ANALISI DELLA CONCENTAZIONE DI SO_2 COMBINANDO I DATI RACCOLTI DA CENTRALINE FISSE E MOBILI: UN MODELLO STATE SPACE

B. Scarpa

2001.8

Dipartimento di Scienze Statistiche
Università degli Studi
Via C. Battisti 241-243
35121 Padova

Aprile 2001

Analisi della concentrazione di SO_2 combinando i dati raccolti da centraline fisse e mobili: un modello *state space*

Bruno Scarpa
Dipartimento di Scienze Statistiche
Università di Padova
Via San Francesco, 33
Padova
Italia
email: scarpa@hal.stat.unipd.it

1 Introduzione

Un'importante applicazione di modellazione e previsione statistica spazio-temporale riguarda la concentrazione di sostanze chimiche presenti nell'atmosfera che costituisce uno degli indicatori principali per la segnalazione di situazioni di inquinamento. Spesso vengono richieste allo statistico, da parte di ricercatori ambientali e di amministratori della cosa pubblica, stime e previsioni della concentrazione degli inquinanti nei diversi luoghi all'interno delle città, ma, tipicamente, le rilevazioni di concentrazioni di variabili ambientali avvengono in un numero fissato e ridotto di punti nello spazio, in quanto le centraline utilizzate per tale rilevazione sono molto costose e l'amministrazione pubblica in generale non ne mantiene più di quattro o cinque in ogni città. L'utilizzo di centraline mobili, che si installano per periodi di tempo ridotti in diversi punti nella regione sotto studio, è un modo per ottenere un maggior numero di informazioni spaziali. L'uso di un'unica centralina per rilevare differenti punti nello spazio non consente, però, di avere informazioni temporalmente complete per ciascun luogo: non è possibile, quindi, l'utilizzo degli usuali modelli presenti in letteratura (cfr. per esempio Cressie, 1993; Haas, 1995; Haas, 1996), ma costringe ad aggiungere ipotesi e a ricercare soluzioni specifiche nella formulazione di un nuovo modello.

Con l'obiettivo di rispondere ad esigenze di questo tipo, questo articolo studia il processo spazio-temporale della concentrazione di biossido di zolfo nella città di Padova. Gli obiettivi dell'articolo sono, quindi: (a) stimare un modello per la concentrazione di biossido di zolfo che colga gli effetti principali del tempo e dello spazio sul valore medio e sulla struttura di covarianza utilizzando osservazioni non complete, caratterizzate cioè da grandi quantità di "dati mancanti" non eliminabili e (b) sviluppare un algoritmo capace di stimare i parametri del processo in esame in maniera sufficientemente flessibile così da poter descrivere la variabilità di grande scala attraverso processi stocastici.

Possibili applicazioni del modello sono poi (i) fornire delle previsioni sia spaziali che temporali (per giorni fissati o per periodi di tempo più lunghi) per il processo nella scala originale e delle misure di variabilità di tali previsioni che non siano particolarmente distorte; (ii) individuare i siti nello spazio dove installare nel futuro le centraline mobili.

I dati utilizzati per le analisi sono stati forniti dall'Amministrazione Provinciale di Padova che, per legge, è tenuta a mantenere una rete di monitoraggio ambientale nella Provincia. I dati si riferiscono alle medie giornaliere della concentrazione di SO_2 rilevata ad intervalli di tempo molto

brevi (alcuni secondi) da cinque centraline per tre anni atmosferici consecutivi (1991-1994). Due di queste centraline erano fisse per tutti i tre anni, una terza è stata installata solo a metà del secondo anno. Le ultime due centraline erano installate su stazioni mobili e durante i tre anni si sono spostate nei diversi quartieri della città, per complessivi 14 siti, rimanendo installate per brevi periodi (alcune settimane) in ogni sito visitato.

L'insieme dei dati può essere quindi visto come un insieme di variabili, una per ogni sito osservato almeno per un giorno, legate da vincoli di carattere spaziale e temporale con una grande quantità di *dati mancanti*, alcuni reali (quelli relativi a siti ed a giorni in cui era collocata una centralina – fissa o mobile – che per qualche ragione non era funzionante) ed alcuni fittizi (quelli relativi ai siti visitati da centraline mobili nei giorni in cui esse erano assenti).

Le osservazioni, effettuate in ogni sito e ogni istante di tempo, allora, possono essere considerate come realizzazioni di un processo stocastico $\{z(s,t)\}$, a due indici, uno, t, legato al tempo e l'altro, s (bivariato), allo spazio. Si suddivide la variabilità di tale processo in una parte che colga principalmente l'andamento dovuto alle differenze globali, spaziali e temporali (la variabilità di grande scala), e in una parte che invece colga sostanzialmente il "rumore" intorno all'altra componente (la variabilità di piccola scala).

L'utilizzo di centraline mobili comporta, ovviamente, la presenza di poche osservazioni contemporanee. Non è quindi possibile l'introduzione di eventuali elementi di interazione tra spazio e tempo nella formulazione di un modello per la componente di grande scala.

In questo articolo, le varie componenti del modello di grande scala e di piccola scala vengono considerate come processi stocastici. Sarà opportuno, quindi, aggiungere alcune ipotesi per poter identificare tutti i processi coinvolti. Per esempio si potrà supporre che i processi che colgono la componente di grande scala, si comportino come processi stocastici non stazionari, lasciando alla componente di piccola scala il processo residuo che, quindi, avrà un comportamento più regolare, per esempio stazionario e isotropico in spazio e tempo. La stima contemporanea di tali processi può avvenire attraverso una formulazione *state space* del modello indicato. È possibile così ottenere delle stime di massima verosimiglianza per tutti i parametri coinvolti attraverso l'applicazione del filtro di Kalman.

L'articolo è suddiviso in tre parti: nella sezione 2 si presenta la specificazione del modello segnalando la struttura di alcuni possibili modelli di processi spaziali non stazionari utilizzabili introducendo la specificazione del modello *state space* per la stima contemporanea dei processi coinvolti; nella sezione 3 viene considerata dettagliatamente l'applicazione ai dati di Padova e infine alcune considerazioni conclusive presentano le linee guida di eventuali sviluppi e ampliamenti che le varie tematiche affrontate richiederebbero.

2 Il Modello e le procedure di stima

2.1 Il Modello

Le osservazioni effettuate in ogni sito, ogni istante di tempo, vengono viste come realizzazioni di un processo stocastico $\{\tilde{z}(s,t)\}_{s\in\mathbb{R}^2,t\in\mathbb{R}}$ che al variare di due indici, uno legato al tempo e l'altro allo spazio (quest'ultimo a sua volta sarà un vettore con le coordinate cartesiane dello spazio) descrive tutte le possibili traiettorie spaziali e temporali della concentrazione di un fissato inquinante.

A partire dalle osservazioni disponibili si vuole fare inferenza su tale processo stocastico, per poter descrivere efficacemente la qualità dell'aria nella città in esame.

In realtà nel seguito si prende in considerazione il processo $\{z(s,t)\}_{s\in\mathbb{R}^2,t\in\mathbb{N}}$ che è un'approssimazione del processo $\{\tilde{z}(s,t)\}_{s\in\mathbb{R}^2,t\in\mathbb{R}}$ dove l'indice temporale è discretizzato e varia nell'insieme dei numeri naturali. Questa scelta è dovuta principalmente al fatto che di solito la dinamica temporale viene rilevata attraverso campioni effettuati per istanti temporali discreti come per esempio le ore del giorno, i giorni della settimana o i mesi dell'anno; inoltre, nelle analisi ambientali, per neutralizzare grosse quantità di osservazioni mancanti o di valori anomali dovuti agli strumenti di rilevazione, le osservazioni che si ottengono dalle centraline, anche se rilevate a una alta frequenza, sono spesso disponibili per le analisi solo in misure aggregate (medie orarie o giornaliere).

Sembra ragionevole esprimere la concentrazione dell'inquinante mediante una combinazione di funzioni in grado di cogliere separatamente la variabilità di *grande scala* e quella di *piccola scala*. Un tale approccio è abbastanza diffuso nella letteratura statistica, nell'ambito dell'analisi di variabili ambientali (vd. per esempio per serie spaziali Korezlioglu e Loubaton, 1986; Cressie e al., 1990; Cressie, 1993, p.112; per serie temporali Nelson e Plosser, 1982; Kunsch, 1986; Bell, 1987).

L'idea consiste nel suddividere la variabilità di $\{z(s,t)\}_{s\in\mathbb{R}^2,t\in\mathbb{N}}$ in una parte che descriva principalmente l'andamento dovuto alle differenze globali spaziali e temporali (la variabilità di *grande scala*), e in una parte che invece colga sostanzialmente il "rumore" intorno all'altra componente (la variabilità di *piccola scala*). Ovviamente, come osserva Cressie (1993, p.113), una tale separazione non è univoca, ma necessita sempre di essere specificata attraverso opportune ipotesi sulle diverse componenti.

Sia, quindi, $\phi(\cdot,\cdot)$ la funzione che descrive la variabilità di *grande scala* e $\epsilon(\cdot,\cdot)$ quella che coglie la componente erratica di *piccola scala*. Una semplice possibilità per combinare queste due funzioni, è considerare il prodotto dei due termini

$$z(s,t) = \phi(s,t)\epsilon(s,t) \tag{1}$$

dove t è una variabile che indica il tempo (espresso, per esempio, in *giorni*) e s è un vettore bivariato composto dalle coordinate spaziali in un sistema di riferimento assegnato (fissati cioè un'origine e una direzione degli assi).

L'utilizzo di centraline mobili implica che osservazioni contemporanee sono poche. Se, infatti, la centralina in un fissato giorno è presente in un sito, non lo potrà essere in un altro. Di conseguenza, se si vogliono utilizzare solo le informazioni disponibili, è impossibile introdurre eventuali interazioni tra spazio e tempo nella formulazione di un modello, per la componente di *grande scala*.

D'altra parte una tale ipotesi, nel caso di variabili come la concentrazione di inquinanti può essere anche sensata, almeno come prima approssimazione. Infatti, se il periodo di riferimento è sufficientemente corto e in tale periodo non ci sono state modificazioni strutturali del territorio (come per esempio nuove urbanizzazioni o apertura di nuove linee di traffico), ci si può aspettare una sostanziale assenza di interazione tra spazio e tempo nella componente che vuole cogliere l'andamento complessivo della concentrazione.

Un semplice modello per la variabilità di *grande scala*, $\phi(\cdot,\cdot)$, capace di descrivere separatamente gli effetti legati alle variazioni nello spazio e quelli legati alle variazioni solo nel tempo, senza considerarne l'interazione, è il seguente:

$$\phi(s,t) = \phi_S(s)\phi_T(t) \tag{2}$$

dove $\phi_S(s)$ è una funzione che rappresenta la componente spaziale assunta fissa nel periodo di riferimento mentre la funzione $\phi_T(t)$ rappresenta la componente di *trend* e stagionale assunta comune per tutta l'aria considerata.

Un modello per la concentrazione degli inquinanti può, quindi, essere scritto

$$z(s,t) = \phi_S(s)\phi_T(t)\epsilon(s,t). \tag{3}$$

La scelta della struttura moltiplicativa fra le diverse componenti del modello (3) deriva dall'uso, diffuso nella letteratura statistica relativa alla modellazione ambientale, della trasformazione logaritmica per il trattamento di variabili di concentrazione (cfr. per esempio Seinfeld, 1986, pp.672).

Se si considerano i logaritmi, infatti, il modello (3) diventa un modello additivo del tipo

$$y(s,t) = \psi_S(s) + \psi_T(t) + a(s,t), \tag{4}$$

dove $y(s,t)$ =log$(z(s,t))$, $\psi_S(s)$ =log$(\phi_S(s))$, $\psi_T(t)$ =log$(\phi_T(t))$ e $a(s,t)$ =log$(\epsilon(s,t))$. Si ipotizza che la componente di *piccola scala*, $\epsilon(\cdot,\cdot)$, descriva la variabilità locale nello spazio e nel tempo, cercando di cogliere quella componente di rumore che si osserva intorno alle componenti di *larga scala*. Quindi $a(\cdot,\cdot)$ può essere pensato come un processo stocastico spazio-temporale con dipendenza nel tempo ridotta, ovvero possiamo assumere che $a(s',t')$ sia praticamente indipendente (o almeno incorrelato) da $a(s'',t'')$ se $\mid t'-t'' \mid$ è sufficientemente grande. Supporemo inoltre che $a(\cdot,\cdot)$ sia stazionario nello spazio e nel tempo e isotropico nello spazio.

Per quanto riguarda le due funzioni $\phi_S(\cdot)$ e $\phi_T(\cdot)$, ci si aspetta che siano abbastanza regolari; si possono pertanto specificare alcune ipotesi relative alla loro struttura e alle loro caratteristiche. Si può ipotizzare, allora, che tali funzioni seguano l'andamento di un processo stocastico non-stazionario; per esempio, si può pensare che $\phi_S(\cdot)$ e $\phi_T(\cdot)$ appartengano alla classe dei processi stazionari intrinseci (Cressie, 1993), classe sufficientemente ampia da contenere un gran numero di possibilità per le traiettorie dei due processi e nello stesso tempo permettere l'utilizzo di alcuni strumenti di immediata interpretazione come ad esempio il variogramma (Cressie, 1993).

In conclusione si vuole, dunque, specificare il modello trasformato (4) in modo tale che:

$a(s,t)$ sia un processo stocastico stazionario e isotropico che colga la variabilità di *piccola scala*;

$\psi_T(t)$ sia un processo stocastico puramente temporale non-stazionario che colga la componente della variabilità di *grande scala* legata al tempo;

$\psi_S(s)$ sia un processo stocastico puramente spaziale non-stazionario che colga la componente della variabilità di *grande scala* legata allo spazio.

Nella sezione 2.2 si presenterà una classe di modelli per la componente $a(s,t)$ che è molto usata in letteratura; nel seguito, poi, si descriveranno alcuni possibili modelli che specificano più concretamente i processi $\phi_T(\cdot)$ e $\phi_S(\cdot)$ (cfr. sezioni 2.3 e 2.4).

Una volta definito dettagliatamente il processo, si descrive una metodologia che può essere utilizzata per stimare contemporaneamente le varie componenti e i relativi parametri (sezione 2.5). Si tratta di costruire un modello *state space* che sia in grado di descrivere il processo complessivo $y(s,t)$ e, attraverso il filtro di Kalman e gli algoritmi ad esso legati, ottenere stime simultanee per i parametri del modello complessivo.

2.2 La componente di *piccola scala*

Il processo $\{\epsilon(s,t) : s \in D(t), t \in T\}$ deve cogliere la variabilità di *piccola scala* nello spazio e nel tempo, può, quindi, essere pensata come un processo stocastico con piccola dipendenza temporale (cioè $\epsilon(t', s')$ è praticamente indipendente da $\epsilon(t'', s'')$ se $\mid t' - t'' \mid$ è sufficientemente grande); una tale ipotesi permette l'interpretabilità del modello. Si assume che la componente di piccola scala $a(\cdot, \cdot)$ sia un processo stocastico stazionario in spazio e tempo e isotropico nello spazio, cioè

$$\text{cov}(a(t', s'), a(t'', s'')) = C(t' - t'', \| s' - s'' \|) \tag{5}$$

dove $\| \cdot \|$ è la distanza Euclidea tra i due punti. In realtà l'ipotesi di stazionarietà spaziale e temporale non sembra particolarmente impegnativa, mentre l'isotropia spaziale, ovvero l'assunzione che la dipendenza nello spazio sia funzione solo della distanza tra due siti e non dalla direzione determinata dalla linea che congiunge i due siti, sembra una richiesta più forte. È comunque possibile rilassare tale ipotesi richiedendo per esempio che la dipendenza sia funzione di qualche particolare direzione (la direzione dei venti dominanti, per esempio).

Per determinare una forma parametrica per tale processo ci si può soffermare sulle caratteristiche di secondo ordine di tale processo, che in quanto stazionario, lo descrivono completamente. Nel seguito ci si sofferma in particolare sul covariogramma.

Si sceglie un modello *separabile* (Cressie 1993, p.85) che è il prodotto di un correlogramma nello spazio e uno nel tempo. La relazione corrispondente tra i covariogrammi sarà

$$C(u, v; \theta) = C_T(u; \theta_T) C_S(v; \theta_S)/\sigma^2$$

dove $C(\cdot, \cdot; \theta)$ è un covariogramma parametrico per il processo spazio temporale, $C_T(\cdot; \theta_T)$ è un covariogramma per il processo puramente temporale e $C_S(\cdot; \theta_S)$ è un covariogramma per un processo puramente spaziale. Inoltre $\theta = \left[\theta_T, \theta_S, \sigma^2\right]^T$ è un vettore di parametri in cui θ_T sono i parametri relativi al processo temporale, θ_S sono quelli legati al processo spaziale, e σ^2 è la varianza di tale processo.

2.3 La componente di *grande scala* temporale

Una semplice e molto usata famiglia di modelli caratterizzata da realizzazioni sostanzialmente regolari, non necessariamente stazionarie, con caratteristiche che bene possono descrivere la richiesta di essere una componente di grande scala temporale per il processo in esame, che sia anche sufficientemente ampia da comprendere una vasta gamma di possibilità diverse è quella dei modelli strutturali (Harvey, 1990) a cui appartengono anche le passeggiate casuali (*random walk*).

Y_t è una *passeggiata casuale*, se μ_t è il suo valore atteso al tempo t, $\mu_t = E(Y_t)$, e

$$\mu_t = \mu_{t-1} + b_t$$

dove b_t, $t = 1, 2, \ldots$, è un processo aleatorio stazionario. Generalmente si assume che b_t sia un rumore bianco (*white noise*) o addirittura che i b_t siano indipendenti e identicamente distribuiti $\mathcal{N}(0, \sigma^2)$.

Un tale processo è *stazionario intrinseco* (non è stazionario, ma i suoi incrementi lo sono; si veda per esempio Chung, 1974, p.250-293 per una serie di risultati delle passeggiate casuali e Cressie, 1993 per la definizione di processo stazionario intrinseco).

Si osservi, inoltre, che la variabilità temporale può essere lenta e graduale, così da riflettere, per esempio, i cambiamenti continui nelle condizioni ambientali, ma può avere anche salti improvvisi che possono riflettere grossi cambiamenti nei principali fattori che influenzano la variabile in esame.

Si può anche specificare ulteriormente tale modello per consentire che l'innovazione del livello del processo possa variare in accordo ad un ulteriore processo stocastico che misura il cambiamento di pendenza nel processo originario. Per il particolare modello degli inquinanti che si sta esaminando il processo stocastico puramente temporale, non-stazionario, può essere del tipo

$$\begin{cases} \phi_T(t) = \mu_t \\ \mu_t = \mu_{t-1} + \beta_{t-1} + b_t \\ \beta_t = \beta_{t-1} + c_t \end{cases} \tag{6}$$

con con b_t e c_t processi aleatori stazionari per esempio b_t e c_t possono essere indipendenti (al variare di t e tra b_t e c_t) e $b_t \sim \mathcal{N}(0, \sigma_b^2)$ e $c_t \sim \mathcal{N}(0, \sigma_c^2)$.

Si osservi che tali modelli strutturali possono essere generalizzati in modo da comprendere componenti stagionali, cicliche e di trend (cfr. Harvey, 1990) sia "deterministiche" che "stocastiche".

Tale processo, se σ_b^2 è piccolo riesce a cogliere bene una traiettoria particolarmente liscia come quella che ci si può attendere dai dati giornalieri sulla concentrazione.

2.4 La componente di grande scala spaziale

Una classe di modelli con realizzazioni che possano bene descrivere la componente di grande scala spaziale può essere trovata in quel gruppo di processi che pur essendo non stazionari, appartengono alla famiglia dei processi stazionari intrinseci. Inoltre pare sensato ipotizzare l'isotropia spaziale per il modello della componente di grande scala; si potranno, infatti, modellare le anisotropie in maniera particolarmente semplice e diretta, attraverso la componente di piccola scala (cfr. Cressie, 1993 e Scarpa, 1997).

Un primo esempio di processo aleatorio non stazionario puramente spaziale può essere costruito generalizzando la passeggiata casuale (cfr. sezione 2.3) al caso in cui la variabile indicizzante il processo appartenga a $\mathbb{R}^2$, sia cioè continua (e non più discreta) e con due componenti (e non una sola). Tale processo, è detto *moto browniano multiparametrico* o *moto browniano di Lévy* (cfr. Lévy, 1954), ed è molto usato in molti campi della probabilità e della statistica. Proprietà probabilistiche e analitiche, rappresentazioni geometriche ed esemplificazioni di tale processo sono trattate in dettaglio in Lévy (1964). Qui è sufficiente ricordare come il processo in esame sia un esempio di processo non stazionario che è invece stazionario intrinseco e isotropico(cfr. Cressie, 1993).

Il *moto browniano multiparametrico* pur essendo interessante per descrivere un processo stazionario intrinseco in $\mathbb{R}^2$, è tuttavia descritto da un solo parametro (la varianza del processo), risultando, quindi, poco flessibile non consentendo di descrivere un gran numero di diverse forme di realizzazioni del processo.

Una possibile generalizzazione ad una classe più ampia di processi stazionari intrinseci, ma non necessariamente stazionari, è costituita dal *moto browniano frazionario.*

Un *moto browniano frazionario* in $\mathbb{R}^d$ è un processo stocastico gaussiano $Z(\cdot)$ caratterizzato dal variogramma della forma

$$\text{var}(Z(s+h) - Z(s)) \propto \|h\|^{2H}, \qquad 0 < H < 1, s, h \in \mathbb{R}^d. \tag{7}$$

La corrispondente funzione di covarianza, $C(s,u)$, dove $u = s + h$, è proporzionale a $\{\|s\|^{2H} + \|u\|^{2H} - \|s-u\|^{2H}\}$. Per dettagli e illustrazioni si veda Mandelbrot e Van Ness (1968).

I modelli visti finora descrivono tutti processi non stazionari che quindi ben risponderebbero alle necessità di modellare la componente di grande scala, a volte, però, sono troppo poco flessibili per potersi adattare alle complesse strutture spaziali che caratterizzano le concentrazioni di inquinanti; il numero ridotto di parametri, anche se molto informativi, non riesce, infatti, a cogliere tutta la variegata struttura dei processi che sono di interesse in questo lavoro.

Per cercare una classe di modelli un po' più ampia e generale, innanzitutto, si può osservare che un modello per il variogramma può essere considerato come una misura di distanza tra due punti nello spazio; è, allora, immediato verificare che se $2\gamma(u-v), u, v \in \mathbb{R}^n$ è un modello per il variogramma di un processo stazionario intrinseco nei punti u e v, il relativo modello per la funzione di covarianza (che non dipenderà solo dalla distanza tra u e v) può essere costruito come

$$C(u,v) = 2\gamma(u-0) + 2\gamma(v-0) - 2\gamma(u-v) \tag{8}$$

che è la funzione di covarianza di un processo non stazionario.

In questo caso è allora sufficiente definire un modello per il variogramma di un processo stazionario intrinseco e un'origine per tale processo ottenendo così il relativo modello per la funzione di covarianza.

Per determinare una classe di modelli per il variogramma sufficientemente ampia è possibile utilizzare la famiglia di modelli per il variogramma di processi stazionari intrinseci proposta da Matérn (1969) che si basa su un approccio spettrale.

Per tali processi, la rappresentazione spettrale del variogramma è

$$2\gamma^0(h) = \int_0^{+\infty} \frac{1 - Y_d(\lambda h)}{\lambda^2} d\Phi(\lambda), \tag{9}$$

dove Φ è una funzione non-decrescente in $(0, +\infty)$ tale che

$$\int_0^{+\infty} (1+\lambda^2)^{-1} d\Phi(\lambda) < +\infty,$$

e $Y_d(t) \equiv \left(\frac{2}{t}\right)^{\frac{d-2}{2}} \Gamma\left(\frac{d}{2}\right) J_{\frac{d-2}{2}}(t)$, e $J_\nu(\cdot)$ è la funzione di Bessel del primo tipo di ordine ν.

Per rendere tale modello utilizzabile in pratica, è sufficiente, per esempio, definire la funzione non-decrescente Φ come una funzione a gradini; il modello (9) diviene allora

$$2\gamma^0(h) = \sum_{i=1}^{g} \frac{1 - Y_d(\lambda_i h)}{\lambda_i^2} \alpha_i. \tag{10}$$

Per adattare il modello (10), una volta scelto il numero g di gradini, basta stimare i parametri α_i, $i = 1, \ldots, g$, che misurano l'ampiezza dei gradini e i parametri $\lambda_i, i = 1, \ldots, g$ che misurano le frequenze della rappresentazione spettrale.

Se si sceglie il modello (9) o il modello (10) come variogramma del processo da stimare, utilizzando la (8) è facile ottenere anche una stima per la funzione di covarianza dello stesso processo.

Si osservi che i modelli per il variogramma e il covariogramma così ottenuti sono *validi* (Cressie, 1993) per costruzione.

Si ha a disposizione così una classe di modelli particolarmente interessante. A seconda delle informazioni che si hanno sulla variabilità del processo che si vuole stimare, è possibile scegliere la sottoclasse di modelli da usare; la scelta di g infatti, permette di definire il numero di parametri necessari per cogliere la variabilità presente nel processo osservato. Una volta determinato g, poi i modelli appartenenti ad ogni sottoclasse sono molti e con forme completamente diverse tra di loro. Sembra, quindi, che questa scelta sia particolarmente interessante per determinare modelli legati a processi di concentrazione di inquinanti nelle città dove difficilmente, a priori, si hanno informazioni sulla variabilità del fenomeno nelle varie zone.

2.5 La stima simultanea del processo spazio-temporale

Si tratta ora di vedere come ottenere delle stime simultanee per i parametri del modello complessivo; una possibilità consiste nel cercare di costruire un modello *state space* che descriva la specificazione del processo in esame.

Per stimare le componenti del modello (4) a partire dalle serie storiche osservate in un certo numero di siti, si può allora costruire un vettore di tutte le osservazioni in ogni istante di tempo, $Y_t = [y(s_1,t),\ldots,y(s_n,t)]^T$, dove possono essere presenti valori mancanti in corrispondenza dei siti in cui, per quell'istante di tempo, non si è avuta la rilevazione (dati mancanti sia reali che fittizi).

Si può scomporre il problema in tre parti distinte che poi possono essere considerate congiuntamente.

Una prima parte riguarda la componente di grande scala spaziale $\psi_S(s)$. Tale funzione non dipende dal tempo, ma unicamente dal sito nello spazio in cui è stata effettuata la rilevazione. Se si indica con $\Psi_{S,t} = \Psi_S$ il vettore di tutte le componenti di grande scala spaziale negli n siti osservati al tempo t che è uguale per ogni t, l'"evolversi" di tale processo nel tempo può essere descritto dal modello

$$\begin{cases} \Psi_{S,t} = \Psi_{S,t-1}, t = 1,2,\ldots \\ \Psi_{S,0} = \delta_S \end{cases} \tag{11}$$

dove il vettore $\delta_S \sim \mathcal{N}(0, C_S)$ con C_S matrice di varianze e covarianze, di dimensione $n \times n$, del processo che descrive la componente di grande scala dei siti per i quali sono disponibili le osservazioni; tale matrice non dipende dal tempo e può essere costruita a partire dalla stima del covariogramma per il processo di grande scala (cfr. sezione 2.4) $C_S = [\mathrm{cov}(\psi_S(s_i), \psi_S(s_j))], (i,j = 1,\ldots,n)$.

La seconda parte del modello (4) è quella riguardante la componente di grande scala temporale. È immediato riscrivere il modello strutturale, visto in sezione 2.3, in forma *state-space*. Per esempio il modello del tipo 6 può essere riscritto come

$$\begin{cases} \psi_T = [1 \;\; 0] \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} \\ \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} b_t & 0 \\ 0 & c_t \end{bmatrix}. \end{cases}$$

dove b_t e c_t sono indipendenti (al variare di t e tra b_t e c_t) e $b_t \sim \mathcal{N}(0,\sigma_b^2)$ e $c_t \sim \mathcal{N}(0,\sigma_c^2)$.

Per quanto riguarda la terza parte del modello (4), si deve assumere che, almeno approssimativamente, il processo degli errori, $\epsilon(s,t)$, possa avere rappresentazione *state space*. Ciò significa che, se si chiama $E_t = [\epsilon(s_1,t),\ldots,\epsilon(s_n,t)]^T$ il vettore di tutte le osservazioni del processo

di piccola scala nei diversi siti al tempo t, si assume che esista un vettore di stato $E_{STATO} = [\epsilon_{STATO}(s_1,t),\ldots,\epsilon_{STATO}(s_g,t)]^T$ per cui si può scrivere

$$\begin{cases} E_t = HE_{STATO} \\ E_{STATO} = \Delta E_{STATO} + \delta_{\epsilon t} \end{cases}$$

dove $\delta_{\epsilon t}$ ha media nulla e matrice di varianze e covarianze Λ_0.

Le matrici H $(n \times g)$, Δ $(g \times g)$ e Λ_0 $(g \times g)$, e il numero g di elementi presenti nel vettore di stato, vengono determinate grazie alle assunzioni fatte sul processo di piccola scala.

Si osservi, comunque, che se il processo $\epsilon(s,t)$ è realmente una componente di piccola scala, cioè un processo che coglie unicamente il rumore intorno alle altre componenti del modello (sostanzialmente ciò significa che nella decomposizione di Wold di tale processo è presente solo la componente stocastica, cfr. Brockwell e Davis, 1991, p.187), è sempre approssimabile con una rappresentazione *state space*, per esempio con un processo autoregressivo troncato a un ritardo opportuno, o con approssimazioni *state space* basate sulle autocorrelazioni fino ad un opportuno *lag* temporale come proposto da Aoki (1990).

Una volta specificate le rappresentazioni per le tre diverse parti del modello (4), è immediato unirle insieme se si assume che ciascuna delle diverse componenti non sia legata alle altre, assunzione peraltro basilare nel modello in esame dove si è interessati a cogliere separatamente le diverse strutture di variabilità della concentrazione degli inquinanti.

Il modello *state space* complessivo può, allora, essere scritto come

$$\begin{cases} Y_t = FX_t \\ X_t = GX_{t-1} + \eta_t \end{cases} \tag{12}$$

con $\eta_t \sim \mathcal{N}(0,W)$; il *vettore di stato* X_t, di dimensione $(n+2+g) \times 1$, è il vettore

$$X_t = [\psi_S(s_1),\ldots,\psi_S(s_n),\mu_t,\beta_t,\epsilon_{\text{STATO},1},\ldots,\epsilon_{\text{STATO},g}]^T$$

costituito dai tre vettori di stato delle tre parti distinte.

Le matrici presenti nel modello (12) sono facilmente determinabili:

F è una matrice fissata di dimensione $n \times (n+2+g)$

$$F = \left[\begin{array}{c|cc|c} & 1 & 0 & \\ & 1 & 0 & \\ I_n & \vdots & \vdots & H \\ & 1 & 0 & \end{array}\right]$$

G è una matrice di dimensione $(n+2+g) \times (n+2+g)$

$$G = \left[\begin{array}{c|cc|c} & 0 & 0 & \\ & \vdots & \vdots & \\ I_n & 0 & 0 & O \\ \hline 0^T & 1 & 1 & 0^T \\ 0^T & 0 & 1 & 0^T \\ \hline & 0 & 0 & \\ O & \vdots & \vdots & \Delta \\ & 0 & 0 & \end{array}\right]$$

W è la matrice di varianze e covarianze si η_t di dimensione $(n+2+g) \times (n+2+g)$

$$W = \left[\begin{array}{c|cc|c} & 0 & 0 & \\ O & \vdots & \vdots & O \\ & 0 & 0 & \\ \hline 0^T & \sigma_b^2 & 0 & 0^T \\ 0^T & 0 & \sigma_c^2 & 0^T \\ \hline & 0 & 0 & \\ O & \vdots & \vdots & \Lambda_0 \\ & 0 & 0 & \end{array}\right].$$

Una volta specificato il modello, si può utilizzare il filtro di Kalman, per ottenere la funzione di verosimiglianza da cui si derivano stime per i parametri, previsioni per il processo, e infine stime di lisciamento (vedi per es. West e Harrison, 1989).

Si osservi che la rappresentazione *state space* (12) permette di trattare senza grosse difficoltà le osservazioni mancanti. Tale caratteristica è fondamentale per il problema che si sta analizzando. La rilevazione tramite centraline mobili, infatti, come già osservato, implica la presenza di dati mancanti fittizi, ma gli algoritmi per il filtraggio, la stima di massima verosimiglianza, il lisciamento e la previsione, sono facilmente modificabili per tener conto di tale fatto. In particolare, per ottenere le stime e le previsioni spaziali e temporali, è sufficiente non considerare l'equazione delle osservazioni per gli istanti di tempo in cui non si hanno osservazioni.

Si osservi inoltre che la letteratura statistica e ingegneristica presenta una serie di algoritmi alternativi per risolvere il filtro di Kalman per far fronte a problemi computazionali specifici. Nel caso in esame un inconveniente che appare spesso riguarda il fatto che, ad ogni passo, le stime delle matrici di varianze e covarianze delle previsioni dovrebbero risultare nonsingolari e definite positive, visto che devono essere invertite ad ogni iterazione, mentre il procedere dell'algoritmo non assicura questa proprietà, soprattutto nelle prime iterazioni dove gli errori computazionali possono influire parecchio nella determinazione delle matrici. L'assenza di questa proprietà rende le stime ottenute inutili nella determinazione del modello che non potrebbe quindi venire utilizzato.

L'algoritmo *square root* (cfr. per esempio Anderson e Moore, 1979) costituisce un possibile strumento per superare il problema, considerando la decomposizione di Cholewski della matrice di interesse, in modo tale da ottenere una sorta di radice quadrata di tale matrice che rende quindi il suo quadrato certamente definito positivo.

2.6 Previsione spaziale

È possibile, cercare di fare previsioni sul comportamento del processo della concentrazione dell'inquinante in punti non osservati, a partire da tutte le informazioni ottenute fino al tempo t. Tale obiettivo per quanto riguarda l'aspetto temporale può essere affrontato con gli usuali algoritmi di previsione legati al modello *state space* (vedi per es. West e Harrison, 1989); sembra invece di particolare interesse, qui, considerare il problema di previsione in siti spaziali non osservati.

Per prevedere l'andamento del processo in un sito non osservato è sufficiente aggiungere nel vettore di stato un paio di nuovi elementi, uno in grado di cogliere la componente di grande scala

puramente spaziale in quel sito e l'altro legato alla componente di piccola scala sempre in quel sito (se necessario la componente di piccola scala può essere caratterizzata da più di un elemento per ogni istante di tempo).

Si applica, quindi, il filtro di Kalman usando le stime dei parametri ottenute massimizzando la verosimiglianza per i siti osservati, e in successione l'algoritmo di lisciamento o di previsione temporale. Si ottiene così, per ogni nuovo sito di interesse, la stima complessiva del processo per ogni istante di tempo che si vuole considerare.

È chiaramente possibile stimare contemporaneamente il processo in diversi nuovi siti, aggiungendo nel vettore di stato non solo una coppia di elementi ma tante coppie quanti sono i nuovi siti su cui si vuole la stima. Tipicamente è interessante ottenere la stima della concentrazione dell'inquinante in esame in un reticolo di siti nella città in esame.

Si osservi che, al crescere del numero di nuovi siti il calcolo delle stime diviene computazionalmente sempre più oneroso, visto che si deve, ad ogni iterazione, invertire la matrice di varianze e covarianze della previsione del vettore di stato, la cui dimensione aumenta al crescere del numero di elementi presenti nel vettore di stato.

Tale problema può essere risolto osservando che, per il modello (12), la previsione in diversi siti spaziali è equivalente ad impostare e risolvere delle equazioni di ***kriging*** sui risultati della stima del vettore di stato nei siti osservati, ottenuta tramite lisciamento. In questa maniera il tempo computazionale si riduce drasticamente in quanto si tratta di risolvere unicamente un sistema lineare.

Si vuole ora mostrare l'uguaglianza dei due procedimenti. Si osservi che lo stimatore di previsione che si ottiene utilizzando in sequenza le due procedure di stima è uno stimatore lineare. Infatti, se si chiama Y il vettore formato dalla giustapposizione di tutti i vettori osservati nei diversi istanti di tempo, S_{OSS} il vettore formato da tutti i vettori di stato dei diversi istanti di tempo e S_{ALTRI} il vettore di stato corrispondente ai nuovi siti su cui si vuole la stima delle varie componenti, si può scrivere lo stimatore ***kriging*** come $\hat{S}_{ALTRI} = AS_{OSS}$, e lo stimatore *state space* come $\hat{S}_{OSS} = BY$.

Utilizzando le stime in successione si può scrivere $\tilde{S}_{ALTRI} = ABY$, da cui si vede che tale stimatore di previsione è lineare.

Per vedere che i due stimatori $\hat{S}_{ALTRI}$ e $\tilde{S}_{ALTRI}$ coincidono, è sufficiente[1], mostrare che

$$E\{(S_{ALTRI} - \tilde{S}_{ALTRI})Y\} = 0. \tag{13}$$

Infatti, è noto che lo stimatore di previsione in nuovi siti del modello *state space* (12) è lineare non distorto e a minima varianza, per cui se è vera la relazione (13) lo stimatore $\tilde{S}_{ALTRI}$ coincide con tale stimatore, mostrando quindi che i due diversi approcci conducono agli stessi risultati.

Per mostrare la relazione (13), è sufficiente sommare e sottrarre al primo termine della (13) la stessa quantità $\hat{S}_{ALTRI}$,

$$E\{(S_{ALTRI} - \tilde{S}_{ALTRI})Y\} = E\{(S_{ALTRI} - \hat{S}_{ALTRI} + \hat{S}_{ALTRI} - \tilde{S}_{ALTRI})Y\}$$

da cui

$$E\{(S_{ALTRI} - \tilde{S}_{ALTRI})Y\} = E\{(S_{ALTRI} - \hat{S}_{ALTRI})Y\} + E\{(\hat{S}_{ALTRI} - \tilde{S}_{ALTRI})Y\}.$$

[1] È noto che se X e Y sono due variabili e $\hat{Y}$ è uno stimatore che dipende linearmente da X, $\hat{Y} = FX$ con F matrice nota, una condizione necessaria e sufficiente affinché tale stimatore sia non distorto a minima varianza è che $E\{(Y - \hat{Y})X\} = 0$, cioè che $E\{(Y - FX)X\} = 0$.

Sostituendo a $\hat{S}_{ALTRI}$ il suo valore in funzione di S_{OSS} e raccogliendo a fattor comune la matrice A si ottiene

$$E\{(S_{ALTRI} - \tilde{S}_{ALTRI})Y\} = E\{(S_{ALTRI} - AS_{OSS})Y\} + E\{AS_{OSS} - ABY)Y\} \quad (14)$$
$$= E\{(S_{ALTRI} - AS_{OSS})Y\} + AE\{S_{OSS} - BY)Y\}. \quad (15)$$

Il secondo termine di quest'ultima espressione è nullo perché $\hat{S}_{OSS} = BY$ è uno stimatore di S_{OSS} che, per definizione, dipende in maniera lineare da Y, soddisfa quindi alla nota proprietà di tali stimatori.

Il primo termine, invece, osservando che, per definizione, nel modello *state space* (12) il vettore di stato è legato al vettore delle osservazioni da una relazione lineare del tipo $Y = MS_{OSS}$, può essere riscritto come

$$E\{(S_{ALTRI} - AS_{OSS})Y\} = E\{(S_{ALTRI} - AS_{OSS})MS_{OSS}\} \quad (16)$$
$$= ME\{(S_{ALTRI} - AS_{OSS})S_{OSS}\} = 0 \quad (17)$$

che è nullo sempre per la stessa relazione. I risultati ottenuti permettono quindi di dire che la relazione (13) è vera per gli stimatori considerati e quindi le due procedure portano allo stesso risultato.

2.7 La rete di monitoraggio

Utilizzando un modello di questo tipo è anche possibile cercare di determinare i nuovi siti in cui installare le centraline mobili per i periodi futuri.

Per esempio, si può cercare di determinare la successione di siti che minimizza in maniera sequenziale una misura complessiva della varianza di previsione del processo di grande scala in tutta la regione spaziale considerata (Arbia e Switzer, 1995).

Sia T il giorno corrente. Fissato il numero di giorni in cui la centralina si ferma nel primo sito, diciamo t_1, determiniamo il primo sito in maniera tale che sia minima la somma, per $t = T+1, \ldots, T+t_1$, degli integrali delle varianze di previsione su tutta la regione.

Tale procedura può venire iterata per determinare il numero di siti che interessano.

Condizionatamente al valore trovato, il secondo sito viene determinato in maniera analoga ma con riferimento al periodo $t = T+t_1+1, \ldots, T+t_1+t_2$, dove t_2 ovviamente è il numero (fissato a priori) di giorni in cui la centralina rimarrà nel secondo sito.

Si osservi che le varianze necessarie possono essere calcolate utilizzando la parte che riguarda le varianze e covarianze dell'algoritmo di filtraggio (a priori si sa unicamente il luogo dove è installata la centralina nel primo periodo, e non il valore della concentrazione, ma ciò è sufficiente). L'integrale può essere ragionevolmente approssimato utilizzando una griglia di punti.

Si osservi che, per iterare la procedura, non è necessario conoscere il valore della concentrazione nei nuovi siti, è quindi possibile programmare gli spostamenti della centralina per un periodo di tempo sufficientemente lungo.

È chiaro, comunque, che ad un certo punto è opportuno fermare l'algoritmo, considerare le nuove osservazioni e ottenere nuove stime di massima verosimiglianza per i parametri sfruttando tutte le informazioni presenti.

Si può pensare anche di utilizzare una combinazione (eventualmente ponderata) delle varianze di previsione per il processo di grande e di piccola scala, o eventualmente dei coefficienti di variazione (che non sono influenzati dall'unità di misura).

3 La concentrazione di biossido di zolfo a Padova

3.1 I dati

Le osservazioni su cui viene applicato il metodo presentato finora riguardano la concentrazione di SO_2 nella zona urbana di Padova negli anni atmosferici 1991, 1992 e 1993 (1/4/1991 - 31/3/1994).

L'area geografica coperta è approssimativamente un rettangolo di 11.5×7.5km, limitata a ovest dalla catena dei Colli Euganei (non inclusi), a sud da un canale artificiale (Canale Scaricatore) e a nord-est dal fiume Brenta.

I dati sono stati raccolti da (*i*) due centraline fisse funzionanti per tutti e tre gli anni, (*ii*) una terza centralina fissa installata il 16 novembre 1993, (*iii*) due centraline mobili che si sono spostate senza alcuna regola specifica in altri 14 siti, stabilendosi per una successione di giorni variabile da sito a sito (va osservato che il numero di siti visitati dalle centraline mobili è ridotto sia perché le centraline stesse sono state utilizzate in altre parti della provincia di Padova, sia perché, a causa di un guasto, una centralina mobile ha rimpiazzato per circa un anno una delle centraline fisse).

Non essendoci, per quanto riguarda l'accuratezza delle misure, differenze sostanziali tra centraline fisse e mobili si possono considerare tutti i 17 siti come luoghi in cui sono insediate delle centraline fisse che nel caso degli ultimi 14 siti sono quasi sempre inoperanti.

L'insieme dei dati viene visto come un insieme di 17 variabili $z(s_j, t_i)$ dove i t_i indicano i giorni (da 1 a 1094) e gli s_j $(j = 1, \ldots, 17)$ denotano i siti, con una grande quantità di *dati mancanti*, alcuni reali ed alcuni fittizi.

I giorni in cui le centraline mobili sono state dislocate in ciascun sito ed il numero di giorni per cui mancano le osservazioni sono riportati nella tabella 1. La posizione geografica dei 17 siti in cui sono state effettuate le rilevazioni è riportata nella mappa della figura 1.

Le caratteristiche tecniche degli strumenti di rilevazione del biossido di zolfo attualmente utilizzati non permettono di determinare il valore delle concentrazioni se queste sono inferiori a 10 $\mu g/m^3$, ma poiché tale forma di censura coinvolge le osservazioni meno interessanti dal punto di vista della rilevazioni degli inquinanti (si tratta infatti di situazioni di "non inquinamento") si è deciso di assegnare a tutte le osservazioni censurate il valore 5 $\mu g/m^3$ (assumendo cioè l'equidistribuzione dei valori tra 0 e 10 $\mu g/m^3$).

Dato il particolare schema di rilevazione, la caratteristica principale di tali dati è che le osservazioni contemporanee sono poche: si hanno infatti un minimo di 2 e un massimo di 4 osservazioni nello stesso giorno.

3.2 La specificazione del modello

Come presentato nella sezione 2 si predispone un modello additivo per descrivere la trasformata logaritmica della concentrazione dell'inquinante. La scelta della trasformata logaritmica, nel caso in esame, è supportata anche dai risultati di alcune semplici analisi descrittive ed esplorative (come per esempio quelle basate sugli *spread versus level plot* presentata da Hoaglin, et al. 1985)

Tabella 1: Siti osservati

	Site	Coord. E-W (km)	Coord. N-S (km)	Period of observation	Number of days observed	Missing observations
A	Ospedale*	0,800	0,550	01.04.91–31.03.94	1096	261
B	S.Gregorio*	3,300	0,675	01.04.91–31.03.94	1096	111
C	Arcella*	0,950	3,425	16.11.93–31.03.94	137	9
D	Via Raggio di Sole	-0,900	1,550	12.04.91–20.05.91	39	0
E	Guizza	-0,600	-2,075	16.05.91–10.06.91	26	6
F	Piazza Castello	-0,550	0,425	04.06.91–29.06.91	26	10
G	Voltabarozzo	2,000	-1,975	11.06.91–21.06.91	11	4
H	Corso Stati Uniti	4,500	-0,450	25.10.91–07.12.91	44	6
I	Piazzale Stanga	2,000	1,150	16.12.91–30.04.92	137	27
L	Chiesa Nuova	-3,725	2,025	11.01.92–29.01.92	19	0
M	Piazza Cavour	0,100	1,075	15.09.92–05.10.92	21	1
N	Mandria	-2,725	-3,125	17.09.92–05.10.92	19	0
O	Ponte di Brenta	5,250	2,950	07.10.92–01.11.92	26	0
P	Selvazzano	-5,425	-0,800	10.11.92–02.12.92	23	0
Q	Bassanello	-0,725	-1,300	11.12.92–31.03.93	111	6
R	Piazzale San Giovanni	-0,950	0,675	06.04.93–13.05.93	38	0
S	Piazza Insurrezione	-0,125	1,225	09.11.93–31.03.94	143	0

* Centraline fisse

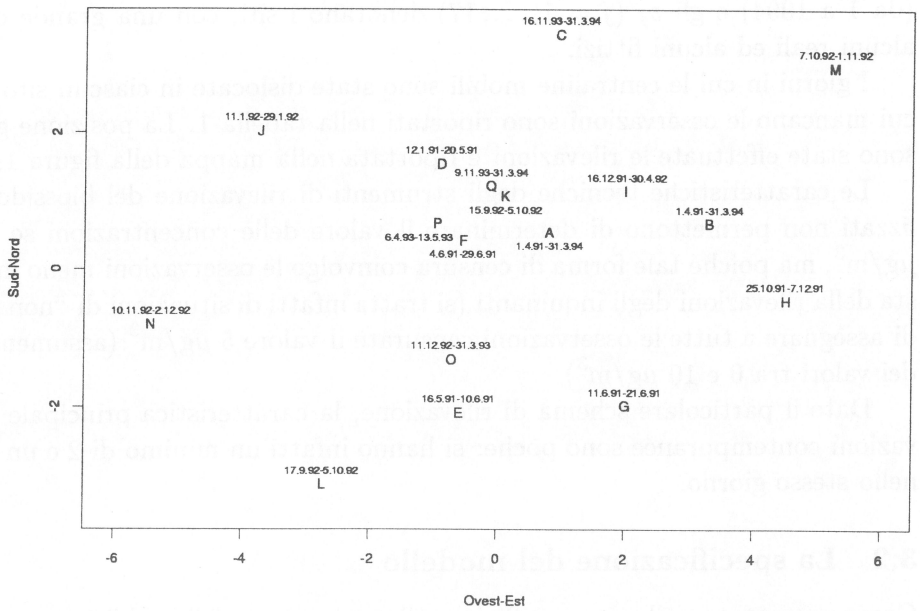

Figura 1: Posizione geografica e periodo di rilevazione delle centraline.

sulla distribuzione della concentrazione dell'SO_2 che hanno aiutato a scegliere tale trasformazione piuttosto che altre.

Si cerca, allora, un modello del tipo

$$y(s,t) = \psi_S(s) + \psi_T(s) + a(s,t). \tag{18}$$

dove $y(s,t)$ =log$(z(s,t))$, $\psi_S(s)$ =log$(\phi_S(s))$ è un processo stocastico puramente spaziale, $\psi_T(t)$ =log$(\phi_T$ un processo puramente temporale e $a(s,t)$ =log$(\epsilon(s,t))$ un processo spazio-temporale.

Nella specificazione ulteriore dei processi coinvolti vengono ora presentate le scelte effettuate.

Per quanto riguarda il processo $a(s,t)$, precedenti analisi esplorative e nonparametriche (cfr. Scarpa, 1997) sul covariogramma della concentrazione di SO_2 al variare di spazio e tempo hanno portato a considerare un processo con correlogramma separabile tra componente spaziale e componente temporale con la parte spaziale che segue un modello sferico e la parte temporale descritto da un processo autoregressivo di ordine uno

$$C(u,v;\theta) = \begin{cases} \sigma^2 a^u & \text{se } v = 0 \\ \sigma^2 a^u b \left\{1 - \frac{3}{2}\frac{v}{c} + \frac{1}{2}\left(\frac{v}{c}\right)^3\right\} & \text{se } 0 < v < c \\ 0 & \text{se } v \geq c. \end{cases} \tag{19}$$

Per quanto riguarda, invece, le ipotesi relative ai processi che descrivono le componenti di grande scala, si è scelto, per la componente puramente temporale, di adattare un modello a passeggiata casuale come visto nella sezione 2.3, mentre per la componente puramente spaziale si è specificato il modello (10) scegliendo $g = 3$ in modo che la funzione di densità spettrale risulti particolarmente semplice (funzione a gradini con tre livelli), ottenendo così il modello per il variogramma

$$2\gamma^0(h) = \sum_{i=1}^{3} \frac{1 - Y_d(\lambda_i h)}{\lambda_i^2}\alpha_i, \tag{20}$$

con $Y_d(t) \equiv \left(\frac{2}{t}\right)^{\frac{d-2}{2}} \Gamma\left(\frac{d}{2}\right) J_{\frac{d-2}{2}}(t)$, e $J_\nu(\cdot)$ è la funzione di Bessel del primo tipo di ordine ν.

La funzione di covarianza per tale processo può essere facilmente calcolata attraverso la relazione (8), da cui si ottiene la matrice C_S.

Le altre matrici per la formulazione *state space* del modello (cfr. sezione 2.5) possono essere costruite come

$$H = I_n, \Lambda_{0,uv} = C(u,v,\theta), \Delta = aI_{17}.$$

Si utilizza l'algoritmo *square root* per effettuare il filtraggio assumendo come valori iniziali per gli elementi diagonali della matrice di varianze e covarianze dell'errore per la componente di grande scala temporale numeri molto elevati (9×10^{10}); in realtà tale assunzione non influisce nelle stime del modello visto che già alla seconda iterazione non resta alcuna traccia dei valori iniziali, è però necessaria per far partire l'algoritmo di filtraggio.

Si determinano così facilmente le stime di massima verosimiglianza dei parametri che sono riassunte nella tabella 2.

È immediato anche implementare gli algoritmi di lisciamento e previsione sia spaziale che temporale.

Tabella 2: Stima di massima verosimiglianza per i parametri.

Parametro	Valore
σ_b^2	0.0714
a	0.8026
b	0,9999
c	1,3750
σ^2	0,0319
λ_1	2,0967
α_1	0,4859
λ_2	0,0052
α_2	0,0013
λ_3	8,6094
α_3	8,8989

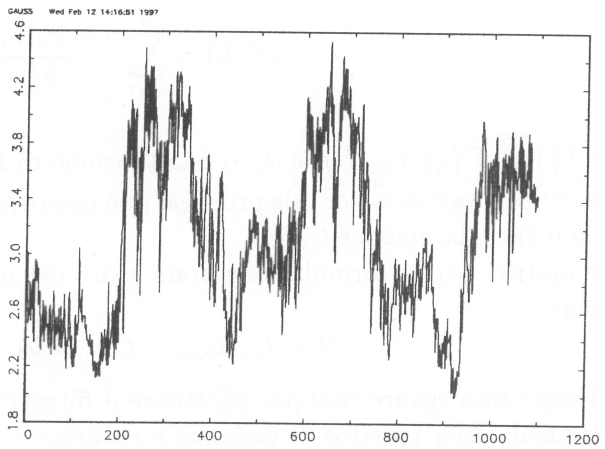

Figura 2: Stima della media del processo che coglie la componente temporale $\psi_T(\cdot)$.

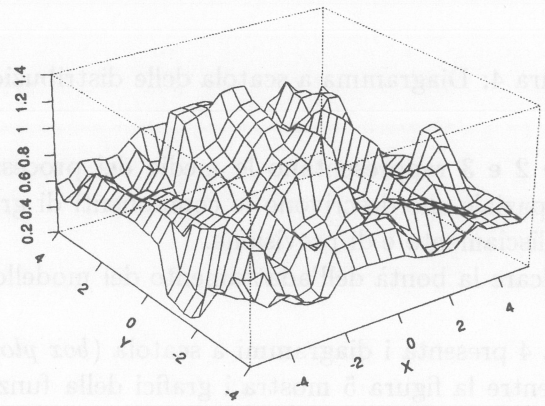

Figura 3: Stima della media del processo che coglie la componente spaziale $\psi_S(\cdot)$.

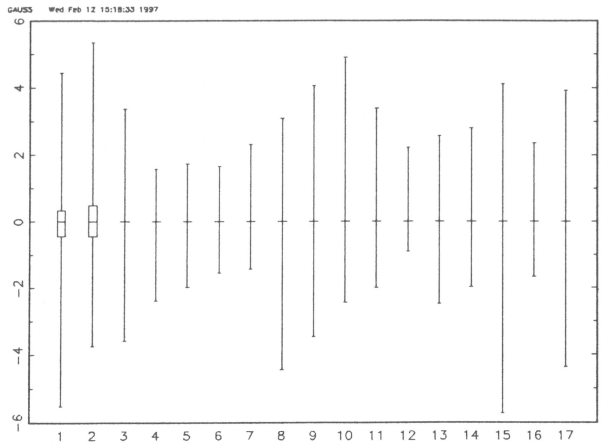

Figura 4: Diagramma a scatola delle distribuzioni dei residui nei 17 siti osservati.

Le figure 2 e 3 rappresentano le medie dei processi rispettivamente puramente temporale e puramente spaziale che descrivono le componenti di grande scala ϕ_T e ϕ_S ottenute attraverso gli algoritmi di lisciamento e di previsione.

Per verificare la bontà dell'adattamento del modello si è effettuata un'analisi dei residui standardizzati.

La figura 4 presenta i diagrammi a scatola (*box plot*) della distribuzione dei residui nei 17 siti osservati, mentre la figura 5 mostra i grafici della funzione di autocorrelazione dei residui con le bande di confidenza rispetto ai *lag* temporali per ogni sito osservato.

3.3 La rete di rilevazione a Padova

È possibile impostare una procedura sequenziale per determinare i nuovi siti su cui installare le centraline nei periodi futuri come è stato mostrato nella sezione 2.7.

Per esempio, se ci si concentra unicamente nel processo di grande scala spaziale, si può determinare un primo sito che minimizzi l'integrale della varianza di *kriging* con le informazioni presenti fino al 31 marzo 1994. Si ricorda che, come visto nella sezione 2.7, tale misura dell'errore di previsione coincide con l'errore di previsione basato sulla stima *state space* del modello. Una volta determinato il primo sito dove installare la centralina, lo si considera nel modello *state space* aggiungendo una coppia di nuovi elementi nel vettore di stato, uno che colga la componente di *grande scala* spaziale nel nuovo sito e uno per la componente di piccola scala. Si aggiornano così le matrici di varianze e covarianze degli errori e quelle di trasferimento tra X_t e Y_t e tra X_{t-1} e X_t.

Si procede poi iterando il filtro, ad esempio, per 7 giorni. L'integrale, calcolato numericamente, su tutto lo spazio considerato delle nuove stime delle varianze di *kriging* con le osservazioni fino al 7 aprile, può essere minimizzato per ottenere un secondo sito.

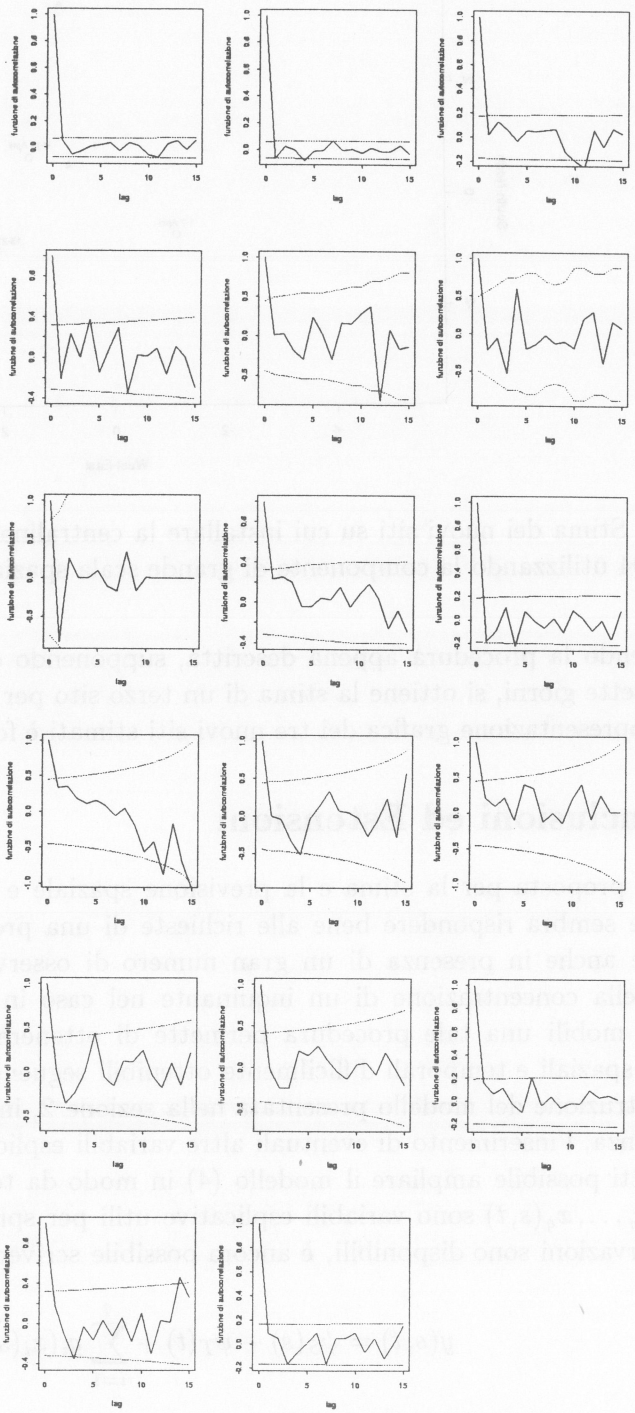

Figura 5: Funzioni di autocorrelazione temporale dei residui nei 17 siti osservati.

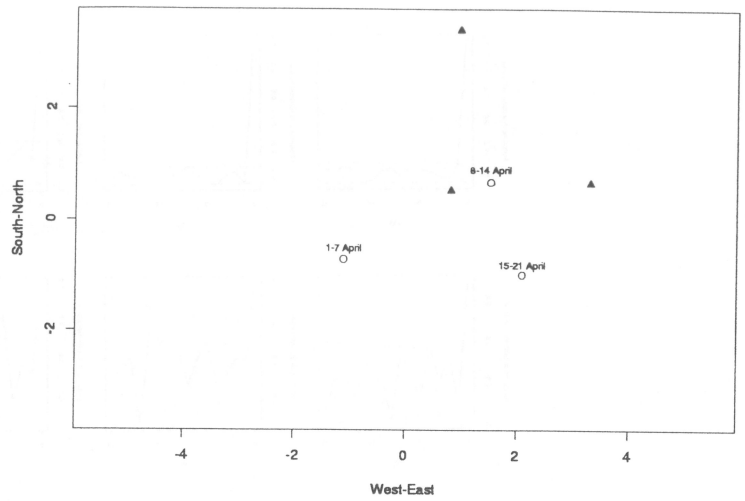

Figura 6: Stima dei nuovi siti su cui installare la centralina mobile nelle tre settimane dal 1 al 21 aprile 1994 utilizzando la componente di grande scala spaziale

Ripetendo la procedura appena descritta, supponendo di fermare la centralina in questo sito per altri sette giorni, si ottiene la stima di un terzo sito per la settimana dal 15 al 21 aprile 1994.

La rappresentazione grafica dei tre nuovi siti stimati è fornita nella figura 6.

4 Conclusioni ed Estensioni

La strada proposta per la stima e la previsione spaziale e temporale della concentrazione di un inquinante sembra rispondere bene alle richieste di una procedura spazio-temporale che riesca a funzionare anche in presenza di un gran numero di osservazioni mancanti. In particolare, per l'analisi della concentrazione di un inquinante nel caso in cui le osservazioni siano raccolte da centraline mobili una tale procedura permette di ottenere stime di massima verosimiglianza e previsioni spaziali e temporali difficilmente ottenibili seguendo altre vie.

La costruzione del modello presentata nella sezione 2, inoltre, può essere estesa, permettendo, all'occorrenza, l'inserimento di eventuali altre variabili esplicative disponibili.

È infatti possibile ampliare il modello (4) in modo da tener conto di eventuali altre variabili. Se $x_1(s,t), \ldots, x_q(s,t)$ sono variabili esplicative utili per spiegare la concentrazione di inquinanti, le cui osservazioni sono disponibili, è ancora possibile scrivere un modello del tipo

$$y(s,t) = \psi_S(s) + \psi_T(t) + \sum_{i=1}^{q} g_i(x_i(s,t)) + a(s,t), \tag{21}$$

dove le $g_i(\cdot)$ sono funzioni diverse una per ciascuna variabile esplicativa a disposizione.

Un tale modello si trasforma in un modello *state space*, analogo al (12), dove il vettore di stato

avrà un gruppo di righe in più per ogni variabile esplicativa inserita, una per ogni sito in cui tale variabile è osservata.

Si osservi che per poter stimare tale modello bisogna conoscere, o quantomeno stimare, la funzione di correlazione incrociata tra le variabili esplicative e la concentrazione di inquinanti. È così possibile inserire nel modello variabili come il traffico automobilistico o la concentrazione di abitazioni o di popolazione nelle varie zone della città.

Bibliografia

- B.D.O. ANDERSON e J.B. MOORE (1979), *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, N.J.
- T. AOKI (1990), *State Space Modelling of Time Series*, Springer-Verlag,
- G. ARBIA e P. SWITZER (1994), Spatial sampling desgns for stratified correlated units with unequal variances, *Working Paper*, N. 94.6, Dipartimento di Scienze Statistiche, Padova.
- W. BELL (1987), A note on overdifferencing and the equivalence of seasonal time series models with monthly means and models with $(0,1,1)_{12}$ seasonal parts when $\theta = 1$, *Journal of Business and Economic Statistics*, **5**, 383–387.
- P.J. BROCKWELL e R.A. DAVIS (1991). *Time Series: Theory and Methods.* Springer-Verlag, New Tork.
- K.L. CHUNG (1974), *A Course in Probability Theory*, Academic Press, New York.
- N. CRESSIE (1990), The origins of kriging, *Mathematical Geology*, **22**, 239–252.
- N. CRESSIE (1993), *Statistics for Spatial Data*, John Wiley & Sons, New York.
- T. HAAS (1995), Local Prediction of a Spatio-Temporal Process with an Application to Wet Sulfate Deposition, *Journal of the American Statistical Association*, **90**, 1189–1199.
- T. HAAS (1996), Multivariate Spatial Prediction in the Presence of Nonlinear Trend and Covariance Nonstationarity, *Environmetrics*, **7**, 145–165.
- A.C. HARVEY (1990), *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, Cambridge.
- D. C. HOAGLIN, F. MOSTELLER e J. W. TUKEY (1985), *Exploring data tables, trends, and shapes*, John Wiley & Sons New York.
- H. KOREZLIOGLU e P. LOUBATON (1986), Spectral factorization of wide sense stationary processes on Z^2, *Journal of Multivariate Analysis*, **19**, 24–47.
- H. R. KUNSH (1986), Discrimination between monotonic trends and long-range dependence, *Journal of Applied Probability*, **23**, 1025–1030.

- P. LÉVY (1954), Le mouvement Brownien, *Memorial des Sciences Mathematiques*, **126**, Gauthier-Villars, Paris.
- P. LÉVY (1964), Processus stochastiques et mouvement Brownien, Gauthier-Villars, Paris.
- B.B. MANDELBROT e J.W. VAN NESS (1968), *Fractional Brownian motions, fractional noises and applications.* SIAM Review, **10**, 422–437.
- B. MATÉRN (1969), Spatial Variation, *Meddelanden fran Statens Skogsforskningsinstitut*, **49**, No. 5. [Seconda edizione (1986), *Lecture Notes in Statistics*, **36**, Springer, New York.]
- C. R. NELSON e C. I. PLOSSER (1982), Trends and random walks in macroeconomic time series, *Journal of Monetary Economics*, **10**, 139–162.
- B. SCARPA (1997), Space-time modelling of SO2 by combining data of fixed and mobile monitoring stations, Università degli Studi di Padova, Dipartimento di Scienze Statistiche, *Technical Report*, **1997.12**.
- J.H. SEINFELD (1986), *Atmospheric Chemistry and Physics of Air Pollution*, John Wiley & Sons.
- M. WEST e J. HARRISON (1989), *Bayesian Forecasting and Dynamic Models*, Springer-Verlag, New York.

Riassunto

La concentrazione di sostanze chimiche presenti nell'atmosfera in una determinata regione geografica e per un fissato intervallo di tempo viene spesso analizzata e studiata al fine di prevedere ed evitare situazioni di inquinamento.

In questo articolo si identifica e si stima un modello statistico per la concentrazione di SO_2 utilizzabile quando le osservazioni disponibili sono raccolte da alcune centraline fisse e da alcune centraline mobili che si installano per periodi di tempo ridotti in diversi punti nella regione sotto studio (nell'esempio la città di Padova).

Il modello proposto coglie separatamente gli effetti di grande scala e di piccola scala, attraverso alcune componenti che vengono considerate come processi stocastici.

La stima contemporanea dei processi coinvolti avviene attraverso una formulazione *state space* del processo.

Si propongono anche un paio di applicazioni del modello nella previsione sia spaziale che temporale per il processo e nell'individuazione di siti nello spazio dove istallare in futuro le centraline mobili.

Summary

Given atmospheric measurement from a network of monitoring sites in the area of a city and over an extended period of time, an important problem is to identify the spatial and temporal structure of data.

In this paper we focus on the identification and estimate of a statistical model to analyse the SO_2 in the city of Padua, where data are collected by some fixed stations and some mobile stations moving without any specific rule in different new locations, staying in every location for a variable number of days.

The proposed method divides the global variability in large scale and small scale using some stochastic process as component of variability.

The estimate is provided using a *state space* formulation of the model.

As applications of the model we propose the spatial and temporal prevision of the concentration of SO_2. Finally, an exercise is proposed to choose an optimal network for the mobiles monitoring stations for a fixed future time.