Università degli Studi di Padova

Facoltà di Scienze Statistiche

Corso di Laurea in Statistica e Informatica

Tesi di Laurea

Modellazione spazio-temporale del livello delle polveri sottili nella pianura veneta

Relatore : Ch.mo Prof. Carlo Gaetan

Laureando : Daniele Levorato

Anno Accademico 2007/2008

Indice

Ringraziamenti

Si ringrazia l'Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto - Osservatorio Regionale Aria per la gentile disponibilità dei dati oggetto dell'analisi.

Un sincero ringraziamento al Professor Carlo Gaetan che con pazienza, partecipazione ed impegno mi ha guidato nella stesura di questa tesi.

Un semplice e profondo grazie, senza ombra di 'aleatorietà' e 'verosimiglianza', ad Alessia che ha condiviso con me le piccole vittorie e soddisfazioni e i molti momenti di tensione, sconforto e difficoltà.

Introduzione

L'inquinamento è un problema che coinvolge l'ambiente da molto tempo, danneggiandolo su vari fronti; da alcuni decenni si è cercato di contenere quanto creava inquinamento visibile - ad esempio i rifiuti urbani e i rifiuti industriali - controllandoli tramite depuratori, inceneritori, discariche gestite secondo normativa, smaltimento, riciclaggio ecc.

In questo ultimo periodo l'attenzione della pubblica opinione si è orientata verso tematiche di sensibilizzazione nei confronti dei problemi derivanti dall'inquinamento non immediatamente riscontrabile e visibile, come quello atmosferico, che desta preoccupazione per la salute degli esseri umani oltre che per l'ambiente in cui essi vivono.

In seguito a questa presa di coscienza, maturata già da tempo in ambito medico-scientifico, gli organi preposti hanno attivato e realizzato una serie di atti normativi destinati al rilevamento, al monitoraggio e all'intervento sui fattori che determinano questo tipo di pericolo.

Questa tesi sviluppa una analisi dei dati relativi all'anno 2006 della concentrazione di polveri sottili nell'atmosfera - ***P**articulate **M**atter* o PM_{10}, forniti dall'Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto, più brevemente ARPAV, ente deputato alla rilevazione degli stessi all'interno della Regione; propone, inoltre, una modellazione del fenomeno mediante tecniche di analisi dei processi spazio-temporali sviluppati nel corso degli ultimi anni.

Nel Capitolo 1 si introduce brevemente la problematica riguardante l'inquinamento ambientale per quanto concerne più specificatamente quello atmosferico, con i diversi agenti che lo originano e la normativa che a vari livelli - comunitario, nazionale e regionale -, negli ultimi anni, ha dato organicità

agli aspetti di limiti, di organizzazione per la rilevazione dei dati oltre che alle procedure e agli strumenti da utilizzare per la loro misurazione

Nel Capitolo 2 si presenta una descrizione dei dati forniti dall'ARPAV, una analisi esplorativa preliminare degli stessi e una prima 'indagine' rispetto alla dimensione temporale.

Nel Capitolo 3 viene sviluppata una analisi geostatistica, secondo la metodologia classica con cui vengono elaborati due modelli per la previsione del livello medio annuale della concentrazione di PM_{10}, tramite il *Kriging* ordinario e per la previsione della probabilità di superamento del livello di soglia, pari a 40 $\mu g/mc$, tramite *Kriging* di funzioni indicatrici.

Nel Capitolo 4 viene proposta una analisi dei dati prendendo in considerazione, oltre all'aspetto spaziale, anche l'aspetto temporale del fenomeno. I dati della concentrazione di inquinante vengono riorganizzati in medie settimali e trasformati mediante logaritmo, al fine di ottenere una distribuzione maggiormente simmetrica e con variabilità analoga nelle varie stazioni. In questa analisi si assume che il processo sia decomponibile in una parte deterministica stimata con tecniche non parametriche - fornite dai modelli additivi - funzione dello spazio, del tempo e di altre covariate, e di una parte aleatoria correlata solo spazialmente e non temporalmente. La prima viene interpretata come componente di larga scala che si evolve nel tempo, la seconda rappresenta la componente di piccola scala, aleatoria, nel solo dominio spaziale e che viene analizzata con gli stessi strumenti della geostatistica visti al capitolo precedente.

Il Capitolo 5 presenta alcuni recenti studi per i processi spazio-temporali che superano i limiti imposti dalla stazionarietà spaziale, come assunta nei Capitoli 3 e 4, e che consentono una modellazione del processo secondo una struttura gerachica bayesiana in cui i parametri del modello sono considerati come variabili casuali tramite la specificazione di una distribuzione a priori che consente di incorporare le conoscenze o le ipotesi del ricercatore. In modo particolare saranno presentati l'approccio di Sampson e Guttorp che consente di trattare processi non stazionari, e la modellazione proposta da Le e Zidek, adatta per la descrizione dei fenomeni di inquinamento atmosferico con dati osservati tramite una rete di rilevamento in cui possono essere presenti valori mancanti. Con l'ausilio della libreria di funzioni da essi stessi sviluppata, si analizzeranno i dati - come organizzati nel Capitolo 4 - per ottenere una previsione del fenomeno nello spazio e nel tempo.

Vengono infine riportate alcune conclusioni sulle analisi realizzate, evidenziati alcuni aspetti che hanno bisogno di un successivo approfondimento e indagine, nonchè alcune modalità ulteriori di analisi dei dati.

Le elaborazioni e i grafici sono stati ottenuti mediante l'utilizzo del *software* statistico `R` e degli opportuni *tools*, in particolare la libreria `mgcv` per l'utilizzo dei modelli additivi, la libreria `geoR` per l'analisi geostatistica dei dati e la libreria di funzioni messa a disposizione da Le e Zidek.

Capitolo 1

L'inquinamento ambientale

Con il termine inquinamento si intende una alterazione di una caratteristica ambientale causata, in particolare, da attività antropica. Il termine è quanto mai generico e comprende molti tipi di inquinamento.

Generalmente si parla di inquinamento quando l'alterazione ambientale compromette l'ecosistema danneggiando una o più forme di vita. Allo stesso modo si considerano atti di inquinamento quelli causati dall'uomo, e non quelli naturali come emissioni gassose naturali, ceneri vulcaniche, aumento della salinità del mare, ecc.

Per quanto riguarda le sostanze inquinanti, solitamente ci si riferisce a prodotti della lavorazione industriale (o dell'agricoltura industriale); tuttavia è bene ricordare che anche sostanze apparentemente innocue possono compromettere seriamente un ecosistema; per esempio sono considerate sostanze inquinanti latte o sale versati in uno stagno o in un corso d'acqua. Gli inquinanti, inoltre, possono essere sostanze presenti in natura e non frutto dell'azione umana. Infine ciò che risulta velenoso per una specie può essere vitale per un'altra: le prime forme di vita immisero nell'atmosfera grandi quantità di ossigeno come prodotto di scarto per esse velenoso.

Una forte presa di coscienza sui problemi causati dall'inquinamento industriale (ed in particolare dalle sostanze cancerogene) è avvenuta nel mondo occidentale a partire dagli anni settanta. Già negli anni precedenti tuttavia si erano manifestati i pericoli per la salute legati allo sviluppo industriale.

Esiste un tipo di inquinamento a livello locale e uno a livello globale. In passato si pensava che solo il primo costituisse un problema; così, per esempio, la combustione del carbone produce un fumo che in concentrazioni

sufficienti può essere un pericolo per la salute. La teoria era che quando l'inquinante fosse sufficientemente diluito non potesse causare danni. Negli ultimi decenni ci si è resi conto che alcuni tipi di inquinamento costituiscono un problema globale. E' importante riflettere sul fatto che già la consapevolezza dei due tipi di inquinamento potrebbe portare allo sviluppo di una cultura ambientalista diffusa tanto da riuscire a coniugare soluzioni economiche, benessere e qualità della vita, insieme ad un limite dell'impatto umano sull'ambiente.

1.1 L'inquinamento atmosferico

Per inquinamento atmosferico si intende l'aspetto che indica tutti gli agenti fisici (particolati), chimici e biologici che modificano le caratteristiche naturali dell'atmosfera.

Gli inquinanti vengono solitamente distinti in due gruppi principali: quelli di origine antropica, cioè prodotti dall'uomo, e quelli naturali.

Relativamente a quelli di origine antropica, finora sono stati catalogati circa 3.000 contaminanti dell'aria, prodotti per lo più dalle attività umane tramite processi industriali, l'utilizzo dei mezzi di trasporto o altre circostanze. Le modalità di produzione e di liberazione dei vari inquinanti sono estremamente varie; allo stesso modo sono moltissime le variabili che possono intervenire nella loro diffusione in atmosfera.

I contaminanti atmosferici, possono anche essere classificati in *primari* cioè liberati nell'ambiente come tali (come ad esempio il biossido di zolfo ed il monossido di azoto) e *secondari* (come l'ozono), che si formano successivamente in atmosfera attraverso reazioni chimico-fisiche.

I principali inquinanti possono essere suddivisi in due classificazioni: gas e particolati. Nella prima classe sono compresi i contaminanti gassosi come:

- *monossido di carbonio (CO)*: emesso principalmente dai processi di combustione e in particolare dagli scarici dei veicoli con motore a idrocarburi. Le concentrazioni maggiori si possono rilevare lungo le arterie stradali. Nel caso di inalazione in gran quantità si riscontrano nei soggetti sintomi come mal di testa, affaticamento e problemi respiratori; è importante evidenziare che sopra una certa soglia diventa letale per gli esseri umani;
- *anidride carbonica (CO_2)*: anche questo gas viene emesso principal-

mente dai processi di combustione, in modo particolare dagli scarichi dei veicoli con motore a idrocarburi, escluso il metano;

• *clorofluorocarburi (CFC)*: sostanze chimiche di sintesi presenti nei refrigeranti dei frigoriferi o negli impianti di condizionamento dell'aria, nei propellenti delle bombolette spray, negli agenti schiumogeni per la produzione di imballaggi e nei prodotti chimici usati per lo spegnimento di incendi. Queste sostanze, pur non avendo degli effetti immediati sull'uomo, a causa delle reazioni chimiche che si sviluppano nella stratosfera, procurano un danno allo strato di ozono dell'atmosfera, che consente di garantire una adeguata protezione dalla radiazione solare;

• *idrocarburi*: provengono sempre dai processi di combustione del traffico veicolare basato sui derivati del petrolio e dal fenomeno dell'evaporazione;

• *piombo e altri metalli pesanti*: sono legati alle attività industriali e di combustione che ne causano la dispersione nell'ambiente. Benchè siano elementi naturalmente presenti nell'ecositema, la loro mobilitazione determinata dalle attività umane ne causa l'accumulo nella biosfera e l'ingresso nella catena alimentare con gravi danni per piante, animali e quindi per l'uomo;

• *ossidi di azoto (NO_x)*: provengono dalla combustione dei fornelli, fumo di sigaretta, fumo di caminetti, scarichi di auto, da lavorazioni industriali e processi di combustione ad elevate temperature. Gli ossidi di azoto reagiscono inoltre con gli idrocarburi presenti nell'atmosfera generando smog fotochimico. Oltre a comportare irritazione agli occhi e problemi all'apparato respiratorio, depositandosi in siti ecologicamente sensibili possono provocare processi di acidificazione e eutrofizzazione dell'ambiente;

• *biossido di zolfo*: viene generato dalla combustione di carburanti contenenti zolfo, principalmente nelle centrali elettriche, durante la fusione di metalli e in altri processi industriali. E' la causa delle cosiddette "piogge acide";

• *ozono*: può essere presente negli strati inferiori dell'atmosfera come inquinante secondario formato da reazioni fotochimiche che coinvolgono gli ossidi di azoto e i composti organici volatili. Infatti, a questi livelli dell'atmosfera risulta un gas irritante che provoca problemi all'apparato respiratorio;

• *composti organici volatili*: con questa dicitura vengono inclusi diversi composti chimici organici, tra cui il benzene, provenienti da vernici, solventi, prodotti per la pulizia e da alcuni idrocarburi. Il benzene è riconosciuto

essere un agente cancerogeno, mentre gli altri sono tra le cause dell'effetto serra.
Alla seconda classe, i "particolati", appartengono invece le piccole particelle solide classificate in base alle loro dimensioni. Le particelle atmosferiche sono solitamente misurate in PTS (**P**olveri **T**otali **S**ospese), le polveri sottili denominate PM (dall'inglese *Particulate Matter*), con la specificazione del diametro aereodinamico PM_{10} per particelle con diametro inferiore ai 10 micron e $PM_{2,5}$ per particelle con diametro inferiore ai 2,5 micron. Ultimamente l'attenzione si sta focalizzando sulle particelle ancora più piccole, le cosidette *nanopolveri*, che tendono ad essere maggiormente pericolose per la salute umana a causa della capacità di rimanere sospese nell'aria per periodi di tempo più lunghi rispetto alle particelle di maggiori dimensioni.

1.2 Le polveri sottili (*Particulate Matter* - PM)

Le particelle sospese sono sostanze allo stato solido o liquido che, a causa delle loro piccole dimensioni, restano sospese in atmosfera per tempi più o meno lunghi. Il particolato nell'aria può essere costituito da diverse sostanze: sabbia, ceneri, polveri, fuliggine, sostanze silicee di varia natura, sostanze vegetali, composti metallici, fibre tessili naturali e artificiali, sali, elementi come il carbonio o il piombo, ecc.

In base alla natura e alle dimensioni delle particelle possiamo distinguere:
- gli *aerosol*, costituiti da particelle solide o liquide sospese in aria e con un diametro inferiore a 1 micron (1 μm);
- le *foschie*, date da goccioline con diametro inferiore a 2 micron;
- le *esalazioni*, costituite da particelle solide con diametro inferiore ad 1 micron e rilasciate solitamente da processi chimici e metallurgici;
- il *fumo*, dato da particelle solide normalmente con diametro inferiore ai 2 μm e trasportate da miscele di gas;
- le *polveri* (vere e proprie), costituite da particelle solide con diametro fra 0,25 e 500 micron;
- le *sabbie*, date da particelle solide con diametro superiore ai 500 μm.

Le particelle primarie sono quelle che vengono emesse come tali dalle sorgenti naturali ed antropiche, mentre le secondarie si originano da una serie

di reazioni chimiche e fisiche in atmosfera. Le particelle fini sono quelle che hanno un diametro inferiore a 2,5 μm, le altre sono dette "grossolane" e sono costituite esclusivamente da particelle primarie. Le polveri PM_{10} costituiscono il particolato che ha un diametro inferiore a 10 micron (10 μm) - un milionesimo di dm, mentre le $PM_{2,5}$, che rappresentano circa il 60% delle PM_{10}, costituiscono il particolato che ha un diametro inferiore a 2,5 micron.

Vengono dette polveri inalabili quelle in grado di penetrare nel tratto superiore dell'apparato respiratorio (dal naso alla laringe). Le polveri toraciche sono quelle in grado di raggiungere i polmoni. Le polveri respirabili possono invece penetrare nel tratto inferiore dell'apparato respiratorio (dalla trachea fino agli alveoli polmonari).

1.2.1 Origine

Le polveri si originano sia da fonti naturali che antropogeniche. Le polveri fini derivano principalmente da processi di combustione (particolato primario, cioè prodotto direttamente) e da prodotti di reazione dei gas (particolato secondario); la frazione grossolana delle polveri si origina in genere da processi meccanici (solo p. primario).

Le principali cause naturali di particolato primario sono le eruzioni vulcaniche, gli incendi boschivi, l'erosione e la disgregazione delle rocce, le piante (pollini e residui vegetali), le spore, lo spray marino e i resti degli insetti.

Il particolato naturale secondario è costituito da particelle fini che si originano in seguito a processi di ossidazione che producono sostanze quali: il biossido di zolfo e l'acido solfidrico, emessi dagli incendi e dai vulcani; gli ossidi di azoto liberati dai terreni; i terpeni (idrocarburi) emessi dalla vegetazione.

Il particolato primario di origine antropica è invece dovuto: all'utilizzo dei combustibili fossili (riscaldamento domestico, centrali termoelettriche, ecc.); alle emissioni degli autoveicoli; all'usura dei pneumatici, dei freni e del manto stradale; a vari processi industriali (fonderie, miniere, cementifici, ecc.). Inoltre grandi quantità di polveri si possono originare in seguito alle attività agricole.

Le polveri secondarie antropogeniche sono invece dovute essenzialmente all'ossidazione degli idrocarburi e degli ossidi di zolfo e di azoto emessi dalle

attività umane.

Spesso il particolato rappresenta l'inquinante a maggiore impatto ambientale nelle aree urbane, tanto da indurre le autorità competenti a disporre dei blocchi del traffico al fine di controllare e ridurre la manifestazione del fenomeno visto che i suoi effetti si riscontrano sull'uomo e sull'ambiente.

1.2.2 Effetti sull'uomo

A prescindere dalla tossicità, le particelle che possono produrre degli effetti indesiderati sull'uomo sono sostanzialmente quelle di dimensioni più ridotte, poiché nel processo della respirazione le particelle maggiori di 15 micron vengono generalmente espulse dal naso. Il particolato che si deposita nel tratto superiore dell'apparato respiratorio (cavità nasali, faringe e laringe) può generare effetti irritativi come l'infiammazione e la secchezza del naso e della gola; tutti questi fenomeni sono molto più gravi se le particelle hanno assorbito sostanze acide (come il biossido di zolfo, gli ossidi di azoto, ecc.).

Per la particolare struttura della superficie, le particelle possono anche adsorbire dall'aria sostanze chimiche cancerogene, trascinandole nei tratti respiratori e prolungandone i tempi di residenza, così da accentuarne gli effetti. Le particelle più piccole penetrano nel sistema respiratorio a varie profondità e possono trascorrere lunghi periodi di tempo prima che vengano rimosse, ed è per questo che sono le più pericolose. Queste polveri aggravano le malattie respiratorie croniche come l'asma, la bronchite e l'enfisema.

Le persone più vulnerabili sono gli anziani, gli asmatici, i bambini e chi svolge un'intensa attività fisica all'aperto, sia di tipo lavorativo che sportivo. Nei luoghi di lavoro più soggetti all'inquinamento da particolato l'inalazione prolungata di queste particelle può provocare reazioni fibrose croniche e necrosi dei tessuti che comportano una broncopolmonite cronica accompagnata spesso da enfisema polmonare.

Per il particolato, studi epidemiologici hanno dimostrato una correlazione tra incremento dei livelli di particolato nell'inquinamento atmosferico e alterazioni del ritmo cardiaco; in altri studi epidemiologici, a partire dagli anni '50, è stata stabilita l'associazione tra incremento delle concentrazioni

di particolato e aumento, nei giorni successivi all'episodio, sia della mortalità, soprattutto per cause respiratorie, sia del numero di ricoveri ospedalieri.

1.2.3 Effetti sull'ambiente

Gli effetti del particolato sul clima e sui materiali sono piuttosto evidenti. Il particolato dei fumi e delle esalazioni provoca una diminuzione della visibilità atmosferica; allo stesso tempo diminuisce anche la luminosità assorbendo o riflettendo la luce solare.

Negli ultimi 50 anni si è notata una diminuzione della visibilità del 50%, ed il fenomeno risulta tanto più grave quanto più ci si avvicina alle grandi aree abitative ed industriali. Le polveri sospese favoriscono la formazione di nebbie e nuvole, costituendo i nuclei di condensazione attorno ai quali si concentrano le gocce d'acqua. Di conseguenza favoriscono il verificarsi dei fenomeni delle nebbie e delle piogge acide, che comportano effetti di erosione e corrosione dei materiali e dei metalli. Il particolato inoltre danneggia i circuiti elettrici ed elettronici, può generare un degrado degli edifici e delle opere d'arte e riduce la durata dei tessuti. Le polveri (ad esempio quelle emesse dai cementifici), possono depositarsi sulle foglie delle piante e formare così una patina opaca che, schermando la luce, ostacola il processo della fotosintesi.

Gli effetti del particolato sul clima della terra sono invece piuttosto discussi. Sicuramente un aumento del particolato in atmosfera comporta una diminuzione della temperatura terrestre per un effetto di riflessione e schermatura della luce solare, in ogni caso tale azione è comunque mitigata dal fatto che le particelle riflettono anche le radiazioni infrarosse provenienti dalla terra. E' stato comunque dimostrato che negli anni immediatamente successivi alle più grandi eruzioni vulcaniche di tipo esplosivo - caratterizzate quindi dalla emissione in atmosfera di un'enorme quantità di particolato - sono seguiti degli anni con inverni particolarmente rigidi.

1.3 La Normativa

Appare evidente come sia stato urgente e fondamentale da parte degli organi politici, amministrativi e sanitari produrre normative e direttive che nel tempo si sono sempre più specificate, talvolta con limiti maggiormente restrittivi; questo al fine di monitorare sia i problemi generati, sia i livelli

raggiunti dagli inquinanti atmosferici, tutelando e riducendo i rischi per la salute e per l'ambiente in cui viviamo.

Il Decreto del Presidente del Consiglio dei Ministri del 28 marzo 1983 fissa i valori limite per le particelle sospese: la media aritmetica delle concentrazioni medie nelle 24 ore rilevate nell'arco di un anno ha il valore limite pari a 150 $\mu g/mc$; il 95° percentile delle concentrazioni medie nelle 24 ore rilevate nell'arco di un anno ha il valore limite pari a 300 $\mu g/mc$.

Il DPR n. 203 del 24 maggio 1988 prevede dei valori guida per le particelle sospese: la media aritmetica delle concentrazioni medie nelle 24 ore rilevate nell'arco di 1 anno ha il valore guida di 40-60 FN equiv/mc; il valore medio nelle 24 ore ha il valore guida di 100-150 FN equiv/mc.

Il Decreto Ministeriale del 25/11/94 fissa il livello di attenzione ed il livello di allarme per quanto riguarda le particelle sospese totali nelle aree urbane: considerando la media dei valori rilevati nell'arco di 24 ore, il livello di attenzione è fissato in 150 $\mu g/mc$, mentre il livello di allarme è posto a 300 $\mu g/mc$. Il DM 25/11/94 prevede anche il monitoraggio della frazione respirabile delle polveri sospese (PM_{10}), prefissando come obiettivo di qualità il valore di 40 $\mu g/mc$ (da raggiungere a partire dal primo gennaio 1999). Il DM 21/04/99 individua i criteri ambientali e sanitari in base ai quali i Sindaci possono limitare la circolazione degli autoveicoli per migliorare la qualità dell'aria nelle aree urbane.

Nel D.M. 60/2002 (art.17, All. III) vengono fissati i valori limite per la protezione della salute umana, rispettivamente in 40 $\mu g/mc$ per la media annuale e 50 $\mu g/mc$ per la media giornaliera da non superare più di 35 volte in un anno, cioè all'incirca il 10% dei giorni in un anno, il che equivale a fissare una soglia di 50 $\mu g/mc$ per il quantile di ordine 0.904. Il Decreto prevede nella Fase 2, a partire dal 2005, che i valori limite diventino rispettivamente 20 $\mu g/mc$ per la media annuale e 50 $\mu g/mc$ per la media giornaliera da non superare più di 7 volte in un anno (cioé la soglia dovrà essere posta sul quantile di ordine 0.981), da raggiungere entro l'1 gennaio 2010, precisando tuttavia che trattasi di valori limite indicativi da rivedere con successivo decreto sulla base della futura normativa comunitaria. Il metodo di riferimento per il campionamento e la misurazione del PM_{10} stabilito dal D.M. 60/2002 è descritto nella norma EN 12341 "*Air quality - Determination of*

the PM_{10} fraction of suspended particulate matter Reference method and field test procedure to demonstrate reference equivalence of measurement methods". Il principio di misurazione si basa sulla raccolta, attraverso un filtro, delle polveri e sulla determinazione della loro massa per via gravimetrica. I test indicati nella norma EN 12341 sono test di riferimento, mentre metodi e sistemi di campionamento e misura diversi, sia manuali sia automatici, devono essere dotati di certificazione di equivalenza rilasciata da enti designati seguendo apposite procedure operative, descritte in Appendice all'Allegato XI. Il D.M. 60/2002 precisa altresì che le regioni sono obbligate a fornire le necessarie giustificazioni durante la trasmissione delle informazioni al Ministero dell'ambiente e della tutela del territorio e al Ministero della salute, per il tramite dell'ANPA. Le regioni devono anche provvedere affinché il pubblico e le categorie interessate siano informati sui livelli di materiale particolato nell'aria oltre che organizzarne l'aggiornamento con frequenza giornaliera. I requisiti per la raccolta minima dei dati e per il periodo minimo di copertura sono rispettivamente il 90% e il 14%, dove quest'ultima percentuale rappresenta la quantità di dati ottenuti con una misurazione in un giorno, scelto a caso, di ogni settimana in modo che le misure siano uniformemente distribuite durante l'anno, oppure mediante 8 settimane di misurazione distribuite in modo regolare nell'arco dell'anno. La Decisione del Consiglio n. 1997/101/CE, con la relativa modifica n. 2001/752/CE, precisa i criteri per l'aggregazione dei dati e per il calcolo dei parametri statistici da comunicare. In particolare i criteri per il calcolo dei valori orari e giornalieri a partire da dati con tempi medi inferiori sono:

- per i valori orari: dati minimi da rilevare: 75%;
- per i valori giornalieri: almeno 13 valori orari disponibili e non più di sei valori orari successivi mancanti;

mentre per il calcolo dei parametri statistici:

- per la media e la mediana: dati minimi da rilevare: 50%;
- per i percentili 98, 99.9 e il massimo: dati minimi da rilevare: 75%.

Inoltre il rapporto tra il numero dei dati validi per le due stagioni dell'anno prese in considerazione non può essere superiore a 2; le due stagioni sono l'inverno (da gennaio a marzo compreso e da ottobre a dicembre compreso)

e l'estate (da aprile a settembre compreso). Gli Allegati VIII e IX del D.M. 60/2002 indicano anche i criteri per l'ubicazione e il numero minimo di punti di campionamento per la misurazione in siti fissi dei livelli degli inquinanti nell'aria.

Nella regione Veneto, con deliberazione n. 902 del 4 aprile 2003, la Giunta Regionale ha adottato il Piano Regionale di Tutela e Risanamento dell'Atmosfera, in ottemperanza a quanto previsto dalla legge regionale 16 aprile 1985, n. 33 e dal Decreto legislativo 351/99. Il Piano Regionale di Tutela e Risanamento dell'Atmosfera è stato infine approvato in via definitiva dal Consiglio Regionale con deliberazione n. 57 dell'11 novembre 2004. Con deliberazione n. 1408 del 16 maggio 2006 la Giunta Regionale ha approvato un Piano Progressivo di Rientro relativo alle polveri PM_{10}.

L'Ente preposto a dar seguito a queste normative dal punto di visto tecnico-realizzativo è l'ARPAV, con l'Osservatorio Regionale Aria, che opera per la prevenzione e promozione della salute collettiva, perseguendo l'obiettivo dell'utilizzo integrato e coordinato delle risorse, al fine di conseguire la massima efficacia nell'individuazione e nella rimozione dei fattori di rischio per l'uomo e per l'ambiente. (art. 1, comma 2, Legge Regionale 18 ottobre 1996, n° 32 istitutiva dell'ARPAV). Nel sito dell'Ente [1] è possibile consultare una pagina Web "Qualità dell'aria - PM_{10} Dati in diretta" dove si possono visionare le concentrazioni di PM_{10} registrate quotidianamente nei Comuni capoluogo del Veneto; è presente inoltre una pagina in cui vengono visualizzate le previsioni, secondo una scala nominale, per i giorni successivi.

1.3.1 Gli strumenti di misurazione

Come accennato sopra, la misurazione del PM_{10} può essere effettuata con diverse tipologie di strumenti ma il D.M.60/2002 e la norma EN 12341 stabiliscono che il metodo di riferimento è quello gravimetrico. In Appendice all'Allegato XI del D.M.60/2002 si prevede: "Il valore di concentrazione di massa del materiale particolato è il risultato finale di un processo che include la separazione granulometrica della frazione PM_{10}, la sua accumulazione sul mezzo filtrante e la relativa misura di massa con il metodo gravimetrico. Un sistema di campionamento, operante a portata volumetrica costante in ingresso, preleva aria, attraverso un'appropriata testa di campionamento e un successivo separatore a impatto inerziale. La frazione PM_{10} così ottenuta

viene trasportata su un mezzo filtrante a temperatura ambiente. La determinazione della quantità di massa PM_{10} viene eseguita calcolando la differenza fra il peso del filtro campionato e il peso del filtro bianco". Tale metodo prevede quindi una operazione di pesata del filtro su cui si è precedentemente accumulato il particolato atmosferico, da cui deriva il valore di concentrazione delle polveri PM_{10}. La necessaria fase preliminare di condizionamento del filtro (portato a 20+1°C e 50+3% di umidità per 48 ore prima del campionamento) e di nuovo immediatamente prima delle operazioni di pesata comporta alcuni giorni di ritardo, solitamente tre, nell'ottenimento del dato.

In sostituzione o accanto a tale metodo manuale possono essere utilizzati dei metodi automatici dotati di certificazione di equivalenza, come specificato dal D.M. 60/02, Allegato XI, punto 2.
Gli strumenti BETA, misurano l'attenuazione di particelle prodotte da una sorgente radioattiva (generalmente ^{14}C o ^{147}Pm) da parte del campione su cui è depositato il particolato. La misura è relativa, vale a dire che viene valutata la differenza tra l'attenuazione del fascio attraverso il filtro bianco e successivamente quella determinata dal particolato atmosferico raccolto sul filtro campionato. Più specificatamente questo tipo di strumenti possono differenziarsi in relazione al tipo di campionamento su *filtro* o su *nastro*.

Capitolo 2

I dati e l'analisi esplorativa

2.1 Studi sperimentali e studi osservazionali

Prima di procedere con la descrizione e l'analisi dei dati oggetto della presente tesi, è importante accennare alla possibile diversità di approccio tra gli studi sperimentali e quelli osservazionali.

Nel primo caso lo statistico, o in generale colui che a vario titolo affronta la fase di analisi dei dati ha potuto, anticipatamente, avere una conoscenza (spesso coadiuvato da uno specialista della materia) del problema; ha potuto programmare la fase di sperimentazione nei modi, nei tempi, nel tipo di campionamento, nella formulazione di ipotesi e/o teorie a priori e di conseguenza ha strutturato la modalità per procedere all'esperimento stesso, alla raccolta dei dati e quindi alla modalità di analisi in maniera "controllata". In altre parole, egli agisce con il controllo e la pianificazione dei fattori sperimentali.

Nel secondo caso molto spesso non si è in presenza di uno studio progettato a monte, ma i dati provengono - proprio come nel caso oggetto di questa tesi - da una rete di stazioni di rilevamento del fenomeno che, solitamente, rispondono ad esigenze di pianificazione della raccolta e conoscenza di dati con finalità a carattere legislativo-conoscitivo e quindi, talvolta, non conformi ad un criterio di campionamento statisticamente inteso; essi possono presentare gli stessi fenomeni misurati con strumenti diversi, o con modalità non omogenee in quanto provenienti da strutture organizzative diverse e/o autonome. Rispetto agli studi sperimentali, questo porta ad una serie di problematiche

aggiuntive, soprattutto per quanto riguarda l'aspetto della distorsione dei risultati, che rendono più difficoltosa la parte di analisi statistica; possono inoltre presentare una diversità di scopi e di analisi per cui sono stati raccolti.

I dati in ambito ambientale, come nel nostro caso, si caratterizzano inoltre per avere carattistiche non standard dovute alla variabilità sia nello spazio sia nel tempo, nell'avere misurazioni non equispaziate ed essere caratterizzati da una significativa presenza di valori mancanti; altre problematiche possono essere dovute al fatto che il fenomeno oggetto di misurazione venga spiegato dalla teoria o da ipotesi che prevedono variabili o misurazioni che possono non essere rilevabili o disponibili.

2.2 La rete di rilevamento

I dati messi a disposizione per questo elaborato si riferiscono al livello di concentrazione delle polveri sottili - PM_{10}, dato medio giornaliero, rilevato in alcuni punti della regione.

Come brevemente descritto nel capitolo precedente, gli agenti atmosferici inquinanti sono oggetto di monitoraggio e tale azione, coordinata a livello nazionale, viene realizzata nella regione Veneto con una rete di stazioni - o centraline - di rilevazione fisse, che misurano i dati analizzati in questa tesi.

Nel corso degli anni l'ARPAV è passata da 2 centraline nell'anno 2002 all'attivazione di 27 centraline nell'anno 2006, raggiungendo una copertura del territorio superiore a quanto previsto dalla normativa nazionale di riferimento (DM 60/2002).

Visto l'impatto organizzativo ed economico conseguente alla costruzione della rete di rilevamento, ogni stazione, solitamente, consente la rilevazione dei dati di una molteplicità di sostanze inquinanti, anche se qui ci si soffermerà esclusivamente sui dati relativi alle polveri sottili.

Gli strumenti installati in ogni centralina appartenente alla rete di rilevazione e che misurano il livello della concentrazione di particolato, hanno modalità diverse di rilevazione: manuale - gravimetrica - e automatica - mediante l'assorbimento della radiazione β con campionamento su nastro o su filtro -. Nel corso dell'anno 2006 risultavano attive 10 centraline con misurazione manuale, 7 con misurazione automatica con campionamento su nastro, 10 con misurazione automatica con campionamento su filtro.

La Figura 2.1 presenta la localizzazione delle stazioni nella regione Veneto, in cui emerge che l'area soggetta a rilevazione risulta essere prevalentemente quella della pianura, ad eccezione di un'area tra Padova e Verona corrispondente ai Colli Euganei e ai Monti Berici; non risultano essere coperte le zone montane del Cadore nel bellunese, l'altopiano di Asiago nel vicentino e la zona dei monti Lessini nel veronese.

Come naturale e ragionevole, la presenza delle centraline appare concentrata nei centri urbani, capolughi di provincia e nei centri abitati con maggiore densità. Le informazioni messe a disposizione dall'ARPAV che carat-

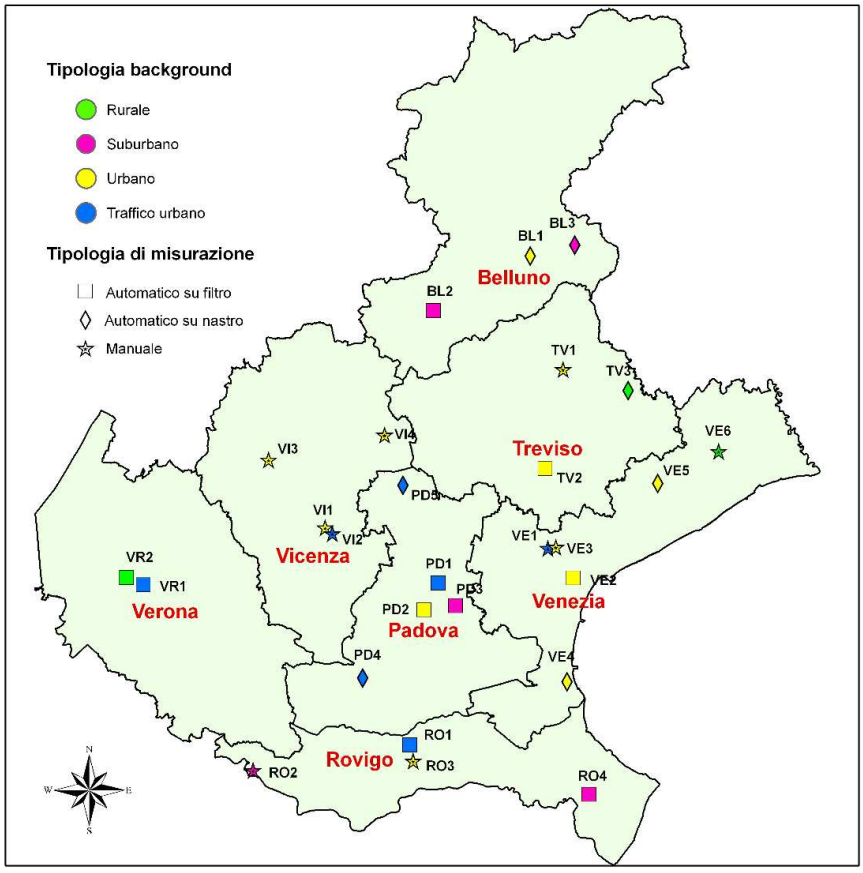


Figura 2.1: Localizzazione e tipo di rilevazione delle stazioni - anno 2006

terizzano ogni stazione, presentate in dettaglio nell'Appendice A (Tabelle A.1 e A.2), riguardano:

le coordinate geografiche : indicano la posizione in cui sono localizzate le centraline nel territorio secondo la proiezione di Gauss-Boaga, siste-

ma di riferimento cartografico italiano che utilizza come meridiano di riferimento quello passante per Roma-Monte Mario, espresse mediante Longitudine (direzione ovest-est) e Latitudine (direzione sud-nord);

il tipo di *background* : descrive la tipologia del contesto territoriale in cui la centralina è posizionata.

La classificazione usata [1] prevede una duplice caratterizzazione tra stazione di traffico, utilizzata prevalentemente per la misura di inquinanti da traffico (CO, NO, benzene, PM_{10}), posizionata in corrispondenza di strade urbane ad elevato flusso di traffico veicolare e stazione di *background* (di fondo), situata in un'area (ad es. parchi, aree verdi, rurali) non direttamente influenzata da sorgenti di traffico, quali strade e autostrade o da sorgenti di tipo industriale, e utilizzata per la misura di tutti gli inquinanti (CO, NO, benzene, PM_{10}, SO_2, NO_2, O_3).

Oltre a questo, l'ambito in cui è presente la stazione viene classificato in base alla tipologia di urbanizzazione distinguendo tra zona urbana, edificata in continuo; suburbana in cui si mescolano parti edificate e aree non urbanizzate; industriale zone a prevalente, se non esclusiva, attività di tipo industriale;rurale zona a bassa edificazione o non compresa nelle categorie precedenti.

il tipo di strumento di misurazione : indica la metodologia adottata dallo strumento presente nella centralina per la rilevazione della concentrazione di PM_{10} distinta tra manuale (gravimetrico) e automatico basata sul principio dell'assorbimento di radiazione β distinto a sua volta tra campionamento su nastro e campionamento su filtro

l'altitudine : indica l'altitudine, sul livello del mare, a cui è situata la centralina di rilevamento

Tutte queste informazioni saranno oggetto di indagine in una analisi preliminare del livello di concentrazione di PM_{10} nel prossimo paragrafo.

L'ARPAV, come previsto dalla normativa, effettua anche delle campagne di monitoraggio attraverso una serie di centraline rilocabili (mobili) in due periodi nell'anno della durata solitamente di due settimane, nella stagione invernale e nella stagione estiva. Grazie alla possibilità di essere facilmente trasportabili, mediante questa tipologia di rilevazione, l'ARPAV riesce a mo-

nitorare diffusamente il territorio regionale, sia pure per periodi di tempo molto limitati.

2.3 Le prime analisi esplorative

Le serie giornaliere della concentrazione media di PM_{10}, fornite dall'ARPAV, presentano, in maniera difforme per ognuna di esse, molti dati mancanti - *missing data* - per cui nessuna delle serie risulta completa con 365 osservazioni. Questo è un problema quasi ineliminabile che si manifesta sistematicamente quando si devono analizzare dati di carattere ambientale. La mancanza del dato può derivare sostanzialmente da due fattori: il primo dipende dal fatto che gli strumenti sono soggetti a guasto, e soprattutto nel caso di quelli automatici, non sempre risulta possibile ricorrere alla riparazione tempestivamente, sia per problemi organizzativi, sia tecnici; il secondo dipende dall'organizzazione stessa della rete di rilevamento che può prevedere o l'attivazione di nuove centraline non necessariamente coincidente con l'inizio dell'anno, o la loro disattivazione in un particolare sito, o il loro spostamento in un nuovo sito, con la conseguente perdita della continuità sia temporale - le operazioni tecniche e di restart possono richiedere un certo intervallo di tempo - sia spaziale, visto che la serie dei dati in una determinata posizione viene interrotta e parte una nuova serie in un diverso punto dello spazio.

Risulta chiaro che i valori mancati (*missing data*) non sono classificabili come mancanti a caso (*missing at random*), per cui con le informazioni a disposizione non risulta possibile adottare tecniche di previsione o interpolazione. La conseguenza di tale scelta avrà necessariamente un impatto sulla distorsione dei risultati.

Nella Tabella 2.1 sono presentate le statistiche descrittive relative all'anno 2006 per le 27 centraline attive nel corso dell'anno.

Come si può notare nella seconda colonna con espressi i giorni in cui ci sono valori rilevati, ognuna delle stazioni di rilevamento presenta dei valori mancanti, per le cause sopra descritte, in qualche caso raggiugendo percentuali oltre il 40% come per le centraline BL3, PD5 e VE6; nella tabella vengono evidenziate le 17 centraline pari al 63% sul totale, che dispongono di

Centralina	gg. oss.	gg. sup.	Media	Dev. Stand.	Q_1	Mediana	Q_3	max
BL1	*364*	*33*	*26,11*	*17,92*	*14,00*	*22,50*	*33,25*	*122*
BL2	*358*	*104*	*39,82*	*29,25*	*19,00*	*32,00*	*55,00*	*182*
BL3	206	1	17,59	9,46	11,00	16,00	23,00	52
PD1	*358*	*174*	*54,98*	*30,03*	*33,00*	*50,00*	*73,00*	*169*
PD2	*354*	*158*	*51,18*	*29,33*	*29,25*	*46,00*	*68,75*	*162*
PD3	*358*	*150*	*49,98*	*28,55*	*28,25*	*44,50*	*65,00*	*149*
PD4	224	84	49,44	31,92	27,75	42,50	61,25	172
PD5	211	58	41,54	24,98	24,00	39,00	53,00	151
RO1	326	119	48,23	27,64	28,25	43,00	60,00	143
RO2	277	62	38,52	23,95	23,00	32,00	46,00	143
RO3	274	73	42,61	30,77	22,25	33,00	52,75	174
RO4	*353*	*73*	*36,86*	*25,07*	*20,00*	*31,00*	*46,00*	*136*
TV1	*338*	*63*	*35,41*	*19,78*	*21,00*	*32,00*	*45,75*	*106*
TV2	*354*	*107*	*40,57*	*24,89*	*22,00*	*35,00*	*54,00*	*123*
TV3	312	59	32,23	23,94	16,00	25,00	44,00	114
VE1	*359*	*171*	*56,55*	*33,60*	*33,00*	*50,00*	*71,00*	*203*
VE2	*362*	*73*	*37,38*	*26,42*	*20,00*	*30,00*	*46,75*	*155*
VE3	*340*	*120*	*46,95*	*31,07*	*26,00*	*39,00*	*60,00*	*182*
VE4	266	49	36,70	25,81	21,00	30,00	46,00	172
VE5	262	43	34,25	24,33	19,00	29,00	42,00	129
VE6	180	31	34,73	18,61	20,75	30,50	45,00	102
VR1	*344*	*188*	*61,22*	*35,05*	*33,00*	*55,00*	*81,00*	*165*
VR2	*340*	*132*	*47,93*	*27,96*	*26,75*	*43,00*	*61,00*	*153*
VI1	*357*	*154*	*50,20*	*30,87*	*25,00*	*43,00*	*68,00*	*158*
VI2	*360*	*173*	*56,37*	*33,05*	*31,00*	*49,00*	*73,25*	*184*
VI3	*344*	*76*	*36,03*	*24,44*	*18,00*	*29,00*	*48,00*	*129*
VI4	283	65	36,57	23,80	19,00	30,00	47,00	132

Tabella 2.1: Statistiche descrittive (in *corsivo* centraline con più del 90% dei dati)

almeno il 90% dei dati come previsto dalla normativa sulla raccolta minima dei dati.

Circa il confronto dei dati, nella terza colonna in cui sono riportati i giorni di superamento, con i limiti previsti per il numero di superamenti nell'anno si osserva, purtroppo, che tutte - con l'eccezione della centralina BL3 (situata nell'Alpago ad una altitudine superiore ai 600 metri) e BL1, appena inferiore al limite - oltrepassano la soglia di 35 giorni nell'anno civile in cui il valore risulta superiore a 50 $\mu g/mc$; è evidente quindi come il problema sia grave, preoccupante e persitente.

Nella Fig. 2.2 vengono presentati i *boxplot* della distribuzione dei valori della concentrazione del PM_{10} per ogni stazione, dalla quale si evince

come le prime tre centraline, site nella provincia di Belluno, abbiano una distribuzione dei valori molto inferiore rispetto a tutte le altre; le centraline che presentano un valore medio più elevato della soglia sono quelle site nelle città di Padova, Verona - dove si trova la centralina con i valori massimi - e Vicenza, mettendo il rilievo, già da ora, come questa sia la zona più critica per questo tipo di inquinamento; da notare come anche la centralina posta a Mestre in Via Circonvallazione abbia valori molto elevati di inquinante. Ciò rafforza l'ipotesi che le aree soggette a maggior traffico, oltre che a maggior urbanizzazione, siano quelle con maggiori livelli di inquinamento.

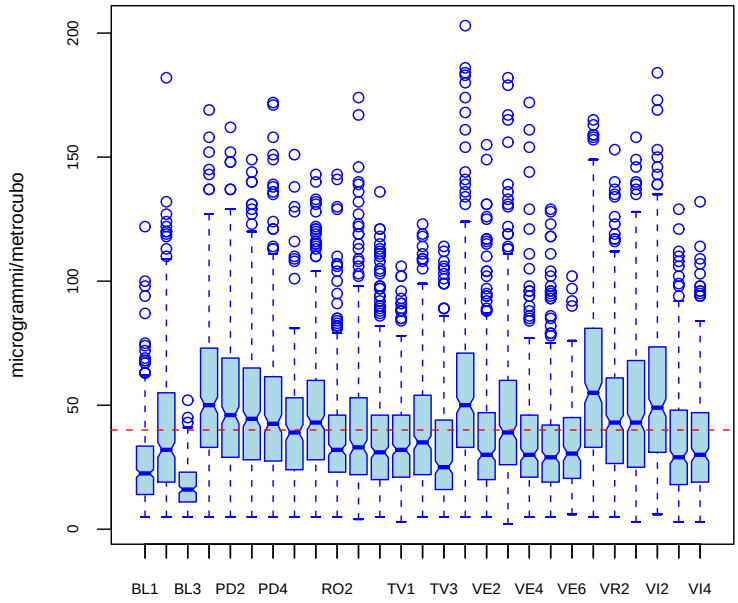


Figura 2.2: Boxplot PM_{10} per stazione - anno 2006

Analizzando i dati dal punto di vista della distribuzione, si vede come i valori siano stati sottoposti a censura, ovviamente dovuta al limite di rilevabilità dell'inquinante a seconda del tipo di strumento impiegato; inoltre si nota come la distribuzione risulti asimmetrica con code molto pesanti e allungate verso valori elevati di concentrazione.

L'analisi dei dati rispetto alla tipologia di misurazione della concentrazione presentata nella Tabella 2.2 e Figura 2.3, fa vedere come le centraline di tipo **Af** (automatico su filtro) abbiano un valore mediano inferiore e una distribuzione più concentrata; le altre due tipologie **An** (automatico su nastro) e **M** (manuale-gravimetrico) si presentano con caratteristiche analoghe sia per il valore mediano sia per la distribuzione dei valori.

Analizzando più in dettaglio le centraline **An** si nota come queste siano in numero minore rispetto alle altre due tipologie; inoltre, in questa classe sono comprese tre stazioni attivate nella tarda primavera - BL3, PD4, PD5 - senza quindi le osservazioni dei primi mesi, periodo in cui si verificano i valori più elevati di concentrazione di polveri - e comprendano anche tre stazioni, BL1, VE4 e VE5, situate in zone 'periferiche' dell'area soggetta a rilevamento.

In base a queste considerazioni e particolarità evidenziate per il gruppo di centraline di tipo **An**, sembra non essere attendibile la differenza presente nei dati, anche se tale considerazione avrebbe bisogno di un maggiore supporto statistico.

Tipologia di misurazione	Numero centraline	Media	Standard Error
Af	10	46,74	29,51
An	7	33,47	24,95
M	10	44,27	29,20

Tabella 2.2: Statistiche descrittive per tipo di misurazione - anno 2006

Passando a considerare i dati secondo il tipo di *background* in cui sono posizionate le centraline - vedi Tab. 2.3 e Fig. 2.4 - si nota subito come siano più elevati i valori per le centraline con tipologia **TU**, la quale ripropone la differenza tra i siti prossimi a zone di traffico, rispetto alle altre tre che presentano valori mediani e dispersione molto simili; secondo questa classificazione può sembrare quindi opportuno aggregare i dati differenziando esclusivamente tra centraline poste in zona di traffico o meno, tralasciando di fatto la distinzione tra tipo di *background* urbano, suburbano e rurale. La centralina PD3 classificata dall'ARPAV secondo la tipologia BS-IND, è unica e quindi è stata considerata come appartenente alla classe BS.

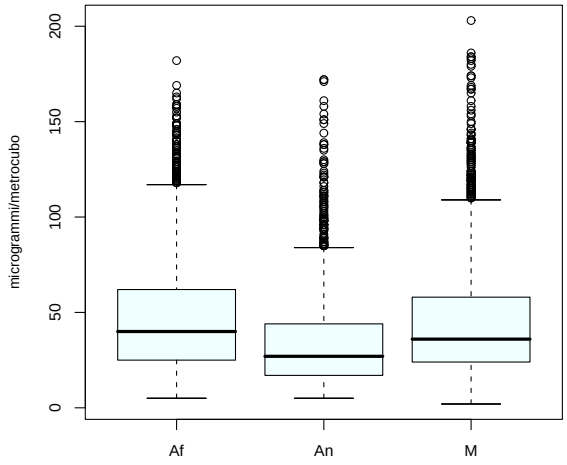


Figura 2.3: Tipo di misurazione - anno 2006

Tipologia di misurazione	Numero centraline	Media	Standard Error
TU - zona traffico/*background* urbano	7	53,57	31,88
BU - zona no traffico/*background* urbano	12	39,64	27,01
BS - zona no traffico/*background* suburbano	5	38,30	27,03
BR - zona no traffico/*background* rurale	3	39,18	25,72

Tabella 2.3: Statistiche descrittive per zona - anno 2006

2.4 Analisi temporale dei dati

Le serie storiche di ogni stazione - presentate nell'Appendice A, Fig. A.1 - permettono di evidenziare i periodi di attivazione di ogni centralina e gli eventuali dati mancanti. Per ogni centralina viene anche evidenziato il valore del limite giornaliero (linea tratteggiata) previsto dalle normative, che consente di mettere in risalto i valori e i periodi di superamento di tale limite.

L'osservazione delle serie storiche mette subito in evidenza il comportamento della concentrazione del PM_{10} nell'anno solare. E' chiaro come le polveri sottili si accumulino maggiormente durante i mesi invernali (inizio e fine delle serie); risulta una tendenza ad un innalzamento dei valori anche nel periodo a metà anno, verosimilmente dovuti ad una componente

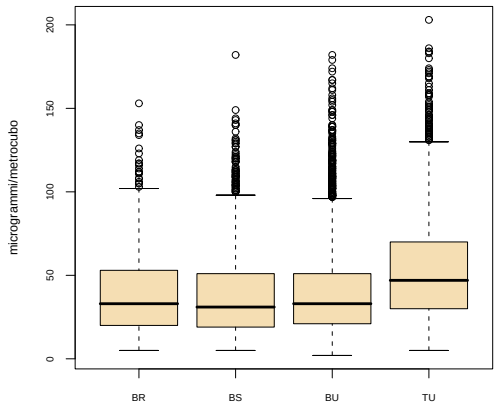


Figura 2.4: Tipo di *background* - anno 2006

legata a processi fotochimici, attivati dalle maggiori radiazioni solari, oltre probabilmente a fenomeni atmosferici che non favoriscono la dispersione del particolato.

L'analisi dei dati aggregati per mese, vedi Fig. 2.5, consente di sottolineare come i mesi di febbraio e novembre siano quelli con i valori di concentrazione più elevati, che risultano invece notevolmente inferiori nei mesi estivi, oltre ad essere più concentrati attorno al valore mediano, fatta eccezione per il mese di giugno. Il mese di agosto, complice, molto probabilmente, la chiusura delle attività lavorative e di conseguenza una diminuzione marcata del traffico, risulta quello con valori più bassi e con minor dispersione.
Passando a valutare l'aggregazione dei valori per giorno della settimana, vedi Fig. 2.6, si evidenzia una non uniformità nella distribuzione con un 'picco' massimo nel giorno di venerdì e una maggiore dispersione a metà settimana (mercoledì). La minore dispersione nel giorno di lunedì, vista la forte correlazione temporale, potrebbe essere dovuta alla diminuzione del traffico del fine settimana.

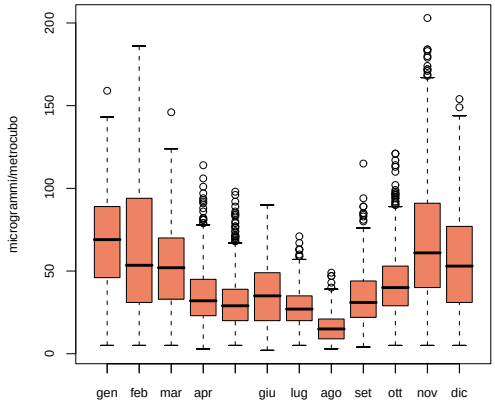


Figura 2.5: Boxplot PM_{10} per mese - anno 2006

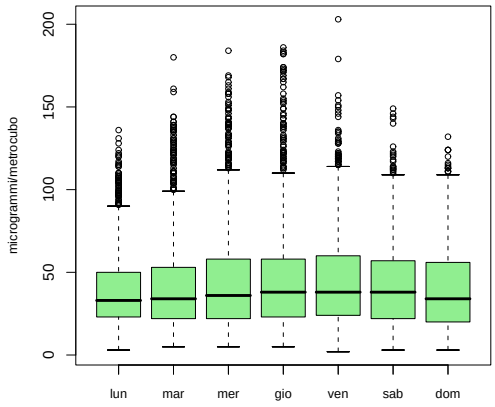


Figura 2.6: Boxplot PM_{10} per giorno della settimana - anno 2006

Capitolo 3

L'analisi spaziale dei valori medi annuali

3.1 La modellazione geostatistica classica

Lo studio di fenomeni che si manifestano nello spazio, la loro analisi, la modellizzazione e la previsione risale a metà del secolo appena trascorso, quando l'ingegnere D. Krige iniziò ad applicare tecniche statistiche per la previsione di riserve di minerali in Sudafrica.

Nel corso degli anni '60, in particolare con il matematico francese G. Matheron [12], vengono formulate le basi teoriche per tali metodi e un modello che tenesse conto della variabilità e del legame tra le variabili osservate, dato dalla correlazione spaziale.

L'insieme di questi metodi prende il nome di geostatistica, che partendo da un insieme di misurazioni di un fenomeno in una regione dello spazio, permette di definire una struttura di correlazione e la previsione del valore nei punti in cui non sono state effettuate misurazioni. I campi di applicazione della disciplina, oltre alla geologia e alle scienze del suolo, - come quanto concerne le riserve minerarie e livello degli inquinanti nel suolo o sottosuolo -, riguardano l'idrogeologia, la meterologia, l'agricoltura, l'ingegneria civile e ovviamente le scienze ambientali, in cui viene compreso lo studio del livello di concentrazione di inquinanti nell'aria.

3.1.1 Le variabili in geostatistica

L'analisi geostatistica classica descrive il meccanismo di generazione assumendo che il fenomeno oggetto di studio sia rappresentato dalla realizzazione di una variabile casuale in determinate localizzazioni dello spazio euclideo d-dimensionale $\mathbb{R}^d$. In quasi tutte le applicazioni citate precedentemente, lo spazio euclideo considerato corrisponde ad una superficie bidimensionale, per cui verrà considerato d'ora in poi lo spazio $\mathbb{R}^2$.

Formalmente, indicando con $\mathbf{s} = (x, y)$ un sito (coppia di coordinate del piano) generico nello spazio $\mathbb{R}^2$ e con $D \in \mathbb{R}^2$ il sottoinsieme, area o regione oggetto di studio, il fenomeno si interpreta mediante l'utilizzo della variabile casuale continua, $\{Z(\mathbf{s}),\ \mathbf{s} \in D\}$, eventualmente multivariata, la cui realizzazione viene denotata con $\{z(\mathbf{s}),\ \mathbf{s} \in D\}$.

Al fine di modellare un insieme di variabili aleatorie, le cui localizzazioni variano in $\mathbb{R}^2$, si deve far ricorso al concetto di processo stocastico.

Un processo stocastico viene usualmente definito attraverso la distribuzione finita dimensionale della sua funzione di ripartizione, ossia per $k \leq +\infty$

$$F_{s_1,\ldots,s_k}(z_1, \ldots, z_k) = Pr(Z(s_1) \leq z_1, \ldots, Z(s_k) \leq z_k) \tag{3.1}$$

Nel caso della geostatistica, $Z(\mathbf{s})$ prende anche il nome di campo aleatorio.

3.1.2 Misure della variabilità per campi aleatori

Covarianza

La misura dell'intensità del legame tra ogni coppia di variabili casuali viene data dalla covarianza. In particolare, la struttura della covarianza in un campo aleatorio indica come il comportamento del valore osservato $z(\mathbf{s}_i)$ delle variabili casuali in siti diversi della regione D risulti analogo o meno.

Si definisce con $\mu(\mathbf{s}) \equiv E[Z(\mathbf{s})]$, momento del primo ordine, il valore atteso del campo aleatorio $Z(\mathbf{s})$, che generalmente dipende dalla localizzazione $\mathbf{s}$, e la covarianza, momento del secondo ordine dalla media, di un campo aleatorio mediante

$$C\big(Z(\mathbf{s}_1), Z(\mathbf{s}_2)\big) \equiv E\Big[\big(Z(\mathbf{s}_1) - \mu(\mathbf{s}_1)\big)\big(Z(\mathbf{s}_2) - \mu(\mathbf{s}_2)\big)\Big] \tag{3.2}$$

per i due siti $\mathbf{s}_1$ e $\mathbf{s}_2$.
Come si può facilmente dedurre, la covarianza calcolata nello stesso sito, $\mathbf{s}_1 = \mathbf{s}_2 = \mathbf{s}$, indica la varianza.

Il variogramma

Nel caso di campi aleatori, Matheron [12] ha introdotto la funzione variogramma, che fornisce informazioni simili a quelle date dalla funzione di covarianza spaziale, la quale viene definita come la varianza tra la differenza delle due v.a. $Z(\mathbf{s}_1)$ e $Z(\mathbf{s}_2)$

$$2\gamma(\mathbf{s}_1 - \mathbf{s}_2) \equiv Var\big[Z(\mathbf{s}_1) - Z(\mathbf{s}_2)\big] \tag{3.3}$$

la funzione $\gamma(\mathbf{s}_1 - \mathbf{s}_2)$, senza il fattore 2, viene definita semivariogramma, che risulta legata alla funzione di covarianza nel caso di campi aleatori aventi caratteristiche particolari come si vedrà in seguito.

Questa funzione esprime il concetto di similarità nel caso di variabili aleatorie nello spazio, ovvero come osservazioni relative a siti vicini tendano ad essere più simili rispetto ad osservazioni relative a siti più distanti.

3.1.3 Processi stazionari

Un campo aleatorio è detto *stazionario in senso forte* se per ogni vettore $\mathbf{h}$ le distribuzioni finite dimensionali di $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ e $\{Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_n + \mathbf{h})\}$ sono identiche per ogni valore di n e di $\mathbf{h}$; ne consegue che il campo aleatorio è invariante rispetto alla posizione ma, raramente, i processi oggetto di analisi geostatistiche possiedono tali requisiti.

Un campo aleatorio è detto *stazionario del secondo ordine* se: (a) esiste il valore atteso non dipendente dalla localizzazione, ossia costante in tutto il dominio $D \subset \mathbb{R}^2$; (b) esiste la funzione di covarianza dipendente solo dal vettore $\mathbf{h}$ che separa i due siti; ossia

(a) $\mu(\mathbf{s}) = E[Z(\mathbf{s})] = \mu$
(b) $C(\mathbf{s}, \mathbf{s} + \mathbf{h}) = C(\mathbf{s} + \mathbf{h} - \mathbf{s}) = C(\mathbf{h})$

La funzione di covarianza $C(\mathbf{h})$ in due siti diversi dipende solo dal vettore $\mathbf{h}$

che separa i due punti. Per $\mathbf{h} = \mathbf{0}$ si ha $C(\mathbf{0}) = Var[Z(\mathbf{s})]$.

La funzione $C(\mathbf{h})$ è detta covariogramma; la funzione $\rho(\mathbf{h}) = C(\mathbf{h})/C(\mathbf{0})$, detta correlogramma, rappresenta la correlazione spaziale e risulta compresa nell'intervallo $[-1, 1]$

Se il campo aleatorio è stazionario del secondo ordine, il variogramma può essere espresso come:

$$\begin{aligned} 2\gamma(\mathbf{h}) &= Var\big[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})\big] \\ &= Var\big[Z(\mathbf{s})\big] + Var\big[Z(\mathbf{s}+\mathbf{h}\big] - 2Cov\big[Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})\big] \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \\ &= 2[C(\mathbf{0}) - C(\mathbf{h})] \end{aligned} \tag{3.4}$$

dove nella terza uguaglianza si è fatto ricorso all'ipotesi $Var[Z(\mathbf{s})] = Var[Z(\mathbf{s})+\mathbf{h}]$. Come conseguenza il semivariogramma può essere espresso nella forma $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$

Una ulteriore forma di stazionarietà definisce un processo *intrinsecamente stazionario* se (a) esiste il valore atteso non dipendente dalla localizzazione e (b) esiste la varianza della differenza tra i due siti separati da $\mathbf{h}$ e dipende esclusivamente da $\mathbf{h}$; in questo caso esiste quindi il variogramma $2\gamma(\mathbf{s}, \mathbf{s}+\mathbf{h}) = 2\gamma(\mathbf{h}) = Var\big[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})\big]$

Mentre un processo stazionario del secondo ordine implica che il processo sia anche intrinsecamente stazionario, il viceversa non vale.

3.1.4 Proprietà delle funzioni di covarianza e variogramma per processi stazionari

La funzione di covarianza definita dalla 3.2 deve soddisfare le seguenti proprietà:

- $|C(\mathbf{h})| \leq C(\mathbf{0}) = Var\big[Z(\mathbf{s})\big]$ è una funzione limitata
- $C(\mathbf{h}) = C(-\mathbf{h})$ è una funzione pari (o simmetrica)
- $C(\mathbf{h})$ è una funzione *definita positiva*

L'ultima proprietà deriva dal fatto che l'utilizzo di $C(\mathbf{h})$ per il calcolo della varianza di una combinazione lineare di v.c. deve essere positiva.

$$var\Big(\sum_i w_i Z(\mathbf{s}_i)\Big) = \sum_i \sum_j w_i w_j C(\mathbf{s}_i - \mathbf{s}_j) = \mathbf{w}^T\ \mathbf{C}\ \mathbf{w} \geq 0$$

$\forall \{s_1, \ldots, s_n\} \in D$ e $\{w_i, \ldots, w_n\}$

Conseguenza delle proprietà viste per la covarianza, il semivariogramma, come definito nella formula 3.3 e a meno del fattore costante pari a 2, deve soddisfare alcune proprietà quali:

- $\gamma(\mathbf{h}) \geq 0$, $\gamma(\mathbf{0}) = 0$ la funzione assume valori positivi ed è continua nell'origine (assume valore 0)
- $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$ è una funzione pari (simmetrica)
- Il variogramma è una funzione *condizionatamente definita negativa*

L'ultima proprietà dipende dal fatto che

$$var\Big(\sum_i w_i Z(\mathbf{s}_i)\Big) = -\sum_i \sum_j w_i w_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \geq 0 \text{ se } \sum_i w_i = 0$$

Anisotropia e Isotropia

Si è visto finora come la covarianza $C(\mathbf{h})$ dipenda dal vettore che unisce i due punti in cui il fenomeno viene osservato, per cui $\mathbf{h}$ specifica sia la direzione che la distanza. Nelle applicazioni di fenomeni variabili nello spazio molto spesso la funzione di covarianza manifesta comportamenti diversi se analizzata in direzioni diverse. Un campo aleatorio con tale caratteristica si definisce *anisotropo*.

Nel caso di applicazioni geostatistiche in cui il campo aleatorio si presenta omogeneo - per esempio in alcuni fenomeni geologici - il comportamento assunto dalla funzione di covarianza può risultare identico, o più realisticamente analogo, in qualunque direzione dello spazio. L'unica dipendenza di

$C(\mathbf{h})$ viene manifestata dalla distanza $||h||$ tra due siti e in questo caso il processo viene detto *isotropo*.

Un caso particolare di anisotropia, chiamata *anisotropia geometrica*, si presenta quando la funzione di covarianza $C(\mathbf{h})$ può essere ricondotta, tramite una trasformazione lineare delle coordinate di h - del tipo $h' = Ah$ - ad una funzione di covarianza isotropica; si cerca di determinare una trasformazione tale per cui $\gamma(h) = \tilde{\gamma}(||h'||)$ dove $\tilde{\gamma}(||h'||)$ risulta un variogramma isotropico.

3.1.5 Modelli di (semi)variogramma isotropici

Diversi studiosi hanno descritto vari modelli di variogramma, utilizzati per rappresentare e interpretare l'andamento della variabilità del campo aleatorio, nel caso di isotropia, in funzione della distanza $||h||$ tra la localizzazione dei siti. Il grafico di $\gamma(h)$ nel caso in cui valgano le ipotesi di stazio-

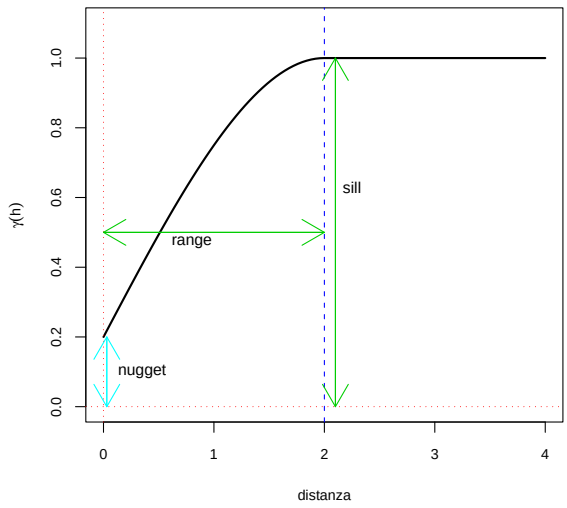


Figura 3.1: Andamento tipo del semivariogramma teorico

narietà del secondo ordine, generalmente, presenta un andamento crescente a partire dall'origine in cui vale 0 ($\gamma(\mathbf{0}) = 0$) e aumenta con l'incremento della distanza $||h||$ per tendere al valore di $C(0)$, visto che al tendere di $||h|| \to +\infty$, il covariogramma assume valori prossimi allo 0; infatti se le due

variabili $Z(s)$ e $Z(s+h)$ non manifestano un legame nello spazio, ossia sono non correlate, allora $C(h) = 0$ e per la 3.4 si ottiene $\gamma(h) = C(0)$.

Il valore $||h||$ per cui non si ha più una correlazione tra le due variabili nello spazio viene denominato *range*, mentre il valore limite raggiunto $C(0)$, talvolta asintoticamente, dal semivariogramma viene denominato *sill*.

In molti casi si è in presenza di una discontinuità nell'origine, ossia $\gamma(0) \neq 0$, e questo fatto viene interpretato come l'effetto di una variabilità esterna al fenomeno e indotta ad esempio dall'errore di misurazione, o alla variabilità spaziale per distanze inferiori a quelle dei siti oggetti di misurazione.

Il semivariogramma si caratterizza, dunque, attraverso tre parametri: l' effetto *nugget* τ che spiega la discontinuità nell'origine dove $h = 0$; un coefficiente ϕ che determina la velocità di decremento dell'intensità del legame tra variabili, che determina il *range* e collegato alla distanza h oltre la quale non si ha correlazione tra due siti nello spazio (dove il covariogramma $C(h) = 0$); il *sill* σ^2 che rappresenta il valore di $C(0)$, ossia la varianza del processo.

Per semplicità, visto il legame a meno del fattore costante pari a 2, si presentano nella Figura 3.2 i modelli di semivariogramma isotropici teorici che soddifano quindi le proprietà viste al paragrafo precedente.

Modello Pepita Caso in cui è assente la correlazione spaziale tra due diversi siti.

$$\gamma(h) = \sigma^2 > 0 \quad \text{per} \quad ||h|| = 0\,; \quad 0 \quad \text{per } ||h|| > 0 \tag{3.5}$$

Modello Esponenziale Presenta un comportamento lineare in prossimità dell'origine, raggiunge il valore del *sill* solo asintoticamente e di conseguenza non ammette *range*

$$\gamma(h) = \sigma^2\Big\{1 - \exp(-\frac{||h||}{\phi}\Big\} \tag{3.6}$$

Modello Gaussiano Presenta caratteristiche simili al modello precedente

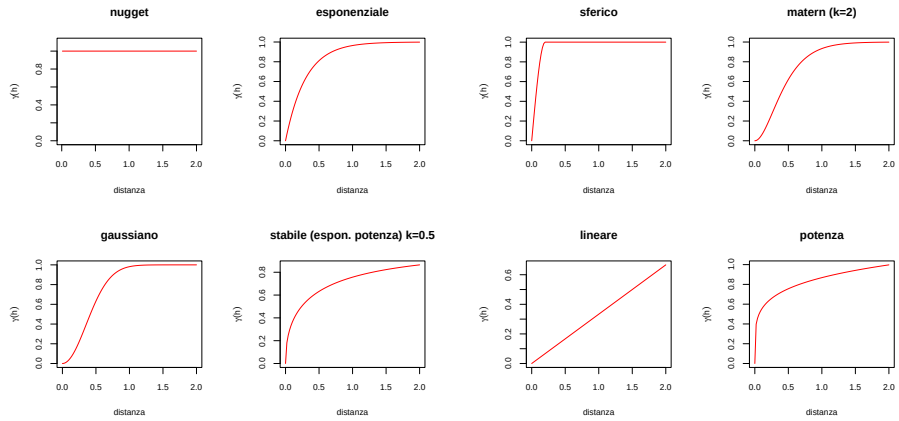


Figura 3.2: Modelli di semivariogramma

per quanto riguarda i parametri di *sill* e *range*, ma ha un comportamento presso l'origine di tipo parabolico che consente di rappresentare un fenomeno più regolare per distanze tra i siti prossime allo 0

$$\gamma(h) = \sigma^2\Big\{1 - \exp\big(-\frac{||h||^2}{\phi}\big)\Big\} \tag{3.7}$$

Modello Esponenziale-Potenza Caso più generale che comprende i due modelli sopradescritti, esponenziale nel caso $k = 1$ e gaussiano nel caso $k = 2$

$$\gamma(h) = \sigma^2\Big\{1 - \exp\big(-\frac{||h||^k}{\phi}\big)\Big\} \text{ con } \phi > 0, 0 < k \leq 2 \tag{3.8}$$

Modello Sferico E' un modello spesso utilizzato nel caso in cui si manifestino con evidenza sia il *sill*, sia il *range*

$$\gamma(h) = \sigma^2\Big\{1 - \frac{3}{2}\frac{||h||}{\phi} + \frac{1}{2}\Big(\frac{||h||}{\phi}\Big)^3\Big\} \text{ se } ||h|| < \phi \tag{3.9}$$

Modello Matérn Questo modello risulta intermedio tra quello esponenziale (di cui è un caso particolare quando $k = 0.5$) e quello gaussiano e consente una certa flessibilità data dal parametro k

$$\gamma(h) = \sigma^2\Big\{1 - \frac{1}{2^{\mathcal{K}-1}\Gamma(\mathcal{K})}\Big(\frac{||h||}{\phi}\Big)^{\mathcal{K}} K_{\mathcal{K}}\Big(\frac{||h||}{\phi}\Big)\Big\} \tag{3.10}$$

dove con $\mathcal{K}$ si indica la funzione di Bessel del secondo tipo.

Sono stati proposti anche modelli di semivariogramma per processi non stazionari, che non ammettono covarianza finita e che sono quindi caratterizzati dall'assenza di *sill*; in questo caso il valore della funzione $\gamma(\cdot)$ aumenta al crescere di h, senza convergere asintoticamente.

Modello Potenza

$$\gamma(h) = \sigma^2 ||h||^{\phi} \text{ con } 0 < \phi < 2, \sigma^2 > 0 \tag{3.11}$$

nel caso in cui $\phi = 1$ si ottiene il **Modello Lineare** in cui $\gamma(h) = \sigma^2 ||h||$

3.1.6 Il variogramma empirico e la stima dei parametri

Considerando l'insieme delle distanze tra tutte le coppie di siti, in cui è nota la determinazione della variabile aleatoria che misura il fenomeno oggetto di interesse, e il quadrato delle loro differenze, si ottiene il grafico, denominato *variogramma nuvola* che permette una analisi esplorativa del fenomeno. Questo grafico, in cui in ascissa viene considerata la distanza tra i siti $|s_i - s_j|$ e in ordinata la quantità $\big(Z(s_i) - Z(s_j)\big)^2$, consente l'individuazione di alcune caratteristiche quali:

- valori anomali globali: sono misurazioni molto lontane dalla maggior parte dei dati e quindi sono valori molto al di fuori da quelli che formano la nuvola;
- valori anomali locali: sono misurazioni molto diverse rispetto a quelle dei siti a loro vicini e possono presentarsi come punti con valori del quadrato delle differenze molto elevati per distanze piccole, possono anche indicare la presenza dell'effetto *nugget*;
- non stazionarietà locale: quando sono presenti insiemi di punti che possiedono maggiore variabilità rispetto ai punti circostanti, può essere inoltre segnalata dalla presenza di valori anomali locali;
- anomalia della distribuzione da quella normale: i punti si dispongono non in maniera omogenea, ma sono aggregati in righe verticali che

possono essere indice del fatto che la variabile osservata sia discreta piuttosto che continua.

Aggregando i dati del variogramma nuvola per intervalli, si può ottenere una stima del variogramma teorico definito nell'equazione 3.3.

Nel caso in cui i dati siano ottenuti da siti disposti secondo una griglia regolare, la stima di questa quantità può essere ottenuta mediante il metodo dei momenti ([12])

$$2\hat{\gamma}(\mathbf{h}) \equiv \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} \Big(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\Big)^2 \text{ con } h \in \mathbb{R}^d \tag{3.12}$$

dove

$$N(\mathbf{h}) \equiv \big\{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}; i, j = 1, \ldots, n\big\}$$

e $|N(\mathbf{h})|$ indica il numero di coppie distinte (cardinalità) in $N(\mathbf{h})$

Una stima del variogramma più robusta rispetto a *outliers* ([8]) risulta essere

$$2\bar{\gamma}(\mathbf{h}) \equiv \frac{\Big\{\frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2}\Big\}^4}{0.457 + 0.494/|N(\mathbf{h})|} \tag{3.13}$$

Qualora i dati non siano ottenuti secondo una griglia regolare si ha

$$N(\mathbf{h}) \equiv \big\{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j \in T(\mathbf{h}); i, j = 1, \ldots, n\big\}$$

in cui $T(\mathbf{h})$ indica una regione di tolleranza di $\mathbf{h}$.

La stima dei parametri del variogramma

La scelta del modello teorico del variogramma, come si è visto precedentemente, comporta, oltre alla scelta del tipo di funzione che interpreta la struttura di variabilità, anche la determinazione dei parametri della funzione $\gamma(\cdot, \boldsymbol{\theta})$ con $\boldsymbol{\theta} = (\phi, \sigma^2, \tau)$, la quale garantisce sempre il rispetto delle proprietà del variogramma.

I metodi di stima utilizzati sono:

Minimi Quadrati La stima di $\boldsymbol{\theta}$ si ottiene tramite

$$\widehat{\boldsymbol{\theta}}_{MQ} = argmin_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\Big\{\sum_{i=1}^{n}\big[\hat{\gamma}(h_i) - \gamma(h_i;\boldsymbol{\theta})\big]^2\Big\}$$

e richiede un valore iniziale per la stima dei parametri del variogramma - ad esempio del variogramma empirico $\widehat{\gamma}(\cdot)$ - ma non richiede nessuna ipotesi sulla distribuzione del campo aleatorio.

Minimi Quadrati Generalizzati La stima è data da

$$\widehat{\boldsymbol{\theta}}_{MQG} = argmin_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\Big\{\sum_{i=1}^{n}|N(h)|\big[\frac{\hat{\gamma}(h_i)}{\gamma(h_i;\boldsymbol{\theta})} - 1\big]^2\Big\} \tag{3.14}$$

Richiede sempre un valore iniziale per la stima dei parametri e non richiede assunzioni sulla distribuzione; diversamente dal precedente, attribuisce un peso variabile agli scarti, ponderandoli mediante la numerosità in ogni intervallo e le stime ottenute risultano non distorte e consistenti. E' il metodo più utilizzato per la stima dei parametri, anche perché nel calcolo delle stime si considera la correlazione presente tra le osservazioni.

Massima Verosimiglianza Questo metodo prevede l'assunzione dell'ipotesi di normalità sulla distribuzione del campo aleatorio, ossia

$$Z = (Z(s_1),\dots,Z(s_n))' \sim N_q(F\beta, \Sigma(\boldsymbol{\theta}))$$

Le stime di massima verosimiglianza per β e $\boldsymbol{\theta}$ si ottengono minimizzando la funzione di verosimiglianza

$$L(\beta,\boldsymbol{\theta}) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma(\boldsymbol{\theta})| + \frac{1}{2}(Z - F\beta)'\Sigma(\boldsymbol{\theta})^{-1}(Z - F\beta)$$

e quindi

$$L(\hat{\beta},\hat{\boldsymbol{\theta}}) = min\{L(\beta,\boldsymbol{\theta}) \;:\; \beta\in\mathbb{R}^q,\;\; \boldsymbol{\theta}\in\boldsymbol{\Theta}\}$$

Il metodo, vista la maggiore informazione sulla distribuzione, fornisce stime più efficienti rispetto a quelle ottenute con gli altri due metodi, anche se

comporta solitamente un maggior carico computazionale.

3.1.7 La previsione del processo spaziale - Il *Kriging*

Uno dei problemi e obiettivi della statistica è quello di riuscire a predire o stimare un fenomeno a partire da un insieme di dati 'limitato' e dare a questa stima, congiuntamente, una misura dell'incertezza. Nel caso della regressione lineare, da un insieme di variabili covariate X, si riesce a formulare la previsione sull'andamento di una variabile casuale Y attraverso l'utilizzo di un modello del tipo $Y = X\beta + \epsilon$.

In geostatistica il modello fondamentale per la descrizione di un campo aleatorio viene specificato nella forma

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{3.15}$$

in cui la variabile casuale $Z(\mathbf{s})$ osservata nel sito $\mathbf{s}$ viene decomposta in una parte deterministica - di larga scala - $\mu(\mathbf{s})$, che descrive e interpreta fisicamente il fenomeno e in una parte o componente stocastica - di piccola scala - $\epsilon(\mathbf{s})$, la quale incorpora e descrive la struttura di covarianza spaziale.

La parte deterministica $\mu(\mathbf{s})$, strutturale, ingloba le caratteristiche globali del processo, e in analogia con quanto avviene per le serie temporali, viene denominata *trend*. Questa componente, se presente nel processo oggetto di analisi, può essere costante in tutta la regione D, oppure essere a sua volta espressa in termini più complessi tramite una dipendenza funzionale data dalla localizzazione $\mathbf{s}$, o da un insieme di altre variabili covariate $\mathbf{X}$.

Nel caso specifico la previsione di una variabile aleatoria il dominio $D \subset \mathbb{R}^d$ in generale, $D \subset \mathbb{R}^2$ per i problemi attinenti alla geostatistica, mediante l'uso di osservazioni rilevate in specifici siti $\mathbf{s}_i \in D$, viene denominata *previsione* o *interpolazione spaziale*. Ad esempio nelle applicazione minerarie, campo in cui questa disciplina ha avuto un importante sviluppo, dal livello di minerale misurato mediante perforazioni in un numero limitato di siti si ricava la previsione del livello su tutta l'area geografica interessata all'analisi; in maniera simile, la previsione della concentrazione del livello di inquinante in una in-

tera regione geografica viene effettuata usando l'informazione derivante da una rete di stazioni che misurano in quei specifici siti il livello di inquinante.

Più specificatamente, le misure effettuate da una rete di rilevazione permettono di disporre, per ogni sito, di intere serie temporali di misurazioni; in questo modo il fenomeno e la sua analisi si estendono, oltre alla dimensione spaziale, anche a quella temporale. Tale impostazione sarà oggetto dell'analisi proposta nei successivi capitoli.

L'approccio classico alla previsione spaziale si denomina *kriging* e viene formalizzato in questi termini: partendo dalle misurazioni del fenomeno, indicate mediante la v.a. $Z(\mathbf{s}_i)$ per ognuno dei siti $\mathbf{s}_i$ con $i = 1, \ldots, n$, l'obiettivo è quello di determinare il livello del fenomeno $Z(\mathbf{s}_0)$ in un sito $\mathbf{s}_0$ diverso da tutti gli $\mathbf{s}_i$, con $\mathbf{s}_0 \in D$.

La tecnica del *kriging* permette di determinare il valore del predittore $\hat{Z}(\mathbf{s}_0)$ mediante una combinazione lineare pesata dei valori osservati $Z(\mathbf{s}_i)$, con la scelta dei coefficienti tali da ottenere che la stima risultante abbia le proprietà di non-distorsione e di minimizzazione dell'errore medio quadratico di previsione. Per questo motivo, il predittore calcolato mediante *kriging* risulta uno stimatore BLUP (*Best Linear Unbiased Predictor*).

Con questo metodo anche la varianza del predittore può essere determinata mediante una specifica funzione dei pesi e delle osservazioni e, conseguentemente, l'applicazione del metodo si riduce alla determinazione dei pesi ottimali della combinazione lineare.

I pesi sono specificati attraverso la determinazione della struttura della covarianza del campo aleatorio tra le localizzazioni $\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_n$. Nel caso in cui il campo aleatorio sia isotropo, ossia invariante per rotazione, allora i pesi sono funzione delle sole distanze di $\mathbf{s}_0$ dai siti $\mathbf{s}_i$ con $i = 1, \ldots, n$. Intuitivamente, i siti vicini a $\mathbf{s}_0$ avranno una influenza maggiore e, conseguentemente, i pesi della combinazione lineare saranno più elevati rispetto ai siti più lontani.

L'assunzione dell'isotropia consente di semplificare il problema riducendolo alla identificazione di una appropriata funzione monotona della distanza $||h||$ tra i siti. Generalmente, tale funzione viene specificata attraverso una

espressione matematica contenente un certo numero di parametri che vengono stimati attraverso le osservazioni. Oltre alla funzione di covarianza $C(\cdot)$ il *kriging* può utilizzare anche la funzione semivariogramma $\gamma(\cdot)$.

Proprietà del *Kriging*

- il *kriging* è un interpolatore esatto ossia, nei siti osservati, si ha $Z^*(s_i) = Z(s_i)$ e la varianza di previsione risulta pari a 0;

- i pesi del *kriging* sono calcolati mediante l'utilizzo del variogramma; oltre alle distanze tra i siti vengono considerate anche le posizioni relative dei punti;

- la somma dei pesi del *kriging* è uguale a 1, e gli stessi pesi possono assumere anche valori negativi; inoltre, non sono influenzati direttamente dai valori osservati. I valori osservati entrano nel calcolo del variogramma;

- punti lontani ricevono pesi del *kriging* inferiori se sono presenti osservazioni a distanze minori;

- oltre al valore della previsione, il *kriging* permette di calcolare anche la varianza della previsione, dando quindi una misura dell'affidabilità della previsione stessa.

Kriging ordinario (*Ordinary Kriging*)

Supposto che il campo aleatorio sia intrinsecamente stazionario e isotropico, ossia per ogni s si ha $E[Z(s)] = \mu$ e $Var[(Z(s) - Z(s+h))] = 2\gamma(||h||)$, il metodo del *kriging* ordinario porta alla determinazione del previsore nella forma

$$Z^*(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) = \boldsymbol{\lambda}' \boldsymbol{Z} \tag{3.16}$$

Al fine di ottenere lo stimatore non distorto

$$\begin{aligned} E\big[Z^*(s_0)\big] &= E\left[\sum_{i=1}^{n}\lambda_i Z(s_i)\right] \\ &= \sum_{i=1}^{n}\lambda_i E[Z(s_i)] \\ &= \sum_{i=1}^{n}\lambda_i \mu \end{aligned} \tag{3.17}$$

la somma dei pesi deve risultare pari a 1, ossia $\sum_{i=1}^{n}\lambda_i = 1$.

La varianza associata allo stimatore risulta

$$\begin{aligned} \sigma^2_{s_0} &\equiv E\big[Z^*(s_0) - Z(s_0)\big]^2 \\ &= E\left[\sum_{i=1}^{n}\lambda_i\big(Z(s_i) - Z(s_0)\big)\right]^2 \\ &= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\big(Z(s_i) - Z(s_j)\big)^2/2 - \sum_{i=1}^{n}\lambda_i\big(Z(s_i) - Z(s_0)\big)^2\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j E\big[Z(s_i) - Z(s_j)\big]^2/2 - \sum_{i=1}^{n}\lambda_i E\big[Z(s_i) - Z(s_0)\big]^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(||h_{ij}||) - 2\sum_{i=1}^{n}\lambda_i\gamma(||h_{i0}||) \end{aligned} \tag{3.18}$$

dove $\gamma(||h_{ij}||)$ è la funzione semivariogramma, dipendente solo dalla distanza tra i siti s_i e s_j

La determinazione dei pesi (o coefficienti) che minimizzano il valore della varianza si trova attraverso la soluzione del problema di minimo condizionato del sistema di $n+1$ equazioni

$$\begin{cases} \sum_{j=1}^{n}\lambda_j\gamma(||h_{ij}||) + m = \gamma(||h_{i0}||) & \text{con } i = 1,\ldots,n \\ \sum_{j=1}^{n}\lambda_j = 1 \end{cases}$$

ottenuta derivando le varie equazioni e applicando il metodo dei moltiplicatori di Lagrange.

Le stime così determinate usando un adeguata funzione semivariogramma stimata $\hat{\gamma}(\cdots)$, sono quindi date da

$$\widehat{Z}^*(s_0) = \sum_{i=1}^{n} \hat{\lambda}_i \, z_i \tag{3.19}$$

$$\widehat{\sigma}^2_{s_0} = \sum_{i=1}^{n}\sum_{j=1}^{n} \hat{\lambda}_i \hat{\lambda}_j \hat{\gamma}(||h_{ij}||) - 2\sum_{i=1}^{n} \hat{\lambda}_i \hat{\gamma}(||h_{i0}||) \tag{3.20}$$

La soluzione, nel caso di processi stazionari del secondo ordine - ma non nel caso di quelli intrinsecamenti stazionari - può essere espressa anche in termini di funzione di covarianza $C(\cdot)$ nella forma

$$\widehat{\sigma}^2_{s_0} = 2\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i, s_j) - 2\sum_{i=1}^{n} \lambda_i c(s_i, s_0) + Var(Z(s_0)).$$

Kriging universale (*Universal Kriging*)

Nel caso visto sopra, il problema del predittore viene risolto considerando il campo aleatorio costante in media. Tuttavia, nella realtà, molti fenomeni non manifestano questo andamento su tutta la regione di interesse D e diventa necessario stimare l'andamento della funzione $\mu(s)$, media del campo, dipendente dalla localizzazione. Il metodo del *kriging* universale dà soluzione a questo problema, assumendo che il campo aleatorio sia una combinazione di due componenti: una deterministica, dipendente dalla localizzazione e un'altra probabilistica, per la quale viene assunta la stazionarietà del secondo ordine.

In questa impostazione la funzione della media può essere rappresentata come combinazione lineare di funzioni note $f_l(s)$ per cui la media e la covarianza del campo aleatorio possono essere espressi come

$\mu(s) = \sum_{l=1}^{k} \beta_l f_l(s)$

$E\big[\big(Z(s_1) - \mu(s_1)\big)\big(Z(s_2) - \mu(s_2)\big)\big] \equiv E\big[\epsilon(s_1)\epsilon(s_2)\big] = C(s_1 - s_2)$

In analogia a quanto visto per il *kriging* ordinario, il predittore, nel caso del *kriging* universale, assume ancora la forma di una media pesata del livello del fenomeno in un intorno del punto da predire

$$\hat{Z}^*(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{3.21}$$

dove i pesi λ_i sono scelti in maniera da minimizzare l'errore quadratico medio di previsione e che lo stimatore sia non distorto.

La condizione di non distorsione è ottenuta imponendo la condizione $E[Z^*(s_0)] = E[Z(s_0)]$ per cui

$$\mu(s_0) - \sum_{i=1}^{n} \lambda_i \mu(s_i) = 0$$

che può essere riscritta come

$$\sum_{l=1}^{k} \beta_l \big(f_l(s_0) - \sum_{i=1}^{n} \lambda_i f_l(s_i) \big) = 0$$

poiché i coefficienti β_l sono generalmente diversi da 0, l'equazione si può esprimere come

$$f_l(s_0) = \sum_{i=1}^{n} \lambda_i f_l(s_i) \quad \text{per } l = 1, \ldots, k \tag{3.22}$$

La varianza dello stimatore calcolato mediante il *kriging* universale, in termini di funzione di covarianza $C(\cdot)$ - per un processo stazionario del secondo

ordine - può essere espressa come

$$\begin{aligned}\sigma^2_{s_0} &\equiv E\big[Z^*(s_0) - Z(s_0)\big]^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(s_i, s_j) - 2\sum_{i=1}^{n}\lambda_i C(s_i, s_0) + Var(Z(s_0)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(s_i - s_j) - 2\sum_{i=1}^{n}\lambda_i C(s_i - s_0) + Var(Z(s_0)) \quad (3.23)\end{aligned}$$

L'equazione 3.23 può essere minimizzata scegliendo dei pesi λ_i appropriati, attraverso il metodo dei moltiplicatori di Lagrange. Il sistema delle $n + k$ equazioni lineari da derivare, si esprime come

$$\begin{cases} \sum_{j=1}^{n}\lambda_j C(s_i - s_j) + m_l f_l(s_i) = C(s_i - s_0) \quad \text{con } i = 1, \ldots, n \\ \sum_{j=1}^{n}\lambda_j f_l(s_j) = f_l(s_0) \quad \text{con } l = 1, \ldots, k \end{cases}$$

Kriging di funzioni indicatrici (*Indicator Kriging*)

Qualora si sia interessati a costruire mappe di rischio in cui si esprime la probabilità di superare un dato valore di soglia, il metodo del *Kriging* di funzioni indicatrici fornisce una previsione della probabilità di non superare tale soglia, e di conseguenza, attraverso il complemento a 1, anche quella di superarla.

In pratica, denotando con c il valore soglia, si considera il processo spaziale ottenuto tramite la funzione indicatrice $\mathbb{I}_{Z(s)\leq c}$ per il quale si ha

$$E[\mathbb{I}_{Z(s)\leq c}] = Pr[Z(s) \leq c] = F_Z(c)$$

Attraverso la trasformazione dei dati $Z(s_i)$ secondo la variabile indicatrice sopradescritta si ottengono i dati binari $\mathbb{I}_{Z(s_1)\leq c}, \ldots, \mathbb{I}_{Z(s_n)\leq c}$ e attraverso l'applicazione del metodo del *kriging* ordinario agli stessi, si ottiene la previsione mediante il predittore lineare

$$I(c, s_0) = \sum_{i=1}^{n}\lambda_i \mathbb{I}_{Z(s_i)\leq c}$$

con il vincolo $\sum_{i=1}^{n} \lambda_i = 1$

Da notare che quest'ultimo fornisce una stima per la media condizionata ai dati binari

$$E[\mathbb{I}_{Z(s_0)\leq c}|\mathbb{I}_{Z(s_1)\leq c}, \ldots, \mathbb{I}_{Z(s_n)\leq c}] \neq E[\mathbb{I}_{Z(s_0)\leq c}|\boldsymbol{Z}] = Pr[Z(S_0) \leq c|\boldsymbol{Z}]$$

e non ai dati originali con una conseguente perdita di informazione.

La fase di convalida del modello

Risulta utile nella scelta del modello poter confrontare quello che consente un miglior adattamento dello stesso al fenomeno reale analizzato. Una tecnica euristica che valuta la bontà del modello, la convalida incrociata, o *cross validation*, consente di ottenere una stima $\widehat{Z}(s_i)$ del valore in un sito osservato s_i tramite l'utilizzo di tutti gli altri valori, e iterando la procedura, si ottengono le stime per ognuno dei siti in cui è stata effettuata la misurazione. Le stime così ottenute vengono confrontate con i valori osservati $Z(s_i)$ per ricavare i residui. Un indice che consente di confrontare i modelli utilizzati è dato da

$$CV = \frac{1}{n} \sum_{i=1}^{n} \frac{e_i^2}{Var(e_i)} \tag{3.24}$$

dove gli $e_i = \widehat{Z}(s_i) - Z(s_i)$. Se il valore ottenuto dell'indice $CV \approx 1$, il modello per il variogramma e le stime ottenute mediante uno dei metodi visti precedentemente, risultano appropriati.

3.2 L'analisi spaziale della concentrazione media annuale

3.2.1 Descrizione dei dati

Viene ora proposta l'analisi del livello di concentrazione di PM_{10}, conformemente a quanto descritto nel paragrafo precedente. Verranno applicati i metodi di interpolazione spaziale al fine di ottenere una previsione dell'an-

damento spaziale del fenomeno anche nei punti in cui il processo non è stato rilevato.

In questa prima fase si analizza il livello medio annuale della concentrazione di PM_{10} procedendo all'aggregazione dei dati validi osservati giornalmente, nel corso dell'anno 2006, tramite la media aritmetica; si prescinde quindi dall'aspetto temporale del processo, che ha bisogno di un approccio più complesso e articolato, che sarà oggetto dei prossimi due capitoli.

L'analisi geostatistica sui valori della concentrazione media annua di PM_{10} comprende esclusivamente le stazioni in cui il numero di valori validi osservati risulta maggiore del 90% (vedi Tab. 2.1), come previsto dalla normativa per la raccolta minima dei dati; quindi la serie storica comprende almeno 328 valori gionalieri.

Le centraline risultanti sono 17 sulle 27 attive nell'anno 2006. Ciò è dovuto in parte a rotture o malfunzionamenti dello strumento di rilevazione, ma soprattutto a causa dell'attivazione, di alcune di queste, solo a partire dalla tarda primavera[1]. Si è ritenuto di comprendere anche la centralina RO1 che dispone di 326 osservazioni giornaliere, solo 2 in meno rispetto al limite.

Le stazioni su cui viene svolta l'analisi geostatistica, risultano più concentrate nella parte centrale della regione Veneto, visto che sono escluse 3 centraline della provincia di Venezia situate a Chioggia (VE4) e nella parte orientale, verso il Friuli, a San Donà di Piave (VE5) e Concordia Saggittaria (VE6). Vengono a mancare anche le centraline collocate in provincia di Padova a Este (PD4) verso sud, e a Cittadella (PD5) verso nord; anche la vicina centralina di Bassano del Grappa (VI4) non soddisfa i requisiti sulla numerosità. Vengono escluse anche 2 centraline in provincia di Rovigo (RO2 e RO3); infine è esclusa anche la centralina situata nell'Alpago, in provincia di Belluno (BL3), la quale comunque, rispetto a tutte le altre situate in zona di pianura, si colloca ad una altitudine molto maggiore e probabilmente anche in un contesto ambientale e territoriale molto diverso.

Riprendendo la formulazione adottata nel paragrafo precedente, i siti in cui sono localizzate le stazioni vengono indicati da s_i con $i = 1, \ldots, 17$ e con $Z(s_i)$ il livello di concentrazione medio annuo di PM_{10}.

[1] L'andamento delle serie storiche di ogni stazione è presente nell'Appendice A

Una prima analisi, presentata nella figura 3.3, consente, nella mappa di

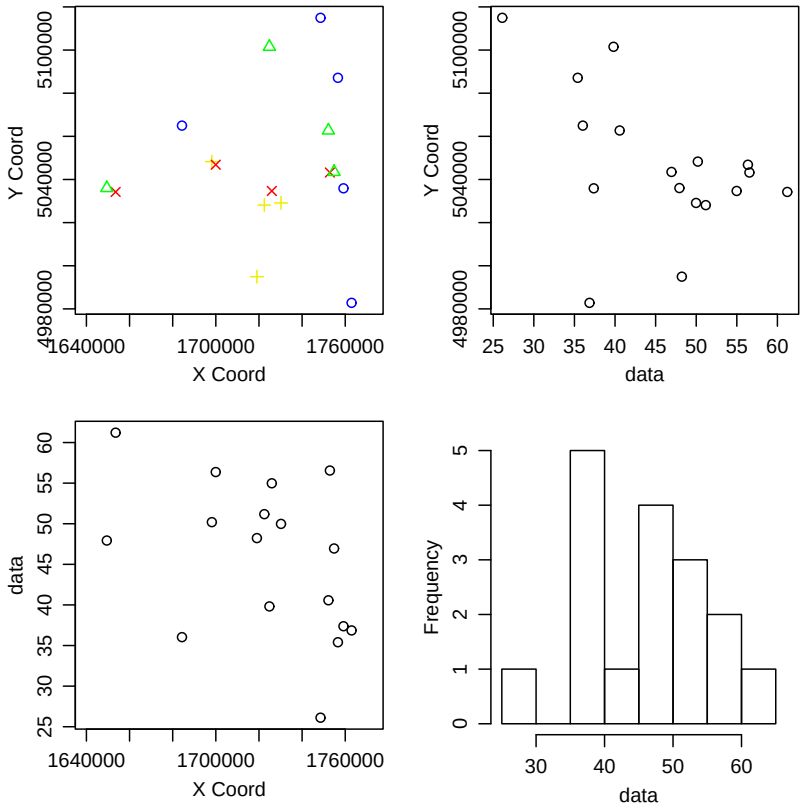


Figura 3.3: Localizzazione e andamento livello PM_{10}

localizzazione in alto a sinistra, di osservare come i valori medi annuali del particolato siano disposti in maniera simmetrica rispetto ad un'asse centrale, corrispondente d'altra parte alle due principali arterie di traffico - l'autostrada A4 e la ferrovia Ve/Mi - che va da Mestre a Verona, passando per Padova e Vicenza (indicati con il simbolo 'x' relativo a valori più elevati, seguito da quelli connotati con '+'), per decrescere man mano che ci si sposta verso la provincia di Rovigo in direzione sud e verso le provincie di Treviso/Belluno/Venezia orientale in direzione nord e nord-est, dove sono presenti concentrazioni di inquinante più bassi (simboli '△' e ∘).

L'istogramma in basso a destra, in cui non viene presa in considerazione la componente spaziale, presenta una distribuzione bimodale, in corrispondenza di valori appena inferiori a 40 e valori appena inferiori a 50, che sembra indicare come il processo sia individuato da una mistura di distribuzioni.

Analisi delle correlazioni

L'analisi delle correlazioni tra le coppie di centraline con almeno il 90% di valori osservati, presentate nella Tab. A.1 dell'Appendice A, consente di osservare come quelle più alte siano riferite a centraline spazialmente più vicine o che siano locate in zone con caratteristiche simili. Infatti, ad esempio, la centralina PD1 (tipo *background* TU) oltre a risultare molto correlata con PD2 e PD3 (sono posizionate nella città di Padova rispettivamente nel quartieri Arcella, Mandria e in prossimità della Zona Industriale a Granze) presenta una alta correlazione anche con TV1 (Conegliano - BU), TV2 (Treviso - BU), VE1 (Mestre Circonvallazione - TU) e VI1 (Vicenza - BU).

Questo potrebbe far presumere che il livello di concentrazione di PM_{10} abbia un comune denominatore indipendentemente da una specifica localizzazione spaziale.

3.2.2 Stima del variogramma

Come visto al paragrafo 3.1.2, per analizzare il legame tra le osservazioni del fenomeno nello spazio si fa riscorso al variogramma. L'analisi grafica della dipendenza spaziale si può avere attraverso la sua stima, ottenuta costruendo la nuvola dei valori per ogni coppia di punti di $[Z(s_i) - Z(s_j)]^2$ per $i, j = 1, \ldots, 17$ con $i \neq j$, presentata in Fig. 3.4, dove non sembrano manifestarsi evidenti valori o andamenti anomali. Nello stesso grafico si evidenziano anche i variogrammi empirici classico (3.12) e 'robusto', (3.13) calcolati mediante il metodo dei momenti.

La distanza massima per il calcolo del variogramma viene fissata poco oltre i 120 km, che è approssimativamente la distanza che intercorre tra Venezia e Verona, ossia l'asse lungo il quale ci sono i valori più elevati e con condizioni ambientali e territoriali simili. L'andamento del variogramma stimato sembra far pensare che le ipotesi sulla stazionarietà di secondo ordine non siano soddisfatte, mentre sembra presentare le caratteristiche tipiche per un processo intrinsecamente stazionario.

L'analisi del variogramma direzionale, per scoprire eventuali comportamenti difformi in direzioni diverse e controllare l'anisotropia, non manifesta evi-

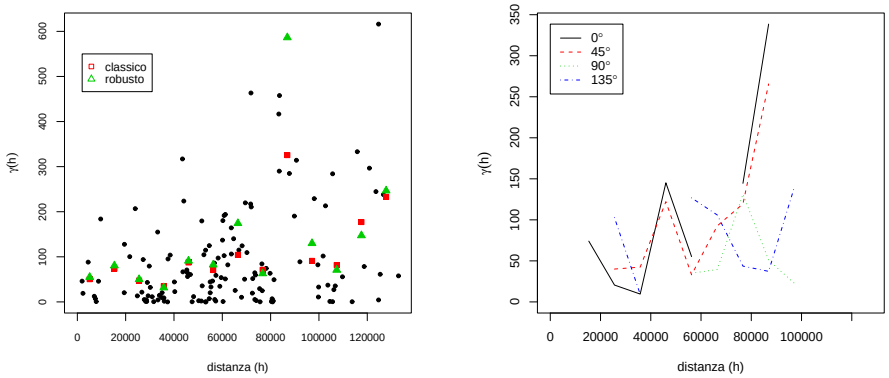


Figura 3.4: (semi)variogrammi nuvola-empirico e direzionale - anno 2006

Coefficienti	Valore stimato	t value
(Intercept)	8.567e+02	2.804
Latitudine (s_x)	-1.125e-04	-2.038
Longitudine (s_y)	-1.223e-05	-2.073
Coeff.correlazione R^2	0.3955	

Tabella 3.1: Stime dei parametri modello lineare

denti differenze di andamento o di aumento di variabilità di alcune direzioni rispetto alle altre, ragion per cui si ritiene di poter affermare che il fenomeno non presenta una componente di trend nello spazio.

Sembra opportuno sottolineare a proposito dei variogrammi empirici sopradescritti, come il numero di coppie di siti, da cui si ricavano le distanze e i valori del quadrato delle differenze siano un numero esiguo e che di conseguenza i valori con cui viene determinata la media all'interno dell'intervallo, soprattutto per le distanze maggiori, sia assolutamente esiguo, con una conseguente limitata affidabilità delle stime stesse.

3.2.3 Costruzione del modello

Completata l'analisi del variogramma empirico, il passo successivo consiste nel costruire il modello per descrivere la correlazione spaziale.

Il modello ipotizzato in 3.15 prevede la scomposizione della realizzazione del fenomeno $Z(\mathbf{s})$ tra componente di larga scala $\mu(\mathbf{s})$ e componente di piccola scala $\epsilon(\mathbf{s})$. Per quanto riguarda la componente deterministica ipotizzando un

trend lineare $\mu(\mathbf{s}) = \beta_0 + \beta_1 s_x + \beta_2 s_y$ e stimando i parametri mediante un modello lineare si ottengono i valori presentati nella tabella 3.1.

Il coefficiente relativo all'intercetta risulta significativo, ma non risultano significativi al livello di confidenza del 5% nessuno dei due coefficienti relativi alle componenti spaziali, Longitudine (direzione ovest-est) e Latitudine (direzione sud-nord); si rafforza così la considerazione per cui la parte deterministica non è influenzata da una presenza di trend lineare. L'analisi di una dipendenza non lineare - assumendo quindi componenti quadratiche - rispetto alle coordinate spaziali non ha portato a nessun miglioramento del modello.

3.2.4 L'identificazione

Per poter procedere ad una previsione nello spazio tramite *kriging* deve essere trovato il modello di variogramma teorico che meglio interpreta la struttura di covarianza spaziale stimata nel paragrafo precedente.

Sono stati analizzati diversi modelli teorici del semivariogramma, con le stime dei parametri calcolate secondo il metodo dei minimi quadrati generalizzati e il metodo della massima verosimiglianza; l'andamento viene visualizzato nella Fig. 3.5. Il modello che meglio sembra interpretare il variogramma empirico sembra essere quello di tipo esponenziale. La selezione del modello viene eseguita mediante il valore AIC e tramite convalida incrociata.

I parametri dei variogrammi calcolati - vedi Tabb. 3.2 e 3.3 - tramite il metodo della massima verosimiglianza con l'assunzione, quindi, della normalità della distribuzione di $Z(\mathbf{s})$, che sembra meglio adattarsi al variogramma empirico, mentre quello calcolato con il metodo dei minimi quadrati generalizzati, pur in assenza di assunzioni circa la distribuzione, viene troppo influenzato dai valori elevati presenti a distanze maggiori.

Covarianza	nugget	sill	range	SQM
esponenziale	32,12	35732,63	$2{,}80e+7$	26,6071
gaussiano	43,10	13192,46	$9{,}46e+6$	21,8866

Tabella 3.2: Parametri dei variogrammi teorici - stime WLS

Nel calcolo delle stime è sembrato opportuno considerare l'effetto nugget, che si può interpretare come componente di fondo dell'inquinamento[2]; il

[2]Cirillo, APAT 2003 stima una media nazionale del livello del PM_{10} per le stazioni di fondo - parchi, isole pedonali, ecc. - pari a 26 $\mu g/m^3$; la media europea risulta, invece,

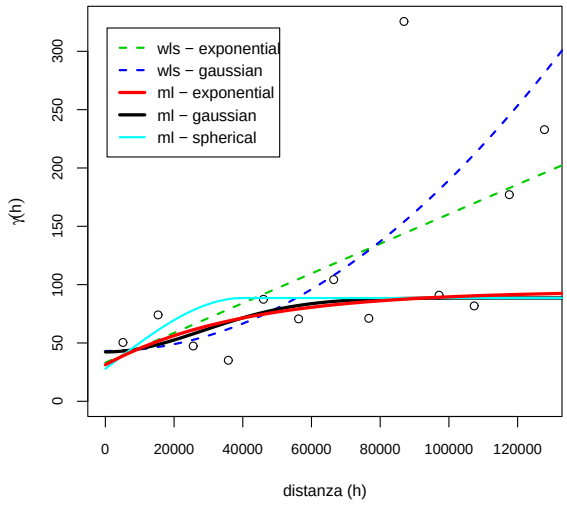


Figura 3.5: (semi)variogrammi teorici

Covarianza	nugget	sill	range	mean	log-Lik	AIC
Esponenziale	31,30	63,64	40000	42,61	-60,62	129,2
Gaussiano	42,41	46,23	40000	43,28	-60,51	129,0
Sferico	27,85	60.69	40000	43,57	-60,96	129,9

Tabella 3.3: Parametri dei variogrammi teorici - stime ML

variogramma di tipo esponenziale consente inoltre di mantenere una influenza sul fenomeno per distanze relativamente maggiori rispetto agli altri tipi.

Previsione

Per la previsione del livello medio della concentrazione di PM_{10} viene applicato il metodo del *kriging* ordinario, visto che l'analisi non evidenzia la presenza di *trend*.

La previsione viene realizzata costruendo una fitta griglia regolare che copre l'intera regione del Veneto e in ognuno di questi siti equispaziati viene determinato il valore dell'inquinante tramite la soluzione del sistema del *kriging* ordinario (vedi par. 3.1.7).

La mappa della media annuale per l'anno 2006 nel Veneto, presentata in

pari a 17 $\mu g/m^3$

Fig. 3.6, mostra come le concentrazioni maggiori siano localizzate nelle aree urbane corrispondenti alle città di Padova, Vicenza e soprattutto Verona, mentre degrada verso valori inferiori quando ci si allontana, in maniera molto più accentuata verso nord-est nelle zone del trevigiano, del bellunese e nella parte nord-orientale della provincia di Venezia; in maniera meno evidente la stessa diminuzione si manifesta anche verso sud nella provincia di Rovigo.

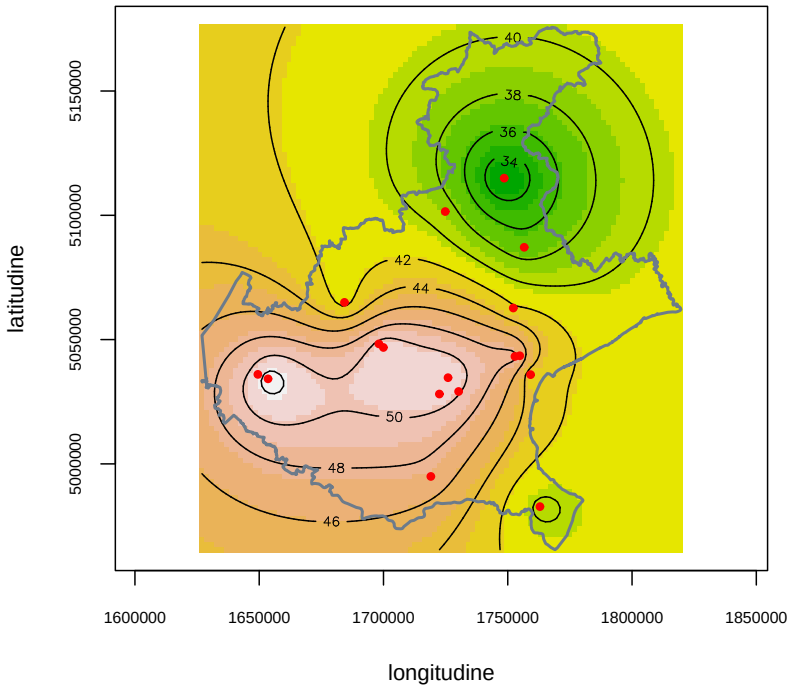


Figura 3.6: Previsione della media annuale - *Ordinary Kriging*

La previsione viene accompagnata anche dalla stima della deviazione standard della stessa in Fig. 3.7, dove si nota la maggior precisione in prossimità dei siti in cui il valore è stato rilevato.

Vista la mancanza di osservazioni e la caratterizzazione orografica molto diversa rispetto all'area situata in pianura, risulta evidente come la previsione non risulti affidabile per le zone montane del veronese e del vicentino e in particolar modo per tutta la parte nord della regione.

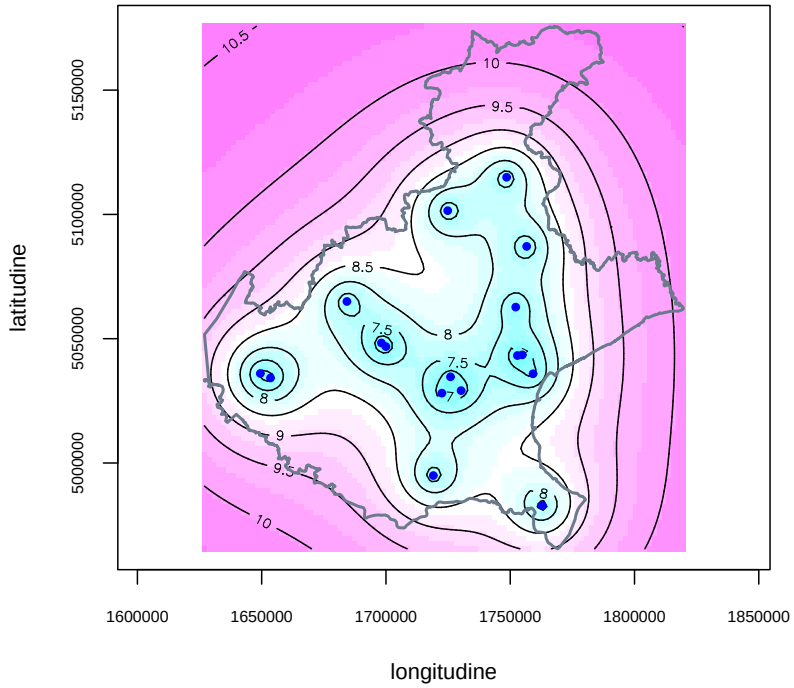


Figura 3.7: Standard error sulla previsione della media annuale - *Ordinary Kriging*

Convalida del modello

L'applicazione della convalida incrociata per la previsione ottenuta mediante l'applicazione del *kriging* ordinario e del modello di variogramma esponenziale porta ad un valore dell'indice CV pari a 1.0096; si è in presenza quindi di una leggera sovrastima. Anche l'analisi diagnostica grafica sembra supportare la scelta effettuata e affermare che il modello, pur senza tenere in considerazione l'aspetto temporale, fornisce una buona indicazione dell'andamento del fenomeno.

Kriging mediante funzioni indicatrici

Per la predizione spaziale sulla probabilità di superamento del limite stabilito dalla normativa sulla media annua di PM_{10}, viene utilizzato il metodo non lineare del *kriging* di funzioni indicatrici, visualizzato nella Fig. 3.8, ottenuto mediante la stessa struttura di covarianza spaziale stimata per il *kriging* ordinario. Come si può notare la probabilità di superamento è risul-

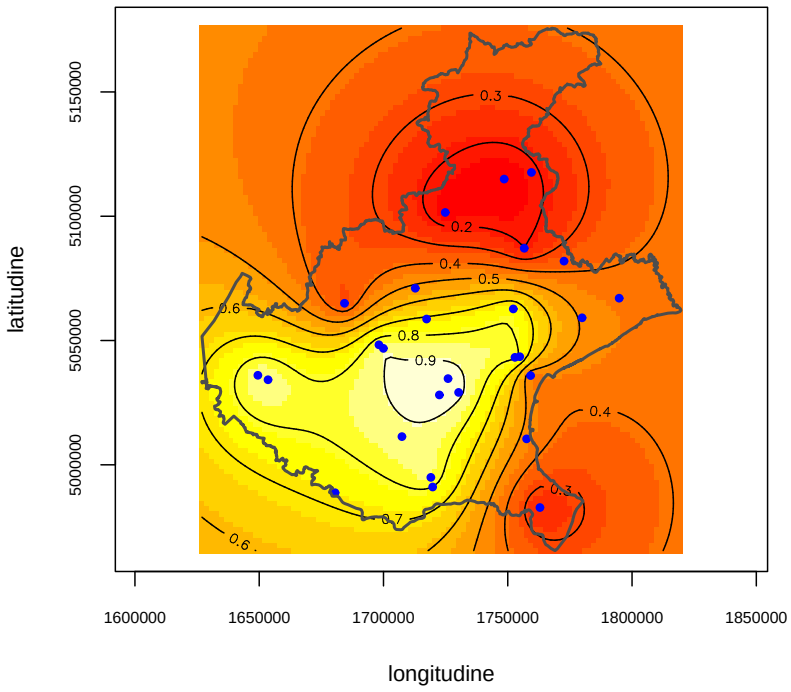


Figura 3.8: *Kriging* di funzioni indicatrici per superamento soglia 40 μg/mc (media annuale)

tata molto alta in tutta la parte della pianura con valori prossimi a 1 nella zona compresa tra Padova e Vicenza, nonchè a Verona.

Capitolo 4

Modellazione spazio-temporale mediante componenti di trend deterministiche

4.1 Introduzione

Per i fenomeni ambientali come la concentrazione di inquinanti in un'area geografica, l'approccio e l'analisi eseguiti mediante gli strumenti visti nei capitoli precedenti appaiono quanto mai limitati; in essi viene a mancare la componente che considera l'andamento del fenomeno anche in dipendenza dell'evoluzione nel tempo, sia per la parte di interpretazione sia per quella che consente di effettuare delle previsioni a proposito del fenomeno stesso.

Nel corso dell'ultimo decennio si è avuto un notevole sviluppo delle tecniche per l'analisi di dati spazio-temporali e dei modelli statistici che le implementano. Tali modelli, sia per la scoperta e descrizione di *pattern* significativi sia per la previsione, vengono utilizzati da varie discipline, come le scienze ambientali, quali l'idrologia, la meterologia, la geologia, estendendosi inoltre a nuovi contesti, come le mappe epidemiologiche o il monitoraggio dei beni immobili in aree geografiche.

Spesso l'interesse principale nell'analisi di tali dati consiste nel catturare l'andamento di fondo e predire l'evoluzione nel tempo di alcune variabili su un definito dominio geografico. Queste previsioni sono effettuate su *dataset* contenenti, assieme ai valori misurati del fenomeno, altre variabili, che a loro

volta possono esprimersi in funzione della posizione nel dominio spaziale e del momento temporale in cui sono rilevate, anche se non caratterizzate da una struttura stocastica.

Al fine di ottenere un alto grado di accuratezza nell'analisi e nella previsione della variabile oggetto di studio, sono necessari modelli matematici che includano la dinamica complessiva del fenomeno, sia spaziale che temporale.

Riassumendo, la modellazione dei dati spazio-temporali è una impresa impegnativa, la quale richiede l'elaborazione di *dataset* di notevoli dimensioni e, inoltre, l'abilità e capacità di adattare, al fenomeno di interesse, modelli complessi ma al tempo stesso realistici. Molto spesso le soluzioni richieste non sono esprimibili matematicamente in forma esplicita e richiedono l'impiego di metodi computazionalmente onerosi. Non da ultimo, lo stesso fenomeno può essere analizzato con metodi e con assunti diversi, dipendenti dagli aspetti che si desiderano mettere in maggior evidenza.

4.2 La modellazione del processo

La teoria della statistica spaziale, vista nel capitolo precedente, rappresenta e modella fenomeni in cui si è in presenza di una singola realizzazione proveniente da un processo stocastico correlato spazialmente e in cui non viene considerata la dipendenza temporale.
Le più recenti ed evolute teorie sui processi spazio-temporali, presentati nel prossimo capitolo, modellano fenomeni in cui i dati sono raccolti in diverse stazioni - solitamente costituenti una rete di rilevazione - e per lunghi periodi di tempo, per cui la struttura della variabilità del processo viene vista in funzione di entrambe le dimensioni.

Tra questi due approcci, esiste una classe di processi nei quali la componente casuale risulta spazialmente, ma non temporalmente, correlata; con questa ipotesi, le osservazioni registrate nel corso del tempo possono essere ipotizzate come replicazioni dello stesso processo spaziale.

Prendendo spunto dal lavoro di Smith, Kolenikov e Cox [19], in cui si ipotizza che la concentrazione del livello di inquinanti possa appropriatamente essere rappresentata da processi di questo tipo, si estende l'analisi dei dati del PM_{10} rilevati nell'anno 2006 dall'ARPAV mediante tale modellazione che consente

di valutare nel corso del tempo l'andamento del processo, rappresentato dalla variabile casuale Z. Il modello di partenza risulta

$$Z(\boldsymbol{s},t) = \mu(\boldsymbol{s},t) + \epsilon(\boldsymbol{s},t) \tag{4.1}$$

in cui la parte deterministica $\mu(\boldsymbol{s},t)$, di larga scala, viene stimata mediante metodi non parametrici che incorporano sia la dimensione spaziale che quella temporale; per gli errori $\epsilon(\boldsymbol{s},t)$, ossia la seconda componenente del modello, si assume, al momento, che provengano da un processo a media nulla e che abbiano varianza finita.

Per i residui del modello $\hat{\epsilon}(\boldsymbol{s}) = z^{oss}(\boldsymbol{s},t) - \hat{\mu}(\boldsymbol{s},t)$, si procede in modo analogo a quanto visto nel Capitolo 3, determinando la struttura di covarianza che meglio li interpreti attraverso l'uso degli strumenti della geostatistica.

Se consideriamo quanto previsto dalla normativa, l'analisi svolta nel capitolo precedente concentra l'attenzione dapprima, su come il fenomeno si comporta nello spazio rispetto al valore medio annuale, in cui non compare nessuna indicazione circa l'aspetto temporale; successivamente rispetto alla probabilità che il limite determinato - pari ad un valore medio annuo di 40 $\mu g/m^3$ - possa essere superato o meno e in quali aree della regione. La stessa normativa considera un aspetto temporale quando indica che il valore medio giornaliero - pari a 50 $\mu g/m^3$ - non deve essere superato per più di 35 giorni nel corso dell'intero anno. Con questa prima estensione si tende ad utilizzare maggiormente il contenuto informativo fornito dalla rete di rilevamento andando a vedere, per esempio, quali possano essere i periodi e le aree geografiche in cui tale limite tende ad essere raggiunto e superato.

Anche con questa impostazione rimangono limiti che attengono fondamentalmente all'assunzione della stazionarietà nello spazio e il considerare la dimensione temporale esclusivamente deterministica, non considerandone la variabilità. Nel prossimo capitolo, tramite approcci più complessi e flessibili, si adotteranno modelli che superano questi limiti.

4.2.1 Metodi non parametrici

Modelli additivi

Al fine di superare ed estendere i vincoli e limiti dati dal modello lineare e di modellare componenti non lineari sono stati introdotte, tra gli anni '80 e '90, classi di modelli più flessibili.

L'utilizzo dei modelli additivi consente di considerare il contributo della generica variabile X_j non più di tipo lineare $\beta_j X_i$, ma di tipo più generale $f_j(X_j)$. Le funzioni f_j possono essere stimate mediante una serie di interpolazioni locali su funzioni lineari (ad esempio il *loess*) o polinominali (ad esempio polinomi di terzo grado nel caso delle *spline cubiche*). Il principio che governa la costruzione di questo metodo è dato dalla segmentazione dei dati, per diversi valori dell'asse delle ascisse, seguita da una interpolazione locale; tanto maggiore è la segmentazione tanto più il risultato dell'interpolazione effettuata cercherà di raggiungere i dati osservati, per cui risulta opportuna la ricerca di un grado corretto di lisciamento che può dipendere anche dagli scopi che l'analisi - dei dati e del fenomeno - si prefigge.

Il modello di regressione lineare classico $Y = f(\mathbf{X}) + \epsilon$ può essere generalizzato mediante $Y = f(X_1 + X_2 + \ldots + X_p) + \epsilon$, dove con f non si indica una specifica, definita e univoca funzione parametrica. L'estensione del modello considera un effetto dei regressori di tipo 'additivo' e le f_j sono funzioni arbitrarie anche diverse per ogni singolo predittore.

Il modello additivo generalmente si esprime come:

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon \tag{4.2}$$

Le ipotesi sottostanti sono le stesse dei modelli parametrici, in cui la componente casuale di errore ϵ risulta indipendente dalle variabili X_j; inoltre $E[\epsilon] = 0$ e $Var[\epsilon] = \sigma^2$.

Al fine di evitare il problema di identificabilità del modello, le varie f_j devono essere centrate sullo 0, ossia risulta implicita nella 4.2 l'assunzione che $E\{f_j(X_j)\} = 0$.

Per arrivare ad una stima delle funzioni relative al modello additivo come definito dalla 4.2 si usa una procedura iterativa che si appoggia ad un metodo di stima non parametrica di funzioni in una variabile per determinare le f_j. Tale procedura, detta di *backfitting* (Buja, Hastie e Tibshirani, 1989) - vedi Tab.4.1 - consente di scegliere tra lisciatori (*smoothers*), indicati con S, e conseguentemente tra metodi di stima diversi, per le diverse funzioni f_j, anche se usualmente si usa lo stesso lisciatore per tutte le f_j, come ad esempio le funzioni *spline* presentate in seguito.

1. Inizializzazione $\hat{\alpha} \leftarrow \sum_i y_i/n,\ \hat{f}_i \leftarrow 0$ per ogni j
2. Ciclo per $j = 1, 2, \ldots, p, 1, 2, \ldots, p, \ldots$:
   (a) $\hat{f}_j \leftarrow S\left[\{y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})\}_1^n\right]$
   (b) $\hat{f}_j \leftarrow \hat{f}_j - n^{-1} \sum_{i=1}^n \hat{f}_j(x_{ij})$

   fino a che le funzioni $\hat{f}_j$ si stabilizzano

Tabella 4.1: Algoritmo di *backfitting*

Spline

Con il termine *spline*, in matematica, si intende una funzione polinomiale a tratti usata per approssimare funzioni di cui si conosce il valore solo in alcuni punti. Considerando funzioni in $\mathbb{R}$, nell'asse delle ascisse vengono scelti K punti $\xi_1 < \xi_2 < \ldots < \xi_K$ detti *nodi*, attraverso i quali la funzione viene vincolata, mentre negli altri punti è libera. Tra i due punti $\xi_i < \xi_{i+1}$ la curva $f(x)$ coincide con un appropriato polinomio di grado prefissato d e si richiede una certa forma di regolarità della $f(x)$ nei punti di 'giunzione' $\xi_i (i = 1, \ldots, K-1)$, ossia che sia continua e che abbia derivate continue fino al grado $d-1$.
Solitamente d è pari a 3, da cui il nome di *spline cubica*, e la funzione $f(x)$ risulta continua con derivata prima e seconda continua.

Spline di regressione

In ambito statistico si studia la relazione tra una variabile esplicativa x e

una variabile risposta y in cui, partendo dal modello $y = f(x;\beta) + \epsilon$, la parte $f(x;\beta)$ viene considerata come una funzione di tipo *spline*. L'asse delle ascisse, o variabile concomitante x, viene suddiviso in $K+1$ intervalli corrispondenti ai punti $\xi_1, \ldots, \xi_K$ - i nodi - e gli n punti - le osservazioni - vengono interpolati mediante un criterio di minimizzazione in cui i β risultano essere i parametri dei $K+1$ polinomi costituenti la curva.

Utilizzando *spline* cubiche, il numero totale di parametri delle cubiche risulta $4(K+1)$ soggetti a $3K$ vincoli di continuità; ne risulta che β ha $K+4$ componenti.

La soluzione del problema di minimizzazione può essere scritta nella forma

$$f(x;\beta) = \sum_{j=1}^{K+4} \hat{\beta}_j h_j(x)$$

dove

$$h_j(x) = x^{j-1} \text{ per } j = 1, \ldots, 4$$

$$h_{j+4}(x) = (x - \xi_j)^3_+ \text{ per } j = 1, \ldots, K$$

e $a_+ = max(a, 0)$. La soluzione è costituita da una opportuna combinazione lineare di una base di funzioni $\{h_j(x), j = 1, \ldots, K+4\}$, costituita in parte da polinomi elementari e in parte da funzioni di tipo $max(0, (x-\xi)^3)$.

Spline di lisciamento

Un diverso utilizzo delle funzioni di tipo *spline* per lo studio della relazione tra variabili è quello che consente di introdurre un approccio alla stima non parametrica.

Si consideri, a tal fine, il criterio dei minimi quadrati penalizzati, rispetto a f,

$$D(f, \lambda) = \sum_{i=1}^{n} \left[y_i - f(x_i)\right]^2 + \lambda \int_{\mathbb{R}} \{f''(t)\}^2 \, \mathrm{d}t$$

in cui λ risulta un parametro di penalizzazione del grado di irregolarità della curva interpolante f ed agisce quindi come parametro di lisciamento.

Per $\lambda \to 0$ non si manifesta nessuna penalità e la soluzione di minimo $\hat{f}(x_i)$ risulta essere la media aritmetica delle y_i corripondenti ad ogni data ascissa; nel caso $\lambda \to \infty$ la penalità risulta massima e il risultato che si ottiene è la retta dei minimi quadrati.

Spline* in due dimensioni: funzioni *Thin-plate Spline

Una estensione delle *spline* cubiche al caso in due dimensioni sono le funzioni *thin-plate spline*, che si ottengono da una generalizzazione delle *spline* di lisciamento viste sopra.

I parametri vengono, calcolati, per un fissato valore di λ, in maniera da minimizzare l'espressione

$$D(f,\lambda) = \sum_{i=1}^{n}\sum_{k=1}^{2}\big(y_{ik} - f_j(s_i)\big)^2 + \lambda\big[J_2(f_1) + J_2(f_2)\big]$$

dove le y_{i1} e y_{i2} sono le due coordinate di y_i e J_2, misura del grado di lisciamento delle funzioni f_j, è definito come

$$J_2(f_j) = \int_{\mathbb{R}^2}\left[\Big(\frac{\partial^2 f_j}{\partial y_1^2}\Big)^2 + 2\Big(\frac{\partial^2 f_j}{\partial y_1 \partial y_2}\Big)^2 + \Big(\frac{\partial^2 f_j}{\partial y_2^2}\Big)^2\right] dy_1\, dy_2 \quad \text{per} \quad j = 1, 2.$$

PROPRIETÀ

La scelta delle *spline* come funzioni usate per la stima delle f_j nei modelli additivi si giustifica sulla base di alcune proprietà:

• Le *spline* cubiche naturali risultano essere gli interpolatori con maggior grado di lisciamento (Green e Silverman - 1994).

• L'interpolazione effettuata tramite le *spline* cubiche risulta ottima (de Boor - 1978) e comunque migliore rispetto ad altre forme funzionali.

4.3 L'analisi del livello di concentrazione del PM_{10}

4.3.1 La riorganizzazione dei dati

I dati messi a disposizione dall'ARPAV presentano alcune problematiche, riscontrabili con una certa regolarità nel caso di fenomeni ambientali, che devono essere considerate prima di procedere con l'analisi. Inoltre, l'utilizzo dei modelli non parametrici, prevede, come per quelli lineari, l'assunzione sugli errori delle ipotesi del secondo ordine ed, eventualmente, sulla normalità della loro distribuzione.

I dati del livello di concentrazione di PM_{10} assumono valori strettamente

positivi e, come già evidenziato nel Capitolo 2, presentano una distribuzione asimmetrica, con lunghe code verso valori elevati; si è anche accennato al fatto che i valori bassi di concentrazione, a causa della precisione degli strumenti di misurazione, sono sottoposti a censura. Oltre a ciò, in ogni stazione il fenomeno presenta sia valori medi diversi, sia diversa variabilità.

Per poter procedere alla modellazione con un andamento dei dati maggiormente simmetrico nella distribuzione e poter avere inoltre, nelle diverse stazioni di rilevazione, una variabilità analoga, si procede alla trasformazione dei dati - mediante le funzioni logaritmo e radice quadrata - che possa rendere la distribuzione maggiormente corripondente ai requisiti sopra descritti.

I dati forniti hanno una cadenza giornaliera e si manifestano con una forte correlazione temporale; nelle varie serie di valori relative alle stazioni di rilevamento ci sono dei valori mancanti, specialmente dovuti ai momenti diversi di attivazione delle centraline di rilevazione.

Al fine di ottenere delle serie di dati con un grado di correlazione nel tempo minore e per ridurre la numerosità di *missing values* nei dati da analizzare, in analogia con quanto effettuato nei lavori di Smith *et al.* [19] e Sahu *et al.* [15], si procede alla loro aggregazione in medie settimanali, effettuata mediante la media aritmetica dei valori intercorrenti tra lunedì e domenica. L'eventuale effetto di distorsione dovuto all'operazione di aggregazione, non viene considerato anche in relazione al fatto che lo scopo della stessa risulta quello di analizzare il fenomeno focalizzando l'attenzione sulla variabilità spaziale.

Le serie storiche dei valori medi settimanali, relativi alla concentrazione di PM_{10} presentano, così aggregati, una numerosità di valori mancanti molto più bassa.
L'andamento delle serie storiche ottenute e la visualizzazione della correlazione temporale, di cui in Fig. 4.1 si riportano due esempi, rende evidente come rispetto alle serie giornaliere originarie ci sia una minore variabilità e come la correlazione decresca rapidamente; per lag > 1, ovvero per osservazioni distanti 2 settimane, si può pensare che il processo sia incorrelato temporalmente.

Per valutare se l'operazione di aggregazione effettuata sui dati consenta,

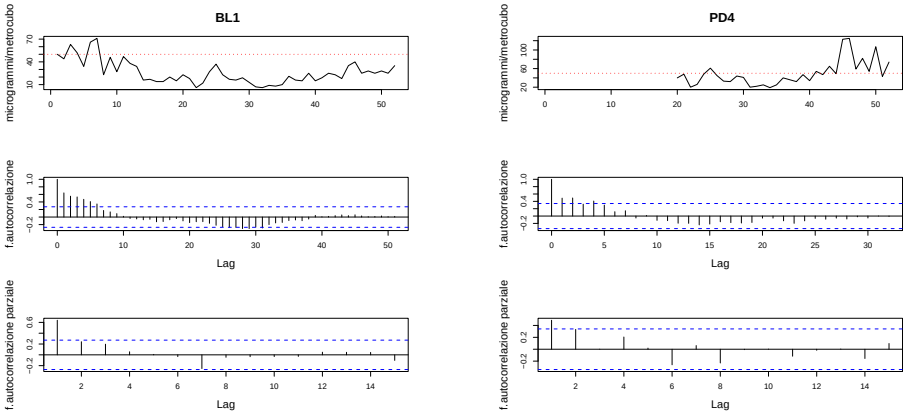


Figura 4.1: Serie storiche dei valori medi medi settimanali di PM_{10}, ACF e Pacf nelle stazioni BL1 e PD4

seppur approssimativamente, di considerare la nuova variabile - media settimanale del PM_{10} - simmetrica e con uguale variabilità nelle diverse stazioni di rilevazione, vengono confrontate due diverse trasformazioni dei dati, vedi la Fig. 4.2, tramite gli istogrammi della distribuzione empirica; appare evidente come la trasformazione dei dati medi settimanali, effettuata con entrambe le funzioni, manifesti una buona simmetria.

Nella Figura 4.3 vengono presentati gli andamenti della varianza rispetto alla media per ogni singola stazione dove, per i dati originari, si nota un incremento della prima all'aumentare della seconda. Grazie alla trasformazione dei dati, in entrambi i casi si evidenzia una stabilizzazione della varianza al modificarsi del valore medio con un andamento più soddisfacente nel caso della trasformazione effettuata tramite il logaritmo e di conseguenza l'analisi e la modellazione verrà eseguita su questa trasformazione dei dati, anche se molti dei lavori citati nella bibliografia, ad esempio Smith *et al.* [19], analizzano i dati secondo la trasformazione effettuata mediante radice quadrata. Analizzando in maggior dettaglio il grafico relativo alla trasformata effettuata mediante il logaritmo, si evidenziano quattro punti, nella parte verso sinistra, che hanno una disposizione difforme rispetto alla nuvola di tutti gli altri; questi punti corrispondono alle stazioni BL3, BL1, TV3 e BL2, rispettivamente da sinistra a destra e dal basso in alto. Per le tre stazioni dislocate nella provincia di Belluno si può notare come presentino situazioni orografiche e di territorio abbastanza diversificate rispetto a tutte le altre; la stazione in provincia di Treviso è situata in una tipologia di *background*

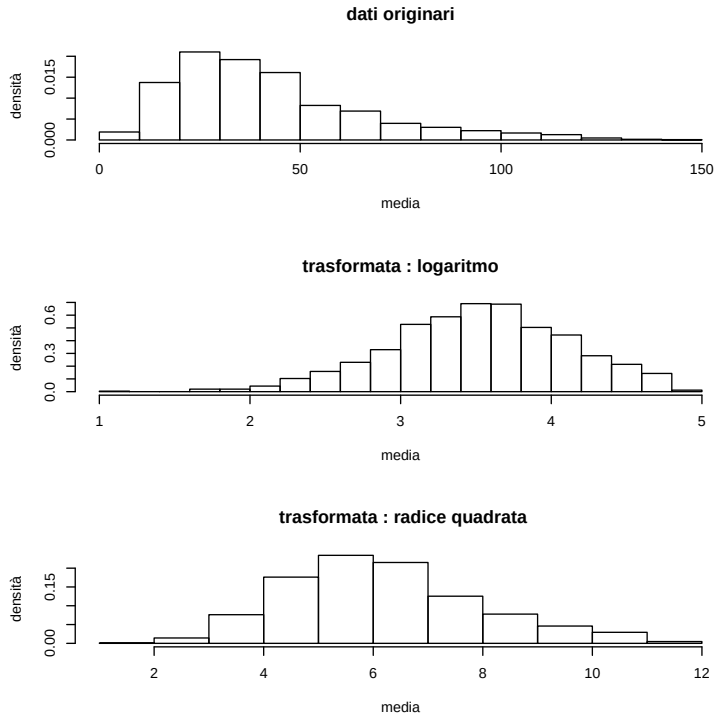


Figura 4.2: Distribuzione empirica delle medie settimanali del PM_{10}

rurale e presenta valori molto bassi nel periodo estivo.

4.3.2 Effetto delle altre variabili covariate

Come descritto nel Capitolo 2, le stazioni di rilevamento sono caratterizzate da una serie di informazioni, ausiliarie per l'analisi statistica, che possono essere utilizzate per la spiegazione del fenomeno. Si descrivono, brevemente, le analisi effettuate su tali variabili in relazione al livello di concentrazione del $\log(PM_{10})$.

Il tipo di *background*

Riprendendo quanto già visto nel Capitolo 2, l'analisi grafica sull'andamento del fenomeno in relazione al tipo di *background* manifesta un andamento simile in media e in variabilità per i primi tre livelli corrispondenti a stazioni site in zone non direttamente influenzate dal traffico veicolare, per cui si

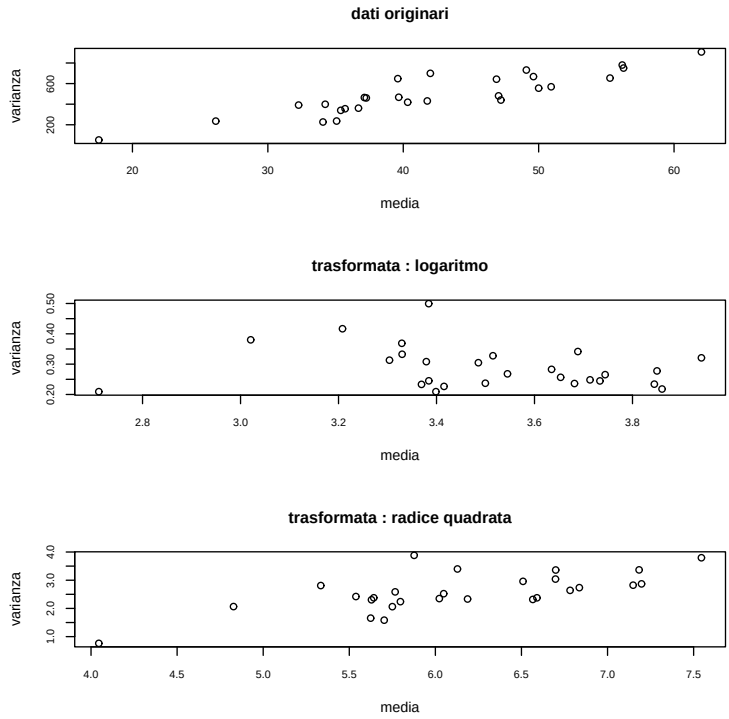


Figura 4.3: VARIANZA versus MEDIA per dati medi settimanali PM_{10} per ognuna delle 27 stazioni

procede ad una riclassificazione di questa variabile - vedi Fig. 4.4 - secondo le due tipologie: B - stazione di fondo (*background*) - e T - stazione in zona di traffico che rende tale variabile discriminante rispetto al fenomeno.

Lo strumento di misurazione

In maniera del tutto analoga a quanto visto sopra, l'analisi grafica della variabile che descrive la modalità di misurazione del livello di concentrazione di inquinante - classificata mediante M, Af e An - e presentata in Fig. 4.5 sembra evidenziare una capacità discriminatoria nella spiegazione del fenomeno per il tipo Af rispetto agli altri due.

Così facendo però, si introduce un problema di disomogeneità di misurazione del fenomeno che dovrebbe essere adeguatamente supportata da una cono-

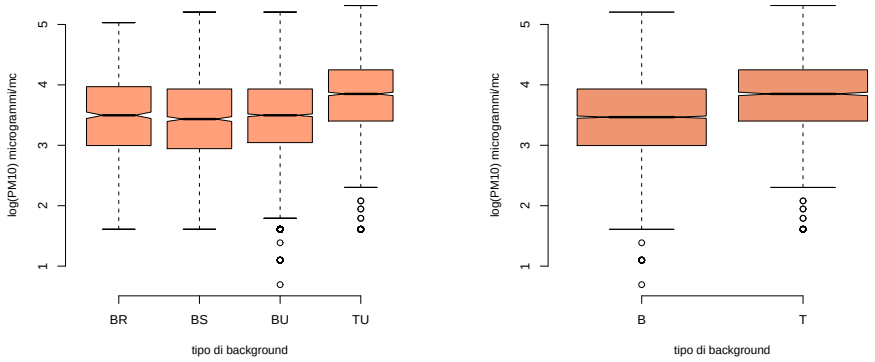


Figura 4.4: Andamento log(PM_{10}) condizionato a prima e dopo la riclassificazione della variabile Tipo di *background*

scenza e uno studio sia tecnico, sia statistico al fine di valutarne l'effetto; per questa ragione si ritiene, a differenza delle altre variabili descritte, di non utilizzare questa variabile per la modellazione successiva.

L'altitudine

L'analisi marginale per la variabile altitudine, condotta mediante modello lineare, presenta una stima del parametro significativa, correlandola negativamente; quindi all'aumentare della altitudine a cui sono situate le stazioni di rilevamento, il livello di concentrazione del log(PM_{10}) tende a diminuire come viene evidenziato nella Fig. 4.6 e dal coefficiente di correlazione pari a $R = -\sqrt{0.057}$

4.3.3 Il modello

In questo paragrafo vengono delineate le modalità con cui si procede alla formulazione e costruzione del modello spazio-temporale che interpreti l'andamento del fenomeno.

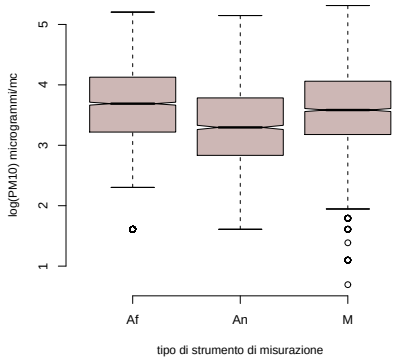


Figura 4.5: Andamento log(PM_{10}) condizionato alla classificazione della variabile Tipo di misurazione

4.3.4 Stima del trend temporale

L'andamento delle serie storiche delle medie settimanali - come esemplificato per due stazioni nella Fig. 4.1 - presenta, per ogni singola stazione, traiettorie simili nel corso dell'anno e non risultando componenti di trend o stagionali, si esclude la possibilità di una analisi secondo la scomposizione usuale delle serie storiche.

Al fine di catturare l'effetto temporale sui dati del logaritmo delle medie settimanali si ipotizza un effetto 'settimana' mediante (1) un modello lineare, con la settimana intesa come livello di una variabile di tipo fattore, e (2) un modello additivo che mediante l'uso di una funzione di tipo *spline*, caratterizzata da un certo grado di lisciamento, interpreti l'evoluzione del livello di inquinante nel corso dell'anno. Nella figura 4.7 vengono visualizzati gli andamenti del trend temporale ottenuti, in cui si può notare la diversa interpretazione del comportamento della stima della componente temporale ottenuta mediante *spline* (curva lisciata); vengono presentati anche i grafici condizionatamente al tipo di *background*, dove si evidenzia come i valori relativi a stazioni situate in zona di traffico (grafico in basso a destra) presentano valori più elevati lungo tutto l'arco temporale rispetto a quelle situate in *background* di fondo.

Questa osservazione conferma quanto visto precedentemente, ossia che il tipo di *background* in cui è sita la stazione è importante per la determinazione

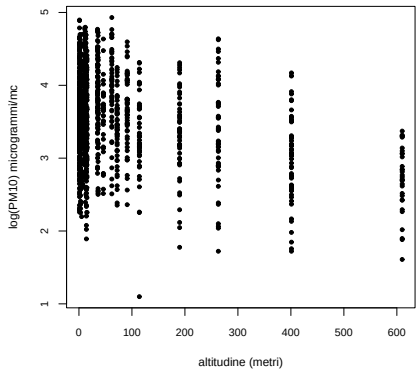


Figura 4.6: Andamento $\log(PM_{10})$ condizionato alla variabile altitudine

del livello di inquinante, e tale risultato è in linea con quanto ottenuto in letteratura.

Sahu, Gelfand e Holland [15] evidenziano che il livello di $PM_{2.5}$ è determinato additivamente da due processi spazio-temporali; il primo che cattura l'effetto di fondo (o rurale) e il secondo dovuto ad un ulteriore effetto presente nelle aree urbane.

4.3.5 Stima del trend spaziale

Per la stima del trend spaziale si procede tramite il metodo non parametrico che consente, attraverso la versione bidimensionale delle *spline*, una rappresentazione 'lisciata' del fenomeno su tutta la regione oggetto di studio, e, di conseguenza, interpretando lo stesso come caratterizzato da una certa gradualità e continuità di intensità nello spazio.

Una ulteriore parte del modello, come descritto precedentemente, sempre attinente ad attributi insiti alla dimensione spaziale del fenomeno, considera sia la variabile relativa all'altitudine sia quella relativa alla tipologia di *background* in cui è situata la stazione.

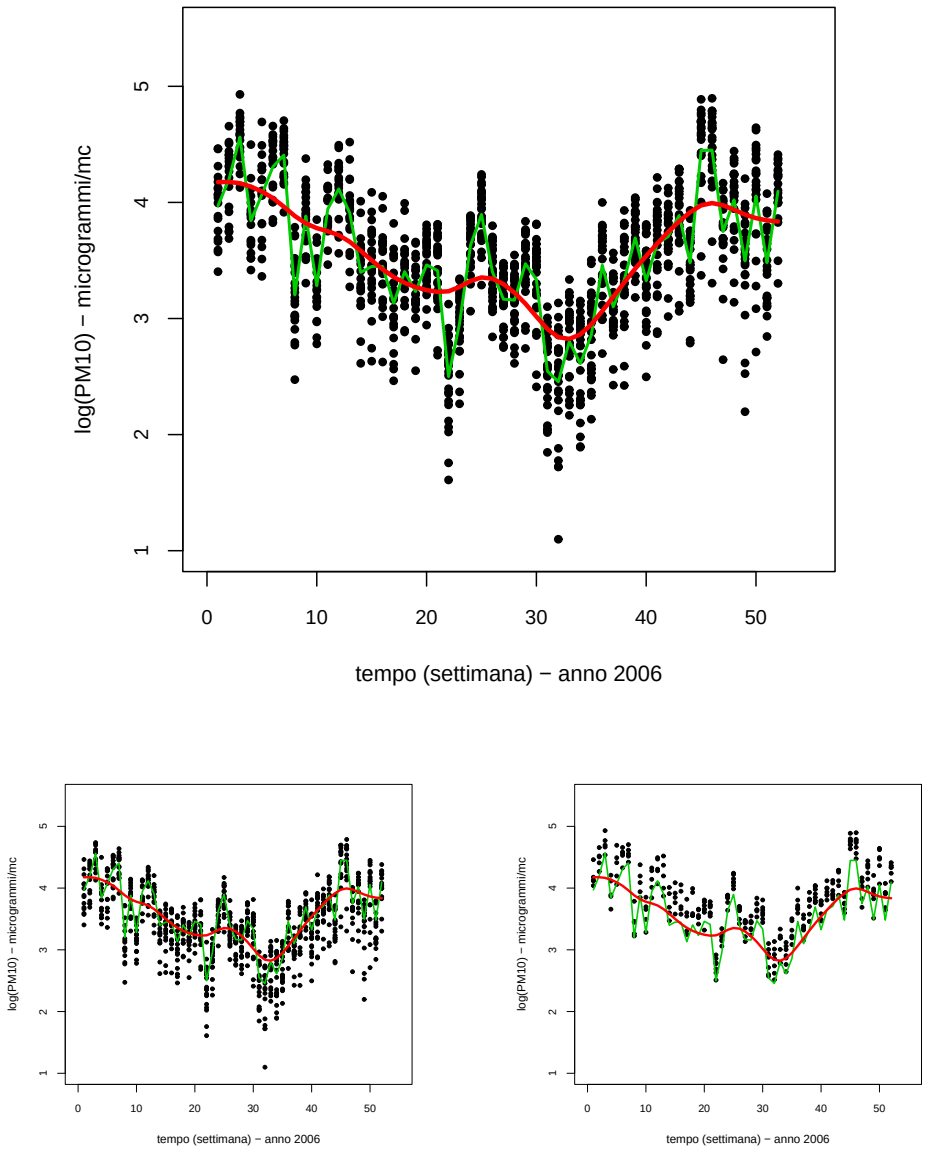


Figura 4.7: Comparazione della stima del trend temporale sulle medie settimanali log(PM_{10}). La curva più irregolare rappresenta l'effetto 'settimana', mentre quella lisciata è costruita mediante *smoothing spline* con 10 g.l.; (in alto) tutte le osservazioni, (in basso-sx) stazioni site nel tipo di *background* di fondo, (in basso-dx) stazioni site nel tipo di *background* di traffico

4.3.6 Il modello di regressione

Indicato con $Y(\boldsymbol{s},t) = \log Z(\mathbf{s},t)$, il modello descrivente il fenomeno oggetto di studio, si può esprimere attraverso

$$Y(\boldsymbol{s},t) = \beta_0 + \omega(t) + \psi(\boldsymbol{s}) + \beta_1 BG(\boldsymbol{s}) + \beta_2 A(\boldsymbol{s}) + \epsilon(\boldsymbol{s},t) \qquad (4.3)$$

dove: $\omega(t)$ denota la parte del modello relativa al trend temporale, $\psi(\boldsymbol{s})$ la parte relativa al trend nello spazio, $BG(\boldsymbol{s})$ e $A(\boldsymbol{s})$ sono le variabili relative al tipo di *background* e all'altitudine in cui sono situate le stazioni di monitoraggio della rete di rilevamento dell'ARPAV.

Per le componenti $\omega(t)$ e $\psi(\boldsymbol{s})$ deve essere determinato il grado di lisciamento in modo da bilanciare la bontà di adattamento del modello, che ne permette una migliore rappresentazione con i dati disponibili, e una parsimonia nei parametri che lo descrivono per mettere a fuoco gli aspetti strutturali del fenomeno.

In letteratura sono stati sviluppati vari criteri che consentono di valutare la perdita di efficacia predittiva, che tende ad essere tanto maggiore quanto meno parsimonioso è il modello utilizzato. I due criteri maggiormente utilizzati basati sul concetto di log-verosimiglianza penalizzata, anche per la loro semplicità di determinazione, sono quelli $AIC = -2logLik + 2p$ e $BIC = -2logLik + p\log(n)$, dove con $logLik$ si indica la log-verosimiglianza, con p il numero di parametri (regressori) presenti nel modello e con n il numero di osservazioni.

Nel nostro caso, la scelta del numero di nodi k da effettuare per le due componenti, temporale e spaziale, stimate tramite *spline* e di conseguenza del grado di lisciamento delle curve interpolanti, è stata effettuata ricorrendo sia all'andamento del valore del criterio AIC sia tramite il metodo di convalida incrociata sulla devianza degli errori di previsione $D^{(*)} = \sum(y_{i(-k)} - \hat{y}_{i(-k)})^2$, calcolata su un sottoinsieme di osservazioni (insieme dati di *test*, di numerosità $n - k$) che non vengono usate nella fase di stima dei parametri del modello (insieme dati di *training*, di numerosità k).

Per la componente temporale la scelta del nodi effettuata risulta pari a 7,

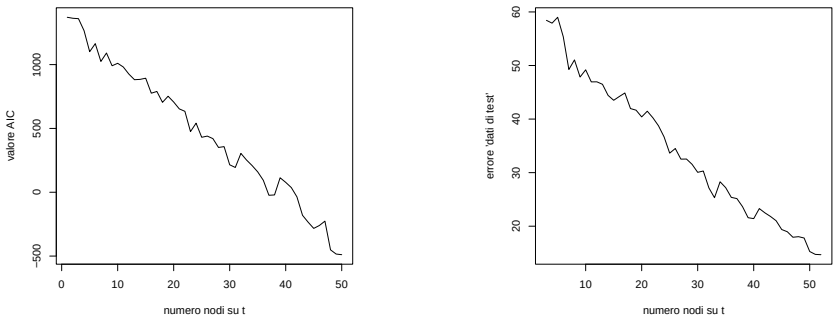


Figura 4.8: Andamento criterio AIC (a sx) e $D^{(*)}$ (a dx) all'aumentare del numero di nodi per la stima della componente temporale

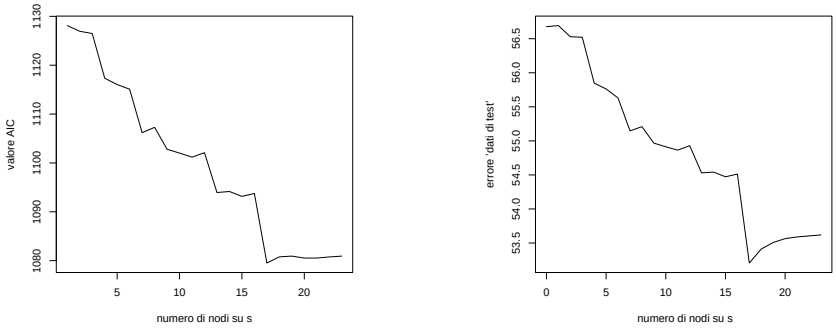


Figura 4.9: Andamento criterio AIC (a sx) e $D^{(*)}$ (a dx) all'aumentare del numero di nodi per la stima della componente spaziale

mentre per la componente spaziale la scelta risulta pari a 17 (vedi Fig. 4.9). In realtà, come presentato in Fig. 4.8, per la *spline* relativa alla componente temporale, il criterio AIC e la devianza $D^{(*)}$ presentano un andamento decrescente all'aumentare del numero di nodi; il valore prescelto si configura essere di minimo relativo. Si ritiene comunque di effettuare la scelta del numero di nodi pari a 7 al fine sia di evitare un sovra-adattamento del modello ai dati specifici, sia di catturare la parte strutturale del fenomeno lungo la dimensione temporale, viste la complessità, particolarità e variabilità dovute alla forte dipendenza del livello di concentrazione di inquinante al manifestarsi delle diverse condizioni metereologiche.

Nella fase di controllo grafico sui residui $\hat{\epsilon}(\mathbf{s}, t)$ - Appendice B, Fig. B.2

- ottenuti dal modello proposto, sono presenti alcuni valori molti difformi dagli altri che sono stati ritenuti *outliers* e si è proceduto ad eliminare. La verifica grafica dell'ipotesi di normalità dei residui - sempre presentata nell'Appendice B, in Fig. B.3 - consente di considerare soddisfacente il modello implementato.

Nella Tabella 4.2 vengono presentati i valori dei coefficienti determinati tramite l'utilizzo della funzione `gam` presente nella libreria `mgcv` di `R`. Come

Coefficienti	Valore stimato	Std. error	t value
(Intercept)	3.6141828	0.02770	130.485
Background tipo:T	0.1850854	0.02906	6.370
Altitudine	-0.0009014	0.00032	-2.781
Fattori di lisciamento	**gradi di libertà stimati**		**F value**
spline(Settimana)	5.986		229.412
spline(Longit., Latit.)	13.924		7.867
Coeff.di correlazione $R^2 = 0.607$			

Tabella 4.2: Stime dei parametri ottenute con il modello (4.3)

si evince, tutti i coefficienti, sia quelli lineari sia quelli calcolati tramite tecniche non parametriche, relativi al modello di regressione risultano altamente significativi con un valore di variabilità spiegata che supera il 60%.

La Fig. 4.10 consente di visualizzare l'andamento stimato delle varie componenti previste dal modello.

4.3.7 Analisi dei residui e modellazione geostatistica

Il passo successivo consiste nella analisi dei residui stimati dal modello $\hat{\epsilon}(\boldsymbol{s}, t)$, al fine di valutare se la componente deterministica stimata dal modello interpreti adeguatamente le dimensioni del processo e quanta parte di correlazione nel tempo e nello spazio sia ancora presente nei residui.

L'autocorrelazione temporale

Per la valutazione della correlazione temporale, i residui $\hat{\epsilon}(\boldsymbol{s}, t)$ vengono considerati come singole serie storiche per $T = 1, \ldots, 52$ in ognuna delle 27 stazioni.

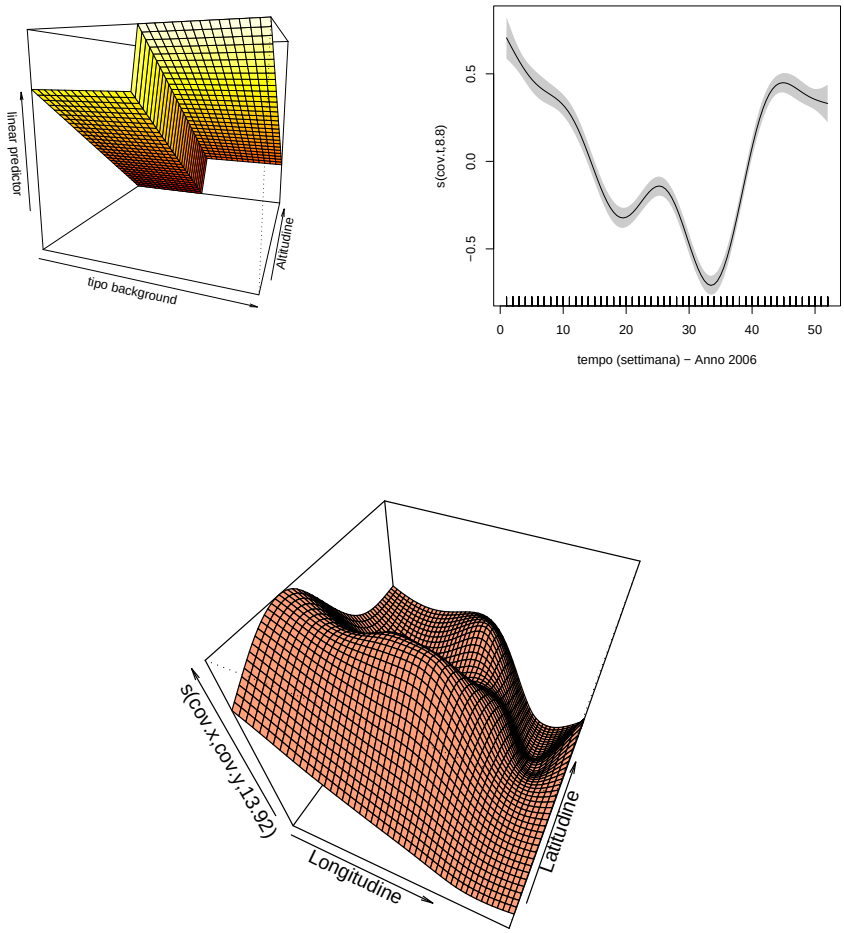


Figura 4.10: Andamento delle componenti del modello (4.3): (in alto-sx) tipo *background* e altitudine; (in alto-dx) temporale - settimana; (in basso) spaziale - coordinate geografiche

Secondo la metodologia standard delle serie storiche, sono stati calcolati i coefficienti di autocorrelazione per ognuna di esse per distanze fino a lag 8, corrispondenti ad una ampiezza temporale di circa due mesi. Per calcolare la significatività dei coefficienti così determinati, si sono messi a confronto con il valore di $\pm 2/\sqrt{T}$, con T pari alla lunghezza della serie temporale (52

nel nostro caso), che identificano, seppur approssimativamente, i limiti per la regione di accettazione in cui sono presenti il 95% dei valori.

In realtà, all'approssimazione sul livello di significatività, si aggiunge anche una possibile distorsione data dal fatto che le serie temporali considerate non hanno la medesima lunghezza poiché non tutte iniziano alla settimana 1 e che alcune presentano dei valori mancanti; si ritiene comunque, che l'indicazione data dalla procedura possa essere sufficientemente attendibile per l'analisi e la modellazione dei dati.

In Figura 4.11 sono rappresentati la distribuzione delle stime dei coefficienti di autocorrelazione e il profilo degli stessi per ogni singola stazione, per distanze fino a lag 8. Come si nota, i valori che fuoriescono dalle bande di variabilità sono molto pochi per cui i residui possono essere considerati temporalmente incorrelati; si ritiene verificata l'ipotesi che la dipendenza temporale sia sufficientemente spiegata dal trend e che i residui risultino ora correlati esclusivamente nella dimensione spaziale.

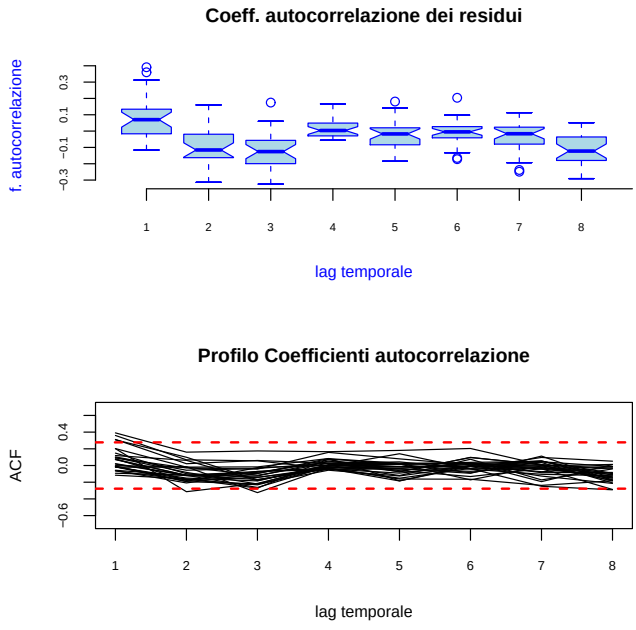


Figura 4.11: Grafici delle distribuzioni dei coefficienti di correlazione (in alto) e dei profili delle autocorrelazioni temporali sui residui $\hat{\epsilon}(\boldsymbol{s})$ delle 27 stazioni con bande di variabilità approssimate al 95% (in basso)

La correlazione spaziale

Quanto finora visto può essere sintetizzato il termini di modello formale come

$$Y(t, \mathrm{s}) = \mu(t, \mathrm{s}) + \epsilon(\mathrm{s}, t) \tag{4.4}$$

dove la componente μ viene stimata tramite il modello 4.3, e rappresenta la parte deterministica in funzione del tempo e della dimensione spaziale di larga scala. Quanto rimane, ossia i residui stimati $\hat{\epsilon}(\boldsymbol{s}, t) = Y(t, \mathrm{s}) - \hat{\mu}(t, \mathrm{s})$, dalle ipotesi formulate, rappresentano realizzazioni di una variabile casuale i.i.d. in ognuna delle 27 stazioni oggetto di rilevamento del fenomeno.

Per la determinazione della correlazione spaziale dei residui $\hat{\epsilon}(\boldsymbol{s})$, ci si avvale delle tecniche della geostatistica, come già visto nel Capitolo 3. Si procede quindi alla determinazione del variogramma empirico ed alla scelta della struttura di correlazione appropriata, determinata mediante la ricerca di un opportuno variogramma teorico.

Il variogramma empirico

Detratta la componente di trend ad ogni stazione $\mathbf{s}$ corrispondono 52 determinazioni (o meno, poiché nella serie si hanno dei valori mancanti) relative ai residui $\hat{\epsilon}_t(\boldsymbol{s})$ con $t = 1, \ldots, T \leq 52$; visto il modello e l'incorrelazione temporale dei residui stessi, al fine di procedere mediante l'analisi geostatistica classica, si provvede alla loro aggregazione, mediante l'utilizzo di un indice sintetico, per ogni stazione.

Per l'aggregazione oltre alla media aritmetica, viene usata la mediana che, visto l'andamento dei valori di partenza in cui sono presenti valori molto elevati, garantisce maggior robustezza.

Proprio per questo motivo si ritiene che l'aggregazione effettuata con la mediana abbia un comportamento meno influenzato dai valori estremi e viene, di conseguenza, preferito per l'elaborazione successiva - si vedano a tal proposito i grafici presentati nell'Appendice B in Fig.6.

Entrando nel merito di quanto visualizzato nel variogramma nuvola, risulta evidente come siano presenti valori elevati della semivarianza per distanze molto piccole; ciò può essere indice della presenza di *outliers*.

Uno specifico controllo rileva che questa anomalia dipende dai residui

corrispondenti alla stazione VE2. Questa stazione è posta nel centro storico di Venezia, e quindi sita in condizioni geografiche ed ambientali estremamente differenziate rispetto alle due stazioni più vicine, una a Mestre nell'area del parco Bissuola e l'altra sempre a Mestre in Via Circonvallazione. Vista questa palese difformità di situazioni, l'analisi della struttura di correlazione spaziale, presentata successivamente, viene eseguita omettendo i valori espressi nella stazione sopracitata.

Il semivariogramma empirico così determinato, presentato in Fig. 4.12, sembra manifestare un andamento crescente fino a valori della distanza approssimativamente pari a 25-30 km, e una stabilizzazione per distanze maggiori. Data la bassa numerosità di punti con cui vengono stimati i valori per le distanze più elevate, si ritiene di considerare la stima del semivariogramma limitando la misura della distanza a 125 km.

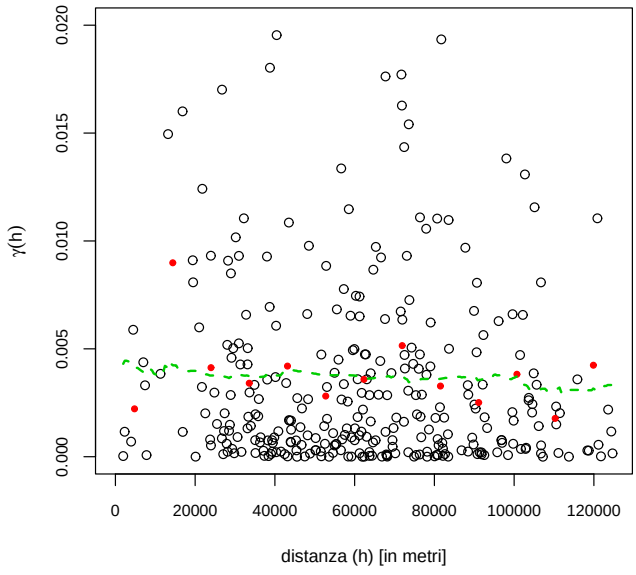


Figura 4.12: (semi)variogramma nuvola, empirico e lisciato dei residui $\hat{\epsilon}(\mathbf{s})$

Per controllare la presenza di anisotropia - ossia di comportamento diverso rispetto alla direzione - dei residui, si è proceduto con la stima del variogramma direzionale - vedi Fig. 4.13 - dal quale non si evidenziano differenze marcate nel comportamento rispetto alle varie direzioni; appare una leg-

gera differenza, per alcune distanze, sulla variabilità in direzione ovest-est (relativa all'angolo $\alpha = 90°$) rispetto alle altre direzioni.

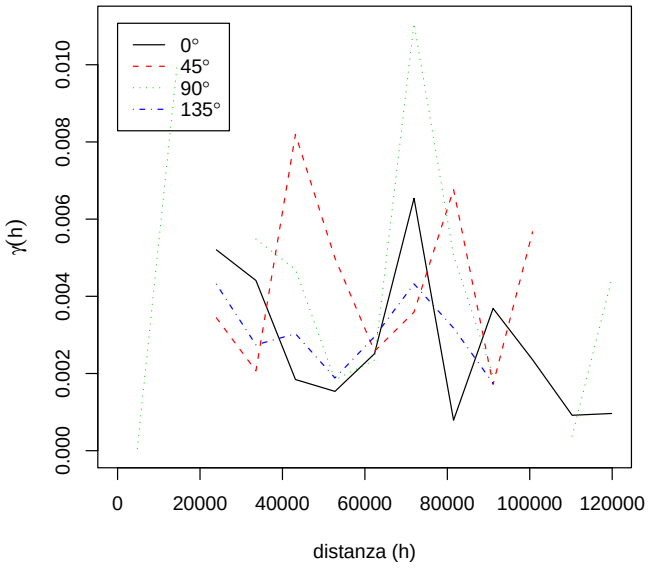


Figura 4.13: (semi)variogramma direzionale per controllo anisotropia

Con tali supposizioni, l'ipotesi di stazionarietà, che sottende al tipo di analisi applicata ai residui del modello, appare soddisfatta, per cui si procede con la determinazione della componente di variabilità espressa nella dimensione spaziale.

La struttura di correlazione spaziale

Le stime del variogramma ottenute sono presentate nelle Tabb. 4.3 e 4.4, mentre nella Fig. 4.14 sono presentati gli andamenti dei variogrammi teorici rispetto ai punti che identificano quello empirico ottenuto direttamente dai residui.
I variogrammi calcolati mediante minimi quadrati pesati (wls), rispetto a quelli calcolati mediante il metodo della massima verosimiglianza (ml), sono caratterizzati da una 'caduta' della correlazione più lenta, il che implica un *range* più elevato, e anche per una variabilità comprendente un effetto pepita

Covarianza	nugget	sill	range	SQM
esponenziale	0,0022	0,0021	15000	26,8438
gaussiano	0,0022	0,0020	15000	27,3153

Tabella 4.3: Parametri dei variogrammi teorici - stime wls

Covarianza	nugget	sill	range	mean	log-Lik	AIC
esponenziale	0,0	0,0035	3691,98	0,0217	-66,33	37,16
gaussiano	0,0	0,0035	4378,76	0,0206	-67,55	37,78
wave	0,02	0,0017	5678,58	0,0199	-65,35	36,68

Tabella 4.4: Parametri dei variogrammi teorici - stime ml

che non risulta essere presente con le stime calcolate con il metodo ml. Per

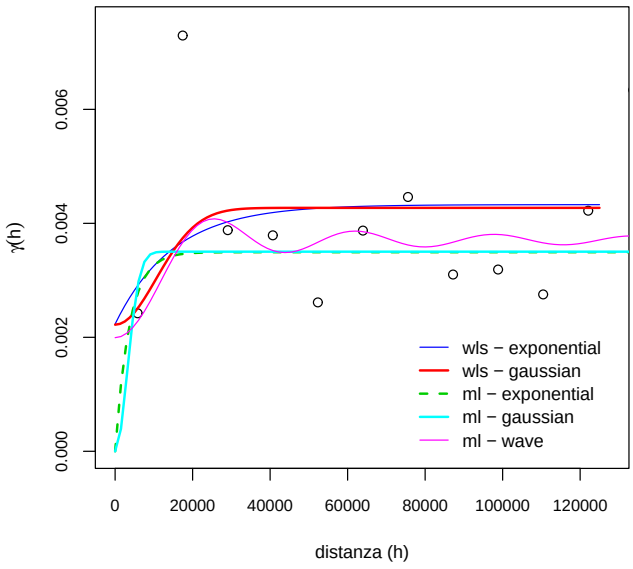


Figura 4.14: (semi)Variogrammi teorici

la scelta tra le varie strutture di correlazione ci si avvale, come nel Capitolo 3, della procedura di convalida incrociata, i cui valori sono presentati nella Tab. 4.5; da ciò si desume che quella di tipo gaussiano con parametri stimati mediante minimi quadrati pesati, pur con una leggera sovrastima, presenta la determinazione più prossima a 1 e di conseguenza preferibile. La validità della scelta, ossia che si debba propendere per i valori più elevati di *sill* e *range*, viene indirettamente confermata anche dall'andamento e dal valore

Modello di covarianza spaziale	Valore CV
esponenziale - wls	0,9894
gaussiano - wls	1,0089
esponenziale - ml	0,9838
gaussiano - ml	0,9838
wave - ml	1,0315

Tabella 4.5: Valori di convalida incrociata

della convalida incrociata ottenuti per il modello *wave* stimato attraverso il metodo della massima verosimiglianza.

4.3.8 La previsione spaziale e temporale

Per la previsione spaziale si ricorre *kriging* ordinario sui residui, su una griglia regolare di punti equispaziati ricoprente la regione Veneto.

Sempre negli stessi punti della griglia viene effettuata la stima della previsione relativa alla parte deterministica di trend, calcolata mediante il modello additivo visto in 4.3. Questa operazione implica l'attribuzione in ognuno dei punti dei valori relativi alle varie componenti del modello, oltre alle coordinate geografiche.

Per la variabile altitudine, partendo dalla conoscenza dei 27 valori noti relativi alle stazioni di monitoraggio, ci si avvale di un modello additivo generalizzato, sempre calcolato tramite *spline*, per predire il valore in tutti i punti della griglia. Nell'Appendice B, viene presentato il modello stimato e l'andamento dell'altitudine (Fig. B.7) nella stessa regione spaziale oggetto di previsione. Pur consapevoli che questa risulta una approssimazione con molti limiti riguardanti la reale orografia della regione Veneto, si ritiene che il risultato ottenuto per questa variabile possa essere adeguato in relazione agli scopi definiti per la nostra analisi.

L'utilizzazione della variabile che interpreta il tipo di *background* avviene assegnando i due livelli definiti sopra - 'B' per tipologia di fondo senza traffico e 'T' relativo alla presenza di traffico - a due previsioni distinte; alla variabile relativa alla dimensione temporale vengono assegnati i valori $t = 1, \ldots, 52$ indicanti la settimana all'interno dell'anno.

Tramite il modello 4.3 si procede alla individuazione dei valori previsti in ogni punto della griglia (corrispondenti a date coordinate di Longitudine

e di Latitudine).

La somma del trend spazio-temporale deterministico e della previsione sui residui, ottenuta mediante il *kriging* ordinario, consente di ottenere due andamenti, a seconda del tipo di *background*, per ognuna delle 52 settimane dell'anno 2006. Ricordando che l'analisi è stata compiuta sui dati trasformati mediante la funzione logaritmo, per ottenere il dato su scala originaria, viene applicata la trasformazione esponenziale.

Il risultato ottenuto, visualizzato per alcuni valori della variabile settimana e per le due diverse tipologie di *background* nella Fig. 4.15, consente di osservare come il livello di concentrazione del PM_{10} si distribuisca nello spazio e come evolva nel tempo conformente al tipo di *background*.

La parte della pianura maggiormente compromessa, ossia che presenta un omogeneo livellamento verso i valori di inquinante elevati, appare essere quella situata lungo l'asse ovest-est comprendente le aree urbane corrispondenti alle città di Padova e Verona, mentre le zone verso nord (provincia di Vicenza e Treviso) e verso sud (provincia di Rovigo) presentano valori relativamente più bassi.

4.3.9 La convalida del modello

Per un controllo sulla bontà della capacità previsiva del modello si opera tramite:

- simulazione spaziale degli errori sui 27 siti appartenenti alla rete di rilevazione, tramite la funzione grf() - ovvero da un processo spaziale aleatorio gaussiano - utilizzando la struttura di covarianza spaziale stimata per il variogramma teorico, per un totale di 40 realizzazioni;

- previsione della componente deterministica, sempre su 27 siti, tramite il modello additivo, utilizzando i valori delle variabili covariate - coordinate, altitudine, tipo di *background*, settimana - relative ad ogni stazione;

- determinazione dell'inviluppo superiore e inferiore dei due effetti - casuale e sistematico - espressi, mediante la trasformazione esponenziale, in scala originaria e confronto con i dati delle medie settimanali risultanti dai dati osservati.

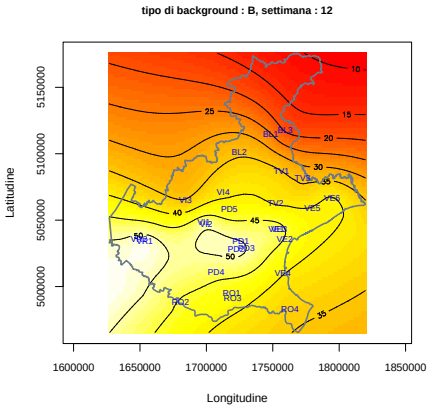
La scelta del numero di realizzazioni simulate pari a 40 e degli inviluppi consente la determinazione di una banda di previsione al cui interno dovrebbero trovarsi il 95% dei valori.

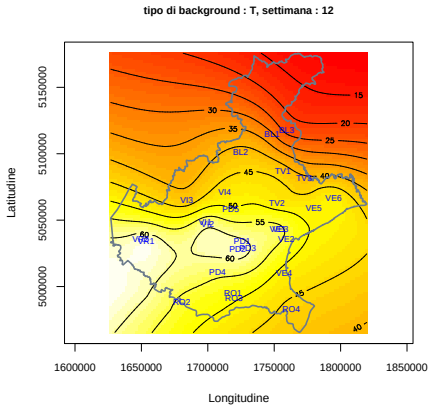
Dalla Fig. 4.16, che visualizza questi andamenti per 4 siti sul totale dei 27, si evincono le limitazioni della modellazione prevista in 4.4, in quanto numerosi sono i punti, rappresentanti i valori osservati, che ricadono al di fuori delle bande di variabilità.

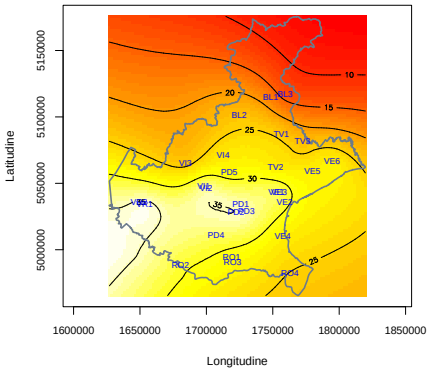
Il problema dell'ampiezza delle bande è conseguenza dell'aggregazione operata, attraverso la mediana, per i residui in ogni stazione che non consente di considerare l'incertezza insita nella replicazione delle osservazioni. In altre parole, la determinazione della componente temporale considerata solo deterministicamente limita le caratteristiche complesse del processo.

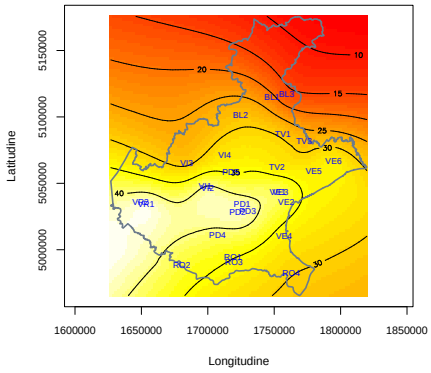
Una conferma proviene dal fatto che se si effettua la simulazione sui residui sulla stessa struttura di covarianza spaziale, ma con stime dei parametri relativi al *range* e soprattutto al *sill*, ossia per la parte di variabilità del processo spaziale, su tutti i residui e non sul singolo valore aggregato, le bande di variabilità sono molto più ampie e ricomprendono una parte maggiore di valori. In Fig. 4.17 vengono presentate le stesse stazioni con la diversa determinazione delle bande di variabilità che risultano troppo ampie e regolari per poter dare una interpretazione sulla bontà del modello.

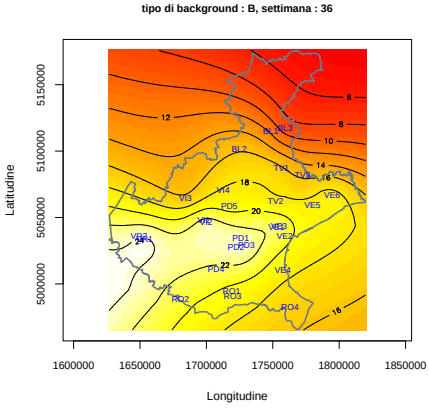
Come parziale conclusione, sembra di poter affermare che il modello descritto in questo capitolo possa esprimere una indicazione qualitativa di massima sull'andamento della previsione nello spazio e nel tempo e che questo modello limita troppo e sembra non interpretare appropriatamente la parte di variabilità spaziale e quella temporale del processo.

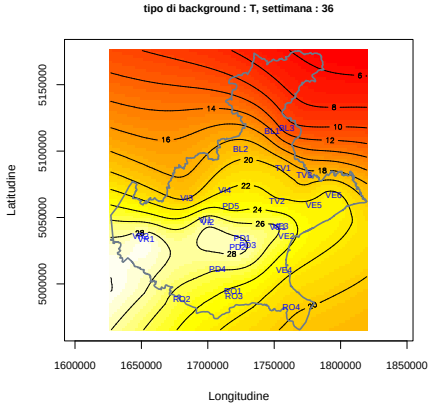

tipo di background : B, settimana : 12
Latitudine
Longitudine
1600000
1650000
1700000
1750000
1800000
1850000
5000000
5050000
5100000
5150000


tipo di background : T, settimana : 12
Latitudine
Longitudine
1600000
1650000
1700000
1750000
1800000
1850000
5000000
5050000
5100000
5150000


tipo di background : B, settimana : 24
Latitudine
Longitudine
1600000
1650000
1700000
1750000
1800000
1850000
5000000
5050000
5100000
5150000


tipo di background : T, settimana : 24
Latitudine
Longitudine
1600000
1650000
1700000
1750000
1800000
1850000
5000000
5050000
5100000
5150000


tipo di background : B, settimana : 36
Latitudine
Longitudine
1600000
1650000
1700000
1750000
1800000
1850000
5000000
5050000
5100000
5150000


tipo di background : T, settimana : 36
Latitudine
Longitudine
1600000
1650000
1700000
1750000
1800000
1850000
5000000
5050000
5100000
5150000

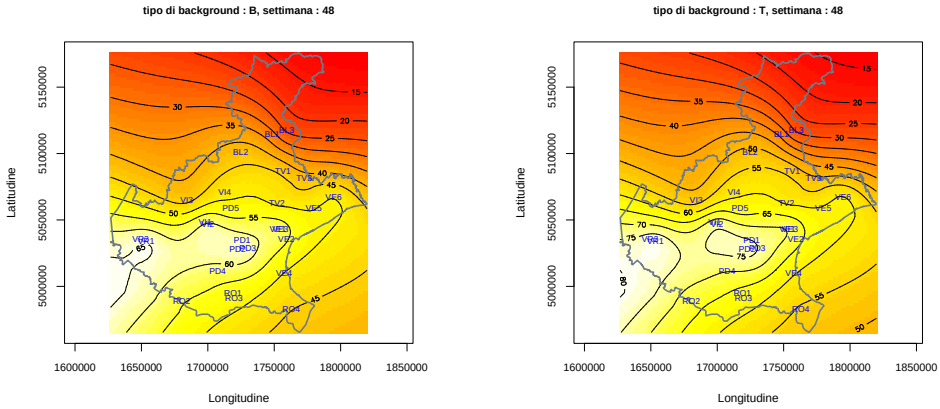


Figura 4.15: Previsione del livello di concentrazione di PM_{10}, in 4 settimane diverse e per diverso tipo di *background* (B/T)

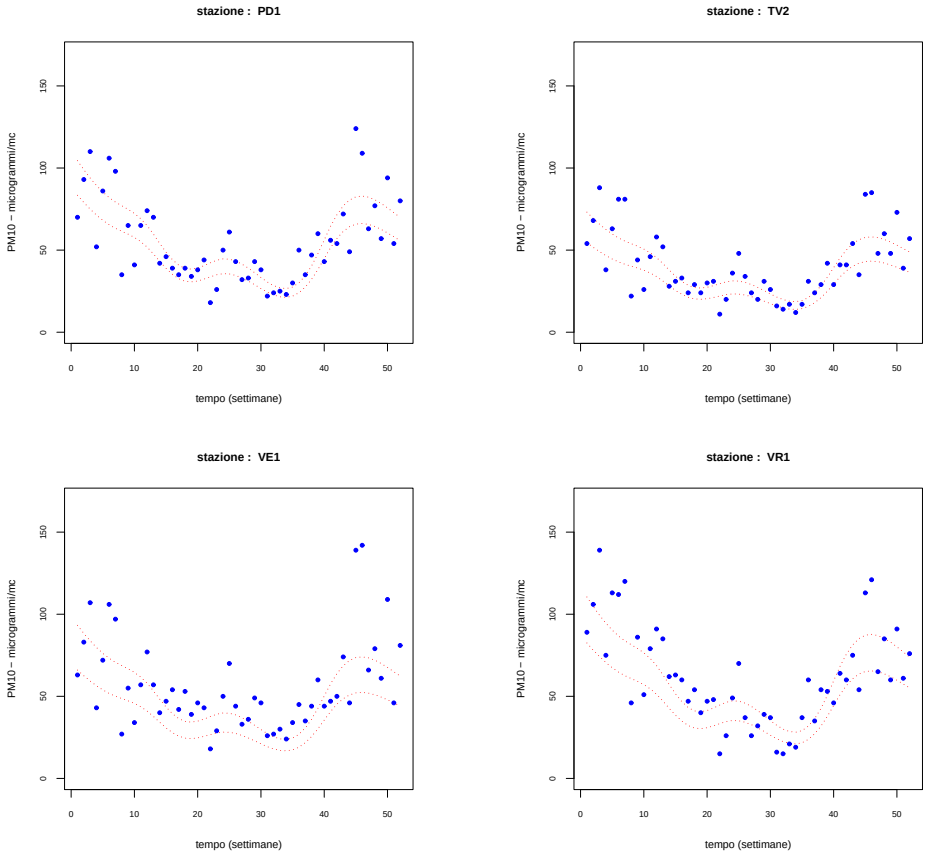


Figura 4.16: Bande di previsione (inviluppi su 40 realizzazioni) e valori osservati per 4 stazioni di rilevamento

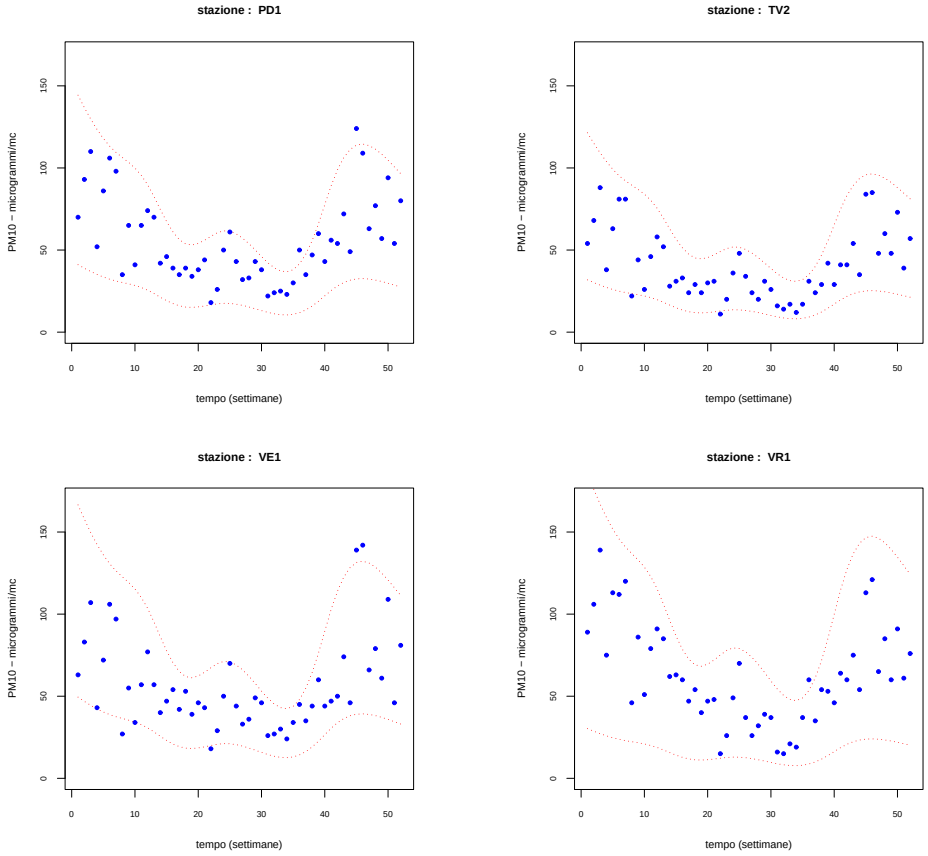


Figura 4.17: Bande di previsione (inviluppi su 40 realizzazioni) determinate con valori del variogramma teorico calcolati sull'insieme dei residui e valori osservati per 4 stazioni di rilevamento

Capitolo 5

Modellazione gerarchica per processi spazio-temporali

5.1 Il modello di riferimento per i processi spazio-temporali

Per una presentazione formale, un singolo dato viene indicato con $Z(\mathbf{s}, t)$, dove $\mathbf{s} \in D \subset \mathbb{R}^2$ e con $t \in \mathbb{R}$ il tempo. In alcuni casi, assieme ai dati specificatamente attinenti al fenomeno, è disponibile un insieme di informazioni supplementari, dette variabili *covariate*, rappresentate con un vettore $\mathbf{x}(\mathbf{s}, t)$. In pratica, la variabile casuale $Z(\mathbf{s}, t)$ viene osservata in alcuni siti $\mathbf{s}_i$, $i = 1, \ldots, n$ siti e in istanti temporali t_j con $j = 1, \ldots, m$.
Per semplicità si suppone che la v.c. è osservata ad istanti regolari per cui $t = 1, \ldots, T$.

Una possibile specificazione del modello avviene secondo una gerarchia a più livelli.
Al primo livello si ha

$$Z(\mathbf{s}, t) = Y(\mathbf{s}, t) + \epsilon(\mathbf{s}, t) \tag{5.1}$$

dove $Y(\mathbf{s}, t)$ è un processo spazio-temporale e $\epsilon(\mathbf{s}, t)$ è un termine di errore di tipo *white noise* indipendente, che segue una distribuzione Normale $N(0, \sigma_\epsilon^2)$.

Al secondo livello il processo spazio-temporale può essere visto come

$$Y(\mathbf{s},t) = \mu(\mathbf{s},t) + w(\mathbf{s},t) \tag{5.2}$$

dove $\mu(\mathbf{s},t)$ è una funzione deterministica che può dipendere da un insieme di variabili covariate $\mathbf{x}(\mathbf{s},t)$ e $w(\mathbf{s},t)$ è un processo spazio-temporale a media nulla.

Per il processo $w(\mathbf{s},t)$ si assume una ben definita struttura di covarianza

$$C(\mathbf{s}_1,\mathbf{s}_2;t_1,t_2) = Cov\big[w(\mathbf{s}_1,t_1), w(\mathbf{s}_2,t_2)\big] \tag{5.3}$$

Il processo spazio-temporale $w(\mathbf{s},t)$, si dice *stazionario in senso debole* se la covarianza risulta esclusivamente funzione della differenza delle coordinate spaziali e temporali

$$C(\mathbf{s}_1,\mathbf{s}_2;t_1,t_2) = C(\mathbf{s}_1-\mathbf{s}_2;t_1-t_2) = C(d;\tau) \tag{5.4}$$

dove $d = \mathbf{s}_1 - \mathbf{s}_2$ e $\tau = t_1 - t_2$. Il processo si dice essere *isotropico* se

$$C(d;\tau) = C(||d||;|\tau|) \tag{5.5}$$

dove $||\cdot||$ è la norma euclidea in $\mathbb{R}^2$.

In letteratura i processi isotropici sono molto usati grazie alla loro semplice espressione funzionale e alla loro interpretabilità. Per questo tipo di modelli la struttura di covarianza è espressa attraverso diverse forme parametriche.

Una ulteriore assunzione, che rende più semplice la forma funzionale della covarianza, è la separabilità. Il processo stazionario $w(\mathbf{s},t)$ si dice *separabile* se risulta essere il prodotto delle due covarianze distinte per le diverse dimensioni, ossia

$$C(||d||;|\tau|) = C_s(||d||)\, C_t(|\tau|) \tag{5.6}$$

dove $C_s(\cdot)$ e $C_t(\cdot)$ sono opportune funzioni di covarianza.

Oltre a quanto visto sopra, sono stati sviluppati anche metodi per la costruzione

di funzioni di covarianza spazio-temporale non separabili e non stazionarie, come ad esempio in Gneiting (2002),

$$C(||d||;|\tau|) = \frac{1}{1+|\tau|}\exp\Big\{-\frac{||d||}{(1+|\tau|)^{\beta/2}}\Big\} \tag{5.7}$$

dove β identifica il parametro di interazione tra le dimensioni spaziali e temporale.

5.2 Alcuni recenti sviluppi

5.2.1 Modelli di covarianza per processi non stazionari

I processi spaziali con struttura di covarianza stazionaria sono stati usati nelle applicazioni geostatistiche per diversi decenni, con risultati spesso incoraggianti. Anche per le applicazioni ambientali, riguardanti i fenomeni di inquinamento dell'aria, le prime analisi sono state condotte mediante gli strumenti classici proposti dalla geostatistica, ma in questo ambito più che in altri, ci si è trovati nella situazione in cui l'assunzione della stazionarietà della covarianza nell'intero dominio del campo aleatorio risulta non essere soddisfatta. Diversi studiosi negli ultimi anni si sono occupati di sviluppare approcci che consentissero di analizzare fenomeni senza considerare l'assunto della stazionarietà del processo in ambito spaziale.

La deformazione dello spazio

Sampson e Guttorp [17] hanno proposto un approccio non parametrico per stimare la struttura di covarianza spaziale per l'intero dominio del campo aleatorio senza assumere l'ipotesi di stazionarietà. Il metodo consiste nella costruzione di una funzione sufficientemente liscia che mappa le localizzazioni dello spazio geografico, *G-space*, in cui non viene assunta la stazionarietà, in un nuovo - virtuale - spazio, *dispersion space* o *D-space*, dove invece viene assunta la stazionarietà della struttura di covarianza. Grazie a questo procedimento risulta possibile stimare un variogramma isotropico usando le correlazioni e le distanze nel nuovo *D-space*.

Le funzioni di lisciamento, stimate dalle correlazioni osservate tra i siti oggetto di monitoraggio del processo, insieme al modello di variogramma isotropico stimato, consentono di trovare la stima della correlazione spaziale

tra ogni coppia di punti del campo aleatorio oggetto di analisi.

Una presentazione più precisa e formale del metodo assume una funzione biunivoca $f : \mathbb{R}^2 \to \mathbb{R}^2$ che fa corrispondere ad ogni punto nel *G-space*, rappresentante le localizzazioni di interesse nello spazio geografico originario, un punto nel nuovo spazio *D-space*.

Per un sito s_i, nel *G-space*, il sito corrispondente nel *D-space* è ottenuto come $z_i = f(s_i)$; equivalentemente, la funzione inversa si denota tramite $s_i = f^{-1}(z_i)$.

Il variogramma del campo aleatorio Y tra le localizzazioni s_i e s_j può essere espresso in termini delle localizzazioni nel *D-space* come

$$2\gamma(s_i, s_j) \equiv Var[Y(s_i) - Y(s_j)] = Var[Y(z_i) - Y(z_j)] = 2g(|h_{ij}^D|),$$

dove z_i e z_j sono le corrispondenti localizzazioni, $|h_{ij}^D| = |z_i - z_j|$ e g rappresenta il semivariogramma nel *D-space*. Visto l'assunto della stazionarietà isotropica, per g possono essere considerate le forme funzionali del variogramma viste al capitolo 3.

Sampson e Guttorp denominano il variogramma tra le localizzazioni nel *G-space*, $2\gamma(s_i, s_j)$, con il termine di *dispersione*, per meglio sottolineare come il campo aleatorio sia non stazionario considerandolo nelle coordinate geografiche.

Per la stima di f e g gli autori hanno proposto una procedura articolata in due passi, usando la dispersione empirica d_{ij}^2 tra le localizzazioni nel *G-space* s_i e s_j con $i, j = 1, \ldots, n$.

Nel primo passo, mediante l'utilizzo delle tecniche di *scaling* multidimensionale, che permettono di dare una rappresentazione geometrica partendo da una matrice di prossimità tra gli n siti tra i quali si assumono relazioni simmetriche, si determina la nuova rappresentazione bidimensionale delle coordinate $(z_i, \ldots, z_n)$ a partire dalle coordinate geografiche $(s_i, \ldots, s_n)$, mediante una trasformazione monotona Δ tale che

$$\Delta(d_{ij}) \equiv \ \Delta_{ij} \approx |z_i - z_j|,$$

La soluzione della relazione sopra espressa porta a esprimere la stima per

g come

$$d_{ij}^2 \equiv [\Delta^{-1}(\Delta_{ij})]^2 \approx g(|z_i - z_j|).$$

La nuova rappresentazione delle coordinate z_i viene scelta in maniera tale che le distanze tra le localizzazioni nel *D-space*, h_{ij}^D, minimizzi il seguente criterio

$$\min_\Delta \Big[\sum_{i<j} \frac{(\Delta(d_{ij}) - h_{ij}^D)^2}{\sum_{i<j}(h_{ij}^D)^2} \Big].$$

Nel secondo passo viene usato l'approccio non parametrico delle *spline* bidimensionali, note anche con il termine di *thin plate spline*, per stimare la funzione f tra le localizzazioni originali s_i e le trasformate z_i. Specificatamente,

$$f(s) = \alpha_0 + \alpha_1 s_x + \alpha_2 s_y + \sum_{i=1}^{n} \beta_i u_i(s)$$

dove $u_i(s) = ||s - s_i||^2 \log ||s - s_i||$ e s_x, s_y indicano le coordinate geografiche di s; i parametri da stimare risultano, quindi, essere α_k e β_i.

Sampson e Guttorp calcolano la funzione f mediante due *spline* f_1 e f_2 per ognuna delle coordinate di z_i mediante l'incorporazione del parametro di lisciamento λ, come visto al capitolo precedente.

Con la stima di $\hat{f}$ e $\hat{g}$, il variogramma tra ogni coppia di siti s_i e s_j nel *G-space* può essere stimato attraverso:

- il calcolo delle localizzazioni corrispondenti nel *D-space* attraverso $z_j = \hat{f}(s_j)$ per $j = 1, 2$;
- il calcolo della distanza nel *D-space* $|h_{12}^D|$ tra z_1 e z_2;
- la determinazione del variogramma $2\gamma(h) = 2\hat{g}(|h_{12}^D|)$.

Equivalentemente, la covarianza tra due localizzazioni può essere stimata attraverso la relazione $C(h) = C(0) - \hat{g}(|h_{12}^D|)$, dove $C(0)$ rappresenta la varianza del processo.

Per tener conto della diversa variabilità all'interno del campo aleatorio, l'approccio sopradescritto viene prima applicato alla matrice di correlazione; successivamente può essere incorporata una stima della varianza del campo aleatorio, in modo tale da garantire che la matrice di covarianza sia definita non negativa.

Questo approccio è stato usato con successo in una grande varietà di problemi in ambito ambientale, grazie alla flessibilità nella modellazione della non stazionarietà del campo aleatorio. Negli ultimi anni sono state proposte alcune varianti basate sul principio della massima verosimiglianza; sono stati inoltre sviluppati modelli gerachici bayesiani con l'inserimento dell'approccio di Sampson e Guttorp per interpretare la variabilità del fenomeno.

Zidek, Le e altri [20] utilizzano questo tipo di approccio per lo sviluppo di una previsione per il campo aleatorio spazio-temporale del livello di PM_{10} giornaliero nella area urbana di Vancouver. Detratta la componente di trend temporale, rappresentata da un processo AR(1), essi utilizzano un approccio gerarchico bayesiano per stimare, assumendo la non stazionarietà del campo spaziale mediante il metodo di Sampson e Guttorp, la struttura di covarianza del processo riguardante i residui.

Con questo metodo, integrato nell'approccio di Le e Zidek formulato in seguito, si procederà per l'analisi dei valori medi settimanali del PM_{10}.

L'approccio mediante medie mobili

Fuentes [4] propone un approccio per la costruzione di un processo spaziale, assumendo che la non stazionarietà sia la media pesata di processi locali stazionari non correlati tra loro; per questo la regione di interesse in cui si manifesta il fenomeno, viene suddivisa in k sub-regioni delimitate dove in ognuna viene assunto un processo stazionario isotropico. La regione geografica viene suddivisa in k sub-regioni $S_1, \ldots, S_k$, in cui $S_i \cap S_j = \varnothing$.

Fuentes rappresenta il processo non stazionario $Y(s)$ come combinazione lineare di processi locali stazionari $Y_i(s)$ mediante

$$Y(s) = \sum_{i=i}^{k} Y_i(s) w_i(s)$$

dove $cov\big(Y_i(s), Y_j(s)\big) = 0$, per $i \neq j$ e $Y_i(s)$ con $s \in S_i$ risulta essere un processo stazionario isotropico specifico della sub-regione S_i con funzione di covarianza $C_{\theta_i}(\cdot)$. I pesi $w_i(s)$ provengono da una funzione *kernel* positiva localizzata nel centroide di S_i.

La covarianza tra la coppia di localizzazioni s_1 e s_2 viene descritta come

$$\begin{aligned} cov\big(Y(s_1), Y(s_2)\big) &= \textstyle\sum_{i=1}^{k} w_i(s_1) w_i(s_2)\, cov\big(Y_i(s_1), Y_i(s_2)\big) \\ &= \textstyle\sum_{i=1}^{k} w_i(s_1) w_i(s_2) C_{\theta_i}(|h|) \end{aligned}$$

dove la covarianza $C_{\theta_i}(|h|)$, dipendente dal parametro θ_i è funzione solo della distanza $|h|$. Visto che il parametro θ_i può essere diverso per le diverse sub-regioni, $cov\big(Y(s_1), Y(s_2)\big)$ risulta essere funzione sia della distanza $|h|$ che della posizione di s_1 e s_2, data la non stazionarietà del processo nell'intera regione.

Per la stima dei parametri, Fuentes ha proposto l'uso della densità spettrale, la quale modifica la funzione di covarianza mediante la trasformata di Fourier e quindi con una sommatoria di funzioni periodiche.

I parametri corrispondenti vengono stimati dai dati e la covarianza può essere ottenuta attraverso l'inversa della trasformata di Fourier con i loro parametri stimati. Tra le possibili funzioni di covarianza proposte, si ricorda

$$C_{\theta_i}(|h|) = b_i \Big(\tfrac{|h|}{a_i}\Big)^{\nu_i} \mathcal{K}_{\nu_i}\Big(\tfrac{|h|}{a_i}\Big)$$

dove $\theta_i = (b_i, \nu_i, a_i)$ è il vettore dei parametri con ν_i, $a_i \geq 0$ e $\mathcal{K}_{\nu_i}$ rappresenta la funzione di Bessel modificata di ordine ν_i. La corrispondente densità spettrale assume la forma

$$f_i(\omega) = g(a_i, \nu_i, b_i)(a_i^{-2} + |\omega|^2)^{(-\nu_i - 1)}$$

in cui ω denota la frequenza nel dominio spettrale e g è una funzione nota di a_i, ν_i e b_i. I parametri di ogni processo locale sono trovati attraverso la stima non parametrica della densità spettrale mediante il *periodogramma* empirico.

5.2.2 L'approccio *Kernel convolution* per processi spazio-temporali

Una classe generale di processi stazionari può essere costruita mediante l'approccio del *kernel convolution* (Ver Hoef e Barry [21]), in maniera simile a quanto visto nell'approccio precedente, anche se in questo caso l'effetto della dimensione temporale viene considerata simultaneamente a quella spaziale.

Il processo spazio-temporale $w(\mathbf{s}, t)$ è pensato essere indotto da un effetto latente $v(\boldsymbol{\omega}_l, t_m)$ dove $\boldsymbol{\omega}_l$ individua una localizzazione spaziale e t_m un istante nel tempo.

Sia $K(d_s, d_t)$ il *kernel* congiunto nello spazio e nel tempo, in cui d_s e d_t rappresentano le distanze delle due dimensioni; $\boldsymbol{\omega}_l$, $l = 1, \ldots, L$ la griglia delle localizzazioni dove saranno associati i vari *kernel* lisciati ed equivalentemente t_m, $m = 1, \ldots, M$ le distanze temporali equispaziate dove saranno associati i *kernel* temporali. Allora si ha

$$w(\boldsymbol{s}, t) = \sum_{l=1}^{L} \sum_{m=1}^{M} K\big(||\boldsymbol{s} - \boldsymbol{\omega}_l||), |t - t_m|\big)\, v(\boldsymbol{\omega}_l, t_m) \tag{5.8}$$

dove $||\boldsymbol{s} - \boldsymbol{\omega}_l||$ rappresenta la distanza tra le localizzazioni $\boldsymbol{s}$ e $\boldsymbol{\omega}_l$.

Una semplice forma del *kernel* può essere espressa mediante il prodotto di due funzioni *kernel*, nello spazio e nel tempo, distinte

$$K(d_s, d_t) = K_s(d_s)\, K_t(d_t) \tag{5.9}$$

Le funzioni K_s e K_t possono essere scelte tra le funzioni di covarianza già viste.

5.2.3 Il *Kriged Kalman Filter*

Il modello sviluppato da Mardia e altri (1998) [10] combina le tecniche geostatistiche date dal *Kriging* con il filtro di Kalman. Questo approccio basato su un modello *state-space* descrive l'evoluzione dei campi spaziali nel tempo mediante un approccio gerarchico.

L'approccio visto all'inizio del capitolo, prevede che per i due livelli gerarchici il modello si esprima mediante 5.1 e 5.2.

La componente sistematica viene assunta essere una combinazione lineare di funzioni nello spazio che evolvono stocasticamente nel tempo. Le funzioni spaziali sono determinate tramite la tecnica del *kriging*, ossia con predizioni lineari non distorte per una definita forma della covarianza spaziale.
Il termine $\mu(\mathbf{s}, t)$ è dato da

$$\mu(\mathbf{s}, t) = H\alpha_t = \Big(\sum_{j=1}^{p} h_{\boldsymbol{s}_1 j}\alpha_{tj}, \ldots, \sum_{j=1}^{p} h_{\boldsymbol{s}_n j}\alpha_{tj} \Big)^T$$

con H matrice di dimensioni $x \times p$ costituita dagli elementi $h_{\boldsymbol{s}_i j}$, per $i = 1, \ldots, n$, $j = 1, \ldots, p$ e $\alpha_t = (\alpha_{t1}, \ldots, \alpha_{tp}$, componente temporale, descritta mediante una forma *state-space* (spazio degli stati), come ad esempio processi autoregressivi e/o a media mobile.

La matrice H moltiplicata per la componente dinamica temporale α_t esprime una combinazione lineare, variabile temporalmente, della superficie di regressione nello spazio, descritta dalle colonne di H rappresentanti funzioni determinate tramite *kriging*.
Le prime q colonne di H rappresentano il trend spaziale determinato dai termini relativi all'intercetta e alle funzioni, lineare e quadratica, delle coordinate $(\mathbf{X}, \mathbf{Y})$ definite tramite la matrice $\mathbf{F} = (\mathbf{1}, \mathbf{X}, \mathbf{Y}, \boldsymbol{X}^2, \boldsymbol{Y}^2)$.

Le rimanenti $p - q$ colonne sono relative alle assunzioni sulla struttura della variabilità del processo descritto da $Z(\cdot)$ distribuita normalmente, data dalla matrice di varianze e covarianze $\Sigma_\gamma = \sigma^2(\boldsymbol{s}_i, \boldsymbol{s}_j)$, non-singolare.

Considerando

$$B = \Sigma_\gamma^{-1} - \Sigma_\gamma^{-1} F (F^T \Sigma_\gamma^{-1} F)^{-1} F^T \Sigma_\gamma^{-1},$$

e applicando la decomposizione spettrale di B, in maniera tale che $B = UEU^T$ e $B\boldsymbol{u}_i = e_i \boldsymbol{u}_i$, dove $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ e $E = diag(e_1, \ldots, e_n)$ e assumendo, senza perdita di generalità, che gli autovalori siano in ordine non decrescente, quindi $e_1 = \ldots = e_q = 0 < e_{q+1} \leq e_{q+2} \leq \ldots \leq e_n$, dove le colonne di F possono essere considerate come gli autovettori associati ai primi q autovettori nulli $e_1, \ldots, e_n$.

La matrice H è, dunque, data da $H = (F, e_{q+1}\Sigma_\gamma \boldsymbol{u}_{q+1}, \ldots, e_p \Sigma_\gamma \boldsymbol{u}_p)$, dove

gli autovettori più piccoli di B sono associati alla variazione spaziale di larga scala e gli autovalori più grandi descrivono la variazione spaziale locale.

La componente temporale viene assunta essere di tipo *random walk state-space*, assumendo $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t$, in cui il termine di errore si distribuisce come $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \Sigma_\eta)$.

Questa impostazione viene sviluppata nell'analisi condotta da Sahu e Mardia [16] per la previsione a breve termine del livello di PM_{10}, grazie alla combinazione della previsione spaziale data dalla tecnica del *kriging* e l'analisi dell'evoluzione temporale data dal metodo del *Kalman filter*, il tutto tramite un approccio bayesiano implementato mediante tecniche MCMC.

5.2.4 Modelli gerarchici bayesiani

Al fine di catturare l'incertezza del fenomeno, negli ultimi anni sono stati estesi i modelli parametrici sulla base dell'approccio bayesiano, consentendo anche ai parametri della distribuzione di variare sulla base della conoscenza già nota o acquisita.

In generale, la struttura gerarchica bayesiana permette ai parametri del modello di essere considerati come variabili casuali, tramite una distribuzione a priori che incorpora le conoscenze o le ipotesi del ricercatore ex ante l'osservazione dei dati relativi al fenomeno oggetto di studio.

Questa tipologia di modelli ha avuto un importante sviluppo in seguito all'incremento della capacità dei computer e alla possibilità di implementare questa tecnica - onerosa dal punto di vista computazionale - tramite i metodi denominati Markov Chain Monte Carlo (MCMC), che hanno permesso di ampliare l'utilizzo delle distribuzioni a priori, prima ristrette alla classe delle distribuzioni coniugate.

La struttura gerarchica prevede di rappresentare il fenomeno in più passi; solitamente, al primo livello della gerarchia vengono incorporati la componente di fondo (o trend) e la covarianza senza specificarne una specifica struttura; la variabilità e le componenti del processo vengono modellate in un secondo livello.

Molti studi sul livello di inquinanti atmosferici effettuati negli ultimi anni sono basati su modelli gerarchici bayesiani per la descrizione del processo.

Il lavoro di Sahu, Gelfand e Holland [15], descrive un modello per la concentrazione del livello di $PM_{2.5}$, in cui vengono introdotti due processi, differenti sia in media che in variabilità; al primo, formulato rispetto al tipo di *background* rurale (di base), viene aggiunto il secondo presente nelle aree urbane. La non stazionarietà spaziale del processo viene introdotta mediante una variabile che rappresenta la densità di popolazione nel territorio. In ognuno dei due processi viene ipotizzata una struttura di covarianza separata nello spazio e nel tempo.

Nell'articolo di Shaddick e Wakefield [18] si considera la modellazione spazio-temporale per 4 agenti inquinanti misurati giornalmente, sempre mediante approccio bayesiano e l'utilizzo di tecniche MCMC. Il modello proposto tiene conto delle interazioni tra gli agenti inquinanti sia nello spazio che nel tempo, assumendo che il processo spaziale sia costante nel tempo, isotropico e stazionario. Per il processo temporale si prevede un modello autoregressivo del primo ordine, non stazionario nel tempo.

5.3 L'approccio di Le e Zidek

Sulla base di quanto sviluppato negli ultimi due decenni per lo studio di fenomeni ambientali, Le e Zidek [9] hanno sviluppato una struttura gerarchica bayesiana al fine di superare i limiti associati alle tecniche del *Kriging*.

Al primo livello, l'approccio prevede che il campo aleatorio, o una opportuna trasformazione della variabile che lo descrive, segua una distribuzione normale multivariata, in cui la funzione che descrive la media dipende da una matrice B di parametri ignoti. La corrispondente matrice di covarianza Σ non si presenta con una struttura preassegnata.

Al secondo livello della gerarchia, gli iperparametri B e Σ, vengono derivati da opportune distribuzioni a priori coniugate. La distribuzione di previsione, condizionata agli iperparametri, viene quindi ricavata dalla distribuzione a posteriori.

La variabilità viene, in questo modo, incorporata nella distribuzione

di previsione a posteriori; il modello consente, inoltre, di catturare, nella funzione che interpreta la media del processo, componenti di trend e/o stagionali.

La struttura proposta consente di considerare campi aleatori non stazionari, in quanto tale caratteristica può essere catturata nella stima della matrice di covarianza degli iperparametri, ad esempio, mediante l'approccio di Sampson e Guttorp.

Data la distribuzione degli iperparametri, la distribuzione di previsione risulta essere il prodotto di distribuzioni t-multivariate. Tale metodo può essere esteso al caso multidimensionale, in cui per ogni sito sono rilevati simultaneamente più fenomeni ambientali.

Visto che la distribuzione a posteriori risulta scarsamente inflenzata dalla scelta delle apriori relative agli iperparametri e la complessità del modello proposto, gli autori utilizzano un approccio più semplice di tipo bayesiano empirico, in cui gli iperparametri sono stimati sulla base della distribuzione marginale dei dati condizionati agli iperparametri stessi.

Si passa ora alla formalizzazione di quanto sopra espresso nel caso unidimensionale.

5.3.1 L'approccio gerarchico

L'approccio a più livelli adottato da Le e Zidek, consente di non definire una struttura per la forma della matrice Σ; infatti, la struttura di covarianza può essere modellata al secondo livello della struttura gerarchica attraverso la definizione di una appropriata distribuzione per gli iperparametri. I dati osservati consentono di aggiornare, grazie alla regola di Bayes, l'andamento degli iperparametri e quindi di tener conto della loro variabilità nella distribuzione a posteriori.

Dal punto di vista formale, si suppongono g assegnate localizzazioni, in cui viene osservato il campo aleatorio al variare del tempo, e u localizzazioni in cui deve essere previsto il valore del campo aleatorio (e quindi non presentano valori osservati).

Con Y_t per $t = 1, \ldots, T$, si denota il vettore riga n-dimensionale (una osservazione per ogni sito $i = 1, \ldots, n$), con $n = u + g$, al tempo t delle determinazioni del campo aleatorio, dove i primi u siti corrispondono a quelli in cui i dati non sono rilevati (siti non osservati e nei quali deve essere previsto il valore del campo aleatorio), e dove le rimanenti g localizzazioni sono quelle in cui è stato rilevato il valore assunto dal fenomeno (siti costituenti la rete di monitoraggio).

$$Y = (Y_1, \ldots, Y_T)^T = \left(\begin{array}{ccc|ccc} y_{11}^{(u)} & \cdots & y_{1u}^{(u)} & y_{1(u+1)}^{(g)} & y_{1(u+2)}^{(g)} & \cdots & y_{1n}^{(g)} \\ y_{21}^{(u)} & \cdots & y_{2u}^{(u)} & y_{2(u+1)}^{(g)} & y_{(2u+2)}^{(g)} & \cdots & y_{2n}^{(g)} \\ \vdots & & \vdots & \vdots & & & \vdots \\ y_{T1}^{(u)} & \cdots & y_{Tu}^{(u)} & y_{T(u+1)}^{(g)} & y_{T(u+2)}^{(g)} & \cdots & y_{Tn}^{(g)} \end{array} \right)$$

Come qui sopra rappresentato, il vettore Y_t può essere partizionato secondo $Y_t \equiv (Y_t^{(u)}, Y_t^{(g)})$.

5.3.2 La specificazione del modello

Si suppone che il vettore Y_t, condizionatamente indipendente, segua una distribuzione Normale multivariata, ossia

$$Y_t \mid z_t, B, \Sigma \sim N_n(z_t B, \Sigma) \tag{5.10}$$

dove con z_t si denota il vettore riga k-dimensionale delle variabili covariate e con B la matrice di dimensioni $(k \times n)$ dei coefficienti di regressione (con $n = u + g$).

Come visto per Y_t, la matrice B può essere partizionata in

$$B = \left(B^{(u)} , B^{(g)} \right).$$

Le covariate z_t possono variare nella dimensione temporale ma sono assunte essere uguali tra tutti i siti.

Anche la matrice di covarianza Σ può essere partizionata in

$$\Sigma = \left(\begin{array}{cc} \Sigma_{uu} & \Sigma_{ug} \\ \Sigma_{gu} & \Sigma_{gg} \end{array} \right)$$

dove le matrici Σ_{gg}, Σ_{uu} e Σ_{ug} rappresentano rispettivamente le covarianze di $Y_t^{(g)}$, di $Y_t^{(u)}$ e tra di loro.

5.3.3 Le distribuzioni a priori per i parametri B e Σ

Le distribuzioni a priori coniugate per i parametri sono assunte essere

$$B \mid \ B_0, \Sigma, F \sim N_{kn}\left(B_0, F^{-1} \otimes \Sigma\right) \tag{5.11}$$

$$\Sigma \mid \ \Psi, \delta \sim W_n^{-1}(\Psi, \delta) \tag{5.12}$$

dove $W_p^{-1}(\Psi, \delta)$ è la distribuzione di Wishart inversa n-dimensionale con matrice di scala Ψ e $m > n$ gradi di libertà.

Con tali ipotesi per il modello a priori le matrici relative agli iperparametri sono: F^{-1} $(k \times k)$ matrice di scala - definita positiva - associata a B_0, rappresentante la variabilità dei parametri $\boldsymbol{\beta}$; B_0 $(k \times 1)$ vettore delle medie dei coefficienti B; Ψ $(n \times n)$ matrice di covarianze stimate tra gli n siti.

Risulta conveniente riparametrizzare Σ in termini di $(\Sigma_{gg}, \Sigma_{u|g}, \tau)$ dove Σ_{gg} è la matrice di covarianza di $Y_t^{(g)}$; $\Sigma_{u|g}$ è una matrice di dimensioni $u \times u$ che rappresenta la covarianza residua di $Y_t^{(u)}$, predittore lineare basato su $Y_t^{(g)}$, dato da $\Sigma_{u|g} \equiv \Sigma_{uu} - \Sigma_{ug}\Sigma_{gg}^{-1}\Sigma_{gu}$; τ è una matrice di dimensioni $u \times g$ e rappresenta i coefficienti angolari del predittore lineare ottimo di $Y_t^{(u)}$ basato su $Y_t^{(g)}$ pari a $\tau \equiv \Sigma_{ug}\Sigma_{gg}^{-1}$. Questa trasformazione viene ottenuta grazie alla decomposizione di Bartlett (vedi Appendice D).

Usando questa nuova parametrizzazione, la distribuzione a priori coniugata per Σ, vista in 5.12, può essere equivalentemente descritta come

$$\Sigma_{gg} \mid \ \Psi, \delta \sim W_g^{-1}(\Psi_{gg}, \delta - u) \tag{5.13}$$

$$\Sigma_{u|g} \mid \ \Psi, \delta \sim W_u^{-1}(\Psi_{u|g}, \delta)$$

$$\tau \mid \ \Sigma_{u|g}, \Psi \sim N_{ug}(\tau_0, \Sigma_{u|g} \otimes \Psi_{gg}^{-1}).$$

Le matrici $\Psi_{gg}, \Psi_{u|g}, \tau_0$ denotano la decomposizione della matrice dei parametri a priori Ψ, in maniera analoga a quanto visto per Σ; quindi $\Psi_{u|g} = \Psi_{uu} -$

$\Psi_{ug}\Psi_{gg}^{-1}\Psi_{gu}$ e $\tau_0 = \Psi_{ug}\Psi_{gg}^{-1}$. Si ha inoltre che Σ_{gg} risulta indipendente da (Σ_{gg}, τ) quando la distribuzione a priori è propria.

5.3.4 La distribuzione di previsione

Definiti con $D = \{(y_1^{(g)}, z_1), \ldots, (y_T^{(g)}, z_T)\}$ i dati osservati, dove con z_t si indicano le variabili covariate, allora, date le z_t, le variabili $y_t^{(g)}$ risultano realizzazioni condizionatamente indipendenti di

$$Y_t^{(g)} \mid z_t, B, \Sigma \sim N_g\left(z_t B^{(g)}, \Sigma_{gg}\right) \tag{5.14}$$

e quindi D rappresenta l'insieme dei dati nelle g localizzazioni oggetto di rilevazione del fenomeno.

La distribuzione spaziale congiunta del vettore aleatorio $Y_f = \left(Y_f^{(u)^T}, Y_f^{(g)^T}\right)$, rappresentante il campo aleatorio al tempo futuro f, condizionata ai dati osservati D, al vettore z_f e ai parametri specificati a priori - B_0 e $(\Psi_{gg}, \Psi_{u|g}, \tau_0)$ - risulta essere

$$Y_f^{(g)} | \; D \sim t_g\Big(\mu^{(g)}, \frac{c}{l}\hat{\Psi}_{gg}, l\Big) \tag{5.15}$$

e quella relativa alle localizzazioni oggetto di previsione

$$Y_f^{(u)} | \; Y_f^{(g)} = y_f^{(g)}, D \sim t_u\Big(\mu^{(u)}, \frac{d}{q}\Psi_{u|g}, q\Big) \tag{5.16}$$

dove t_m indica la distribuzione *t-multivariata* con m gradi di libertà - vedi Appendice D - dove le costanti, indicanti i gradi di libertà, assumono i valori

$l = \delta + n - u - g + 1$;
$c = 1 + z\big(A + F\big)^{-1} z^T$;
$d = 1 + z\, F^{-1} z^T + \big(y_f^{(g)} - z_f\, B_0^{(g)}\big)\Psi_{gg}^{-1}\big(y_f^{(g)} - z_f\, B_0^{(g)}\big)^T$;
$q = \delta - u + 1$

con

$\mu^{(g)} = (I - W)\hat{B}^{(g)} + W B_0^{(g)}$
$\mu^{(u)} = z_f B_0^{(u)} + \tau_0\big(y_f^{(g)} - z_f B_0^{(g)}\big)$
$\hat{\Psi}_{gg} = \Psi_{gg} + S + (\hat{B}^{(g)} - B_0^{(g)})^T (A^{-1} + F^{-1})^{-1} (\hat{B}^{(g)} - B_0^{(g)})$

Le quantità sopra riportate sono espresse da

$$\hat{B}^{(g)} = A^{-1}\,C \qquad \text{con} \quad A = \sum_{t=1}^{n} z_t'\, z_t \quad \text{e} \quad C = \sum_{t=1}^{n} z_t'\, y_t^{(g)}$$

$$S = \sum_{t=1}^{n} \left(y_t^{(g)} - z_t \hat{B}^{(g)}\right)^T \left(y_t^{(g)} - z_t \hat{B}^{(g)}\right)$$

$$W = (A + F)^{-1} F^{-1}.$$

In questo modo, la matrice di covarianza a posteriori $\hat{\Psi}_{gg}$, include le informazioni della distribuzione a priori data da Ψ_{gg}, delle osservazioni S e del modello. Le medie a posteriori $\mu^{(g)}$ e $\mu^{(u)}$ incorporano i contributi sia della distribuzione a priori sia delle osservazioni.

Quanto visto qui sopra, può essere riassunto e puntualizzato come segue:

• La distribuzione a posteriori si caratterizza come prodotto di due distribuzioni t di Student, noti gli iperparametri $\{B_0, F, \Psi, \delta\}$. All'istante temporale f, dove $1 \leq f \leq T$, la distribuzione nei siti non osservati, ossia dove deve essere eseguita la previsione, è data dalla 5.16 e risulta avere code più pesanti di una distibuzione Normale visto l'inserimento nella distribuzione a priori dell'incertezza associata ai parametri del modello.
• I contributi del modello, dei dati osservati - presenti nei parametri della distribuzione di previsione - e della conoscenza disponibile, si riscontrano nella distribuzione a posteriori. La media $\mu^{(g)}$ della distribuzione di previsione, nei siti oggetto di rilevazione, è la media pesata della stima lineare ottima basata sui dati osservati e la media a priori; la media $\mu^{(u)}$ associata alle localizzazioni in cui non ci sono osservazioni, risulta il previsore lineare ottimo basato sui dati osservati e sulla conoscenza a priori. La matrice $\hat{\Psi}_{gg}$, che riflette la covarianza tra i siti oggetto di rilevazione, è la somma della corrispondente matrice a priori, della somma dei quadrati dei residui e del modello a priori.
• Non viene imposta nessuna condizione sulla stazionarietà del processo per la matrice di covarianze Σ. La struttura di covarianza del campo aleatorio può essere modellata attraverso la struttura di Ψ, e si vedrà nell'analisi dei dati come per tale struttura si farà ricorso al metodo non parametrico di Sampson e Guttorp.
• Il modello specificato in 5.10 consente l'utilizzo di variabili covariate; questo

permette di incorporare eventuali componenti di trend e di stagionalità nella distribuzione a posteriori senza la loro rimozione nella parte di analisi preliminare.

• La struttura proposta può essere modificata per incorporare in z_t trasformazioni delle coordinate spaziali di tipo $f_l(s_i)$, $i = 1, \ldots, n$ e $l = 1, \ldots, L$ dove le $f_l(\cdot)$ sono funzioni specificate e s_i le coordinate relative al sito i-esimo. Queste coordinate sono usate per la previsione mediante l'approccio del *kriging* universale, come visto nel Capitolo 3.

5.3.5 Modellazione dei dati mancanti

Il problema dei dati mancanti è inevitabilmente connaturato allo studio dei fenomeni ambientali per una molteplicità di cause.

Le reti di rilevamento per i fenomeni ambientali sono spesso soggette ad aggiornamenti incrementali in cui, conformemente all'evolversi delle disposizioni normative e alle disponibilità tecnologiche ed economiche, vengono aggiunte e attivate, via via nel corso del tempo, le stazioni di monitoraggio.

Le e Zidek hanno formulato una soluzione, nel caso in cui tale mancanza sia dovuta all'attivazione in tempi diversi dei sistemi di rilevamento e/o misurazione del fenomeno, per cui la matrice dei dati si presenta come raffigurata in Tab. 5.1, dove il simbolo (o) indica i dati rilevati e (x) i dati mancanti a causa del differimento temporale del momento di attivazione della centralina $(u+1, \ldots, g)$ o perché relativi alle localizzazioni in cui devono essere eseguite le previsioni $(1, \ldots, u)$. La matrice dei dati viene rappresentata con una struttura monotona a 'scalini' - eventualmente dopo una riorganizzazione in maniera appropriata delle colonne $1, \ldots, g$ - dove l'insieme delle stazioni oggetto di rilevazione viene raggruppato in blocchi. Tale struttura viene denominata dai due studiosi con il termine di *staircase pattern*.

Per esemplificare la trattazione formale riportata qui sotto, si considera il caso con $k = 2$ blocchi di osservazioni e in cui con g_1 e g_2 si indicano il numero di stazioni in cui vengono rilevate le misurazioni del fenomeno nei due blocchi, ovviamente si ha $g = g_1 + g_2$; con m_1 ed m_2 vengono indicati la numerosità dei valori mancanti in ogni blocco.

Le variabili che descrivono il fenomeno nei siti oggetto di previsione e di rilevazione si rappresentano mediante

tempo																	
1	x	x	...	x	x	x	x	x	x	...	x	x	x	x	o	o
2	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	o	o
														x	o	o
														o		
														o		
											x	x	x			
											o	o	o			
											o	o	o			
							x	x	x							
							o	o	o							
					x	x	o	o	o							
					o	o										
					o	o										
	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮
T	x	x	...	x	o	o	o	o	o	...	o	o	o	o	o	o
	1	2	...	u	u+1	u+2	u+3				...				n-1	n

Tabella 5.1: Struttura della matrice dei dati

$$Y \equiv \left[Y^{[u]}, Y^{[g]}\right] \equiv \left[Y^{[u]}, Y^{[g_1]}, Y^{[g_2]}\right]$$

in cui
•$Y^{[u]}$ risulta essere la matrice $(n \times u)$ della variabile risposta nei siti oggetto di previsione e quindi non osservati;
•$Y^{[g]}$ risulta essere la matrice $(n \times g)$ della variabile risposta nei siti oggetto di misurazione;
•$Y^{[g_j]}$ risulta essere la matrice $(n \times g_j)$ della variabile risposta nei siti oggetto di misurazione per il j-esimo blocco con $j = 1, 2$.

Partizionando il blocco della variabile risposta nelle componenti contenenti i dati mancati e quelli osservati si ottiene

$$Y = \left[Y^{[u]}, \binom{Y^{[g_1^m]}}{Y^{[g_1^o]}}, \binom{Y^{[g_2^m]}}{Y^{[g_2^o]}}\right]$$

in cui in ognuno dei j-esimo blocco:
•$Y^{[g_j^m]}$ risulta essere la matrice $m_j \times g_j$ dei valori mancanti - *missing value* - della variabile risposta nei siti oggetto di misurazione;
•$Y^{[g_j^o]}$ risulta essere la matrice $m_j \times g_j$ dei valori osservati della variabile

risposta nei siti oggetto di misurazione.

L'insieme delle l variabili covariate o concomitanti, in ogni istante di osservazione t, può essere espresso come $Z_t = \left(z_{t1}, \ldots, z_{tl}\right)^T$, con l'assunzione che siano costanti rispetto alla dimensione spaziale.

Partizionando la matrice dei coefficienti $\boldsymbol{\beta}$ - di dimensione $l \times (u+g)$ - e la matrice di covarianza Σ - di dimensione $(u+g) \times (u+g)$ - nei siti oggetto di previsione (u) e di misurazione (g) si ottiene

$$\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{[u]}, \boldsymbol{\beta}^{[g_1]}, \boldsymbol{\beta}^{[g_2]}\right)$$

$$\Sigma = \begin{pmatrix} \Sigma^{[u,\,u]} & \Sigma^{[u,\,g]} \\ \Sigma^{[g,\,u]} & \Sigma^{[g,\,g]} \end{pmatrix}$$

Analogamente partizionando la sottomatrice $\Sigma^{[g,g]} = \begin{pmatrix} \Sigma^{[g_1,\,g_1]} & \Sigma^{[g_1,\,g_2]} \\ \Sigma^{[g_2,\,g_1]} & \Sigma^{[g_2,\,g_2]} \end{pmatrix}$

Mediante la trasformazione di Bartlett (vedi Appendice D) della matrice Σ si ottiene la seguente semplificazione

$$\Sigma_{22} = \Sigma^{[g_2,\,g_2]}$$

$$\Gamma_1 = \Sigma^{[g_1,\,g_1]} - \Sigma^{[g_1,\,g_2]}\left(\Sigma^{[g_2,\,g_2]}\right)^{-1}\Sigma^{[g_2,\,g_1]}$$

$$\tau_1 = \left(\Sigma^{[g_2,\,g_2]}\right)^{-1}\Sigma^{[(g_2,g_2),\,g_1]}$$

5.3.6 Specificazione del modello *Staircase*

La variabile risposta Y viene assunta essere distribuita secondo un modello gaussiano-Wishart inverso generalizzato; per cui, usando la notazione sopra riportata

$$\begin{cases} Y \mid \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, I_n \otimes \Sigma) \\ \boldsymbol{\beta} \mid \Sigma, \boldsymbol{\beta}_0, F \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma) \\ \Sigma \sim GIW(\Psi, \delta) \end{cases} \tag{5.17}$$

dove $\boldsymbol{\beta}_0$ è la matrice delle medie degli iperparametri di $\boldsymbol{\beta}$ di dimensione $l \times (g+u)$; F^{-1}, di dimensione $l \times l$, è una matrice definita positiva rap-

presentante la varianza dei componenti di $\boldsymbol{\beta}$ all'interno delle l righe; Z è la matrice delle variabili covariate.

GIW rappresenta la distribuzione inversa generalizzata di Wishart con Ψ insieme degli iperparametri e con $\delta = (\delta_0, \delta_1, \delta_2)^T$ che rappresenta i gradi di libertà (vedi Appendice D).

Attraverso la decomposizione di Bartlett, la distribuzione GIW, sopra presentata, viene ridefinita ad un successivo livello mediante:

$$\begin{cases} \Sigma^{[g,\,g]} \sim GIW(\Psi^{[g]}, \delta^{[g]}) \\ \\ \Gamma^{[u]} \sim IW(\Psi_0, \delta_0) \\ \\ \tau^{[u]} \mid \Gamma^{[u]} \sim N(\tau_{00}, H_0 \otimes \Gamma^{[u]}) \end{cases} \tag{5.18}$$

dove $\Gamma^{[u]} = \Sigma^{[u|g]} - \Sigma^{[u,u]}\left(\Sigma^{[g,g]}\right)^{-1}\Sigma^{[g,u]}$; $\tau^{[u]} = \left(\Sigma^{[g]}\right)^{-1}\Sigma^{[gu]}$.

IW definisce la distribuzione di Wishart inversa con iperparametri (ψ_0, δ_0); la matrice τ_{00} è l'iperparametro di $\tau^{[u]}$ e la matrice H_0 è la componente della varianza di τ_u tra le righe.

Proseguendo con la specificazione della distribuzione, la matrice $\Sigma^{[g,g]}$, sempre tramite la decomposizione di Bartlett, può essere vista come funzione del nuovo insieme di variabili $(\Sigma_{22}, \Gamma_1, \tau_1)$ la cui distribuzione è data da:

$$\begin{cases} \Sigma_{22} \sim IW(\Psi_2, \delta_2) \\ \\ \tau_1 \mid \Gamma_1 \sim N(\tau_{01}, H_1 \otimes \Gamma_1) \\ \\ \Gamma_1 \sim IW(\Psi_1, \delta_1) \end{cases} \tag{5.19}$$

Gli iperparametri coinvolti in questo modello possono essere descritti come

$$\mathcal{H} = \{\boldsymbol{\beta}_0, F, \Psi, \delta\} \text{ dove } \Psi = \{\Psi_0, \tau_{00}, H_0, \Psi_1, H_1, \tau_{01}, \Psi_2\} \text{ e } \delta = (\delta_0, \delta_1, \delta_2)^T.$$

La distribuzione *GIW* (Brown e altri, 1994), generalizza la distribuzione *IW* permettendo diversi gradi di libertà per una matrice casuale definita positiva. Tale distribuzione inoltre, risulta essere una apriori coniugata rispetto alla Normale e consente di gestire opportunamente la struttura a scalini dei dati osservati, attraverso diversi gradi di libertà per i vari blocchi, espressi dal vettore degli iperparametri δ.

5.3.7 La distribuzione di previsione

Se vengono indicate con $Y_{noss} = \left\{Y^{[u]}, Y^{[g_1^m]}, Y^{[g_2^m]}\right\}$ le risposte non osservate in tutte le localizzazioni e con $D = \left\{Y^{[g_1^o]}, Y^{[g_2^o]}\right\}$ le rilevazioni osservate nelle localizzazioni oggetto di misurazione, nell'ipotesi del modello 5.17, la distribuzione sui siti non osservati, condizionata ai dati osservati e all'insieme degli iperparametri $\mathcal{H}$, è data da

$$\left(Y_{noss}|\ D, \mathcal{H}\right) \sim \left(Y^{[u]}|\ Y^{[g_1^m]}, Y^{[g_2^m]}, D, \mathcal{H}\right)\left(\ Y^{[g_1^m]}|\ Y^{[g_2^m]}, D, \mathcal{H}\right)\left(Y^{[g_2^m]}|\ D, \mathcal{H}\right) \tag{5.20}$$

in cui le tre componenti delle distribuzioni condizionate sono specificate secondo:

$$\left(Y^{[g_2^m]}|\ D, \mathcal{H}\right) \sim t_{m_2 \times g_2}\left(\mu^{[2]}_{(u|g)}, \Phi^{[2]}_{(u|g)} \otimes \Psi^{[2]}_{(u|g)}, \delta^{[2]}_{(u|g)}\right);$$

$$\left(Y^{[g_1^m]}|\ Y^{[g_2^m]}, D, \mathcal{H}\right) \sim t_{m_1 \times g_1}\left(\mu^{[1]}_{(u|g)}, \Phi^{[1]}_{(u|g)} \otimes \Psi^{[1]}_{(u|g)}, \delta^{[1]}_{(u|g)}\right);$$

$$\left(Y^{[u]}|\ Y^{[g_1^m]}, Y^{[g_2^m]}, D, \mathcal{H}\right) \sim t_{n \times u}\left(\mu^{[u|g]}, (\delta_0 - u + 1)\Phi^{[u|g]} \otimes \Psi_0, (\delta_0 - u + 1)\right)$$

dove $t_{m_j \times g_j}$ indica la distribuzione *matric-t* (vedi Appendice D).

Le medie a posteriori $\mu^{[j]}_{(u|g)}$ per $j = 1, 2$ che combinano i dati osservati e la conoscenza a priori, rappresentano il miglior predittore lineare per i dati mancanti nei siti oggetto di rilevazione; similmente, $\mu^{[u|g]}$ rappresenta il miglior predittore lineare nei siti non rilevati, nel caso in cui Σ sia nota.

Le matrici $\Phi^{[j]}_{(u|g)}$, $\Psi^{[j]}_{(u|g)}$ rappresentano la struttura di covarianza per la distribuzione di previsione.

Per i siti in cui effettuare la previsione (u), i dati osservati contribuiscono tramite la matrice $\Phi^{[u|g]}$.

5.4 Analisi delle medie settimali del PM_{10} secondo l'approccio gerarchico di Le e Zidek

Come già descritto nel Capitolo 4, negli ultimi anni, per lo studio e l'analisi del comportamento dei processi ambientali, sono state sviluppate metodologie evolute per superare i limiti posti dai modelli derivanti dalla geostatistica e usati per l'interpretazione del processo nella dimensione spaziale.

Questa nuova classe di modelli supera il limite dovuto all'ipotesi della stazionarietà del processo dal punto di vista spaziale ed espande la struttura temporale, per la quale vengono proposti metodi più complessi di analisi della dinamica del processo.

In questa parte verrà utilizzato un approccio gerarchico, secondo quanto formulato da Le e Zidek, che consente di determinare la distribuzione della previsione spaziale in siti in cui il processo non è stato rilevato.

L'analisi è stata sviluppata utilizzando la libreria di funzioni sviluppata dagli stessi autori, presentata nell'Appendice C. La struttura gerarchica consente di pervenire ai risultati attraverso una serie di passi, più sotto descritti, secondo quanto visto a livello teorico nel precedente paragrafo.

5.4.1 La riorganizzazione dei dati

L'utilizzo dell'approccio delineato da Le e Zidek prevede che i dati abbiano, almeno approssimativamente, una distribuzione gaussiana e che le determinazioni siano indipendenti. Per tale ragione si continua ad operare con i dati relativi alle medie settimanali della concentrazione di PM_{10} trasformati mediante la funzione logaritmo, che tende a rendere i dati osservati simili alla distribuzione richiesta con buona approssimazione.

Un altro requisito richiesto prevede che le serie dei dati presentino valori mancanti (NA) esclusivamente all'inizio della serie, e conseguentemente, per l'analisi, si utilizzano i dati di 15 stazioni che dispongono delle serie complete dei dati e di altre 5 che hanno dei valori mancanti all'inizio della serie. Dall'analisi per la costruzione del modello sono escluse le serie che contengono valori mancanti all'interno o alla fine dell'anno.

Al fine di catturare la componente temporale del processo, sono state eseguite diverse prove utilizzando la variabile settimana $Z_t = 1, \ldots, 52$ (come

nell'approccio visto all'inizio del capitolo); per problemi riscontrati nell'utilizzo della funzione che determina le stime iniziali degli iperparametri, probabilmente dovuti ad una sovraparametrizzazione, si è optato per definire la variabile covariata temporale secondo il mese, per cui $Z_t = 1, \ldots, 12$.

La matrice dei dati viene organizzata secondo una struttura monotona decrescente - *staircase pattern* - relativamente al numero di dati mancanti presenti in ogni stazione, come previsto dagli autori delle funzioni; si procede quindi all'ordinamento delle stazioni conformemente a tale assunto.

5.4.2 Stima degli iperparametri sui siti osservati

Come già visto nel capitolo precedente, il modello che descrive la distribuzione della variabile risposta Y - indicante il valore del logaritmo della concentrazione di PM_{10} - viene assunto essere

$$\begin{cases} Y \mid \ \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, I_n \otimes \Sigma) \\ \\ \boldsymbol{\beta} \mid \ \Sigma, \boldsymbol{\beta}_0, F \sim N(\boldsymbol{\beta}_0, F^{-1} \otimes \Sigma) \\ \\ \Sigma \sim GIW(\Psi, \delta) \end{cases} \tag{5.21}$$

La distribuzione di previsione nei siti in cui non viene rilevato il fenomeno, come visto in 5.20, risulta essere condizionata ai dati osservati $Y^{[g]}$ e ai valori degli iperparametri, associati ai siti osservati, organizzati in k blocchi,

$$\mathcal{H}_g = \left\{F, \beta_0, \Omega, (\tau_{0,1}, H_1, \Lambda_1, \delta_1), \ldots, (\tau_{0,k-1}, H_{k-1}, \Lambda_{k-1}, \delta_{k-1}), (\Lambda_k, \delta_k)\right\}$$

e quelli associati alle localizzazioni in cui viene effettuata la previsione e nei quali non ci sono osservazioni, ovvero $\Lambda^{[u]}$, $\tau_0^{[u]}$, $H^{[u]}$, $\delta^{[u]}$.
La prima operazione da effettuare consiste nel determinare, a partire dai dati osservati, le stime degli iperparametri $\mathcal{H}_g$ secondo un approccio bayesiano empirico, ossia massimizzando la verosimiglianza, ovvero la funzione di densità valutata sulle osservazioni rilevate.

L'algoritmo EM implementato (Dempster, Laird, Rubin - 1977) consente la

massimizzazione di una funzione obiettivo secondo una procedura iterativa in due passi: **E-step** valuta il valore atteso condizionatamente ai valori osservati e alle stime dei parametri calcolate per la corrente iterazione; **M-step** massimizza la funzione obiettivo e aggiorna le stime dei parametri ottenute nella iterazione precedente (si ottiene quindi la stima aggiornata $\mathcal{H}_g^{(j+1)}$ di $\mathcal{H}_g$ alla $j+1$-esima iterazione).

L'output della funzione mette a disposizione:

- il numero di blocchi k in cui sono aggregate le stazioni che presentano la stessa numerosità di valori osservati;
  I primi 5 riguardano le stazioni dove si trovano valori mancanti (NA)

blocco	1	2	3	4	5	6
numero stazioni	1	1	1	1	1	15

  di numerosità diversa in ogni serie - e l'ultimo con 15 stazioni in cui sono disponibili le serie complete di 52 osservazioni;

- la stima dei coefficienti β_{0t} con $t = 1, \ldots, 12$, assunti ugualmente distribuiti e di conseguenza gli stessi valori sono associati ad ogni stazione;

β_{0_1}	β_{0_2}	β_{0_3}	β_{0_4}	β_{0_5}	β_{0_6}
4.716	0.0341	-0.239	-0.927	-1.448	-1.150
β_{0_7}	β_{0_8}	β_{0_9}	$\beta_{0_{10}}$	$\beta_{0_{11}}$	$\beta_{0_{12}}$
-1.305	-1.829	-1.358	-0.981	-1.035	-0.829

Tabella 5.2: Stime degli iperparametri β_{0t}, $t = 1, \ldots, 12$

- la stima di Λ che rappresenta la matrice di covarianza condizionata relativa alle stazioni presenti in ogni blocco (Appendice C - Tab. C.1);

- la stima di Ψ (Appendice C - Tab. C.2) ovvero la matrice di covarianza non condizionata tra tutte le stazioni, che misura il grado di legame tra le osservazioni delle varie stazioni per la dimensione spaziale. Il risultato ottenuto, trasformato mediante la relazione $2\gamma(h) = 2C_\Psi(0) - 2C_\Psi(h)$ e ottenendo di conseguenza i valori stimati del variogramma, espressi in funzione della distanza, ossia il variogramma nuvola, viene presentato nella Fig. 5.1, in cui viene aggiunto il variogramma di tipo esponenziale.

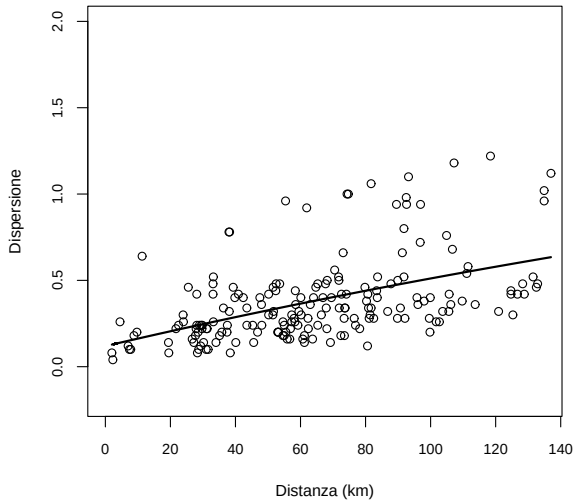


Figura 5.1: Dispersione tra le varie stazioni e variogramma stimato mediante struttura di correlazione esponenziale (nel *G-space*)

5.4.3 Stima della covarianza spaziale tramite il metodo di Sampson e Guttorp

La matrice di covarianza non condizionata Ψ, stimata tra le stazioni oggetto di osservazione, viene estesa ai siti in cui non si hanno osservazioni e nei quali si vuole ottenere una previsione attraverso il metodo non parametrico implementato da Sampson e Guttorp, come visto precedentemente. In questo modo si ottiene una stima della struttura di variabilità spaziale che supera l'ipotesi di stazionarietà del processo in tutto il campo aleatorio.

La procedura di stima avviene tramite una sequenza di operazioni:

Step 1 Il primo passo consiste nel determinare la nuova configurazione delle coordinate relative alle stazioni in cui è stato rilevato il fenomeno, ossia passare dalle localizzazioni nel *G-space* a quelle del *D-space* in cui sia verificata l'ipotesi di stazionarietà del fenomeno. Il termine di 'dispersione' usato da Sampson e Guttorp, al posto dell'usuale 'variogramma' mette l'accento sul fatto che la struttura di correlazione nello spazio originario (spazio geografico) può non seguire un modello isotropico.
La funzione, utilizzando un algoritmo iterativo, permette di deter-

minare le nuove coordinate, usando tecniche di *scaling* multidimensionale, stimando il variogramma, nel caso specifico, di tipo esponenziale. Nella Fig. 5.2 vengono visualizzate le situazioni di partenza (*G-space*) e di arrivo (*D-space*) per la disposizione delle coordinate riguardanti le stazioni, la nuvola del variogramma empirico e il variogramma stimato di tipo esponenziale. Come si può osservare, risulta evidente

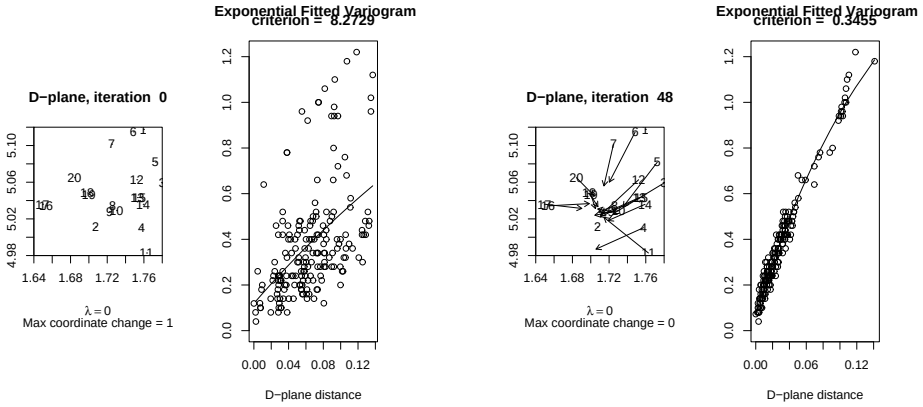


Figura 5.2: Coordinate e stima del variogramma nel *G-space* (a sx) e nel *D-space* (a dx)

come il variogramma esponenziale calcolato tra le coordinate espresse nel *D-space*, interpreti il fenomeno in maniera più adeguata, rendendo evidente come le ipotesi di stazionarietà e di isotropia usate finora, possano essere previste solo in prima approssimazione quando si opera nello spazio geografico canonico.

Step 2 Dopo aver trovato le nuove coordinate nel *D-space*, viene determinato il parametro di lisciamento per le *thin-plate spline*, al fine di evitare un effetto, indotto nel caso di forte deformazione, per cui localizzazioni lontane possono presentare valori di correlazione maggiore rispetto a quelle collocate geograficamente tra queste.
Utilizzando la funzione a disposizione si procede alla scelta di λ in modo interattivo, mediante la valutazione dell'esito, dato dal valore scelto per il parametro di lisciamento, sulla deformazione dello spazio. La scelta, risultante da un *trade-off* tra la stima del variogramma e il grado di lisciamento della deformazione dello spazio - vedi Fig.5.3 - porta

ad un valore del parametro λ pari a 0.001, con cui il *D-space* risultante si configura come un insieme convesso.

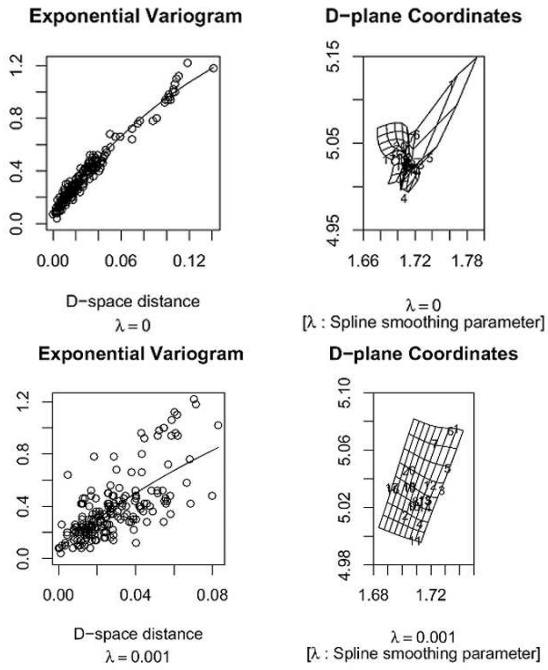


Figura 5.3: Definizione del parametro di lisciamento λ per la *thin-plate spline*: senza lisciamento ($\lambda = 0$) (in alto) e con lisciamento ($\lambda = 0.001$) (in basso)

Step 3 In questo passo vengono combinati i risultati ottenuti nei due step precedenti per creare la *thin-plate spline* che consente di mappare le coordinate dallo spazio geografico al *D-space*.
La Fig. 5.4 presenta, su griglia, il comportamento relativo alle coordinate dal quale si evidenzia una contrazione dello spazio nella direzione sud-nord (linea continua) e una espansione dello spazio nella direzione ovest-est (linea tratteggiata) della *thin-plate spline*.

Step 4 Vengono ora utilizzati i risultati del passo 3 e del modello di variogramma esponenziale - visto al passo 1 - al fine di ottenere la struttura spaziale della variabilità, sia tra i siti in cui sono presenti le stazioni, sia tra le nuove localizzazioni nelle quali effettuare la previsione spaziale.
Per ottenere questo viene creata una griglia di 400 (20x20) punti equispaziati, in modo da ricoprire uniformemente l'area geografica contenente le stazioni.

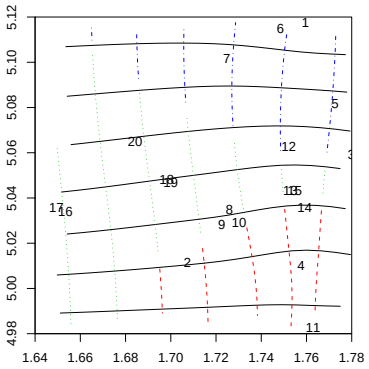


Figura 5.4: Griglia con contrazione (linea continua) - espansione (linea tratteggiata) per la *thin-plate spline*

Una funzione opportuna consente di determinare la correlazione stimata per tutte le 420 localizzazioni - 20 relative alle stazioni in cui sono rilevate i valori e 400 in cui effettare la previsione. Nella Tab.5.3 vengono visualizzate le correlazioni stimate per le prime 10 localizzazioni.

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	l_{10}
l_1	1.00	0.95	0.94	0.93	0.93	0.92	0.91	0.90	0.89	0.88
l_2	0.95	1.00	0.95	0.94	0.93	0.93	0.92	0.91	0.90	0.89
l_3	0.94	0.95	1.00	0.95	0.94	0.94	0.93	0.92	0.91	0.90
l_4	0.93	0.94	0.95	1.00	0.95	0.94	0.94	0.93	0.92	0.91
l_5	0.93	0.93	0.94	0.95	1.00	0.95	0.94	0.94	0.93	0.92
l_6	0.92	0.93	0.94	0.94	0.95	1.00	0.95	0.95	0.94	0.93
l_7	0.91	0.92	0.93	0.94	0.94	0.95	1.00	0.95	0.95	0.94
l_8	0.90	0.91	0.92	0.93	0.94	0.95	0.95	1.00	0.95	0.95
l_9	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.95	1.00	0.95
l_{10}	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.95	1.00

Tabella 5.3: Valori della correlazione stimata tra le prime 10 localizzazioni

Step 5 In questo passo viene stimata la varianza relativa a tutte le localizzazioni e combinata con la matrice di correlazione al fine di ottenere la matrice di covarianza stimata, attraverso la usuale relazione $Cov(\cdot) = Corr(\cdot) \times \sqrt{Var(\cdot)}$ sull'insieme delle 420 localizzazioni. Dal momento che la variabilità del campo spaziale non figura essere omogenea tra le stazioni, la varianza viene stimata utilizzando la fun-

zione della libreria con gli stessi parametri relativi alla *thin-plate spline* determinati al passo 3.

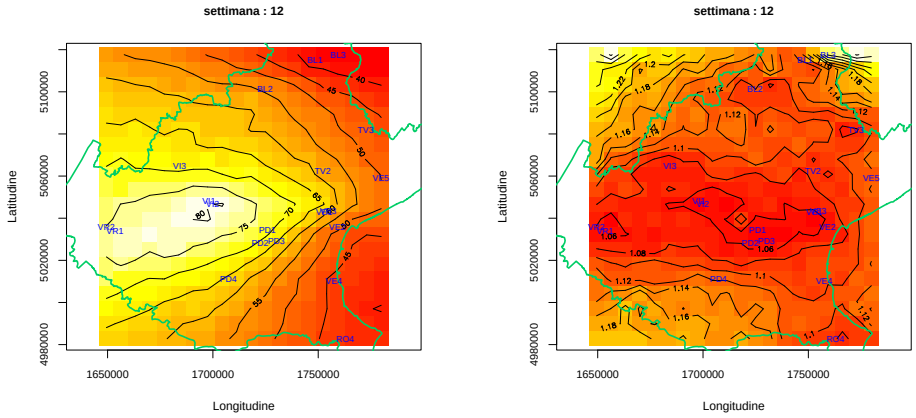
Con questi 5 passi la procedura relativa al metodo di Sampson e Guttorp per la stima della matrice di covarianza spaziale è completata.

5.4.4 Stima dei parametri nei siti oggetto di previsione

Integrando le stime iniziali degli iperparametri nei siti oggetto di rilevazione e la struttura di covarianza calcolata non parametricamente come sopra descritto, si perviene alla stima degli iperparametri nei siti oggetto di previsione attraverso la funzione che combina i valori iniziali, determinati all'inizio della procedura, con il risultato ottenuto mediante il metodo di Sampson e Guttorp, al fine di determinare le stime degli iperparametri relative alle 400 localizzazioni uniformemente distribuite nello spazio.

5.4.5 La previsione spaziale

Tramite la funzione di simulazione, si procede alla generazione di un campione di 1000 valori per ognuna delle 400 localizzazioni in cui prevedere il fenomeno in 4 diverse settimane nel corso dell'anno (rispettivamente 12, 24, 36, 48). La previsione dell'andamento viene ottenuta tramite il valore medio delle 1000 determinazioni campionarie; sempre tramite i valori simulati può essere determinato lo standard error per dare una misura dell'incertezza della previsione. In Fig.5.5 viene rappresentato l'andamento della concentrazione di PM_{10} nella scala originaria.



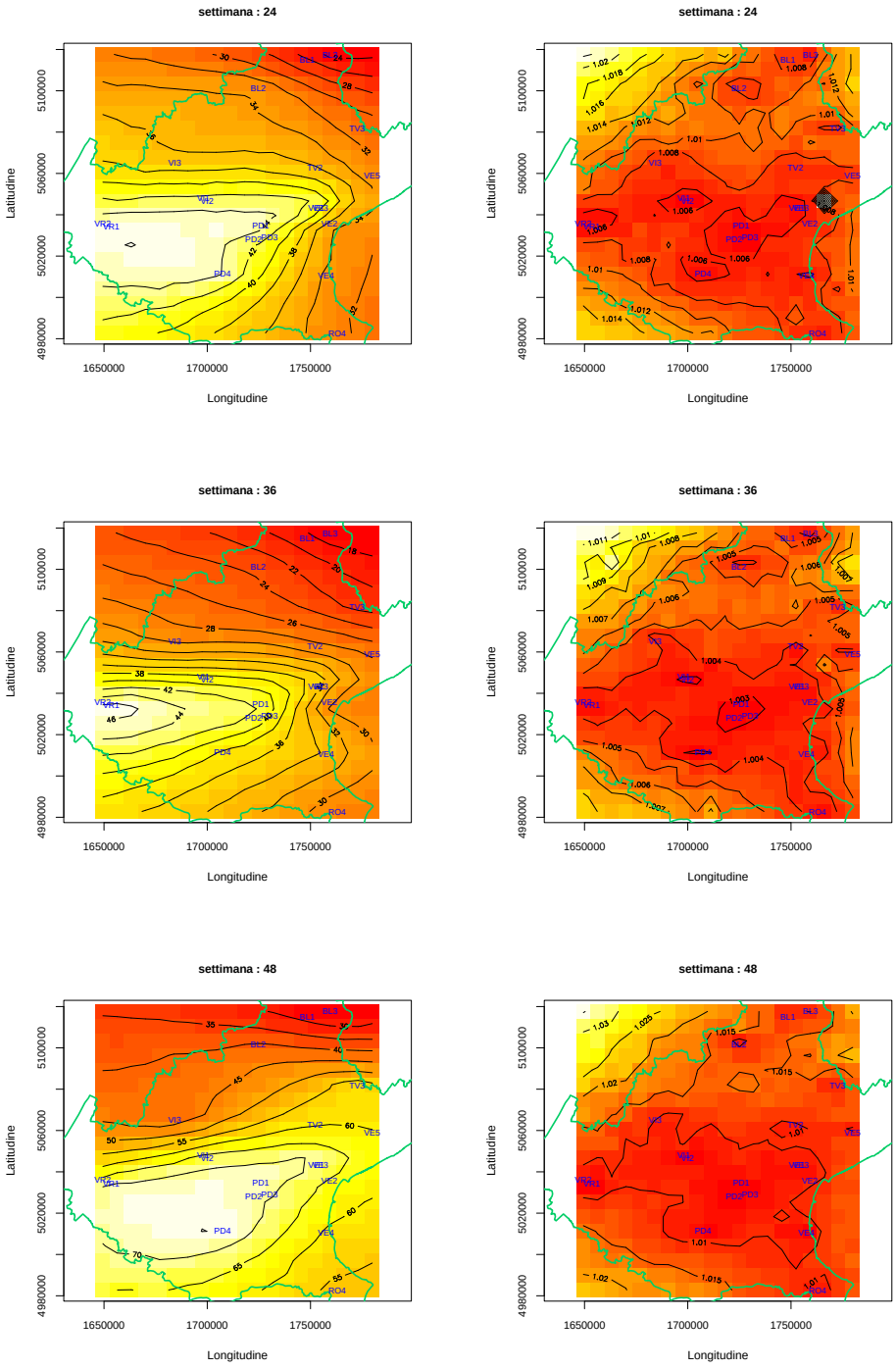


Figura 5.5: Previsione (a sx) e standard error (a dx), per 4 diverse settimane, del livello di concentrazione di PM_{10} mediante l'approccio gerarchico proposto da Le e Zidek

Dall'analisi dei grafici, si può notare come le previsioni relative alle varie settimane diano l'idea di come il fenomeno si evolva nel tempo, ma anche di come la caduta dell'ipotesi di stazionarietà, renda l'andamento del fenomeno nello spazio diverso e più verosimile di quanto presentato nell'approccio spazio-temporale con stima del trend non parametrica visto al Capitolo 4.

Per gli standard error, si nota come nell'area prossima alle stazioni e comunque in tutta quella comprendente la pianura veneta risultino sullo stesso livello, mentre si incrementano man mano che ci si situa nella parte montana, oltre che al di fuori dei confini regionali.

Un confronto tra i risultati di questo approccio all'analisi dei dati e quello sviluppato precedentemente propone una similarità per la struttura spaziale del fenomeno.

5.4.6 La convalida del modello

Per poter dare una indicazione qualitativa di quanto questa modellazione permette di descrivere il fenomeno, per ogni stazione è stato elaborato un modello, con gli stessi passi visti sopra, conformemente al metodo di convalida incrociata *one-leave-out*, in cui le stime degli iperparametri sono determinate utilizzando i valori delle rimanenti e calcolando i valori per le 52 settimane tramite simulazione.

Le bande rappresentano i quantili empirici, ricavati dalla distribuzione dei dati simulati con 1.000 replicazioni, rispettivamente di probabilità pari a 0.025 e 0.975, in modo da rappresentare una banda di variabilità di previsione approssimata al 95%.
Nella Fig. 5.6 sono presentati, per quattro stazioni il grafico che individua i due limiti sopracitati e i valori delle medie settimali della concentrazione di PM_{10}, ottenuti tramite simulazione, su scala originaria.

I valori osservati (punti) si trovano, per quasi tutte le stazioni e per quasi tutte le 52 settimane, all'interno della banda, e questo consente di valutare l'approccio implementato, al fine di dare una rappresentazione del processo spazio-temporale, in termini abbastanza soddisfacenti.

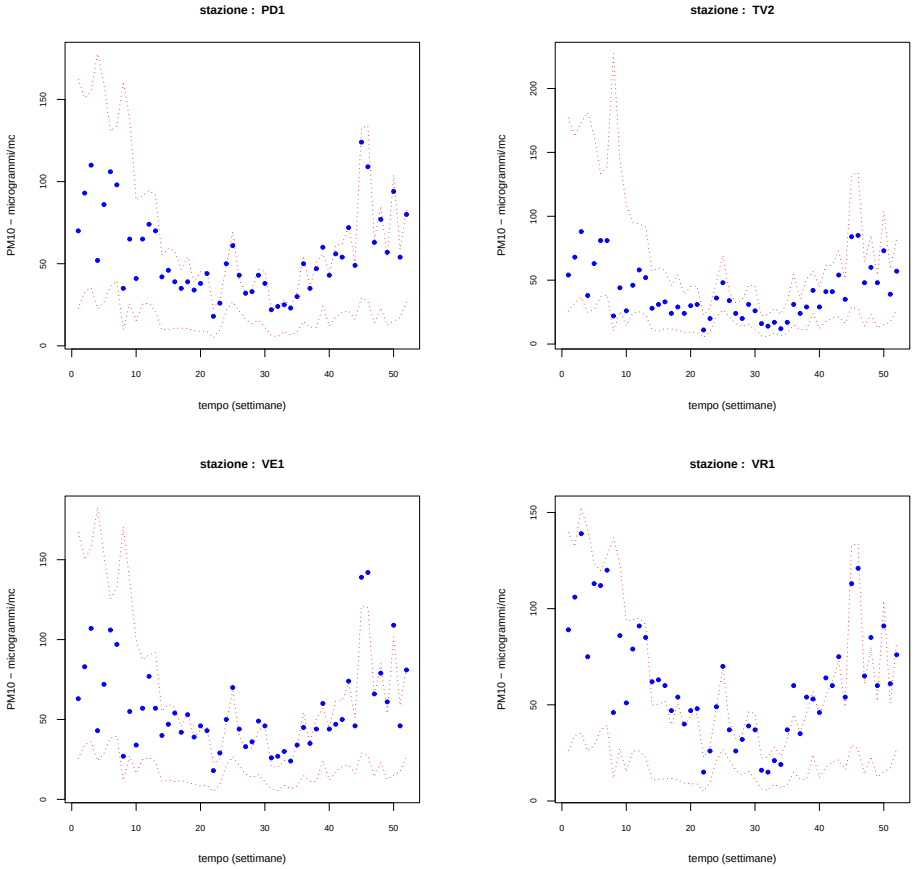


Figura 5.6: Andamento, in quattro stazioni, dei valori osservati e della banda di variabilità $\approx 95\%$

Per valori di t bassi, ossia le prime settimane dell'anno, la banda di variabilità risulta più ampia e questo riflette la minore informazione dovuta ai valori mancati presenti nelle prime 5 stazioni.

Nella stazione VE1, collocata a Mestre in Via Circonvallazione - zona molto trafficata - il modello non dà buoni risultati, in effetti sono diversi i valori osservati che risultano elevati di quanto non venga previsto dalla simulazione. Una possibile spiegazione può essere la vicinanza della stessa a due stazioni posizionate in contesti molto diversi (Mestre presso il parco Bissuola e a Venezia a Sacca Fisola) che non consentono al modello di catturare precisamente, pur senza l'assunto della stazionarietà, il fenomeno. Un'altra

possibile interpretazione può essere dovuta al fatto che con questo approccio non è stata utilizzata l'informazione riguardante la tipologia di *background* in cui è posizionata la stazione di rilevamento dal momento che, come visto nel modello presentato nel capitolo precedente, questa variabile può spiegare livelli diversi della concentrazione di inquinante.

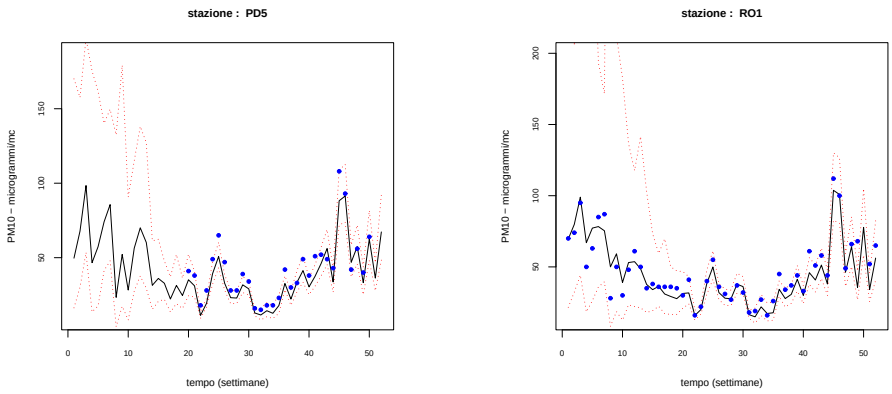
Convalida del modello sui 7 siti non considerati nell'analisi

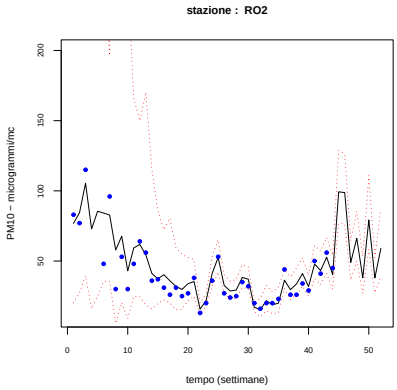
Si è accennato all'inizio dell'analisi come l'approccio per la costruzione del modello non abbia consentito di considerare i dati delle 7 stazioni che hanno valori mancanti lungo la serie o alla fine.

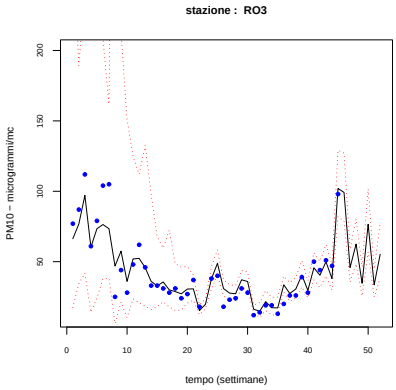
Si è ritenuto quindi di utilizzare il modello ottenuto con l'intento di valutarne l'efficacia e la possibilità di utilizzo ai fini della previsione e/o interpolazione dei dati mancanti.

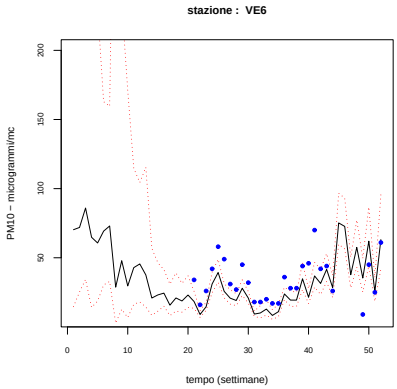
Per queste 7 stazioni sono stati previsti i dati per tutte le 52 settimane, sulla base delle 20 stazioni utilizzate per la costruzione del modello, mediante simulazione con 1.000 replicazioni.

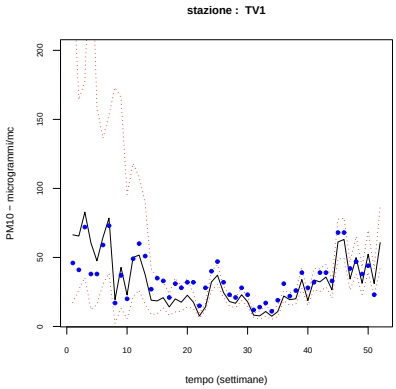
Nei grafici viene presentato l'andamento del valore medio mediante la linea continua, delle bande di variabilità approssimate al 95%, determinate tramite i quantili empirici al livello 0.025 e 0.975, nonchè i valori osservati.

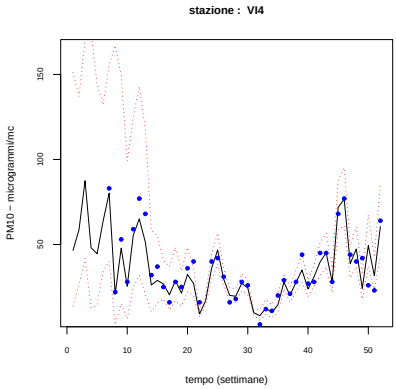



stazione : RO2
PM10 – microgrammi/mc
tempo (settimane)


stazione : RO3
PM10 – microgrammi/mc
tempo (settimane)


stazione : VE6
PM10 – microgrammi/mc
tempo (settimane)


stazione : TV1
PM10 – microgrammi/mc
tempo (settimane)


stazione : VI4
PM10 – microgrammi/mc
tempo (settimane)

Capitolo 6

Conclusioni

La modellazione e la previsione spaziale e temporale rendono evidente come la situazione del livello di concentrazione di PM_{10} presenta valori elevati, rispetti ai limiti normativi, nella regione Veneto e soprattutto nella parte relativa alla pianura.

La rappresentazione delle mappe condotta con le diverse tecniche, modelli e livelli di complessità, pur con i limiti riportati nei vari capitoli, rende evidente come l'area geografica comprendente le tre città di Padova, Vicenza e Verona presenti una sostanziale uniformità rispetto al fenomeno esaminato. Le aree geografiche disposte verso il mare Adriatico - provincia di Venezia - e quelle verso sud -Rovigo- e verso nord -Treviso e Belluno- presentano una concentrazione di inquinante meno elevata.

A prescindere dagli aspetti statistici e di modellazione, questo comportamento può trovare spiegazione nel fatto che l'area centrale della pianura è un continuum di terreni edificati, con una maggior concentrazione di attività e insediamenti produttivi, di persone e di conseguenza di agenti inquinanti. La minor concentrazione nella provincia di Venezia potrebbe essere determinata dalla favorevole influenza dei venti, che hanno una capacità di dispersione dell'inquinante a danno delle zone più interne. La provincia di Rovigo presenta una tipologia di utilizzo del territorio meno intensiva rispetto alle altre aree.

Risulta meno attendibile la previsione nella parte montana delle province di Verona e Vicenza, come la provincia di Belluno.

Valutando quanto visto nel corso dell'analisi appare evidente che la tipologia di *background* influenza il livello di PM_{10}; quale poi sia il livello di fondo potrebbe essere oggetto di analisi ulteriori.

La previsione effettuata con l'approccio di Le e Zidek, con la determinazione della struttura di covarianza spaziale non stazionaria, meglio interpreta i valori osservati, anche se l'approccio risulta più complesso e meno immediato degli altri due.

Un positivo riscontro del modello e dell'approccio sembra consentire, oltre la stima della previsione del livello di inquinante, anche la possibilità di stimare i valori medi settimanali di PM_{10} nella stazioni in cui sono presenti valori mancanti all'interno o alla fine della serie storica.

Per superare i limiti di questa prima analisi, le linee di approfondimento e di sviluppo per il futuro potrebbero riguardare:

- lo studio di una eventuale componente di fondo, dovuta al *background*;
- la modellazione della componente temporale sui dati giornalieri;
- il trattamento dei valori mancanti;
- l'analisi per l'ottimizzazione, ai fini previsivi, oltre che normativi, della localizzazione delle stazioni nel territorio.

Appendice A

Analisi esplorativa

A.1 Rete di monitoraggio ARPAV degli inquinanti atmosferici

Codice	Località	Tipologia background	Tipologia rilevamento	Indirizzo
BL1	Belluno	BU	An	Parco Città di Bologna
BL2	Feltre	BS	Af	Via Colombo
BL3	Pieve d'Alpago	BS	An	
PD1	PD-Arcella	TU	Af	Via T. Aspetti
PD2	PD-Mandria	BU	Af	Via Cà Rasi
PD3	PD-Granze	BS-IND	Af	Via Terranegra
PD4	Este	TU	An	Via Versori
PD5	Cittadella	TU	An	Via Pilastroni
RO1	Rovigo	TU	Af	Largo Martiri
RO2	Castelnovo Bariano	BS	M	Via Emilia
RO3	Rovigo Borsea	BU	M	Via Grotto
RO4	Porto Tolle	BS	Af	Via Campion
TV1	Conegliano	BU	M	Via Kennedy
TV2	Treviso	BU	Af	Via Lancieri di Novara
TV3	Mansué	BR	An	Via Cornaré
VE1	Venezia Mestre	TU	M	Via Circonvallazione
VE2	Venezia	BU	Af	Sacca Fisola
VE3	Venezia Mestre	BU	M	Parco Bissuola
VE4	Chioggia	BU	An	Orti ovest
VE5	San Donà di Piave	BU	An	Via Turati
VE6	Concordia Sagittaria	BR	M	
VR1	Verona	TU	Af	Corso Milano
VR2	Verona Cason	BR	Af	Via Ferrarin
VI1	Vicenza	BU	M	Via N. Tommaseo
VI2	Vicenza	TU	M	Via Spalato
VI3	Schio	BU	M	Via Vecellio
VI4	Bassano del Grappa	BU	M	Via Muhlacker

Tabella A.1: Localizzazione delle stazioni - Anno 2006

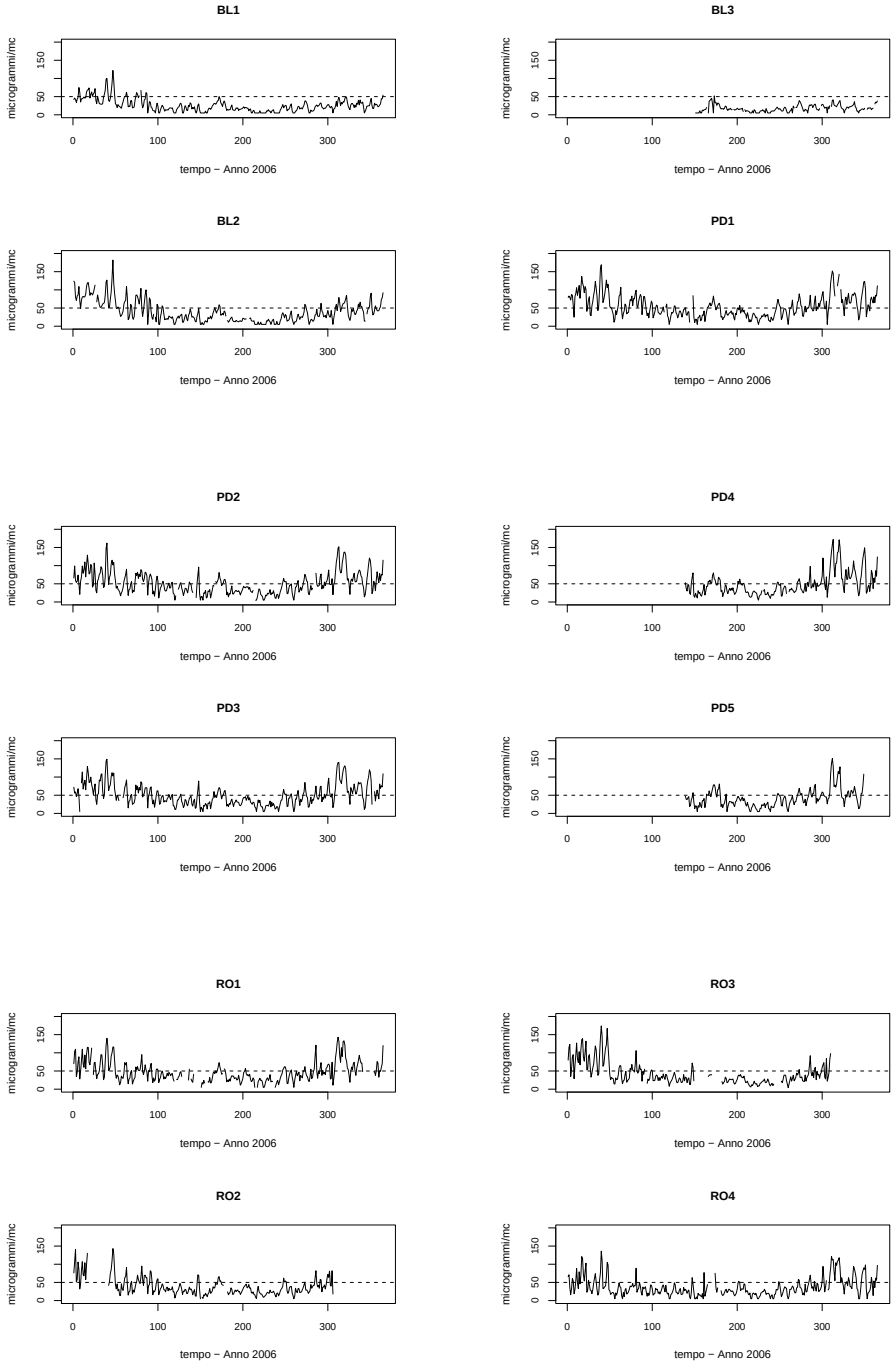
Codice	Località	Longitudine	Latitudine	Altitudine
BL1	Belluno	1748542,000	5114945,000	401
BL2	Feltre	1724841,765	5101513,263	263
BL3	Pieve d'Alpago	1759413,000	5117620,000	610
PD1	PD-Arcella	1725961,605	5034672,101	12
PD2	PD-Mandria	1722487,212	5028105,894	13
PD3	PD-Granze	1730222,600	5029118,510	13
PD4	Este	1707421,344	5011343,759	11
PD5	Cittadella	1717334,167	5058641,760	46
RO1	Rovigo	1719049,807	4994940,300	7
RO2	Castelnovo Bariano	1680516,554	4988762,259	12
RO3	Rovigo Borsea	1719789,444	4991070,397	3
RO4	Porto Tolle	1762948,094	4982810,018	1
TV1	Conegliano	1756609,839	5087129,234	72
TV2	Treviso	1752210,928	5062705,386	15
TV3	Mansué	1772628,710	5081943,010	14
VE1	Venezia Mestre	1752889,228	5043228,151	1
VE2	Venezia	1759183,852	5035901,051	1
VE3	VE-Mestre	1754826,108	5043492,095	1
VE4	Chioggia	1757576,761	5010430,857	2
VE5	San Donà di Piave	1779895,496	5059132,165	3
VE6	Concordia Sagittaria	1794818,211	5067021,208	5
VR1	Verona	1653542,500	5034215,000	62
VR2	Verona Cason	1649457,500	5036015,000	91
VI1	Vicenza	1698156,838	5048282,415	36
VI2	Vicenza	1699963,137	5046834,377	35
VI3	Schio	1684307,204	5064971,199	190
VI4	Bassano del Grappa	1712780,000	5071035,000	114

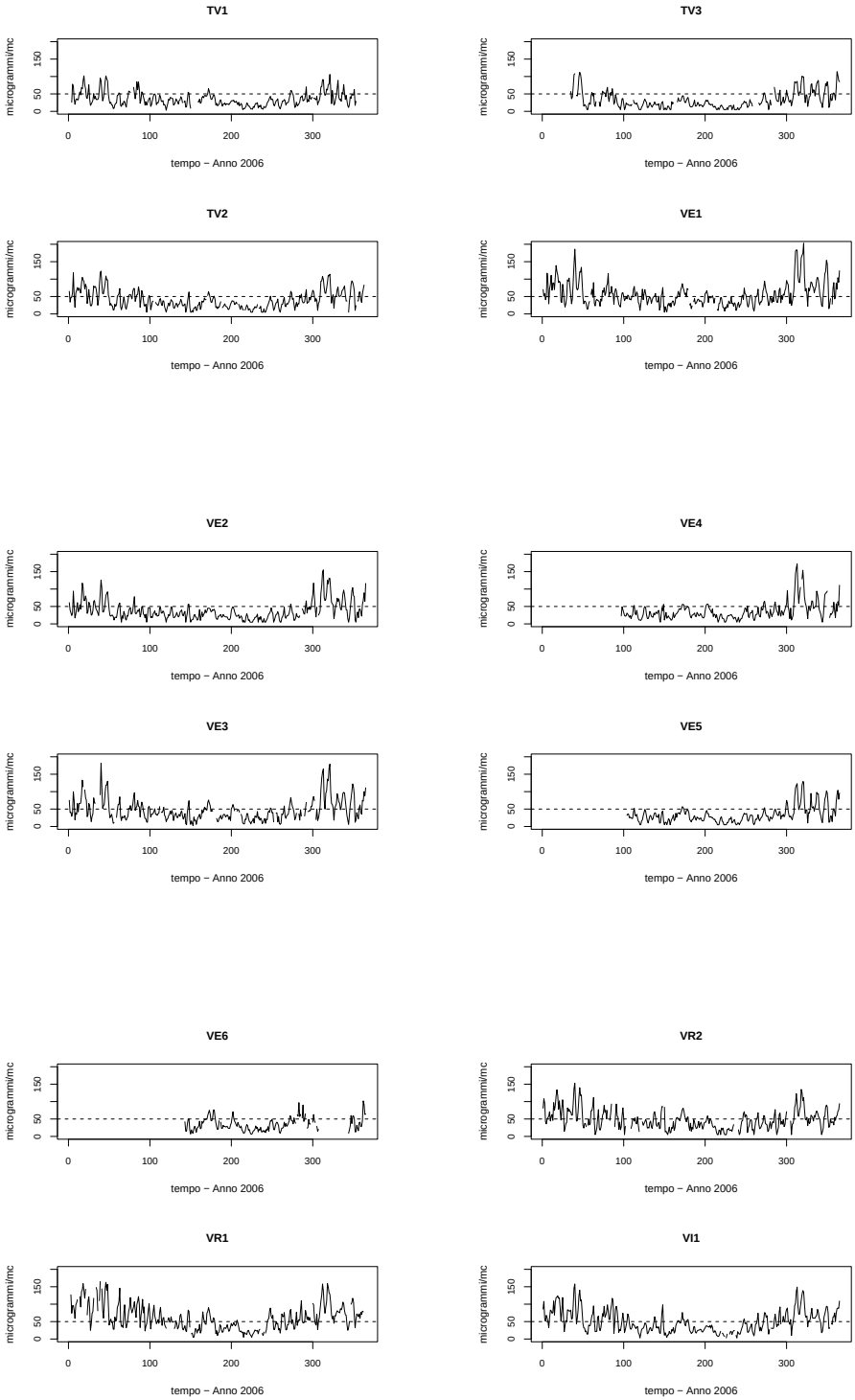
Tabella A.2: Coordinate geografiche e altitudine delle stazioni - Anno 2006

	BL1	BL2	PD1	PD2	PD3	RO4	TV1	TV2	VE1	VE2	VE3	VR1	VR2	VI1	VI2	VI3
BL1	100															
BL2	92	100														
PD1	79	84	100													
PD2	86	89	94	100												
PD3	86	87	95	94	100											
RO4	66	65	80	74	81	100										
TV1	90	92	90	94	89	67	100									
TV2	85	90	94	95	93	74	94	100								
VE1	79	83	93	91	91	82	89	93	100							
VE2	70	72	89	85	88	85	80	88	87	100						
VE3	66	71	84	80	77	79	80	82	92	82	100					
VR1	71	75	89	90	89	75	83	84	83	81	74	100				
VR2	69	61	76	76	81	77	67	75	76	79	68	78	100			
VI1	84	89	91	94	89	70	96	94	87	82	78	87	68	100		
VI2	76	81	87	89	83	74	89	86	84	77	77	88	68	92	100	
VI3	88	90	87	92	87	64	97	91	85	76	75	79	66	96	86	100

Tabella A.3: Coefficienti di correlazione (x 100) delle stazioni con più del 90% dei dati

A.2 Serie storiche delle medie giornaliere della concentrazione di PM_{10}




TV1
TV3
TV2
VE1
VE2
VE4
VE3
VE5
VE6
VR2
VR1
VI1
microgrammi/mc
0
50
150
0
100
200
300
tempo – Anno 2006

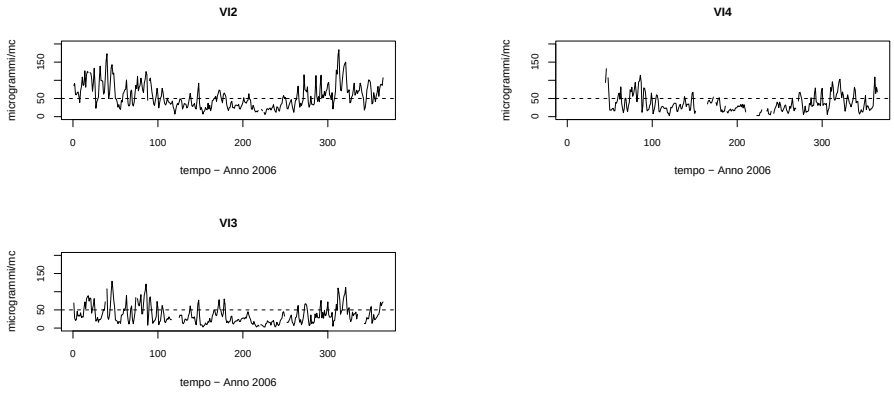
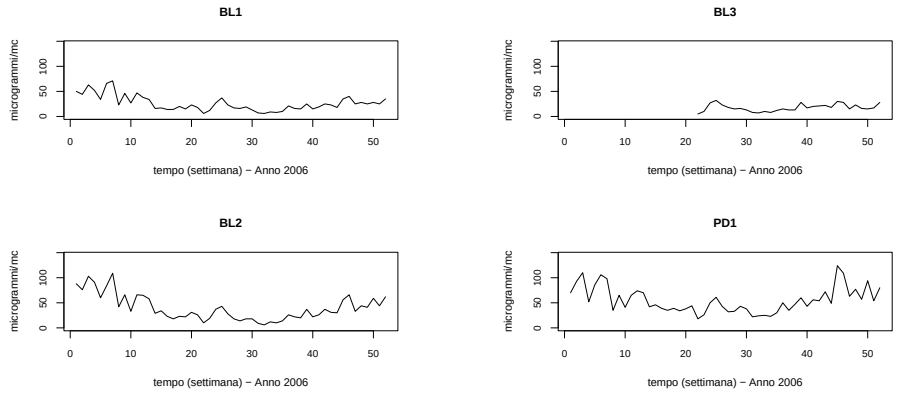


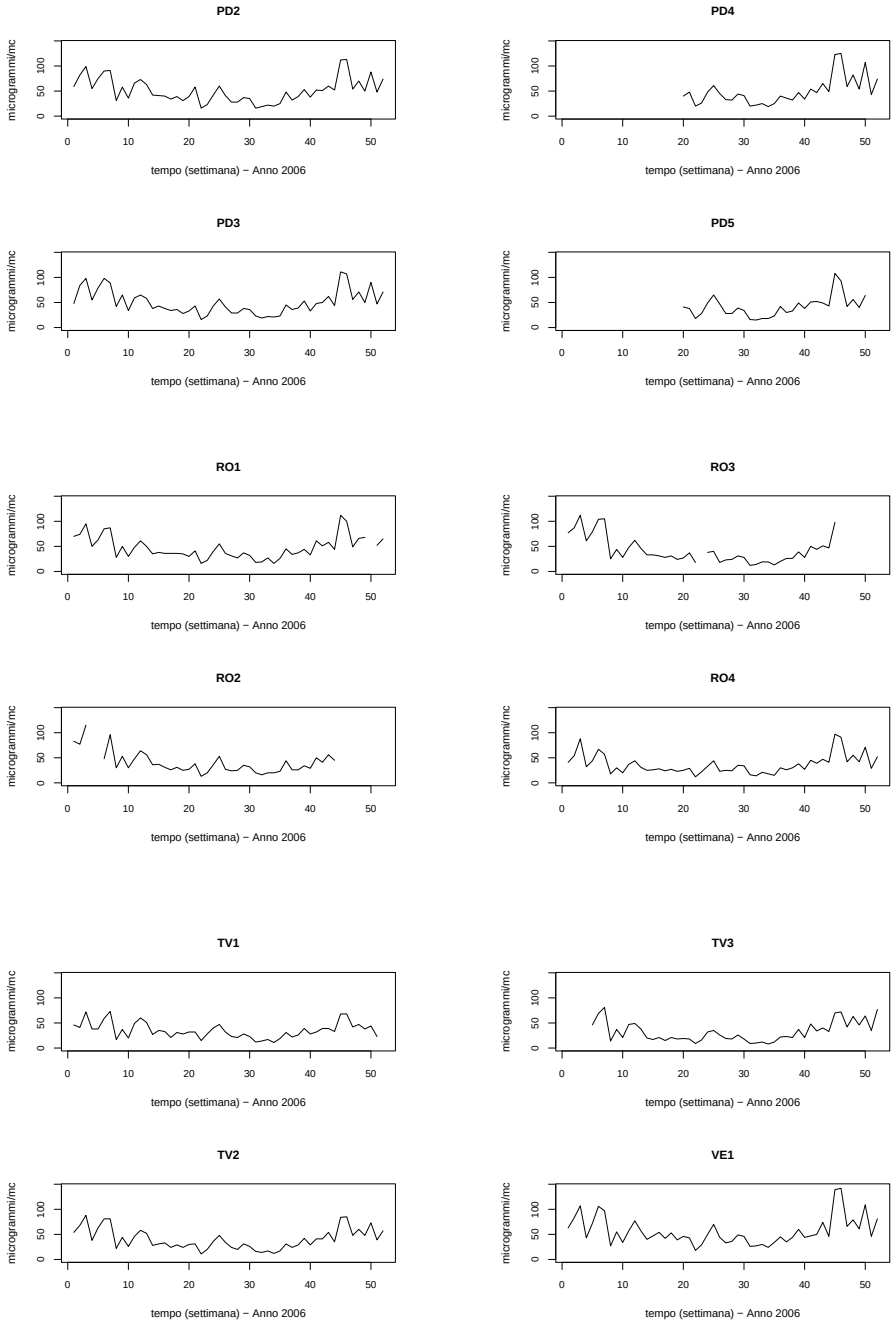
Figura A.1: Serie storiche dei valori medi giornalieri del PM_{10} in ogni stazione - anno 2006

Appendice B

Analisi del capitolo 4

B.1 Serie storiche delle medie settimanali della concentrazione di PM_{10}




PD2
PD4
PD3
PD5
RO1
RO3
RO2
RO4
TV1
TV3
TV2
VE1
microgrammi/mc
tempo (settimana) − Anno 2006

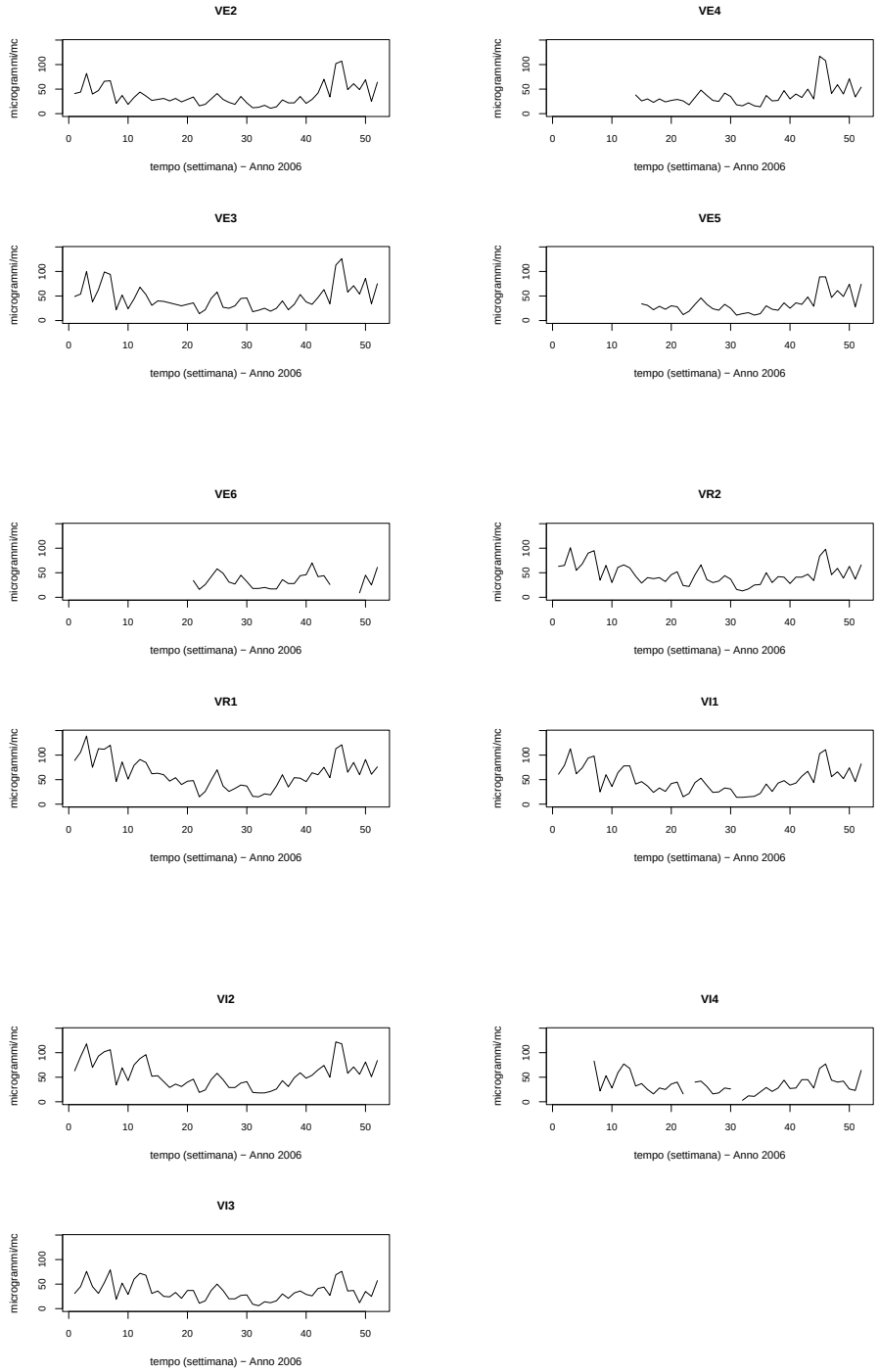


Figura B.1: Serie storiche dei valori medi settimanali del PM_{10} per ogni stazione - anno 2006

B.2 Modello additivo per la stima parametri della componente di trend

B.2.1 Analisi dei residui del modello

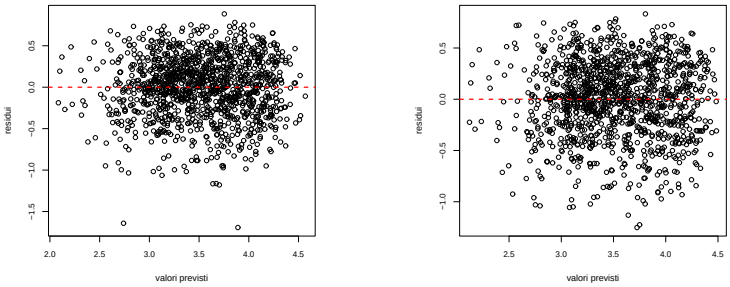


Figura B.2: Residui prima (a sinistra) e dopo (a destra) l'eliminazione dei due valori *outliers*

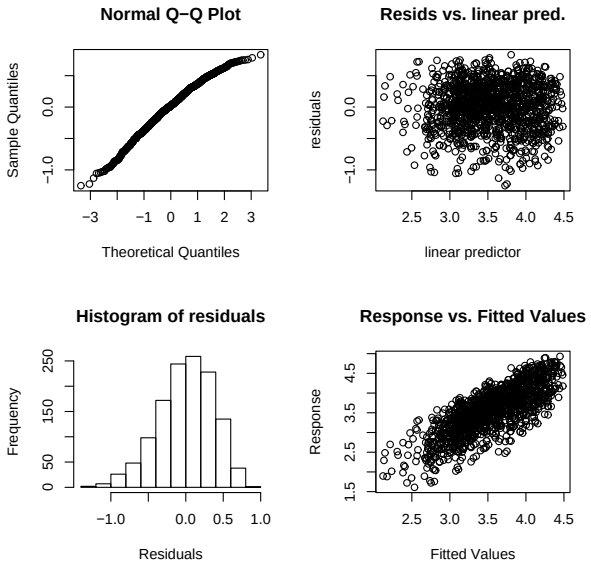


Figura B.3: Grafici per controllo andamento dei residui del modello GAM stimato

B.3 Analisi della correlazione spaziale dei residui

In questa sezione vengono presentati i grafici relativi ai (semi)variogrammi empirici, alla presenza di anisotropia, alla stima ottenuta tramite *Kriging* ordinario sui residui del modello GAM e alla convalida del modello di correlazione nello spazio mediante metodo grafico.

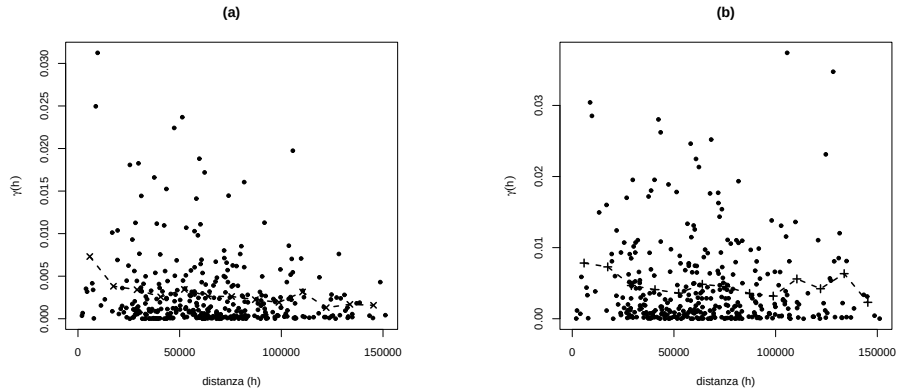


Figura B.4: Semivariogrammi nuvola ed empirici dei residui aggregati in ogni stazione tramite: (a) media; (b) mediana

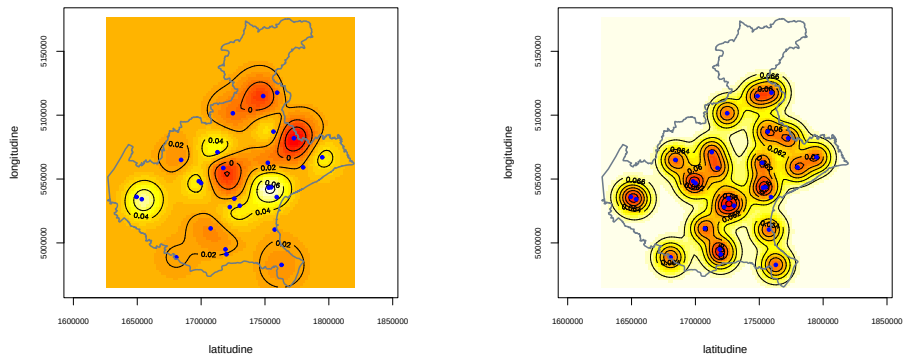


Figura B.5: Previsione (a sx) e Standard Error (a dx) sui residui mediante *kriging* ordinario

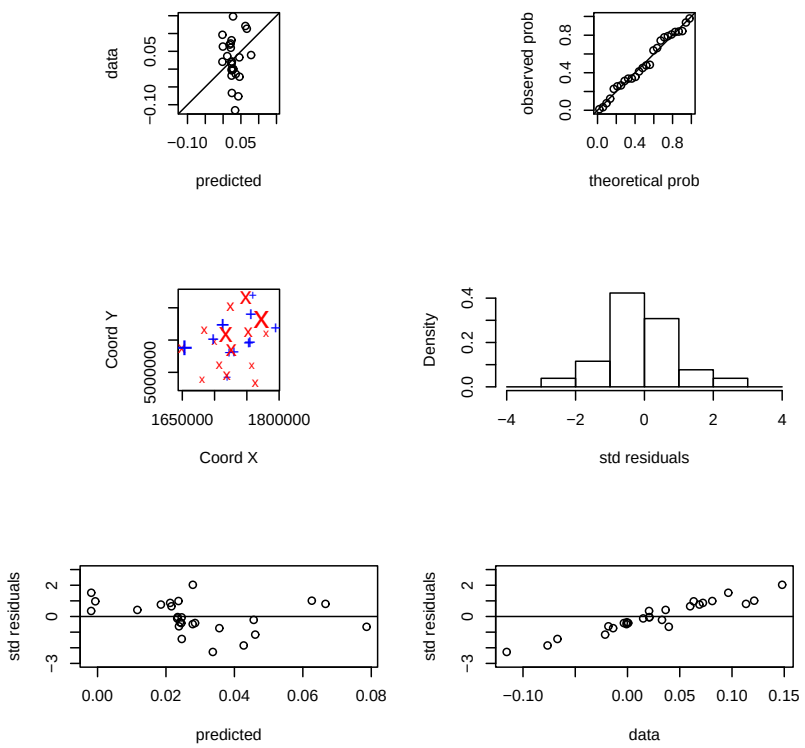


Figura B.6: Grafici per la validazione del modello di covarianza spaziale di tipo gaussiano - stime wls

B.4 Previsione spazio-temporale del fenomeno

Stima dei valori di altitudine sulla griglia di valori usati per la previsione spazio-temporale del fenomeno
Stima del modello

```
Model: A ~ s(x, y, k = 20)
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   75.333      0.992   75.94 1.55e-11 ***
---
Approximate significance of smooth terms:
         edf Est.rank    F  p-value
s(x,y) 18.95       19 1032 2.54e-10 ***
---
R-sq.(adj) =  0.999   Deviance explained =  100%
GCV score = 101.72   Scale est. = 26.571    n = 27
```

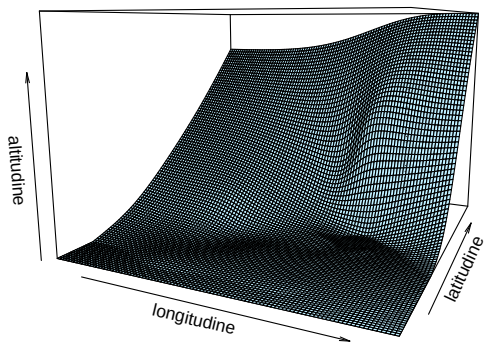


Figura B.7: Andamento della variabile altitudine stimata con modello GAM

Appendice C

Analisi del capitolo 5

C.1 La libreria utilizzata

Procedura e funzioni per l'interpolazione spaziale

Al fine di poter implementare l'analisi dei dati relativi a fenomeni spazio-temporali, secondo le ipotesi e la struttura proposta da Le e Zidek, gli stessi hanno sviluppato un pacchetto software - disponibile per i sistemi operativi maggiormente utilizzati - reperibile all'indirizzo `http://enviRo.stat.ubc.ca` comprendente un insieme di funzioni `R` e di librerie dinamiche.

In questa Appendice vengono presentate le funzioni principali utilizzate per l'analisi dei dati come descritto al Capitolo 5.

C.1.1 La stima iniziale degli iperparametri

```
staircase.EM(data, p=1, block=NULL, covariate=NULL, B0=NULL,
             init=NULL, verbose=F, maxit=20, tol=1e-6)
```

Questa funzione consente, sulla base delle ipotesi sulla distribuzione (vedi 5.17), di ottenere le stime dei valori degli iperparametri $\mathcal{H}_g$ relativi ai dati osservati, tramite la rete di monitoraggio, del fenomeno oggetto di analisi.
Argomenti in input:
`data`: matrice dei dati, raggruppata in blocchi in cui per ognuno si hanno stazioni con lo stesso numero di valori mancati all'inizio della serie. I blocchi sono organizzati in ordine decrescente rispetto al numero di valori mancanti. Ogni colonna della matrice rappresenta le osservazioni relative ad una

stazione di rilevamento; le righe indicano le osservazioni relative ai momenti temporali
`p`: numero di variabili rilevate in ogni stazione (nel caso multivariato indica il numero di diversi agenti inquinanti rilevati in ogni stazione)
`block`: vettore contenente il numero di stazioni presenti in ogni blocco; se non presente, la funzione procede automaticamente alla determinazione dei blocchi sulla base dei valori mancanti relativi alle osservazioni di ogni stazione
`covariate`: matrice contenente le variabili covariate, per la dimensione temporale, categorizzati come fattori
`B0`: valori per β_0 se noti (e non stimati, di conseguenza, dai dati osservati).
`init`: valori iniziali per gli iperparametri
`verbose`: flag per la visualizzazione dei risultati di ogni iterazione
`maxit`: numero massimo delle iterazioni per l'algoritmo
`tol`: valore che determina la condizione di convergenza dell'algoritmo.

Dati in Output:
`block`: blocchi e numero di stazioni contenute in ognuno
`Delta`: gradi di libertà stimati per ognuno dei blocchi (list)
`Omega`: matrice di covarianza stimata per i vari inquinanti (nel caso multivariato)
`Lambda`: matrice di covarianza tra le stazioni appartenenti allo specifico blocco condizionata alle osservazioni presenti nelle stazioni con minor numero di dati NA
`Xi0`: coefficienti di regressione stimati tra le stazioni in ogni blocco
`Beta0`: coefficienti per le variabili covariate (assunti essere gli stessi tra le stazioni per ogni singolo inquinante)
`Finv`: matrice di scala associata ai parametri Beta0
`Hinv`: iperparametri stimati - inversa di Hj
`Psi`: matrice di covarianze (marginali) stimate tra tutte le stazioni
`covariate`: come gli argomenti in input.

Stime ottenute dalla funzione staircase.EM()

Λ_1	Λ_2	Λ_3	Λ_4	Λ_5
0.025	0.0072	0.013	0.049	0.041

Tabella C.1: Covarianza residua per ogni singola stazione presente nei primi 5 blocchi

	Λ_6							
	s1	s2	s3	s4	s5	s6	s7	s8
s1	10.80	9.82	7.94	9.13	7.70	8.71	8.90	8.75
s2	9.82	10.39	7.97	9.25	7.81	8.83	8.93	8.85
s3	7.94	7.97	9.54	9.83	8.95	9.92	10.30	10.56
s4	9.13	9.25	9.83	11.11	9.72	10.73	10.73	11.04
s5	7.70	7.81	8.95	9.72	9.45	9.89	9.86	10.11
s6	8.71	8.83	9.92	10.73	9.89	12.17	11.31	11.57
s7	8.90	8.93	10.30	10.73	9.86	11.31	11.99	11.90
s8	8.75	8.85	10.56	11.04	10.11	11.57	11.90	12.76
s9	9.48	9.58	10.10	11.01	10.24	11.76	11.85	12.05
s10	9.97	10.01	11.70	12.32	11.33	13.03	13.17	14.06
s11	8.29	8.45	9.04	10.01	8.70	9.42	9.72	10.38
s12	9.42	9.34	9.69	10.98	9.59	10.76	10.33	10.95
s13	9.41	9.67	10.48	11.27	9.86	10.75	11.40	11.59
s14	7.56	7.99	9.18	9.77	8.73	9.37	9.78	10.00
s15	11.83	11.86	11.16	12.74	11.29	12.14	12.14	12.50

	Λ_6						
	s9	s10	s11	s12	s13	s14	s15
s1	9.48	9.97	8.29	9.42	9.41	7.56	11.83
s2	9.58	10.01	8.45	9.34	9.67	7.99	11.86
s3	10.10	11.70	9.04	9.69	10.48	9.18	11.16
s4	11.01	12.32	10.01	10.98	11.27	9.77	12.74
s5	10.24	11.33	8.70	9.59	9.86	8.73	11.29
s6	11.76	13.03	9.42	10.76	10.75	9.37	12.14
s7	11.85	13.17	9.72	10.33	11.40	9.78	12.14
s8	12.05	14.06	10.38	10.95	11.59	10.00	12.50
s9	14.07	13.89	9.53	11.02	11.55	9.88	12.92
s10	13.89	16.70	11.46	12.47	12.98	11.21	14.34
s11	9.53	11.46	10.38	10.20	10.35	8.83	11.97
s12	11.02	12.47	10.20	13.24	10.99	9.57	13.53
s13	11.55	12.98	10.35	10.99	12.69	10.87	13.49
s14	9.88	11.21	8.83	9.57	10.87	9.68	11.52
s15	12.92	14.34	11.97	13.53	13.49	11.52	17.95

Tabella C.2: Covarianza residua tra le 15 stazioni presenti nel blocco 6

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
s1	1.00	0.39	0.54	0.41	0.61	0.68	0.61	0.53	0.53	0.45
s2	0.39	1.00	0.84	0.79	0.82	0.71	0.74	0.88	0.88	0.88
s3	0.54	0.84	1.00	0.76	0.85	0.80	0.80	0.88	0.88	0.86
s4	0.41	0.79	0.76	1.00	0.67	0.62	0.64	0.80	0.77	0.79
s5	0.61	0.82	0.85	0.67	1.00	0.79	0.77	0.85	0.83	0.76
s6	0.68	0.71	0.80	0.62	0.79	1.00	0.93	0.78	0.83	0.76
s7	0.61	0.74	0.80	0.64	0.77	0.93	1.00	0.80	0.86	0.79
s8	0.53	0.88	0.88	0.80	0.85	0.78	0.80	1.00	0.95	0.94
s9	0.53	0.88	0.88	0.77	0.83	0.83	0.86	0.95	1.00	0.95
s10	0.45	0.88	0.86	0.79	0.76	0.76	0.79	0.94	0.95	1.00
s11	0.49	0.89	0.89	0.79	0.86	0.76	0.79	0.92	0.92	0.92
s12	0.52	0.89	0.89	0.76	0.88	0.78	0.80	0.96	0.93	0.93
s13	0.50	0.90	0.89	0.76	0.83	0.75	0.77	0.96	0.93	0.92
s14	0.47	0.87	0.89	0.77	0.82	0.77	0.79	0.87	0.88	0.89
s15	0.50	0.85	0.88	0.74	0.80	0.74	0.76	0.93	0.90	0.90
s16	0.52	0.82	0.79	0.66	0.76	0.78	0.81	0.91	0.93	0.88
s17	0.44	0.82	0.77	0.73	0.74	0.79	0.80	0.86	0.91	0.86
s18	0.53	0.84	0.86	0.72	0.83	0.80	0.84	0.95	0.95	0.90
s19	0.51	0.83	0.83	0.75	0.79	0.74	0.80	0.95	0.94	0.91
s20	0.60	0.78	0.80	0.67	0.75	0.85	0.87	0.85	0.90	0.87

	s11	s12	s13	s14	s15	s16	s17	s18	s19	s20
s1	0.49	0.52	0.50	0.47	0.50	0.52	0.44	0.53	0.51	0.60
s2	0.89	0.89	0.90	0.87	0.85	0.82	0.82	0.84	0.83	0.78
s3	0.89	0.89	0.89	0.89	0.88	0.79	0.77	0.86	0.83	0.80
s4	0.79	0.76	0.76	0.77	0.74	0.66	0.73	0.72	0.75	0.67
s5	0.86	0.88	0.83	0.82	0.80	0.76	0.74	0.83	0.79	0.75
s6	0.76	0.78	0.75	0.77	0.74	0.78	0.79	0.80	0.74	0.85
s7	0.79	0.80	0.77	0.79	0.76	0.81	0.80	0.84	0.80	0.87
s8	0.92	0.96	0.96	0.87	0.93	0.91	0.86	0.95	0.95	0.85
s9	0.92	0.93	0.93	0.88	0.90	0.93	0.91	0.95	0.94	0.90
s10	0.92	0.93	0.92	0.89	0.90	0.88	0.86	0.90	0.91	0.87
s11	1.00	0.94	0.93	0.90	0.91	0.84	0.85	0.86	0.86	0.82
s12	0.94	1.00	0.96	0.91	0.93	0.87	0.82	0.92	0.91	0.83
s13	0.93	0.96	1.00	0.90	0.96	0.90	0.84	0.91	0.90	0.83
s14	0.90	0.91	0.90	1.00	0.91	0.79	0.81	0.86	0.85	0.81
s15	0.91	0.93	0.96	0.91	1.00	0.87	0.84	0.89	0.88	0.83
s16	0.84	0.87	0.90	0.79	0.87	1.00	0.87	0.90	0.88	0.88
s17	0.85	0.82	0.84	0.81	0.84	0.87	1.00	0.85	0.84	0.88
s18	0.86	0.92	0.91	0.86	0.89	0.90	0.85	1.00	0.98	0.89
s19	0.86	0.91	0.90	0.85	0.88	0.88	0.84	0.98	1.00	0.87
s20	0.82	0.83	0.83	0.81	0.83	0.88	0.88	0.89	0.87	1.00

Tabella C.3: Correlazione marginale tra tutte le 20 stazioni

C.1.2 Stima della covarianza spaziale

Le funzioni descritte in seguito consentono di determinare la covarianza spaziale nei siti oggetto di previsione implementando il metodo non parametrico proposto da Sampson e Guttorp.

```
Falternate3(disp, coords, model=1., a0=0.1, t0=0.5,
            max.iter=50., max.fcal=100., alter.lim=50.,
            tol=1e-05, prt=0., dims=2., lambda=0.)
```

Funzione per la determinazione delle nuove coordinate nel D-space e stima del variogramma (esponenziale o gaussiano).

Argomenti in input

`disp`: matrice di dispersione spaziale tra i siti con i dati osservati

`coords`: matrice delle coordinate geografiche ($n \times 2$)

`model`: modello di variogramma 1-esponenziale, 2-gaussiano

`a0,t0`: stime iniziali dei parametri del variogramma

`max.iter, max.fcal`: parametri di controllo per l'utilizzo della funzione di ottimizzazione `nlmin()`

`lambda` : parametro di lisciamento - opzionale.

```
Ftransdraw(disp, Gcrds, MDScrds, gridstr, sta.names,
           lambda=0., lsq=F, eye, model=1., a0=0.1, t0=0.5)
```

Funzione interattiva che permette la selezione del valore opportuno del parametro di lisciamento λ per la *thin-plate spline* che determina le coordinate nel *D-space*.

Argomenti in input:

`disp`: matrice di dispersione spaziale tra i siti con i dati osservati

`Gcrds`: coordinate geografiche ($n \times 2$)

`MDScrds`: coordinate relative al *D-space*, determinate tramite la funzione `Falternate3()`

`gridstr`: griglia regolare sul piano dello spazio geografico.

`sinterp(coords, ncoords.sg.est, lambda)`

Funzione che mappa le coordinate dallo spazio geografico al nuovo spazio (*D-space*) utilizzando i risultati ottenuti mediante le funzioni `Falternate3()` e `Ftransdraw()`.
Argomenti in input:
`coords` : coordinate geografiche
`ncoords.sg.est` : coordinate relative al *D-space*
`lambda` : parametro di lisciamento della *spline*.

`corrfit(crds, Tspline, sg.fit, model = 1)`

Funzione per determinare la correlazione tra le localizzazioni in cui ci sono i dati osservati e quelle in cui effettuare la previsione.
Argomenti in input:
`crds` : coordinate sull'insieme delle localizzazioni su cui effettuare la previsione
`Tspline`: *thin-plate spline* stimata con il metodo SG
`sg.fit`: coordinate risultanti dall'applicazione del metodo SG
`model`: modello di variogramma; 1: esponenziale 2: gaussiano
Dati in output:
`cor`: matrice di correlazione tra le localizzazioni oggetto di previsione.

`seval(allcrds,Tspline.var)`

Funzione per determinare la varianza stimata per le localizzazioni, usando la *thin-plate spline*.
Argomenti in input:
`allcrds` : coordinate sull'insieme delle localizzazioni
`Tspline.var` : matrice di varianze e covarianze determinate tramite la funzione `sinterp()`.

C.1.3 Stima degli iperparametri della distribuzione di previsione

`staircase.hyper.est(emfit,covfit,u,p,g)`

Funzione che combina i risultati ottenuti tramite `staircase.EM()` e i risultati ottenuti tramite il metodo di Sampson e Guttorp, per ottenere la stima degli iperparametri associati a tutte le localizzazioni in cui effettuare la previsione.

Argomenti in input:

`emfit` : output della funzione `staircase.EM()`

`covfit` : matrice di covarianza tra tutte le localizzazioni

`u`: numero delle localizzazioni su cui effettuare la previsione

`p`: dimensione della variabile risposta (nel caso multivariato)

`g`: numero di stazioni in cui sono stati osservati i dati.

C.1.4 Funzione per la simulazione dei valori della distribuzione di previsione

`pred.dist.simul(hyperest, tpt, include.obs = T, N =1)`

Questa funzione consente di ottenere N replicazioni dalla distribuzione di previsione nelle n localizzazioni, in uno specifico valore della variabile tempo t.

Argomenti in input:

`hyperest`: output ottenuto dalla funzione `staircase.hyper.est()`, contenente le stime degli iperparametri

`tpt`: valore relativo alla variabile relativa al tempo t

`include.obs`: Nel caso 'TRUE', vengono restituiti i valori osservati al tempo t

`N`: numero di valori simulati.

Dati in output:

Una matrice con N righe; contenente il numero di valori di previsione N per ogni localizzazione

Appendice D

Distribuzioni di probabilità e Complementi

D.1 Distribuzione *Matric-t* multivariata

Distribuzione *t-multivariata* : Un vettore casuale p-dimensionale X ha una distribuzione t-multivariata con ν gradi di libertà se la sua distribuzione ha la forma

$$f(X) = \frac{\Gamma(\frac{p+\nu}{2})\sqrt{|A|}}{\Gamma(\nu/2)\sqrt{2\pi p}} \times \left[1 + \frac{1}{\nu}(X-\mu)^T A(X-\mu)\right]^{-(p+\nu)/2}$$

per ogni vettore μ e A matrice definita positiva. La distribuzione si denota con $X \sim t_p(\mu, A, \nu)$ in cui A viene chiamata *matrice di precisione* e si ha

$$E(X) = \mu \text{ e } Cov(X) = \frac{\nu}{\nu-2}A$$

Distribuzione *Matric-t* : Una matrice aleatoria X di dimensioni $n \times m$ ha una distribuzione *matric-t* con δ gradi di libertà se la sua funzione di densità ha la forma

$$f(X) \propto |A|^{-\frac{m}{2}}\ |B|^{-\frac{n}{2}}|I_n + \delta^{-1}[A^{-1}(X-\mu)][(X-\mu)B^{-1}]^T|^{-\frac{\delta+n+m-1}{2}}$$

per A e B matrici definite positive di dimensioni rispettivamente $n \times n$ e

$m \times m$ e μ è con una matrice di dimensione $n \times m$. La costante di normalizzazione è pari a

$$K = (\delta)^{-(mn)/2} \frac{\Gamma_{n+m}[(\delta+n+m-1)/2]}{\Gamma_n[(\delta+n-1)/2]\ \Gamma_m[(\delta+m-1)/2]},$$

dove $\Gamma_p(t) = \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma[t-(i-1)/2]$ è la funzione gamma multivariata. La distribuzione si denota con $X \sim t_{n\times m}(\mu, A \otimes B, \delta)$ e gode delle seguenti proprietà:
- $E(X) = \mu$
- $X \sim t_{n\times m}(\mu, A \otimes B, \delta)$ se e solo se $X^{'} \sim t_{m\times n}(\mu^{'}, B \otimes A, \delta)$
- Se $n = 1$ e $A = 1$, X ha una distribuzione t multivariata, ossia $X \sim t_m(\mu, B, \delta)$
- Se $m = 1$ e $B = 1$, X ha una distribuzione t multivariata, ossia $X \sim t_n(\mu, A, \delta)$

D.2 Distribuzione di Wishart e Wishart inversa

Distribuzione di *Wishart* : Una matrice definita positiva S di dimensioni $p \times p$ ha una distribuzione di Wishart con m gradi di libertà se la sua funzione di densità ha la forma

$$f(S) = \left[2^{(mp)/2}\Gamma_p(m/2)\right]^{-1} |A|^{-m/2}|S|^{(m-p-1)/2}\ e^{tr(A^{-1}S)/2}$$

con A matrice definita positiva e dove Γ_p è la funzione gamma multivariata vista sopra. La distribuzione si denota con $S \sim W_p(A, m)$.

Distribuzione di *Wishart inversa* : Una matrice definita positiva Σ di dimensioni $p \times p$ ha una distribuzione di Wishart inversa con δ gradi di libertà se la sua funzione di densità ha la forma

$$f(\Sigma) = \left[2^{(mp)/2}\Gamma_p(m/2)\right]^{-1} |\Psi|^{-\delta/2}|\Sigma|^{-(\delta+p+1)/2}\ e^{tr\Sigma^{-1}\Psi/2}$$

con Ψ matrice definita positiva.
La distribuzione si denota con $\Sigma \sim W_p^{-1}(\Psi, \delta)$ e possiede le seguenti pro-

prietà:
- $Y \sim W_p^{-1}(\Psi, \delta)$ se e solo se $Z = Y^{-1} \sim W_p(\Psi^{-1}, \delta)$
- Se $Z \sim W_p(\Sigma, \delta)$ allora $E(Z) = \delta\,\Sigma$ e $E(Z^{-1}) = \frac{\Sigma^{-1}}{\delta - p - 1}$ per $\delta - p - 1 > 0$
- Se $Y \sim W_p^{-1}(\Psi, \delta)$ allora $E(Y) = \frac{\Psi}{\delta - p - 1}$ e $E(Y^{-1}) = \delta\,\Psi^{-1}$

D.3 Distribuzione di Wishart inversa generalizzata

Sia Σ una matrice positiva di dimensioni $g \times g$ avente una struttura a k-blocchi per cui possa essere scritta come

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \cdots & \Sigma_{1,k} \\ \vdots & \ldots & \vdots \\ \Sigma_{k,1} & \cdots & \Sigma_{k,k} \end{pmatrix}$$

con $\Sigma_{i,j}$ avente dimensioni $g_i \times g_j$ e con $g = g_1 + \ldots + g_k$.

Si denota con $\Sigma^{[j,\cdots,k]}$ la sottomatrice corrispondente al j-esimo dei k blocchi

$$\Sigma^{[j,\cdots,k]} = \begin{pmatrix} \Sigma_{j,j} & \cdots & \Sigma_{j,k} \\ \vdots & \ldots & \vdots \\ \Sigma_{k,j} & \cdots & \Sigma_{k,k} \end{pmatrix}$$

Siano $\Sigma^{[j(j+1)]} = \left(\Sigma_{j,j+1}, \ldots, \Sigma_{j,k}\right)$ e $\Sigma^{[(j+1)j]} = \left(\Sigma_{j+1,i}, \ldots, \Sigma_{k,j}\right)$

Si denota con Ψ una matrice definita positiva, di dimensioni $g \times g$, avente la stessa struttura a blocchi, e in maniera analoga a quanto visto sopra, $\Psi^{[j,\cdots,k]}$ denota la sottomatrice. Sia $\delta = (\delta_j, \cdots, \delta_k)$ un vettore k-dimensionale a valori positivi e si denota $\delta^{j,\cdots,k} = (\delta_j, \cdots, \delta_k)$.

Σ si dice avere distribuzione di Wishart inversa generalizzata, denotata con $GIW(\Psi, \delta)$ se

$$\begin{cases} \Sigma^{[2,\cdots,k]} \sim GIW(\Psi^{[2,\cdots,k]}, \delta^{[2,\cdots,k]}) \\ \\ \Gamma_1 \sim IW(\Psi_1, \delta_1) \\ \\ \tau_1 |\, \Gamma_1 \sim N(\tau_{01}, H_1 \otimes \Gamma_1) \end{cases}$$

La distribuzione GIW viene definita ricorsivamente per cui all'ultimo step si ha

$\Gamma_k \sim IW(\Psi_k, \delta_k)$ con $\Gamma_k = \Sigma_{k,k}$ e $\Psi_k = \Psi_{k,k}$

D.4 Decomposizione di Bartlett

Sia la matrice di covarianza Σ rappresentabile mediante una decomposizione a blocchi, quindi

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

La matrice Σ può essere decomposta come $\Sigma = T\Delta T^T$ dove

$$\Delta = \begin{pmatrix} \Sigma_{1|2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \quad \text{e} \quad T = \begin{pmatrix} I & \tau \\ 0 & I \end{pmatrix}$$

in cui $\Sigma_{1|2} \equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ e $\tau \equiv \Sigma_{22}^{-1}\Sigma_{21}$

Tramite questa decomposizione, comunemente conosciuta come la decomposizione 1-1 di Bartlett, la matrice Σ può essere scritta come

$$\Sigma = \begin{pmatrix} \Sigma_{1|2} + \tau\Sigma_{22}\tau^T & \tau\Sigma_{22} \\ \Sigma_{22}\tau^T & \Sigma_{22} \end{pmatrix}$$

Elenco delle tabelle

Elenco delle figure

Bibliografia

[1] Sito Web ARPAV, *http://www.arpa.veneto.it*.

[2] A. Azzalini and Scarpa B., *Analisi dei dati e* D*ata* M*ining*, Springer, Milano, 2004.

[3] N.A.C. Cressie, *Statistics for spatial data - revised edition*, John Wiley and Son, New York, 1993.

[4] M. Fuentes, *Spectral methods for non-stationary spatial process*, Biometrika **89** (2002), 197–210.

[5] F. Greco, *Hierarchical space-time modelling of* PM_{10} *pollution in the* E*milia*-R*omagna region*, GRASPA Working paper **22** (2005).

[6] T.J. Hastie and R.J. Tibshirani, *Generalized additive models*, Chapman and Hall, London, 1990.

[7] T.J. Hastie, R.J. Tibshirani, and J. Friedman, *The elements of statistical learning : data mining, inference and prediction*, Springer, New York, 2001.

[8] D.M. Hawkins and N.A.C. Cressie, *Robust kriging - a proposal*, Journal of the International Association for Mathematical Geology **17** (1984), 563–586.

[9] N.L. Le and J.V. Zidek, *Statistical* A*nalysis of* E*nvironmental* S*pace-*T*ime* P*rocesses*, Springer, 2006.

[10] K.V. Mardia, C. Goodall, E.J. Redfern, and F.J. Alonso, *The* K*riged* K*alman* F*ilter (with discussion)*, Test **7** (1998), 217–252.

[11] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate analisys*, Academic Press, London, 1979.

[12] G. Matheron, *Traité de géostatistique appliquée*, Editions Technip, Paris, 1962.

[13] L. Pace and A. Salvan, *Introduzione alla statistica* II - *inferenza, verosimiglianza, modelli*, Cedam, Padova, 2001.

[14] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2006.

[15] S.K. Sahu, A.E. Gelfand, and D.M. Holland, *Spatio-temporal modeling of fine particulate matter*, (2004).

[16] S.K. Sahu and K.V. Mardia, *A Bayesian Kriged-Kalman filter model for short-term forecasting of air pollution levels*, Journal of the Royal Statistical Society **Series C - 54** (2005), 223–244.

[17] P. Sampson and P. Guttorp, *Nonparametric estimation of nonstationary spatial covariance structure*, Journal of the American Statistical Association **87** (1992), 108–119.

[18] G. Shaddick and J. Wakefield, *Modelling daily multivariate pollutant at multiple sites*, Journal of the Royal Statistical Society **Series C - 51** (2002), 351–372.

[19] R.L. Smith, S. Kolenikov, and L.H. Cox, *Spatiotemporal modeling of* $PM_{2.5}$ *data with missing value*, J. Geophys. Res. **108(D4)** (2003).

[20] L. Sun, J.V. Zidek, N.D. Le, and H. Ozkaynak, *Interpolating Vancouver's Daily Ambient* PM_{10} *Fields*, Environmetrics **11** (2000), 651–663.

[21] J.M. Ver Hoef and R.D. Barry, *Constructing and fitting models for cokriging and multivariate spatial prediction*, Journal of Statistical Plannig and Inference **69** (1998), 275–294.

[22] H. Wakernagel, *Multivariate geostatistics : An introduction with applications*, Springer, Berlin, 2003.

[23] S.N. Wood, *Generalized additive models an introduction with R*, Chapman and Hall, 2006.