

UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

TESI DI LAUREA

ANALISI SPAZIO TEMPORALE DEL BISSIDO DI ZOLFO
NELLA CITTA' DI PADOVA. UN APPROCCIO MEDIANTE
MODELLI STATE-SPACE

Relatore: Ch.mo Prof. Carlo Gaetan

Laureando: Antonio Fantinato

Anno accademico: 2007-2008

Indice

Introduzione	3
1. L'inquinamento	
1.1.L'inquinamento dell'aria	7
1.2.Effetti dell'inquinamento sull'atmosfera	10
1.3. Effetti dell'inquinamento sugli esseri viventi	11
1.4. I composti dello zolfo e del biossido di zolfo	13
1.5.La rilevazione degli inquinanti atmosferici	16
2. La convoluzione dal continuo al discreto	
2.1.Metodo della convoluzione	19
2.2.La convoluzione dal continuo al discreto	24
2.2.1.Costruzione di un modello spazio-temporale	25
2.3.Il modello spazio-temporale	27
2.4.Previsione, filtro e lisciamiento	29
2.5.Stima di massima verosimiglianza	38
3. Specificazione e risultati	
3.1.I dati: descrizione della base di dati	41
3.2. Analisi preliminare	46
3.3.Rappresentazione nello stato-spazio	51
3.4.Specificazione dei nuclei di convoluzione	53
3.5.Selezione del numero di nuclei	58
3.6.Risultati	61
3.7.Brevi considerazioni finali	74
Bibliografia	77
Ringraziamenti	81

Introduzione

Uno dei fenomeni di grande attualità in questi ultimi periodi è sicuramente l'inquinamento dell'atmosfera. Studiare metodi in grado di ricavare informazioni utili al fine di essere in grado di produrre stime su un'area specifica, date rilevazioni puntuali dei principali inquinanti diventa, oggi sempre più frequentemente, uno dei compiti più rilevanti dello statistico ambientale.

Tradizionalmente un approccio geostatistico considera solo la dimensione spaziale, ma ovviamente nell'analisi degli agenti inquinanti, l'aspetto temporale ha una sua rilevanza. Proprio per questo, in questo ultimo periodo si è notato un fiorire di contributi legati alla modellazione spazio-temporale degli inquinanti.

Obiettivo di questa tesi è lo studio dell'inquinamento da biossido di zolfo (anidride solforosa, SO_2) nel territorio relativo al comune di Padova e ad esso limitrofo, per il periodo che va dal 1 Aprile 1991 al 31 Marzo 1994.

Il biossido di zolfo è un gas dall'odore pungente. In natura viene immesso nell'atmosfera in seguito alle eruzioni vulcaniche, mentre le principali sorgenti antropiche sono costituite dagli impianti per il riscaldamento e la produzione di energia alimentati a gasolio, carbone e oli combustibili.

Sono vari i rischi legati ad un'esposizione da SO_2 . Esso provoca nell'uomo irritazione e lesione al tratto superiore dell'apparato respiratorio e aumenta la predisposizione ed episodi infettivi acuti e cronici come le tracheiti e le bronchiti.

I danni alla vegetazione ai materiali sono dovuti essenzialmente alle precipitazioni di questo inquinante tramite la formazione delle cosiddette "piogge acide".

Le uniche informazioni disponibili provengono dall'uso di apposite centraline di rilevazione sia mobili che fisse, predisposte dalla Provincia di Padova. Questo tipo di misurazione produce però una serie di dati mancanti per due ordini di motivi: per la diversa collocazione delle centraline durante differenti periodi e per il malfunzionamento delle stesse una volta posizionate.

Proprio per questo motivo la modellazione non risulta semplice in quanto si richiede che il modello sia in grado di:

- cogliere dinamiche temporali, ma che allo stesso tempo sia in grado di trattare i valori mancanti
- adattarsi con sufficiente flessibilità a cambiamenti nello spazio
- prevedere sia nello spazio che nel tempo.

Per questo si è ritenuto di creare una sorta di rete di nuclei di liscio bidimensionali dislocati opportunamente nell'area, combinandoli poi linearmente, tecnica che prende il nome di **convoluzione**. Questa combinazione aveva le necessarie caratteristiche di flessibilità nello spazio.

Tuttavia questa non poteva rimanere statica e qui si è scelto di supporre che i pesi della combinazione evolvessero nel tempo in maniera stocastica, secondo un processo di tipo *random walk*.

Si è quindi formalizzato un modello di tipo *state-space* e, tramite l'utilizzo del **filtro di Kalman** si è arrivato a ricostruire un processo latente alla base del modello state-space di cui prima, in grado di trattare i dati mancanti e di fare previsioni nel tempo e nello spazio.

Il modello sopra menzionato si è distinto per una buona capacità previsiva su tutto lo spazio di interesse, sia nella zone dove sono state rilevate molte osservazioni (tipicamente vicino alle centraline fisse), sia nella periferia, e si è rilevato capace di cogliere fluttuazioni temporali del fenomeno di interesse.

La struttura della tesi è la seguente:

Nel primo capitolo vi saranno delucidazioni sull'inquinamento, si parlerà quindi di cosa si intende per inquinamento, quali sono i suoi effetti su cose, animali e persone e in che direzione sono tese le politiche per contingentare il fenomeno.

Nel secondo capitolo vi sarà una trattazione teorica sul metodo della convoluzione, dei modelli stato spazio e del filtro di Kalman.

Nel terzo, dopo una breve analisi preliminare dei dati, vi saranno illustrazioni dei risultati ottenuti e considerazioni sulla parametrizzazione del modello, oltre a trarre delle conclusioni finali sui risultati ottenuti.

Capitolo 1

L'inquinamento

1.1 L'inquinamento dell'aria

Il lavoro di questa tesi, sarà quello di specificare un modello per l'interpretazione e la previsione di un fenomeno di largo interesse che ha attratto, e che continua a calamitare, l'attenzione dell'opinione pubblica negli ultimi anni e che ha, in certi casi condizionato in maniera diretta, o indiretta, il rapporto del cittadino con l'ambiente (in particolare quello urbano) in cui vive; quello cioè, dell'inquinamento atmosferico.

L'inquinamento dell'aria è definito come una condizione atmosferica dove determinate "sostanze" sono presenti in misura superiore rispetto alle condizioni normali e tale da causare effetti sull'uomo, animali, vegetazione o materiali.

Per sostanze si intende un qualsiasi elemento o composto, sia esso di derivazione chimica o naturale, presente nell'aria in forma di gassosa, liquida o di particelle solide.

L'inquinamento è dato dalle fonti di emissione che, come detto, modificano la composizione dell'atmosfera, e sono proprio le modificazioni ritenute rilevanti ad essere segnalate dai recettori.

Il valore che rileva la concentrazione di sostanze chimiche nell'atmosfera è uno dei principali indicatori per la segnalazione di situazioni di inquinamento, tuttavia risulta molto difficile trovare una relazione diretta tra la presenza di sostanze inquinanti e i danni da esse causate; motivo per cui l'azione di politiche pubbliche atte a limitare il fenomeno si rivelano spesso poco considerate e inefficaci.

Nel linguaggio comune, tutte le sostanze che possono rivelarsi dannose per la salubrità dell'ambiente e dell'uomo in particolare, si identificano con il nome di *polveri sottili*. Queste stesse sostanze vengono raggruppate con un'identificazione più scientifica con la sigla PM₁₀, sigla che comprende buona parte delle sostanze inquinanti, e che identifica il materiale presente nell'atmosfera in forma di particelle microscopiche di diametro non superiore ai 10 $\mu\text{m}/\text{m}^3$, vale a dire 10 millesimi di millimetro per metro cubo.

È costituito da polvere, fumo e microgocce di sostanze liquide.

Le fonti principali di PM₁₀, non sono solo, come verrebbe spontaneamente da pensare, quelle dovute all'attività dell'uomo, come lo smaltimento dei rifiuti e i processi di combustione (in particolare dei motori a scoppio, dei riscaldamenti, in molte attività industriali, e nelle centrali termoelettriche), ma vi sono anche cause naturali come quelle derivanti dai fumi rilasciati dagli incendi boschivi, dalle eruzioni vulcaniche, o persino dalla dispersione dei pollini, solo per citarne alcuni.

In effetti, solo una parte del PM₁₀ presente nell'atmosfera deriva dalla trasformazione in particelle liquide di alcuni gas (in particolare composti dell'azoto e dello zolfo) emesse dalle attività umane, si stima attorno al 29%, ciò nonostante non si può negare che questo non rappresenti un problema.

È comunque vero che, la parte dell'inquinamento indotta dall'uomo è l'unica parte elidibile o per lo meno riducibile del fenomeno, e attuare politiche di riduzione e miglioramento delle fonti di emissione è l'unica azione attuabile in questo senso.

Vi sono numerose tipologie di sostanze inquinanti e vi sono molti modi per classificarle, ad esempio a seconda della composizione chimica si possono specificare:

- Composti che contengono Zolfo
- Composti che contengono Azoto
- Composti che contengono Carbonio
- Composti che contengono Alogeni
- Sostanze tossiche
- Composti radioattivi.

Un'altra classificazione riguarda il modo in cui le polveri sottili raggiungono l'atmosfera, abbiamo quindi:

- Inquinanti primari: cioè emessi direttamente dalle fonti di inquinamento
- Inquinanti secondari: che si formano nell'atmosfera mediante reazione chimica tra inquinanti primari, oppure tra gli stessi e i componenti naturali dell'atmosfera.

La prima tipologia è quella in cui rientra, quello che si ritiene sia uno dei principali inquinanti, vale a dire il biossido di zolfo (SO_2), che è accompagnato molto spesso da un suo derivato, il particolato di zolfo, che rientra nella seconda categoria appena sopra elencata.

L'atmosfera è costituita per buona parte da azoto, ossigeno e altri gas nobili con concentrazioni che sono rimaste pressoché invariate nel tempo, ma vi sono altri gas che, se pur in piccole quantità, presenti in quantità variabile nella troposfera come vapore acqueo, biossido di carbonio e ozono.

In realtà, i gas prodotti dall'attività biologica, vulcanica, dai processi di combustione dovuti all'azione dell'uomo, dal decadimento radioattivo vengono eliminati dall'atmosfera attraverso reazioni chimiche, attività biologiche, assorbimento e adsorbimento dagli oceani e dalla terra; il lasso temporale che caratterizza questi cosiddetti "cicli gassosi" viene chiamato tempo di residenza, perciò le problematiche legate all'inquinamento non sarebbero tali, se non fosse che, per alcune sostanze inquinanti questo ciclo

dura anche qualche milione di anni.

L'inquinante principale, poiché uno dei maggiori agenti scatenanti dell'inquinamento atmosferico, l' SO_2 , di cui si parlerà nel prosieguo della tesi, presenta un ciclo di residenza relativamente basso, tuttavia la pericolosità di questa sostanza non risiede nel tempo di eliminazione, ma nella quantità dello stesso che quotidianamente viene riversata nell'ambiente.

1.2 Effetti dell'inquinamento sull'atmosfera

Il fatto che l'inquinamento produca effetti dannosi su persone, animali, vegetazione e cose è ormai sotto gli occhi di tutti, ma non è altrettanto vero che gli effetti siano tutti facilmente misurabili. La salute degli esseri viventi, in particolare, non risulta affatto banale da misurare, soprattutto perché, ad inficiare sulle condizioni di vita possono essere molte concause a loro volta difficilmente stimabili o stimabili con grossa approssimazione.

I principali effetti che sortiscono gli inquinanti atmosferici sulla troposfera sono:

- Riduzione della visibilità
- Formazione di nebbie e precipitazioni
- Riduzione della radiazione solare
- Alterazione della distribuzione della temperatura e dei venti.

La riduzione della visibilità, viene definita come la maggiore distanza a cui è possibile vedere un oggetto scuro e prominente alla luce diurna oppure una fonte di luce sfocata di notte. La riduzione di visibilità è causata dalla deviazione di luce a causa delle sostanze presenti nell'aria che disperdono la luce solare o artificiale nel secondo caso.

Una particolare diffusione dell' SO_2 , inoltre agevola la formazione di nebbie, dove le goccioline di acido solforico generate dall'ossidazione dell' SO_2 , si comportano come nuclei di condensazione che agevolano la formazione di piccole gocce di nebbia.

Anche la diffusione e l'assorbimento di radiazioni solari e infrarossi subiscono una riduzione quanto l'ambiente risulta inquinato. Alte concentrazioni di agenti inquinanti riducono la radiazione solare diretta specialmente quando il sole è basso rispetto la linea dell'orizzonte, quando la luce deve compiere un percorso maggiore per attraversare la troposfera.

1.3 Effetti dell'inquinamento sugli esseri viventi

L'inquinamento riesce ad agire oltre che sulla salute degli organismi viventi anche sui materiali tramite una reazione chimica che porta al deterioramento. Alti livelli di particolato sono in grado di intaccare vestiti e strutture, particelle acide e alcaline contenute nei solfuri sono in grado di corrodere anche vernici, murature, materiale elettrico; l'ozono è capace di deteriorare facilmente la gomma.

Più rilevante per l'alimentazione umana è capire come agiscono sostanze inquinanti come l' SO_2 , il nitrato di perossiacetile e l'etilene che sono in grado di distruggere la clorofilla presente nelle foglie dei vegetali e interrompere il processo di fotosintesi, causando danni ingenti alle colture.

Ultimo, ma non certo in ordine di importanza, sono gli effetti che comportano sostanze inquinanti sull'uomo. Esposizione sistematiche e di tempo prolungato causano gravi danni all'apparato respiratorio in particolar modo a fasce sociali più esposte a problemi di questo carattere, come gli anziani e coloro che soffrono di malattie respiratorie.

In particolare, le esposizioni prolungate ad agenti inquinanti in ambiente urbano, soprattutto in assenza di vento, comportano una stabilizzazione di quelle che viene chiamata “isola di calore”, fenomeno metropolitano causato dal calore sprigionato degli edifici cittadini, che viene a collidere con l’ambiente della campagna tipicamente più fresco, questo crea una calotta di calore, piena di inquinanti, che opprime chi vi abita.

Nonostante quanto detto, è tutto sommato ancora difficile tracciare una dipendenza tra l’inquinamento e le condizioni di salute del cittadino, anche perché un’analisi di questo tipo non può prescindere dagli stili di vita, dai comportamenti personali e culturali del soggetto.

Bisogna anche saper distinguere quali sono le reazioni croniche (o congenite) e le reazioni acute. Tipiche affezioni possono essere le bronchiti e l’enfisema polmonare (di carattere tipicamente cronico).

La nocività delle polveri sottili dipende oltre che dalla loro natura chimica, anche dalle loro dimensioni e dalla loro capacità di raggiungere diverse parti dell’apparato respiratorio:

- oltre i $7 \mu\text{g}/\text{m}^3$: cavità orale e nasale
- fino a $7 \mu\text{g}/\text{m}^3$: laringe
- fino a $4,7 \mu\text{g}/\text{m}^3$: trachea e bronchi primari
- fino a $3,3 \mu\text{g}/\text{m}^3$: bronchi secondari
- fino a $2,1 \mu\text{g}/\text{m}^3$: bronchi terminali
- fino a $1,1 \mu\text{g}/\text{m}^3$: alveoli polmonari.

Anche la mortalità dovuta all’inquinamento da polveri sottili è oggetto di studio, e le prime stime stillate dall’OMS, su base di uno studio¹ condotto su 8 città nel mondo, stima che già lo 0,5% dei decessi registrati nell’anno sono causati da inquinamento da polveri sottili.

¹ Fonte: Electric Power Research Institute; riportato da Seinfeld (1986)

1.4 I composti dello zolfo e il biossido di zolfo

Lo zolfo, uno dei principali inquinanti conosciuti ha un ciclo globale che coinvolge principalmente H_2S , SO_2 , SO_3 e altri solfati. In effetti ogni anno entrano nell'atmosfera all'incirca 100 milioni di tonnellate di zolfo emesso in forma di SO_2 . Ciò che ancora non è molto chiaro agli studiosi è come rientri nel ciclo dello zolfo l' H_2S , emesso principalmente da fonti vulcaniche o decadimento biologico di sostanze organiche, che viene comunque convertito in SO_2 .

È comunque noto che, la fonte primaria di biossido di zolfo generato dall'uomo è legata al carbone, da cui lo stesso biossido dovrebbe essere rimosso per limitare le situazioni inquinanti. In realtà circa metà del biossido presente nel carbone è sotto forma di pirite, FeS_2 , e l'altra metà come zolfo organico. In sostanza comunque, quasi tutto lo zolfo viene convertito in SO_2 tramite i processi di combustione, e solo una piccola parte, circa l'1-2% viene tramutato in SO_3 , cioè anidride solforosa.

I due composti precedentemente elencati sono i principali imputati dell'inquinamento atmosferico da ossidi di zolfo e sono caratterizzati da: assenza di calore, odore pungente, alta reattività dell' SO_3 , mancata combustione dell' SO_2 .

Lo zolfo che è presente nell'atmosfera è proveniente per circa due terzi da fonti naturali e, solo per un terzo è dato dalle attività dell'uomo.

Il problema dell'inquinamento globale in effetti, non è tanto dovuto all'eccesso di immissioni di inquinante da parte dell'uomo, ma più che altro ad una sua pessima distribuzione concentrandosi in zone urbane che spesso sono limitrofe, se non coincidenti con le zone caratterizzate da alta attività industriale.

Tabella 1.1: Cause dell'inquinamento atmosferico.

Sorgente di emissione	10 ml tonnellate/anno	%
Trasporti	0,7	2,3
Impianti termoelettrici	19,1	61,8
Combustione in altri impianti	5,2	16,8
Processi industriali	5,7	18,5
Rifiuti solidi	0,1	0,3
Altri	0,1	0,3
Totale	30,9	100

Come possiamo osservare dalla tabella, i principali indiziati per la produzione di inquinanti sono gli impianti di combustione fissi.

Per quanto le case automobilistiche cerchino di produrre veicoli sempre meno inquinanti e per quanto la pubblica amministrazione a tutti i livelli cerchi di limitare il traffico, questo può rivelarsi un problema solo nei grossi centri abitati lontani da attività industriali.

Sono quindi gli impianti termoelettrici i principali indiziati dell'inquinamento atmosferico, seguiti dagli impianti industriali fra cui le fonderie, le raffinerie di petrolio, gli impianti di acido solforico e gli impianti di conversione del carbone fossile.

Gli effetti degli ossidi di zolfo possono essere di assoluta rilevanza quando ad una concentrazione elevata di questo inquinante, si accompagnano fenomeni di inversione termica che determinano uno stazionamento per un periodo prolungato di queste sostanze, su un'area limitata e densamente abitata. Casi di rilevante entità si sono registrate In Pennsylvania nel 1948 e nel 1952 a Londra dove si sono avuti aumenti di decessi e ricoveri per malattie respiratorie che possono essere spiegati solo dalla coincidenza appena spiegata di fenomeni di inquinamento e atmosferici.

In particolare, il biossido di zolfo, data la sua alta solubilità viene subito

assorbito nei primi passaggi del sistema respiratorio, e in soggetti affetti da malattie respiratorie riscontrano broncocostrizione e secrezione di muco anche a concentrazione non particolarmente elevate.

Altri inquinanti come il CO, l'NO₂ e l'O₃ sono relativamente insolubili e penetrano più in profondità nei polmoni. Il biossido di azoto e l'ozono sono la causa di edema polmonare, mentre il monossido di carbonio è in grado di interferire con il trasporto di ossigeno da parte dell'emoglobina. Bisogna anche dire che, lo stesso sintomo può essere figlio di cause scatenanti diverse. Parecchi inquinanti sono presenti nello stesso tempo e i loro effetti sono attribuibili all'azione combinata di più sostanze.

L'inquinante SO₂ sortisce effetti anche sui materiali, in particolare su vernici, metalli e materiali da costruzione.

In presenza di concentrazioni di SO₂ le vernici si essiccano e si induriscono in tempi meno rapidi, per cui la pellicola lasciata dalle vernici rischia di diventare più molle o più fragile, il che influisce sulla loro durata.

I metalli come il ferro, l'acciaio e lo zinco vengono corrosi più rapidamente se in ambienti saturi di biossido di zolfo e questo in particolare nei mesi freddi caratterizzati dall'inquinamento dovuto all'SO₃, emesso a causa del riscaldamento domestico.

I danni più evidenti vengono comunque apportati ai materiali da costruzione, che vengono intaccati dall'acido solforico prodotto dalla reazione dell'anidride solforosa con il vapore acqueo presente nell'atmosfera.

Come per l'uomo, anche per la vegetazione, esposizioni prolungate di SO₂ causano danni alla vegetazione e alle colture. Esposizioni brevi causano danni acuti, dove le foglie deperiscono e cambiano colore poiché assorbendo SO₂, riescono a trasformarlo in un solfato, l'H₂SO₄, che si deposita ai margini delle foglie stesse, che aggiunto a quello prodotto naturalmente dalla pianta ne determina la caduta.

Esposizioni prolungate provocano invece danni cronici, come la clorosi, vale

a dire un graduale ingiallimento delle foglie dovuto al mancata formazione di clorofilla. Le piante presentano anche il caratteristico color giallo anche quando i vapori di acido solforico (derivato dal biossido di zolfo) si posa sulle stesse foglie bagnate dalla rugiada o dalla bruma; in effetti le piante ricevono inquinamento proprio quanto i loro stomi devono intercambiare con l'ambiente i gas (prevalentemente nelle ore diurne) presenti nell'atmosfera.

D'altro canto anche lo zolfo (e quindi il biossido di zolfo) è un elemento indispensabile per la crescita delle piante ed in ambiente moderatamente inquinato le piante contribuiscono a purificare l'atmosfera.

1.5 La rilevazione degli inquinanti atmosferici

Le città italiane sono tenute a rilevare le concentrazioni di sostanze pericolose presenti nell'aria del tessuto urbano di riferimento per tutelare l'ambiente ed in particolar modo la salute dei cittadini.

La normativa di riferimento (D.P.R. 203/88) stabilisce per inquinamento atmosferico “ogni modificazione della normale composizione o stato fisico dell'aria atmosferica, dovuta alla presenza nella stessa di una o più sostanze in quantità e con caratteristiche tali da alterare le normali condizioni ambientali e di salubrità dell'aria; da costituire pericolo ovvero pregiudizio diretto o indiretto per la salute dell'uomo; da compromettere le attività ricreative e gli altri usi legittimi dell'ambiente; alterare le risorse biologiche e gli ecosistemi ed i beni materiali pubblici e privati”.

È dunque, inteso che questa attività deve essere tesa non solo alla semplice realizzazione di tecniche in grado di rappresentare il fenomeno, ma anche all'individuazione delle principali fonti inquinanti e alla previsione in senso spaziale e temporale delle concentrazioni inquinanti, in modo da poter essere

in grado di prendere provvedimenti in merito.

I valori limite sono definiti in Italia dal D.L. n. 60 del 2 aprile 2002; tale decreto fissa due limiti accettabili di PM_{10} in atmosfera:

- Il primo è un valore limite di $50 \mu g/m^3$ come valore medio misurato nell'arco di 24 ore da non superare più di 35 volte/anno.
- Il secondo come valore limite di $40 \mu g/m^3$ come media annuale.

È comunque noto che, la fonte primaria di biossido di zolfo generato dall'uomo è legata al carbone, da cui lo stesso biossido dovrebbe essere rimosso per limitare le situazioni inquinanti. In realtà, come detto precedentemente, circa metà del biossido presente nel carbone è sotto forma di pirite, FeS_2 , e l'altra metà come zolfo organico. In sostanza comunque, quasi tutto lo zolfo viene convertito in SO_2 tramite i processi di combustione.

Causa principale della presenza di polveri sottili nell'atmosfera oltre agli scarichi relativi al trasporto urbano e nell'immediata periferia, sono anche i residui gassosi delle installazioni industriali, che però almeno per quanto concerne il comune di Padova, sembra non siano così rilevanti come lo sono i gas di scarico relativi al trasporto soprattutto privato, e in particolar modo i residui dati dal riscaldamento domestico.

Per la rilevazione degli inquinanti è previsto l'uso, nell'ambito del territorio di riferimento, di speciali centraline di rilevazione, che si possono distinguere in quattro categorie.

Le stazioni di tipo **A** sono stazioni di base e riferimento in aree non direttamente interessate all'inquinamento in grado di misurare:

•Inquinanti primari e precursori degli inquinanti secondari:

-In fase gassosa: concentrazioni di CO , SO_2 , NO_2 , idrocarburi e volatili

-In fase di particelle: concentrazione in massa di particolato spesso e di piombo e altri metalli pesanti nel particolato stesso

•Inquinanti secondari:

-In fase gassosa: concentrazioni di O₂, O₃, NO

-In fase di particelle: concentrazioni di prodotti di trasformazione degli ossidi di azoto e dell'anidride solforosa

Le stazioni di tipo **B**, sono da collocare in zone ad alta intensità abitativa, in gradi di misurare concentrazioni di NO₂, SO₂, idrocarburi.

Le stazioni di tipo **C** sono da collocare in zone ad elevato traffico e che sono appunto in grado di misurare gli inquinanti prodotti dalla combustione autoveicolare: CO, idrocarburi volatili.

Le stazioni di tipo **D** si collocano in zone periferiche e sono finalizzate alla rilevazione dell'inquinamento fotochimico come l'NO₂, ozono, periossiacetilnitrato.

A chi usualmente lavora nel campo della geostatistica, viene chiesto, in particolar modo da parte dall'amministrazione pubblica, di stilare stime e previsioni sulla concentrazione degli agenti inquinanti nei diversi luoghi del tessuto urbano.

L'esigenza di avere stime precise e raffinate si scontra molto spesso, e questo caso non esula certo dalla norma, con il fatto che, chi è tenuto a dare queste informazioni ha a disposizione dei fondi limitati.

Oltre a questo è anche opportuno dire che le centraline usate per la rilevazione sono molto costose e che un comune delle dimensioni di quello di Padova, difficilmente si può permettere più di 4-5 centraline di rilevazione.

Capitolo 2

I modelli state-space

2.1 Metodo della convoluzione

È possibile costruire modelli che consentano di studiare un fenomeno di tipo geostatistico nello spazio e nel tempo, analizzandolo sia tramite la dinamica spaziale sia quella temporale.

Nella fattispecie infatti, abbiamo dati che, come specificato in precedenza, si riferiscono a rilevazioni delle polveri sottili riguardanti il territorio relativo al Comune di Padova, nell'arco evolutivo di 3 anni. Risulta pertanto interessante, non solo scattare una fotografia del fenomeno in un preciso momento, ma è oltremodo utile, capire l'evoluzione del fenomeno nell'intero lasso temporale di cui disponiamo di rilevazioni e questo, anche e soprattutto, per poter stilare proiezioni sul fenomeno in atto e prendere eventualmente misure atte a contingentare o ridurre questa problematica di carattere tipicamente ambientale.

Il modello che verrà utilizzato, e che farà da presupposto per le successive elaborazioni si basa sull'idea della *convoluzione*, e quindi sull'uso di un numero di nuclei o *kernel* di lisciamiento che formeranno una rete di punti, tramite i quali si arriverà a studiare, in due dimensioni, in che misura si manifesta l'inquinamento da polveri sottili nell'intero territorio comunale.

In un contesto di regressione non parametrica, si utilizza un metodo di lisciamiento basato sul nucleo quando si desidera trovare, nella fattispecie un piano, che sia in grado di interpolare le osservazioni sparse in R^2 .

La tecnica consiste nel pesare le osservazioni, solitamente in misura

proporzionale alla distanza (qui in termini euclidei) dal punto in cui si vuole stimare tale funzione. L'ampiezza dell'influenza esercitata da tale nucleo si chiama larghezza di banda.

Questo approccio è stato proposto in letteratura (Ickstadt e Wolpert, 1999), (Haas, 1995), in antitesi al tradizionale covariogramma, poiché comporta una semplificazione dal punto di vista dell'elaborazione dei dati ed una facile estendibilità attraverso semplici modelli stazionari.

Modellare dati spaziali con processi Gaussiani è in realtà il filo conduttore di tutte le analisi geostatistiche.

I processi gaussiani sono completamente caratterizzati dai suoi primi due momenti, e per questi, stazionarietà forte e debole coincidono.

Un approccio molto comune, sarebbe stato quello di modellare la dipendenza spaziale attraverso un covariogramma, cosicché la covarianza tra due punti qualsiasi di rilevazione dipende solo dalla distanza tra di essi.

Il covariogramma o funzione di covarianza stazionaria, è la rappresentazione della relazione che intercorre tra un variabile $Z(s)$ e $Z(s+h)$, dove h è la distanza (euclidea) che intercorre tra le due variabili.

Il metodo alternativo di cui parlerò in seguito, implica dunque, la costituzione di una griglia di kernel dislocati sul territorio, che potremo, per semplicità considerare di tipo Gaussiano. In realtà, se trattiamo un griglia con kernel di tipo Gaussiano, i due metodi messi a confronto, non evidenziano particolari differenze, tuttavia il metodo della convoluzione risulta essere molto più flessibile specialmente quando si ha a che fare con condizioni “non standard” come, uno su tutti, la non stazionarietà. Inoltre, il suddetto metodo risulta essere più duttile con modelli spazio-temporali alternativi.

È possibile quindi costruire un processo Gaussiano, andando ad ottenere il valore $Z(s)$ su una regione, “convolvendo” un processo white noise (ossia una successione di variabili casuali indipendenti, identicamente distribuite, di

media nulla e varianze costante), $Z(s)$ con un kernel di lisciamiento $k(s)$ (cioè una funzione di densità che pesa le osservazioni vicine ad essa a seconda della sua forma e di un fattore di scala che è chiamato ampiezza di banda)

$$Z(s) = \int_S k(u-s) X(u) du \quad \text{per } s \in S \quad (2.1.1)$$

La covarianza di $Z(s)$ rispetto ad un altro $Z(s')$ risulta pertanto essere dipendente dal solo vettore $d = s-s'$ è pertanto data da

$$Cov(Z(s), Z(s')) = \int_S k(u-s) k(u-s') du = \int_S k(u-d) k(u) du = c(d)$$

Se $k(s)$ è un nucleo di tipo isotropico (dove l'influenza della funzione di densità centrata nel nucleo è proporzionale solo alla distanza che intercorre tra del nucleo e l'osservazione stessa) per cui $k(s) = g(\|s\|)$, anche $Z(s)$, il processo stocastico lo è, questo perché, l'influenza (data la definizione di isotropia) che esercita il nucleo di lisciamiento sulle rilevazioni dipende solo dalla distanza dal nucleo e non dalla posizione rispetto quest'ultimo. Proprio per questo motivo si ottiene l'ultima uguaglianza ed è anche per questo, che un nucleo con queste caratteristiche può definire un cosiddetto processo isotropico.

Con questo tipo di approccio il modello è determinato non solo dal processo latente $X(s)$ ma anche dal nucleo di lisciamiento $k(s)$.

Portata a compimento la scelta di questi due ingredienti, la modellazione presenta differenti caratteristiche:

•modelli a covarianza non-parametrica

Tramite la rappresentazione della media mobile, il nucleo $k(s)$ si presta ad una miglior flessibilità rispetto al covariogramma $c(s)$, che deve essere definito positivamente. È possibile infatti, specificare modelli flessibili tramite l'utilizzo di un nucleo per costruire covarianze non standard per processi Gaussiani, anche se vi sono approcci alternativi che specificano classi di covarianze definite positivamente (Gelfand ed Ecker, 1997).

•non normalità di $x(s)$

Il modello può essere anche specificato gestendo la variabile latente $X(s)$, assumendo così la non normalità del processo, in questo modo il processo $Z(s)$ non potrà essere normale, e conseguentemente ciò produrrà un tipo di stazionarietà che non potrà essere definita di tipo Gaussiano. Si veda ad esempio Wolpert e Ickstadt (1999) che specificarono un processo di tipo Lèvy.

•restrizione del dominio di $X(s)$

Invece di specificare il processo latente come un *white noise* in ambito continuo, è possibile restringere il dominio di $X(s)$ per eliminare le dipendenze che si vengono a creare in un modello, per esempio ai margini dell'area presa in considerazione. È inoltre possibile modificare anche la forma del kernel nella zone prossime ai confini di interesse.

•riduzione della dimensione

Può essere d'aiuto ridurre i siti dove vengono centrati i kernel in modo da favorire una minor parametrizzazione del modello e conseguentemente da diminuire il tempo di elaborazione dei dati.

•covarianza spaziale non stazionaria

Il kernel di lisciamiento può essere fatto variare a seconda della sua collocazione spaziale

$$Z(s) = \int_S k_s(u) X(u) du \quad \text{per } s \in S$$

dove $k_s(u)$ stabilisce che la forma del nucleo dipende da s . Risulta quindi possibile creare una dipendenza tra s e la struttura del kernel.

In tal caso si introducono delle non stazionarietà.

•modelli spazio-temporali

Può essere interessante studiare un fenomeno non solo in relazione alla variabile spaziale, ma anche all'interno di una dinamica temporale, cosicché il processo latente possa essere definito con entrambe S e T . Quest'ultimo risulterà quindi essere definito tramite $X(s,t)$ e i corrispettivi nuclei tramite $k(s,t)$ che evolvono nel tempo.

La sua formulazione risulterà pertanto essere:

$$Z(s,t) = \int_S k(u-s,t) X(u,t) du \quad \text{per } s \in S, t \in T$$

In realtà il modello che verrà proposto in seguito non terrà conto di una dinamica temporale legata ai nuclei di lisciamiento, che saranno di tipo $k(s)$, ma si terrà conto che il processo latente $X(s,t)$, dovrà comunque evolvere oltre che nello spazio anche nel tempo.

•processi spaziali dipendenti

Il processo di convoluzione dà anche la possibilità di costruire processi

spaziali dipendenti (Ver Hoef e Barry, 1998) e (Ver Hoef, Cressie e Barry, 2000). È possibile quindi costruire un processo latente $Z_j(s)$ che condivide parte del dominio (in senso spaziale) di un altro processo latente:

$$Z_1(s) = \int_{S_0 \cup S_1} k(u-s) X(u) du$$

$$Z_2(s) = \int_{S_0 \cup S_2} k(u-s) X(u) du$$

Nella fattispecie S è l'unione di domini disgiunti S_0 , S_1 ed S_2 , vale a dire un'approssimazione discreta dello stesso spazio S .

2.2 La convoluzione dal continuo al discreto

Nella realtà si è usata un'approssimazione discreta al modello in 2.1.1. che si è proposto in precedenza.

I nuclei di lisciamiento, che ora indicheremo η_1, \dots, η_m , sono stati collocati su una griglia che si estende sull'area sotto esame, tramite criteri che verranno esposti in seguito.

Inoltre, su questi nuclei di lisciamiento si sviluppa un processo di tipo *white noise* $X(s)$.

È necessario identificare il processo latente $X(s)$ ad essere diverso da zero nei siti della griglia sopramenzionati η_1, \dots, η_m oltre che in S .

Avremmo quindi

$$Z(s) = \sum_{j=1}^m X_j k(s-\eta_j)$$

dove

$$X \sim N(0, \sigma_x^2 I_m)$$

Qui $X_j = X(\eta_j)$ e $k(\cdot - \eta_j)$ è il nucleo centrato in η_j . Dobbiamo anche specificare che il nucleo $k(\cdot)$ è un nucleo “radialmente” simmetrico, ovvero che esercita la sua influenza solo proporzionalmente alla distanza che intercorre tra il centro del nucleo di lisciamiento e l’osservazione, influenza che sarà inversamente proporzionale alla distanza.

Supponendo ora di osservare n osservazioni $y(s_1), \dots, y(s_n)$

$$Y(s) = \mu + Z(s) + \varepsilon$$

dove μ rappresenta il livello medio del fenomeno ed ε rappresenta il vettore degli errori che si distribuisce

$$\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n)$$

Il modello così specificato appare come un modello lineare ad effetti casuali.

2.2.1 Costruzione di un modello spazio-temporale

In questa sezione si mostrerà come costruire un modello spazio-temporale sfruttando il metodo della convoluzione.

Il principio è infatti quello costruire un modello semplice e lisciarlo con una quantità e una tipologia di kernel da definire, in base ai vincoli indotti dal fenomeno e dal modello.

Come sottolineato precedentemente i vantaggi di questo approccio sono molteplici (estensione alla non stazionarietà, ai modelli non strettamente di tipo Gaussiano, alla non separabilità delle dinamiche spaziali e temporali etc...) senza contare che il carico computazionale non può certo dirsi gravoso almeno in relazione ad altri metodi più tradizionali.

In sostanza il processo di convoluzione si può riassumere nell'assegnazione di variabili in uno spazio s e un arco temporale t e il lisciamiento di queste stesse variabili tramite dei kernel definiti nello spazio.

Tenendo conto di un'evoluzione temporale del fenomeno in analisi, è anche possibile dislocare i kernel di lisciamiento evolvendo nel tempo, spostando così, la griglia formata dai nuclei.

Posto $X_{jt}=X(\eta_j, t)$, ogni sequenza $\{X_{jt}\}$ con $t=1, \dots, 1096$ (giorni di rilevazione) è supposta essere un processo temporale gaussiano di tipo *random walk*.

Supponiamo ora che

$$Z(s, t) = \int_s k_s(u-s) X(u, t) + \varepsilon(s, t) \, du \quad (2.2.1)$$

e dandone una rappresentazione discreta

$$Z(s,t) = \sum_j k(s-\eta_j)X_{jt} + \varepsilon(s,t) \quad (2.2.2)$$

dove $\varepsilon(s,t)$ è il vettore degli errori.

Considerando sia la (2.2.1) che la (2.2.2) si ottiene un modello utile per descrivere la dinamica spazio-temporale.

Il modello può anche essere espresso in termini matriciali tramite queste due equazioni evolutive.

$$Y_t = K_t X_t + \varepsilon_t$$

$$X_t = X_{t-1} + v_t$$

dove l'elemento K_t , matrice di dimensione $k_t * m$ (dove k_t è il numero dei siti di rilevazione al tempo t ed m è il numero dei nuclei di lisciamiento), è meglio specificato dall'equazione

$$K_{ij}^t = k(s_{it} - \eta_j)$$

dove $y_t = (y_{1t}, \dots, y_{nt})$ sono le rilevazioni nei siti s_{1t}, \dots, s_{nt} ed inoltre

$$\varepsilon_t \sim N(0, \sigma_\varepsilon^2), \quad t=1, \dots, n_t$$

$$v_t \sim N(0, \sigma_v^2) \quad t=1, \dots, m$$

2.3 Il modello state-space

Il modello precedente è un caso particolare di una specificazione particolare nello spazio degli stati di un modello state-space.

La formulazione nella sua forma più generale è data da un vettore di valori osservati di dimensione ad esempio k del tipo $y_t=(y_{t1},y_{t2},\dots,y_{tk})'$ e rilevati per ogni $t=1,2,\dots,n$.

Il vettore y_t è funzione lineare di un vettore $x_t=(x_{1t},x_{2t},\dots,x_{mt})'$, detto vettore latente, ed eventualmente di un vettore di r covariate $z_t=(z_{t1},z_{t2},\dots,z_{tr})'$.

La relazione è specificata attraverso l'equazione di osservazione scritta nella forma più generale.

$$Y_t=B_tZ_t+G_tX_t+\omega_t \quad (2.3.1)$$

dove B_t è una matrice $q*r$ e la matrice G_t è la matrice di dimensione $k*m$, che descrive la relazione tra il vettore di stato non osservato X_t e le osservazioni Y_t .

Il vettore degli errori ω_t è assunto essere indipendente da X_t , e dai suoi valori passati, gaussiano con media zero e matrice di covarianza R_t .

Il vettore latente ha un sua dinamica specificata dall'equazione di stato

$$X_t=\Phi_tX_{t-1}+v_t \quad \text{per ogni } t=1,2,\dots,n \quad (2.3.2)$$

Anche qui supporremo che, per l'inizializzazione del vettore x si assume x_0

distribuito normalmente con media 0 e matrice di covarianza degli errori Ω_0 .

La transizione tra $t-1$ e t è regolata dalla matrice di transizione $p \times p$ Φ e dalla componente di errore additivo v_t , assunto indipendente da ω_t con media 0 e matrice di covarianza Q_t .

La specificazione *random walk* non è unica ma si possono creare anche altre specificazioni.

Le componenti non osservate nel vettore x_t devono essere anch'esse stimate.

Siamo anche interessati a prevedere il vettore x_t e la sua varianza; ciò è possibile con il Kalman *filter* ed il Kalman *smoother*, come vedremo in seguito.

2.4 Previsione, filtro e lisciamiento

Il modello *state-space* definito in precedenza presenta un vettore di valori non osservati x_t .

Il filtro di kalman (Shumway e Stoffer, 2000; Brockwell Davis, 1991), è l'algoritmo che permette la valutazione ricorsiva dei momenti della distribuzione di x_t , dato il vettore dei valori osservati y_1, y_2, \dots, y_s .

Si denoti

$$x_{t/s} = \mathbf{E} [x_t | y_1, y_2, \dots, y_s] \tag{2.4.1}$$

- se s è uguale a t si compie un aggiornamento dei dati,
- se s è uguale a $t-1$ si compie una previsione ed infine
- se s è uguale ad n si crea un lisciamiento dei dati

Analogamente di denoti la matrice di covarianza

$$\Omega_{t/s} = E [(x_t - x_{t/s})(x_t - x_{t/s})' | y_1, y_2, \dots, y_{t-1}] \quad (2.4.2)$$

Se s , a scopo esemplificativo, fosse uguale ad n , questo risulterà essere il lisciamento di Kalman, se invece $s = t-1$ definisce il **previsore ad un passo di Kalman**, se ancora $s = t$ si parlerebbe di **filtro di Kalman**.

Il problema principale nel calcolo dei momenti delle variabili casuali condizionate è il carico di elebarazione dati, infatti nella formula multivariata condizionale, il carico computazionale sarebbe stato molto elevato considerando la necessità di invertire le matrici in questione.

Questo rendeva inutilizzabile il tradizionale approccio, quando le capacità di memorizzazione e velocità di calcolo degli elaboratori intorno agli anni '60, non erano elevate come quelle odierne.

Il primo algoritmo di Kalman infatti, che fu elaborato proprio in questo ambito, implicava una ricorsione che elaborava le previsioni condizionali, (Kalman, 1960; Kalman e Bucy 1961) ed operava su matrici di dimensioni ben più ridotte.

In particolare il filtro di Kalman elabora ricorsivamente nell'ordine sotto indicato

$$x_{1/0} \rightarrow x_{1/1} \rightarrow x_{2/1} \rightarrow x_{2/2} \rightarrow \dots \rightarrow x_{n-1/n-1} \rightarrow x_{n/n-1} \rightarrow x_{n/n} \quad (2.4.3)$$

in sintesi la ricorsione alterna ad una previsione, un passo di filtraggio dove appunto, $x_{t/t-1}$ è la previsione ad un passo del filtro di Kalman, mentre $x_{t/t}$ è

l'aggiornamento che compie l'algoritmo successivamente ad ogni previsione, cioè il filtro di Kalman.

Esiste poi, anche una ricorsione per il lisciatore di Kalman che computa successivamente nell'ordine

$$X_{n/n} \rightarrow X_{n-1/n} \rightarrow X_{n-1/n-1} \rightarrow X_{n-2/n-2} \rightarrow \dots \rightarrow X_{1/n} \rightarrow X_{0/n}$$

Le stime lisciate e filtrate sono dipendenti dalle matrici di covarianza $\Omega_{t,s}$, a loro volta lisciate e filtrate elaborate ricorsivamente in questo ordine

$$\Omega_{1/0} \rightarrow \Omega_{1/1} \rightarrow \Omega_{2/1} \rightarrow \Omega_{2/2} \rightarrow \dots \rightarrow \Omega_{n-1/n-1} \rightarrow \Omega_{n/n-1} \rightarrow \Omega_{n/n} \quad (2.4.4)$$

e

$$\Omega_{n/n} \rightarrow \Omega_{n-1/n} \rightarrow \Omega_{n-1/n-1} \rightarrow \Omega_{n-2/n-2} \rightarrow \dots \rightarrow \Omega_{1/n} \rightarrow \Omega_{0/n}$$

Si inizia con il considerare la previsione ad un passo

$$\hat{X}_t = P_{t-1}(X_t) = X_{t/t-1} \quad (2.4.5)$$

dove $P_{t-1}(X_t) = P(X_t/Y_0, Y_1, \dots, Y_{t-1})$ indica la proiezione ortogonale sullo spazio delle osservazioni Y_0, Y_1, \dots, Y_{t-1} .

Nel caso del modello gaussiano $P_s(X_t)$ coincide con $E[X_s/Y_1, \dots, Y_s]$.

Per questo $X_t - \hat{X}_t$ è incorrelato con Y_s con $s=1, \dots, t$.

$$[X - P_t(X)] \perp Y_s \quad s=0,1,\dots,t$$

ed inoltre la sua matrice di covarianza degli errori

$$\Omega = E [(X_t - \hat{X}_t) (X_t - \hat{X}_t)'] \quad (2.4.6)$$

che sono rispettivamente inizializzati da

$$\begin{aligned} \hat{X}_1 &= P(X_1/X_0) \quad \text{e} \\ \Omega_1 &= E [(X_1 - \hat{X}_1) (X_1 - \hat{X}_1)'] \end{aligned}$$

successivamente l'algoritmo, continua calcolando per ogni $t=1,2,\dots,n$

$$\hat{X}_{t+1} = F_t \hat{X}_t + \Theta_t \Delta_t^{-1} (Y_t - G_t \hat{X}_t) \quad \text{e} \quad (2.4.7)$$

$$\Omega_{t+1} = F_t \Omega_t F_t' + Q_t - \Theta_t \Delta_t^{-1} \Theta_t' \quad (2.4.8)$$

dove

$$\Delta_t = G_t \Omega_t G_t' + R_t$$

e

$$\Theta_t = F_t \Omega_t G_t'$$

Per verificare quanto precedentemente affermato dobbiamo in primo luogo specificare l'uso di quelle che chiameremo innovazioni I_t

$$I_t = Y_t - P_{t-1} Y_t = Y_t - G_t \hat{X}_t = G_t (X_t - \hat{X}_t) + W_t \quad t=1,2,\dots,n$$

(2.4.9)

ricordando l'ortogonalità del vettore Y_t , rispetto la proiezione del vettore X_t
Quindi, date le proprietà lineari delle matrici di proiezione e considerato che

$$P(X/Y) = MY$$

dove M è una matrice tale per cui $M = E(XY') [E(YY')]^{-1}$, si ha che

$$P_t(\cdot) = P_{t-1}(\cdot) + P(\cdot/I_t) \quad \text{o anche} \tag{2.4.10}$$
$$P_t X_t = P_{t-1} X_t + M I_t$$

da cui otteniamo che la previsione *one-step* per il vettore X_t è

$$\begin{aligned} \hat{X}_{t+1} &= P_{t-1}(X_{t+1}) + P(X_{t+1}/I_t) \\ &= P_{t-1}(F_t X_t + V_t) + \Theta_t \Delta_t^{-1} I_t \\ &= F_t \hat{X}_t + \Theta_t \Delta_t^{-1} I_t \end{aligned} \tag{2.4.11}$$

dove

$$\begin{aligned} \Delta_t &= E(I_t I_t') \\ &= G_t \Omega_t G_t' + R_t \quad e \end{aligned}$$

$$\begin{aligned} \Theta_t &= E(X_{t+1} I_t') \\ &= E[(F_t X_t + V_t) ([X_t + \hat{X}_t]' G_t' + W_t')] \\ &= F_t \Omega_t G_t' \end{aligned}$$

Per quanto concerne invece la matrice di covarianza degli errori Ω_{t+1} osserviamo che

$$\Omega_{t+1} = E(X_{t+1}X'_{t+1}) - E(X_{t+1}\hat{X}'_{t+1}) \quad (2.4.12)$$

e usando la precedente espressione (2.4.10) di uguaglianza per \hat{X}_{t+1}

$$\begin{aligned} \Omega_{t+1} &= F_t E(X_t X'_t) F'_t + Q_t - F_t E(X_t \hat{X}'_t) F'_t - \Theta_t \Delta_t^{-1} \Theta'_t \\ &= F_t \Omega_t F'_t + Q_t - \Theta_t \Delta_t^{-1} \Theta'_t \end{aligned} \quad (2.4.13)$$

Ad ogni passo t l'algoritmo, oltre a fornire previsioni *one-step* sul vettore \hat{X}_{t+1} e sulla sua matrice di covarianza Ω_{t+1} , compie anche un filtraggio dei dati, detto per l'appunto **Kalman filter**, dove si calcola $X_{t|t} = P_t(X_t)$ e la sua matrice di covarianza degli errori $\Omega_{t|t} = E[(X_t - X_{t|t})(X_t - X_{t|t})']$, che sono determinate dalle equazioni

$$P_{t+1} X_{t+1} = P_t X_{t+1} + \Omega_{t+1} G'_{t+1} \Delta_{t+1}^{-1} (Y_{t+1} - G_{t+1} \hat{X}_{t+1}) \quad (2.4.14)$$

e

$$\Omega_{t+1|t+1} = \Omega_{t+1} - \Omega_{t+1} G'_{t+1} \Delta_{t+1}^{-1} G_{t+1} \Omega'_{t+1}. \quad (2.4.15)$$

Entrambe derivano dalla proiezione ortogonale richiamata nell'equazione (2.4.10) $P_t X_t = P_{t-1} X_t + M I_t$ dove però

$$\begin{aligned} M &= E(X_t I_t') [E(I_t I_t')]^{-1} \\ &= E[X_t (G_t (X_t - \hat{X}_t) + W_t)'] \Delta_t^{-1} \end{aligned}$$

$$= \Omega_t G'_t \Delta_t^{-1} \quad (2.4.16)$$

L'equazione (2.4.15) deriva in particolare da

$$X_t = P_{t-1} X_t = X_t - P_t X_t + P_t X_t - P_{t-1} X_t = X_t - P_t X_t + M I_t$$

e usando la (2.4.16) abbiamo

$$\Omega_{t+1|t+1} = \Omega_{t+1} - \Omega_{t+1} G'_{t+1} \Delta_{t+1}^{-1} G_{t+1} \Omega'_{t+1}$$

La prima equazione (2.4.14) può anche essere scritta come

$$\begin{aligned} X_{t+1|t+1} &= X_{t+1|t} + K_{t+1} (Y_{t+1} - G_{t+1} X_{t+1|t}) \quad \circ \\ X_{t+1} &= \hat{X}_{t+1} + K_{t+1} (Y_{t+1} - G_{t+1} \hat{X}_{t+1}) \end{aligned}$$

dove $K_{t+1} = \Omega_{t+1} G'_{t+1} \Delta_{t+1}^{-1}$

Si può provare la formulazione del previsore di Kalman sostituendo la precedente formulazione all'equazione del predittore (2.4.5) $\hat{X}_t = X_{t-1|t} = F_t X_{t|t}$, otterremmo così

$$\begin{aligned} \hat{X}_{t+1} &= F_t (\hat{X}_t + K_t (Y_t - G_t \hat{X}_t)) \\ &= F_t (\hat{X}_t + \Omega_t G'_t \Delta_t^{-1} (Y_t - G_t \hat{X}_t)) \\ &= F_t \hat{X}_t + \Theta_t \Delta_t^{-1} (Y_t - G_t \hat{X}_t) \end{aligned}$$

Anche la (2.4.14) può essere sostituita nell'equazione della matrice di previsione degli errori (2.4.8)

$$\Omega_{t+1} = F_t \Omega_t F_t' + Q_t - \Theta_t \Delta_t^{-1} \Theta_t'$$

nella sua versione semplificata

$$\Omega_{t+1} = F_t \Omega_t F_t' + Q_t$$

dove andiamo a sostituire al posto di Ω_t quanto precedentemente trovato $\Omega_{t+1|t} = \Omega_{t+1|t} - K_{t+1} G_{t+1}' \Omega_{t+1|t}'$ ottenendo

$$\begin{aligned} \Omega_{t+1} &= F_t (\Omega_t - K_t G_t' \Omega_t') F_t' + Q_t \\ &= F_t \Omega_t F_t' - F_t K_t G_t' \Omega_t' F_t' + Q_t \\ &= F_t \Omega_t F_t' - F_t \Omega_t G_t' \Delta_t^{-1} G_t \Omega_t' F_t' + Q_t \\ &= F_t \Omega_t F_t' + Q_t - \Theta_t \Delta_t^{-1} \Theta_t' \end{aligned}$$

come si voleva dimostrare.

Come detto in precedenza l'algoritmo di Kalman si avvale anche di una procedura di lisciamiento chiamata appunto lisciamiento di Kalman (a punto fisso) che calcola $X_{t|n} = P_n X_t$ e la relativa matrice di covarianza degli errori $\Omega_{t|n} = E[(X_t - X_{t|n})(X_t - X_{t|n})']$.

Le relative equazioni sono date da:

$$P_n X_t = P_{n-1} X_t + \Omega_{t,n} G_n' \Delta_n^{-1} (Y_n - G_n X_n) \quad (2.4.17)$$

$$\Omega_{t,n+1} = \Omega_{t,n} [F_n - \Theta_n \Delta_n^{-1} G_n'] \quad (2.4.18)$$

$$\Omega_{t|n} = \Omega_{t|n-1} - K_n G_n \Omega_n'$$

(2.4.19)

con $K_n = \Omega_{t,n} G_n' \Delta_n^{-1}$

In effetti usando l'equazione $P_n X_t = P_{n-1} X_t + C I_n$, dove qui $G_n(X_n - X_n^{\wedge}) + W_n$ si ha che

$$C = E [X_t (G_n (X_n - X_n^{\wedge}) + W_n)'] = \Omega_{t,n} G_n' \Delta_n^{-1} \quad (2.4.20)$$

in cui

$$\Omega_{t,n} = E[(X_t - X_t^{\wedge}) (X_n - X_n^{\wedge})']$$

Pertanto, prendendo in considerazione l'ortogonalità di V_n e W_n con $X_n - X_n^{\wedge}$, quanto appena detto su $\Omega_{t,n}$ deriva dalla previsione h passi in avanti che usa la ricorsione di Kalman

$$P_t X_{t+1} = F_t P_{t-1} X_t + \Theta_t \Delta_t^{-1} (Y_t - P_{t-1} Y_t)$$

e da questa abbiamo l'equazione che prova (2.4.18)

$$\begin{aligned} \Omega_{t,n+1} &= E[(X_t - X_t^{\wedge}) (X_n - X_n^{\wedge})'] (F_n - \Theta_n \Delta_n^{-1} G_n)' = \\ &= \Omega_{t,n} [F_n - \Theta_n \Delta_n^{-1} G_n]' \end{aligned}$$

Per provare invece l'equazione (2.4.19), ricorriamo alla scrittura

$$X_t = P_n X_t = X_t - P_{n-1} X_t - C I_n$$

rimarcando l'ortogonalità di $X_t - P_n X_t$, e richiamando quanto enunciato per C nella (2.4.19)

$$\Omega_{t/n} = \Omega_{t/n-1} - K_n G_n \Omega'_n$$

2.5 Stima di massima verosimiglianza

Ora si specificherà l'approccio che sviluppano procedure per calcolare stime di massima verosimiglianza.

Considerando il modello dato dalle formule (2.3.1) e (2.3.2) e supponendo che il modello si possa definire parametrizzato dal vettore θ che comprende le stime per le matrici Q ed R specificate in 2.3.1 e 2.3.2.

La stima di massima verosimiglianza per θ è basata sulla massimizzazione della funzione di verosimiglianza delle osservazioni y_1, y_2, \dots, y_n rispetto alle componenti del vettore θ .

Se la densità condizionata di y_t , date le osservazioni y_1, y_2, \dots, y_{t-1} , è indicata con $f(\cdot | y_{t-1}, y_{t-2}, \dots, y_1, \theta)$ la verosimiglianza per Y_t , $t=1, 2, \dots, n$ può essere scritta come

$$L(\theta) = \prod_{t=1}^n f_t(y_t | y_{t-1}, y_{t-2}, \dots, Y_1) \quad (2.5.1)$$

Il calcolo della verosimiglianza per θ potrebbe essere in generale di notevole complessità ma nel caso di un modello state-space gaussiano abbiamo che, le densità condizionate possono essere così espresse

$$f_t(y_t | y_{t-1}, \dots, y_1) = (2\pi)^{-\omega/2} (\det \Delta_t)^{-1/2} \exp[-1/2 I'_t \Delta_t^{-1} I_t]$$

così che la verosimiglianza delle osservazioni vale

$$L(\theta; Y_1, Y_2, \dots, Y_n) = (2\pi)^{-n\omega/2} (\prod_{j=1}^n \det \Delta_t)^{-1/2} \exp[-1/2 \sum_{j=1}^n I' \Delta_t^{-1} I_t] \quad (2.5.2)$$

Le stime per i parametri in questione possono essere ottenute sfruttando una procedura di ottimizzazione, nota come Nelder-Mead (Nelder-Mead, 1965).

Capitolo 3

Specificazione e risultati

3.1 I dati: Descrizione della base dati

Solitamente, per la rilevazione degli agenti inquinanti vengono usate centraline fisse in concomitanza con alcune mobili. Quelle fisse vengono installate in punti ritenuti tipici per la rilevazione dell'inquinamento.

Nella fattispecie la stazione presente in zona Ospedale (di tipo **A**) e quelle collocate in zona San Gregorio e l'Arcella (di tipo **B**) sono quelle fisse, mentre quelle mobili, sono installate su appositi pulmini e vengono spostate di periodo in periodo di rilevazione, per ottenere importanti informazioni in regioni, relative alla zona di ricerca, che altrimenti non sarebbe coperta.

L'utilizzo di un'unica centralina per diversi punti di rilevazione, non consente però, l'utilizzo di modelli molto usati in letteratura (Cressie, 1993; Haas, 1995; Haas, 1996), e costringe a ricercare metodi alternativi e soluzioni specifiche per la formulazione di un modello adatto a dare stime efficienti.

Obbiettivi di questa tesi di laurea sono pertanto:

- Elaborare un modello per la concentrazione di biossido di zolfo nella città di Padova che colga le dinamiche principali del tempo e dello spazio sul valore medio e sulla struttura della covarianza, utilizzando informazioni incomplete come quelle che si presentano nei dati a disposizione
- Formulare previsioni in senso sia spaziale che temporale.

I dati che si utilizzeranno sono stati forniti dall'Amministrazione Provinciale di Padova, che è tenuta a mantenere una rete di monitoraggio degli agenti inquinanti sul territorio provinciale.

Sono relativi alle medie giornaliere di SO₂, rilevati nell'arco dello stesso giorno. Le stazioni in realtà sono in grado di misurare l'inquinante in un lasso temporale molto breve (pochi secondi), ma vengono fornite solo le medie orarie. Per comodità, dati sono comunque relativi alla media giornaliera.

Le rilevazioni sono effettuate da cinque centraline: tre sono fisse, di cui due operanti sin dall'inizio della periodo di interesse mentre la restante è stata installata verso la fine del periodo di riferimento; le restanti due invece, sono state spostate di quartiere in quartiere per la durata di qualche settimana, per un totale di 14 siti.

Le caratteristiche tecniche della strumentazione che rileva la concentrazione di biossido di zolfo utilizzati al momento della rilevazione non permettono di rilevare il valore, quando questo non supera i 10 $\mu\text{g}/\text{m}^3$, ma dato che queste rilevazioni sono le meno interessanti in termini di segnalazioni di situazioni inquinanti, si è deciso di assegnare convenzionalmente a tutte le osservazioni censurate, il valore 5 $\mu\text{g}/\text{m}^3$ (cioè l'equidistribuzione dei valori tra 0 e 10 $\mu\text{g}/\text{m}^3$).

Il vero problema dei dati forniti in realtà è il numero preponderante di dati mancanti che si vengono a creare per effetto, in primo luogo, della mobilità delle centraline che, rilevano solo per un periodo limitato di tempo rispetto all'intero arco temporale la concentrazione di inquinanti, e anche a causa di fenomeni meno preventivabili, come il malfunzionamento delle centraline dovuto a cause tra le più disparate, come la normale staratura o atti vandalici.

Le osservazioni sono state rilevate nella zona del comune di Padova nel periodo che va da inizio Aprile 1991 a fine Marzo del 1994 (1/4/1991-31/3/1994), per un totale quindi di 1094 giorni complessivi.

L'area geografica che le rilevazioni andranno a coprire è vasta all'incirca 11,5*7,5 km, limitata ad Ovest dalla catena di Colli Euganei, a sud dal canale scaricatore, a nord e a est dal fiume Brenta.

Visto il grado di accuratezza e le modalità delle misurazioni, si possono, per facilitare la comprensione, considerare tutti i complessivi 17 siti come centri di rilevazione permanente di cui però, 14 dei quali presentano rilevazioni per solo brevi periodi di tempo.

Si denoterà il singolo dato con $Z(s_j, t_i)$, dove t_i indica il giorno che varia da 1 a 1096 e s_j indica i siti a cui appartengono le rilevazioni.

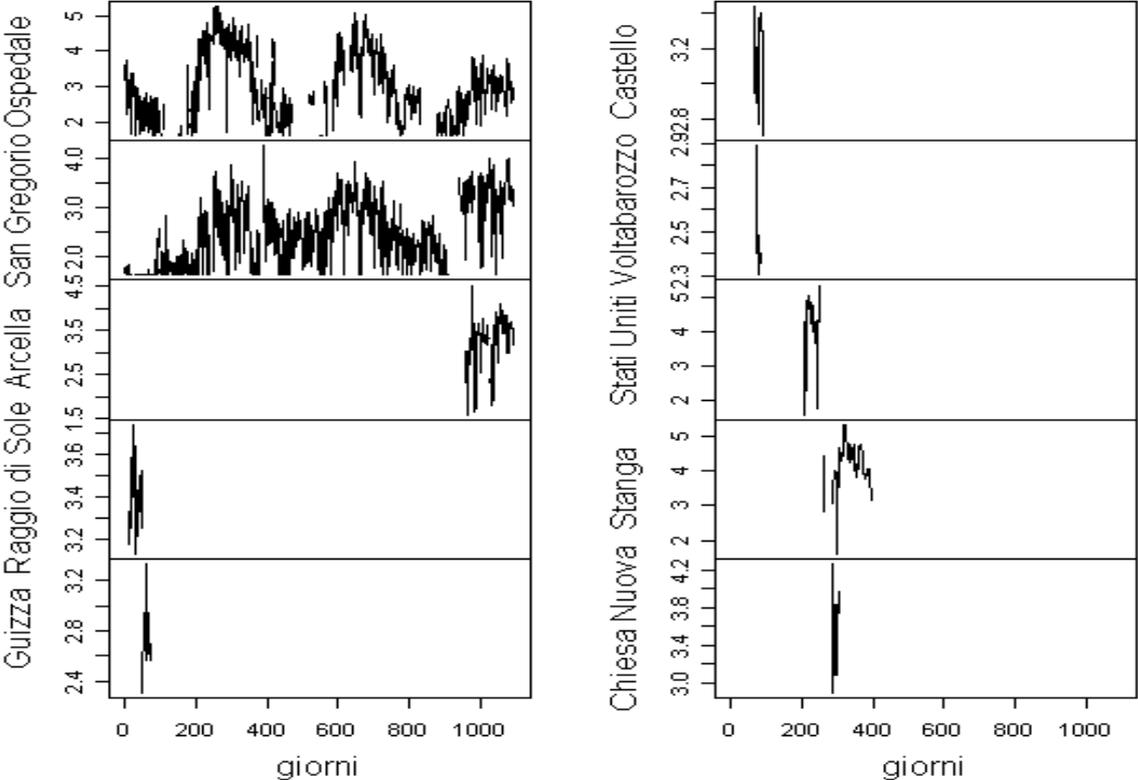
Si presenterà ora una tabella che evidenzia oltre alla posizione geografica, il periodo di osservazione, il numero di giorni in cui la centralina è stata predisposta per la rilevazione su un determinato sito e il numero di giorni in cui si sono riscontrati problemi di malfunzionamento che hanno comportato necessariamente la mancanza di rilevazione.

	Sito	Coord E-O	Coord N-S	Periodo di osservazione	Numero di giorni di osservazione	Giorni mancanti
A	Ospedale	0,800	0,550	01.04.91-31.03.94	1096	261
B	S.Gregorio	3,300	0,675	02.04.91-31.03.94	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	16.05.91-10.16.91	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	-1.125	1.225	09.11.93-31.03.94	143	0

Tabella 3.1: Siti di rilevazione, collocazione, numero di osservazioni, dati mancanti.

Per avere una visione d'insieme relativa alla presenza dei molti valori mancanti si riportano di seguito i grafici delle serie temporali che si sono venute a creare nei vari posizionamenti delle centraline, siano esse fisse o mobili.

Andamento delle rilevazioni



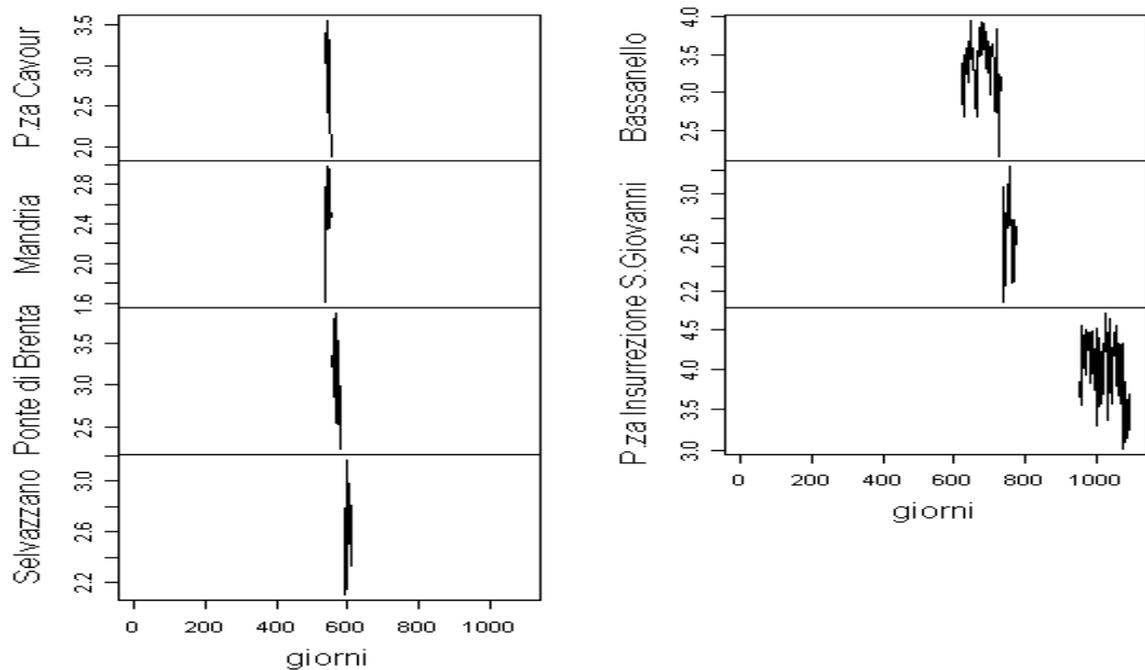


Figura 3.1: Serie storiche dei siti di rilevazione.

Dai grafici presentati, e in particolar modo dai primi due (che coprono l'intero arco temporale) è evidente un forte componente stagionale, che come vedremo in seguito porta la concentrazione di biossido di zolfo a livelli elevati proprio in corrispondenza dei mesi invernali per poi tornare verso valori relativamente più bassi in quelli estivi.

Si riporta di seguito anche la collocazione geografica delle centraline di rilevazione, evidenziando in rosso quelle fisse.

In particolare si è ritenuto opportuno rappresentare il luogo geografico in esame su un piano di tipo cartesiano, dove l'origine si trova in un punto che potremo approssimare con il centro storico di Padova, le ordinate rappresentano la dimensione "Ovest-Est" e le ascisse rappresentano quella "Nord-Sud".

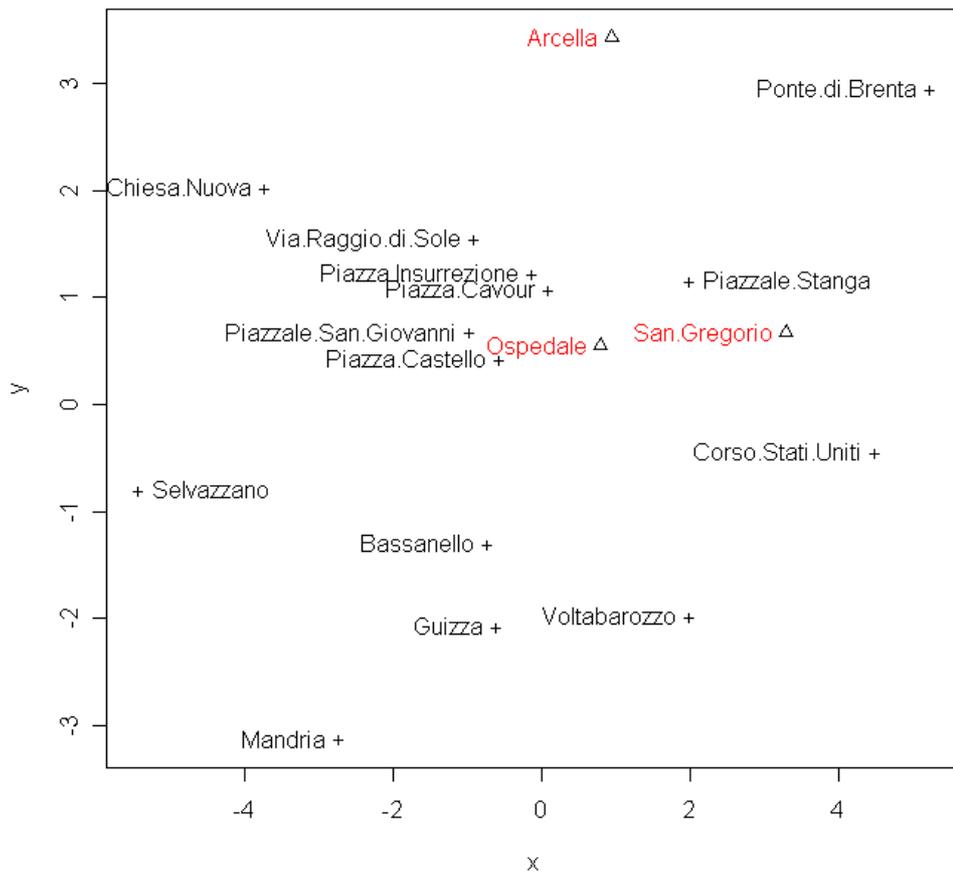


Figura 3.2: Collocazione geografica delle centraline di rilevazione.

3.2 Analisi preliminare

I dati a disposizione sono un elenco di 2571 righe in cui ogni misurazione

presenta informazioni sulla localizzazione della centralina che ha rilevato il dato.

Nelle figure che seguiranno (3.3-3.4), si mostra l'andamento dell'inquinamento nei siti di rilevazione che hanno coperto con le loro informazioni tutto l'arco di osservazione del fenomeno, ciò relativi rispettivamente alle centraline in zona Ospedale e in zona S.Gregorio. Si riportano solo questi due esempi perché la serie temporale registrata dalla terza centralina fissa in zona Arcella è entrata in funzione a fine del periodo di osservazione, e le altre mobili, sebbene possano aiutare a coprire il territorio e fornire informazioni in questo senso, non sono graficamente di aiuto, poiché apportano informazioni molto frammentarie rispetto all'intero arco temporale. Sono riportati di seguito i grafici delle serie temporali riguardanti le due centraline fisse.

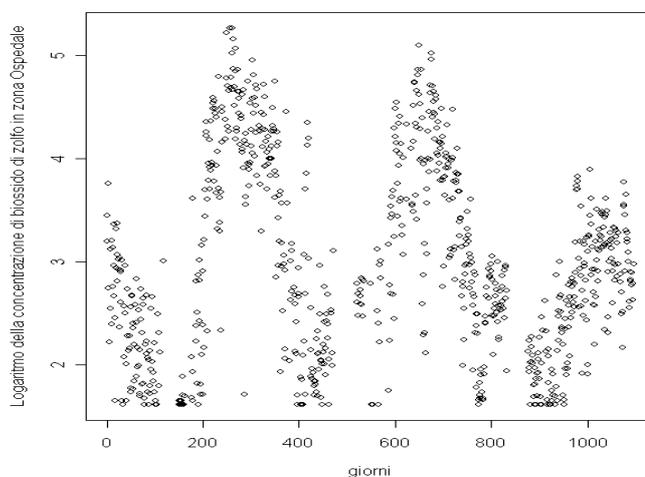


Figura 3.3: Andamento della concentrazione di biossido di zolfo rilevato dalla centralina fissa situata vicino all'Ospedale.

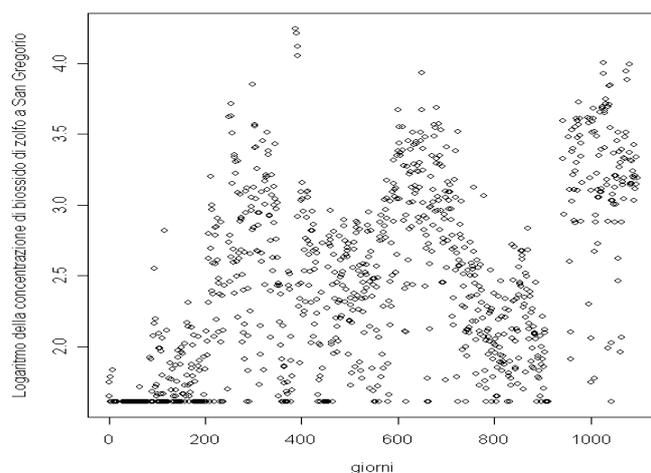


Figura 3.4: Andamento della concentrazione di biossido di zolfo rilevato dalla centralina fissa situata in zona San Gregorio.

I grafici 3.3 e 3.4, e in particolare quelli relativi alla zona dell’Ospedale, evidenziano un marcato andamento stagionale che si manifesta con picchi di inquinamento intensi in relazione ai mesi freddi dell’anno.

Ciò non dovrebbe sorprendere poiché, come si diceva in precedenza, l’innalzamento sistematico della concentrazione di inquinante è causato dalla dispersione nell’ambiente di sostanze inquinanti derivanti dal riscaldamento domestico.

Bisogna notare che, i picchi di inquinamento registrati in zona Ospedale sono in genere, molto più elevati di quelli registrati a S.Gregorio, questo probabilmente perché quest’ ultima, è meno densamente abitata e quindi meno soggetta ad inquinamento causato da riscaldamento domestico.

Nei due grafici si è evidenziato un congruo numero di rilevazioni attorno al valore $5 \mu\text{g}/\text{m}^3$, questo perché, come si diceva nel capitolo introduttivo in

relazione al caso della rete di rilevazione del comune di Padova, si assume che i valori inferiori ai $10 \mu\text{g}/\text{m}^3$ siano equidistribuiti sull'intervallo 0-10, pertanto il valore 5 è stato preso come valore medio di misurazioni non relative a situazioni inquinanti, e quindi di poca rilevanza.

Le rilevazioni sono però riportate in scala logaritmica, questo probabilmente perché si voleva limitare l'influenza di misurazioni (in particolare quelle riferibili ai periodi invernali) che possono essere state influenzate da condizioni meteorologiche che avrebbero portato alle rilevazione di misure, per quanto precise, troppo fuorvianti.

Si riporta di seguito il grafico relativo al comportamento del fenomeno di interesse dove ogni punto è la media giornaliera delle rilevazioni effettuate nello stesso giorno; bisogna rimarcare che il numero giornaliero di osservazioni non è costante (varia da 0 a 5), perciò le medie presenti sono state calcolate con un numero di elementi variabile (ovviamente i giorni senza osservazioni non sono stati rappresentati).

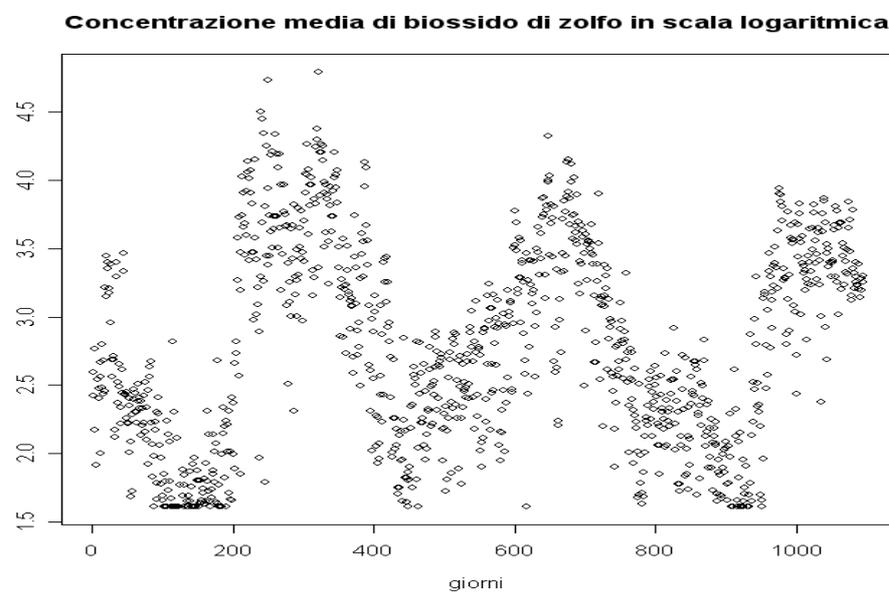


Figura 3.5: Comportamento delle medie delle rilevazioni giornaliere (scala logaritmica).

Anche qui, il comportamento della serie evidenzia una netta stagionalità, che porta i valori medi giornalieri su livelli più alti proprio in concomitanza con i mesi più freddi dell'anno.

Preso atto del comportamento che la concentrazione media di biossido di zolfo, nonché quella riferita ai singoli luoghi di rilevazione, assume in relazione alla sua stagionalità, è opportuno depurare i dati dall'effetto del suddetto fenomeno.

Si è preferito quindi, cercare di eliminare la stagionalità tramite un modello di tipo additivo.

Si ipotizza la singola realizzazione dipendente da tempo e spazio $Z(s,t)$ (ricordando che questa è il logaritmo dei dati originali), come dipendente da due fattori tra loro indipendenti:

$$Z(s,t) = \varphi_S(s) + \varphi_T(t) + \varepsilon(s,t)$$

dove $\varphi_S(s)$ definisce la dinamica legata allo spazio e $\varphi_T(t)$ definisce quella legata al tempo, corredata ovviamente dalla componente di errore $\varepsilon(s,t)$ (cioè la differenza in scala logaritmica tra la realizzazione e la stima per la singola osservazione).

In sostanza, $\varphi_T(t) = s(t)$ dove s è una funzione non nota del trend temporale mentre

$$\varphi_S(s) = \sum_{j=1}^{17} \beta_j I_{sj}(s)$$

dove

$$I_{sj}(s) = \begin{cases} 1 & \text{se } s=s_j \quad \text{per } j=1,\dots,17 \\ 0 & \text{altrimenti} \end{cases}$$

Si cercherà dunque, come si diceva in precedenza, di spiegare parte della dinamica spaziale tramite l'uso di un modello additivo generalizzato tramite l'utilizzo della libreria *mgcv* presente in *R*, ottenendo così i residui direttamente dall'output del modello specificato.

I residui del modello in questione pertanto sono il risultato della differenza:

$$Z^d(s,t) = Z(s,t) - \hat{Z}(s,t)$$

dove otteniamo i dati depurati dalla componente stagionale $\hat{Z}(s,t)$ stimata dal modello.

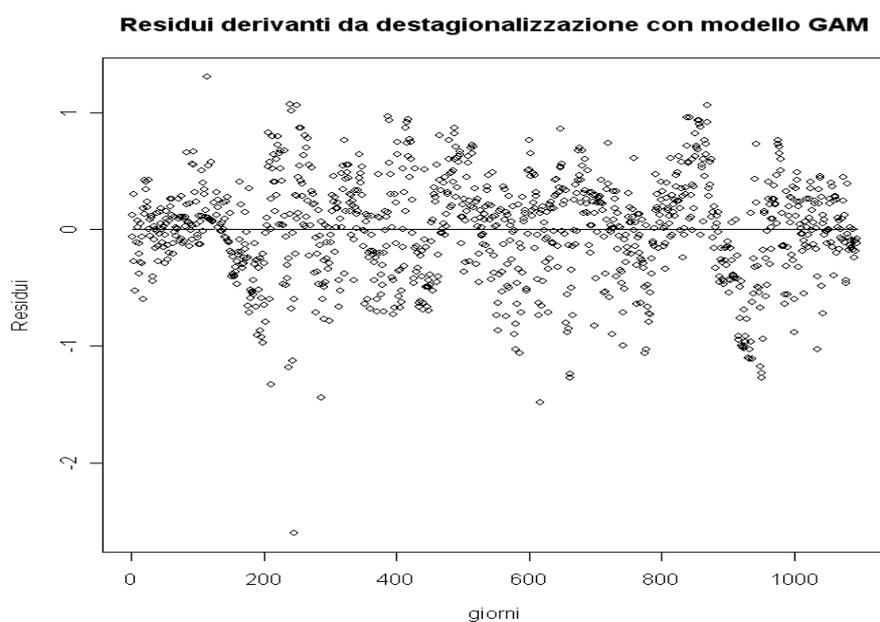


Figura 3.6: Residui della serie destagionalizzata della concentrazione media giornaliera di biossido di zolfo.

3.3 Rappresentazione nello stato-spazio

La matrice di trasposizione Φ definita in 2.3.2, che nel programma è stata definita con la lettera F , è stata supposta diagonale ed identica di ordine m (poiché m è il numero di nuclei di lisciamento), questo perché si è supposto che il generico vettore X_t differisce da X_{t-1} solo a causa del vettore degli errori v_t , descrizione di un processo *random walk* di ordine 1

$$X_t = X_{t-1} + v_t$$

con v che si distribuisce

$$v_t \sim Wn(0, Q)$$

dove la matrice di covarianza Q è stata supposta di tipo diagonale per la non correlazione degli errori di previsione nei nuclei di convoluzione

$$Q = (\sigma^2_1, \dots, \sigma^2_m)$$

Anche la matrice di covarianza R del vettore degli errori ω_t , che ha dimensioni variabili nel tempo $k_t * k_t$, viene supposta di tipo diagonale ed inoltre scalare, visto che è stata ipotizzata la non correlazione degli errori di rilevazione

$$R = (\sigma^2_1, \dots, \sigma^2_k)$$

Implicitamente si assume quindi che gli errori rispetto alla parte deterministica del modello

$$Y_t = G_t X_t + \omega_t$$

siano omoschedastici con varianza comune σ^2_{ω} .

Si è poi deciso di inizializzare x_0 come un vettore di m elementi pari a 0.

La matrice di covarianza degli errori Ω_0 , è stata inizializzata come una matrice diagonale, $\Omega_0 = 1000 * I_m$, questo per la mancata conoscenza del valore iniziale x_0 .

La matrice G_t , per ogni istante t è formata da un numero k_t di centraline (poiché varia a seconda del numero di rilevazione pervenute in un determinato giorno) e un numero m (fissato) di nuclei di lisciamiento.

Per ottenere la matrice di dimensionalità esatta per ogni istante t , bisognerà dunque, operare una selezione delle righe relative a rilevazioni non nulle per ogni giorno, di cui comunque si parlerà più approfonditamente in seguito.

Ogni elemento della matrice, che supporremo già selezionata, G_{ij} (con i indice dell' i -esimo sito di rilevazione e j indice del j -esimo nucleo di lisciamiento) sarà

$$G_{ij} = (2/\pi) \det (\Phi)^{-1/2} * \exp(-1/2 * (||m_j - x_i||^T * \Phi^{-1} * ||m_j - x_i||))$$

dove $||m_j - x_i||$ è la distanza euclidea dal j -esimo nucleo di lisciamiento al i -esimo sito di rilevazione e Φ qui è una matrice simmetrica e definita positiva, quadrata di dimensione $2*2$, che si è supposto, diagonale e che presenta le larghezze di banda del nucleo nella direzione 'Ovest-Est' e di quella 'Sud-Nord'.

Se i due valori sulla diagonale principale sono uguali, allora il nucleo

gaussiano bivariato che si viene a creare, si può definire di tipo isotropico, nel senso che l'influenza che il nucleo esercita sulla rilevazione è proporzionale alla distanza euclidea tra i due, mentre se i due valori sono diversi allora avremo che l'influenza che esercita il nucleo può essere maggiore o minore a seconda, nella direzione "Sud-Nord" oppure in quella "Ovest-Est", e ciò è sufficiente a definire un nucleo di tipo anisotropico, dove cioè, la sua influenza varia a seconda della direzione.

3.4 Specificazione dei nuclei di convoluzione

Come si diceva in precedenza, per costruire un algoritmo in grado di fornire previsioni tramite il modello stato-spazio è necessario arrivare alla costruzione della matrice G_t , dell'equazione in (2.3.3). Questa è matrice di dimensione $k*m$, dove k è il numero di localizzazioni (nella fattispecie 17) ed m il numero di nuclei.

Nella specificazione della matrice si deve tener conto oltre che del numero di nuclei di convoluzione, anche della loro posizione.

È possibile dislocare i nuclei in modo che questi riescano a coprire l'area di interesse ed è ovvio che bisognerebbe cercare dei criteri guida per cui questi nuclei debbano essere posti per poter catturare in modo esaustivo l'informazione che proviene dalle rilevazioni.

In particolare, bisognerebbe cercare di dislocare dei nuclei proprio dove la concentrazione di siti di rilevazione è maggiore, nel caso di Padova è la zona relativa al centro storico (circa 1,5 km di raggio con centro approssimativamente in piazza Cavour), ma nello stesso tempo bisognerebbe cercare di non isolare eccessivamente zone che presentano centraline di rilevazione molto più sparse nel piano, e qui ovviamente ci riferiamo alla periferia.

È quindi opportuno collocare nuclei in misura proporzionale alla concentrazione dei siti di rilevazione posti nelle zone limitrofe al centro, in maniera che anche l'informazione proveniente da queste aree venga esaminata dal modello.

È ovvio peraltro che, un numero di nuclei elevato produrrebbe nel campione un buon adattamento a scapito però della capacità previsiva. Inoltre con un numero eccessivo di nuclei finirebbe per appesantire il modello dal punto di vista computazionale.

Seguendo questi criteri, si è cercato di collocare un numero limitato di nuclei di lisciamiento nella spazio di riferimento.

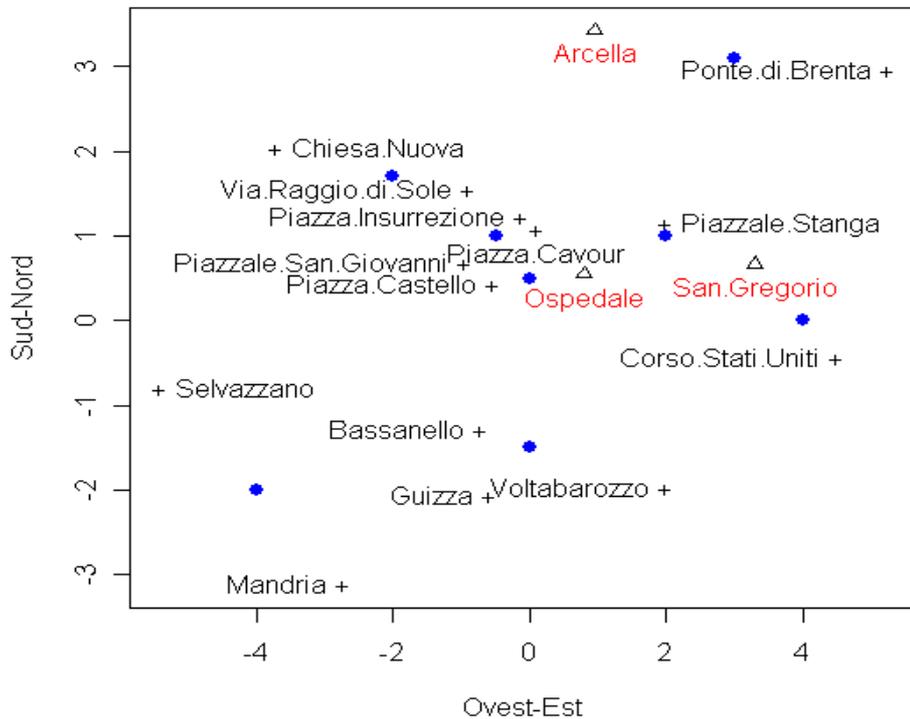


Figura 3.7: Possibile collocazione dei nuclei di lisciamiento.

Nella figura 3.7 sono rappresentati oltre ai siti di rilevazione, anche i nuclei contrassegnati con un pallino blu. Questi siti sono stati allocati solo seguendo i criteri guida sopra elencati e lasciandone la discrezionalità all'esaminatore, ma è anche possibile trovare una procedura di ottimizzazione che ora si andrà a descrivere.

Nel primo passo si è scelto di circoscrivere tutti i siti di rilevazione all'interno di un poligono in grado di contenere gli stessi siti.

Questo poligono, e l'algoritmo che regola la sua costruzione (*comp.poly* incluso nella libreria *splancs* di *R*) dipende dalla scelta che si è fatto riguardo la griglia di cui sopra, infatti se si specifica una certa numerosità n , la procedura produrrà una griglia fatta di n linee che tagliano l'area di rilevazione senso orizzontale ed altre n linee che tagliano l'area in senso verticale, producendo $n*n$ punti di intersezione.

Il poligono avrà quindi, come vertici alcuni di questi punti di intersezione, e cercherà di racchiudere tutti i siti di rilevazione al suo interno; i nuclei di lisciamento giaceranno, come vedremo, all'interno di questo poligono.

Nello stabilire il livello di dettaglio della griglia è pertanto opportuno dare una numerosità adatta a far sì che tutti i punti vengano racchiusi all'interno di questo poligono, ma è altrettanto importante che la griglia non sia eccessivamente articolata poiché, in questo modo, il nucleo verrebbe centrato in un punto prossimo a quello di rilevamento.

In realtà, non è di interesse che il nucleo sia vicino al luogo di rilevazione, ma che, in un'ottica di compromesso, il minor numero di nuclei riesca a coprire e a raccogliere in modo esaustivo informazioni su una zona in cui nessun sito di rilevazione venga trascurato.

Infatti la logica del metodo della convoluzione applicato alla matrice G_t , imporrebbe un peso basso ad osservazioni rilevate in siti lontani da nuclei di lisciamento.

Si riporta nella figura 3.8 il caso di un griglia inadatta a coprire i siti.

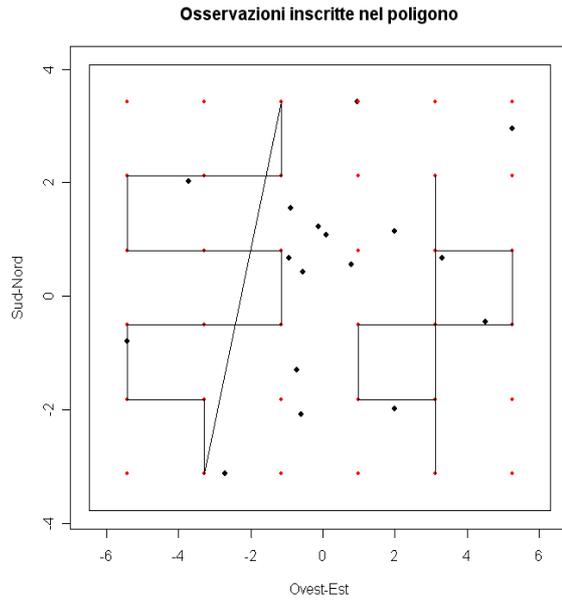


Figura 3.8: Specificazione di una griglia troppo rada, inadatta a racchiudere i siti dentro un poligono.

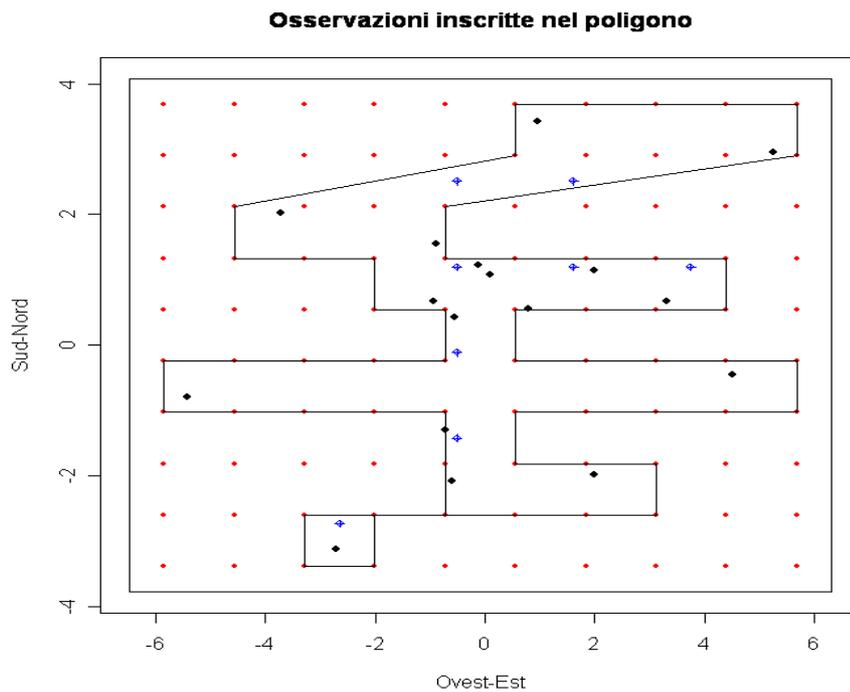


Figura 3.9: Buona specificazione di una griglia.

Nella figura 3.9 si presenta la creazione di un poligono adatto allo scopo; in effetti il poligono vede inscritti al suo interno tutto i siti contrassegnati con un pallino nero, mentre quelli rossi sono i punti di intersezione che si sono venuti a creare dall'incrocio tra linee orizzontali e verticali della griglia e che sono andati a comporre il poligono.

Tramite la funzione *gridpts*, implementata nella libreria *splancs* del programma R, si sono poi potuti trovare le localizzazioni dei nuclei di lisciamiento, dove, questa volta, il dominio è stato ristretto al poligono trovato in precedenza.

La funzione (si veda Rowlingson e Diggle, 1993) richiede, oltre al poligono in cui si devono inscrivere i nuclei di lisciamiento, anche una valutazione non approssimata della loro numerosità, cosicché la funzione possa posizionare i nuclei in un numero prossimo a quello suggerito.

È pertanto frequente che il numero di nuclei che si desiderano introdurre nel modello non corrisponda a quelli che la funzione, essendo vincolata dal poligono, effettivamente crea.

Si sono creati in questo modo i nuclei contrassegnati in figura 3.9 con il pallino di colore blu.

3.5 Selezione del numero di nuclei

Un importante indice della bontà di stima del modello è data dall'errore quadratico medio che si può dividere in 2 componenti (Azzalini A. e Scarpa B., 2004)

$$E(Y_i(\hat{\theta}) - E(Y_i|Y_{t-1}))^2 = [E(Y_i(\hat{\theta})) - E(Y_i|Y_{t-1})]^2 + \text{Var}(Y_i(\hat{\theta}))$$

la prima che potremmo chiamare devianza del modello e la seconda che potremmo invece chiamare varianza, dove $Y_t(\hat{\theta})$ è la stima degli Y_t date le stime dei parametri del modello ed Y_{t-1} è uguale a $Y_{t-1}Y_{t-2}, \dots, Y_1$.

Ebbene, quanto nello stesso poligono di riferimento aumentiamo il numero di nuclei, è come se introducessimo dei nuovi parametri, pertanto il modello si adatta in maniera sempre migliore ai dati, facendo calare la devianza, ma un aumento rilevante dei nuclei di liscio comporta necessariamente un incremento contestuale della varianza, poiché il modello sovrapparametrizzato non sarebbe in grado di adattarsi a dei nuovi dati.

In sostanza, con l'aumento dei nuclei, si crea nel modello una dipendenza dai dati, che andrebbe a scapito di un'adattabilità dello stesso rispetto ad altri dati. È buona norma quindi in modelli come questo, cercarne il numero di nuclei che minimizza l'errore quadratico medio, cioè la somma di entrambi gli elementi. In questo senso infatti, il totale delle due componenti dovrebbe tracciare una forma concava e il minimo corrisponderà quindi, al numero ottimo di parametri.

La selezione del numero di nuclei avviene attraverso il metodo della convalida incrociata.

In primo luogo si divide il campione di 1096 giorni in uno che comprende i primi 800, detto campione di *training*, ed un altro comprendente i restanti 296, detto insieme di verifica.

Successivamente, per un numero fissato di m nuclei di liscio, si arriva alla stima del modello, sui primi 800 giorni, tramite l'algoritmo di Kalman, che in seguito si specificherà nel dettaglio.

Nello stesso tempo si ricava la devianza del modello calcolata nel seguente modo

$$D = \sum_{t=1}^T D_t = \sum_{t=1}^T \sum_{i=1}^{nt} (I_{it}^*)^2$$

in cui le I_{it}^* sono

$$I_{it}^* = \Delta_t^{-1/2} I_{it} \tag{3.5.1}$$

e dove Δ_t non è che

$$\Delta_t = \text{Var} (Y_t - \hat{Y}_t)$$

Successivamente, tramite la funzione di ottimizzazione sui parametri (descritta in 2.5.2), si sono trovate la stima del parametro della matrice R (matrice di varianza/covarianza del vettore degli errori ω , supposta diagonale scalare) e gli m valori della matrice Q (matrice di varianza/covarianza del vettore degli errori v , supposta diagonale).

A questo punto non resta che da sostituire i suddetti valori al modello stimato sui restanti 296 giorni di osservazione.

Anche qui, contestualmente alla stima del modello si ricaverà la varianza dello stesso, in modo simile a quanto precedentemente specificato

$$V = \sum_{t=1}^T V_t = \sum_{t=1}^T \sum_{i=1}^{nt} (I_{it})^2$$

Sia la devianza che la varianza sono state calcolate per un numero limitato di nuclei, in particolar modo perché l'ottimizzazione dei parametri richiede tempi molto elevati, e nello specifico si sono calcolati solo per un numero di

nuclei pari a 5, 10, 15 e 20.

Bisogna poi dire che, il numero di nuclei per cui sono stati calcolati varianza e devianza è solo un numero suggerito dall'operatore, che può variare a causa della funzione *gridpts*, che ottimizza la numerosità sul poligono di riferimento.

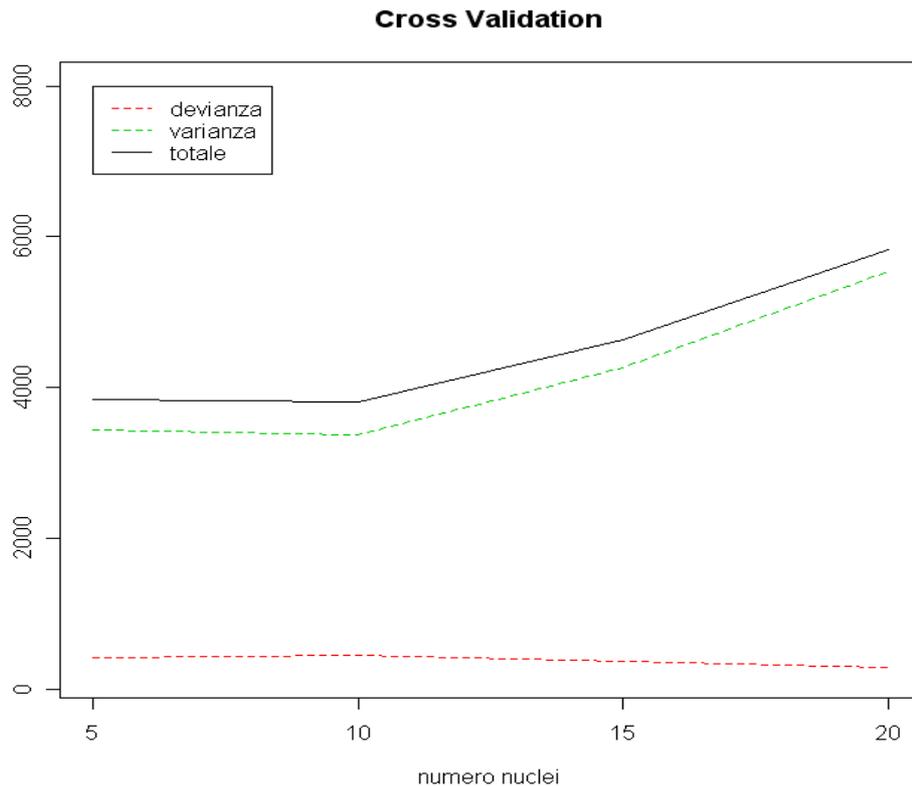


Figura 3.10: Risultato della cross validation.

Nella figura emerge che il numero ideale di nuclei di liscio è un numero prossimo ai 10.

3.6 Risultati

Dopo aver stimato il modello, con l'aiuto del filtro di Kalman utilizzando il numero idoneo di nuclei di lisciamiento stabilito in predentenza e le stime per i parametri della matrici di varianza/covarianza Q e R si è arrivati all'ottenimento di altri parametri relativi ai nuclei di lisciamiento, che, tramite una generalizzazione del metodo della convoluzione sono serviti a produrre previsioni su tutto il piano di interesse.

Si sono prodotti quindi, una serie di grafici in grado illustrare la concentrazione di biossido di zolfo nell'area di Padova e dintorni, nell'arco temporale di rilevazione, cercando di focalizzare l'attenzione sui giorni invernali ed estivi (si è scelto Luglio per quelli estivi e Gennaio per quelli invernali), che dovrebbero essere contraddistinti da oscillazioni rispetto il livello medio di inquinamento di maggior rilevanza.

Bisogna rimarcare che, la scala di colori caldi che contraddistingue il livello di inquinamento rispetto quello medio del periodo, sebbene illustri differenze apparentemente marcate nella zona di interesse, il *range* in cui spaziano i livelli di inquinamento sono piuttosto limitati; proprio per questo motivo si è deciso di corredare i grafici con un prospetto in tre dimensioni in modo da informare il lettore sulla scala di valori che si sta usando.

Figura 3.11: Concentrazione di biossido di zolfo nel Luglio 1991.

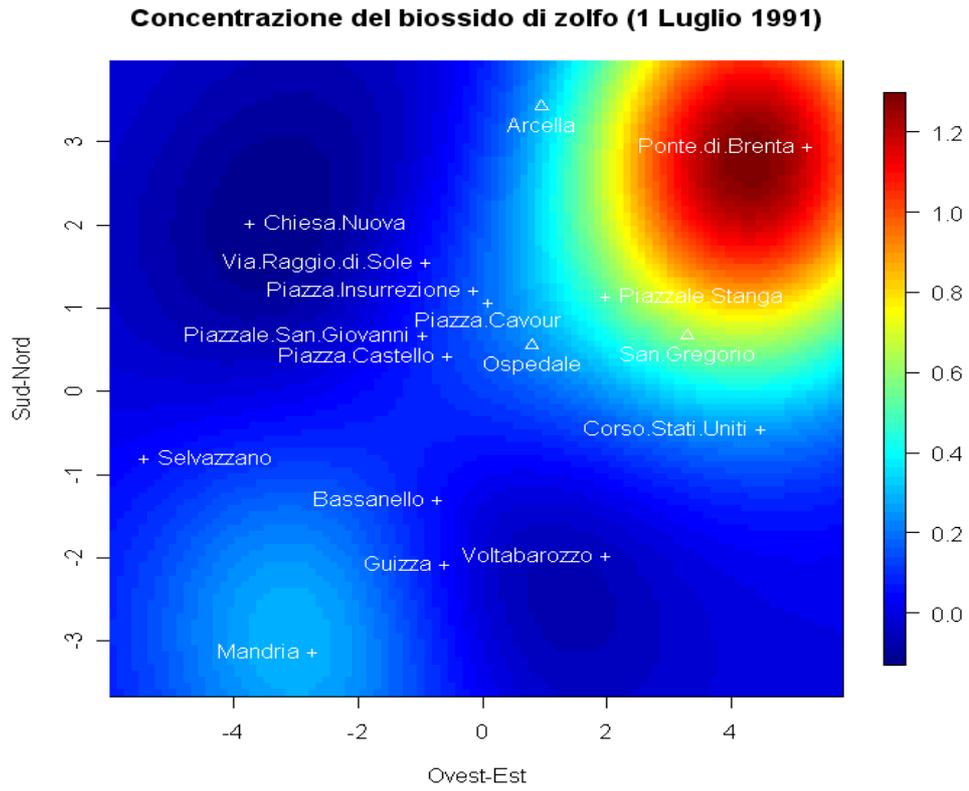


Figura 3.12: Concentrazione di biossido di zolfo nel Gennaio 1992.

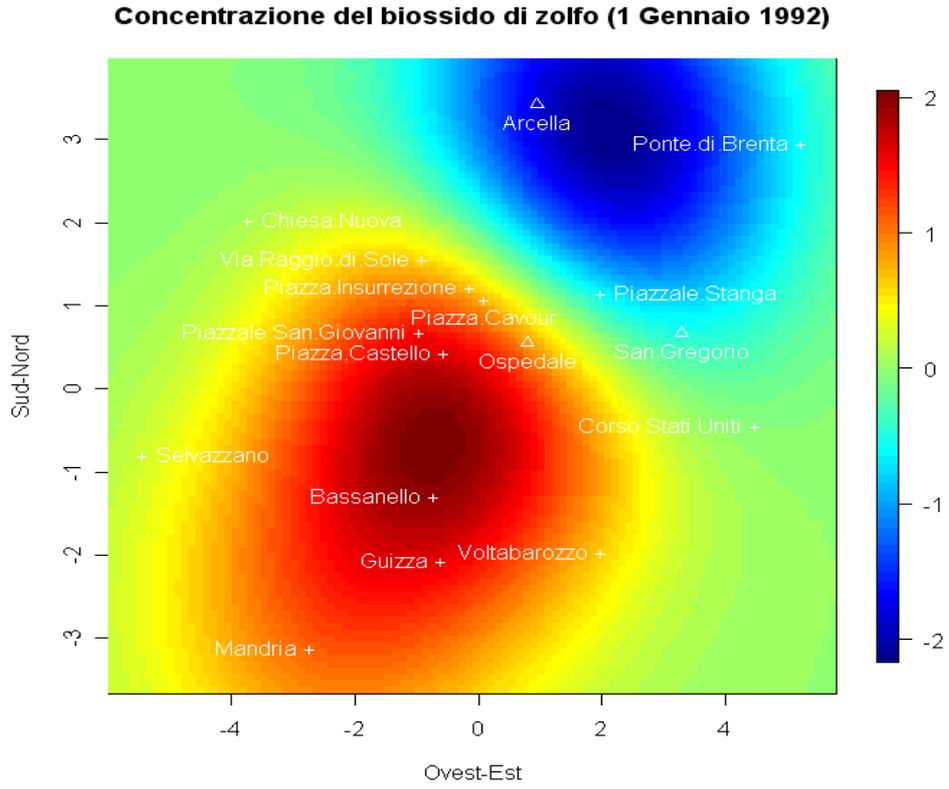


Figura 3.13: Concentrazione di biossido di zolfo nel Luglio 1992.

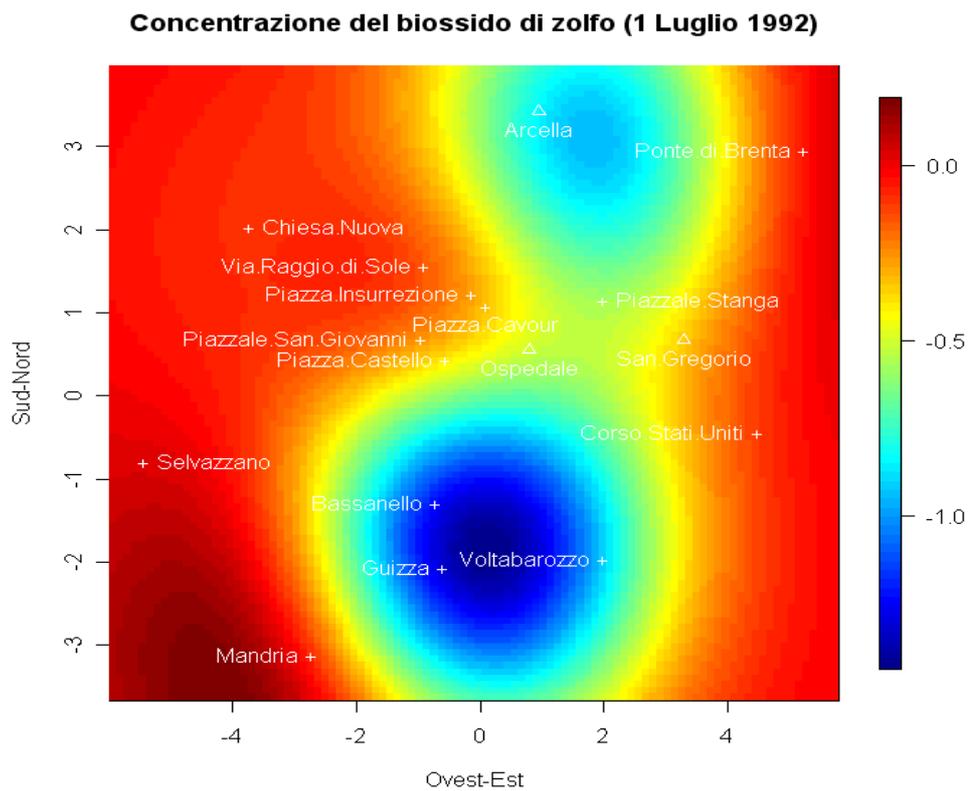


Figura 3.14: Concentrazione di biossido di zolfo nel Gennaio 1993.

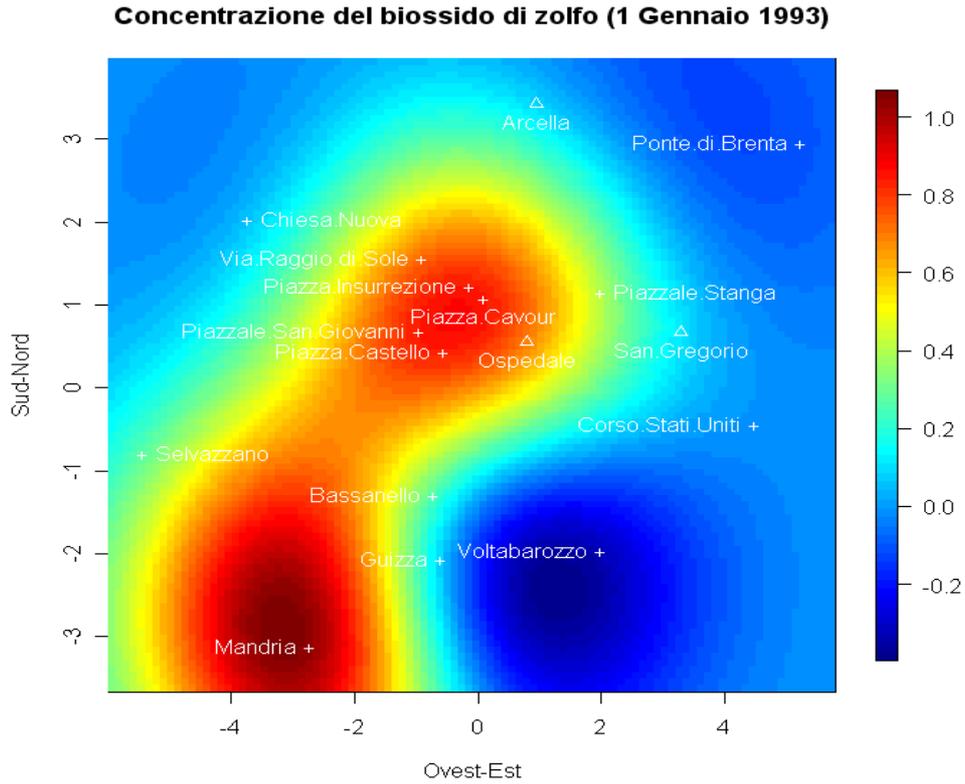


Figura 3.15: Concentrazione di biossido di zolfo nel Luglio 1993.

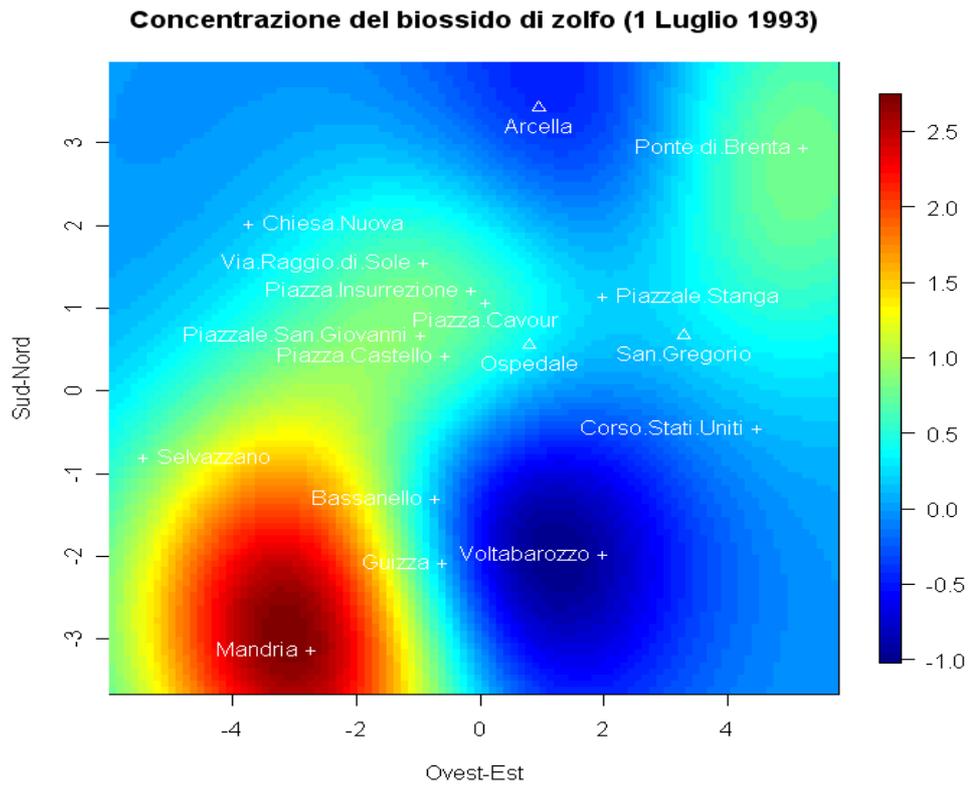
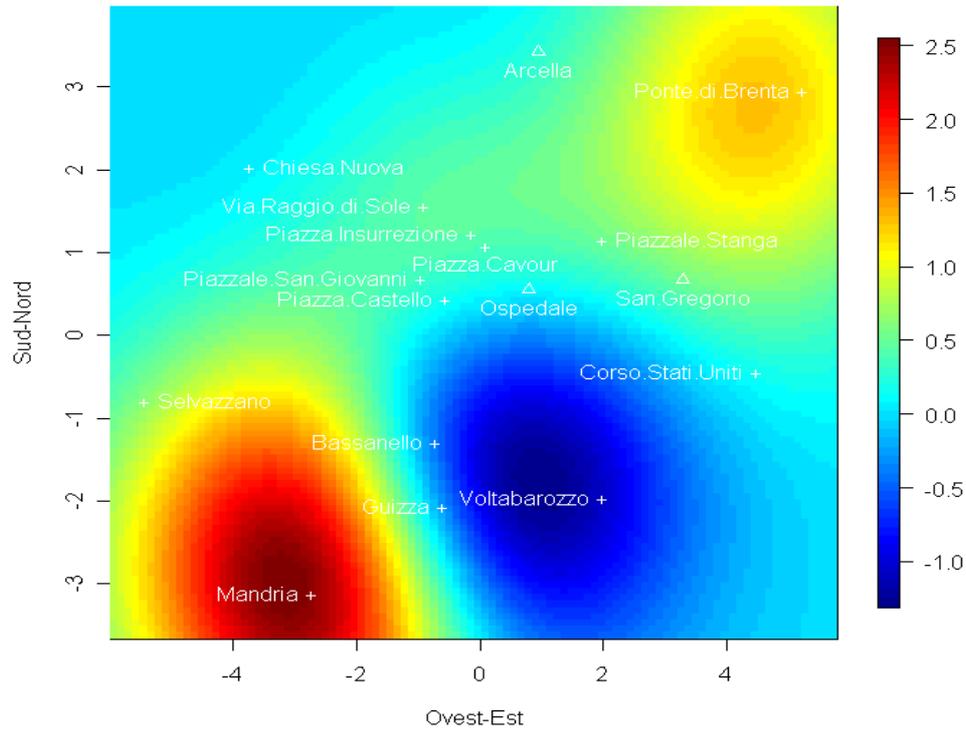


Figura 3.16: Concentrazione di biossido di zolfo nel Gennaio 1994.

Concentrazione del biossido di zolfo (1 Gennaio 1994)



Le previsioni che sono state ottenute con questo metodo, mettono in risalto quale sia il comportamento del fenomeno nello spazio di interesse rispetto alla media del periodo.

Nella zona relativa al centro di Padova e in quella a Sud-Ovest rispetto al centro stesso il livello di inquinamento sembra essere inferiore rispetto il livello generale di inquinamento, mentre la zona a Sud-Est di Padova sembra nello stesso lasso temporale sembra verificarsi il contrario, dove cioè l'inquinamento sembra essere quasi costantemente superiore rispetto il livello

medio del periodo; a Nord di Padova invece, ha luogo un comportamento altalenante della stessa entità nel periodo di interesse.

Per comprendere se il modello riesca a prevedere in maniera adeguata, la concentrazione di biossido di zolfo, si è deciso di mettere a confronto le rilevazioni di cui si dispone, con le previsioni ottenute per lo stesso punto. In primo luogo, si è pervenuti alla scelta di due siti di rilevazione come quello in zona Ospedale e in zona San Gregorio, in modo da capire quanto il modello riesca a produrre dati vicini a quelli reali, proprio dove disponiamo della quasi totalità dei dati della serie storica.

La scelta è stata fatta perché le serie in questione sono le più complete rispetto all'arco temporale di riferimento, sebbene ci siano comunque delle osservazioni mancanti.

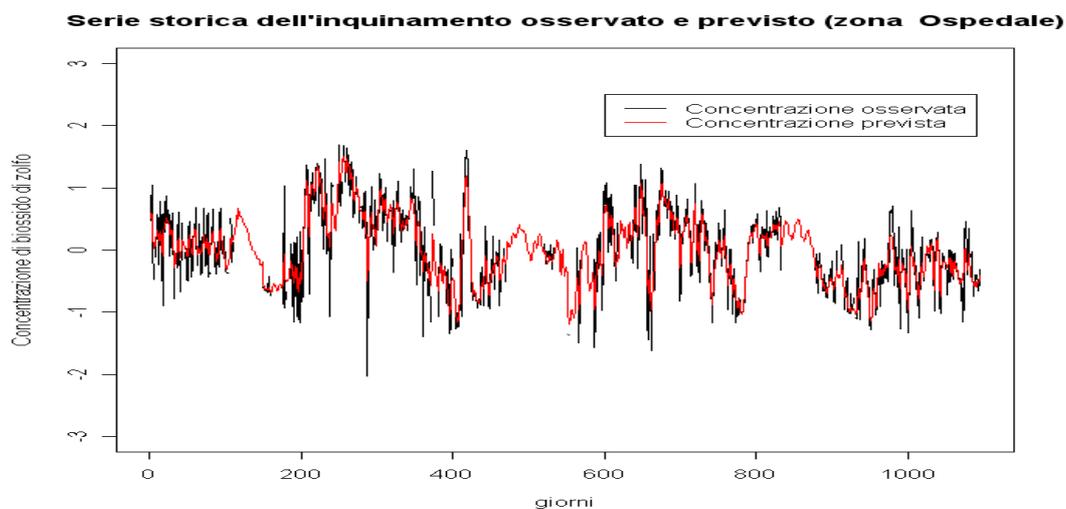


Figura 3.17: Concentrazione osservata e prevista di biossido di zolfo per zona Ospedale.

Serie storica dell'inquinamento osservato e previsto (San Gregorio)

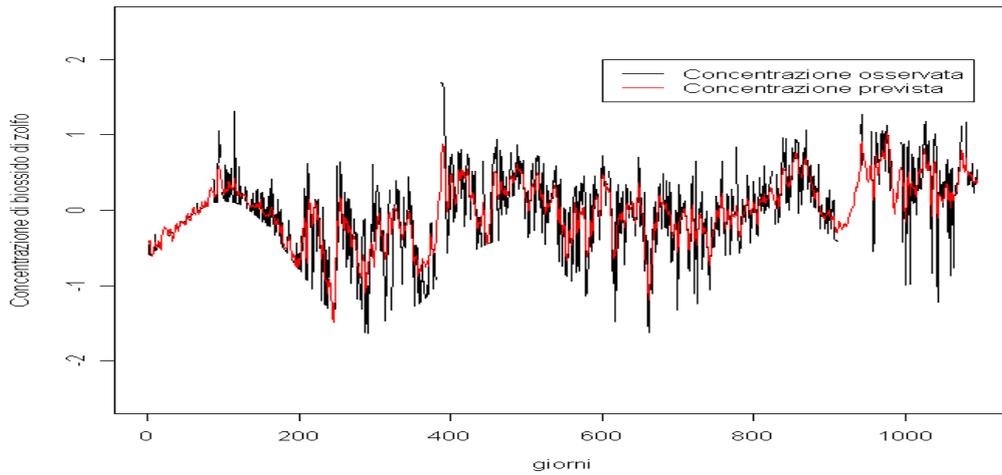


Figura 3.18: Concentrazione osservata e prevista di biossido di zolfo per San Gregorio.

Per i due luoghi di rilevazione si riportano anche gli andamenti degli errori del modello di previsione rispetto alle osservazioni e l'autocorrelazione degli stessi.

Residui zona Ospedale

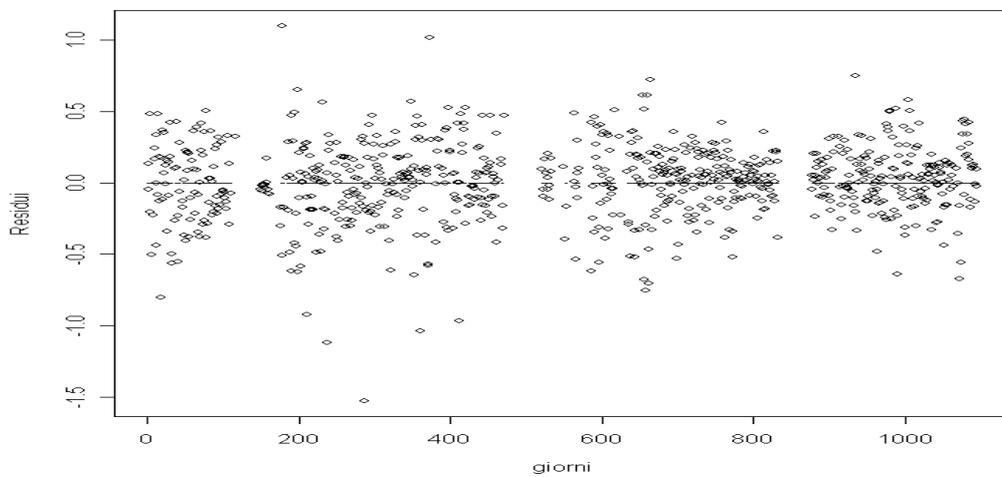


Figura 3.19: Andamento degli errori di previsione per zona Ospedale.

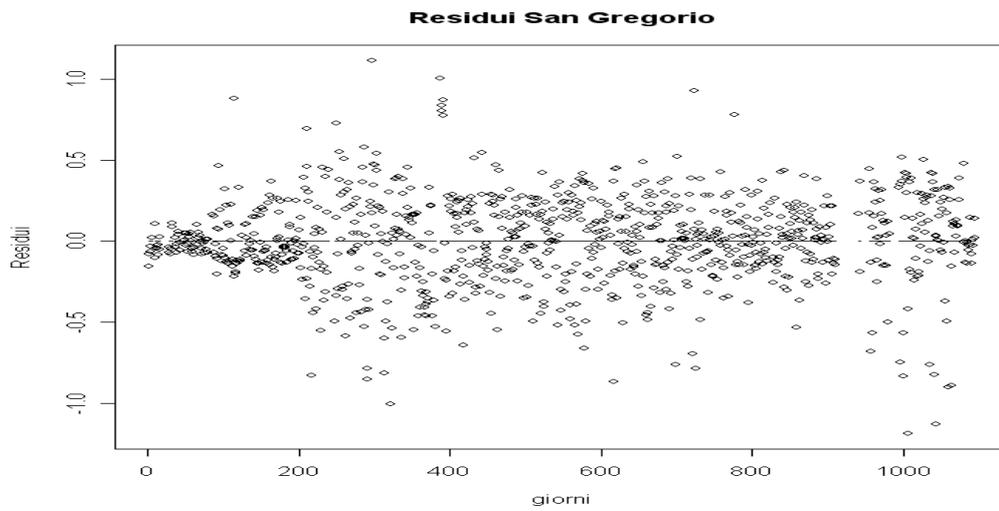


Figura 3.20: Andamento degli errori di previsione per San Gregorio.

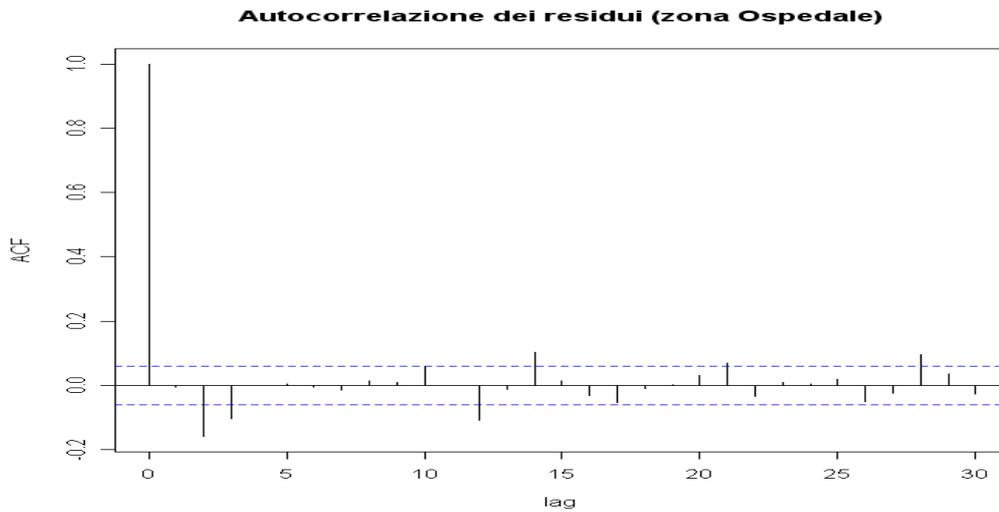


Figura 3.21: Autocorrelazione dei residui in zona Ospedale.

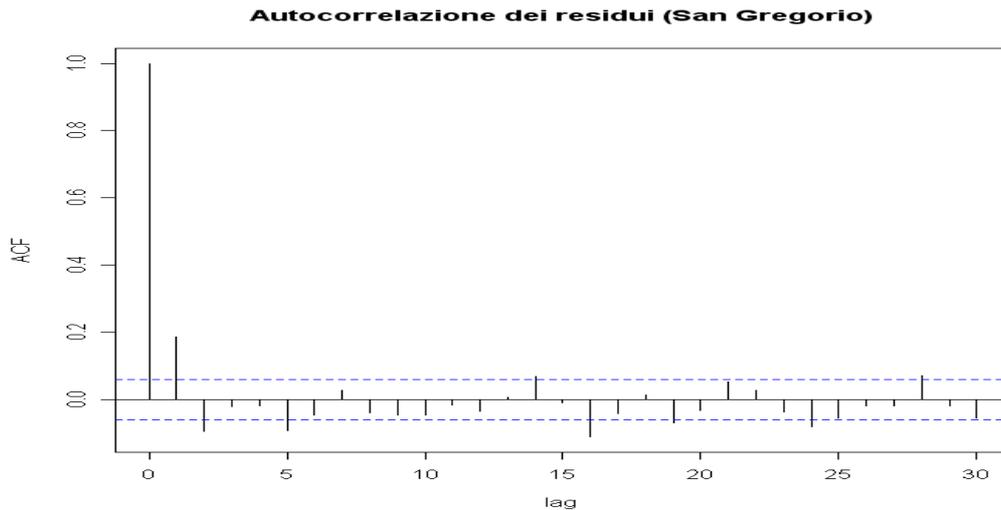


Figura 3.22: Autocorrelazione dei residui per San Gregorio.

Per entrambe le serie si nota come la capacità previsiva del modello sia, almeno a colpo d’occhio, abbastanza buona in queste due postazioni. Le previsioni date dal modello sembrano seguire ed approssimare in maniera soddisfacente le fluttuazioni dei dati destagionalizzati.

È possibile notare che, anche per i periodi in cui non sono presenti i dati, sono comunque state prodotte le previsioni per quello stesso luogo del piano di riferimento.

I residui sono relativamente bassi, e pressoché centrati attorno allo 0, tranne alcuni “*outliers*” che comunque non destano preoccupazione visto il loro numero esiguo.

Anche l’autocorrelazione degli stessi non sembra destare particolari allarmi rispetto alla bontà del modello, infatti, se escludiamo i ritardi di lunghezza 1, che comunque non sembrano essere particolarmente elevati, la situazione sembra rimarcare una buona capacità previsiva del modello.

Si è anche deciso di riportare lo stesso tipo di grafici per stazioni di

rilevazione in cui sono state presenti centraline di tipo mobile, in particolare per la zona di Ponte di Brenta (collocata a Nord-Est di Padova) e Selvazzano (Sud-Ovest) di Padova.

Data la limitatezza dell'arco temporale in cui sono state effettuate le rilevazioni (meno di 30 giorni) si è ritenuto opportuno restringere la rappresentazione ai giorni in cui sono state effettuate le rilevazioni.

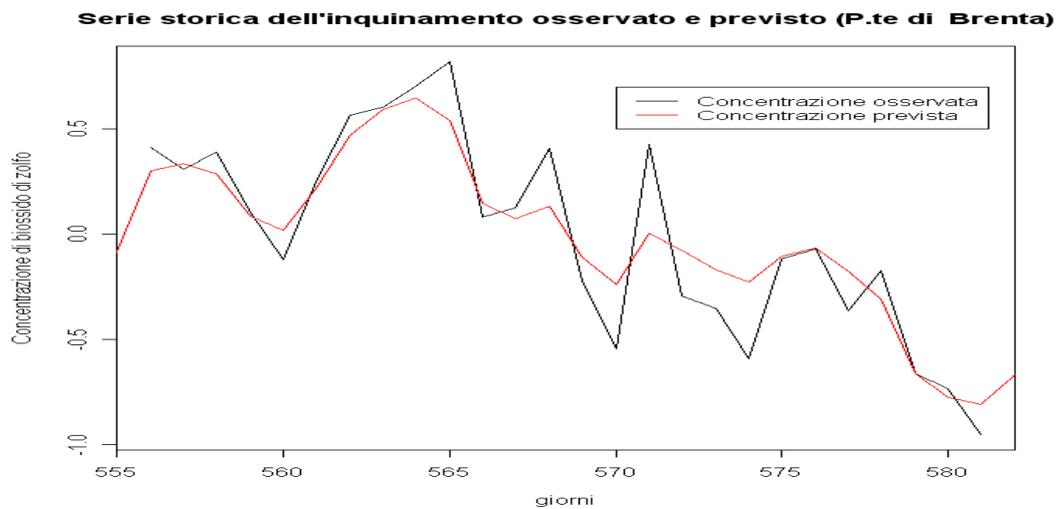


Figura 3.23: Concentrazione osservata e prevista di biossido di zolfo per Ponte di Brenta.

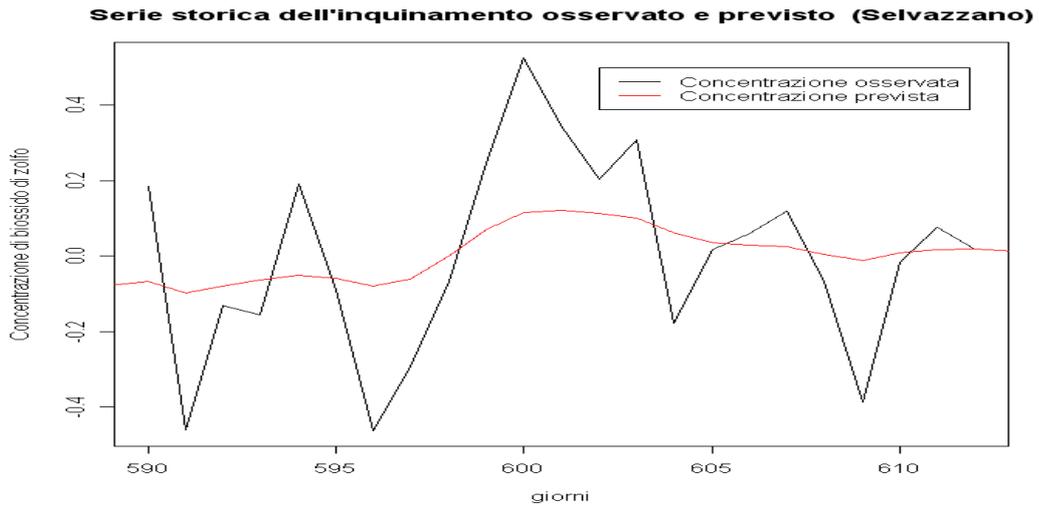


Figura 3.24: Concentrazione osservata e prevista di biossido di zolfo per Selvazzano.

Si riportano le serie dei residui e l'autocorrelazione degli stessi anche per Ponte di Brenta e Selvazzano.

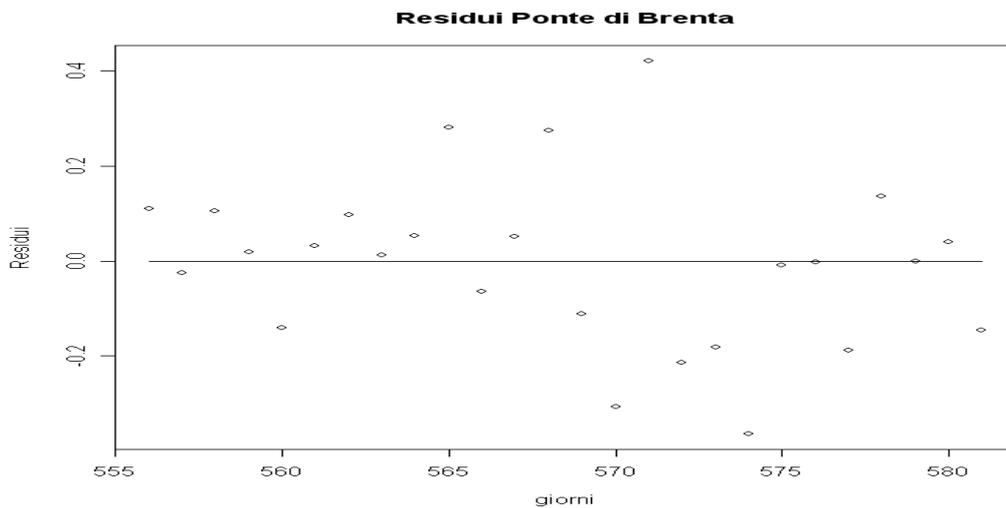


Figura 3.25: Andamento degli errori di previsione per Ponte di Brenta.

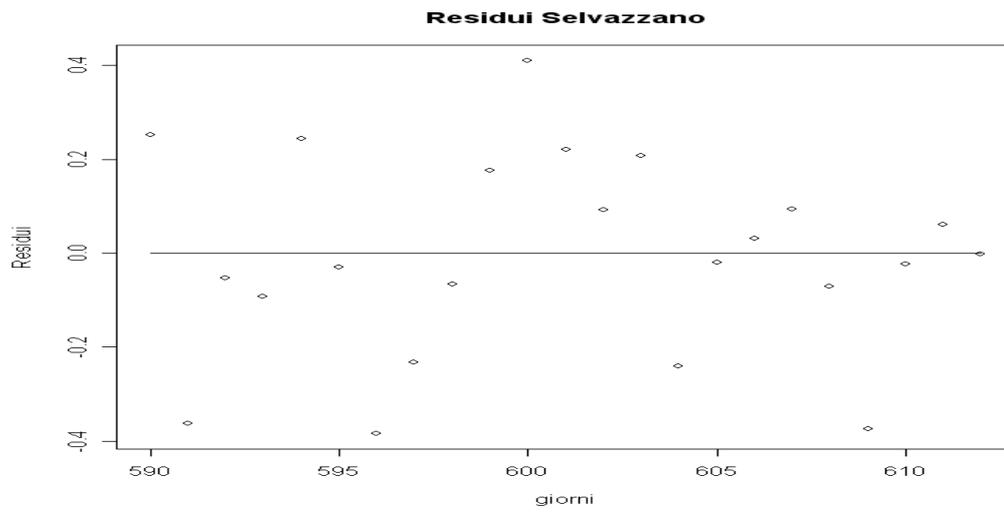


Figura 3.26: Andamento degli errori di previsione per Selvazzano.

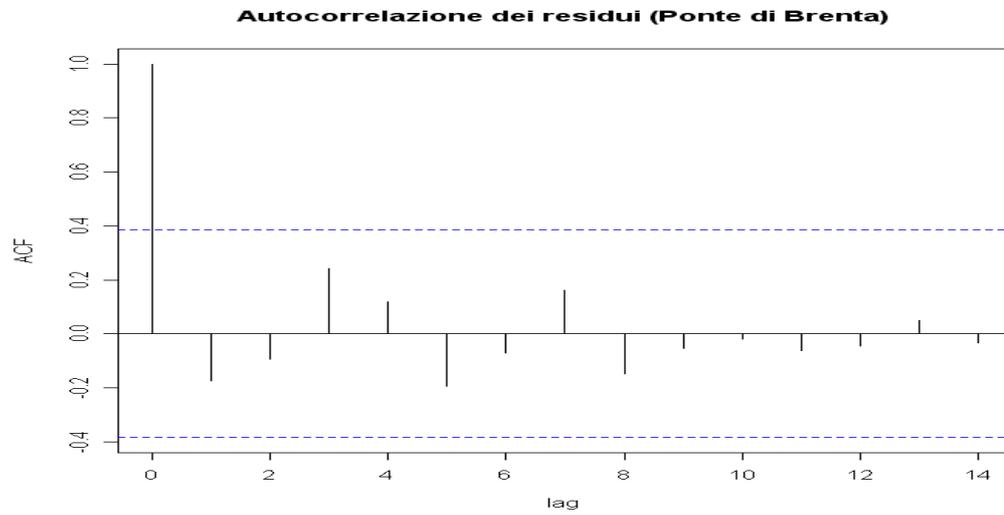


Figura 3.27: Autocorrelazione dei residui per Ponte di Brenta.

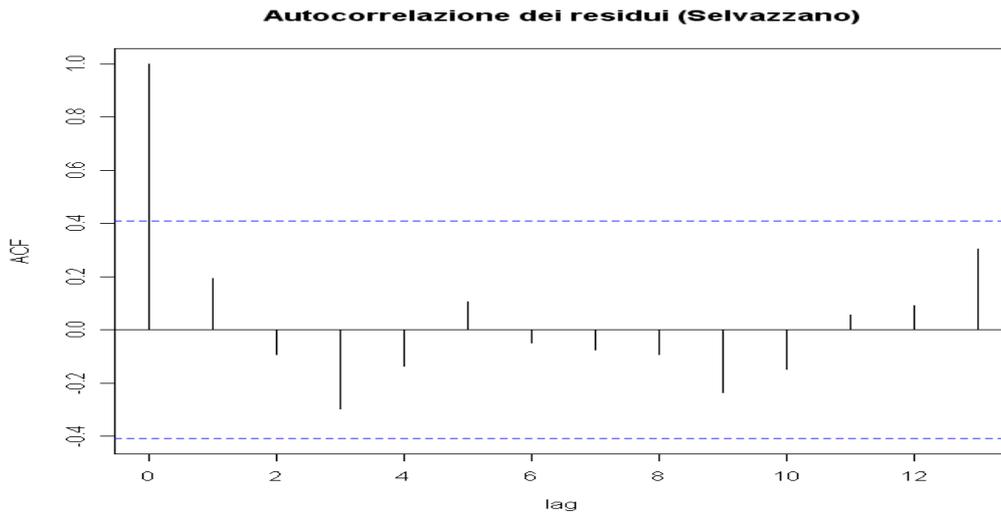


Figura 3.28: Autocorrelazione dei residui per Selvazzano.

Anche per queste due ultime stazioni di rilevazione il modello sembra adattarsi in maniera più che soddisfacente ai dati, i residui sembrano distribuirsi attorno al valore 0, e anche se per quelli di Ponte di Brenta attorno a metà del periodo di rilevazione l'omoschedasticità sembra violata, non si può negare la bontà del modello. L'autocorrelazione degli stessi residui per queste due stazione è ottima.

Si riportano gli stessi grafici per una zona molto meno periferica come Piazza Insurrezione. Per quanto concerne la serie storica e la sua previsione, oltre alla suddetta figura ristretto all'arco temporale di interesse, è presente anche un altro grafico che illustra le previsioni nell'intero periodo rilevazione messe a confronto con i dati osservati.

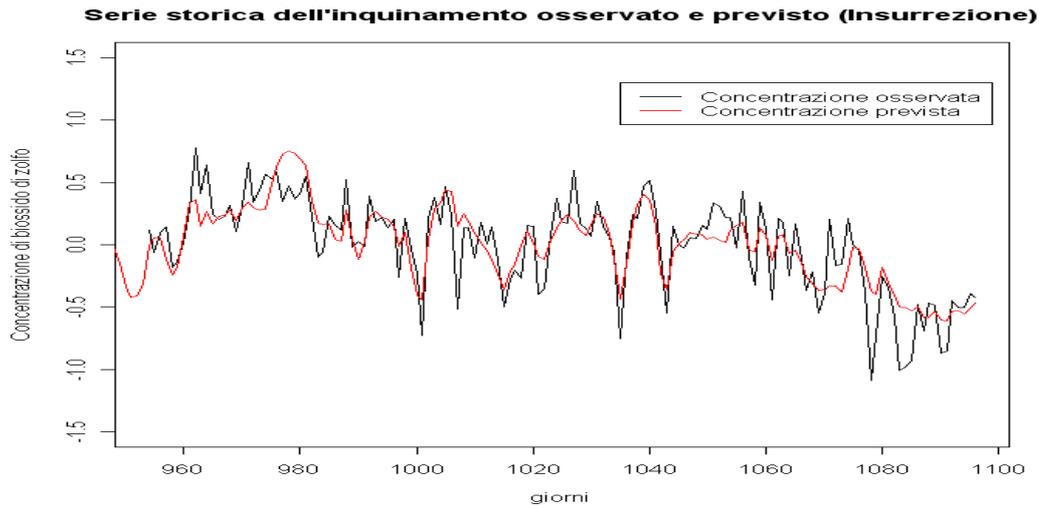


Figura 3.29: Concentrazione osservata e prevista di biossido di zolfo per Piazza Insurrezione.

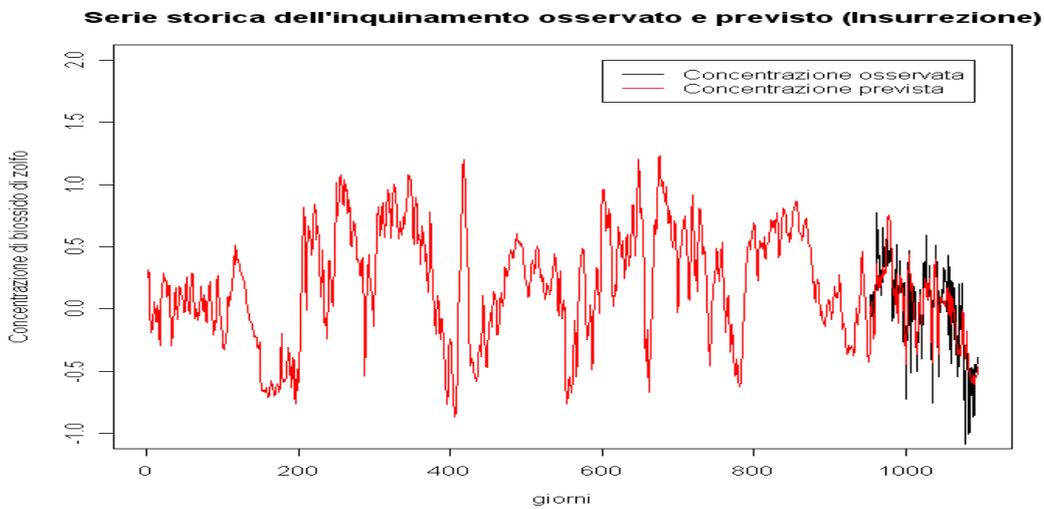


Figura 3.30: Concentrazione osservata e prevista di biossido di zolfo per Piazza Insurrezione (previsioni per tutto l'arco temporale).

Anche in questo caso le previsioni sembrano essere più precise e sembrano adattarsi in modo migliore alle osservazioni.

Si riporta di seguito anche l'andamento dei residui per Piazza Insurrezione.

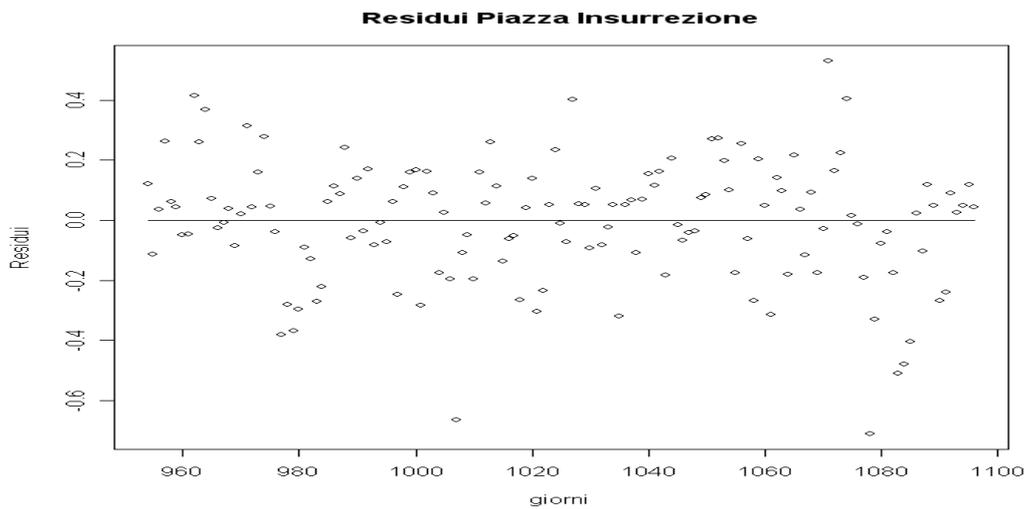


Figura 3.31: Andamento degli errori di previsione per Piazza Insurrezione.

Si riporta di seguito anche l'autocorrelazione dei residui.

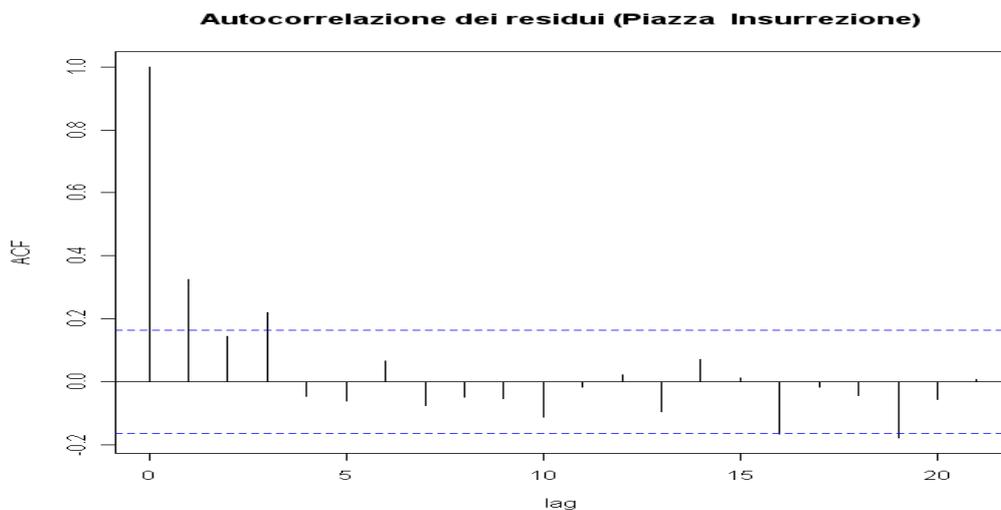


Figura 3.32: Autocorrelazione dei residui per Piazza Insurrezione.

L'autocorrelazione dei residui evidenzia al ritardo 2 e 4 una correlazione significativa, ma nel complesso non si può dire che questo vada ad inficiare su quanto di buono detto precedentemente sulla capacità previsiva del modello.

3.7 Brevi considerazioni finali

La destagionalizzazione operata con la creazione di un modello additivo nella fase preliminare ha contribuito ad eliminare la naturale fluttuazione dei dati, che vede un'aggravarsi della situazione complessiva dell'inquinamento tipicamente nei mesi invernali, ed ha inoltre contribuito ad attenuare l'influenza di rilevazioni contaminate da una tipicità di tipo spaziale rispetto al luogo in cui sono state rilevate.

Per quanto si è detto nei capitoli precedenti la scarsità e l'irregolarità, in senso temporale, di informazioni sono stati la vera motivazione che ha portato all'utilizzo di procedure come il filtro di Kalman e il metodo della convoluzione, applicati ai modelli di tipo spazio-temporale.

Il filtro è stato di particolare aiuto a studiare e a ricostruire il processo autoregressivo che si è supposto potesse soggiacere alla modellazione di un fenomeno, quale l'inquinamento, che potesse presupporre come matrice di regressione una serie ordinata di coefficienti ricostruiti con il metodo della convoluzione.

Il modello che si è costruito si è contraddistinto per una buona capacità previsiva su tutto lo spazio di interesse, sia nella zone dove sono state rilevate molte osservazioni (tipicamente vicino alle centraline fisse), sia nella periferia

e capace di cogliere le fluttuazioni temporali del fenomeno di interesse.

BIBLIOGRAFIA

- Aoki T. (1990), *State Space Modelling of Time-Series*, Springer-Verlag, Berlin.
- Azzalini A. (1983), Maximum likelihood estimation of order m for stationary stochastic processes, *Biometrika*, **70**, 381-87.

- Barry R.P. and Ver Hoef J.M. (1996), Blackbox Kriging: spatial prediction without specifying variogram models. *Journal of Agricultural, Biological and Environmental Statistics*, **1**, 297-322.
- Barry R.P. and Ver Hoef J.M. (1998), Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, **69**, 275-94.
- Brockwell P.J. and Davis R.A. (1991), *Time Series: Theory and Method*, Springer-Verlag, New-York.
- Brockwell P.J. and Davis R.A. (1996), *Introduction to Time Series and Forecasting*, Springer-Verlag, New-York.
- Clark I. (1979), *Practical Geostatistics*, Applied Science Publishers, Essex.
- Cressie N. (1985), Fitting variogram models by weighted least squares, *Journal of the International Association for Mathematical Geology*, **17**, 563-586.
- Cressie N. (1989), The many faces of spatial predictions. *Geostatistics*, **1**, 163-76.
- Cressie N. (1990), The origins of Kriging, *Mathematical Geology*, **22**, 239-52.
- Cressie N., Gotway C.A. e Grondona M.O. (1990), Spatial prediction from networks, *Chemometrics and Intelligent Laboratory Systems*, **7**, 251-71.
- Cressie N. (1993), *Statistics for Spatial Data*, John Wiley & Sons, New York.
- Cressie N. e Hawkins D.M. (1980), Robust estimation of the variogram, *Journal of the International Association for Mathematical Geology*, **12**, 115-125.

- Di Fonzo T. e Lisi F. (2001), *Complementi di statistica economica: analisi delle serie storiche univariate*, Cleup Editrice, Padova.
- Finzi G. e Brusasca G. (1991), *La qualità dell'aria, Modelli previsionali e gestionali*, Masson, Milano.
- Gelfand A.E. and Ecker M.D. (1997), Bayesian variogram modeling for an isotropic spatial process, *Journal of Agricultural, Biological and Environmental Statistics*, **2**, 347-69.
- Georgii H.W. (1969), The effects of air pollution on urban climates, *Bull. World Health Organization*, **40**.
- Haas T. (1995), Local prediction of a spatio-temporal process with an application to wet sulfate deposition, *Journal of the American Statistical Association*, **90**, 1189-99.
- Haas T. (1996), Multivariate spatial prediction in the presence of nonlinear trend and covariance nonstationarity, *Environmetrics*, **7**, 145-65.
- Harvey A.C. (1990), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.
- Harvey A.C. e Pierce R.G. (1984), Estimating missing observation in economic time series, *Journal of the American Statistical Association*, **79**, 320-40.
- Higdon D.M. (1998), A process convolution approach to modelling temperatures in north Atlantic ocean. *Journal of Environmental and Ecological Statistics*, **5**, 173-90.
- Higdon D.M., Swall J. and Kern J. (1999), Non stationary spatial modelling. *Bayesian Statistics 6. Proceedings of the Sixth Valencia International Meeting*, 761-68. Oxford University press, Oxford.
- Ickstadt K. and Wolpert R.L. (1999), Spatial regression for marked point

- process. *Bayesian statistics 6. Proceedings of the Sixth Valencia International Meeting*, 323-41.
- Lèvy P. (1964), *Processus stocastique et mouvement Brownien*, Gauthier-Villars, Paris.
 - Nelder J.A. and Mead R. (1965), A simplex algorithm for function minimization, *Computer Journal*, **7**, 148-54.
 - Perelli M. e Franzin F. (1984), Effetti delle piogge acide sui terreni e sui vegetali. *Atti del convegno "Città e ambiente, Inquinamento atmosferico"*, Padova, aprile 1984, 307-26.
 - Rowlingson, B. and Diggle, P. (1993) Splancs: spatial point pattern analysis code in S-Plus. *Computers and Geosciences*, **19**, 627-55.
 - Scarpa B. (1997), *Modelli per la concentrazione di inquinanti atmosferici combinando i dati di centraline fisse e mobili*, Tesi di dottorato, Padova.
 - Scarpa B. (2001), Analisi della concentrazione di SO₂ combinando i dati raccolti da centraline fisse e mobili: un modello state-space. *Statistica*, **61**, 707-30.
 - Seinfeld J.H. (1986), *Atmospheric Chemistry and Physics of air pollution*, John Wiley & Sons, New York.
 - Shumway R.H. and Stoffer D.S. (2000), *Time Series Analysis and its Applications*, Springer, New York.
 - Thiébaux H. and Pedder M.A. (1987), *Spatial Objective Analysis: with applications in atmospheric science*, San Diego: Academic Press.
 - Ver Hoef J., Cressie N. and Barry R. (2000), Flexible spatial models based on the fast Fourier transform (FFT) for cokriging. *Technical Report, Department of Statistics, Ohio state University, Ohio*.
 - Wolpert R. and Ickstadt K., (1998), Poisson/Gamma random fields models

for spatial statistics, *Biometrika*, **85**, 251-67.

Ringrazio

Vorrei ringraziare il prof. Carlo Gaetan per la grande disponibilità e competenza dimostrata nel corso della preparazione di questa tesi, ringrazio i miei genitori Angelo e Paola, mia sorella Sara, che mi sono stati d'aiuto e sprono per continuare e concludere il cammino universitario che mi ero prefisso di compiere,

ringrazio tutti i miei compagni universitari per aver condiviso momenti difficili e gioiosi di questi ultimi anni,
ringrazio i miei amici per avermi regalato un sorriso quando ne avevo bisogno.