale del processo Y utilizzando un modello autoregressivo a ritardo unitario:

$$\begin{matrix} y_t & = & hy_{t-1} & + & \eta_t \\ (1 \times 1) & & (1 \times 1)(1 \times 1) & & (1 \times 1) \end{matrix} \quad (3)$$

Per completezza si assume che il processo latente inizi al tempo $t = 0$ e che sia definito da una distribuzione gaussiana univariata del tipo $y_0 \sim N_1(\mu_0, \sigma_0^2)$. Le distribuzioni di probabilità degli errori utilizzati nelle equazioni (1), (2) e (3), da considerarsi indipendenti l'uno dall'altro, sono di seguito specificate:

$$\varepsilon_t \sim N_n(0, \Sigma_\varepsilon) \quad (4)$$

$$\omega_t \sim N_n(0, \Sigma_\omega) \quad (5)$$

$$\eta_t \sim N_1(0, \Sigma_\eta) \quad (6)$$

In particolare per gli errori (4) e (6) viene utilizzata una struttura di tipo white noise e, quindi, le matrici di varianze e covarianze sono date, rispettivamente, da $\Sigma_\varepsilon = \sigma_\varepsilon^2 I_n$ e $\Sigma_\eta = \sigma_\eta^2$, dove I_n rappresenta la matrice identità n -dimensionale. L'errore (5), invece, è definito da un processo puramente spaziale. Utilizzando, in particolare, una funzione di covarianza spaziale C_θ di tipo esponenziale (Cressie, 1993, pag 61), il generico elemento della matrice di varianze e covarianze Σ_ω è dato da $\Sigma_\omega(ij) = \sigma_\omega^2 [\exp(-\theta^2 d_{ij})] = \sigma_\omega^2 C_\theta(ij)$ dove d_{ij} è la distanza euclidea tra due generiche centraline i e j .

Sostituendo l'equazione (2) nella (1) si ottiene un'unica equazione di misura n -dimensionale:

$$z_t = X_t\beta + k_ty_t + e_t \quad (7)$$

dove e_t è il processo dato dalla somma dei processi ε_t e ω_t . Esso, in particolare, ha una distribuzione gaussiana n -dimensionale con vettore delle medie nullo e matrice di varianze e covarianze Σ_e , il cui generico elemento è così definito:

$$\Sigma_e(ij) = \begin{cases} \sigma_\omega^2 + \sigma_\varepsilon^2 & i = j \\ \sigma_\omega^2 [\exp(-\theta^2 d_{ij})] & i \neq j \end{cases} \quad (8)$$

Si noti che la varianza dell'errore di misura ε contribuisce alla varianza del processo e solamente in termini di microscala, ovvero per distanze nulle tra le centraline; ne consegue, quindi, che la funzione di covarianza spaziale che genera la matrice Σ_e non è continua nell'origine (effetto nugget, Cressie (1993,

pag 59)).

Così come ampiamente descritto in Fassò *e altri* (2006), la rete di monitoraggio può essere eterogenea; può, quindi, accadere che i processi Z e Y siano definiti su una scala spaziale e/o temporale differente. In termini spaziali, ad esempio, il processo latente Y è n -dimensionale analogamente alla rete di monitoraggio mentre, al contrario, il processo Z in un generico istante t può avere dimensione inferiore, $n_t \leq n$, per la presenza di valori mancanti. In questo caso, quindi, il vettore k_t della (7) è un vettore indicatore che mappa il processo Z sul processo Y . Considerando, invece, la risoluzione temporale può accadere che le centraline abbiano frequenze di campionamento diverse (ad esempio, oraria, bioraria, giornaliera). Ipotizzando che il processo latente Y sia definito con la più alta risoluzione temporale (frequenza oraria), k_t è il vettore di pesi costruito in modo da definire il processo Z come media giornaliera del processo Y . Ovviamente, in assenza di valori mancanti e con una frequenza di campionamento comune a tutta la rete di monitoraggio, il vettore k_t è un vettore unitario di dimensione $(n \times 1)$.

3 La stima dei parametri

Il modello gerarchico di tipo state-space definito dalle equazioni (7) e (3) del paragrafo 2 dipende da un insieme di parametri ignoti Ψ così definito:

$$\Psi = \{\beta, \sigma_\varepsilon^2, \sigma_\omega^2, \theta, h, \sigma_\eta^2\} \quad (9)$$

La procedura di stima di Ψ viene svolta in due passi: in un primo tempo, viene stimato il vettore dei coefficienti $\hat{\beta}$ attraverso un modello di regressione lineare multiplo. Successivamente, i residui del modello $r_t = z_t - X_t \hat{\beta}$ vengono utilizzati come variabile proxy di $k_t y_t$ e i restanti parametri vengono stimati con l'algoritmo **EM** (McLachlan e Krishnan (1997); Schafer (1997)). L'algoritmo **EM** è una procedura iterativa per la stima di massima verosimiglianza particolarmente utilizzata nelle situazioni soggette alla presenza di valori mancanti: si intuisce facilmente come un modello gerarchico come quello presentato nel paragrafo 2 e caratterizzato dalla presenza di una variabile latente Y , possa essere ricondotto ad una situazione di dati mancanti e, quindi, ben si presti all'utilizzo dell'algoritmo **EM**.

La procedura di stima, opportunamente adattata per un modello di tipo state-space, analogamente a quanto fatto nei lavori di Shumway e Stoffer (1982) e Xu e Wikle (2006), alterna il passo **E** (Expectation) al passo **M** (Maximization). In particolare, dato il vettore dei parametri $\Psi^{(j-1)}$, al passo **E** dell'iterazione j l'algoritmo calcola la funzione Q , ovvero il valore atteso della funzione di logverosimiglianza completa $\log L_c(\Psi)$, cambiata di segno

e moltiplicata per due, condizionatamente alle osservazioni disponibili fino all'istante T :

$$\begin{aligned}
Q(\Psi; \Psi^{(j-1)}) &= E[-2 \log L_c(\Psi) \mid z_1, \dots, z_T] = \\
&T \log |\Sigma_\varepsilon| + tr \left[\Sigma_\varepsilon^{-1} \sum_{t=1}^T \left(k_t P_t^T k_t' + (z_t - k_t y_t^T) (z_t - k_t y_t^T)' \right) \right] + \\
&\log(\sigma_0^2) + \frac{1}{\sigma_0^2} \left(P_0^T + (y_0^T - \mu_0)^2 \right) + \\
&T \log(\sigma_\eta^2) + \frac{1}{\sigma_\eta^2} \sum_{t=1}^T \left(P_t^T + h^2 P_{t-1}^T - 2h P_{t,t-1}^T + (y_t^T - h y_{t-1}^T)^2 \right)
\end{aligned} \tag{10}$$

I termini $y_t^T = E[y_t \mid z_1, \dots, z_T]$ e $P_t^T = Var[y_t \mid z_1, \dots, z_T]$ rappresentano, rispettivamente, il valore atteso e la varianza della variabile di stato Y condizionatamente a tutte le osservazioni disponibili fino al tempo T . Essi sono ottenuti come output del *Kalman Smoother* (Durbin e Koopman, 2001, pag 70) implementato utilizzando come valori iniziali dei parametri il vettore $\Psi^{(j-1)}$. Inoltre, $P_{t,t-1}^T$ rappresenta la covarianza a ritardo unitario tra le variabili di stato, condizionatamente al vettore delle osservazioni $(z_1, \dots, z_T)$; essa è così definita:

$$P_{t,t-1}^T = E \left\{ (y_t - y_t^T) (y_{t-1} - y_{t-1}^T)' \mid z_1, \dots, z_T \right\}$$

Al passo **M** dell'algoritmo viene definita una stima aggiornata del vettore dei parametri $\Psi^{(j)}$ in modo tale che la funzione Q , definita dalla (10) sia decrescente ovvero soddisfi la seguente disuguaglianza:

$$Q(\Psi^{(j)}; \Psi^{(j-1)}) \leq Q(\Psi^{(j-1)}; \Psi^{(j-1)}) \tag{11}$$

Questa condizione è sufficiente per garantire che la sequenza delle funzioni di logverosimiglianza $\log L(\Psi^{(j)})$ sia monotona non decrescente. Se la stima $\Psi^{(j)}$ è anche il minimo della (10) si ha il classico algoritmo **EM**; se, invece, $\Psi^{(j)}$ soddisfa solamente la (11) l'algoritmo è noto come Generalized EM (**GEM**) (McLachlan e Krishnan, 1997, pag 84).

In relazione al modello considerato, dalla minimizzazione della (10), sotto l'ipotesi che le varianze degli errori ε e ω siano uguali, ovvero $\sigma_\varepsilon^2 = \sigma_\omega^2 = \sigma^2$, si ottiene che la stima aggiornata di $\Psi^{(j)} = \{\sigma^{2(j)}, \theta^{(j)}, h^{(j)}, \sigma_\eta^{2(j)}\}$ è data dalle

seguenti espressioni ($j = 1, 2, \dots$):

$$\sigma^{2(j)} = \frac{\text{tr}(d)}{Tn} \quad (12)$$

$$\theta^{(j)} = \theta^{(j-1)} - a \frac{f'(\theta^{(j-1)})}{f''(\theta^{(j-1)})} \quad (13)$$

$$h^{(j)} = \frac{\sum_{t=1}^T (P_{t,t-1}^T + y_t^T y_{t-1}^T)}{\sum_{t=1}^T P_{t-1}^T + (y_{t-1}^T)^2} \quad (14)$$

$$\sigma_\eta^{2(j)} = \frac{\sum_{t=1}^T [P_t^T + h^{2(j)} P_{t-1}^T - 2h^{(j)} P_{t,t-1}^T + (y_t^T - h^{(j)} y_{t-1}^T)^2]}{T} \quad (15)$$

dove

$$d = (I_n + C_\theta)^{-1} \sum_{t=1}^T B_t \quad (16)$$

$$B_t = k P_t^T k' + (z_t - k y_t^T) (z_t - k y_t^T)' \quad (17)$$

Il termine C_θ utilizzato nella (16) è la matrice generata dal processo spaziale di tipo esponenziale, il cui elemento generico è dato dalla (8) ponendo $\sigma_\varepsilon^2 = \sigma_\omega^2 = \sigma^2$.

Si noti che la (13) utilizza un algoritmo di ottimizzazione numerica di tipo *Newton-Raphson* (McLachlan e Krishnan, 1997, pag 28) dove a è una costante compresa nell'intervallo $[0, 1]$ che assicura che la funzione (10) soddisfi la condizione (11) di monotonicità decrescente. Inoltre, con la notazione $f'(\theta^{(j-1)})$ e $f''(\theta^{(j-1)})$ viene indicata, rispettivamente, la derivata prima e seconda della funzione (10) rispetto al parametro θ , entrambe valutate per θ pari a $\theta^{(j-1)}$.

4 Il caso della Regione Piemonte

In questo paragrafo viene presentato lo studio dell'applicazione del modello gerarchico e della procedura di stima presentati nei paragrafi 2 e 3 per la Regione Piemonte per l'anno 2004. La scelta della Regione Piemonte è dettata dal fatto che essa è da anni attiva nella promozione e nella attuazione di iniziative e di politiche per la riduzione delle emissioni inquinanti e per il miglioramento progressivo della qualità dell'aria per la salvaguardia della salute dell'uomo e dell'ambiente.

4.1 La struttura della rete di monitoraggio

Le informazioni sulla qualità dell'aria della Regione Piemonte derivano dalle misure rilevate dal Sistema regionale di rilevamento della qualità dell'aria (S.R.Q.A.) gestito dall'ARPA Piemonte¹. In particolare, la rete di monitoraggio per il PM_{10} nell'anno considerato era costituita da 22 centraline dotate di strumenti gravimetrici e posizionate sul territorio regionale così come mostrato nella figura 1. In riferimento alla notazione introdotta nel paragrafo 2, si ha che il vettore z_t è costituito dalle concentrazioni di PM_{10} misurate nelle 22 centraline nel generico giorno t ($t = 1, \dots, 366$).

Si evidenzia che ogni centralina ha una frequenza di campionamento gior-

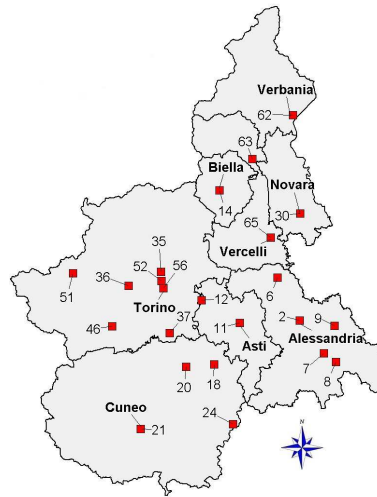


Figura 1: La rete di monitoraggio per il PM_{10} della Regione Piemonte.

naliera e che la percentuale annua di valori mancanti varia da un minimo dello 0.3% (Verbania) ad un massimo del 9.8% (Pinerolo e Buttigliera Alta). Come si evince dalla mappa della figura 2, quasi tutte le centraline della rete nell'anno 2004 hanno superato in media il valore limite annuale per la protezione della salute umana, fissato pari a $40 \mu g/cm^3$, ad eccezione di alcune stazioni non situate in prossimità dei grandi centri urbani. Emerge, inoltre, la già nota stagionalità del fenomeno che, per semplicità, in questo lavoro viene dicotomizzata in un fattore del tipo Estate/Inverno. Il boxplot della figura 2, infatti, evidenzia un livello medio e una variabilità inferiori per la stagione estiva.

¹<http://www.sistemapiemonte.it/ambiente/srqa/>

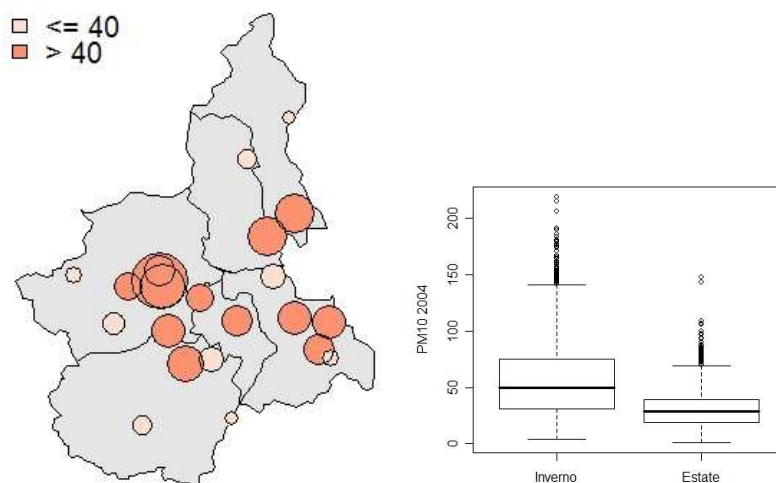


Figura 2: *A sinistra*: concentrazioni medie di PM_{10} in riferimento alla soglia limite (anno 2004); *A destra*: boxplot della concentrazione di PM_{10} con un fattore stagionale (anno 2004).

4.2 La catena modellistica

L'Area Previsione e Monitoraggio Ambientale dell'ARPA Piemonte² implementa e gestisce una complessa catena modellistica per la previsione della dispersione di alcuni inquinanti su scala temporale e spaziale ad alta risoluzione (Muraro e De Maria, 2005). Gli output dei singoli modelli che costituiscono la catena sono orari e definiti su una griglia di 4 km per 4 km che copre una superficie di $220\text{ per }284\text{ km}^2$ che, come mostrato nella figura 3, va oltre i confini della regione Piemonte. La catena modellistica è composta da tre moduli (meteorologico, emissivo e chimico) ognuno dei quali, a sua volta, utilizza specifici modelli deterministici. Ai fini del lavoro, sono state selezionate alcune variabili di interesse, ritenute importanti per l'analisi del fenomeno e ottenute come output intermedio della catena modellistica. Le variabili selezionate, contenute nella tabella 1, verranno successivamente utilizzate come variabili esogene (per mezzo della matrice X_t) nel modello di regressione presentato nel paragrafo 4.3.

Per quanto precedentemente detto, la risoluzione spaziale e temporale delle variabili provenienti dalla catena modellistica è differente rispetto a quella delle variabili misurate per mezzo della rete di monitoraggio; per questo motivo è necessario riallinearle su una scala spazio-temporale comune. In

²<http://extranet.regione.piemonte.it/ambiente/aria/>

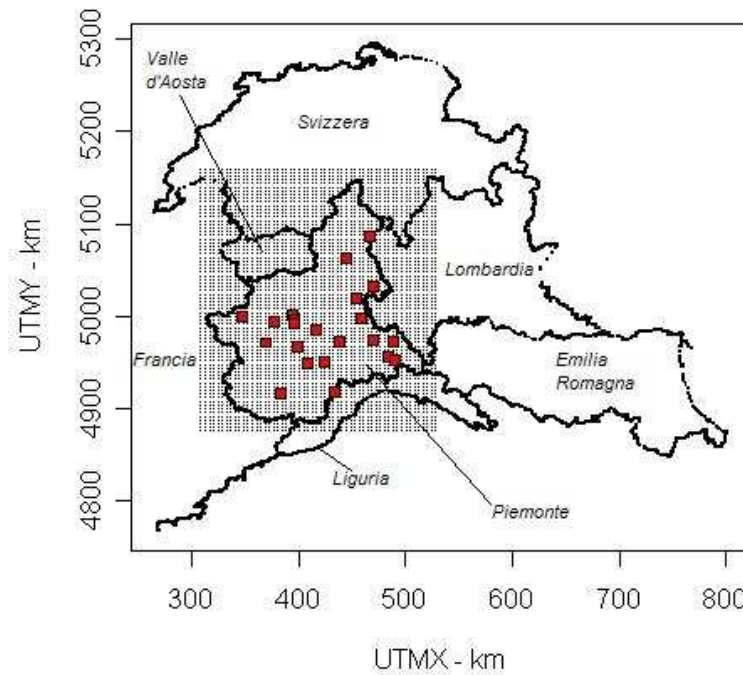


Figura 3: La risoluzione spaziale per la catena modellistica.

prima istanza abbiamo, quindi, trasformato i dati orari in dati giornalieri utilizzando un'opportuna funzione di sintesi, così come indicato nell'ultima colonna della tabella 1. Ad esempio, in riferimento alle variabili relative alle emissioni primarie di particolato atmosferico, è stata calcolata la somma giornaliera. Successivamente, utilizzando il metodo del centroide, si è passati dalla risoluzione spaziale della griglia (4032 punti) alla quella della rete (22 centraline) per ogni variabile esogena selezionata: per ogni centralina della rete, in particolare, è stato associato il valore corrispondente al punto centrale della maglia della griglia alla quale la centralina appartiene.

Successivamente le variabili qualitative prescelte sono state rese dicotomiche: per la variabile "Classi di stabilità" sono state scelte due modalità, 4 e $\neq 4$, che corrispondono alla presenza o all'assenza di una situazione neutra per l'atmosfera in termini di stabilità e convezione. Per la variabile "Uso del suolo", invece, viene utilizzata una suddivisione del tipo Urbano/Non urbano (modalità pari a 1 o $\neq 1$). Infine, per introdurre una struttura di vicinato sulla rete, sono state definite 4 nuove variabili (con il suffisso Nord, Sud, Est, Ovest) a partire dalle emissioni primarie di particolato. Così, ad esempio, la variabile denominata "Emisize01 Ovest" definisce le emissioni primarie di particolato di diametro fine nelle maglie a ovest dei punti della rete.

Etichetta	Variabile	Unità di misura	Funzione
Meteo			
<i>HMX</i>	Altezza strato di rimescolamento	m	Media
<i>PGT</i>	Classi di stabilità di Pasquill, Gifford, Turner	Qualitativa da 1 a 6	Moda
Emissioni primarie di particolato			
<i>EMISIZE01</i>	Diametro aerodinamico inferiore a 2.5 micron	g/s	Somma
<i>EMISIZE02</i>	Diametro aerodinamico compreso tra 2.5 e 10 micron	g/s	Somma
Variabili geografiche			
<i>QUOTA</i>	Orografia (DTM 250m)	m	–
<i>LANDUSE</i>	Uso del suolo (Corine 250m)	Qualitativa da 1 a 10	–

Tabella 1: Variabili selezionate.

4.3 Il modello di regressione lineare multipla

La prima parte della procedura di stima dei parametri del modello gerarchico si è avvalsa di un modello di regressione lineare multipla, dove le concentrazioni di *PM10*, sottoposte a trasformazione di tipo radice quadrata, sono state espresse come funzione lineare delle variabili esogene prescelte. Il modello finale utilizzato, frutto di un'accurata selezione basata su principi statistici e di coerenza rispetto al fenomeno considerato, è definito nell'equazione (18), dove per ogni variabile vengono utilizzate le etichette introdotte nella tabella 1.

$$z_t = \alpha_0 + \alpha_{estate} + \sum_{i=1}^2 \left[\begin{array}{l} \beta_{1i}Emisize01 \text{ Est} + \beta_{2i}Emisize01 \text{ Ovest} \\ + \beta_{3i}Emisize02 + \beta_{4i}Emisize02 \text{ Ovest} + \beta_{5i}Hmix \\ + \beta_{6i}Landuse \text{ Urbano} + \beta_{7i}PGT \text{ D} + \beta_{8i}Quota \end{array} \right] \quad (18)$$

Si noti che il modello è di tipo grey-box con una componente stagionale: esso è costituito da un'intercetta comune α_0 e da un'intercetta specifica per il periodo estivo, α_{estate} . Le variabili esogene, inoltre, sono ponderate con dei coefficienti che sono diversi a seconda che si consideri la stagione estiva ($i = 1$) o la stagione invernale ($i = 2$). Ciò permette al modello di utilizzare i dati dell'intero anno, catturandone così la dinamica temporale, ma di fornire stime differenziate in funzione della stagione.

Nella tabella 2 sono contenute le stime dei minimi quadrati dei coefficienti del modello con le corrispondenti deviazioni standard. L'indice di determinazione multipla del modello, ovvero la quota di varianza dei dati spiegata, è pari a 0.36; questo valore è indice di una scarsa performance del modello ma non desta alcuna preoccupazione in quanto la procedura di stima dei parametri non è conclusa e la struttura di correlazione spaziale dei dati deve essere ancora considerata (si veda il successivo paragrafo). L'analisi dei residui del

	Stima		Std error	
α_0	8.42		0.08	
α_{estate}	-2.86		0.14	
	Estate		Inverno	
	Stima	Std error	Stima	Std error
Emisize01 Est	0.015	0.002	0.013	0.001
Emisize01 Ovest	-0.021	0.004	-0.003	0.002
Emisize02	0.08	0.023	0.181	0.012
Emisize02 Ovest	0.231	0.043	0.045	0.023
Hmix	-0.0004	0.0002	-0.004	0.0003
Landuse Urbano	-0.846	0.0108	-1.2	0.117
PGT4	-0.192	-0.068	-0.819	0.005
Quota	-0.001	0.0002	-0.003	0.0002

Tabella 2: Stime dei minimi quadrati dei coefficienti del modello di regressione lineare multipla.

modello di regressione lineare multiplo conferma questa evidenza: come si può osservare nella figura 4, i residui hanno una distribuzione normale centrata sullo zero anche se con delle code un pò pesanti. L'analisi della funzione di autocorrelazione dei residui, svolta singolarmente per ogni centralina della rete, mostra che dopo il primo lag l'autocorrelazione diventa trascurabile, come si può, ad esempio, vedere nella figura 5, relativa alla centralina di Vercelli. Ciò supporta l'utilizzo di un modello a ritardo unitario nell'equazione di stato (3) per la dinamica temporale del processo latente. Inoltre, l'analisi delle componenti principali per la matrice di varianze e covarianze dei residui, mostra che il primo autovalore spiega il 60% della variabilità residua. Questo risultato avvalorava l'ipotesi di utilizzare una variabile di stato unidimensionale per interpretare il vero livello di inquinamento comune a tutta la regione.

4.4 L'algoritmo EM

Dopo la stima del vettore dei coefficienti $\hat{\beta}$ (tabella 2), i residui $r_t = z_t - X_t \hat{\beta}$ sono stati utilizzati come proxy della variabile di stato $k_t y_t$ ($t = 1, 2, \dots, 366$), dove k_t è un vettore unitario costante nel tempo. La procedura di stima dei restanti parametri è stata realizzata utilizzando l'algoritmo **EM**, utilizzando le espressioni (12), (13), (14), (15). E' da evidenziare che l'algoritmo è stato testato attraverso un'intensa analisi di sensitività per verificare la sua robustezza rispetto alla scelta dei valori iniziali e ai criteri di arresto.

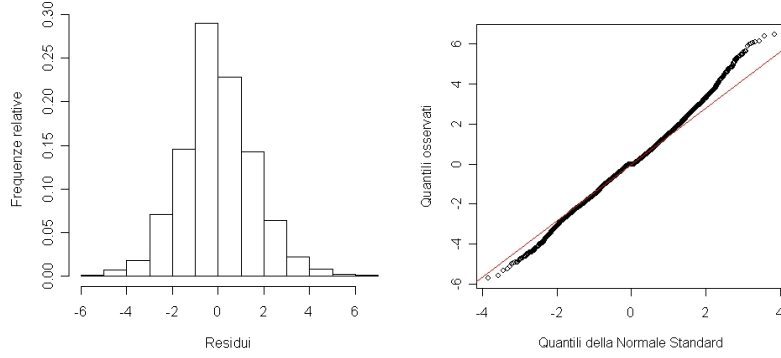


Figura 4: Istogramma e qqplot dei residui del modello di regressione lineare multipla.

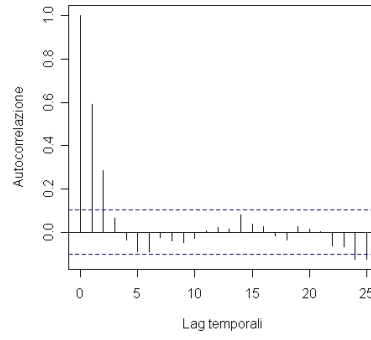


Figura 5: Funzione di autocorrelazione: centralina di Vercelli.

Le stime finali dei parametri sono, quindi, risultate le seguenti:

$$\hat{\Psi} = \left\{ \hat{\sigma}^2 = 0.66, \hat{\theta} = 0.13, \hat{h} = 0.75, \hat{\sigma}_{\eta}^2 = 0.51 \right\} \quad (19)$$

La stima ottenuta per il parametro θ esprime una discreta covarianza spaziale, che non era stata catturata dal modello di regressione lineare, che corrisponde a funzione di covarianza esponenziale simile a quella rappresentata nella figura 6. Dalla figura si può notare che la covarianza spaziale tende ad annullarsi poco prima dei 50 *km* di distanza: ciò porterebbe a concludere, in prima approssimazione, che l'inquinamento atmosferico, al netto dell'influenza delle variabili considerate, sia caratterizzato da una persistenza spaziale abbastanza marcata. Il parametro di persistenza temporale, stimato pari a 0.7, evidenzia una discreta persistenza temporale tra un istante temporale e

il precedente e ciò è in accordo con quanto presente in letteratura.

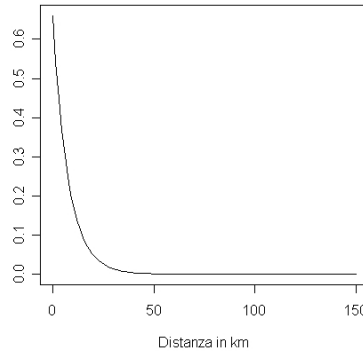


Figura 6: Funzione di covarianza spaziale esponenziale con $\hat{\theta} = 0.13$.

4.5 Le mappe di concentrazione

L'utilizzo di un modello statistico permette di stimare la concentrazione di polveri sottili su tutta la regione di interesse, anche nei punti ove non è presente una centralina per il monitoraggio. Nell'ambito di questo lavoro abbiamo, in prima approssimazione, costruito mappe giornaliere (si veda, ad esempio, la figura 7) utilizzando i parametri stimati con il modello di regressione multipla presentato nel paragrafo 4.3. Dalla mappa, la cui risoluzione spaziale corrisponde a quella della griglia della catena modellistica, emergono chiaramente i tre capoluoghi di regione, Torino, Milano e Genova, per la più alta concentrazione di polveri.

5 Lavori in corso

La metodologia utilizzata in questo lavoro si presenta come innovativa dal punto di vista dell'approccio integrato per l'analisi del fenomeno della concentrazione di polveri sottili. Il modello e la procedura di stima utilizzati sono risultati robusti ed efficienti anche se parzialmente incompleti. Innanzitutto è necessario implementare una procedura che stimi simultaneamente tutti i parametri all'interno del modello gerarchico con l'algoritmo EM, senza dover utilizzare la regressione multipla. Secondariamente, allo stato attuale l'algoritmo **EM** non restituisce le deviazioni standard delle stime dei parametri. A tale proposito, è in fase di progettazione l'implementazione di un algoritmo di ricampionamento parallelizzato su un cluster di computer, così

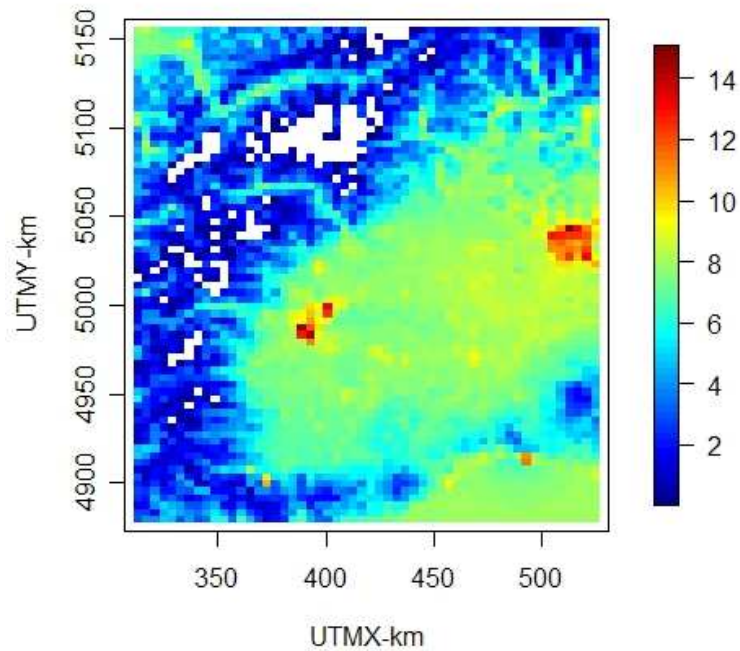


Figura 7: Mappa di concentrazione per il 30 gennaio 2004.

da poter ottenere una stima di tipo bootstrap delle deviazioni standard. Una volta conseguiti questi obiettivi, sarà possibile costruire mappe di concentrazione basate sulle stime finali dell'algoritmo **EM**, insieme alle corrispondenti mappe per la valutazione dell'incertezza delle previsioni.

Infine, allo scopo di migliorare la flessibilità del modello si potranno utilizzare nuove funzioni di covarianza spaziale, con un numero maggiore di parametri e che considerino la natura anisotropica del fenomeno.

In conclusione, i risultati ottenuti in questo lavoro, anche se parziali rispetto alle potenzialità della metodologia proposta, sembrano essere un buon inizio per lo sviluppo di un'analisi integrata del fenomeno dell'inquinamento atmosferico e della valutazione della qualità dell'aria.

Riferimenti bibliografici

- Cressie N. (1993). *Statistics for Spatial Data*. John Wiley and Sons.
- Durbin J.; Koopman S. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, New York.
- Fassò A.; Cameletti M.; Nicolis O. (2006). Air quality monitoring using heterogeneous networks. *Environmetrics*. in stampa.
- Finzi G.; Brusasca G. (1991). *La qualità dell'aria. Modelli previsionali e gestionali*. Masson – Milano.
- Finzi G.; Pirovano G.; Volta M. (2000). *Gestione della qualità dell'aria. Modelli di simulazione e previsione*. McGraw Hill – Milano.
- McLachlan G. J.; Krishnan T. (1997). *The EM Algorithm and Extensions*. John Wiley and Sons, New York.
- Muraro M.; De Maria R. (2005). Modelling applications and developments for air quality assessment in piemonte. In *Air quality assessment and management in the Piemonte Region according to European Legislation*. Torino, 28 ottobre 2005.
- Schafer J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.
- Shumway R.; Stoffer D. (1982). An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series Analysis*, **3**, 253–264.
- Xu K.; Wikle C. K. (2006). Estimation of parameterized spatio-temporal dynamic models. *Journal of Statistical Inference and Planning*. In stampa.