UNIVERSITÀ DEGLI STUDI DI TRENTO

Facoltà di Scienze Matematiche, Fisiche e Naturali

Scuola di Dottorato in Fisica
XXI ciclo

Tesi di Dottorato

La Distribuzione del gas Radon Indoor: Analisi con Moderne Tecniche Statistiche

Relatori:
prof. Antonio Miotello
dott. Luca Verdi

Dottorando:
Stefano Pegoretti

Anno Accademico 2007–2008

... se impari la strada a memoria,
non troverai certo un granché,
se invece smarrisci la rotta,
il mondo è li tutto per te ...

Il Viaggiatore > La musica dei poveri
Lorenzo Monguzzi_ Mercanti di Liquore

Indice

CAPITOLO 1

Introduzione e Contestualizzazione

Il radon è un gas nobile radioattivo naturalmente presente nei terreni, dai quali diffonde con relativa facilità mescolandosi con gli altri gas presenti in atmosfera e raggiungendo valori di concentrazione tipici pari a 10 $\text{Bq}\cdot\text{m}^{-3}$. Benché stime recenti affermino che esso contribuisca per il 50% alla dose media annuale di radiazione naturale di fondo cui un individuo è esposto, questi valori tipicamente presenti in atmosfera non sono preoccupanti per la salute.

Diversa invece la situazione in cui questo gas si accumula in ambienti chiusi[1]: si parla allora di Radon Indoor, per porre l'accento sulle caratteristiche peculiari di questo fenomeno. In tali situazioni, infatti, le concentrazioni possono raggiungere valori anche superiori ai 4000 $\text{Bq}\cdot\text{m}^{-3}$, e le ripercussioni sulla salute risultare importanti [cfr. §2.4]. Per questo, il fenomeno radon indoor e la sua analisi hanno assunto negli ultimi anni una forte rilevanza socio-sanitaria, soprattutto in relazione agli effetti cancerogeni di questo gas sulla salute umana: basti ricordare che il radon è inquadrato al secondo posto, dopo il fumo, come causa di insorgenza di tumore polmonare.

Tuttavia, come spesso accade per i problemi di tipo ambientale, l'importanza del fenomeno va di pari passo con la sua *complessità*, rendendo quindi difficile tanto una sua comprensione esaustiva e dettagliata quanto una sua efficace modellizzazione. Peculiarità del fenomeno radon indoor, rispetto alla situazione in cui questo gas possa disperdersi liberamente in atmosfera, è la stretta e complessa *interazione* con l'edificio nel quale il gas si può accumulare. Benché i principali meccanismi di diffusione siano noti e correttamente modellizzati [cfr. §2.2.1], questo purtroppo non si traduce in previsioni del valore di concentrazione misurato in ambienti chiusi altrettanto affidabili. Verosimilmente, i principali problemi che in incontrano in questo contesto sono relativi a:

- *individuazione* e riconoscimento delle variabili secondarie che possono influenzare il valore di concentrazione di radon misurato negli ambienti chiusi ↦ Quali e quante sono queste possibili variabili? Come le si possono identificare in maniera affidabile? Quali hanno una significativa influenza sul valore di concentrazione? Tale influenza si manifesta in maniera costante o dipende invece dal valore di concentrazione?

[1] Questi aspetti verranno trattati nel dettaglio nel paragrafo §2.2, a pagina 15.

- *interazione* tra le possibili variabili secondarie significative, che tipicamente non è semplicemente di tipo additivo e lineare, ma può risultare molto più complessa (ad esempio, interazioni di tipo moltiplicativo e/o in funzione del valore delle variabili stesse o di quella di riferimento); queste caratteristiche possono di conseguenza rendere anche molto difficile e articolata tanto l'individuazione delle variabili secondarie significative quanto una loro corretta modellizzazione (presenza di "fattori confondenti");
- *interazione* tra l'insieme delle covariate e la variabile di riferimento (concentrazione di radon indoor), che anche in questo caso può non essere (e spesso non è!) di tipo lineare e costante in funzione del valore della variabile stessa.

Da un lato si ha quindi la necessità di avere a disposizione un modello sufficientemente raffinato per l'individuazione delle cosiddette Radon Prone Areas, ovvero "zone a rischio" per la popolazione che vi risiede stabilmente; dall'altro, questo presuppone però una conoscenza approfondita del fenomeno, alla quale, in virtù dei problemi che sono stati brevemente descritti, non si ha, nella pratica, facile accesso ↦ Quali possono essere quindi gli strumenti utili per affrontare questo tipo di problema?

1.1 Motivazioni

Sia a livello nazionale che internazionale, i primi tentativi cui si è rivolta l'attenzione sono stati quelli relativi a modelli statistici lineari e a metodi di interpolazione basati sulla distanza euclidea (come quello basato sull'inverso del quadrato della distanza); altri tentativi sono stati fatti suddividendo il territorio in esame in celle o unità amministrative all'interno delle quali condure analisi statistiche convenzionali (ad esempio, calcolo di media o quartili) volte a una semplice classificazione. Tuttavia, i risultati che si possono ottenere con questo tipo di approcci non sono pienamente soddisfacenti.

Metodi e modelli più raffinati, su cui la comunità scientifica ha iniziato a porre l'attenzione in maniera significativa, sono quelli offerti dalla *geostatistica*, che fornisce gli strumenti teorici e pratici per una trattazione esplicita della componente spaziale del fenomeno di interesse: risulta così possibile modellizzare efficacemente la correlazione spaziale tra le misure di radon indoor in funzione non tanto della distanza euclidea, ma ricorrendo a una misura della distanza di tipo statistico. Tuttavia, ci sono situazioni, come quella relativa al territorio dell'Alto Adige cui il presente lavoro fa riferimento, per le quali la complessità del fenomeno è tale da rendere questo tipo di analisi di difficile applicazione, benché si possano ottenere risultati interessanti, come brevemente discusso nel paragrafo §1.2.

Infine, altro problema cui fare riferimento è quello relativo al *fine* cui le analisi saranno volte, ovvero se lo scopo principale sarà quello di una *descrizione* dei dati finalizzata a una loro comprensione più o meno approfondita o piuttosto quello di una *previsione* dei valori di concentrazione sul territorio (mappatura volta ad esempio all'individuazione delle citate radon prone areas). Va da sé che si dovranno individuare e valutare strumenti differenti in funzione dell'ambito al quale verranno applicati. Uno schema di massima che descrive sinteticamente i principali soggetti e le loro possibili interazioni nel contesto appena descritto è riportato in figura 1.1.

> *In un contesto generale i cui contorni risultano ancora non ben definiti e i cui principali problemi non hanno ancora trovato soluzioni certe o ricette "standard", il lavoro descritto in questa tesi vuole presentarsi come il tentativo di testare ed esplorare approcci differenti e diversificati — sia in relazione allo scopo dell'analisi, sia in relazione*

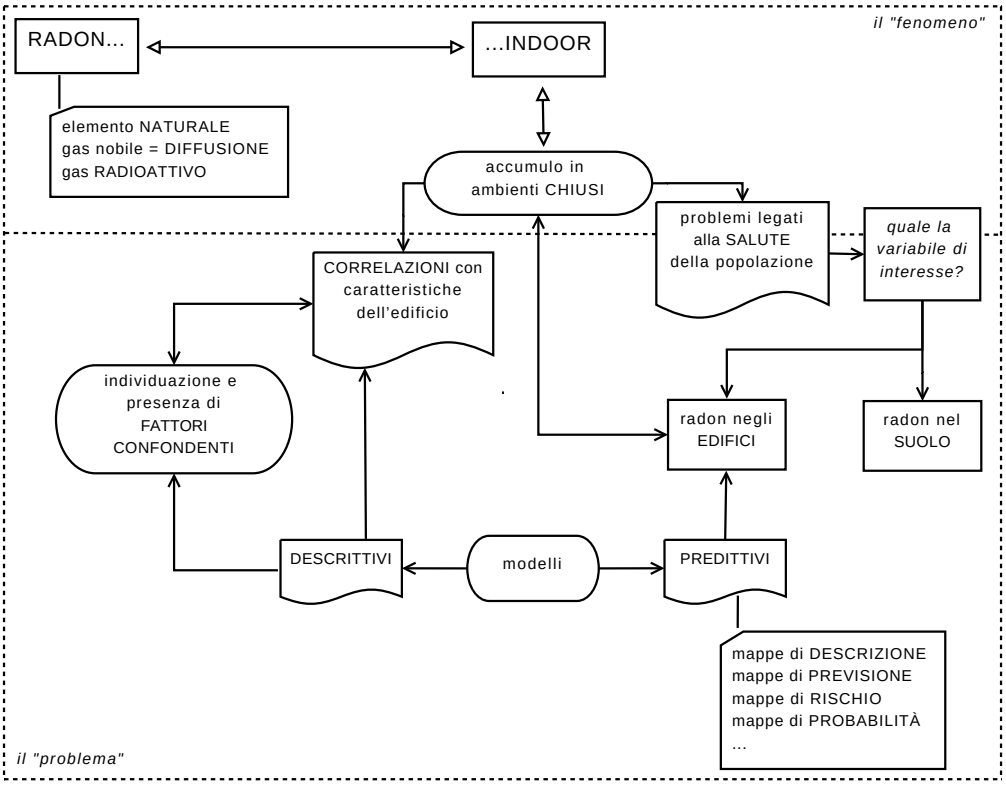


Figura 1.1: Diagramma di massima che descrive i principali soggetti e le loro possibili interazioni nell'ambito del problema affrontato e presentato in questo lavoro di dottorato.

alle fondamenta teoriche cui fanno riferimento — per affrontare il fenomeno e il problema radon indoor*: l'intenzione è stata quella di ricercare punti di vista alternativi e complementari dai quali poter osservare un problema comune secondo prospettive differenti.*

Nel labirinto delle possibilità oggi offerte dalle numerose tecniche statistiche disponibili, che possono venir adattate al problema in questione mutuandole da ambiti di applicazione anche molto diversi, si sono forzatamente dovute operare delle scelte che hanno privilegiato alcune tecniche rispetto ad altre, potenzialmente altrettanto utili o efficaci.

Queste scelte sono state dettate dalle complesse caratteristiche dei dati su cui le analisi sono state condotte [cfr. §1.3] e dalla disponibilità di software (principalmente di tipo Open Source) atto all'implementazione delle tecniche selezionate; inoltre, utili spunti per privilegiare una tecnica rispetto ad altre sono emersi partecipando a numerosi congressi di carattere nazionale ed europeo, frequentando scuole e corsi specifici, e discutendo e confrontandosi con esperti del settore con background e formazioni differenti.

Anche in relazione a questo, alcuni approcci si sono rivelati complementari, altri hanno fornito risultati in accordo tra loro pur partendo da punti di vista differenti, altri ancora hanno portato alla luce informazioni e chiavi di lettura alternative per interpretare quanto ottenuto in ambiti apparentemente slegati tra loro; infine, alcuni non hanno portato a nessuna conclusione, per

problemi legati alla complessità del dataset di riferimento e/o all'implementazione software delle tecniche in esame.

1.1.1 La struttura della tesi

Tutto il materiale sviluppato e raccolto durante il dottorato è stato suddiviso in tre parti, privilegiando non tanto l'aspetto cronologico del loro sviluppo quanto piuttosto in base alle caratteristiche teoriche che sottendono i vari modelli implementati — e più o meno esplicitamente, in base al fine delle analisi stesse.

Dopo un capitolo introduttivo relativo alla descrizione e contestualizzazione specifiche del problema e del fenomeno Radon Indoor (cap. 2), a ogni tecnica indagata verrà dedicato uno specifico capitolo, cercando di mantenere uno sviluppo logico comune. Dopo un iniziale, breve inquadramento del problema affrontato, delle motivazioni che lo hanno ispirato e dell'orizzonte verso il quale ci si è mossi, si forniranno dei brevi, sintetici richiami teorici relativi alla tecnica in esame, con lo scopo di definire l'ambito di applicazione e le fondamenta su cui la teoria generale si basa — se ritenuto opportuno, si forniranno riferimenti bibliografici per eventuali approfondimenti[2]. Seguirà la descrizione specifica del dataset impiegato nell'analisi e la descrizione cronologica della stessa, cercando di evidenziare i risultati parziali via via ottenuti. Al termine di ogni capitolo si cercheranno infine di riassumere i principali risultati ottenuti, sottolineandone di volta in volta la reale spendibilità nella pratica operativa, nonché eventuali pregi e difetti.

Si sottolinea fin d'ora come le conclusioni andranno comunque lette in relazione al dataset impiegato, che rispetto ad altri, relativi a misure di radon indoor condotte in regioni differenti, risulta essere peculiare per quanto riguarda l'elevata variabilità dei dati stessi e la presenza di numerosi valori di concentrazione particolarmente elevati rispetto alla situazione media nazionale[3].

I capitoli relativi alle varie tecniche esaminate sono stati raggruppati nel seguente modo:

Parte I: Geostatistica — in questo contesto, l'attenzione è focalizzata sulla sola *componente spaziale* del fenomeno, che riceve una trattazione raffinata e dettagliata attraverso una opportuna analisi del variogramma sperimentale, in grado, in linea di principio, di modellizzare in maniera ottimale la correlazione spaziale espressa dai dati;

- il capitolo 3 descrive l'applicazione di una particolare tecnica di stima mediante kriging che prevede l'assunzione (discutibile) di un modello log-normale dei dati per produrre una mappa dei valori di concentrazione di radon indoor sul territorio;
- i capitoli 4 e 5 sono invece dedicati a studi specifici volti all'interpretazione di alcune caratteristiche legate tanto al dataset operativo quanto al fenomeno stesso che rendono la struttura del variogramma sperimentale di difficile lettura e interpretazione;

Parte II: Statistica "classica" — in questo contesto, la parte spaziale del fenomeno viene trascurata, concentrando gli sforzi sull'implementazione di tecniche statistiche volte principalmente all'analisi esplorativa delle variabili secondarie, e nello specifico:

- a una più efficace modellizzazione della p.d.f. dei dati (rispetto ai modelli cui comunemente si fa ricorso) [cfr. cap. 6];

[2] Per quanto riguarda gli argomenti relativi alla geostatistica, che ricorrono in più capitoli, è sembrato più utile riassumere i principali concetti utilizzati in questo lavoro in un capitolo a parte, riportato in appendice B a pagina 217.

[3] Per una descrizione generale della situazione a livello nazionale, si può far riferimento a quanto riportato ad esempio da Bochicchio *et al.* {4}

- a una approfondita analisi descrittiva delle variabili secondarie, delle loro caratteristiche e delle loro correlazioni [cfr. cap. 7];
- a un studio dell'influenza e del peso predittivo delle variabili secondarie sul valore di concentrazione, anche in funzione del valore di concentrazione stesso; questo approccio consente inoltre di costruire dei modelli di abitazione di riferimento rispetto ai quali valutare l'influenza delle varie classi delle variabili secondarie prese in esame [cfr. cap. 8];

Parte III: Machine Learning — in base all'esperienza acquisita in ambito geostatistico [cfr. {42}], appare evidente comela sola componente spaziale del fenomeno, almeno in relazione al dataset di riferimento, non sia sufficiente per una trattazione e comprensione esaustive del problema che si vuole affrontare, sia esso di natura descrittiva o predittiva; per questo, si è tentato un primo e superficiale approccio alle tecniche sviluppate nell'ambito generale del *data mining*, e che nello specifico vengono etichettate come "machine learning"; tali tecniche consentono di trattare, accanto alla componente spaziale, anche l'informazione eventualmente contenuta nelle covariate. Nello specifico:

- il capitolo 9 descrive una tecnica volta all'individuazione delle variabili secondarie che mostrano un maggior peso predittivo sul valore di concentrazione;
- i capitoli 10 e 11 descrivono invece due tecniche alternative per la costruzione di modelli teoricamente in grado di stimare il valore di concentrazione ricorrendo anche alle informazioni relative alle caratteristiche proprie dell'edificio sede della misura; i risultati ottenuti per questi modelli verranno confrontati con quanto si può ottenere con un approccio di tipo geostatistico nel capitolo 12;
- infine, nel capitolo 13 verrà affrontato il problema della classificazione binaria per i valori di concetrazione di radon indoor.

1.2 Il precedente lavoro di tesi

Il lavoro di dottorato presentato e descritto nei capitoli che seguiranno si fonda in gran parte sull'esperienza che in questo campo ho accumulato nel corso della preparazione della tesi per la laurea specialistica in fisica {42}, e vorrebbe pertanto configurarsi come un tentativo di sviluppare e approfondire tematiche, problemi e questioni che sono solamente state portate in luce in precedenza, ma che non avevano ancora ottenuto l'attenzione che probabilmente meritano — non solo in un'ottica puramente *accademica*, ma anche in relazione alle ricadute socio-ambientali del problema radon.

Per questo mi sembra utile, ai fini di una più semplice comprensione e di una più snella trattazione dei risultati che verranno descritti, riassumere brevemente le principali conclusioni che sono emerse durante la stesura della citata tesi.

Dopo aver speso del tempo per rendere affidabile da un punto di vista sperimentale il dataset operativo impiegato per l'intero studio e descritto in §1.3, il lavoro di tesi si è focalizzato sull'applicabilità dell'approccio di tipo *geostatistico* [cfr. appendice B, a pagina 217] al Fenomeno Radon Indoor, caratterizzato nel caso specifico da:

- elevata variabilità dei dati;
- complessità orografica del territorio;

entrambi questi fattori hanno manifestato *pesanti ripercussioni* sulle stime che si possono ottenere applicando gli strumenti della geostatistica [cfr. §B.4 e B.6.1], *limitando* le potenzialità dei metodi impiegati — o, se si preferisce, configurandosi come un'interessante e difficile sfida ancora aperta.

La geostatistica offre una vasta gamma di strumenti, algoritmi e varianti degli stessi, tra i quali se ne sono scelti tre, in relazione i) alla relativa *semplicità* che li caratterizza (almeno dal punto di vista teorico), ii) alla *robustezza* che hanno manifestato in ambiti anche differenti tra loro e iii) agli obiettivi che di volta in volta ci si è posti. Quelli che hanno rilevanza in questo contesto, sono i seguenti:

1. l'approccio dell'**Ordinary Kriging** [cfr. §B.4.3], estremamente robusto anche in situazioni di non-stazionarietà — come quella in esame — si è rivelato utile per produrre una *mappa* in grado di rappresentare graficamente la *situazione media* del fenomeno sull'intero territorio altoatesino, fornendo così informazioni qualitative[4] per l'individuazione delle zone con maggior rischio di esposizione;

2. l'approccio delle **Simulazioni Stocastiche**, e in particolare quelle Gaussiane Sequenziali (sGs) [cfr. §B.6.1], ha consentito di *simulare* realizzazioni alternative ed equiprobabili del fenomeno sotto studio, mantenendo in linea di principio l'intera variabilità manifestata dai dati stessi — evitando così il tipico effetto di smoothing che caratterizza inevitabilmente ogni tipo di algoritmo di interpolazione, tra cui anche quello del kriging; l'elaborazione di queste realizzazioni ha consentito di ottenere delle *mappe di probabilità* per il superamento di determinati valori di soglia di concentrazione di radon indoor, uno strumento estremamente utile nell'ottica dell'individuazione delle zone a rischio.

Accanto ai problemi legati all'utilizzo e/o alla comprensione dei risultati ottenuti con software differenti — spesso, infatti, anche assegnando gli stessi parametri, si sono ottenuti output diversi —, le principali difficoltà incontrate nel precedente lavoro di tesi sembrano essere intrinsecamente legate alle caratteristiche proprie dei dati stessi, che si caratterizzano per:

- una natura *multivariata*, con la conseguente difficoltà di individuare le variabili che manifestano l'influenza maggiore sul valore della concentrazione;
- la presenza di *clustering*, che non sempre può essere adeguatamente trattato;
- una rete di monitoraggio necessariamente *non-omogenea*, in quanto legata alle localizzazioni degli edifici sede delle misure;
- una *elevata variabilità* manifestata *su scale spaziali differenti*, che inevitabilmente si ripercuote in tutte le fasi di analisi e previsione, dalla comprensione dell' eventuale correlazione spaziale, alla sua modellizzazione, fino all'interpretazione dei risultati ottenuti;
- un *contesto geografico* piuttosto variabile e complesso, caratterizzato dalla presenza di numerose zone montuose e valli che contribuiscono in maniera sensibile a introdurre direzioni privilegiate — a scale inferiori rispetto a quelle dell'intero territorio esaminato — che possono mascherare o confondere strutture spaziali a livello globale.

[4] Va da sé che la procedura di stima fornisce anche valori quantitativi, come il singolo valore di concentrazione e la relativa varianza, ma tale strumento non si è rivelato molto utile per una affidabile stima puntuale della variabile in esame.

Nonostante tutto questo, concludo sottolineando come in base ai risultati ottenuti, si possa affermare che il Fenomeno Radon Indoor può essere studiato e affrontato sfruttando gli strumenti messi a disposizione dalla geostatistica, ricorrendo eventualmente a misure della continuità spaziale non-standard, come ad esempio variogrammi relativi[5].

1.3 Il dataset di riferimento: descrizione generale

Una conoscenza approfondita e una buona familiarità con la variabile o le variabili coinvolte in indagini di tipo statistico risultano fondamentali sia in relazione alla scelta della strategia di analisi, sia in relazione a una corretta interpretazione dei risultati che le analisi forniranno. È quindi buona norma spendere del tempo per studiare a fondo le caratteristiche delle variabili di cui si dispone, alla ricerca di eventuali errori di inserimento, valori anomali, correlazioni tra due o più variabili, ecc...

L'acquisizione di tale familiarità con i dati può risultare ancora più importante nel caso si intenda ricorrere a un approccio di tipo *geostatistico* o di *machine learnign*, in quanto la consapevolezza della strategia di campionamento e la conoscenza dettagliata delle variabili correlate a quella sotto esame devono essere sempre tenute in considerazione nella fase di interpretazione dei risultati: ci sono infatti molte situazioni nelle quali risultati apparentemente "inspiegabili" o "anomali" possono ad esempio trovare una giustificazione in un campionamento non uniforme sul territorio (in ambito geostatistico, ad esempio), o ancora casi in cui una forte correlazione tra due variabili può mettere in crisi un algoritmo fornendo a esso informazioni ridondanti (in ambito di machine learnign).

In quest'ottica, sembra quindi importante nonché utile, prima di passare a una trattazione specifica delle analisi che le hanno viste protagoniste, fornire una descrizione di massima delle variabili che caratterizzano l'intera banca dati fornita dall'Agenzia Provinciale per la Protezione dell'Ambiente di Bolzano; il numero totale di misurazioni a disposizione è pari a **4064**.

Ogni misura di concentrazione di attività di Radon Indoor è accompagnata da una ricca serie di altre variabili relative alle caratteristiche dell'edificio, dell'ambiente e delle condizioni in cui la misura è stata condotta, al fine di caratterizzare il fenomeno sotto esame in relazione al *contesto* in cui lo stesso si manifesta: la *complessità* del fenomeno risulta infatti tale da richiedere una conoscenza il più dettagliata e variegata possibile delle caratteristiche geografiche, strutturali, geologiche e temporali che accompagnano la singola misura di attività.

Per questo, la banca dati raccoglie specifiche informazioni relative a:

caratteristiche geografico-spaziali dell'edificio : ogni misurazione è collocata nell'unità amministrativa e nello spazio tridimensionale, in modo che risulti così *georeferenziata* — caratteristica fondamentale per qualsiasi tipo di analisi geostatistica; le variabili sono nello specifico:[6]

- altitudine (0): indica la quota sul livello del mare espressa in metri;
- longitudine (0);
- latitudine (0);

[5] Per una panoramica di queste misure alternative per la correlazione spaziale, si può fare riferimento a {42} o a {16, pag. 45–46}

[6] Le variabili di tipo *quantitativo* sono indicate con il simbolo "○", quelle di tipo *qualitativo* con il simbolo "●"; inoltre, in parentesi viene riportato il numero di *missing values* — non tutti i valori di concentrazione sono infatti accompagnati da un set completo di variabili secondarie.

  - pendenza (0): indica la pendenza del terreno sul quale è stato costruito l'edificio sede della misura;
  - curvatura (0): è un indice della curvatura totale del terreno sul quale sorge l'edificio in cui è stato esposto il dosimetro;
- esposizione (1): sono state definite 8 classi di esposizione suddividendo l'angolo giro in settori di 45°, dove il Nord corrisponde a 0°; è stata inoltre introdotta una nona classe, denominata "flat", nel caso in cui non fosse riscontrabile una esposizione evidente;
- comune (0): sono stati esposti dosimetri in 118 comuni, tutti quelli presenti nella provincia di Bolzano;
- frazione (0);
- comprensorio (0);

caratteristiche specifiche dell'edificio : la concentrazione di Radon presente all'interno delle abitazioni è in stretta relazione con le caratteristiche non solo strutturali, ma anche con quelle di utilizzo dello stesso; le variabili prese in considerazione sono:

- data di costruzione (127): la maggior parte degli edifici campionati (50% ca.) sono stati costruiti tra il 1941 e il 1985;
- tipo di costruzione (45): sono state individuate 5 categorie, quali mattoni, cemento, sassi, prefabbricato e legno;
- qualità degli infissi (1330): sono state definite 3 differenti classi, quali bene, medio e scarso;
- contatto con il terreno (1336): indica se il locale sede della misura si trova in contatto con il terreno (anche contatto di tipo "laterale");
- piano dell'abitazione (13);
- utilizzo (5): sono state introdotte 4 classi che identificano l'uso dell'edificio sede della misura, quali abitazione, scuola, lavoro/scuola (luoghi di lavoro in un edificio scolastico), lavoro; come ci si può aspettare, il 75% ca. delle misure sono state condotte in abitazioni;
- tipo di locale (427): sono state individuate 10 principali tipologie di locale, quali camera da letto, salotto, cucina, corridoio, cantina, stanza, sala, aula, ufficio, negozio, altro;

caratteristiche geologiche : l'ambiente in cui l'edificio sede della misura è inserito è stato caratterizzato dalle seguenti variabili di tipo geologico:

- PERS (Potenziale di Esalazione di Radon dal Suolo) (827): in base a un precedente progetto di tipo geologico promosso dall'ANPA (Agenzia Nazionale per la Protezione dell'Ambiente), sono state introdotte 4 classi, quali altissimo, alto, medio, basso; questo progetto non ha però coinvolto l'intero territorio altoatesino, da cui l'elevato numero di missing value per questa e le altre variabili geologiche legate al progetto stesso;
- spessore (della roccia sottostante) (827);
- fratturazione (della roccia sottostante) (827);
- radioattività (827): indica la concentrazione di isotopi radioattivi nelle rocce;
- radioattività sottostante (827): indica la concentrazione di isotopi radioattivi nelle rocce che si trovano sotto una copertura di depositi continentali;

- tipo filone (827): i filoni — falde di rocce eruttive — possono avere un'importante influenza locale, in quanto lungo gli stessi si possono trovare arricchimenti o accumuli localizzati di elementi radioattivi;
- influenza filone (827): per ogni filone, è stata definita una fascia di influenza, l'ampiezza della quale risulta funzione della fratturazione delle rocce circostanti; sono stati definiti due livelli di influenza (alta, nulla), a seconda che l'edificio sede della misura ricada o meno all'interno della fascia stessa; ben il 98% degli edifici si trovano esterni alla fascia di influenza;
  - distanza filone (827);
- tipo faglia (827): una faglia — frattura della costa terrestre accompagnata dallo spostamento di una delle due parti lungo un piano — può creare delle vie preferenziali di risalita di gas dal sottosuolo;
- influenza faglia (827): analogo a "influenza filone"; in questo caso, il 73% degli edifici risulta esterno alla fascia di influenza;
- importanza faglia (827): indica se la faglia più vicina è di tipo principale o secondaria;
  - distanza faglia (827);

caratteristiche temporali della misurazione la notevole variabilità del fenomeno è legata anche alle caratteristiche temporali della misura; nello specifico, si sono raccolte informazioni riguardanti:

- anno di esposizione (0);
- mese di esposizione (428);
- stagione di esposizione (0): sono state introdotte due classi, in relazione al semestre tipico di esposizione del dosimetro, ovvero estate o inverno;
  - ore di esposizione (0): il valor medio per questa variabile corrisponde a circa 5 mesi e mezzo.

Parte di queste informazioni, come ad esempio le caratteristiche strutturali dell'edificio o il mese di esposizione, sono raccolte al momento dell'installazione del dosimetro nell'edificio; altre, come quelle relative alle caratteristiche geologiche del sottosuolo, sono raccolte invece con l'ausilio di ArcView©, un apposito software di tipo *GIS* (Geographical Information System) in grado di gestire ed elaborare varie informazioni tematiche associate al territorio.

CAPITOLO 2

Il Problema/Fenomeno Radon Indoor

Recentemente, e con maggior frequenza rispetto a qualche anno fa, capita a volte che i mezzi di comunicazione si occupino di RADON; si tratta di un elemento presente in natura con cui dobbiamo convivere, ma da cui dobbiamo anche proteggerci poiché, essendo un gas *radioattivo*, durante il suo processo di decadimento emette radiazioni ionizzanti (in particolare, particelle α, ovvero nuclei di elio composti da due protoni e due neutroni), cioé radiazioni in grado di innescare, quando interagiscono con la materia, un trasferimento di energia al materiale bersaglio, con generazione di ioni e/o atomi eccitati: l'effetto è quello di un potenziale *danneggiamento* del materiale stesso che può risultare pericoloso nel caso si tratti ad esempio di un tessuto biologico.

La quantità che viene utilizzata in campo radioprotezionistico per descrivere questo trasferimento di energia è la DOSE ASSORBITA (DA), definita come: DA

$$\mathrm{DA} = \frac{dE}{dm} \tag{2.1}$$

dove dE indica l'energia ceduta dalla radiazione ionizzante alla massa dm del volume di materia attraversato dalla radiazione stessa; l'unità di misura attuale è il *gray* (Gy) che corrisponde a 1 J kg^{-1}.

Accanto alla dose assorbita, si introducono anche:

- la `Dose Equivalente` (DE) (o equivalente di dose), misurata in *sievert* (1 Sv = 1 J kg^{-1}), che corrisponde alla DA *corretta* per i fattori di qualità della radiazione; DE
- la `Dose Equivalente Efficace` (DEE) (o equivalente di dose efficace) che corrisponde alla DE *corretta* per l'efficacia biologica relativa (RBE) della radiazione — non tutti i tessuti biologici, a parità di DE, rispondono manifestando lo stesso danno. DEE

L'uomo subisce *quotidianamente* — e da sempre — l'interazione con il campo di radiazione presente sulla Terra, che viene chiamato *radiazione di fondo*[1]; a tale contributo naturale si

[1] Si può dire che dall'alba dei tempi fino ad oggi, gli esseri viventi sono immersi in un vero e proprio "bagno di

aggiunge quello artificiale legato all'utilizzo delle radiazioni da parte dell'uomo. Nel caso delle *radiazioni ionizzanti*, il *danno biologico* avviene tipicamente a livello del DNA cellulare, con una probabilità P che dipende da numerosi fattori:

$$P = P_I \cdot P_{DNA} \cdot P_{NR} \cdot P_{NRI} \cdot P_{NA} \simeq 10^{-17} \tag{2.2}$$

dove P_I indica la probabilità di interazione radiazione-materia, P_{DNA} quella che il DNA subisca un danno in seguito a questa interazione, P_{NR} che tale danno non venga riparato[2], P_{NRI} la probabilita che la cellula mutante (danneggiata) non venga riconosciuta come tale, P_{NA} la probabilità che tale cellula, una volta riconosciuta, non venga distrutta.

Come riportato nell'equazione (2.2), si stima che solo 1 caso su 10^{17} dia origine a un fenomeno tumorale. In Italia, si stima anche che la dose di radioattività naturale cui è sottoposto annualmente ciascun individuo sia pari approssimativamente alla dose associata ad una radiografia del torace moltiplicata per venti.

La `radioattività naturale` deriva da tre diverse tipologie di sorgenti:

- *raggi cosmici* — radiazione proveniente dall'universo e filtrata in gran parte dall'atmosfera;
- *nuclidi cosmogenici* — generati dall'impatto dei raggi cosmici sull'atmosfera;
- *radiazione terrestre* — dovuta alla presenza sulla Terra di materiali radioattivi.

La `radioattività artificiale` deriva invece prevalentemente:

- dall'*uso pacifico delle radiazioni ionizzanti* — impiegate nella diagnostica medica, nella produzione di energia, nelle lavorazioni industriali;
- da *particolari attività o manufatti* — come viaggi in aereo, materiali da costruzione, combustione del carbon fossile, ...;
- dalla *ricaduta di frammenti radioattivi* generati a seguito di esplosioni e/o incidenti nucleari.

Si riporta a questo proposito in tabella 2.1 un quadro sintetico riassuntivo derivato dai valori proposti nel 1993 dall'UNSCEAR (Organismo delle Nazioni Unite che si occupa della radioprotezione) per le sorgenti naturali. Occorre far presente che i dati riportati sono *valori medi annuali* di equivalente di dose (DE) pro-capite: questo significa che vi sono alcune aree geografiche ove i valori medi possono essere sensibilmente più alti — ad esempio nelle città poste a grandi altitudini il contributo dei raggi cosmici può essere maggiore di un fattore da 5 a 10.

Come si evince chiaramente dai valori riportati in tabella 2.1, il radon è responsabile di circa il 50% del contributo naturale alla radiazione di fondo: esso è dunque l'elemento di gran lunga più importante per i potenziali effetti che può indurre sull'uomo, i più importanti dei quali saranno descritti in §2.4.

L'entità del rischio correlato all'esposizione al Radon e ai suoi prodotti di decadimento è ancora oggetto di discussione: per avere un termine di confronto facilmente comprensibile si fa spesso riferimento al rischio equivalente dovuto al fumo di sigaretta; per avere un qualche termine di paragone, si può fare riferimento a quanto riportato in tabella 2.2.

radioattività", la cui origine è del tutto naturale; per fare qualche esempio — e qualche confronto — un kg di granito ha un'attivita naturale pari a ca. 1000 Bq·m^{-3}, un litro di latte a ca. 80 Bq·m^{-3}, un litro di acqua di mare pari a ca. 10 Bq·m^{-3}e una persona di 70 kg... pari a ca. 8000 Bq·m^{-3}! — causata dalla presenza, nel corpo umano, di isotopi radioattivi naturali, il più importante dei quali risulta sicuramente il ^{40}K. Questi valori sono riportati dal sito Internet `http://www.sinanet.anpa.it`

[2] Il corpo umano, come ogni altro sistema biologico, essendosi sviluppato in un ambiente caratterizzato dalla presenza della radiazione di fondo, ha sviluppato sistemi di auto-riparazione e controllo delle cellule "anomale" molto efficienti.

Sorgente di esposizione	*Equivalente di dose annua* [mSv/anno]
Raggi cosmici	0,38
Radionuclidi cosmogenici	0,01
Radiazione terrestre (esposizione esterna)	0,46
Radiazione terrestre (esp.interna escluso Rn)	0,23
Inalazione ed ingestione di radon	1,205
Inalazione di ^{220}Rn	0,07
TOTALE	*2,355*

Tabella 2.1: Equivalente di dose media annua pro-capite derivante da sorgenti naturali di radiazioni ionizzanti in zone con fondo naturale normale; i valori sono stati proposti nel 1993 dall'UNSCEAR.

Conc. di Rn	N_d	*Rischio equivalente di contrarre tumore ai polmoni*
7400	440–770	60 volte il rischio di un non-fumatore
3700	270–630	rischio di un fumatore da 4 pacchetti/giorno
1480	120–380	4000 radiografie/anno al torace
740	60–210	rischio di un fumatore da 2 pacchetti/giorno
370	30–120	rischio di un fumatore da 1 pacchetto/giorno
148	13–50	5 volte il rischio di un non-fumatore
74	7–30	200 radiografie/anno al torace
37	3–13	rischio di un non-fumatore di contrarre tumore ai polmoni
7.4	1–3	20 radiografie/anno al torace

Tabella 2.2: Valutazione del rischio da esposizione al radon; i dati sono riportati da EPA in "A citizen guide to Radon"; la concentrazione di radon è espressa in $Bq \cdot m^{-3}$, mentre N_d rappresenta il numero stimato di decessi (su 1000) per tumore ai polmoni imputabile a esposizione a radon.

2.1 Proprietà Fisico-Chimiche

Il radon è un gas radioattivo appartenente alla catena di decadimento dell'Uranio-238, riportata in figura 2.1.

L'^{238}U è un elemento molto comune della crosta terrestre, con una concentrazione media pari a 4 ppm, anche se può variare considerevolmente in relazione alle caratteristiche geologiche del terreno[3]. La composizione isotopica *naturale* dell'Uranio assegna all'isotopo 238 un 99,3%, nella cui catena di decadimento, che termina con l'isotopo stabile ^{206}Pb, si trovano vari radionuclidi radioattivi. I primi cinque figli dell'^{238}U, chimicamente reattivi, tendono a fissarsi nella matrice solida del materiale: non creano quindi problemi di tipo radioprotezionistici, in quanto le particelle α emesse da alcuni di loro depositano tutta l'energia nella matrice stessa.

Diverso è invece il caso del radon (^{222}Rn).

A temperatura ambiente, è un gas incolore, insapore, inodore e quasi inerte, che quando viene raffreddato sotto il punto di fusione, pari a -71 °C, acquista una brillante luminescenza il cui colore tende al giallo al calare della temperatura per diventare rosso-arancione all'equivalente

[3]Nei filoni minerari cui si ricorre per l'estrazione di materiale da impiegare nella fabbricazione di combustibile nucleare, si può trovare una concentrazione di Uranio prossima allo 0.5%, ma capita spesso, come nel caso di graniti o altri materiali cementizi impiegati nelle costruzioni, di misurare concentrazioni anche pari a 20 ppm.

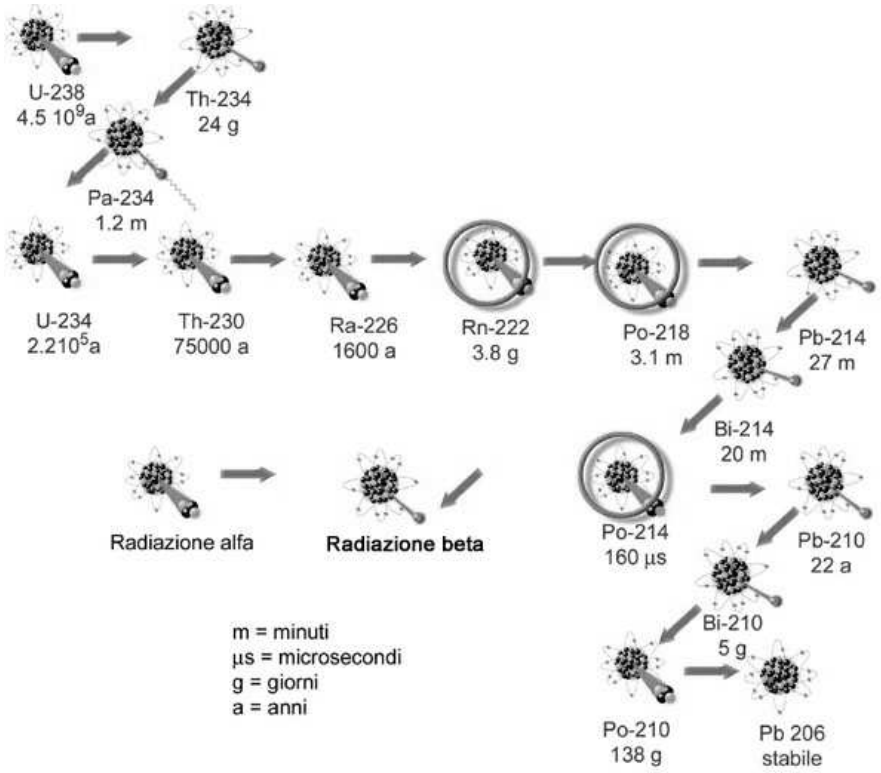


Figura 2.1: Catena di decadimento di ^{238}U.

temperatura dell'aria liquida. Chimicamente, è un gas nobile, e come tale non crea legami chimici[4] e tende quindi a *migrare* all'interno del materiale in cui si è formato, muovendosi in relazione a gradienti di concentrazione, pressione e temperatura; così, perché da un minerale comune possa verificarsi una efficiente emanazione di radon, questo deve formarsi nei primi 0.02–0.07 μm dalla superficie[5]; il radon che si forma più in profondità resta imprigionato nel materiale e decade in tempi brevi nei suoi sotto-prodotti solidi. Per questo, si può affermare che

> *la concentrazione di Uranio nel materiale, le caratteristiche meccaniche dello stesso — come porosità e granulosità — e le condizioni ambientali — come pressione, temperatura e umidità — caratterizzano l'*esalazione *di radon dal materiale stesso.*

Come riportato in tabella 2.3, il radon è poco solubile in acqua, mentre lo è molto di più nel caso di liquidi organici — per l'olio d'oliva, il coefficiente di solubilità è pari a 29.0 (a 18 °C); questo implica che il radon fuoriesce facilmente dall'acqua, ad esempio facendovi gorgogliare dell'aria o semplicemente scuotendola.

Il radon ha un tempo di dimezzamento, ovvero un tempo caratteristico trascorso il quale sono

[4] Si è scoperto che il Radon reagisce con il fluoro, dando origine a RnF_2; tuttavia, non si è ancora riusciti a caratterizzare questo composto.

[5] La migrazione avviene infatti con dei tempi caratteristici che sono *limitati* dal tempo di dimezzamento del radon, pari a 3.824 giorni [cfr. fig. 2.1].

numero atomico	86
massa atomica	222 g mol^{-1}
densità	9.96×10^{-3} g cm^{-3} a 20 °C
coeff. di solubilità in acqua	0.25 a 20 °C
punto di fusione	- 71 °C
punto di ebollizione	- 62 °C
configurazione elettronica	[Xe] $4f^{14}$ $5d^{10}$ $6s^2$ $6p^6$
energia di prima ionizzazione	1037 kJ mol^{-1}
data della scoperta	1898 da Fredrich Ernst Dorn

Tabella 2.3: Alcune proprietà fisico-chimiche del radon.

rimasti la metà degli atomi di partenza, pari a 3.8 giorni[6]. Si sottolinea in questo contesto che la conoscenza dei tempi tipici di decadimento del radon e dei suoi figli [cfr. fig. 2.1] diventa *fondamentale* sia per l'analisi del comportamento del radon all'interno di ambienti chiusi [cfr. §2.2], sia in relazione ai suoi effetti di tipo biologico [cfr. §2.4].

2.2 Il Radon negli Ambienti Chiusi

Le concentrazioni di radon nell'aria sono variabili in funzione, oltre che della presenza di Uranio nel sottosuolo, anche di numerosi parametri fisici o meteorologici — come la geomorfologia del sito, la pressione atmosferica, la temperatura, l'umidità , la stagione dell'anno, ... Il radon, liberandosi dal suolo in forma gassosa e attraversando il terreno, raggiunge la superficie e si mescola rapidamente con l'atmosfera — in una concentrazione tipica di 10 Bq m^{-3}.

Ben diversa è invece la situazione per i *locali chiusi* — situazione cui si farà riferimento col termine di Radon Indoor — dove si raggiungono normalmente valori molto superiori (anche di 2 ordini di grandezza). Ovviamente non è possibile realizzare edifici totalmente schermati dal radon, mentre è possibile, pur evitando costi elevati, progettare edifici con caratteristiche tali da minimizzare l'ingresso del radon o effettuare, in maniera relativamente semplice, il monitoraggio della presenza del radon in edifici già esistenti al fine di pianificare eventuali interventi.

Radon Indoor

All'interno di un ambiente lavorativo o abitativo, la concentrazione di radon può essere ricondotta in prima approssimazione a pochi fattori principali, quali:

Esalazione dal suolo : questo contributo è essenzialmente legato sia alle *caratteristiche geologiche* della zona, sia alle *condizioni atmosferiche*, rendendo così difficile una previsione accurata della quantità di radon emessa e del suo rateo di emissione; le rocce con maggior contenuto di Uranio e Radio — come *tufi*, *graniti* e *porfidi* — possono emanare maggiori quantità di radon, soprattutto se *permeabili* e/o fratturate: spesso sono proprio le *fratture* e le faglie a essere associate a elevate concentrazioni di radon, in quanto in corrispondenza di tali formazioni l'acqua trasporta, accumulandolo, l'Uranio; dal suolo, il radon diffonde negli ambienti chiusi passando dalle fondamenta e/o dalle penetrazioni delle tubature: particolarmente a rischio risultano quindi i locali *interrati* o *seminterrati*, specialmente se il pavimento è in terra battuta e le pareti non sono intonacate; dato che la diffusione del gas

[6] Se al momento dell'imbottigliamento di un'acqua minerale fosse presente del radon, dopo un mese di stoccaggio "praticamente" tutto il gas inizialmente presente sarebbe decaduto!

è limitata dal suo tempo di dimezzamento, ci si aspetta che questo fenomeno sia sempre meno rilevante allontanandosi dal suolo[7];

Materiali da costruzione : i materiali impiegati per la costruzione degli edifici possono contenere Radio (precursore del radon) in concentrazioni anche abbastanza elevate, che possono portare il materiale ad avere un rateo di esalazione di radon pericoloso; il rateo di esalazione risulta tanto maggiore quanto più alta è la concentrazione di Radio e la porosità del materiale: particolarmente elevati possono risultare i rischi associati ad alcuni ti pi di cementi, materiali argilloso e/o tufi.

Altri fattori possono essere dati dall'*aria esterna* o l'*acqua corrente*. In condizioni normali, non è riscontrato rischio da radon in seguito all'aerazione dei locali, che invece può configurarsi come una contromisura provvisoria per abbassare la concentrazione di radon nella stanza — nei casi in cui questa non sia comunque troppo elevata. Anche il secondo fattore citato generalmente non porta contributi significativi, a meno che la falda acquifera non scorra in prossimità o all'interno di rocce o suoli con un elevato rateo di emanazione: in questo caso, l'acqua funge da vettore per il radon e lo trasporta con tempi caratteristici molto inferiori a quello di dimezzamento[8]; per questo, in alcuni casi piuttosto rari si possono avere dei contributi alla concentrazione legati alle acque potabili.

2.2.1 Meccanismi di Diffusione

Per quanto riguarda l'Alto Adige, il meccanismo principale attraverso cui il radon penetra nelle case [cfr. fig. 2.2] è la *diffusione diretta dal suolo* attraverso fessure, crepe, pavimentazione naturale delle cantine, tubazioni, ...; il contributo legato alla diffusione dai materiali edilizi ben sigillati è quasi trascurabile. La causa principale dell'afflusso di questo gas all'interno delle abitazioni è da ricercare nella `differenza di pressione` che si viene a creare tra l'esterno e l'interno degli edifici — quest'ultimo risulta essere in depressione.

Sono due i principali fenomeni responsabili di tale *depressione* [cfr. ad esempio {2, pagg. 92–106}]:

Effetto Camino : è dovuto alla *differenza di temperatura* tra interno ed esterno dell'abitazione, a seguito della quale si forma un differenza di pressione ΔP; quanto più caldo è l'interno della casa, tanto maggiore sarà il flusso di aria fredda risucchiata dal terreno; il ΔP può essere calcolato in questo modo:

$$\Delta P = \mu \left(\frac{1}{t_{est} + 273} - \frac{1}{t_{int} + 273} \right) \tag{2.3}$$

dove μ è una costante pari a 3462 Pa·K, t_{est} e t_{int} la temperatura rispettivamente interna ed esterna, misurata in °C.

In base alla (2.3), con una differenza di temperatura pari a 30 °C, la differenza di pressione sarà pari a 1.3 Pa; questo implica che attraverso una fessura larga 1 mm e lunga qualche metro, possono venir aspirati diversi metri cubi di aria all'ora. È infine importante sottolineare che il funzionamento di una stufa o di sistemi di aspirazione in bagni o cucine possono produrre un effetto di risucchio e far quindi aumentare le concentrazioni di radon all'interno degli edifici;

[7] In base a questo fatto, la legge prescrive infatti che vengano monitorati tutti gli ambienti di lavoro che abbiamo almeno tre pareti su quattro interrate.

[8] Esempi tipici di questo fenomeno sono riscontrabili nelle *acque termali*, tanto che la legge italiana tiene conto in maniera esplicita del rischio radiologico da radon negli stabilimenti termali.

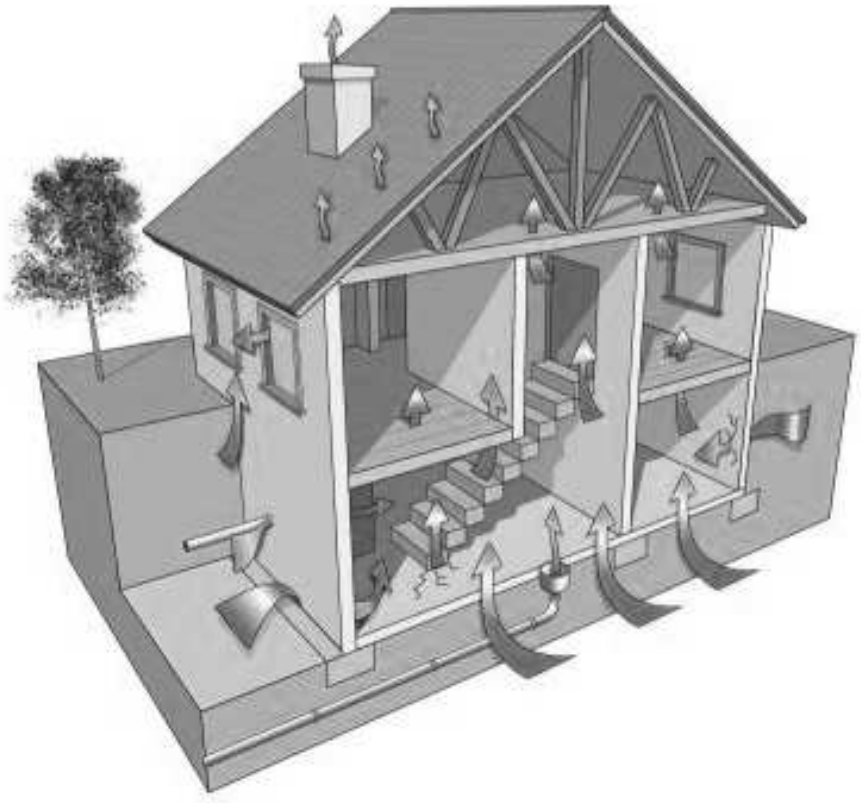


Figura 2.2: Principali meccanismi di diffusione del radon nelle abitazioni.

Effetto Vento : è dovuto alla *differenza di velocità* dell'aria tra esterno e interno dell'abitazione; la pressione P esercitata su una parete si può calcolare in questo modo:

$$P = P_0 + C_P \left(\frac{1}{2} \rho v^2 \right) \tag{2.4}$$

dove P_0 indica la pressione statica del vento, v la sua velocità , ρ la densità dell'aria e C_P un coefficiente di pressione che va determinato sperimentalmente in galleria del vento e dipende da una moltitudine di fattori (direzione del vento, presenza di aperture, forma dell'edificio).

Per fare un esempio tipico, alla velocità di 5 m/s e con una temperatura di 10 °C, viene indotta una depressione pari a $\Delta P = -5$ Pa.

In base a quanto esposto, è chiaro che la concentrazione di radon nelle abitazioni può subire notevoli variazioni, sia giornaliere che stagionali; in generale, valori più elevati si osservano al mattino e in inverno — quando cioè la differenza di temperatura è maggiore.

Inoltre, in Alto Adige si è osservato che la concentrazione aumenta quando il terreno *ghiaccia*: è ragionevole aspettarsi infatti che lo strato di ghiaccio ostacoli il normale flusso dal terreno di gas, che quindi diffonde attraverso zone non gelate, come il pavimento delle cantine o attraverso crepe.

2.3 La Normativa di Legge in Italia e in Europa

Il radon è un gas ubiquitario e naturale, e come tale non può essere eliminato completamente. Per questo, tutte le normative in materia manifestano come caratteristica generale quella di fissare dei cosiddetti livelli di azione o intervento, ovvero valori di soglia per la concentrazione di radon superati i quali *raccomandare* o *imporre* delle *azioni di rimedio* per ridurre la concentrazione sotto tali livelli — valori per i quali il rischio potenziale è ritenuto accettabile.

A livello internazionale, vari organi competenti hanno fissato indicazioni in relazione ai livelli di azione; tra questi, ricordiamo l'**ICRP** (Commissione Internazionale per la Protezione Radiologica), l'**IAEA** (International Atomic Energy Agency) e la **CE** (Commissione Europea).

Nella maggior parte dei paesi europei, si riscontrano delle distinzioni per il tipo di normativa proposta tra *ambienti di lavoro* e *ambienti residenziali* (abitazioni).

2.3.1 Ambienti di Lavoro

In Italia esiste una recente normativa che offre un quadro di riferimento sulla problematica del radon negli ambienti di lavoro, in particolare il

`DECRETO LEGISLATIVO 26 maggio 2000, n. 241`

> Attuazione della direttiva 96/29/EURATOM in materia di protezione sanitaria della popolazione e dei lavoratori contro i rischi derivanti dalle radiazioni ionizzanti. (pubblicato sul Supplemento Ordinario n. 140/L alla Gazzetta Ufficiale n. 203 del 31 agosto 2000)

Vengono individuate come attività di lavoro a rischio quelle svolte in luoghi quali tunnel, catacombe, grotte, terme e tutti i luoghi interrati.

Il decreto stabilisce inoltre, all'articolo 37, che

> *"La prima individuazione delle zone ad elevata probabilità di alte concentrazioni di attività di radon... avviene comunque entro cinque anni dalla data di pubblicazione del presente decreto nella Gazzetta Ufficiale."*

Questo limite temporale corrisponde al 31 agosto 2005, e in relazione all'articolo 10-sexies, comma 1, del decreto legislativo 17 marzo 1995, n. 230, introdotto dall'articolo 5, comma 1, del decreto n. 241, l'individuazione di tali zone ad elevata probabilità di alte concentrazioni di attività di radon spetta alle province e alle regioni autonome.

Il livello di azione fissato dal decreto è pari a $\mathbf{500\ Bq\ m^{-3}}$ — concentrazione di radon mediata su un anno.

2.3.2 Ambienti Residenziali

In Italia, attualmente *non* esiste alcuna normativa nazionale o regionale che fissi un limite di legge per gli ambienti residenziali.

Tuttavia, a livello europeo è in vigore la raccomandazione 90/143/EURATOM emanata il 21 febbraio 1990 sulla tutela della popolazione contro l'esposizione al radon in ambienti chiusi.

Questa raccomandazione consiglia, per gli edifici residenziali, una soglia di intervento pari a $\mathbf{400\ Bq\ m^{-3}}$ per gli edifici esistenti, e pari a $\mathbf{200\ Bq\ m^{-3}}$ per quelli ancora in fase di progetto. In caso di superamento di tali valori, si raccomanda di adottare *provvedimenti semplici ma efficaci volti a ridurre il livello di radon.*

2.3.3 La Normativa in Alto Adige

Nonostante le concentrazioni relativamente elevate di radon Indoor che si sono misurate in Alto Adige, non esiste a riguardo una normativa provinciale. Tuttavia, a partire dal 1998, su iniziativa del Laboratorio di Chimica Fisica dell'APPA di Bolzano in collaborazione con il Consorzio dei Comuni, è stato introdotto nel formulario di richiesta per la concessione edilizia un passo dedicato nello specifico al radon, invitando il cittadino a informarsi in merito presso il proprio comune o presso APPA.

In relazione agli ambienti di lavoro, viene adottata la regolamentazione in vigore a livello nazionale (Dlgs 241/2000); vengono regolarmente svolti controlli presso le Terme di Merano e sono state eseguite misurazioni in edifici pubblici — scuole, asili, biblioteche — e ambienti di lavoro particolari, come centrali idroelettriche e tunnel.

In diverse scuole e asili sono stati eseguiti con successo interventi atti alla riduzione del valore di concentrazione.

2.4 Effetti Sanitari e Rischio Associato

Poiché la principale sorgente di radon è generalmente il suolo, non stupisce che i primi effetti sanitari legati a questo gas e alla sua inalazione siano stati messi in evidenza sin dal XVI secolo sui minatori che operavano in miniere sotterranee, dove la concentrazione poteva raggiungere valori anche prossimi ai 100 kBq m^{-3}; si era notato che tali minatori erano affetti da una patologia polmonare cronica, detta "malattia dei minatori".

All'inizio del XX secolo venne scoperto l'elemento radon, e vennero condotti anche i primi esperimenti che consentirono di allargare la conoscenza scientifica al campo della *radioattività*: in particolare, si scoprì col tempo che le radiazioni ionizzanti (radiazioni α, per quanto interessa in questo contesto) potevano provocare tumori. Solo negli anni '50 si svelò il "mistero" dei minatori dello Schneeberg: indagini epidemiologiche condotte sui lavoratori delle miniere di Uranio mostrarono che il radon e i suoi prodotti di decadimento erano potenzialmente in grado di provocare il cancro polmonare.

Non a caso, quindi, le prime *indagini epidemiologiche* furono in larga scala condotte sugli operai delle miniere uranifere, a partire dagli anni '60, e vennero quindi introdotte la prime normative di protezione. L'attenzione sull'esposizione al radon nelle abitazioni è invece più recente, e tali anche gli studi (di vario tipo) per valutare il *rischio associato*.

2.4.1 Radon e Tumore Polmonare

Il principale effetto sanitario dei prodotti di decadimento del radon è l'`aumento di rischio di tumore polmonare`.

Per quanto riguarda l'evidenza di questo effetto cancerogeno, basti ricordare che a partire dal 1988 nella classificazione dei cancerogeni effettuata dalla IARC (International Agency for Research on Cancer) per conto dell'OMS (Organizzazione Mondiale della Sanità), i prodotti di decadimento del radon sono classificati nel gruppo 1 — sostanze per le quali l'evidenza di cancerogenicità è maggiore, e precisamente tale evidenza è ritenuta *sufficiente* in base a studi condotti su esseri umani (nello specifico, studi condotti su coorti di minatori).

Il Processo Fisico-Biologico

Il processo fisico-biologico che lega il radon al tumore polmonare si può sinteticamente descrivere in questo modo:

1. il radon è un *gas nobile* che, dopo essere stato emesso dal suolo e dai materiali di costruzione, si accumula nell'aria degli ambienti chiusi (case, luoghi di lavoro, ...), dove *decade* producendo altri radionuclidi, detti `prodotti di decadimento`;

2. quando si respira, l'aria inalata contiene sia radon sia i suoi prodotti di decadimento: il primo aspetto da prendere in considerazione è che *non tutti i prodotti di decadimento danno contributi significativi alla dose efficace* (DE); il contributo del radon alla DE è legato alla sua concentrazione media nei polmoni, ma non si accumula in essi, in quanto viene inspirato ed espirato assieme all'aria;

3. i prodotti di decadimento interagiscono invece con le pareti interne dell'apparato respiratorio attaccandosi a esse; qui decadono, emettendo `radiazioni ionizzanti`, in particolare particelle α e β; le più pericolose sono le prime, in quanto le seconde sono più penetranti, e quindi depositano la loro energia in strati molto maggiori di tessuto biologico; nelle cellule del nostro organismo, le molecole più vulnerabili sono quelle di grandi dimensioni e quindi statisticamente più esposte alla probabilità di essere danneggiate: le più grosse molecole presenti sono quelle del DNA e dell'RNA, responsabili della trasmissione del patrimonio genetico e della sintesi dei vari componenti cellulari; un danneggiamento di queste macromolecole porta nel migliore dei casi alla semplice morte della cellula, ma di frequente l'esito è invece una trasformazione tumorale, con la cellula che inizia a replicarsi in modo incontrollabile. Fortunatamente l'organismo è in grado, entro certi limiti, di individuare e distruggere le cellule cancerose, ed il rischio di contrarre tumori polmonari causati dalla progenie del radon è in media decisamente basso, ma in ogni caso, proporzionale alla dose accumulata nel tempo;

4. il `danno` biologico è quindi collegato ai figli del radon α-emettitori il cui tempo di dimezzamento fisico sia minore di quello biologico — il contributo maggiore alla DE è quindi dato dagli ioni ^{214}Po e ^{218}Po, che si attaccano facilmente al pulviscolo presente nell'aria e sono quindi trasportati, depositati (con tempi di dimezzamento biologico molto lunghi) e accumulati nell'apparato respiratorio al pari del pulviscolo[9] che li trasporta.

Quindi, *il radon si comporta come una sorta di "trasportatore" dei suoi prodotti di decadimento, i quali sono i* veri *responsabili del danno biologico*; tuttavia, è consuetudine fare riferimento al problema col termine 'rischio radon', intendendo in realtà il rischio connesso principalmente all'attività dei sui figli.

2.4.2 Fattore di Rischio per il Tumore Polmonare

La correlazione diretta tra concentrazione di radon negli ambienti chiusi e rischio associato di contrarre un tumore ai polmoni e tutt'oggi oggetto di studio.

Importanti fattori limitanti e di incertezza in quest'area di studio possono essere identificati in:

- *metodi per la stima* del fattore di rischio, che sono essenzialmente di tre tipi:

[9] Tutti i modelli per la valutazione della dose dovuta al radon e ai suoi prodotti di decadimento partono dalla considerazione che la dose dipende fortemente dalle *condizioni dell'aria*, dalla sua umidità , dalla concentrazione di pulviscolo e dalle dimensioni delle particelle che costituiscono il pulviscolo. In generale, si afferma che il pulviscolo con le dimensioni granulometriche inferiori penetra più in profondità nell'apparato respiratorio. La gran parte dei modelli elaborati è concorde nell'affermare che la frazione più rilevante dell'energia emessa dai figli del radon viene liberata nel tratto tracheo-bronchiale; meno di un decimo è ceduta nella parte alveolare dei polmoni. Se l'aria è ricca di fumo di sigaretta — le cui dimensioni granulometriche sono maggiori rispetto al normale particolato presente in atmosfera — una parte dei figli del radon viene depositata anche nel tratto faringeo.

  - nell'*approccio dosimetrico* viene determinata la DE in base a opportuni modelli (sia per l'organo di interesse che per il tipo di radiazione coinvolta) e quindi il fattore di rischio ricorrendo al fattore *rischio/dose* ottenuto in base agli studi epidemiologici condotti sulle due coorti di sopravvissuti di Hiroshima e Nagasaky: il problema è che queste persone sono state esposte per brevissimi tempi a radiazioni γ e neutronica, mentre il radon espone la popolazione a radiazione α in maniera continuata;
  - l'*approccio dosimetrico sui minatori*, da cui vengono estrapolati i fattori da riferire alla popolazione generale; il problema è che le condizioni ambientali dei due gruppi possono essere anche molto differenti; un recente studio condotto su un campione di 68000 minatori (2700 tumori polmonari) ha evidenziato una relazione lineare tra aumento di rischio relativo ed esposizione al radon — cui si attribuiscono il 40% dei tumori polmonari riscontrati;
  - l'*approccio epidemiologico residenziale* è il più recente e prevede lo studio condotto direttamente sulla popolazione generale nelle abitazioni; il problema è in questo caso legato alla bassa potenza di queste indagini e al basso numero di soggetti coinvolti; i risultati, tenendo conto delle incertezze statistiche, sono compatibili con quelli ottenuti da studi sui minatori;

- *effetto del fumo*: la difficoltà principale in questo tipo di studi è quello di controllare tutti i fattori che possono contribuire alla comparsa di un tumore al polmone (confounding effects), e in particolare il fattore legato al fumo di sigaretta; BEIR VI stabilisce un fattore di rischio per il cancro polmonare legato al fumo pari a [10–20], mentre quello legato al radon pari a [0.2–0.3]: è chiaro che questa notevole differenza richiede che le analisi condotte siano estremamente accurate nel raccogliere tutte le informazioni legate al fattore "fumo" — azione non sempre possibile; Conrady *et al.* {11}, hanno condotto uno studio volto alla determinazione del rischio relativo per il cancro polmonare legato al radon con lo scopo preciso di eliminare il bias introdotto dal fattore fumo (l'indagine è limitata a donne non-fumatrici); l'intervallo di esposizioni è ampio, e pari a [50–3000] Bq m^{-3}; tale studio conclude che gli studi precedenti *sovrastimano* il rischio relativo anche di un fattore 2 per concentrazioni fino ai 1000 Bq m^{-3}; oltre tale valore, il rischio relativo, sempre secondo i risultati di Conrady *et al.* {11}, sale rapidamente da 1 a oltre 7 per valori di concentrazione pari a ca. 3000 Bq m^{-3} (estrapolando con una retta i valori ottenuti da studi precedenti, e quindi assumendo valido il modello LNT[10], si ottiene un corrispondente rischio relativo pari a ca. 4.5); infine, si sostiene che questi risultati non sono compatibili con il modello LNT, il quale non dovrebbe più essere ritenuto adeguato nel caso di esposizione in ambienti abitativi;

- *effetto della "storia dell'esposizione"*: il cancro al polmone è caratterizzato da un alto periodo di latenza, per cui il più importante fattore legato alla dose risulta essere l'esposizione totale, che consideri almeno alcune decadi passate rispetto al presente; tuttavia, per ragioni di natura pratica, gli studi vengono condotti in relazione al valore di concentrazione misurata al momento dell'indagine stessa; Mc Laughlin {40} afferma che poiché il rischio legato al radon e ai suoi figli è il risultato dell'esposizione cumulativa che ha coinvolto le decadi passate piuttosto che l'esposizione attuale, è necessario ricostruire la storia dell'esposizione di ogni soggetto coinvolto negli studi al fine di ottenere un risultato *corretto* e *affidabile*.

[10] Il cosiddetto *radiation paradigm* sostiene che tutte le radiazioni a tutte le dosi risultano dannose; in relazione al problema radon, questo paradigma implica anche che il rischio di contrarre un cancro polmonare cresce linearmente con la concentrazione, senza la presenza di *valori di soglia* — il cosiddetto modello LNT (Linear No-Threshold).

Da un lato, in base a vari studi epidemiologici si è accertato che esiste una relazione *lineare* tra dose legata al radon e ai suoi prodotti di decadimento e l'insorgenza di tumori ai polmoni nei minatori; dall'altro, l'estrapolazione di tale risultato al caso di esposizione in ambienti abitativi è tutt'ora oggetto di indagini.

> *Si può comunque ragionevolmente affermare che esiste un* `rischio` *da tenere in considerazione per quelle persone che abitano o frequentano ambienti con concentrazioni di radon molto superiori a quelle previste dai limiti di legge; diventa quindi importante monitorare tutti quegli ambienti che presentano delle caratteristiche per le quali si potrebbero verificare degli* accumuli eccessivi *di radon e classificare le eventuali zone e rischio.*

Parte I

Geostatistica

CAPITOLO 3

L'approccio del kriging log-normale

L'idea che ha guidato questo approccio si basa sul fatto che spesso, in letteratura, si assume che la distribuzione dei valori di concentrazione di radon indoor sia di tipo log-normale*; per questo, in alcuni casi si suggerisce di ricorrere a una trasformazione logaritmica dei dati in modo da avere a disposizione un dataset di tipo* gaussiano*, con numerosi vantaggi sia di natura teorica che pratica.*
Nell'ambito di questo dibattito, che risulta ancora 'controverso', si è deciso di testare l'algoritmo del kriging log-normale *al fine di valutarne la reale efficacia, applicandolo a dati reali che hanno manifestato caratteristiche complesse.*

La parte computazionale è stata svolta principalmente in ambiente `R`{44} ricorrendo ai packages *geoR*{30} e *lattice*{45}; le analisi relative alla cross-validation [cfr. §3.4.2] sono state invece ottenute con `Geostat Office`©.

3.1 Brevi richiami teorici

L'algoritmo del kriging log-normale appartiene alla famiglia degli interpolatori nota come kriging non lineari [cfr. §B.4, pag. 232], nel senso che prevedono di condurre sui dati di partenza $Z(\mathbf{u})$ (dati 'raw') una *trasformazione non lineare* e ricorrere successivamente agli algoritmi lineari convenzionali del kriging applicandoli però alla nuova variabile trasformata $Y(\mathbf{u})$; questo è appunto l'approccio seguito in questo lavoro. In altre parole, tutti gli algoritmi di kriging non lineari altro non sono che kriging lineari (siano essi SK o OK) applicati a specifiche trasformazioni non lineari dei dati originali.

Il vantaggio principale di tali procedure risiede nel fatto che normalmente la trasformazione viene attuata al fine di rendere le analisi successive più semplici e di facile interpretazione; tuttavia, come sottolineato da chi critica questo tipo di approccio, il prezzo da pagare consiste nei numerosi problemi, sia di natura analitica che pratica, che accompagnano la delicata fase che caratterizza la trasformazione inversa — assolutamente necessaria se si desidera ottenere

un risultato finale che rispecchi i valori della variabile di partenza![1] Infatti, è noto che, pur applicando degli algoritmi di tipo lineare quali appunto quelli del kriging a una variabile che è stata preventivamente sottoposta a una trasformazione non lineare $F(Z)$, non è sufficiente applicare semplicemente $F^{-1}(Y)$ per riportare i risultati ottenuti nella scala di valori di partenza — a dire, la media dei logaritmi *non* è il logaritmo della media!

Formalmente, sia $z(\mathbf{u})$ la variabile di interesse (quale ad esempio il valore di concentrazione di radon indoor) e sia $y(\mathbf{u}) = \ln[z(\mathbf{u})]$ la variabile log-trasformata, con $y(\mathbf{u}) \in N(m, \sigma^2)$, ovvero caratterizzata da una distribuzione di probabilità normale con media m nota e costante sull'intero dominio di studio D. Sfortunatamente, una buona stima di $y^*(\mathbf{u})$ non conduce in maniera diretta a una altrettanto buona stima di $z(\mathbf{u})$; in particolare, una semplice (e apparentemente ovvia) trasformazione del tipo $e^{y^*(\mathbf{u})}$ risulta essere uno stimatore polarizzato di $z(\mathbf{u})$.

Sotto le ipotesi di trasformazione appena descritte, si può dimostrare {5, pag. 191} che, applicando l'algoritmo del simple kriging per ottenere la stima $y^*(\mathbf{u})$, lo stimatore non-polarizzato $z^*(\mathbf{u})$ per la variabile $z(\mathbf{u})$ e per la relativa varianza sono dati rispettivamente da:

$$z^*(\mathbf{u}) = \exp\left[y^*(\mathbf{u}) + \frac{\sigma^2_{sk}(\mathbf{u})}{2}\right] \tag{3.1}$$

$$\mathrm{var}\left[Z^*(\mathbf{u})\right] = \left[Z^*(\mathbf{u})\right]^2 \left[e^{\sigma^2_{sk}(\mathbf{u})} - 1\right] \tag{3.2}$$

dove $\sigma^2_{sk}(\mathbf{u})$ rappresenta la varianza ottenuta mediante simple kriging sui dati log-trasformati.

Sembra importante far notare come, in base all'esponenziale presente nell'equazione (3.1), la trasformazione inversa risulti particolarmente delicata in quanto 'esponenzialmente' sensibile a qualsiasi errore che coinvolga tanto la stima log-normale $y^*(\mathbf{u})$ quanto la sua varianza $\sigma^2_{sk}(\mathbf{u})$; inoltre, quest'ultima è a sua volta legata, oltre che alla distribuzione spaziale dei campionamenti attorno al punto di stima, al valore di sella del modello di variogramma utilizzato [cfr. §B.4.4, pag. 238]; vari autori insistono pertanto sull'importanza di una stima corretta del valore di sella del variogramma — e questo è a volte considerato un limite intrinseco del kriging log-normale.

Tuttavia, se le ipotesi di partenza sono verificate (o comunque si ritiene siano sufficientemente compatibili con il dataset utilizzato), la procedura di trasformazione inversa è analiticamente corretta e garantisce la non polarizzazione dello stimatore dato dall'equazione (3.1).

3.2 Descrizione del dataset utilizzato

Il dataset impiegato per lo studio presentato in questo lavoro si compone di **2312** valori georeferenziati di concentrazione di radon indoor riferiti a media annuale mediante una opportuna procedura di conversione dai valori misurati per una durata tipica pari a un semestre [cfr. {53}]; per ragioni di omogeneità, si sono selezionate le sole misure condotte al piano zero. Tali valori di concentrazione sono stati estratti dal dataset di riferimento descritto dettagliatamente in §1.3.

[1] Non è raro trovare infatti in letteratura articoli nei quali viene ad esempio applicata una trasformazione di tipo logaritmico ai dati di concentrazione, e successivamente i risultati (mappe, analisi statistiche, ecc.) discussi senza operare la trasformazione inversa — discutendo quindi di logaritmi dei valori di concentrazione e non di *reali* valori di concentrazione.
Questo tipo di analisi non mi sembra del tutto soddisfacente né del tutto corretta (anche da un punto di vista per così dire 'etico'), visto che in ambito applicativo/sanitario, quello che importa è il valore di concentrazione e non il suo logaritmo.

	$z(\mathbf{u})$	$y(\mathbf{u})$
N	2312	2312
media	200	4.75
σ	294	0.985
min	2	0.838
I quartile	56	4.030
mediana	102	4.622
III quartile	212	5.356
max	2924	7.981
skewness	4.40	0.506
kurtosis	27.1	0.074

Tabella 3.1: Alcuni parametri statistici riferiti ai valori di concentrazione di radon indoor prima e dopo la trasformazione log-normale; i valori sono riportati in $\mathrm{Bq{\cdot}m^{-3}}$.

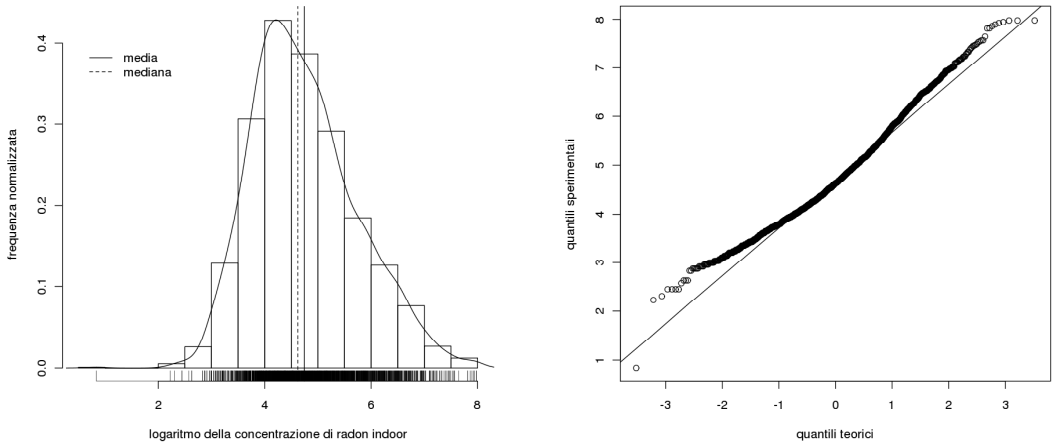
Dall'analisi di una recente panoramica delle campagne di misura condotte in 32 paesi europei [cfr. {18}], emerge come vari paesi abbiano optato per tecniche di campionamento e mappatura diversificate tra loro, manifestazione esplicita delle difficoltà incontrate nel modellizzare e conseguentemente predire i valori di concentrazione di radon indoor [cfr. {19}]. I valori misurati all'interno delle abitazioni risultano infatti caratterizzati da distribuzioni molto asimmetriche (tipicamente, in letteratura è abitudine assumere quasi implicitamente che siano di tipo log-normale[2]) e pesantemente influenzati sia dalle caratteristiche proprie dell'edificio/stanza sede della misura (come tipo di costruzione, abitudini di vita degli occupanti, effetti legati alla stagione e/o alle condizioni meteorologiche) sia dalle caratteristiche del contesto geologico e del sottosuolo. Anche in relazione a questi aspetti, sovente il variogramma sperimentale manifesta un *elevato effetto nugget*, indice di una marcata variabilità su corta scala: misure di concentrazione condotte in abitazioni anche molto vicine tra loro posso differire in maniera sensibile.

Nell'ottica di ridurre parzialmente questi problemi e facilitare l'analisi di tipo geostatistico, si è pensato di ricorrere a un algoritmo di interpolazione che appartiene alla famiglia dei cosiddetti kriging non lineari, e nello specifico all'algoritmo del kriging log-normale. Come descritto in aperture di capitolo, questo approccio prevede una preliminare trasformazione dei dati: la variabile di studio $y(\mathbf{u})$ diventa quindi il logaritmo naturale dei valori di concentrazione di radon indoor $z(\mathbf{u})$ che compongono il dataset originale; d'ora in poi, si farà sempre riferimento alla variabile trasformata $y(\mathbf{u}) = \ln[z(\mathbf{u})]$.

Alcuni parametri statistici riferiti ai valori di concentrazione prima e dopo la trasformazione sono riportati in tabella 3.1; si noti tra l'altro come i valori di skewness e kurtosis si avvicinino notevolmente a zero, valori caratteristici di una distribuzione di tipo normale [cfr. eq. (A.11) e (A.13) a pag. 214].

Da quanto riportato in figura 3.1a, si nota come la trasformazione logaritmica renda l'istogramma sperimentale piuttosto simmetrico e prossimo a una gaussiana, anche se media e mediana non risultano perfettamente coincidenti — come ci si aspetterebbe per una distribuzione di probabilità simmetrica; all'istogramma è stata sovrapposta anche la stima a kernel della corrispondente p.d.f. La figura 3.1b riporta invece un qq-plot per la variabile trasformata $y(\mathbf{u})$, i cui quantili sono confrontati con quelli di una distribuzione normale di riferimento: se si eccettuano le zone

[2]Per una trattazione dettagliata di questi aspetti e di quanto tale assunzione trovi reale riscontro nella pratica, si faccia riferimento a quanto discusso nel Capitolo 6.



(a) (b)

Figura 3.1: Istogramma (a) e qq-plot (b) per la variabile utilizzata nello studio geostatistico — logaritmo della concentrazione di radon indoor riferita a media annuale.

relative alle code, l'accordo si può considerare abbastanza buono. Infine, il test di Shapiro per la normalità risulta essere significativo con un p-value pari a $p \leq 2.2 \times 10^{-16}$.

In base a quanto emerso in questa fase di analisi, si assume che la variabile trasformata $y(\mathbf{u})$ ammetta una distribuzione di tipo normale.

3.3 Analisi esplorativa spaziale

Accanto a una prima analisi statistica di tipo convenzionale, l'approccio geostatistico prevede una serie di indagini che prendano in considerazione in modo esplicito le caratteristiche spaziali del fenomeno.

In questo contesto, risulta innanzitutto importante visualizzare la distribuzione spaziale dei punti di misura con un grafico che fornisca anche un'idea di massima della distribuzione dei valori della variabile in esame — ad esempio un grafico nel quale la dimensione del punto che identifica la localizzazione risulti proporzionale al valore della variabile di studio, come quello riportato in figura 3.2. Si noti inoltre come la distribuzione dei punti campionati sia tutt'altro che omogenea, ma segua forzatamente l'andamento delle valli che caratterizza il complesso contesto orografico altoatesino e si concentri principalmente nelle zone abitate.

Tuttavia, un'analisi specifica della distribuzione spaziale dei quartili non ha evidenziato particolari fenomeni di clustering, a dire che le distribuzioni spaziali di ogni singolo quartile sono in linea con quella dell'intero dataset; questo risulta importante sia in relazione alle ipotesi di stazionarietà che sottendono l'impiego del simple kriging [cfr. §3.1 e §3.5], sia per escludere la presenza significativa di clustering preferenziale [cfr. §B.1.2, pag.221].

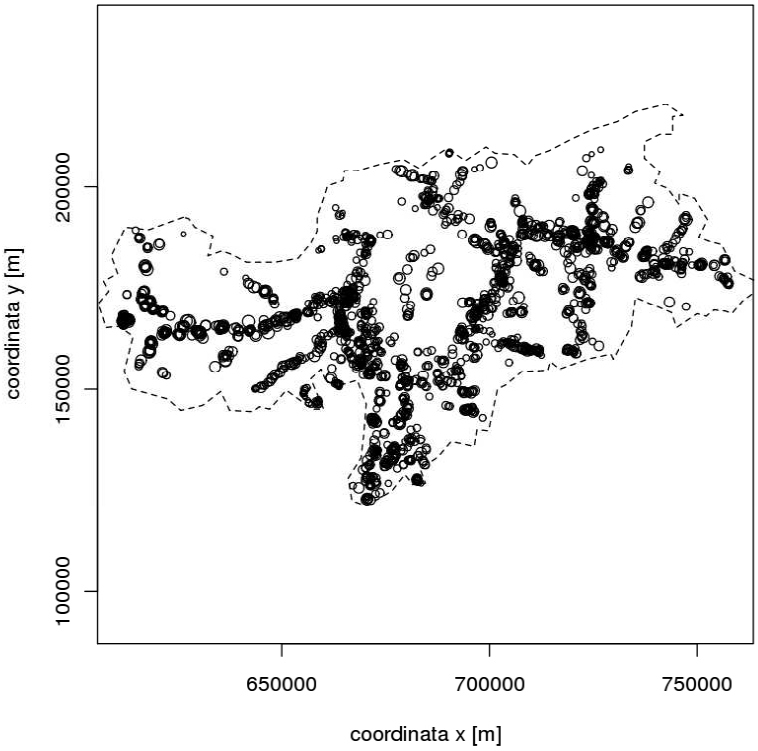


Figura 3.2: post plot dei valori della variabile $y(\mathbf{u})$; la dimensione del punto è proporzionale al valore della variabile stessa; la linea tratteggiata rappresenta i confini dell'Alto Adige.

3.3.1 Analisi della rete di monitoraggio

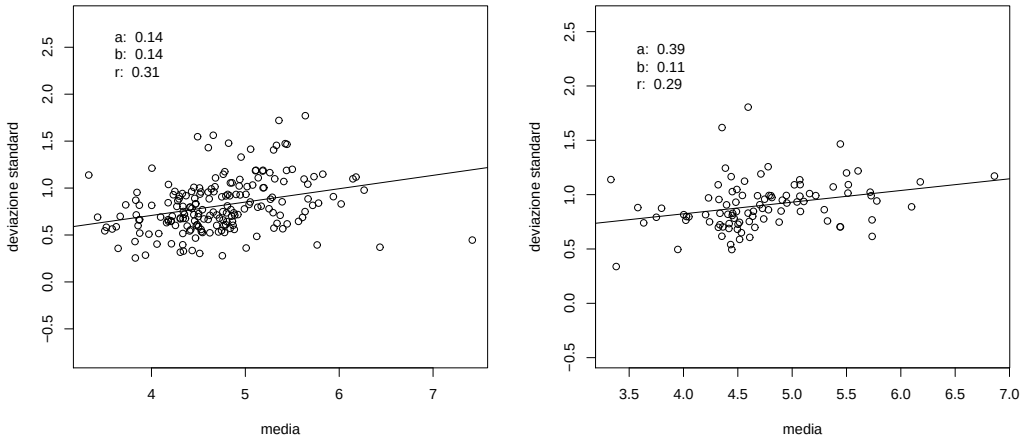
Data l'evidente disomogeneità della rete di monitoraggio evidenziata dalla figura 3.2, ho ritenuto opportuno determinare alcuni parametri quantitativi atti alle descrizione della rete stessa [cfr. §B.1.1]:

indice di Morishita : in linea con le precedenti analisi visive, questo indice evidenzia la *non-omogeneità* della rete di monitoraggio;

dimensione frattale : ricavata mediante box-counting, risulta pari a 1.7; nonostante la riconosciuta scarsa uniformità, i campionamenti sembrano essere sufficientemente ben distribuiti per una descrizione bi-dimensionale del fenomeno in esame.

3.3.2 Analisi statistica con finestre mobili

Ho condotto un'analisi mediante finestre mobili non sovrapposte sull'intero dominio di studio; le dimensioni scelte per le finestre sono state pari a (5×3) km^2 e (10×6) km^2, in rispetto delle



(a) (5×3) km^2 (b) (10×6) km^2

Figura 3.3: Risultati dello studio condotto mediante finestre mobili; 'a', 'b' e 'r' si riferiscono rispettivamente a intercetta, pendenza e fattore di correlazione relativi al fit lineare condotto mediante minimi quadrati.

proporzioni spaziali dell'area di studio, che risultano essere pari a $\Delta x/\Delta y = 1.69$.

> *In base ai grafici riportati in figura 3.3 e ai relativi parametri statistici, mi sembra ragionevole affermare che l'eventuale presenza di* effetto proporzionale *sulla variabile log-trasformata risulta trascurabile*[3].

3.4 Analisi variografica

Tutta l'analisi variografica *omnidirezionale* è stata condotta in ambiente R{44} ricorrendo al *variogramma tradizionale* e al pacchetto *geoR*{30}.

Inizialmente, ho deciso di considerare come distanza massima tra le coppie di punti quella che caratterizza l'intero dominio di studio (pari a circa 146 km), al fine di valutare la presenza di eventuali sotto-domini di stazionarietà; in tutti i casi, la tolleranza sul lag è stata scelta come metà del lag-step, in modo da utilizzare tutte le coppie disponibili.

Conducendo analisi dettagliate al variare del numero di bin (lag-step) e quindi della risoluzione spaziale dei variogrammi stessi, si nota:

- la presenza di un plateau tra i 60 e i 110 km, che si stabilizza su un valore maggiore di σ^2 (varianza a priori del dataset di riferimento);
- l'assenza di un segnale evidente della presenza di un trend;

[3] Questo probabilmente in relazione alla trasformazione che è stata applicata, visto che diverso è invece il caso in cui venga considerata la variabile raw $z(\mathbf{u})$: in questa situazione, come del resto risulta prevedibile anche da un punto di vista teorico [cfr. §B.1.3, pag. 222], analisi analoghe evidenziano in maniera significativa la presenza di effetto proporzionale.

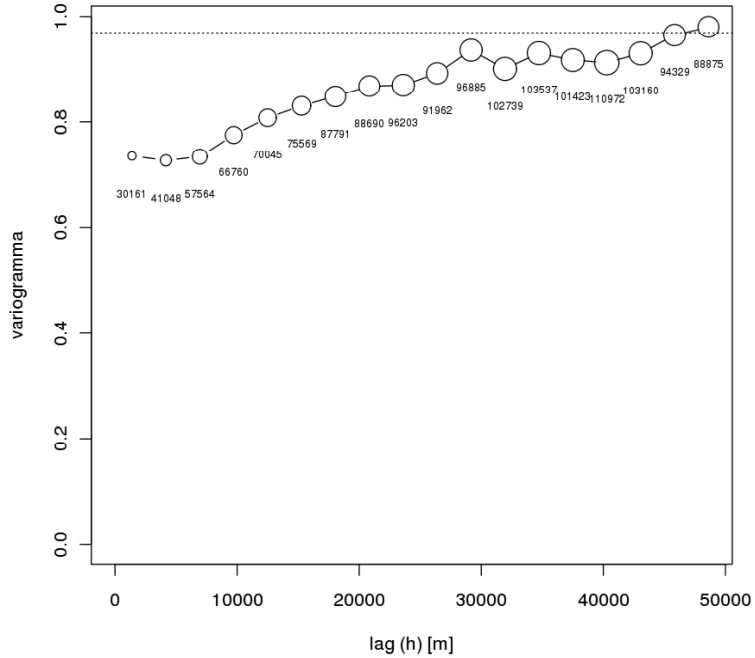


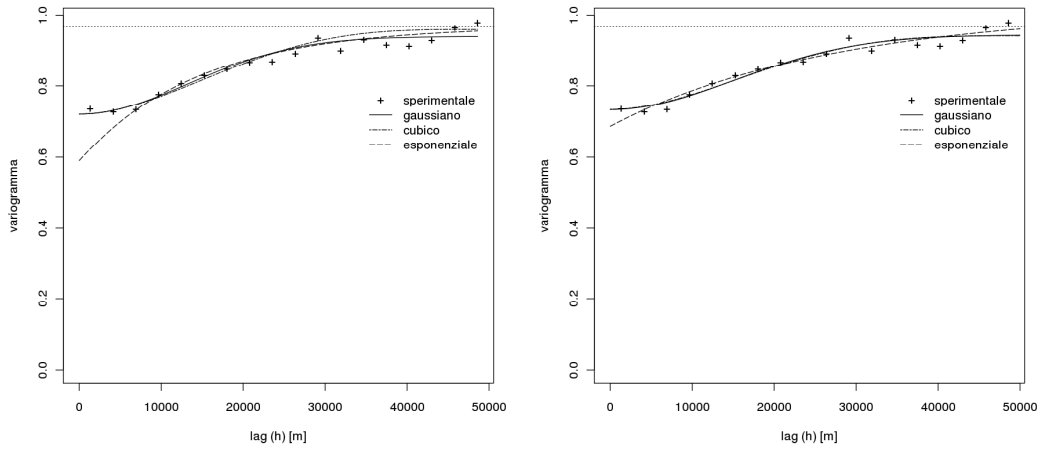
Figura 3.4: Variogramma sperimentale omnidirezionale $\gamma(h)$ per la variabile $y(\mathbf{u})$; il lag-step è pari a 3 km ca.; la dimensione dei punti risulta proporzionale al numero di coppie (riportato sotto il punto stesso) che caratterizza ogni lag; la linea tratteggiata rappresenta la varianza a priori sperimentale σ^2.

- all'aumentare del numero di bin, la comparsa di un secondo plateau tra i 30 e i 50 km che si stabilizza su un valore minore di σ^2;

In relazione di questi primi risultati, sembrerebbe di poter individuare la presenza di due strutture annidate, che operano su due scale spaziali differenti; tuttavia, alla luce di quanto emerso calcolando anche altri variogrammi al variare della distanza massima, ho deciso di limitare l'analisi ai primi 50 km, interpretando il primo plateau come segnale della presenza di un primo dominio di stazionarietà. A sostegno di questa scelta, il fatto che in relazione alle dimensioni del dominio di analisi e delle precedenti esperienze acquisite sui dati in esame, considerare correlazioni spaziali su distanze maggiori appare francamente poco ragionevole.

Le analisi successive si sono quindi focalizzate sullo studio variografico per i primi 50 km, variando la risoluzione spaziale e determinando anche variogrammi direzionali; questi ultimi non hanno evidenziato particolari anisotropie geometriche e/o zonali. Il lag-step che rende più leggibile la struttura variografica sperimentale è risultato essere pari a 3 km, ed è quello che caratterizza il variogramma riportato in figura 3.4.

Il variogramma sperimentale presenta un elevato effetto nugget, ma, fatto piuttosto 'curioso', l'andamento nei pressi dell'origine risulta molto dolce, indice di un fenomeno continuo — anche in questo caso, credo sia imputabile alla trasformazione introdotta, che ha un notevole effet-



(a) fit manuale

(b) fit WLS

Figura 3.5: Variogramma sperimentale omnidirezionale e relativi modelli teorici.

to di smoothing su dati raw, caratterizzati invece da una distribuzione di probabilità molto asimmetrica e ricca di outlier.

3.4.1 Stima e costruzione dei modelli

Per la costruzione del modello di variogramma da impiegare nelle successive fasi di interpolazione, sono ricorso a:

- un fit di tipo *manuale* — ovvero, ricavando i parametri dei vari modelli testati in modo interattivo, confrontando visivamente modello e variogramma sperimentale; in questo caso, ho scelto tre differenti modelli, e nello specifico:
  1. modello *gaussiano*
  2. modello *cubico*
  3. modello *esponenziale*

  i primi due modelli sono stati scelti in modo da riprodurre 'al meglio' il valore dell'effetto nugget e la continuità su corta scala; il terzo,in modo da introdurre un effetto nugget minore concentrando invece l'attenzione sulla zona di crescita;
- un fit di tipo *automatico* (WLS) — basato sulla minimizzazione dei minimi quadrati pesata sul numero di coppie che caratterizza ogni singolo lag.

I vari modelli ottenuti sono riportati in figura 3.5, mentre i relativi parametri nella tabella 3.2. Si nota come ci siano differenze significative tra stima manuale e stima WLS solo per quanto riguarda il modello esponenziale.

3.4.2 Cross validation: scelta dei modelli e dei parametri

Per valutare in maniera quantitativa il comportamento dei vari modelli di variogramma introdotti e l'influenza del parametro N_{max} coinvolto nella procedura di stima (descritta nel dettaglio in

modello	*gaussiano*		*cubico*		*esponenziale*	
fit	manuale	WLS	manuale	WLS	manuale	WLS
nugget	0.72	0.733	0.72	0.733	0.59	0.686
range	18904	21408	49961	51109	14853	27616
sella	0.22	0.212	0.24	0.209	0.38	0.329

Tabella 3.2: Valori relativi ai modelli di variogramma indagati; per maggiori dettagli, si faccia riferimento a §3.4.1; il valore del range è riportato in metri.

§ 3.5), che corrisponde la numero massimo di primi vicini utilizzati per la stima stessa, è stata condotta una serie di cross-validation (CV) [cfr. § B.5, pag. 239], ricorrendo al software `Geostat Office`©. *CV*

> ***Problema:*** *In seguito ad alcuni problemi legati all'implementazione software del modello cubico, l'analisi è stata 'forzatamente' limitata ai modelli di variogrammi* gaussiano *ed* esponenziale. *Va comunque sottolineato come questo inconveniente non risulti particolarmente significativo, visto che per i modelli WLS, il modello gaussiano e quello cubico sono 'visivamente' equivalenti* [cfr. fig. 3.5b], *mentre per quelli fittati manualmente si hanno differenze minime solo per quanto riguarda il valore di sella* [cfr. fig. 3.5a].
>
> *Per questo, ho deciso di limitare le cross-validation ai modelli di variogrammi* gaussiano *ed* esponenziale *fittati* manualmente *e al modello* esponenziale *fittato con* WLS — *quest'ultimo introdotto per valutare eventuali marcate sensibilità della procedura di stima al valore dell'effetto nugget.*

Tutte le cross-validation sono state condotte ricorrendo all'algoritmo del *simple kriging*, assumendo come media m quella campionaria, pari a 4.75 [cfr. tab. 3.1]. La griglia di simulazione ha una dimensione di (2000×1200) m^2, corrispondente a 74×73 punti di stima sull'intero dominio (rettangolare) di studio. Ho optato per un raggio di ricerca per i punti del vicinaggio pari a $S_r = \infty$ in quanto, visti i valori di range in gioco, anche limitando S_r in distanza il risultato sarebbe comunque quello di coprire l'intera area rettangolare — ho quindi preferito ricorrere al parametro N_{max} per limitare il numero di punti coinvolti nella stima, e conseguentemente valutare l'influenza di questo unico parametro (trascurando quindi S_r).

Descrizione dei risultati ottenuti

Se le ipotesi che sottendono il modello log-normale sono corrette, e se l'impiego del SK risulta accettabile, la teoria prevede che i residui standardizzati alla varianza del SK σ^2_{sk} debbano comportarsi come una variabile aleatoria con *media nulla* e *varianza unitaria*.

Questi e altri aspetti sono stati indagati sia con procedure statistiche convenzionali (determinazione dei parametri statistici relativi alle distribuzioni di probabilità dei residui standardizzati) sia ricorrendo a grafici del tipo:

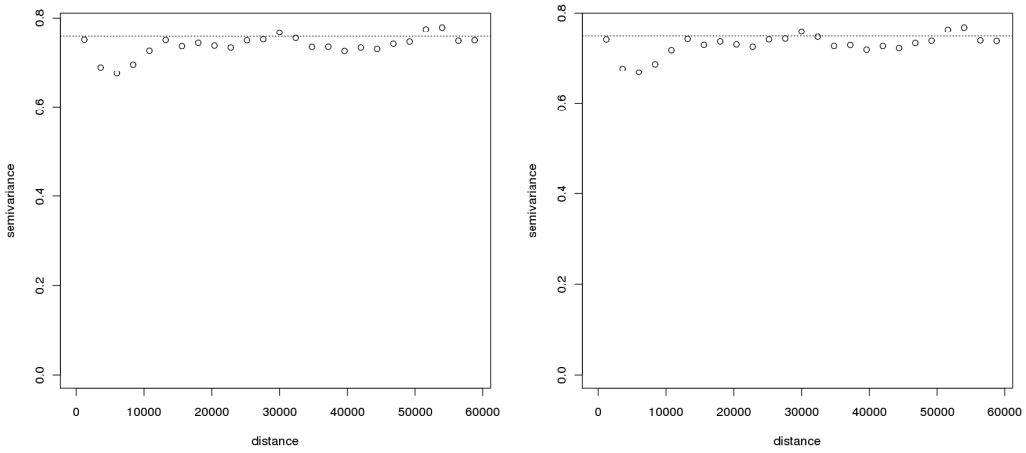
- 'valore stimato' *vs.* 'valore reale' per valutare l'entità dell'eventuale effetto di smoothing introdotto dalla procedura di stima e la correlazione tra queste due variabili;
- 'residuo standardizzato' *vs.* 'valore stimato' per valutare l'ortogonalità (che dovrebbe esserci dal punto di vista teorico) tra queste due variabili.

Dalle analisi numeriche (riassunte nella tabella 3.3) e grafiche condotte, si può concludere che:

id CV	*11*	*12*	*13*	*21*	*22*	*23*
N	2312	2312	2312	2312	2312	2312
min	-5.14	-5.08	-5.24	-5.83	-5.77	-5.85
media	-3.17e-5	-5.30e-3	-3.40e-3	-1.94e-3	-1.04e-3	-1.52e-3
mediana	-0.15	-0.16	-0.16	-0.18	-0.17	-0.17
max	3.88	4.15	4.56	4.56	4.42	4.46
σ^2	1.24	1.29	1.34	1.55	1.58	1.60
skewness	0.48	0.49	0.52	0.45	0.47	0.48
kurtosis	0.20	0.15	0.22	0.22	0.18	0.21
ρ *res vs. stima*	0.029	0.020	-0.012	-0.013	-0.017	-0.027
ρ *stima vs. reale*	0.465	0.466	0.460	0.471	0.479	0.481
intercetta	3.77	3.76	3.72	3.63	3.62	3.59
pendenza	0.20	0.21	0.22	0.24	0.24	0.24
id CV	*31*	*32*	*33*			
N	2312	2312	2312			
min	-5.23	-5.18	-5.30			
media	-5.44e-4	-2.91e-3	-7.98e-4			
mediana	-0.15	-0.16	-0.16			
max	4.02	4.05	4.30			
σ^2	1.26	1.30	1.32			
skewness	0.46	0.48	0.51			
kurtosis	0.21	0.16	0.21			
ρ *res vs. stima*	-0.009	0.004	-0.005			
ρ *stima vs. reale*	0.470	0.476	0.478			
intercetta	3.68	3.68	3.65			
pendenza	0.23	0.23	0.23			

Tabella 3.3: Risultati numerici relativi alle analisi statistiche condotte sui residui standardizzati ottenuti in fase di cross-validation; *id CV* = $\alpha\beta$ identifica la particolare coss-validation con la seguente codifica: α per il modello di variogramma (1 = gaussiano fit manuale, 2 = esponenziale fit manuale, 3 = esponenziale fit WLS), β per il parametro N_{max} (1 = 10, 2 = 20, 3 = 80); ρ indica il coefficiente di correlazione ricavato da fit lineare mediante minimi quadrati e 'intercetta' e 'pendenza' sono riferiti al fit per valore stimato vs. valore reale.

- *i valori di skewness e kurtosis risultano prossimi a zero, a indicare che gli istogrammi ottenuti sono di tipo gaussiano; in questo contesto, i variogrammi sotto esame risultato equivalenti;*
- *non si hanno evidenti problemi di polarizzazione (media prossima allo zero e varianza prossima all'unità); in base a questi parametri, il modello esponenziale con fit manuale risulta essere il peggiore;*
- *per quanto riguarda il parametro N_{max}, non si notano evidenti effetti di smoothing al suo aumentare, come evidenziato dal valore della pendenza della retta del fit lineare; la sua influenza sugli errori non appare pertanto evidente;*
- *in relazione al range degli errori (minimo e massimo), il modello gaussiano con fit manuale appare come migliore (range più basso);*
- *per quanto riguarda l'ortogonalità tra errore e stima corrispondente, il mo-*



(a) CV 12 (b) CV 32

Figura 3.6: Variogrammi omnidirezionali dei residui standardizzati relativi alle cross-validation 12 e 32 (per l'identificazione dei relativi parametri, si faccia rifermento alla didascalia della tab. 3.3); il lag-step è pari a 2.4 km.

dello esponenziale con fit WLS risulta essere il migliore (ma i coefficienti di correlazione sono comunque molto bassi in tutti i casi);

- *la distribuzione spaziale delle sovra- e sotto-stime risulta omogenea su tutto il dominio di studio, a testimonianza del fatto che la procedura di stima non è affetta, da questo punto di vista, da particolari fenomeni di polarizzazione o artefatti.*

Questi risultati mi sembra non consentano, limitatamente al tipo e al numero di analisi condotte, di poter scegliere un particolare modello di variogramma e un particolare valore di N_{max} su una base solida e/o certa, in quanto *globalmente* la varie cross-validation non manifestano differenze sostanziali; ho comunque deciso di ricorrere ai parametri relativi a *CV 12* e a *CV 32* per la stima della concentrazione di radon indoor mediante kriging log-normale.

Il modello 2 (esponenziale con fit manuale) è stato introdotto per abbassare 'sensibilmente' — comunque nel rispetto del variogramma sperimentale — l'effetto nugget, ma in base a quanto riportato in tabella. 3.3 non sembra rispettare molto i dati di riferimento (se paragonato globalmente agli altri modelli introdotti): per questo, ho deciso di non considerarlo per le stime successive.

Come ulteriore analisi, ho condotto uno studio variografico relativamente ai residui standardizzati con lo scopo di valutare la presenza di una eventuale componente residua di correlazione spaziale. In linea di principio, i residui dovrebbero essere completamente scorrelati tra loro, per cui i corrispondenti variogrammi dovrebbero essere dei 'puri' effetti nugget oscillanti attorno al valore della varianza a priori dei residui stessi.

Questo tipo di analisi è stata condotta per tutti i modelli analitici al variare del parametro N_{max}: come per le analisi precedenti, i risultati non manifestano particolari differenze in funzione dei parametri indagati. In tutti i casi (i variogrammi relativi alle CV i cui parametri sono stati

impiegati in fase di stima sono riportati, a titolo d'esempio, in fig. 3.6) i variogrammi sperimentali mostrano un andamento a effetto nugget oscillante attorno al valore della varianza a priori, e manifestano una parziale correlazione residua tra i 3 e i 10 km.

Per indagare nello specifico questa inattesa caratteristica, sono state condotte ulteriori analisi variografiche (sia direzionali che non) limitando la distanza massima d'indagine a 20 km e ricorrendo a lag-step pari a 1 km — sempre al variare del modello teorico di variogramma e al valore di N_{max}: la parziale correlazione permane in ogni caso. Questo sembrerebbe indicare che gli errori, benché scorrelati su piccole e grandi distanze, siano parzialmente correlati su quelle brevi-medie. Questa caratteristica è comune anche ai variogrammi della variabile log-trasformata (con un effetto meno pronunciato), e si rende ancora più marcata per le direzioni N e $N45E$. Ulteriori analisi non hanno però portato alla luce spiegazioni plausibili del fenomeno in base a caratteristiche riconducibili alle proprietà fisico/geologiche della situazione reale.

Un'ipotesi che è stata quindi presa in considerazione è stata quella per la quale la parziale correlazione residua — che ricordo risulta essere 'praticamente' indipendente sia dal modello di variogramma sia dal valore di M_{max} — potesse essere imputabile a una caratteristica *comune* a tutte le CV; si è pertanto indirizzata l'attenzione sulla rete di monitoraggio, che come evidenziato in fig. 3.2 e discusso in §3.3.1, risulta essere piuttosto disomogenea e tutt'altro che regolare.

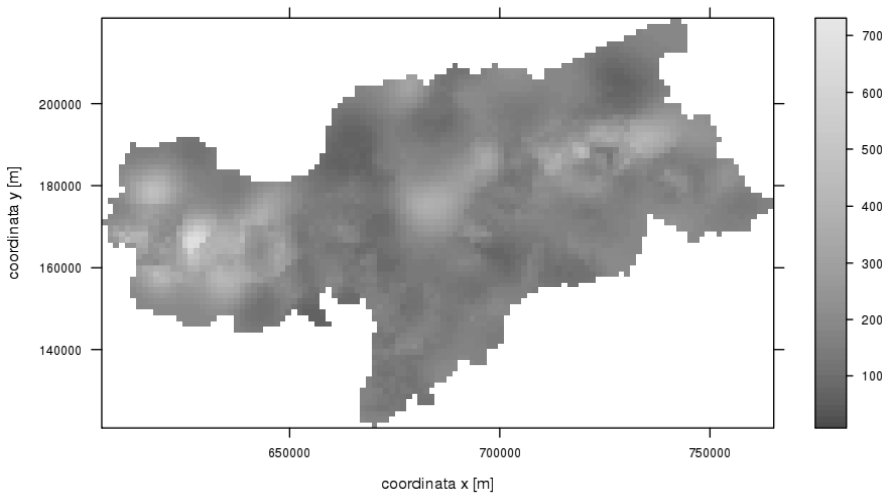
La plausibilità/veridicità di tale ipotesi è stata valutata con una operazione di *shake* sui residui, ovvero ridistribuendo in maniera *casuale* i loro valori mantenendo però fisse le singole localizzazioni: in questo modo, la rete di monitoraggio rimane la medesima, ma il risultato è quello di 'distruggere' l'eventuale correlazione spaziale della variabile sotto esame. Ho calcolato variogrammi per numerose operazioni di shake, su entrambe le scale di lavoro adottate in precedenza (20 e 60 km): in tutti i casi, si ottengono andamenti a puro effetto nugget sul valore della varianza a priori, in linea con la previsione teorica di una assenza di correlazione. La natura della correlazione residua va quindi ricercata altrove, eventualmente in un qualche artefatto legato all'implementazione dell'algoritmo di stima.

3.5 Stima della concentrazione mediante kriging log-normale

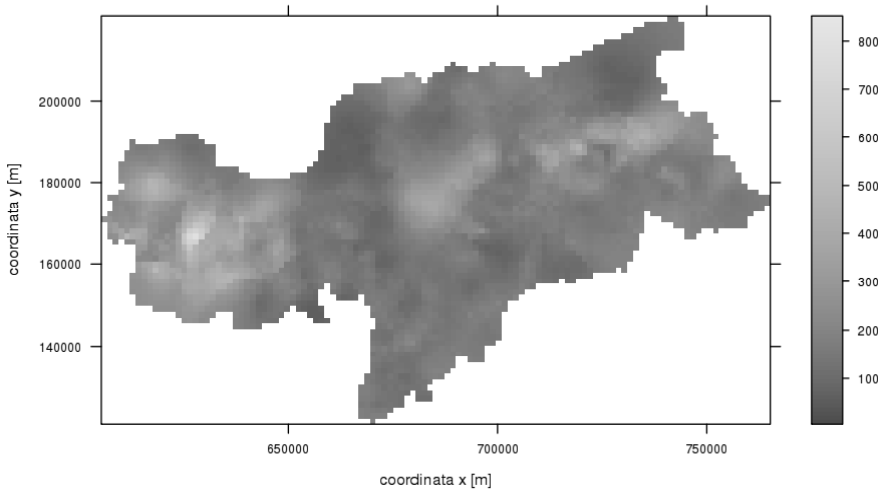
In figura 3.7 sono riportate due mappe per il valore della concentrazione di radon indoor riferito a media annuale, ottenute mediante simple kriging sui dati trasformati $y(\mathbf{u})$, a loro volta back-trasformati secondo le equazioni (3.1) e (3.2) allo scopo di ritornare alla scala di valori che caratterizza il dataset di riferimento. È stata scelta una griglia quadrata di simulazione con passo pari a 1.4 km ca. e un valore di $N_{max} = 20$.

Dall'analisi e dal confronto delle mappe 3.7a e 3.7b, che differiscono unicamente per il modello di variogramma impiegato, si può osservare che:

- *entrambe le mappe riproducono in modo corretto la* situazione media *che caratterizza il territorio, identificando in modo corretto le zone con i valori elevati di concentrazione (anche in relazione a quanto ottenuto in precedenza con altre procedure di stima, quali ordinary kriging e simulazioni gaussiane sequenziali* [cfr. {52} e §1.2]*);*
- *risulta evidente l'effetto di smoothing tipico della procedura di interpolazione (nessuno dei due modelli è in grado di fornire valori per la stima superiori a 800 Bq·m^{-3}, a fronte di un valore massimo del dataset pari a 2924 Bq·m^{-3});*



(a) modello gaussiano, fit manuale



(b) modello esponenziale, fit WLS

Figura 3.7: Mappe per il valore della concentrazione di radon indoor corrette a media annuale; i valori sono espressi in $\mathrm{Bq \cdot m^{-3}}$; il numero massimo di punti di vicinaggio N_{max} è per entrambe le mappe pari a 20.

id	*CV 12*	*CV 32*	*riferimento*
N	4007	4007	2312
min	59	57	2
I quartile	139	137	56
mediana	169	168	102
media	188	192	200
III quartile	217	230	212
max	687	799	2924
σ	77	85	294

Tabella 3.4: Alcuni parametri statistici riferiti alle stime per la concentrazione di radon indoor riferita a media annuale; i valori sono riportati in $\mathrm{Bq \cdot m^{-3}}$; 'riferimento' indica la statistica relativa al dataset relativo ai dati reali.

- *il modello gaussiano produce una mappa che propone un fenomeno più* continuo *e* morbido*, come evidenziato anche dalle curve di livello: questo non stupisce, visto che tale modello è di solito utilizzato per descrivere situazioni estremamente regolari e dolci.*

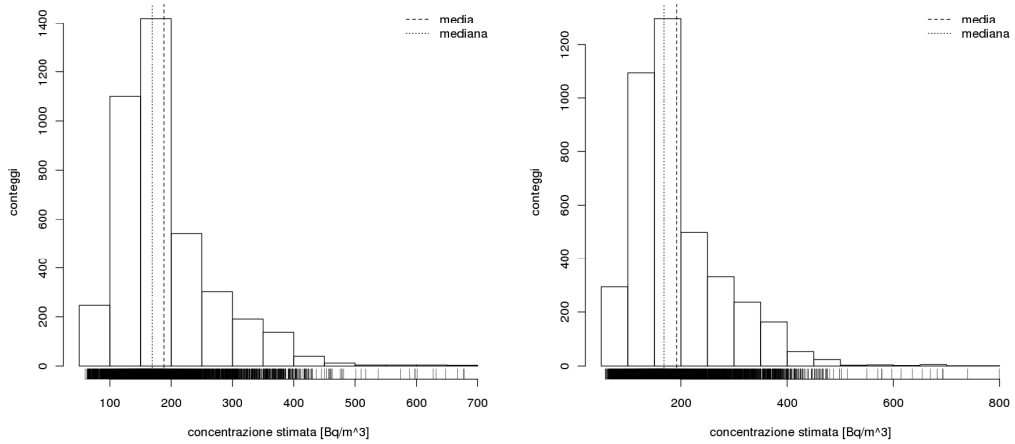
La tabella 3.4 riporta alcuni parametri statistici relativi alle stime ottenute mediante l'algoritmo del kriging log-normale; i valori su cui si basa l'analisi sono quelli che compaiono nelle mappe di figura 3.7, dai quali si è però avuta l'accortezza di rimuovere quelli esterni al territorio di interesse (con lo scopo di evitare possibili e probabili fenomeni di polarizzazione dei parametri). Non si notano particolari e significative differenze tra i due modelli di variogramma impiegati, se non, come prevedibile, una distribuzione leggermente più larga nel caso del modello esponenziale (CV 32). Ciò che risulta invece particolarmente evidente, è il significativo effetto di *smoothing* che caratterizza la procedura di stima, evidenziato tra l'altro dalla riduzione del valore di σ rispetto a quella che caratterizza il dataset reale.

Le distribuzioni di probabilità delle stime risultano asimmetriche e di tipo log-normale [cfr. fig. 3.8], in linea con quanto atteso: la procedura di stima dovrebbe infatti riprodurre un insieme di valori che rispecchino le caratteristiche statistiche (almeno globalmente) del dataset di riferimento.

Come ultima analisi, ho generato 6 serie di 3000 localizzazioni casuali all'interno del dominio di interesse e stimato il relativo valore di concentrazione, seguendo sempre la medesima procedura, con lo scopo di valutare eventuali ripercussioni sulla statistica delle stime — in relazione sia alla numerosità del dataset di riferimento, sia alla sua distribuzione spaziale disomogenea. Dalle 6 estrazioni differenti condotte, non si sono ottenute differenze significative: tutti i risultati sono in linea con il caso di griglia regolare discusso in precedenza.

3.6 Influenza dei parametri: nugget e sella

Anche in relazione a quanto espresso nel paragrafo 3.1, ho ritenuto opportuno e utile concludere la serie di analisi descritte in precedenza analizzando nello specifico l'eventuale influenza dei parametri caratteristici del variogramma (nugget e sella) sulle stime prodotte. Le discussioni dei risultati ottenuti si basano su cross-validation e mappe determinate in linea con quanto esposto nei paragrafi 3.4.2 e 3.5.



(a) modello gaussiano, fit manuale

(b) modello esponenziale, fit WLS

Figura 3.8: Istogrammi relativi alle stime per i valori di concentrazione di radon indoor riferiti a media annuale ottenuti mediante kriging log-normale; altri parametri statistici sono riportati in tabella 3.4.

effetto nugget : partendo dal variogramma sperimentale relativo alla variabile $y(\mathbf{u})$, ho costruito un modello esponenziale omnidirezionale fittando in maniera interattiva i dati empirici, cercando per quanto possibile di mantenere un compromesso accettabile tra andamento globale del variogramma ed effetto nugget *artificialmente* basso (per questo modello, ho scelto un valore pari a 0.3, da confrontare con quanto riportato in tab. 3.2).

Per quanto riguarda l'analisi dei risultati ottenuti in fase di CV, si può affermare che:

- *benché media e mediana dei residui siano comunque prossime a zero, la varianza risulta essere più alta rispetto a quanto ottenuto in §3.4.2 di un fattore 4;*
- *i residui sono di un ordine di grandezza più correlati alla stima;*
- *il variogramma dei residui manifesta anche in questo caso una correlazione residua tra i 6 e i 10 km;*
- *la distribuzione spaziale delle sovra- sotto-stime risulta omogenea sull'intero territorio d'indagine.*

Relativamente alle mappe ottenute, i risultati globali sono in linea con quanto esposto in §3.5, anche se in questo caso si ottiene una struttura spaziale più dettagliata: questo credo sia imputabile a una riproduzione dei valori di concentrazione che segue in maniera più fedele i campionamenti stessi (del resto, questo è in linea con quanto ci si aspetta dalla teoria [cfr. §B.4.4, pag. 238]) — in questo caso, è da aspettarsi però una minor capacità predittiva del modello, come confermato dai risultati globalmente peggiori che si ottengono in fase di CV.

valore di sella : in questo caso, sono partito dal modello gaussiano con fit manuale moltiplicando per un fattore 10 il valore di sella; la teoria prevede che in fase di stima, il valore dei pesi

assegnati dall'algoritmo del SK non cambi, mentre dovrebbe scalare dello stesso fattore 10 la σ^2.

Dalle analisi di CV e delle mappe prodotte, si nota che:

- gli istogrammi dei residui standardizzati sono più stretti di un fattore 10 ca.;
- sulle stime ottenute, non si ottengono differenze rispetto a quanto prodotto con i parametri impiegati in CV 12;

Si può quindi concludere che

> *l'influenza del valore di sella σ^2 risulta evidente e in linea con quanto previsto da un punto di vista teorico*

e sembra importante a questo punto sottolineare come questo effetto sia da non sottovalutare nel caso in cui si ricorra all'algoritmo del kriging log-normale, visto che in base alle equazioni (3.1) e (3.2) il suo valore viene 'pesato' in maniera esponenziale.

Problema: *Al fine di valutare in maniera sistematica l'influenza del valore si sella, si è provato a creare una* VARIOGRAM ENVELOPE *sul variogramma sperimentale, mediante la tecnica del* bootstrap*: fissato un modello teorico di variogramma, questa procedura prevede di determinare come variano i parametri del modello stesso ri-campionando di volta in volta il variogramma sperimentale; al posto di una singola linea univoca che rappresenta il modello teorico, si ottiene invece una* fascia *di possibili modelli attorno ai punti del variogramma sperimentale. Si ha così modo di valutare, anche visivamente, quale sia l'entità della variabilità dei parametri del modello stesso.*

Purtroppo, i risultati ottenuti sono stati poco chiari e difficilmente interpretabili in seguito a numerosi bachi relativi all'implementazione software della procedura; per questo, tale via (per quanto interessante) è stata abbandonata.

3.7 Conclusioni

Risulta quasi una consuetudine, in letteratura, assumere 'a priori' che la funzione densità di probabilità per i valori di concentrazione di attività di radon indoor sia di tipo *log-normale*. In virtù di tale assunzione, spesso si ricorre a una trasformazione di tipo logaritmico in modo da ottenere un insieme di dati la cui p.d.f. risulti di conseguenza di tipo *normale*, con vantaggi sia di natura pratica (ad esempio, leggibilità del variogramma sperimentale) che teorica. Sembra tuttavia importante sottolineare come non tutta la comunità sia concorde sulla bontà e validità di questa assunzione, e come il dibattito in merito risulti sempre più vivace. Analisi specifiche e differenti approcci di modellizzazione sono trattati nel dettaglio nel capitolo 6, a pagina 75.

In questo contesto, si è deciso quindi di applicare l'algoritmo del *kriging log-normale*, appartenete alla famiglia dei cosiddetti kriging non lineari: nella pratica, altro non sono che kriging lineari (simple o ordinary kriging, solitamente) applicati però a dati che hanno subito una preventiva trasformazione di tipo non lineare (come quella logaritmica, nello specifico). Il principale problema di questo approccio è quello legato alla fase di trasformazione inversa — non basta infatti calcolare l'esponenziale del valore ottenuto in fase di stima per ritornare alla scala di valori di partenza, quelli che sono rappresentativi della situazione reale. Se però si assume valido il modello log-normale, le equazioni (3.1) e (3.2) forniscono una soluzione analiticamente corretta per la fase di trasformazione inversa.

Dopo aver log-trasformato la variabile di interesse e assunto per valido il modello normale per tale variabile [cfr. §3.2], ho condotto analisi spaziali e variografiche, in base alle quali si può affermare che:

- la presenza di effetto proporzionale per la variabile log-trasformata risulta trascurabile (mentre non lo è per la variabile originale);
- la struttura dei variogrammi sperimentali risulta più leggibile e quindi di più facile modellizzazione;

questi aspetti, conseguenza della trasformazione introdotta, ricadono tra i vantaggi dell'approccio del kriging log-normale.

Sulla base dell'analisi delle mappe ottenute [cfr. figg. 3.7a e 3.7b] e dei parametri statistici relativi alle stime [cfr. tab.3.4], si può concludere che l'applicazione dell'algoritmo del kriging log-normale:

- consente di ottenere delle mappe che riproducono in modo corretto la *situazione media* del fenomeno sull'intero territorio di studio (anche in relazione a quanto ottenuto con approcci differenti, quali ordinary kriging e simulazioni gaussiane sequenziali [cfr. {52} e §1.2]);
- non è esente da un evidente *effetto di smoothing* sulle stime ottenute (nessuno dei modelli testati risulta infatti in grado di fornire valori per la stima superiori a 800 $\mathrm{Bq{\cdot}m^{-3}}$, a fronte di un valore massimo del dataset pari a 2924 $\mathrm{Bq{\cdot}m^{-3}}$);
- permette di lavorare con un dataset che rende le analisi di tipo spaziale e variografico più semplici.

Se posti a confronto con quanto si può ottenere applicando ad esempio un kriging ordinario (OK), i risultati ottenuti non mostrano differenze apprezzabili, probabilmente anche alla luce della natura estremamente complessa dei dati stessi, per i quali verosimilmente la sola correlazione spaziale non è sufficiente per una modellizzazione e comprensione esaustiva del fenomeno nella sua totalità — per approfondire questi aspetti, si può fare riferimento a quanto discusso nel capitolo 5 o nella parte III.

Concludendo, dalle analisi condotte sembra che l'unico apprezzabile vantaggio che si ottiene dall'applicazione del kriging log-normale sia quello di una più facile modellizzazione del variogramma, che non si traduce però in significativi miglioramenti in fase di stima. Inoltre, sembra importante sottolineare come errori o eventuali imprecisioni anche minime in fase di modellizzazione del variogramma, relative al valore di sella e/o di nugget, possano avere ripercussioni non trascurabili sui valori stimati, come discusso nel dettaglio nel paragrafo §3.6.

Per questo, e anche in relazione alla forte assunzione che si è costretti ad accettare in fase di pre-processing dei dati, l'approccio del kriging log-normale non si è dimostrato un'alternativa valida ed efficace rispetto ad altri algoritmi precedentemente testati, quali kriging ordinario e simulazioni gaussiane sequenziali.

CAPITOLO 4

Effetto nugget: analisi sulla possibile origine

L'idea che ha guidato questa analisi di approfondimento è stata quella di valutare la possibilità di trovare una base solida e ragionevole per ridurre *l'elevato effetto nugget che accompagna il variogramma sperimentale; le linee principali di indagine che mi è sembrato di individuare sono state sostanzialmente due:*

a) *indagine che ho identificato come* CASE-SPECIFIC*: una volta individuati i cosiddetti* punti anomali[a] *nell'analisi della* variogram cluod*, procedere a una loro identificazione e analisi* caso-per-caso*, al fine di ottenere informazioni specifiche sia sulla natura di tali punti sia sui punti che costituiscono il loro vicinaggio, con lo scopo finale di ottenere un qualche parametro decisionale in base al quale poterli eventualmente escludere dal dataset* ↦ questa esclusione porta a un miglioramento del comportamento del variogramma nei pressi dell'origine?

b) *indagine del (possibile) legame tra effetto nugget e* RETE DI MONITORAGGIO*, ovvero alla sua irregolarità e/o alla presenza di clustering; questo si è tradotto nell'implementazione di simulazioni ad hoc di fenomeni con un variogramma noto e quindi nel campionare il fenomeno stesso sia in modo regolare sia come è stato fatto nella realtà (MN disomogenea)* ↦ il tipo di campionamento ha ripercussioni apprezzabili e/o evidenti sul comportamento del variogramma nei pressi dell'origine?

[a] Punti che risultano per varie ragioni *diversi* rispetto ai punti che costituiscono il loro vicinaggio.

La parte computazionale è stata svolta in ambiente `R`{44} ricorrendo ai packages *geoR*{30}, *gstat*{41}, *RandomFields*{46} e *lattice*{45}.

4.1 Indagine case-specific

punti anomali

Questa prima parte del lavoro ha previsto l'identificazione dei cosiddetti PUNTI ANOMALI ricorrendo alla visualizzazione fornita dalla *variogam cloud* [cfr. §B.3, pag. 226]; tali punti giocano il ruolo di outliers nella fase di stima del variogramma, e se presenti per brevi distanze (piccoli lag h), si crede possano avere ripercussioni sull'elevato valore di $\gamma(h)$ nei pressi dell'origine, ovvero sull'effetto nugget.

4.1.1 Descrizione del dataset utilizzato

dataset 'all'

Il dataset operativo utilizzato si compone di **2578** valori georeferenziati di concentrazione di radon indoor estratti dal dataset di riferimento descritto nel capitolo 1, paragrafo §1.3; questo dataset sarà identificato con 'all'. Accompagnano il valore di concentrazione anche alcune variabili qualitative che caratterizzano gli edifici, con lo scopo di poter valutare eventuali *peculiarità* di un punto anomalo rispetto a quelli che compongono il suo vicinaggio. Ho infine deciso di non operare alcuna correzione sui valori di concentrazione, e pertanto, nell'ottica di mantenere una certa uniformità dei dati, ho preventivamente selezionato le misure condotte in *inverno* e al *piano zero*.

4.1.2 Analisi variografica

Inizialmente, ho calcolato dei variogrammi omnidirezionali esplorativi al variare della risoluzione di lag per una distanza massima pari a 80 km; un esempio di tali variogrammi è riportato in figura 4.1. Successivamente, ho determinato altri variogrammi al variare della massima distanza di lag e aumentando anche la risoluzione di lag — su una distanza massima pari a 10 km: come nei casi precedenti, i variogrammi sperimentali ottenuti risultano tutti molto rumorosi, con un elevatissimo effetto nugget (che si attesta sopra la varianza a priori) e una successiva discesa e stabilizzazione a valori inferiori alla varianza a priori. Come ci si poteva aspettare, anche in relazione a precedenti esperienze di analisi condotte sui medesimi dati (come discusso nel paragrafo 1.2), il variogramma tradizionale non è in grado di portare in luce una struttura di correlazione spaziale evidente, mentre lo sono altre misure alternative, come ad esempio il Pairwise Relative Variogram (PWR)[1].

Essendo questa analisi focalizzata sullo studio dell'effetto nugget, e riscontrata comunque una debole struttura nel variogramma, ho deciso di concentrare le successive fasi di studio a distanze massime pari a 3 km, visto che da osservazioni specifiche dei variogrammi la parte principale dell'effetto che si intende indagare si limita i primi 1000–1500 m.

4.1.3 Analisi della variogram cloud

Con lo scopo di identificare la presenza significativa ed evidente di eventuali coppie di punti anomali — eventualmente da mettere in relazione all'elevato effetto nugget riscontrato nella prima fase di analisi esplorativa — ho calcolato la variogram cloud sui primi 2000 m, ottenendo quanto riportato in figura 4.2.

Mediante una procedura interattiva, sono quindi state identificate le coppie di campionamenti che nella variogram cloud danno origine ai punti che sono stati considerati *anomali*, ovvero quelli contenuti nel riquadro indicato in figura 4.2; analizzando nello specifico le varie coppie di campionamenti che danno origine a tali punti, si riconoscono solo 4 localizzazioni ricorrenti

[1] Qualche informazione in merito è descritta da Deutsch e Journel {16}, pag. 45.

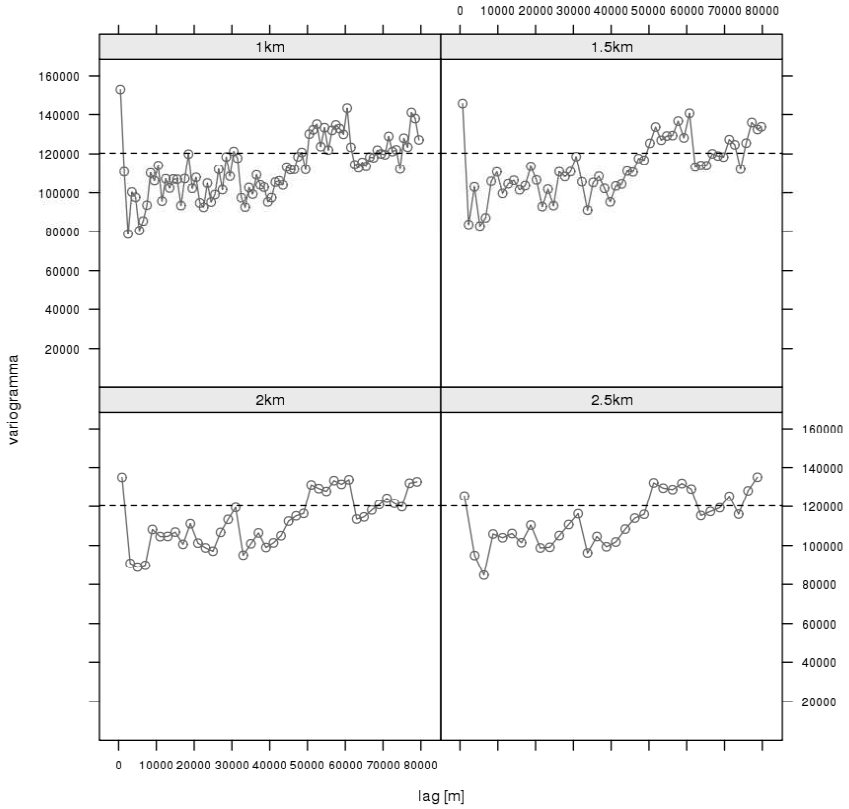


Figura 4.1: Alcuni variogrammi sperimentali relativi al dataset completo al variare della risoluzione di lag, riportata nella strip superiore di ogni grafico; la linea tratteggiata rappresenta la varianza a priori.

e comuni, la cui collocazione spaziale[2] è riportata nella figura 4.3. Una volta identificati questi 4 punti anomali, ho condotto delle analisi specifiche analizzando caso-per-caso la distribuzione spaziale dei vicini (determinando anche le relative distanze) e le specifiche caratteristiche dell'edificio sede della misura; lo scopo è stato quello di valutare la presenza di eventuali significative differenze che rendano l'esclusione di tali punti anomali sufficientemente ragionevole e motivata.

I risultati ottenuti si possono riassumere in questi termini:

case nr 884 : la sua concentrazione è pari a 3592 $Bq \cdot m^{-3}$, mentre il primo vicino manifesta una concentrazione di 1330 $Bq \cdot m^{-3}$(distanza pari a 1800 m ca.) e i seguenti una attorno ai 500 $Bq \cdot m^{-3}$; le distanze in gioco sono comunque relativamente elevate — per questo contesto d'indagine — e attorno ai 2 km ca.;

case nr 1026 : la sua concentrazione è pari a 3210 $Bq \cdot m^{-3}$, mentre i primi vicini manifestano una concentrazione piuttosto omogenea attorno ai 100 $Bq \cdot m^{-3}$; le distanze del vicinaggio

[2]Non stupisce che si vadano a posizionare proprio in quelle che sono note per essere le "zone calde" della regione di studio.

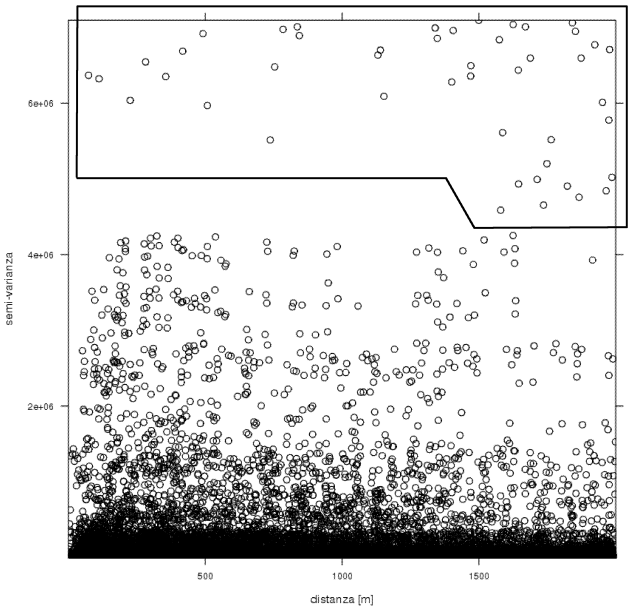


Figura 4.2: Variogram cloud per il dataset descritto in §4.1.1; il riquadro indica la zona cui si è fatto ricorso per l'individuazione dei punti anomali.

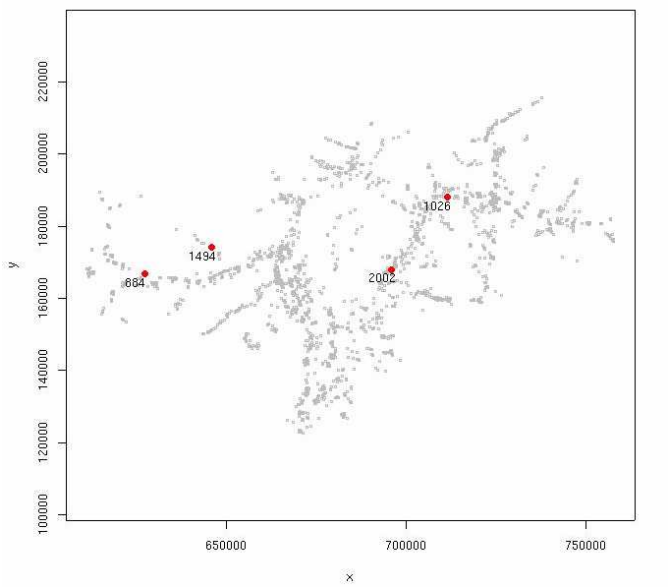


Figura 4.3: Localizzazione spaziale dei punti anomali identificati dall'analisi della variogram cloud di figura 4.2; i valori riportati si riferiscono al relativo *case number*; per maggiori dettagli, si faccia riferimento a quanto discusso in §4.1.3.

sono attorno ai 1500 m; il materiale da costruzione del punto anomalo è *sassi* e l'abitazione risale al 1700; mediamente, gli edifici vicini sono della seconda metà del 1900 e costruiti con *mattoni*;

case nr 1494 : la sua concentrazione è pari a 3787 $Bq \cdot m^{-3}$, mentre i primi vicini manifestano una concentrazione piuttosto omogenea attorno ai 200 $Bq \cdot m^{-3}$; le distanze dei vicini (5 campionamenti) sono molto piccole, dai 70 ai 500 m; l'intorno risulta molto omogeneo per quanto concerne la tipologia di edificio (buoni gli infissi, case recenti e di mattoni — tutte costruite dopo il 1950) e diverso dal punto anomalo (casa di sassi e risalente al 1550);

case nr 2002 : la sua concentrazione è pari a 3794 $Bq \cdot m^{-3}$, mentre i primi vicini manifestano una concentrazione piuttosto omogenea attorno ai 200 $Bq \cdot m^{-3}$; i primi 6 vicini coprono distanze dai 70 agli 800 m; in questo caso, le caratteristiche edilizie del punto anomalo rispetto ai primi vicini non sono molto diverse, eccettuato il fatto che l'anomalo è in contatto con il terreno[3].

In base alla serie di analisi riassunta in precedenza, ho deciso di eliminare dal dataset i 4 punti anomali identificati dall'analisi della variogram cloud di figura 4.2; questo ha condotto alla creazione di un nuovo dataset, che d'ora in poi sarà identificato con 'S01'.

dataset S01

4.1.4 Confronto variografico per differenti dataset

Ho quindi sviluppato l'analisi procedendo con un confronto visivo tra i variogrammi ottenuti per il dataset completo (all) e quello da cui sono stati tolti i 4 punti anomali (S01): per quanto riguarda l'effetto nugget (che ricordo è il parametro che in questo contesto risulta di maggior interesse), non si nota alcuna riduzione significativa, come si evince dalla figura 4.4; risulta invece più interessante il fatto che i due variogrammi siano molto simili per quanto riguarda *l'andamento globale*, con un evidente shift verso valori più bassi per $\gamma(h)_{S01}$, in una maniera apparentemente indipendente dal valore di lag. Ho quindi provato a incrementare la risoluzione sulla distanza di lag, con lag step pari a 200/400/600/800 m: lo shift individuato sembra non dipendere in alcun modo da questo parametro.

In base a una successiva analisi della variogram cloud relativa al dataset S01 e condotta in perfetta analogia a quella precedentemente descritta, ho deciso di eliminare un ulteriore punto anomalo; identifico questo nuovo dataset con S02: anche in questo caso, come riportato in figura 4.4, quello che si ottiene è un effetto di shift del variogramma.

dataset S02

> ***Nota 1:*** *Sembra a questo punto importante sottolineare come in tutti i casi descritti siano stati eliminati punti anomali caratterizzati da concentrazioni molto elevate e da un vicinaggio con concentrazioni molto basse: le anomalie riscontrate sono tutte e sole di questo tipo — a dire, nessun caso in cui si abbia un punto con concentrazione bassa circondato da punti con valori invece molto elevati.*

Riassumendo, mi sembra di poter concludere che per quanto concerne questa prima fase dell'indagine, *quello che è stato fatto credo possa venir interpretato come un* taglio *sulla coda superiore della distribuzione originaria*; a conferma di questa ipotesi, i risultati riportati nella tabella 4.1 mostrano una progressiva riduzione della varianza dei vari dataset.

Con lo scopo di valutare se il riscontrato effetto di shift fosse dovuto effettivamente all'eliminazione di outliers piuttosto che alla presunta riduzione della variabilità si corta scala, è stato creato un nuovo dataset, identificato con S03, sul quale ho imposto un cut a valori di concentra-

dataset S03

[3] Vale la pena di sottolineare che, anche se è l'unico parametro, tra quelli disponibili, che rende differenti punto anomalo e punti del vicinaggio, la sua influenza sul valore di concentrazione misurato si è rivelata, in base ad altre analisi, significativa [cfr. quanto discusso nei capitoli 8 e 9.]

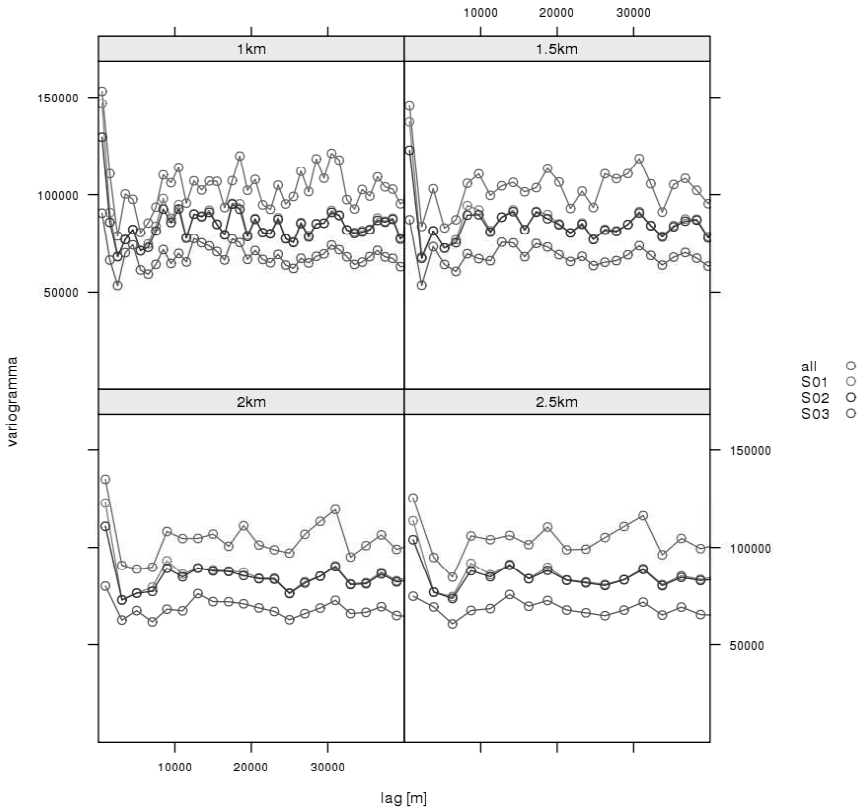


Figura 4.4: Variogrammi omnidirezionali sperimentali ottenuti al variare della risoluzione spaziale (riportata nella parte superiore di ogni grafico) e del dataset di riferimento; per maggiori dettagli sulla natura dei vari dataset, si faccia riferimento a §4.1.1, §4.1.3 e §4.1.4.

zione superiore ai 2000 Bq·m^{-3}; questo nuovo dataset si compone di **2561** cases. Un confronto variografico per tutti i dataset descritti fino a questo punto è riportato in figura 4.4.

Anche per il dataset S03 si può affermare, in linea con quanto ottenuto in precedenza, che:

- su larga scala è confermato l'effetto di shift verso il basso (legato alla riduzione della varianza, come evidente da quanto riportato in tabella 4.1);
- su piccola scala, si notano leggere differenze nel comportamento sul primo km circa, ma nulla di apprezzabilmente significativo — inoltre, l'analisi della variogram cloud non evidenzia la presenza di ulteriori punti anomali, per cui non mi sembra ci sia una motivazione solida per utilizzare il dataset S03.

A conclusione di questa fase di analisi e dei risultati riportati in figura 4.4 e in tabella 4.1, mi sembra di poter riassumere quanto emerso affermando che

> *l'elevato effetto nugget riscontrato nel variogramma sperimentale* non *si può imputare alla presenza di* isolati punti anomali, *ma piuttosto si manifesta come una sorta di*

dataset	*varianza*
all	120334
S01	102915
S02	100032
S03	77595

Tabella 4.1: Valori della varianza relativa ai vari dataset costruiti e descritti in dettaglio nei paragrafi §4.1.1, §4.1.3 e §4.1.4.

caratteristica intrinseca *del fenomeno stesso; in linea con questa ipotesi, il fenomeno di* shift *dei variogrammi risulta strettamente legato alla varianza globale del dataset — i variogrammi scalano infatti secondo questo parametro.*

4.2 Simulazioni per lo studio dell'effetto nugget

Con lo scopo di verificare le ipotesi e le conclusioni tratte nel paragrafo precedente, ho deciso di condurre delle simulazioni ad hoc mediante le quali poter controllare i vari parametri in gioco che ho ritenuto potessero essere rilevanti in questo contesto — in particolare, modificare il fenomeno simulato introducendo valori anomali con valori e intorni specifici e sotto controllo diretto[4].

È stata inizialmente creata una griglia regolare costituita da (100×100) punti con passo pari a 1; di conseguenza, risulterà regolare anche il campionamento[5]. Il fenomeno è stato simulato imponendo il seguente modello di variogramma — modello sferico con range pari a 30, sella pari a 0.8 e nugget pari a 0.2:

$$\gamma(h) = 0.2 + 0.8 \cdot \mathrm{sph}\left(\frac{h}{30}\right) \tag{4.1}$$

e imponendo una media globale pari a 10; questo fenomeno sarà identificato con 'N01'.

fenomeno N01

Sono state condotte 9 differenti simulazioni; in figura 4.5 sono riportati i corrispondenti variogrammi "sperimentali": si nota come le deviazioni rispetto a quello di riferimento (linea continua) siano in alcuni casi piuttosto marcate[6]; tuttavia, in base al comportamento nei pressi dell'origine — zona di maggior rilevanza per questo tipo di analisi — o comunque su corta scala, ho deciso di ricorrere alla simulazione numero 9 per le successive analisi; il dataset di riferimento per il fenomeno sarà d'ora in poi identificato con N01S9.

dataset N01S9

A questo punto, si è intervenuti sul dataset N01S9 introducendo 3 punti di tipo UP e 3 punti di tipo DOWN — con *up* indico punti il cui valore è stato aumentato rispetto all'originale, con *down* punti il cui valore è stato ridotto; l'entità di tali variazioni sono in linea con quelle che caratterizzano i punti anomali identificati nel dataset operativo [cfr. §4.1.1]; sono quindi stati modificati 6 punti della griglia di simulazione, e il nuovo dataset ottenuto identificato con N01S9a.

dataset N01S9a

[4] In relazione sia al valore numerico di concentrazione (più o meno elevato rispetto alle caratteristiche dell'intorno), sia alla situazione locale, ovvero introdurre situazioni con un singolo valore elevato circondato da valori inferiori ma anche situazioni opposte, che non sono presenti nei vari dataset reali presi in esame nelle analisi descritte nel paragrafo §4.1.

[5] Questa potrebbe essere una limitazione dell'analisi rispetto al caso reale, caratterizzato invece da un campionamento tutt'altro che regolare, ma alla luce dei risultati descritti in §4.3, tale aspetto non sembra essere molto rilevante.

[6] Non è risultato del tutto chiaro se questo sia imputabile a un qualche difetto nell'algoritmo di simulazione o sia piuttosto in linea con le fluttuazioni statistiche legate a una dimensione del dataset forse esigua.

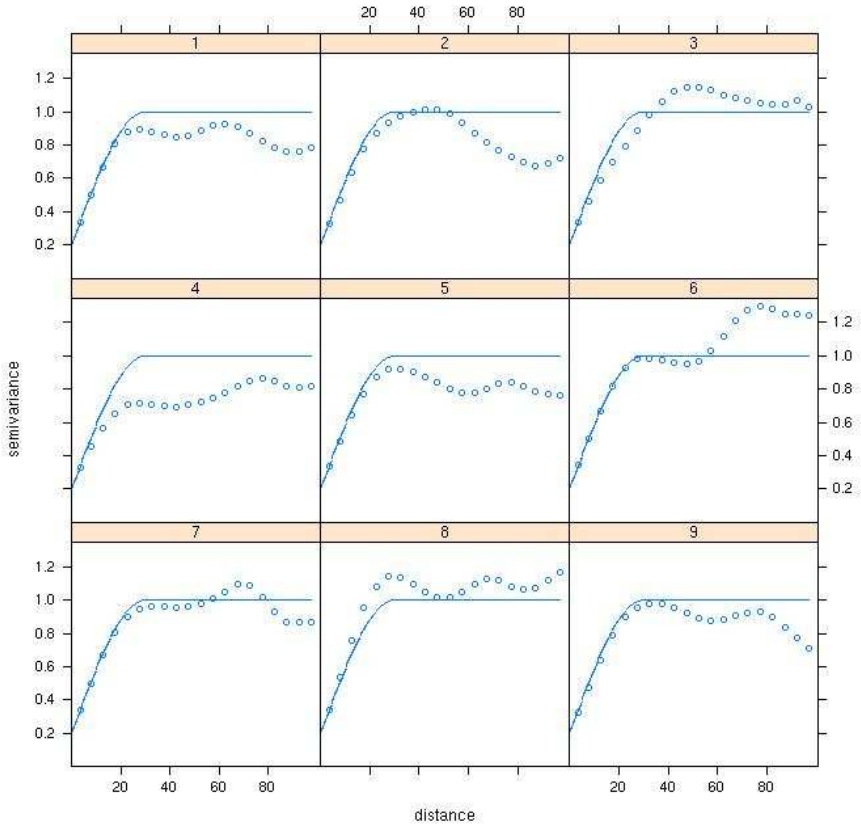


Figura 4.5: Variogrammi calcolati per le 9 simulazioni del fenomeno N01; la linea continua rappresenta il modello di variogramma di riferimento [cfr. (4.1)]; per le successive analisi, ho deciso di ricorrere alla simulazione numero 9.

Dall'analisi specifica della variogram cloud per i due dataset appena descritti, e riportata in figura 4.6, si evince che:

- i punti di tipo up sono identificati nel modo corretto e la loro influenza risulta evidente;
- i punti di tipo down non manifestano un'influenza evidente nella variogram cloud (mentre aggiungono un po' di rumore nel variogramma); per accorgersi della loro presenza, è necessaria un'indagine opportuna andando a variare (e nello specifico, diminuire) i valori della scala dell'asse y — a dire, vanno cercati di proposito.

Da questi primi risultati, mi sembra di poter affermare che

> *l'analisi della variogram cloud rende facilmente riconoscibile la presenza di una situazione del tipo* "valore elevato con vicinaggio caratterizzato da valori bassi"*, ma non quella di una situazione opposta, che va in qualche modo "ricercata"*[7].

[7] Peraltro, questo è in linea con un ragionamento legato al valore della varianza, che forzatamente tende a evidenziare situazioni/punti di tipo up.

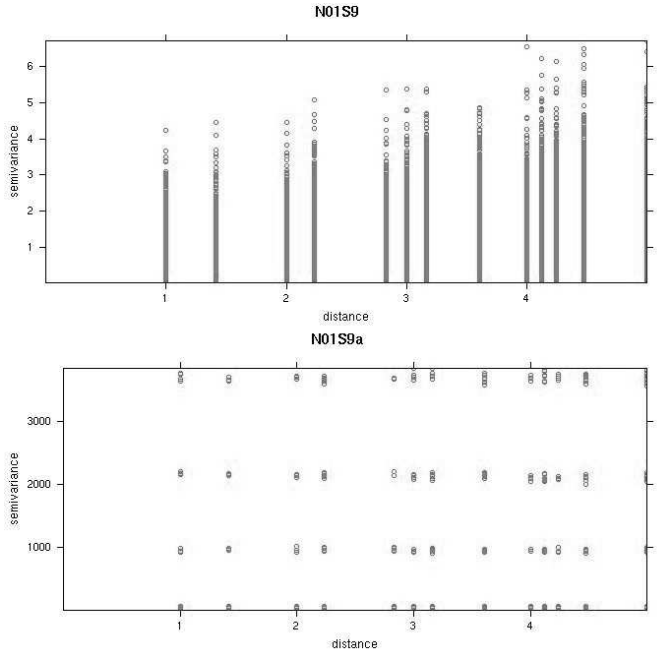


Figura 4.6: Confronto della variogram cloud per il dataset di riferimento N01S9 e quello modificato con 3 punti up e 3 punti down; per maggiori dettagli, si faccia riferimento al testo.

È stata quindi definita una ulteriore serie di dataset per valutare in maniera più sistematica l'eventuale influenza sul variogramma di punti di tipo up, down o della presenza di entrambi; una descrizione dettagliata di tali dataset è riportata nella tabella 4.2.

Per ognuno dei nuovi dataset sono stati calcolati i variogrammi (tradizionali e omnidirezionali, e anche normalizzati alla varianza a priori) e posti a confronto con quello di riferimento; da quanto riportato in figura 4.7 e da ulteriori analisi più dettagliate e specifiche, si può concludere che:

- γ_{N01S9} e γ_c risultano praticamente identici nella forma, con γ_c leggermente shiftato verso l'alto — perfettamente in linea con quanto ci si aspetta in relazione alla varianza di N01S9c rispetto a quella di N01S9 [cfr. tab. 4.2] e con quanto discusso in §4.1.4;

- γ_a e γ_b risultano praticamente identici, con effetto di shift in linea con le rispettive varianze [cfr. tab. 4.2]; la forma non è però sovrapponibile a quella di γ_{N01S9};

- γ_d e γ_e risultano sovrapposti a γ_{N01S9};

- γ_f e γ_g hanno forme diverse (effetto più evidente su larga scala che non su piccola); l'effetto di shift è in linea con la previsione basata sul valore della varianza a priori riportata in tab. 4.2;

- per quanto riguarda i variogrammi normalizzati alla varianza a priori, sembra di poter concludere che quelli caratterizzati dalla presenza di punti di tipo up manifestino range

dataset	*descrizione*	*varianza*
a	3 punti up e 3 punti down	2.21
b	solo gli up di a	2.19
c	solo i down di a	0.90
d	N01S9 senza gli up di a (3 punti in meno)	0.88
e	N01S9 senza gli up e down di a (6 punti in meno)	0.87
f	3 nuovi punti up (diversi da quelli di a)	2.72
g	f con i 3 up di a	4.04

Tabella 4.2: Descrizione dei dataset impiegati per lo studio sistematico dell'influenza di punti anomali sul variogramma; la varianza per il dataset di riferimento N01S9 è pari a 0.88.

mediamente più corti e nugget più elevati — sempre rispettando l'effetto di shift legato alla varianza; si evidenzia inoltre una maggior rumorosità rispetto al variogramma di riferimento.

A conclusione di questa ulteriore fase di analisi, credo si possa affermare che

> *l'eventuale presenza di* hot spots *manifesta la sua influenza sull'intero variogramma (larga scala) e non solo sull'entità dell'effetto nugget (corta scala); punti di tipo* up *possono eventualmente portare a variazioni nella forma del variogramma (rumorosità) rispetto a quello privo di tali punti.*

4.2.1 Influenza dell'entità dell'anomalia rispetto alla variabilità globale del fenomeno

L'idea di base che ha guidato questa nuova fase dell'analisi è stata quella di introdurre dei punti anomali andando nello specifico a controllare l'*entità* di tale anomalia, a dire introducendo punti il cui valore numerico risulti:

- *interno* al campo di variabilità globale del dataset di riferimento;
- *esterno* al campo di variabilità globale del dataset di riferimento;

in quest'ottica, ho deciso di limitare l'analisi a una singola porzione del dataset completo N01S9, ovvero ai primi (30×20) punti della griglia; identifico questo nuovo dataset con 'N02'.

dataset N02

In linea con quanto fatto nelle fasi precedenti, modifico il dataset N02 introducendo:

- 2 punti anomali interni (dataset *a*) — per uno, ho aumentato il valore di concentrazione, per l'altro, diminuito;
- 2 punti anomali esterni, nelle stesse localizzazioni di a (dataset *b*).

dataset N02a

dataset N02b

Anche in questo caso, ho determinato la variogram cloud (su range limitato) per i singoli dataset e anche i relativi variogrammi (su tutto il range); da quanto riportato in figura 4.8, si può affermare che:

- per quanto riguarda il dataset N02a, la presenza di punti anomali interni al campo di variabilità si manifesta su corta scala, come mostrato dalla regione evidenziata nel grafico corrispondente; i due punti anomali interni sono correttamente identificati da un'analisi interattiva delle coppie che caratterizzano questa regione della variogram cloud;

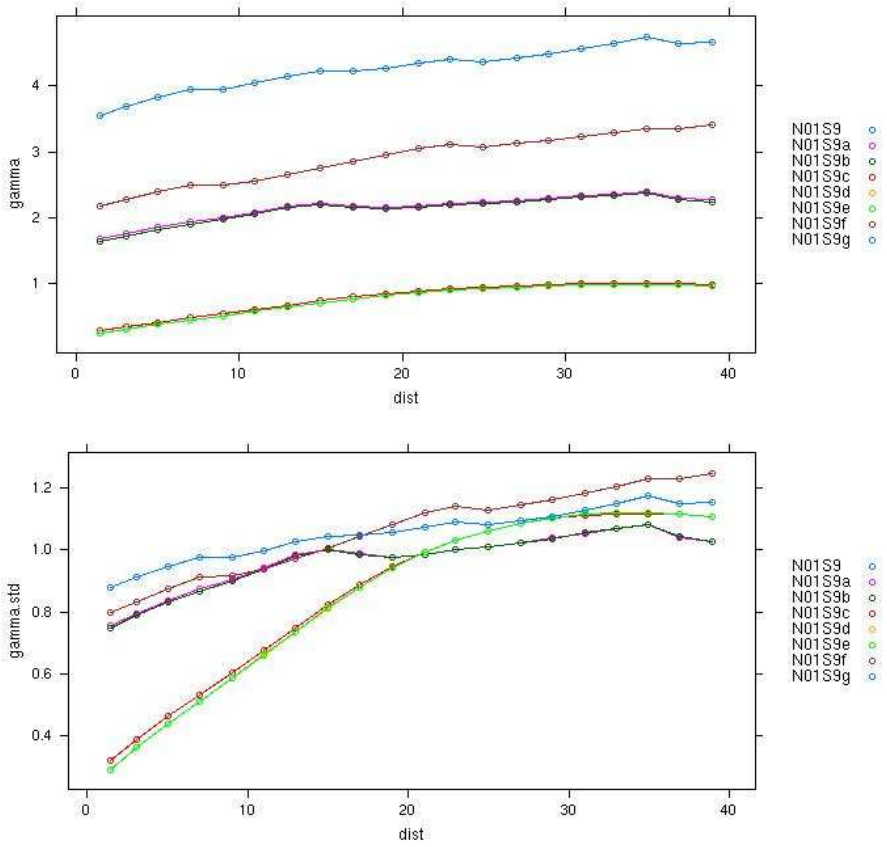


Figura 4.7: Variogrammi e variogrammi standardizzati al variare del dataset impiegato; per la descrizione di tali dataset, si faccia riferimento alla tabella 4.2; si tenga presente che il colore blu è riferito al N01S9g, mentre N01S9 risulta essere sovrapposto a N01S9c.

- per quanto riguarda il dataset N02b, sono ben evidenti le due strisce orizzontali che caratterizzano in maniera uniforme l'intero range della variogram cloud; ancora, l'identificazione interattiva dei due punti anomali esterni è corretta.

Concludendo, mi sembra di poter affermare che:

i) *nel caso siano presenti anomalie il cui valore numerico si colloca all'*interno *del campo medio di variabilità che caratterizza il dataset completo, queste si rendono visibili nella variogram cloud solo su piccola scala;*

ii) *nel caso siano presenti anomalie il cui valore numerico si colloca all'*esterno *del campo medio di variabilità che caratterizza il dataset completo (possibili outliers), queste si rendono visibili come una sorta di* striscia orizzontale *che coinvolge l'intero range della variogram cloud e la cui posizione lungo l'asse y risulta proporzionale all'entità dell'anomalia stessa.*

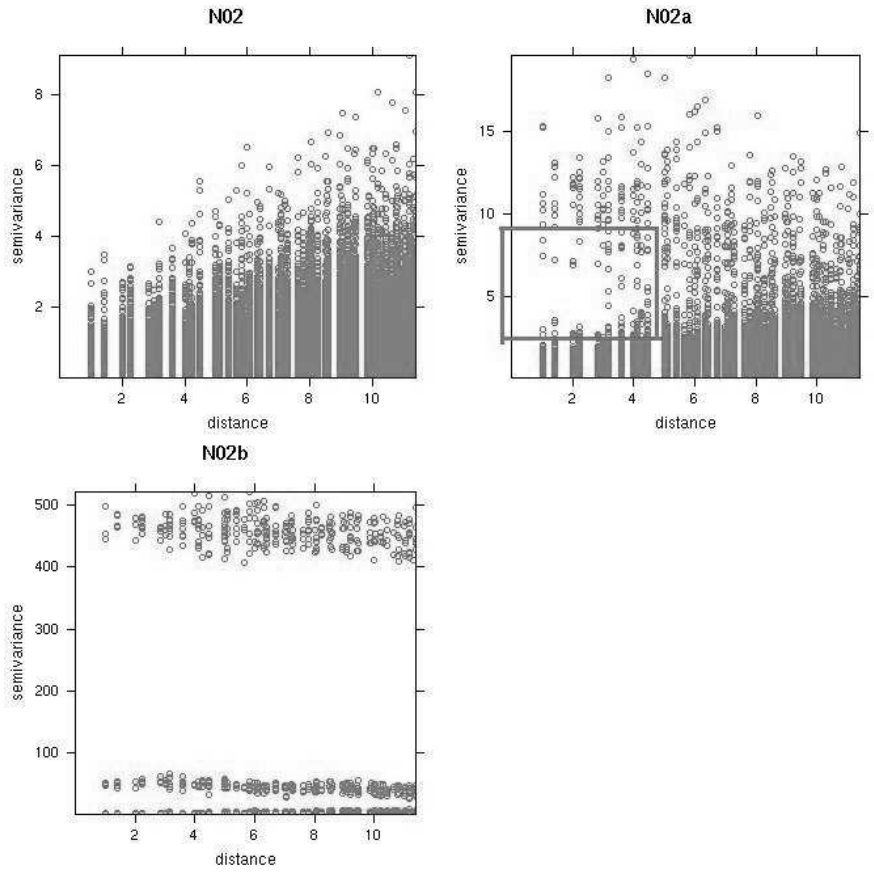


Figura 4.8: Variogram cloud per i dataset indicati nella strip posta sopra ogni grafico e descritti in §4.2.1; per una descrizione dei risultati, si faccia riferimento al testo.

4.3 Analisi dell'influenza delle caratteristiche della rete di monitoraggio

In quest'ultima fase di indagine, ho deciso di condurre delle ulteriori simulazioni con lo scopo di valutare l'eventuale influenza del *tipo di campionamento* sulla presenza di effetto nugget e sulla relativa entità: verranno perciò condotti campionamenti differenti del fenomeno simulato per valutare se e come varia il corrispondente variogramma "sperimentale" rispetto a quello teorico di riferimento [cfr. eq. (4.2)].

La parte computazionale è stata svolta in ambiente `R`{44} ricorrendo ai packages *gstat*{41} e *lattice*{45} e a `Geostat Office`© per le analisi specifiche condotte sulle proprietà della rete di monitoraggio (§4.3.1).

Per fare questo, è stata generata una griglia regolare di (100×60) punti su cui è stato simulato un fenomeno secondo il seguente modello di variogramma:

$$\gamma(h) = 4 + 40 \cdot \exp\left(\frac{h}{20}\right) \tag{4.2}$$

Anche in questo contesto sono state condotte differenti simulazioni, e scelta quella che meglio ha riprodotto:

- il modello di variogramma (4.2), soprattutto nei pressi dell'origine;
- i parametri statistici di media e varianza imposti all'intero fenomeno.

Ho quindi campionato il fenomeno simulato ricorrendo a differenti tipi di reti di monitoraggio (MN), variando sia il numero di punti di campionamento, sia il tipo di campionamento stesso, secondo queste modalità:

MN

numero di punti : le varie MN verranno identificate con delle lettere nel seguente modo:

- a: 600 punti
- b: 300 punti
- c: 2000 punti
- d: 738 punti

tipo di campionamento : per ogni tipo di MN (a,b,c) ho impiegato tre differenti tipi di campionamento, e in particolare sono stati utilizzati:

- 3 differenti campionamenti di tipo random, identificati con 'rnd'
- 2 differenti campionamenti di tipo non-aligned[8], identificati con 'nal'
- 1 campionamento di tipo regular, identificato con 'reg'

La MN d è stata costruita, per quanto possibile, in modo da rispettare la distribuzione spaziale che caratterizza quella reale dei campionamenti condotti in Alto Adige; un esempio relativo alla MN a e alla MN d è riportato in figura 4.9.

Per ognuna delle MN descritte, sono stati calcolati i variogrammi "sperimentali" e messi a confronto con quello di riferimento, dato dall'equazione (4.2). I risultati ottenuti sono riportati in figura 4.10: non mi sembra si possa notare nulla di significativo in relazione all'influenza del tipo di campionamento sul valore dell'effetto nugget; si evince invece come all'aumentare del numero di punti che costituiscono la rete di monitoraggio, le fluttuazioni statistiche decrescano sensibilmente e i vari $\gamma(h)$ tendano a quello di riferimento. Sempre concentrando l'attenzione sul comportamento su piccola scala, anche per $\gamma_{\mathrm{d}}(h)$, se messo in relazione con $\gamma_{\mathrm{a}}(h)$ e con quello di riferimento, non si apprezza alcuna differenza significativa.

In base a quanto appena mostrato e discusso, credo di poter concludere questa ulteriore fase dell'analisi affermando che

> *eventuali differenze sui variogrammi al variare del* tipo *di campionamento si rendono significative su larga scala, ma molto meno si piccola — ovvero, nei pressi dell'origine; il tutto risulta comunque significativamente legato al numero di punti che caratterizza le rete di monitoraggio — a dire, alle fluttuazioni statistiche legate alla determinazione del variogramma stesso.*

[8] Questo tipo di campionamento si colloca, per quanto riguarda la regolarità, tra uno di tipo random e uno di tipo regular.

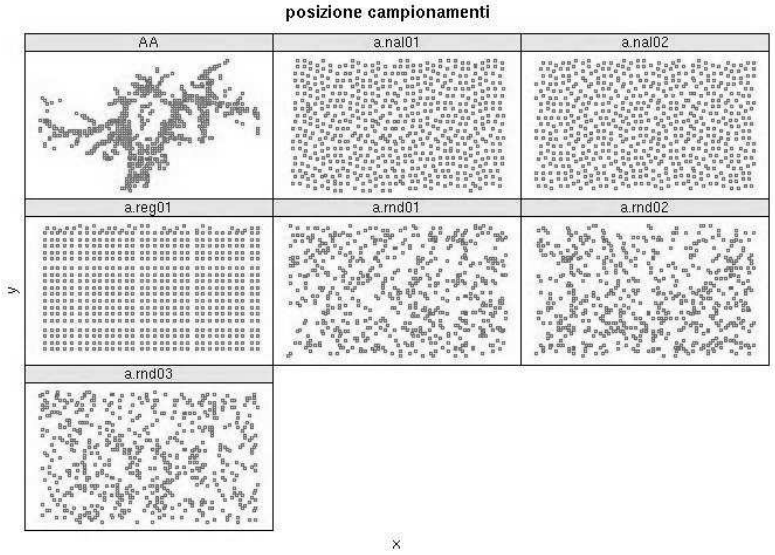


Figura 4.9: Esempio di distribuzione spaziale dei campionamenti impiegati per lo studio dell'influenza della MN sul variogramma al variare del tipo di campionamento; per maggiori dettagli, si faccia riferimento al testo; la distribuzione spaziale "AA" è quella relativa alla MN *d*, che rispecchia, per quanto possibile, la reale distribuzione dei campionamenti del dataset reale.

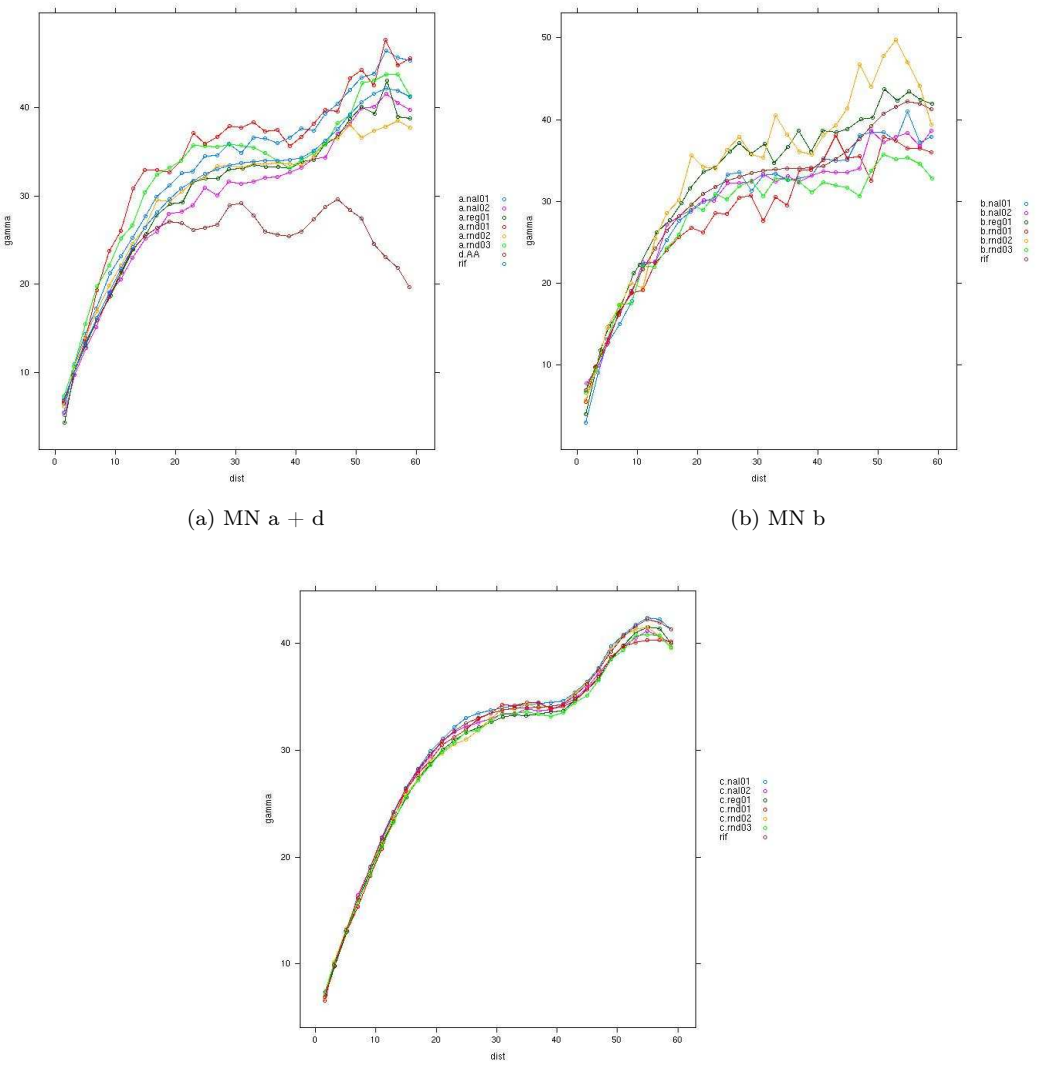
4.3.1 Confronto quantitativo delle differenti reti di monitoraggio

Per valutare in maniera quantitativa e oggettiva la differenza tra le varie reti di monitoraggio impiegate nelle analisi appena descritte, ne ho calcolato la dimensione frattale [cfr. §B.1.1, pag. 221] ricorrendo a box-counting con 20 step; i risultati ottenuti sono i seguenti:

	600 punti	*300 punti*	*2000 punti*
nal01	1.995	1.962	2.000
nal02	1.996	1.963	2.000
reg01	1.988	1.961	2.000
rnd01	1.937	1.928	1.999
rnd02	1.942	1.937	1.999
rnd03	1.941	1.946	1.999

Come si può notare, non ci sono differenze apprezzabili tra le varie MN[9] e risulta inoltre ragionevole il fatto che i campionamenti di tipo random (rnd) abbiano valori leggermente inferiori rispetto agli altri; queste conclusioni sono del resto in linea con quanto è possibile dedurre anche con una semplice occhiata alla distribuzione spaziale dei campionamenti, come quelli riportati in figura 4.9. Per completezza, la dimensione frattale della MN d (quella che rispecchia il campionamento adottato in Alto Adige) è pari a 1.696: non sorprende che il valore sia inferiore rispetto alle altre, molto più regolari nello spazio bi-dimensionale, mentre è soddisfacente il fatto che non risulti molto distante da 2.

[9] Tutte adatte a un campionamento di tipo 2D, come è ragionevole aspettarsi!



(a) MN a + d

(b) MN b

(c) MN c

Figura 4.10: Variogrammi "sperimentali" al variare della MN e del tipo di campionamento: risulta evidente come all'aumentare del numero di punti di campionamento, le fluttuazioni statistiche si riducano; per maggiori dettagli e per la descrizione delle MN, si faccia riferimento al testo.

Sempre ricorrendo a `Geostat Office`©, ho calcolato anche i diagrammi di Morishita per le varie MN: anche questa analisi non ha messo in luce significative differenze, e gli andamenti ottenuti sono in perfetto accordo con quanto previsto dalla teoria a seconda del tipo di campionamento (regolare o meno). Per quanto riguarda invece la MN d, il diagramma parte da valori elevati per stabilizzarsi poi attorno all'unità: anche questo è in linea con la previsione teorica e indica, come ci si poteva del resto aspettare, che la rete di monitoraggio risulta affetta da possibili fenomeni di clustering.

4.4 Conclusioni

Come risulta evidente dalla figura 4.1, il variogramma sperimentale che caratterizza il dataset di riferimento è caratterizzato da un effetto nugget estremamente elevato. In virtù del lavoro di controllo e correzione dei dati che è stato in precedenza condotto su tale dataset [cfr. {42}], si crede che un valore così elevato non si possa imputare a errori di misura, inserimento o posizionamento delle varie misure condotte; ancora, non sembra verosimile che la variabilità su corta scala risulti tale da spiegare in maniera esaustiva questa indesiderata caratteristica del variogramma sperimentale.

Con lo scopo di individuare una ragione plausibile e fondata in grado di rendere conto dell'elevato valore per l'effetto nugget, si è deciso di indagare nello specifico alcune situazioni anomale[10] che caratterizzano il dataset sperimentale, nonché di valutare l'eventuale influenza delle caratteristiche della rete di monitoraggio (principalmente, la sua disomogeneità) sul parametro in esame. Entrambi questi aspetti sono stati esaminati sia ricorrendo ai dati sperimentali, sia conducendo delle simulazioni ad hoc in modo da poter controllare i parametri che nelle precedenti analisi sperimentali sono emersi come significativi (o quantomeno presunti tali).

In base alle numerose e diverse analisi condotte, a conclusione di questa parte del lavoro d'indagine si può affermare che:

- in relazione al dataset sperimentale, la presenza di un elevato effetto nugget non si può legare alla presenza di *isolati punti anomali*, ma sembra piuttosto manifestarsi come una *caratteristica intrinseca* del fenomeno stesso: infatti, anche eliminando, sulla base di opportune e mirate analisi, tali punti, il valore dell'effetto nugget non viene ridotto;
- lo studio della *variogram cloud* risulta utile ed efficace per l'identificazione di situazioni anomale del tipo "valore elevato di concentrazione con vicinaggio caratterizzato da valori bassi", mentre non è in grado di rendere evidente la presenza di situazioni opposte;
- la presenza di *hot spots* manifesta la sua influenza sulla forma dell'intero variogramma sperimentale (larga scala) e non solo sull'entità dell'effetto nugget (corta scala);
- nel caso siano presenti anomalie il cui valore numerico si colloca all'*interno* del campo medio di variabilità che caratterizza il dataset completo, queste possono essere identificate correttamente analizzando la variogram cloud su brevi distanze;
- nel caso siano presenti anomalie il cui valore numerico si colloca all'*esterno* del campo medio di variabilità che caratterizza il dataset completo (anomalie che possono essere identificate come possibili outliers), queste si rendono invece visibili sull'intero range della variogram cloud; visivamente, appaiono come delle *strisce orizzontali* la cui posizione lungo l'asse y risulta proporzionale all'entità dell'anomalia stessa;

[10] Situazioni nelle quali un campionamento manifesta caratteristiche sensibilmente diverse rispetto a quelle degli altri punti che costituiscono il suo vicinaggio, e che risultano invece tra loro piuttosto omogenei.

- per quanto riguarda le caratteristiche della rete di monitoraggio, eventuali differenze sulla forma del variogramma al variare del *tipo* di campionamento (più o meno regolare) si rendono visibili su larga scala, ma non su corta (nei pressi dell'origine): per questo, non sembra quindi che l'eventuale disomogeneità del campionamento del fenomeno possa rendere conto in maniera evidente della presenza di un elevato effetto nugget; inoltre, sembra importante sottolineare come tali conclusioni risultino sensibilmente legate al numero di punti che costituiscono la rete di monitoraggio — ovvero, alle fluttuazioni statistiche legate alla determinazione del variogramma stesso: al crescere delle dimensioni del dataset, le differenze tra i vari tipi di campionamento si riducono.

CAPITOLO 5

Analisi variografica: pre-selezione delle classi

A seguito dei risultati ottenuti dalle analisi riportate nel capitolo 9, l'idea che ha ispirato quanto descritto in seguito è stata quella di condurre delle analisi variografiche in funzione delle variabili secondarie che accompagnano il valore di concentrazione; in questo modo si ottengono dei subset di dimensioni ridotte, in linea di principio più omogenei *rispetto al dataset globale, con la possibilità di valutare se, fissata una variabile secondaria, le varie classi che la caratterizzano diano origine a variogrammi differenti.*
La speranza è stata quella di riuscire a individuare variabili e/o classi che potessero rendere la struttura del variogramma tradizionale *più* leggibile *rispetto a quella che si ottiene dal dataset globale, che risulta caratterizzata da un elevato effetto nugget e un andamento molto rumoroso.*

La parte computazionale è stata svolta in ambiente `R`{44} ricorrendo anche ai packages *GeoXp*{51}, *lattice*{45}, *geoR*{30}, *gstat*{41} e *maptools*{35}; ulteriori analisi spaziali e verifiche sono state eseguite con lo strumento GIS `QGis`{1} e con il tool esplorativo `Mondrian`{50}. Le analisi relative alle reti di monitoraggio si sono invece avvalse delle possibilità computazionali offerte da `Geostat Office`©.

5.1 Descrizione del dataset utilizzato

In analogia a quanto descritto nel paragrafo §9.2, con lo scopo di poter eventualmente condurre dei confronti con i risultati ottenuti nel capitolo 9, ho deciso di considerare come dataset di partenza sia quello di riferimento descritto in §1.3, che si compone delle misure georeferenziate di radon indoor riferite a una esposizione tipica della durata di un semestre, sia quello che si compone delle misure convertite invece a media annuale mediante opportuni fattori di correzione {53}.

dataset 'inverni' e 'annuali'

In entrambi i casi, che saranno identificati come 'inverni' e 'annuali' ho preparato i dati in

questo modo:

- rimozione della classe *no data* per la variabile *esposizione*;
- rimozione delle classi *prefabbricato* e *corridoio* per la variabile *tipo locale*;
- estrazione delle sole misure condotte al *piano zero*;
- estrazione delle sole misure per le quali sono disponibili *tutti* i valori per le variabili secondarie prese in considerazione nell'analisi, che nello specifico sono:
  - coordinate spaziali;
  - utilizzo;
  - tipo locale;
  - tipo di costruzione;
  - classe data di costruzione;
  - qualità degli infissi;
  - contatto;
  - esposizione.

Infine, per quanto riguarda il dataset inverni, come suggerisce il nome stesso ho considerato solamente le misure condotte nel semestre invernale; dopo la fase di preparazione appena descritta, il dataset inverni si compone di **1821** cases, il dataset annuali di **1827**.

5.2 Analisi variografica esplorativa

Inizialmente, per entrambi i dataset operativi ho condotto delle analisi variografiche esplorative al variare della distanza massima tra le coppie di punti e della risoluzione spaziale (lag-step), ricorrendo al *variogramma tradizionale* e *omnidirezionale* $\gamma(h)$ [cfr. eq. (B.24), pag. 230]; l'intento è stato quello di trovare un insieme di parametri che rendessero la struttura spaziale leggibile, e inoltre quello di osservare se i due dataset mostrassero già in questa fase iniziale analogie o differenze.

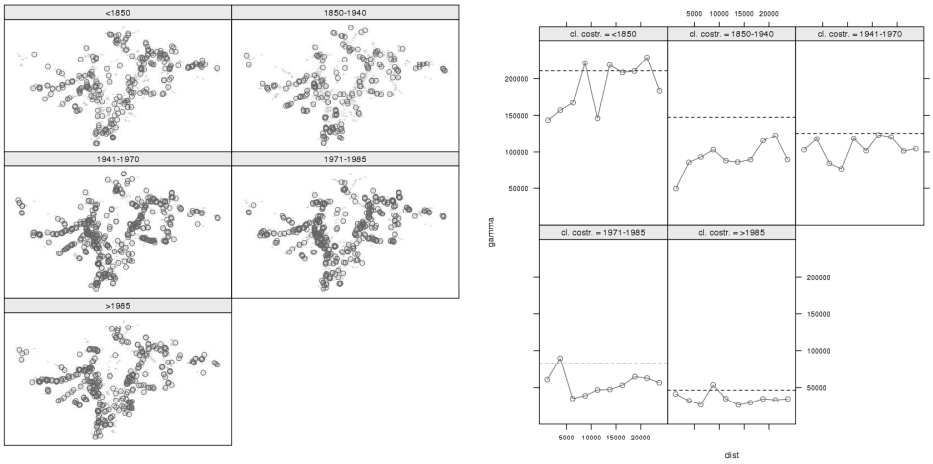
Operativamente, si nota un primo dominio di stazionarietà attorno ai 40 km (in analogia a quanto già riscontrato in lavori precedenti, come ad esempio in {42}) e nessuna differenza significativa tra inverni e annuali. Per le successive analisi esplorative in funzione delle singole variabili secondarie, alla luce di quanto ottenuto in questa fase preliminare, ho quindi deciso di limitare la distanza massima a 25 km e ricorrere a un lag-step pari a 2.5 km.

Sottolineo che in questo modo, i confronti per le variabili secondarie *non* dipendono da questi parametri operativi — a dire, fissata una variabile secondaria e analizzati i variogrammi in funzione delle singole classi che la caratterizzano, l'intento non è quello di ricercare la leggibilità del variogramma stesso, quanto piuttosto le eventuali differenze legate alla maggior *uniformità* dei subset.

5.2.1 Analisi per le singole variabili secondarie

Per ognuna delle variabili secondarie riportate in §5.1, si è analizzata dal punto di vista grafico:

a) la distribuzione spaziale dei campionamenti al variare della singola classe [cfr. fig. 5.1a], in modo da escludere eventuali subset caratterizzati da fenomeni di clustering eccessivi che potrebbero compromettere o quantomeno rendere meno affidabile l'analisi stessa;



(a) post plot (b) variogrammi

Figura 5.1: Esempio di visualizzazioni utilizzate per lo studio dei variogrammi al variare delle singole classi di una fissata variabile secondaria, in questo caso *classe data costruzione*; le distanze sono riportate in metri e la linea orizzontale in (b) rappresenta la varianza a priori dello specifico subset; il dataset di riferimento è 'inverni'.

b) l'andamento dei variogrammi categorizzati in funzione della singola classe [cfr. fig. 5.1b], mantenendo fissa la scala per l'asse y e confrontando anche con la varianza a priori propria di ogni singolo subset;

Un esempio di quanto descritto, per la variabile *classe data costruzione* e il dataset inverni, è riportato nella figura 5.1. Questo tipo di analisi è stata condotta per entrambi i dataset operativi.
Da questo tipo di analisi, si può concludere che:

- *globalmente, i due dataset operativi danno risultati del tutto analoghi;*
- *tutti i variogrammi si stabilizzano su un valore di sella attorno al valore della varianza a priori dello specifico subset: la presenza di un primo dominio di stazionarietà identificata per il dataset completo è confermata anche per ogni singolo subset (classe di ogni variabile secondaria);*
- *le variabili (feature) al cui interno si trovano classi per le quali si evidenziano variogrammi con una struttura più leggibile rispetto a quella del dataset di riferimento sono le stesse che sono emerse come significative — dal punto di vista del potere predittivo sul valore di concentrazione — dall'analisi di* Feature Selection [cfr. cap. 9], *e nello specifico*[1]*:*
  - tipo di costruzione*;*
  - contatto*;*
  - esposizione*.*

[1] A queste, a rigore, andrebbe aggiunta anche la variabile *classe data costruzione*, ma come sarà discusso più avanti, nel paragrafo §5.3.2, l'informazione portata da questa variabile (classe *<1850*) è per così dire "contenuta" in quella portata da *tipo di costruzione* (classe *sassi*).

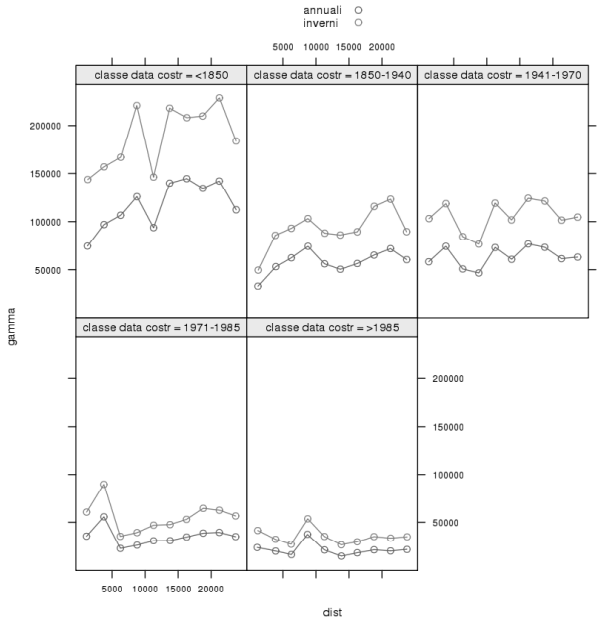


Figura 5.2: Esempio di visualizzazione grafica impiegata per un confronto diretto tra i variogrammi ottenuti per i due dataset operativi; le distanze sono riportate in metri.

5.3 Confronti tra 'inverni' e 'annuali'

A questo punto, ho ritenuto opportuno condurre delle analisi più specifiche sulle eventuali differenze variografiche tra i due dataset operativi impiegati, sempre al variare delle variabili secondarie di riferimento e delle relative classi. Per fare questo, i variogrammi relativi a inverni e annuali, fissata variabile secondaria e relativa classe, sono stati posti a confronto graficamente mantenendo costante la scala dei valori y (valore del variogramma); un esempio di tale visualizzazione è riportato nella figura 5.2.

Questo tipo di analisi ha messo in luce che:

- *al variare della variabile secondaria e delle relative classi, gli andamenti dei variogrammi per i due dataset operativi sono praticamente* sovrapponibili*;*
- *i $\gamma(h)$ relativi a inverni sono costantemente shiftati verso l'alto rispetto ad annuali: questo risulta in linea con la varianza a priori che caratterizza i due dataset (e i relativi subset) e con quanto discusso in dettaglio nel paragrafo §4.1.4 a pagina 47; globalmente, infatti, inverni ha una varianza a priori pari a $\sigma^2_{inv} = 115241$, mentre per annuali $\sigma^2_{ann} = 68583$;*
- *annuali risulta essere, da un punto di vista statistico, un dataset più 'smooth' rispetto a inverni, ma questa caratteristica (eccettuato il fenomeno di shift)* non *sembra avere effetti rilevanti sulla forma dei variogrammi sperimentali.*

D'ora in poi, se non espressamente detto, tutte le analisi sono sempre state condotte in maniera parallela su entrambi i dataset operativi inverni e annuali, senza che queste abbiano portato alla

luce differenza significative — se non che mediamente annuali risulta essere meno rumoroso di inverni: la procedura stessa di conversione del valore di concentrazione da misura semestrale (reale) ad annuale (mediante opportuno fattore correttivo) [cfr. quanto riportato in {42}] attenua infatti la presenza di outliers, rendendo quest'ultimo dataset statisticamente più "morbido".

5.3.1 Analisi variografiche specifiche

A questo punto, limitando l'analisi alle variabili secondarie che hanno manifestato delle classi "interessanti" dal punto di vista variografico [cfr. §5.2.1], ho condotto delle analisi variografiche di tipo esplorativo, ricorrendo sempre al *variogramma tradizionale* e *omnidirezionale*, alla ricerca dei parametri operativi di massima distanza e lag-step che rendessero la struttura della correlazione spaziale il più leggibile e strutturata possibile.

Dall'analisi di questa ricca serie di variogrammi, si può concludere che

> *si è riscontrata la presenza di classi per le quali il variogramma sperimentale manifesta un insperato quanto inatteso* basso effetto nugget *e un buon andamento per piccole distanze — in linea con un andamento teorico di salita.*
> *Nello specifico, si tratta della classe* `sassi` *per la variabile* tipo costruzione *e la classe* `<1850` *per la variabile* classe data costruzione.

Ho quindi deciso di approfondire la questione conducendo dei confronti (sia variografici che statistici) più mirati tra inverni e annuali 'categorizzati' per specifiche classi rispetto ai relativi dataset 'completi', ovvero composti di tutti i valori di concentrazione disponibili.

Questo ulteriore tipo di analisi ha permesso di concludere che, rispetto ai corrispondenti dataset completi,

> *categorizzare i dataset operativi per le classi* SASSI *e* $<$1850 *si ripercuote in un notevole* abbassamento *dell'effetto nugget e in un sensibile* miglioramento *dell'andamento del variogramma sperimentale entro una distanza pari a circa 7 km* [cfr. quanto riportato in fig. 5.3.]

5.3.2 Analisi specifiche su 'sassi' e '<1850'

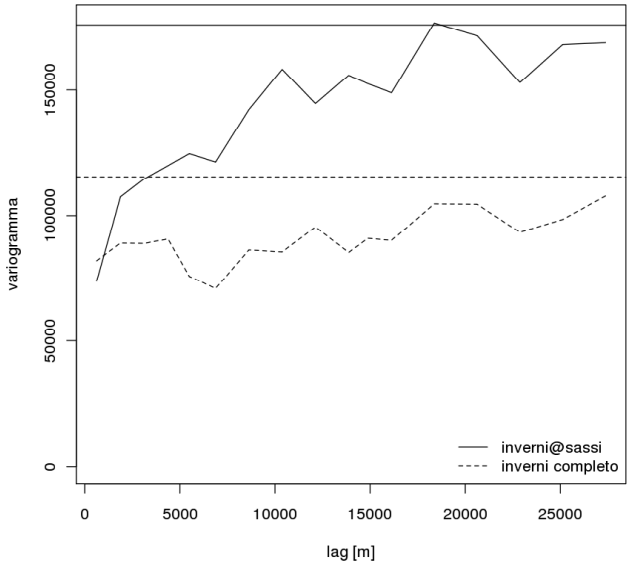
Ho successivamente provato a mettere a confronto i due subset relativi alle due classi per le quali i variogrammi sperimentali $\gamma(h)$ manifestano una buona struttura secondo differenti punti di vista, con lo scopo di individuare una qualche peculiarità di questi rispetto ai dataset completi in grado di rendere conto dell'abbassamento dell'effetto nugget.

Riporto brevemente i risultati ottenuti:

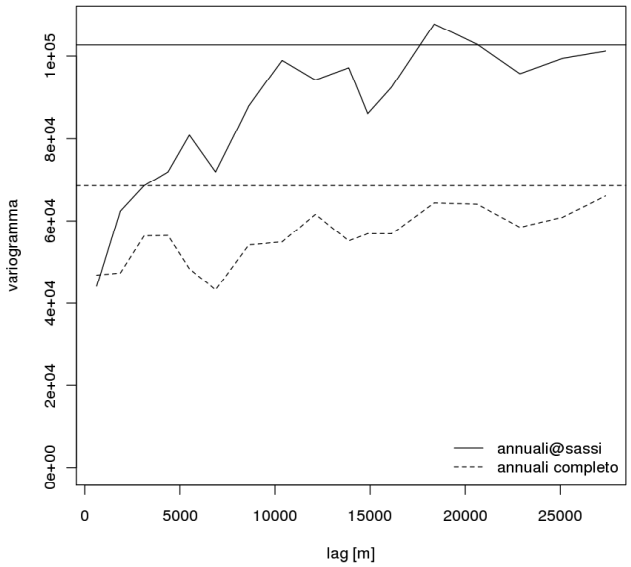
correlazione tra le due classi : applicando le tecniche descritte nel capitolo 7, ho riscontrato una forte correlazione[2] tra le due classi prese in esame — anche alla luce di analisi della distribuzione spaziale per i relativi campionamenti, sembra di poter affermare che studiare il subset *<1850* equivalga nella pratica a studiare un sotto-campionamento di *sassi*.

analisi variografica specifica : focalizzando l'attenzione su corta scala — primi 10-15 km — ho deciso di valutare nel dettaglio il comportamento del variogramma introducendo dei lag-step non uniformi; gli andamenti sono molto buoni e, cosa interessante, emerge una

[2] Il valore del residuo di Pearson [cfr. eq. 7.2, pag. 102] è pari a $\simeq 19$, mentre per le altre coppie i valori massimi sono mediamente ≤ 5



(a) inverni



(b) annuali

Figura 5.3: Variogrammi sperimentali per i due dataset operativi, determinati per la classe *sassi* della variabile *tipo costruzione* e messi a confronto con il relativo dataset completo; per maggiori dettagli, si faccia riferimento a quanto discusso in §5.3.2; le linee orizzontali rappresentano il valore della varianza a priori.

	N	*min*	*max*	*varianza*	*media*	*mediana*
annuali	1827	2	2924	68583	183	94
annuali@sassi	573	9	2918	102682	245	133
annuali@<1850	272	12	2918	123108	253	143
inverni	1821	3	3794	115241	233	120
inverni@sassi	555	12	3787	176026	315	171
inverni@<1850	263	15	3787	210141	320	177

Tabella 5.1: Alcuni parametri statisti relativi ai dataset riportati nella prima colonna.

seconda struttura[3] (la cui presenza può essere intuita anche osservando la figura 5.3a) nei primi 5-7 km, del tutto mascherata nei variogrammi relativi ai dataset completi;

presenza di punti anomali : entrambi i subset in esame hanno al loro interno due *punti anomali*, ovvero punti con un valore di concentrazione molto elevato circondato da punti con valori notevolmente inferiori; nonostante questo, l'effetto nugget è basso, a corroborare le conclusioni che a tal proposito sono state ottenute nel capitolo 4 (i punti anomali individuati in questa analisi sono infatti tra quelli descritti nel dettaglio nel paragrafo §4.1.3): l'effetto nugget elevato *non* è legato alla presenza di singoli punti anomali;

analisi statistiche : dai risultati di semplici e convenzionali analisi statistiche condotte tanto sui subset che sui relativi datatset completi, non emerge nulla di significativo [cfr. tab. 5.1]; gli istogrammi risultano molto simili, e le code superiori delle distribuzioni di probabilità — importanti in un contesto di valutazione dei rischio associato all'esposizione a valori elevati di concentrazione — non vengono stravolte dalla procedura di categorizzazione; ancora, è evidente anche in questo caso come annuali sia un sorta di versione smooth di inverni.

5.4 Analisi spaziali specifiche

Da quanto emerso fino a questo punto, non è chiaro quale sia l'origine del basso effetto nugget, poiché le analisi condotte non sono state in grado di portare alla luce significative differenze tra i subset categorizzati e i relativi dataset completi. Per questo, l'ipotesi che si è proposta è stata che la natura di questi risultati fosse da ricercare in una qualche *caratteristica spaziale* dei subset; con questo obiettivo, ho deciso di estendere l'analisi anche alla classe `nord` per la variabile *esposizione*, in quanto anche il variogramma che le compete manifesta un basso valore per l'effetto nugget — anche se poi, per distanze maggiori, la struttura diventa piuttosto rumorosa e meno leggibile rispetto a quelle considerate e analizzate in precedenza.

Accanto ad analisi visuali condotte sia in ambiente `R`{44} che con lo strumento GIS `QGis`{1}, ho utilizzato `Geostat Office`© per confronti di tipo quantitativo relativi alle reti di monitoraggio [cfr. §B.1.1, pag. 219] che caratterizzano i vari subset e dataset completi presi in esame. Di seguito, le conclusioni che si possono trarre da questa serie di analisi:

istogrammi delle distanze tra coppie : non si notano differenze tra inverni e annuali, e gli istogrammi hanno andamenti molto simili in funzione delle variabili secondarie e delle relative classi, evidenziando un picco attorno ai 40-50 km;

[3] Sembra che si possa identificare quindi un ulteriore sotto-dominio di stazionarietà; tuttavia, non avendo trovato una spiegazione fondata per questo tipo di struttura, nelle analisi successive non ho ritenuto opportuno introdurre modellizzazioni che ne tenessero conto.

dimensione frattale : questo parametro è stato calcolato mediante box-counting [cfr. {31, pagg. 35–38}] e i grafici — una volta corretti per le differenti numerosità — utilizzati per la sua determinazione mostrano tutti andamenti molto simili in funzione di dataset e subset; in tutti i casi, la dimensione frattale ha valori tipici pari a $\simeq 1.9$, valore ottimale per la descrizione di un fenomeno nello spazio bi-dimensionale;

indice di Morishita : anche nel caso di quest'ultimo parametro considerato [cfr. eq. (B.1), pag. 220], non si notano differenze significative; in tutti i casi, i grafici mettono in luce la presenza di reti di monitoraggio affette da clustering — risultato che del resto non stupisce.

Concludo sottolineando che purtroppo, anche dopo aver rivisto tutte le analisi precedenti cercando di ricorrere a differenti punti di vista,

> *la serie di analisi condotte* non *è stata sufficiente ed efficace nel portare alla luce un qualche parametro significativo in grado di rendere conto del* basso effetto nugget *mostrato dalle classi* `sassi` *e* `<1850` *rispetto al dataset completo.*
> *Riassumendo, la riduzione dell'effetto nugget* non *è imputabile a una maggior* uniformità *dei subset, né a effetti legati alla* rete di monitoraggio*, né a effetti legati a indirette procedure di* declustering.

5.5 Ulteriori analisi esplorative e geologia

Poiché, come ampiamente discusso in precedenza, le variabili di tipo antropogenico prese in considerazione non sembrano essere utili per interpretare in modo sicuro il basso effetto nugget mostrato da alcuni subset di dati, ho deciso di estendere l'analisi considerando anche altre variabili secondarie[4] che per ragioni concettuali ho diviso in due gruppi distinti di lavoro, e in particolare:

a) *variabili* non *direttamente legate alle caratteristiche dell'edificio*, quali anno di esposizione, saturazione del dosimetro, tipo di dosimetro, altitudine, pendenza e curvatura;

b) *variabili di tipo geologico*, quali spessore della roccia sottostante, influenza e tipo di filoni e e faglie, fratturazione, PERS, radioattività (rocce e sottosuolo), milonite, sepoltura, distanza filoni e faglie.

Il dataset utilizzato per questa analisi è *inverni* (in questo contesto, il valore di concentrazione non ha rilevanza), al quale sono stati aggiunti i valori per le variabili appena citate; da sottolineare come per le variabili di tipo geologico — gruppo b) — non per tutti i campionamenti siano disponibili questi valori [cfr. §1.3]; l'analisi relativa alla parte geologica sarà pertanto limitata a una parte di inverni, pari a circa il 73%.

Lo scopo di questa analisi è stato quindi quello di valutare se i tre subset individuati in precedenza — tipo di costruzione = sassi, classe data di costruzione = <1850 ed esposizione = nord — manifestassero qualche correlazione con le altre variabili secondarie prese ora in considerazione, oppure mostrassero delle 'prevalenze' o sproporzioni nella distribuzione sulle classi delle altre variabili secondarie rispetto alla distribuzione completa — a dire, se si "categorizza" per la classe <1850, si nota nel barchart per la variabile 'tipo di costruzione' che tali campionamenti vanno

[4] Per una descrizione completa di queste ulteriori variabili disponibili, si faccia riferimento a quanto riportato nel paragrafo §1.3, a pagina 7.

quasi tutti a cadere nella classe sassi, come del resto ci si poteva aspettare da quanto discusso fin qui.

Questo tipo di analisi visuali è stata svolta con il tool esplorativo `Mondrian`{50}, che consente di visualizzare contemporaneamente più rappresentazioni grafiche e statistiche per le variabili di interesse, consentendo tra l'altro di selezionare una particolare classe ed evidenziare come la selezione si distribuisce nelle altre rappresentazioni attive — avendo così modo di valutare rapidamente eventuali sproporzioni o 'prevalenze' anomale.

Per quanto riguarda le variabili del gruppo a), non è emerso nulla di significativo che differenziasse i tre subset in esame rispetto a una qualche classe delle variabili indagate.

Qualche lieve prevalenza è stata invece evidenziata per alcune classi delle variabili del gruppo b), e in particolare tutti e tre i subset mostrano uno sbilanciamento verso una *fratturazione bassa* e un *PERS basso*; sembrerebbe esserci anche qualche ulteriore evidenza di un possibile legame con altre variabili di tipo geologico, come *radioattività* e *sepoltura*, ma la scarsa numerosità e la mancanza di parametri oggettivi mi sembra non garantiscano di poter trarre conclusioni fondate a riguardo.

A conclusione di questa ulteriore analisi, credo si possa affermare che

> *i subset che manifestano un* basso effetto nugget *nel variogramma sperimentale hanno delle caratteristiche comuni in relazione ad alcune variabili di tipo* geologico*, in particolare con fratturazione e PERS, che li differenziano rispetto ai relativi dataset completi.*
> *L'ipotesi che si può quindi proporre è che tali edifici, tipicamente vecchi e poco isolati dall'ambiente circostante, siano in una certa misura* meno *legati alla parte antropogenica del fenomeno e risentano quindi meno anche dei fattori confondenti ad essa legati — risultano per così dire più 'in linea' con la parte* geologica *del fenomeno radon, teoricamente di più facile interpretazione in quanto caratterizzata, almeno in linea di principio, da una maggior continuità spaziale.*

A sostegno di quest'ultima ipotesi, il fatto che gli edifici che appartengono alla classe *sassi* risultino per la maggior parte costruiti prima del 1850, in contatto con il terreno e abbiano degli infissi di scarsa qualità.

5.5.1 Analisi variografica e geologia

Con lo scopo di valutare se l'effetto di abbassamento dell'effetto nugget fosse realmente legato a fattori di tipo geologico, dal dataset di riferimento ho estratto le misure condotte nel semestre invernale e al piano zero, accompagnandole con tutte le informazioni di tipo geologico disponibili (che ricordo, non coprono tutto il territorio dell'Alto Adige, ma solo una fascia centrale). Si ottiene così un dataset operativo composto di **1901** cases.

Ho quindi utilizzato questo ulteriore dataset per calcolare dei variogrammi tradizionali omnidirezionali pre-selezionando alcune variabili di tipo geologico, e concentrando in particolare l'attenzione su quelle che in precedenza avevano mostrato un qualche legame con i subset relativi a 'sassi' e '<1850'. Purtroppo, la scarsa numerosità dei nuovi subset e la loro distribuzione spaziale clusterizzata e/o particolarmente disomogenea non consente di ottenere strutture facilmente interpretabili, né risultati affidabili.

Rimane comunque il fatto che se l'influenza della sola parte geologica fosse prevalente, questa dovrebbe allora manifestarsi in variogrammi caratterizzati da una struttura ben leggibile, cosa che non accade; sembra quindi che la complessità del fenomeno radon indoor rimanga legata a una interazione di non facile interpretazione tra parte geologica e antropogenica.

5.6 Conclusioni

Come più volte messo in evidenza, ad esempio nel capitolo 4, il variogramma sperimentale relativo al dataset operativo risulta affetto da un elevato effetto nugget, la cui origine non può essere collegata in maniera diretta, evidente ed esaustiva alla presenza di errori di misura e/o posizionamento, alla presenza di hot spot o alle caratteristiche di disomogeneità della rete di monitoraggio[5]. Per questo, e anche in parte in relazione a quanto ottenuto dalle analisi descritte nel capitolo 9, si è deciso di condurre delle analisi variografiche suddividendo il dataset operativo in base al valore delle variabili secondarie che accompagnano il singolo valore di concentrazione: in questo modo si ottengono subset di dimensioni ridotte, ma almeno in linea di principio più omogenei per quanto riguarda le caratteristiche delle misure di concentrazione. Lo studio è quindi volto alla ricerca di variabili secondarie in grado di rendere la struttura di correlazione spaziale più leggibile, in particolare in relazione all'entità dell'effetto nugget — il cui valore particolarmente elevato ha inevitabili ripercussioni sulla bontà delle stime che si possono ottenere.

Le analisi sono state condotte su due differenti dataset operativi, opportunamente preparati al fine di eliminare le classi che sono risultate essere numericamente troppo esigue o non rilevanti (in maniera evidente) sul valore di concentrazione; si sono considerate da un lato le misure condotte nel semestre invernale (nessun intervento correttivo sul valore misurato), dall'altro le misure opportunamente convertite a una media annuale.

Alla luce dei risultati ottenuti dalle numerose serie di analisi variografiche e spaziali descritte in questo capitolo, si può concludere che:

- globalmente, in relazione alla forma del variogramma sperimentale, i due dataset operativi danno risultati del tutto analoghi; il dataset composto di misure riferite a media annuale risulta essere, dal punto di vista statistico, più smooth rispetto a quello costituito da misure riferite al semestre invernale; tuttavia questo non sembra avere ripercussioni significative sulla forma dei variogrammi sperimentali;

- fatto interessante, le variabili al cui interno si trovano classi che rendono la struttura variografica più leggibile sono quelle che sono emerse come le più predittive per il valore di concentrazione in base alle analisi di Feature Selection condotte e descritte nel capitolo 9; nello specifico, *tipo di costruzione*, *contatto* ed *esposizione*;

- introdurre una categorizzazione per le classi 'tipo di costruzione = sassi' e 'data di esposizione = <1850' porta a variogrammi sperimentali con un effetto nugget sensibilmente inferiore rispetto a quello che caratterizza le altre classi e/o variabili (nonché ovviamente quello relativo ai dataset globali) e a un miglioramento del comportamento dello stesso per distanze fino ai 7 km circa [cfr. fig. 5.3];

- conducendo analisi specifiche in relazione ai parametri geologici disponibili, i subset che manifestano un *basso effetto nugget* nel variogramma sperimentale hanno delle caratteristiche comuni in relazione ad alcune variabili di tipo *geologico*: si può quindi ipotizzare che tali edifici, tipicamente vecchi e poco isolati dall'ambiente circostante, siano in una certa misura *meno* legati alla parte antropogenica del fenomeno e risentano quindi meno anche dei fattori confondenti ad essa legati — detto in altre parole, risultano per così dire più "aderenti" alla parte geologica del fenomeno radon, teoricamente di più facile interpretazione in quanto caratterizzata, almeno in linea di principio, da una maggior continuità spaziale.

[5] Una trattazione dettagliata di questi aspetti è riportata nel capitolo 4, a pagina 43.

Anche se le diverse analisi condotte non sono state sufficienti per portare alla luce un qualche parametro significativo ed evidente in grado di rendere conto del *basso effetto nugget* mostrato dalle classi 'sassi' e '<1850' rispetto al variogramma relativo ai dataset completi, questo studio ha comunque consentito di escludere che la riduzione dell'effetto nugget sia imputabile in maniera evidente a una maggior *uniformità* dei subset, a effetti legati alla *rete di monitoraggio*, o ancora a effetti legati a indirette procedure di *declustering*.

Parte II

Statistica “Classica”

CAPITOLO 6

Modellizzazione della p.d.f.: la Teoria dei Valori Estremi

Per una corretta identificazione delle cosiddette radon prone areas *(zone a rischio), punto chiave risulta essere una corretta analisi e successiva modellizzazione della distribuzione statistica delle misure di radon indoor. Attualmente, l'approccio più diffuso consiste nel ricorrere a modelli di tipo Normale/log-Normale; tuttavia, questo tipo di assunzione può* non *rivelarsi la più efficace nell'ambito dell'analisi del rischio, in quanto è noto come possa limitare se non addirittura trascurare l'importanza di eventuali outliers, ovvero di quei valori di concentrazione così elevati da poter apparire, almeno in una prima e superficiale fase dell'analisi, apparentemente inspiegabili.*
In questo contesto, la Teoria dei Valori Estremi *sembra configurarsi quindi come un valido e appropriato strumento statistico in grado di prendere in considerazione nella maniera opportuna quegli eventi poco probabili che popolano la coda superiore della distribuzione, e sui quali è bene porre la dovuta attenzione.*

La parte computazionale è stata svolta in ambiente `R`{44}, ricorrendo ai packages *lattice*{45} ed *evd*{49}.

6.1 Brevi richiami teorici

La maggior parte dei metodi statistici concentrano la loro attenzione sul comportamento dei dati sperimentali in relazione alla parte *centrale* della distribuzione di probabilità empirica, dando poca importanza alle code della stessa. Ancora, la filosofia che guida la vasta area della statistica che si occupa di *metodi robusti* di stima è quella per la quale i metodi statistici non dovrebbero essere influenzati in maniera eccessivamente sensibile da valori estremi — siano essi appartenenti alla coda bassa o alta della distribuzione.

Tuttavia, ci sono situazioni per le quali sono proprio i VALORI ESTREMI che giocano la parte più importante del problema/fenomeno che si intende affrontare: è in questo contesto che si colloca la
EVT EXTREME VALUE THEORY, o *Teoria dei Valori Estremi*, che storicamente si fa risalire al primo lavoro di Gumbel {23}, del 1935.

Lo scopo di questa teoria e quello di trovare il modello migliore per la stima della distribuzione di probabilità dei valori estremi; in altre parole, questa teoria ricerca distribuzioni — note come GEV o GPD, [cfr. §6.1.1 e §6.1.2] — adatte alla modellizzazione dei valori estremi, e che nei contesti citati rivestono il ruolo di predominio che ha la distribuzione Normale nella statistica classica, ma le cui code decadono troppo rapidamente per una trattazione adeguata dei valori estremi.

Semplificando, si possono distinguere due principali tipi di modello per i valori estremi[1]:

- il più vecchio insieme di modelli — ovvero la teoria "classica" che trae origine dai primi lavori di Gumbel — è costituito dai cosiddetti BLOCK MAXIMA MODELS: sono modelli per i valori massimi[2] raccolti da una numerosa serie di osservazioni IID (indipendenti e identicamente distribuite); in questo ambito, nasce la distribuzione nota in letteratura come **GEV** [cfr. §6.1.1];
- un insieme più moderno di modelli, noti come PEAKS-OVER-THRESHOLD (POT), che vengono applicati a un gran numero di osservazioni che superino un determinato valore di soglia; questi modelli sono generalmente considerati i più utili nelle applicazioni pratiche, in quanto i) usano i dati disponibili — che, essendo relativi a valori estremi e quindi per loro stessa natura poco probabili, spesso sono limitati in numero — in maniera più efficiente, e ii) possono essere facilmente estesi a situazioni nelle quali si richieda uno studio di come i valori estremi di una data variabile dipendano dai valori di una seconda variabile (o serie di variabili); in questo ambito, nasce la distribuzione **GPD** (Generalized Pareto Distribution).[cfr. §6.1.2]

6.1.1 La distribuzione GEV

Questa distribuzione nasce formalmente come una distribuzione limite per i massimi (o minimi) di una distribuzione di variabili aleatorie.

Si considerino $X_1, X_2, \ldots, X_n$ variabili aleatorie indipendenti con una funzione di distribuzione di probabilità comune F tale che

$$F(x) = \Pr\{X_j \leq x\}$$

Si può allora dimostrare che, nel limite $n \to \infty$, la distribuzione di probabilità H per i massimi $M_n = \max\{X_1, \ldots, X_n\}$, assume la seguente forma:[3]

$$H(x; \mu, \psi, \xi) = \exp\left\{-\left(1 + \xi\frac{x-\mu}{\psi}\right)^{-1/\xi}\right\} \tag{6.1}$$

definita nella regione per la quale $1 + \xi(x-\mu)/\psi > 0$ — altrove, $H(x)$ vale 0 o 1. Nella (6.1), μ è noto come *location parameter*, ψ è noto come *scale parameter* e ξ è noto come *shape parameter*

[1] Per una trattazione introduttiva ma comunque dettagliata della teoria generale, si può far riferimento al testo di Coles {10}, che riporta anche una ricca bibliografia per eventuali approfondimenti.
[2] Tipicamente, possono essere massimi giornalieri, annuali, ecc.
[3] A questa si fa riferimento col termine di *generalized extreme value distribution* (GEV) — combina tutti e tre i tipi di distribuzioni che nascono dall'approccio classico.

— quest'ultimo è quello più importante in quanto controlla/determina la natura delle code della distribuzione.

In relazione al valore di ξ, si distinguono le tre distribuzioni della teoria classica, e in particolare:

- se $\xi > 0$, si ottiene la distribuzione di *Fréchet*:

$$H(x;\xi) = \begin{cases} 0 & \text{se } x < 0 \\ \exp\left(-x^{-1/\xi}\right) & \text{se } 0 < x < \infty \end{cases} \tag{6.2}$$

- se $\xi = 0$, si ottiene la distribuzione di *Gumbel*:

$$\lim_{\xi \to 0} H(x;\mu,\psi,\xi) = \exp\left\{-\exp\left(-\frac{x-\mu}{\psi}\right)\right\} \quad \text{per } -\infty < x < \infty \tag{6.3}$$

- se $\xi < 0$, si ottiene la distribuzione di *Weibull*:

$$H(x;\xi) = \begin{cases} \exp\left\{-(-x)^{-1/\xi}\right\} & \text{se } -\infty < x < 0 \\ 1 & \text{se } x > 0 \end{cases} \tag{6.4}$$

Alcuni andamenti relativi alle equazioni (6.2), (6.3) e (6.4) sono riportati in figura 6.1.
In generale, è bene sottolineare che non tutti i momenti dell'equazione (6.1) esistono: in particolare, il k-esimo momento esiste se $\xi < k^{-1}$.

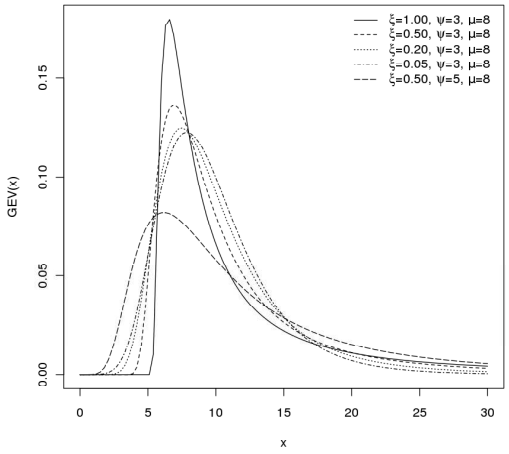
Soprattutto nel campo dei processi ambientali, si può obiettare con ragione che i fenomeni sotto studio raramente danno luogo a osservazioni rigorosamente IID; tuttavia, esistono teorie dei valori estremi per processi non-IID nelle quali emergono le stesse distribuzioni della teoria appena esposta.
Sembra infine utile porre l'accento sul fatto che:

- questo tipo di distribuzioni si può utilizzare per modellizzare l'*intero* dataset di valori e non solo la coda — come accade invece per la famiglia di distribuzioni GPD [cfr. §6.1.2];
- le distribuzioni GEV sono valide anche per valori negativi della variabile x — questo può risultare rilevante nel caso vengano applicate a variabili il cui valore è per loro stessa natura positivo e debba pertanto sottostare a questo vincolo, ad esempio in fase di stima.[4]

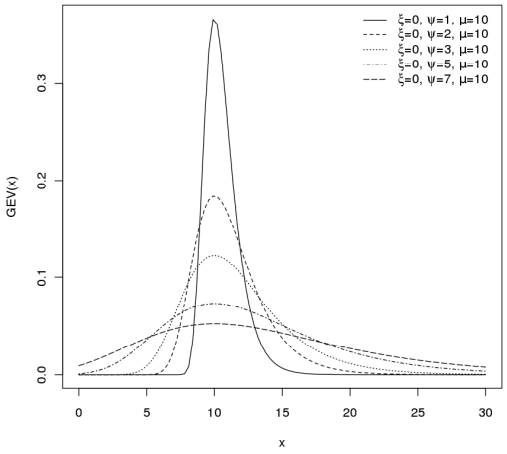
6.1.2 La distribuzione GPD

L'idea di base è quella di considerare in questo caso un valore di soglia sufficientemente elevato u e limitare lo studio a tutte le osservazioni che superano u.
È bene notare che la scelta del valore di soglia u può risultare un *punto critico* per l'applicazione pratica di questo tipo di approccio, in quanto dovrà essere una scelta di *compromesso* tra i) un valore *sufficientemente elevato* in modo che il teorema asintotico su cui la teoria si basa si possa considerare valido, e ii) un valore *sufficientemente basso* in modo che le osservazioni disponibili siano numericamente sufficienti per garantire la potenza statistica richiesta per la stima dei parametri dell'equazione (6.6).

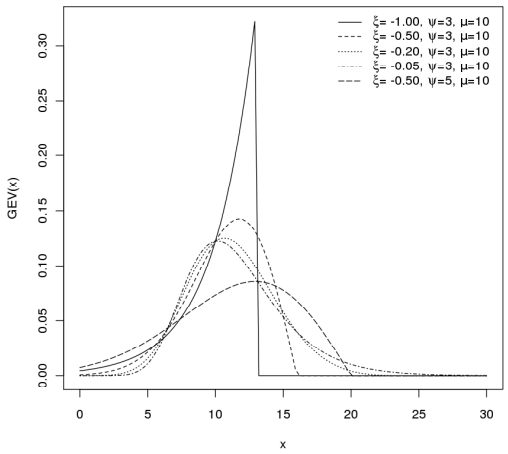
[4] Si pensi ad esempio al caso del valore di concentrazione di radon indoor: in alcune applicazioni, bisognerà prestare attenzione al fatto che i valori estratti da tali distribuzioni risultino sempre positivi! Per una trattazione reale del problema, si può fare riferimento ad esempio a quanto esposto nella nota 2, a pagina 93.



(a) distribuzione di Fréchet



(b) distribuzione di Gumbel



(c) distribuzione di Weibull

Figura 6.1: Andamenti della funzione GEV [cfr. eq (6.1)] al variare dei vari parametri che la caratterizzano.

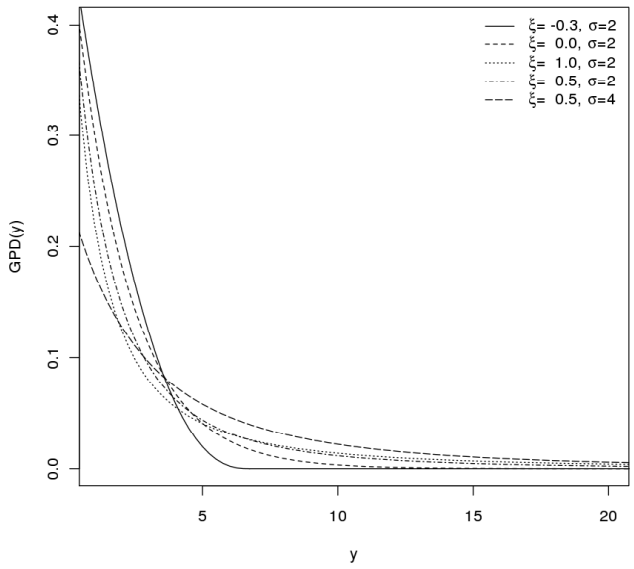


Figura 6.2: Andamenti della funzione GPD [cfr. eq (6.6)] al variare dei parametri che la caratterizzano; il valore per la soglia è $u = 0$.

Sia X una variabile aleatoria con funzione di distribuzione di probabilità $F(x)$, e si definisca una nuova variabile aleatoria $Y = X - u$ condizionata al fatto che $X > u$; allora, la funzione di distribuzione di probabilità $F_u(y)$ per i valori che superano la soglia u può essere scritta come:

$$F_u(y) = \Pr\{Y \leq y\} = \Pr\{X \leq u + y | X > u\} = \frac{F(u+y) - F(u)}{1 - F(u)}$$

La cosa interessante è che esiste un teorema (asintotico) in base al quale, per una vasta classe di funzioni di distribuzioni di probabilità — che comprende tutte quelle più usate nella statistica classica, quali Normale, log-Normale, t, Gamma, Esponenziale, χ^2, Beta, Uniforme —, al crescere di u (ovvero nel caso in cui $u \to \omega_F$, con ω_F l'upper endpoint di F) vale che:

$$\lim_{u \to \omega_F} \sup_{0 \leq y < \omega_F - u} |F_u(y) - G(y; \sigma_u, \xi)| = 0 \tag{6.5}$$

dove $G(y; \sigma_u, \xi)$ è nota come *generalized pareto distribution* (GPD), ed è data da:

$$G(y; \sigma_u, \xi) = 1 - \left(1 + \xi \frac{y}{\sigma_u}\right)^{-1/\xi} \tag{6.6}$$

dove σ_u è noto come *scale parameter* e ξ è noto come *shape parameter*; come nel caso della GEV, è quest'ultimo il parametro fondamentale che caratterizza il decadimento/tipo della coda. In figura 6.2 sono riportati alcuni andamenti dell'equazione (6.6).

Il significato della (6.5) è quindi il seguente: per un valore di soglia u sufficientemente elevato, esistono σ_u e ξ (che non dipende da u), per i quali la GPD risulta essere un'ottima approssimazione per la funzione di distribuzione di probabilità F_u per i superamenti.

Come accade per la distribuzione GEV, al variare del segno del parametro ξ si ottengono tre differenti casi:

- se $\xi > 0$, allora la (6.6) è valida per $0 < x < \infty$ e la coda della distribuzione soddisfa

  $$1 - G(y; \sigma_u, \xi) \sim cy^{-1/\xi}, \quad c > 0$$

  questa è tradizionalmente nota come *Pareto tail*

- se $\xi = 0$, si ottiene

  $$G(y; \sigma_u, 0) = 1 - \exp\left(-\frac{y}{\sigma_u}\right)$$

  ovvero una distribuzione di tipo Esponenziale con media pari a σ_u

- se $\xi < 0$, allora la $G(y; \sigma_u, \xi)$ ammette un upper endpoint a $\omega_G = \sigma/|\xi|$ e si ottiene una distribuzione simile alla (6.4) della teoria classica

Anche per l'equazione (6.6), il k-esimo momento esiste se $\xi < k^{-1}$.

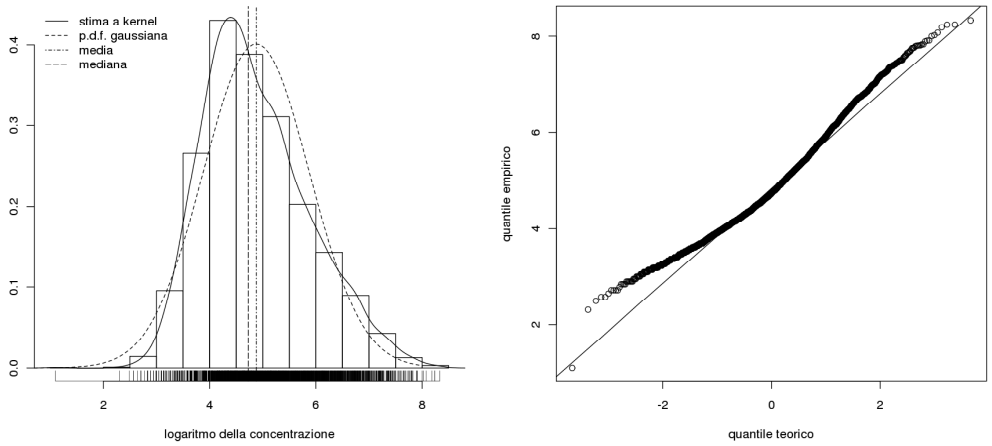
6.2 Descrizione del dataset utilizzato

Il dataset utilizzato per questa analisi si compone di **4050** valori di concentrazione di attività di radon indoor estratti dal dataset di riferimento [cfr. §1.3]; ho deciso di non operare alcuna correzione sul valore di concentrazione, mantenendo quindi quello relativo all'esposizione reale (nessuna correzione per ottenere un valore riferito a media annuale), e di non considerare alcuna variabile in base alla quale rendere il dataset più omogeneo.

Dal dataset di riferimento ho estratto tutti e soli i cases per i quali fossero disponibili tutti valori relativi ad alcune variabili secondarie che caratterizzano la singola misura, in modo da poterle eventualmente sfruttare per analisi successive; nello specifico, ogni valore di concentrazione è accompagnato da informazioni relative a[5]:

- piano di esposizione
- stagione di esposizione
- comprensorio
- utilizzo
- tipo locale
- tipo di costruzione
- data di costruzione (e relativa classe)
- qualità degli infissi
- contatto con il terreno

[5] Per una descrizione delle variabili citate, si faccia riferimento al paragrafo §1.3, a pagina 7.



(a) (b)

Figura 6.3: Istogramma (a) e qq-plot (b) per il logaritmo della concentrazione di radon indoor, espressa in $Bq \cdot m^{-3}$; la distribuzione di riferimento è di tipo Normale, i cui parametri sono stati stimati a partire dai dati empirici.

Sembra importante infine sottolineare come per questo tipo di analisi, la parte *spaziale* del fenomeno, ovvero la informazioni relative alla *localizzazione* delle singole misure, è stata ignorata.

Prima di intraprendere le analisi specifiche, mi è sembrato opportuno valutare quanto la distribuzione di probabilità dei dati presi in esame si potesse considerare di tipo log-Normale, ipotesi che spesso in letteratura viene assunta quasi 'a priori'. Per questo, sono state condotte delle semplici analisi statistiche e grafiche sui dati log-trasformati, con lo scopo di facilitare i test e le visualizzazioni prendendo come riferimento una distribuzione di tipo Normale.

La figura 6.3a riporta l'istogramma sperimentale, da confrontare con una stima a kernel della p.d.f. e con la distribuzione gaussiana i cui parametri sono stati stimati a partire dai dati sperimentali; il grafico riporta anche le stime per la media e la mediana, con lo scopo di rendere più evidente l'eventuale asimmetria della distribuzione stessa. La figura 6.3b riporta invece un qq-plot, la cui distribuzione di riferimento è anche in questo caso la medesima gaussiana. È facile notare come:

- la distribuzione della variabile log-trasformata non sia simmetrica, come evidenziato dalla stima a kernel e da media e mediana non sovrapposte;
- la distribuzione gaussiana non modellizzi al meglio la distribuzione empirica (sia in relazione alla parte centrale, sia in relazione alle code), essendo tra l'altro una distribuzione perfettamente simmetrica;
- entrambe le code non vengano modellizzate nel modo corretto: risulta evidente come la distribuzione gaussiana sovrastimi la coda inferiore e sottostimi quella superiore[6] — quest'ultimo effetto risulta di particolare rilevanza in un contesto di *prevenzione* e tutela, in

[6]Fatto inizialmente piuttosto curioso, i coefficienti di skewness e kurtosis sono rispettivamente pari a 0.542 e -0.02, valori che farebbero propendere per una distribuzione simmetrica e gaussiana; tuttavia, un'analisi grafica

soglia u	*scale* σ_u	*shape* ξ
500	432 ± 32	0.13 ± 0.06
1100	630 ± 83	-0.02 ± 0.09
1100	630 ± 58	0

Tabella 6.1: Valori dei parametri per il modello GPD ottenuti mediante *maximum likelihood*; $\xi = 0$ indica che il valore per questo parametro è stato imposto a priori; i valori di soglia sono riportati in Bq·m^{-3}.

quanto si corre il rischio di sottostimare i valori elevati di concentrazione, quelli che invece dovrebbero ricevere la maggior attenzione e risultare affidabili.

Da queste e altre analoghe analisi condotte, concludo sostenendo che

la distribuzione Normale/log-Normale non sembra assere in grado di riprodurre in maniera opportuna non solo la coda superiore della distribuzione dei dati sperimentali, ma nemmeno il comportamento globale.

6.3 Analisi GPD

Come discusso nel paragrafo §6.1.2, il primo passo per la determinazione del modello di distribuzione della coda superiore della p.d.f. empirica — a dire, il modello stocastico per i superamenti del valore di soglia u — consiste nel ricavare, in base ai dati in possesso, il valore di soglia appropriato.

Per fare questo, ho utilizzato una procedura grafica basata sui cosiddetti *tc-plot*, grafici che mostrano la stabilità dei parametri shape ξ e scale σ_u [cfr. eq. (6.6)] in funzione del valore di soglia u; un esempio di tali grafici è riportato in figura 6.4. Buoni valori di soglia si manifestano in corrispondenza di evidenti discontinuità nell'andamento riportato nei grafici stessi, ovvero se il valore di soglia u è un valore valido per il modello POT, la stima del parametro in esame dovrebbe essere approssimativamente costante oltre il valore u.

Conducendo inizialmente analisi esplorative su tutto il range dei valori, e limitando successivamente l'attenzione sui valori di u più significativi, ho individuato due possibili valori di soglia, e in particolare:

- $u = 500$ Bq·m^{-3}, con $\xi > 0$ (Pareto tail) e $N_{500} = 441$ [cfr. fig. 6.4]
- $u = 1100$ Bq·m^{-3}, con $\xi \leq 0$ e $N_{1100} = 117$

avendo indicato con N_u il numero di valori sperimentali sopra la soglia u.

I parametri relativi alle distribuzioni GPD sono quindi stati determinati mediante una stima basata sull'approccio *maximum likelihood*[7], controllando di volta in volta anche la simmetria dell'intervallo di confidenza attorno al valore stimato per larghezze pari a 1σ (68%) e 2σ (98%); in tutti i casi, gli intervalli risultano molto simmetrici. La tabella 6.1 riporta i risultati ottenuti.

della stessa mette in luce come questo risultato sia molto probabilmente imputabile a un "effetto di compensazione" sul comportamento opposto per le due code.

[7] Con questo metodo si ha il vantaggio di ottenere anche degli intervalli di confidenza che accompagnano la stima del valore puntuale, e di poterne verificare graficamente la simmetria — forti asimmetrie attorno al valore centrale potrebbero infatti essere indicazioni di eventuali problemi di natura prettamente computazionale o di anomalie relative al dataset, che andrebbero quindi indagate nello specifico al fine di ottenere delle stime affidabili.

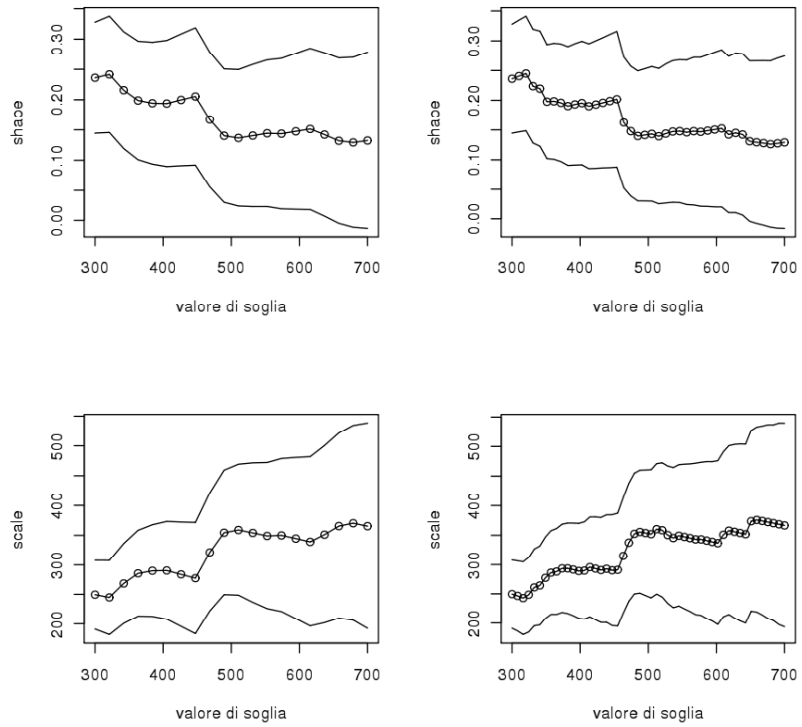


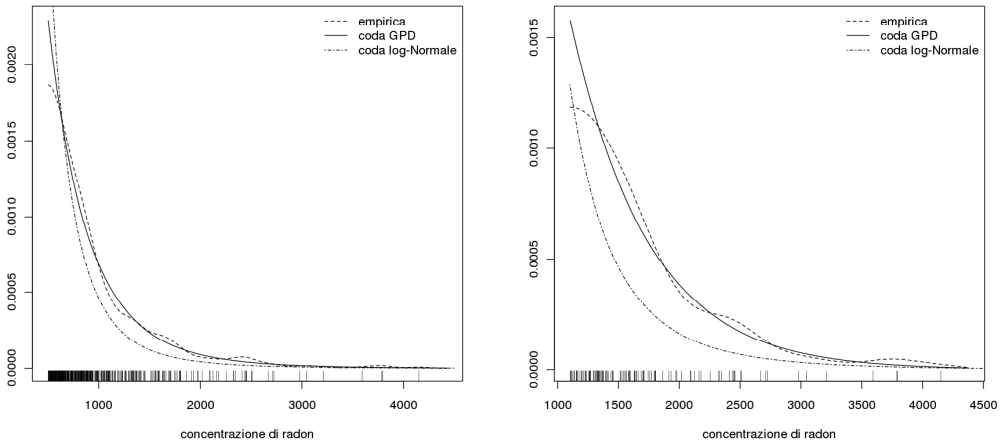
Figura 6.4: Esempio di *tc-plot* utilizzati per la ricerca del valore ottimale per il valore di soglia u; in questo caso, si è limitato il range di visualizzazione attorno a $u = 500$, uno dei valori di soglia individuati nelle analisi condotte: risulta evidente la discontinuità nei pressi di tale valore; le concentrazioni sono espresse in $\mathrm{Bq \cdot m^{-3}}$.

In relazione a quanto ottenuto per $u = 1100$, situazione compatibile con un decadimento di tipo esponenziale — per altro uno dei più ricorrenti in letteratura —, ho provato a imporre al modello anche un valore di shape $\xi = 0$: come ci si poteva aspettare, il decadimento esponenziale risulta compatibile con quanto emerso dalla precedente analisi più "libera".

In figura 6.5 sono riportati gli andamenti per la p.d.f. empirica per la coda superiore della distribuzione dei valori di concentrazione, messi a confronto con il modello GPD corrispondente e anche con una distribuzione log-Normale i cui parametri sono stati ricavati da fit sui dati che costituiscono il dataset operativo: risulta del tutto evidente come l'approccio log-Normale non sia in grado di riprodurre l'andamento reale della coda superiore, con una pesante sottostima dello stesso — ancora più evidente se si analizza la figura 6.5b.

Si può quindi concludere questa prima fase di analisi affermando che

> *una volta individuati il/i valore/i di soglia corretto/i u, il modello GPD è in grado di riprodurre nella maniera corretta il comportamento della coda superiore della distribuzione dei valori di concentrazione di radon indoor, evitando le pesanti sottostime che caratterizzano invece un approccio di tipo log-Normale.*



(a) $u = 500$ (b) $u = 1100$

Figura 6.5: Confronti per la coda superiore della distribuzione dei valori di concentrazione di radon indoor per i due modelli GPD per i due valori di soglia individuati; le concentrazioni sono espresse i $\mathrm{Bq \cdot m^{-3}}$.

6.3.1 Analisi su singoli comprensori

Con lo scopo di valutare l'applicabilità dei risultati ottenuti per l'intero dataset a zone con caratteristiche diverse, ho deciso di condurre, in maniera del tutto analoga, le stesse procedure e analisi descritte in precedenza scegliendo però alcuni comprensori dell'Alto Adige in base agli istogrammi sperimentali e ai valori di concentrazione che li caratterizzano; altro parametro da prendere in considerazione per la scelta, il numero N di campionamenti disponibili, che dovrà essere sufficiente per garantire la potenza statistica richiesta.

Gli istogrammi sperimentali per i tre comprensori selezionati — Burgraviato, Val Pusteria e Val Venosta — sono riportati in figura 6.6, mentre le prime colonne della tabella 6.2 riportano alcuni parametri statistici utili per caratterizzare la distribuzione dei valori di concentrazione.

Per ogni comprensorio, ho determinato il valore di soglia u ricorrendo ad analisi visuali dei tc-plot, anche se gli andamenti, probabilmente in relazione alla scarsa numerosità dei subset[8], risultano di più difficile lettura, o quantomeno la struttura risulta molto meno evidente. I risultati del fit per i parametri σ_u e ξ sono stati eseguiti con il metodo della *maximum likelihood*, ma in questo caso gli intervalli di confidenza risultano meno simmetrici rispetto a quelli per il dataset completo. I valori numerici sono riportati in tabella 6.2.

Poiché in tutti e tre i casi il parametro ξ risulta prossimo a zero, ho provato a imporre un modello di decadimento della coda di tipo esponenziale per valutare l'influenza di questa "scelta a priori": potrebbe infatti manifestarsi come un modello valido per tutti i casi, con il vantaggio che si avrebbe di conseguenza un solo parametro da stimare — caratteristica da non sottovalutare nel caso in cui si avessero a disposizione pochi campionamenti sperimentali.

[8] In una prima fase, è stato incluso anche Bolzano, con lo scopo di testare il modello GPD su una zona caratterizzata da valori piuttosto bassi, in relazione agli altri comprensori scelti; tuttavia, la scarsa numerosità che caratterizza questo subset ($N = 138$) ha reso praticamente illeggibile la struttura del corrispondente tc-plot e pertanto l'analisi su Bolzano non ha avuto seguito.

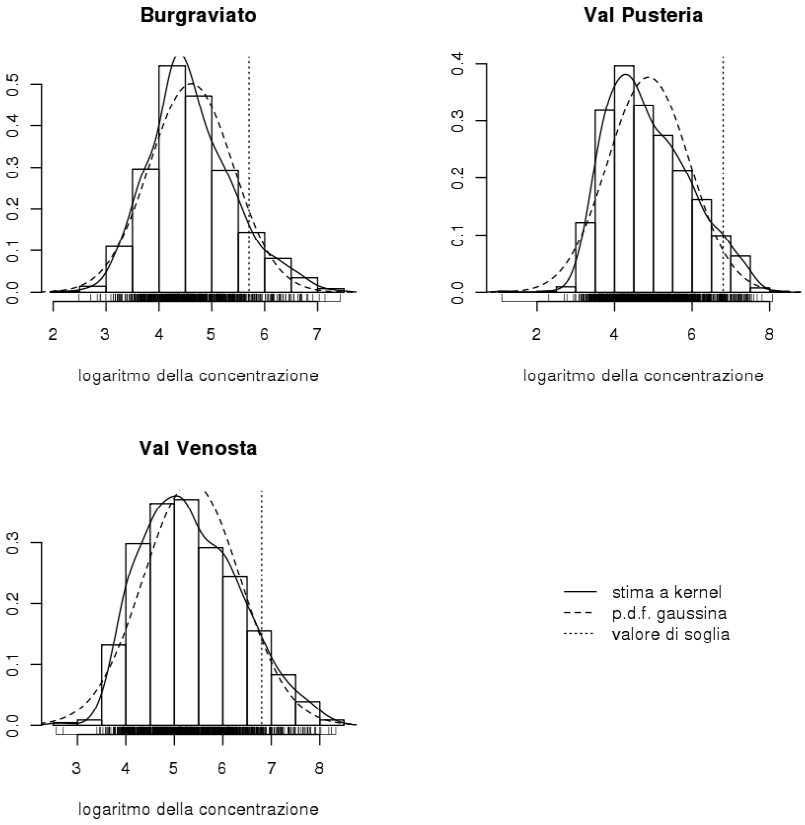


Figura 6.6: Istogrammi sperimentali relativi ai 3 comprensori presi in esame per analisi specifiche; per maggiori dettagli, si faccia riferimento al testo.

comprensorio	N	$N > u$	*media*	*mediana*	*max*	*soglia* u	*scale* σ_u	*shape* ξ
Burgraviato	695	65	144	92	1687	300	240±43	0.03±0.1
							245±30	0
Val Pusteria	1049	56	242	111	3210	900	496±86	-0.08±0.1
							469±63	0
Val Venosta	912	78	364	190	4154	900	736±124	-0.03±0.1
							733±83	0

Tabella 6.2: Parametri statistici descrittivi per i valori di concentrazione (espresse i $\mathrm{Bq \cdot m^{-3}}$) dei 3 comprensori considerati per analisi specifiche e relativi parametri per i modelli GPD stimati mediante *maximum likelihood*; $\xi = 0$ indica che il valore per questo parametro è stato imposto a priori; per maggiori dettagli, si faccia riferimento al testo.

Come per l'analisi precedente, un confronto visivo tra coda empirica, modello GPD e modello log-Normale rende evidente l'inadeguatezza (che si traduce in una sottostima) di quest'ultimo nel descrivere il comportamento della coda superiore della distribuzione di probabilità.

> *I risultati ottenuti, sia per l'intero dataset che per alcuni comprensori (subset), mostrano come un modello di decadimento esponenziale della coda superiore per la p.d.f. dei valori di concentrazione di radon indoor risulti compatibile con quello sperimentale, evitando il pesante effetto di sottostima che caratterizza invece l'approccio di tipo log-Normale, ampiamente usato in letteratura; inoltre, si ottiene anche il vantaggio* operativo *di avere un solo parametro da stimare — proprietà utile nel caso in cui si disponesse di un esiguo numero di campionamenti.*

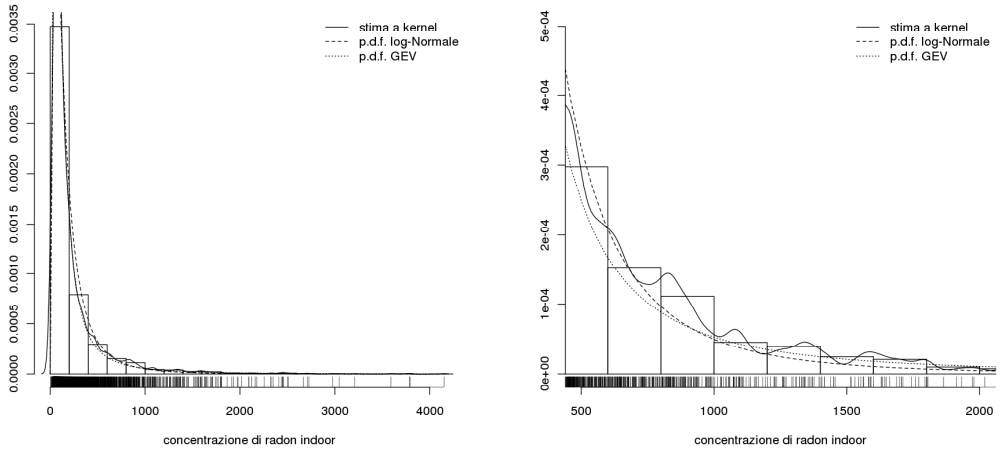
6.4 Analisi GEV

Inizialmente, ho deciso di applicare il modello classico proposto da Gumbel all'intero dataset descritto in precedenza, senza operare alcuna trasformazione sui dati: lo scopo è stato quello di trovare una distribuzione di probabilità adatta a riprodurre l'*intero* istogramma sperimentale (e non solo la coda superiore, come nel caso GPD), che si ricorda risulta essere asimmetrico e con la coda superiore piuttosto estesa. Per questo, è stato fittato un modello di tipo GEV [cfr. eq. (6.1)] sui dati sperimentali ricorrendo al metodo della *maximum likelihood*; in linea di principio, rispetto a una distribuzione log-Normale, la distribuzione GEV dovrebbe essere in grado di riprodurre in maniera più efficace tanto l'asimmetria dell'istogramma sperimentale quanto il comportamento delle code — in particolare, in un contesto di prevenzione, l'attenzione sarà focalizzata su quella superiore.

In figura 6.7 sono riportati l'istogramma sperimentale e i risultati ottenuti per i due modelli (log-Normale e GEV), per un confronto visivo. A una prima analisi, sembra che il modello log-Normale riproduca meglio il comportamento della coda superiore della distribuzione, ma focalizzando l'attenzione solo su quest'ultima, come evidenziato nella figura 6.7b, si può notare invece come per valori elevati, il modello GEV sia in grado di riprodurre meglio il comportamento del dataset sperimentale, ovvero manifesti un decadimento più lento rispetto al modello log-Normale e più in linea con quanto evidenziato dai dati stessi. Nell'ottica di un approccio di tipo cautelativo e di prevenzione, risulta quindi importante la capacità del modello impiegato di non sottostimare questa parte della distribuzione.

Successivamente, ho deciso di applicare la stessa procedura appena descritta sempre al medesimo dataset, ma operando in fase di preparazione dei dati una trasformazione di tipo logaritmico, come spesso viene peraltro suggerito in letteratura; quella che si ottiene è una distribuzione più simmetrica, e in linea di principio di tipo Normale — assunto per valido il modello log-Normale, ipotesi che come si è visto può risultare discutibile. I risultati ottenuti per questo nuovo modello GEV sono riportati in figura 6.8.

Appare evidente anche in questo caso come l'assunzione di una distribuzione di tipo Normale non sia adatta al dataset di riferimento, in quanto non è in grado di riprodurne l'asimmetria che lo caratterizza — e che permane anche a seguito di una trasformazione di tipo logaritmico. Il modello GEV riproduce invece in maniera sufficientemente buona entrambe le caratteristiche peculiari dei dati di radon indoor sotto studio, ovvero la già citata asimmetria della distribuzione e il decadimento delle code. Si noti infine come la distribuzione Normale, in conseguenza della sua simmetria intrinseca, sovrastimi la coda inferiore e sottostimi quella superiore, configurandosi quindi come non ottimale per lo studio dei valori elevati, quelli cui prestare maggior attenzione nel campo della prevenzione.



(a) (b)

Figura 6.7: Confronto tra istogramma sperimentale e modelli di distribuzioni di tipo log-Normale e GEV fittati sui dati sperimentali; l'istogramma (a) riporta l'intero campo di variabilità dei valori di concentrazione di radon, l'istogramma (b) uno zoom sulla coda superiore per meglio evidenziare le differenze tra i due modelli proposti; i valori di concentrazione sono riportati in $Bq \cdot m^{-3}$.

Nota 1: *Conducendo analisi specifiche sul modello fittato sui dati raw (non trasformati), è emerso che il decadimento della coda superiore potrebbe risultare troppo* lento *se paragonato a quello log-Normale, che tuttavia risulta al contrario troppo* rapido*; in particolare, estraendo dalla p.d.f. GEV un insieme di valori numericamente elevato (indicativamente, $N > 500$), non è raro ottenere valori dell'ordine di 20000 $Bq \cdot m^{-3}$ o superiori, che risultano eccessivi se considerati come valori di concentrazione di radon — sottolineo comunque che si tratta di pochi casi, nell'ordine dell'unità, che sembrano più che altro configurarsi come outliers, facilmente identificabili rispetto al resto delle estrazioni.*

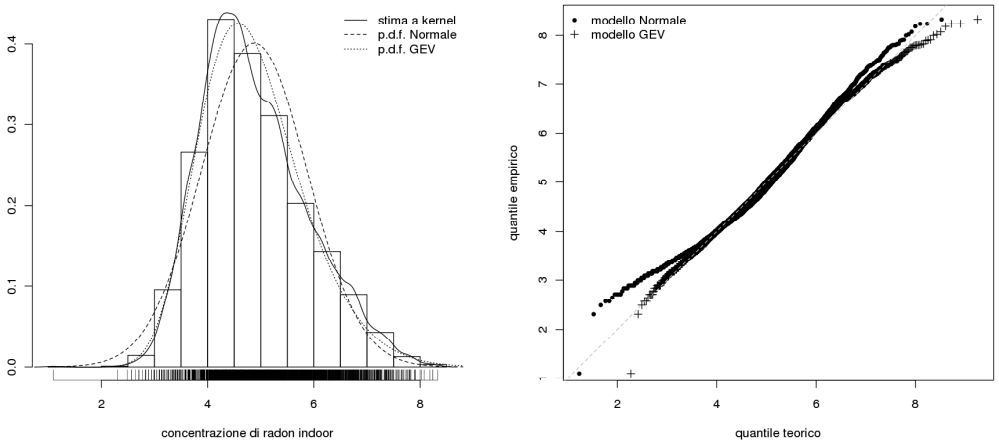
Probabilmente, questo aspetto è legato alla forte asimmetria della distribuzione empirica, che potrebbe influenzare il modello nella direzione di una coda superiore 'troppo' estesa. Tale problema non si verifica invece nel caso in cui il modello GEV venga fittato su dati log-trasformati, a parziale conferma dell'ipotesi espressa — la trasformazione, pur non eliminando l'asimmetria, attenua il peso della coda superiore. Per un esempio pratico degli effetti appena descritti, si può far riferimento a quanto esposto in §6.5.3.

6.4.1 Analisi su singoli comprensori

In perfetta analogia a quanto descritto in §6.3.1, è sembrato utile valutare l'applicabilità del modello GEV su scale spaziali ridotte e a subset con caratteristiche differenti in relazione ai valori caratteristici di radon indoor. I comprensori scelti per questa fase dell'analisi sono gli stessi cui si è fatto ricorso in precedenza, e i cui parametri statistici sono riportati in tabella 6.2.

Anche in questo caso, il modello GEV si è rivelato migliore rispetto a quello Normale, riuscendo a modellizzare in maniera più efficace tanto l'asimmetria della distribuzione quanto il comportamento di entrambe le code.

Infine, la tabella 6.3 riporta i valori dell'errore quadratico medio per i due modelli messi a confronto, calcolato sulla base di una stima a kernel della distribuzione di probabilità empirica:



(a) istogramma (b) qq-plot

Figura 6.8: Istogramma sperimentale e qq-plot per un confronto tra i modelli GEV e Normale fittati sui valori di concentrazione di radon sottoposti a una preventiva trasformazione di tipo logaritmico.

dataset	*modello GEV*	*modello Normale*
completo	1.49	3.76
Burgraviato	1.89	3.66
Val Pusteria	2.51	4.07
Val Venosta	2.14	3.28

Tabella 6.3: Valori dell'errore quadratico medio per i due modelli proposti al variare del dataset di riferimento e calcolato sulla stima a kernel della p.d.f. empirica; i valori vanno moltiplicati per un fattore 10^{-2}.

in tutti i casi presi in esame, si nota come il modello GEV risulti migliore anche in base a questo parametro quantitativo.

Riassumendo quanto esposto per la modellizzazione dell'intera distribuzione empirica dei valori di concentrazione di radon indoor mediante un modello di tipo GEV, si può affermare che

> *rispetto al più convenzionale e diffuso approccio di tipo Normale/log-Normale, la distribuzione GEV derivata dalla teoria dei valori estremi è in grado di descrivere in maniera più aderente alla realtà tanto l'evidente asimmetria che caratterizza la distribuzione empirica dei valori di radon indoor quanto il comportamento di entrambe le code della stessa.*
> *La validità di questo approccio risulta verificata anche al variare della numerosità del campione e delle sue caratteristiche di tipo statistico (quali ad esempio media e valori massimi), come dimostrato dall'analisi condotta su specifici comprensori dell'Alto Adige che manifestano differenze evidenti nelle caratteristiche del fenomeno preso in esame.*

6.5 Una possibile applicazione: EVT e sGs

Nell'ambito delle applicazioni di tipo geostatistico per la stima di una data variabile in localizzazioni non campionate, le Simulazioni Gaussiane Sequenziali (sGs) [cfr. §B.6, pag. 240] si configurano come una valida e consolidata alternativa all'approccio più standard del kriging, superando, almeno in parte, i due principali difetti che caratterizzano questo tipo di interpolatore: l'inevitabile quanto intrinseco effetto di smoothing (sovrastima dei valori bassi di concentrazione e sottostima di quelli elevati) e l'incapacità di riprodurre le caratteristiche globali del fenomeno. Questo tipo di simulazioni richiede una fase di pre-trattamento dei dati che si traduce nella pratica in una trasformazione di tipo *Nscored* — i dati $z(\mathbf{u})$ devono essere trasformati in dati $y(\mathbf{u})$ in modo tale che appartengano a una distribuzione Normale con media nulla e varianza unitaria, ovvero $y(\mathbf{u}) \in \mathcal{N}(0,1)$; i modi per ottenere questo risultato possono essere svariati. *sGs*

In questo contesto, e alla luce dei risultati descritti in questo capitolo, ho deciso di testare la distribuzione di tipo GEV operando la trasformazione Nscored in via del tutto analitica, avendo a disposizione l'equazione (6.1) e gli strumenti per la stima dei parametri che la caratterizzano. Il medesimo processo analitico è stato quindi applicato anche in fase di trasformazione inversa, per ottenere un valore di concentrazione di attività di radon indoor reale. Ulteriori dettagli sulla procedura e sui modelli implementati saranno discussi nel paragrafo §6.5.1

La parte computazionale è stata svolta in ambiente `R`{44} e ricorrendo ai packages *Hmisc*{24}, *evd*{49}, *lattice*{45}, *gstat*{41} e *geoR*{30}; la parte relativa alle simulazioni è stata invece eseguita ricorrendo a `Geostat Office`©.

6.5.1 Descrizione del dataset utilizzato

Per questa applicazione, il dataset operativo è costituito dai valori di concentrazione di radon indoor che costituiscono il dataset di riferimento [cfr. §1.3], dal quale sono stati pre-selezionati i dosimetri esposti nel semestre invernale e al piano zero, per questioni di omogeneità; i valori disponibili sono 2578. Successivamente, questo dataset è stato suddiviso in due parti, con lo scopo di avere a disposizione un nuovo dataset su cui condurre delle validazioni ad hoc [cfr. §6.5.4]; la procedura di divisione è stata eseguita con `Geostat Office`© e un algoritmo di declustering a celle: in questo modo, si ottengono due distinti subset che preservano la distribuzione spaziale di quello di partenza. Il dataset operativo si compone di **707** cases, mentre quello di validazione dei rimanenti **1871**.

Su questo dataset ho quindi condotto dei fit per la determinazione dei parametri del modello GEV [cfr. eq. (6.1)], sia sui dati raw che su quelli log-trasformati, in perfetta analogia a quanto discusso nel paragrafo §6.4 e ottenendo dei valori in linea con quelli che caratterizzano il dataset impiegato in quella fase dell'analisi.

Nella pratica, ho deciso di mettere a confronto 4 differenti tipi di trasformazione, con lo scopo di valutare l'effetto in una applicazione reale di alcuni aspetti emersi nelle analisi precedenti; brevemente, le trasformazioni implementate sono state le seguenti (l'identificativo *Ns'n'* sarà usato nel corso dell'intera discussione):

trasformazione Ns01 : partendo dai dati raw (non trasformati), trasformazione a dati $\mathcal{N}(0,1)$ mediante una distribuzione di probabilità cumulativa stimata da quella empirica; con questo approccio, si può però correre il rischio di non avere una stima corretta delle code, in quanto per loro stessa natura poco popolate [cfr. ad esempio quanto discusso da Deutsch e Journel {16}, a pagina 134 e seguenti]: nonostante questa parziale limitazione, è uno degli approcci più diffusi e consolidati;

trasformazione Ns02 : trasformazione a dati $\mathcal{N}(0,1)$ mediante un modello GEV (c.d.f. di tipo analitico) fittato sui dati raw;

trasformazione Ns03 : trasformazione a dati $\mathcal{N}(0,1)$ mediante un modello GEV (c.d.f. di tipo analitico) fittato sui dati log-trasformati;

trasformazione Ns04 : trasformazione a dati $\mathcal{N}(0,1)$ mediante un modello log-Normale (c.d.f. di tipo analitico) fittato sui dati raw.

In figura 6.9 sono riportati gli istogrammi per i valori trasformati secondo quanto appena descritto, e messi a confronto con una stima a kernel e con una distribuzione Normale di riferimento; per ogni istogramma, sono riportati anche media e varianza dei corrispondenti dataset.

6.5.2 Analisi variografica

Poiché le sGs, basandosi sull'algoritmo del simple kriging, richiedono un modello per il variogramma $\gamma(\mathbf{h})$, ho condotto una serie di analisi variografiche esplorative per i 4 dataset operativi al variare della distanza massima, della risoluzione spaziale (lag-step) e della direzione; in tutti i casi, ho fatto ricorso al variogramma tradizionale, controllando anche il numero minimo di coppie per ogni lag. In base ai risultati ottenuti, non ho ritenuto opportuno introdurre eventuali anisotropie, che l'analisi delle variogram maps non ha peraltro reso particolarmente evidenti.

La struttura dei variogrammi sperimentali, come è del resto prevedibile aspettarsi a seguito della trasformazione introdotta, risulta in tutti i casi leggibile (rispetto a quella che caratterizza invece i dati raw, come si può ad esempio vedere osservando i variogrammi della figura 4.1 a pagina 45) e non evidenzia particolari differenze tra un dataset e l'altro per quanto riguarda l'andamento globale.

Analisi più dettagliate hanno però messo in luce che:

- per quanto riguarda i dataset Ns01–02–03, gli andamenti sono praticamente sovrapponibili, leggermente scalati in relazione al valore della varianza a priori; per questo, ho deciso di ricorrere a un unico modello comune (la figura 6.10a riporta tutti e tre i variogrammi sperimentali per una visione d'insieme) rappresentato dall'equazione (6.7);

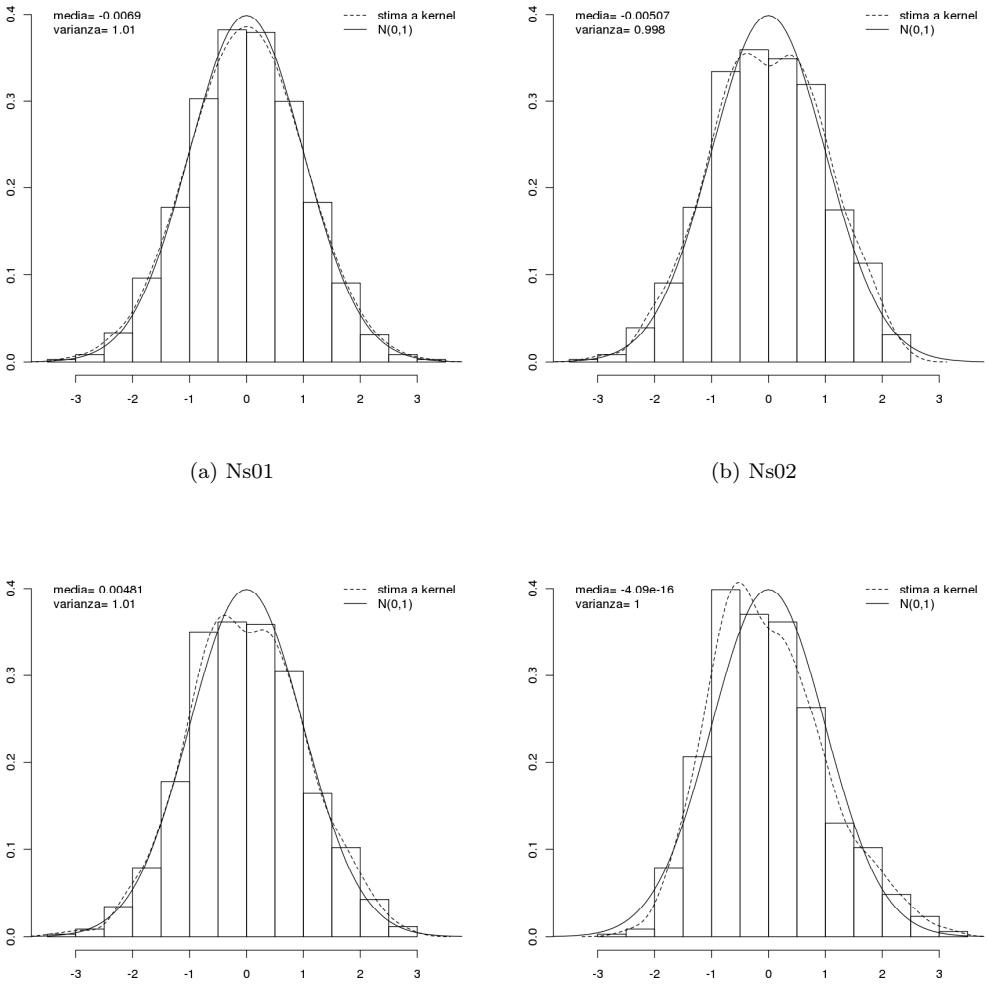
- per quanto riguarda invece il dataset Ns04, questo manifesta un nugget più elevato e un andamento leggermente differente, soprattutto in relazione al range; per questo, ho deciso di ricorrere a un modello specifico per questo dataset [cfr. eq. (6.8)].

I modelli teorici di variogramma sono stati determinati mediante un fit interattivo sfruttando il package *geoR*, focalizzando l'attenzione sulla forma "globale" del variogramma sperimentale; i modelli ottenuti e utilizzati per le successive fasi di simulazione sono i seguenti:

$$\gamma(h)_{Ns0123} = 0.62 + 0.42 \cdot \exp\left(\frac{h}{10830}\right) \tag{6.7}$$

$$\gamma(h)_{Ns04} = 0.68 + 0.34 \cdot \exp\left(\frac{h}{12200}\right) \tag{6.8}$$

dove i valori del range sono riportati in metri e 'exp' indica un modello di tipo esponenziale.



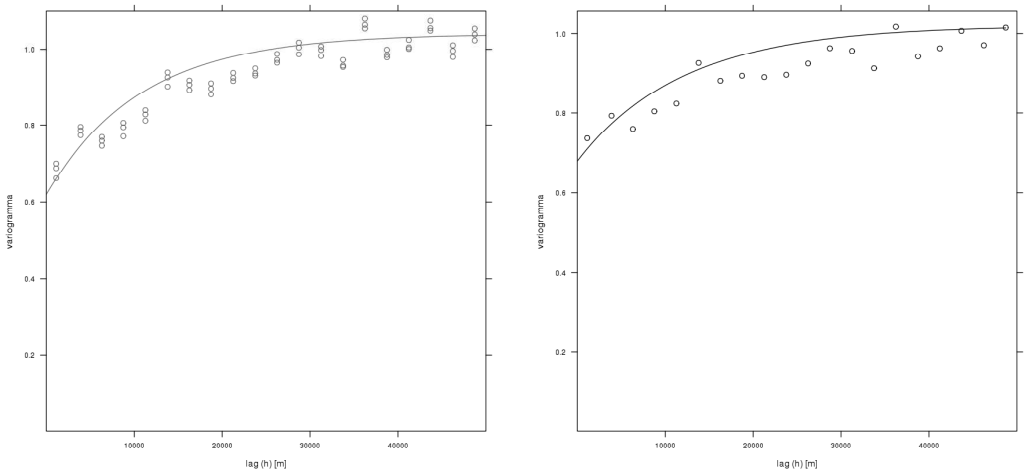
(a) Ns01

(b) Ns02

(c) Ns03

(d) Ns04

Figura 6.9: Istogrammi per i valori di concentrazione di radon ottenuti in seguito alle varie trasformazioni Nscored descritte in §6.5.1; ogni grafico riporta anche media e varianza relative alla specifica trasformazione.



(a) Ns01–02–03 (b) Ns04

Figura 6.10: Variogrammi sperimentali e relativi modelli (linea continua) per i dataset descritti in §6.5.1.

6.5.3 Simulazioni Gaussiane Sequenziali

Ho inizialmente definito una griglia regolare di passo (2×2) km^2 attorno ai campionamenti, per un totale di 4187 punti di stima sull'intero territorio altoatesino. Tutte le simulazioni gaussiane sequenziali sono state condotte con il software `Geostat Office`©, impostando i seguenti parametri:

- N (numero di simulazioni) pari a 200
- algoritmo di kriging: ordinary kriging — in questo modo, possono essere prese in considerazione anche situazioni di stazionarietà locale
- raggio di ricerca S_r: 6 km — vengono così utilizzati per la fase di stima solo i punti che si trovano all'interno di un cerchio di raggio S_r centrato sul punto di simulazione
- numero massimo di punti da utilizzare in fase di stima N_{max}: 10 reali e 3 simulati — l'algoritmo di simulazione prevede infatti di considerare anche i punti precedentemente simulati come facenti parte del dataset operativo [cfr. §B.6.1]
- numero minimo di punti da utilizzare in fase di stima N_{min}: 3 — se all'interno del cerchio di raggio S_r non ci sono almeno 3 punti, il valore per quel nodo della griglia non viene stimato
- utilizzo dell'approccio *multigrid* durante la visita dei punti della griglia di simulazione, al fine di ridurre al minimo la comparsa di possibili artefatti legati alla parte computazionale [cfr. {31, pag. 152}]

Le sGs forniscono in uscita dati Nscored, che vanno pertanto riportati alla scala originale dei valori di partenza (raw); eccettuato il dataset Ns01, per il quale la trasformazione inversa è stata

eseguita in automatico da `Geostat Office`©, in tutti gli altri casi sono ricorso alle equazioni analitiche delle varie c.d.f. coinvolte (GEV e log-Normale) elaborando i risultati in ambiente `R`{44} e considerando solo le localizzazioni per le quali fossero disponibili tutti i 200 valori delle singole simulazioni — in questo modo, si ha la garanzia che tutte le analisi statistiche sono basate sullo stesso numero di punti per ogni nodo della griglia di simulazione.

I dati ottenuti sono stati elaborati e preparati opportunamente in funzione delle analisi successive, e nello specifico:

analisi 'E-type' : questo tipo di analisi prevede di sfruttare i risultati ottenuti con sGs per produrre mappe di tipo per così dire "convenzionale" (tipicamente, mappe per il valor medio di concentrazione), basandosi su un approccio statistico di tipo classico; per ogni nodo della griglia, ho calcolato media, mediana, varianza, valore minimo e massimo;

analisi 'Probability Map' : questo tipo di analisi prevede invece di determinare delle mappe che riportino la probabilità di superare un determinato valore di soglia, sfruttando in questo caso le fluttuazioni statistiche associate alle 200 simulazioni condotte; per questo, si è creata una opportuna routine in ambiente `R`{44}.

Si descrivono quindi i risultati ottenuti per i due differenti approcci esaminati.

Analisi 'E-type'

Un primo controllo della bontà della simulazione, rispetto ai risultati ottenuti mediante kriging, lo si può ottenere valutando contemporaneamente la mappa per il valor medio e quella della relativa varianza; nel caso delle sGs, alla luce del fatto sperimentale che, assumendo o meno un modello di tipo log-normale per i dati raw, le zone caratterizzate da valori elevati di concentrazione manifestano anche una elevata varianza (fenomeno noto come *effetto proporzionale*), le due mappe dovrebbero riprodurre visivamente andamenti simili — non così nel caso di mappe prodotte mediante kriging, in quanto la varianza dipende *unicamente* dalla distribuzione spaziale dei campionamenti attorno al punto di stima [cfr. §B.4.5, pag 239]. Questa preliminare verifica ha portato i risultati attesi per tutte le 4 tipologie di simulazioni condotte: globalmente, il fenomeno simulato rispetta la caratteristica base di quello reale. In figura 6.11 sono riportati gli istogrammi per i valori medi di concentrazione che costituiscono i quattro dataset simulati.

> ***Nota 2:*** *Prima di proseguire con la descrizione dei risultati, mi sembra importante nonché interessante discutere brevemente un aspetto che caratterizza la simulazione* Ns02*: in questo caso, si sono ottenuti valori di concentrazione anche molto elevati* [cfr. tab. 6.4]*, e soprattutto anche valori di concentrazione negativi, chiaramente privi di senso in questo contesto — la distribuzione GEV* non *è limitata infatti a valori positivi.*
>
> *A titolo d'esempio, se si estrae da una distribuzione* $\mathcal{N}(0,1)$ *un valore* $x < -5.123$*, in fase di trasformazione inversa si ottiene un valore negativo di concentrazione! Bisogna anche però dire che la probabilità di estrarre tale valore risulta pari a* 1.51×10^{-7}.
>
> *Per questo,* nel dataset ottenuto dalla simulazione Ns02 ho deciso di sostituire i 5 valori negativi con un valore nullo.

Visivamente, gli istogrammi riproducono delle distribuzioni asimmetriche analoghe a quelle che caratterizzano i dataset reali; ancora, come ci si aspetta nel caso Ns02 e Ns03 la coda superiore risulta più popolata, e in particolare *troppo* (anche in relazione allo specifico valore numerico di concentrazione) per la simulazione basata su GEV fittata sui dati raw (Ns02). Questo risulta evidente anche da quanto riportato nella tabella 6.4, dalla quale si evince come per alcune localizzazioni di stima si ottengano valori medi e/o massimi decisamente troppo elevati in relazione al fenomeno in esame.

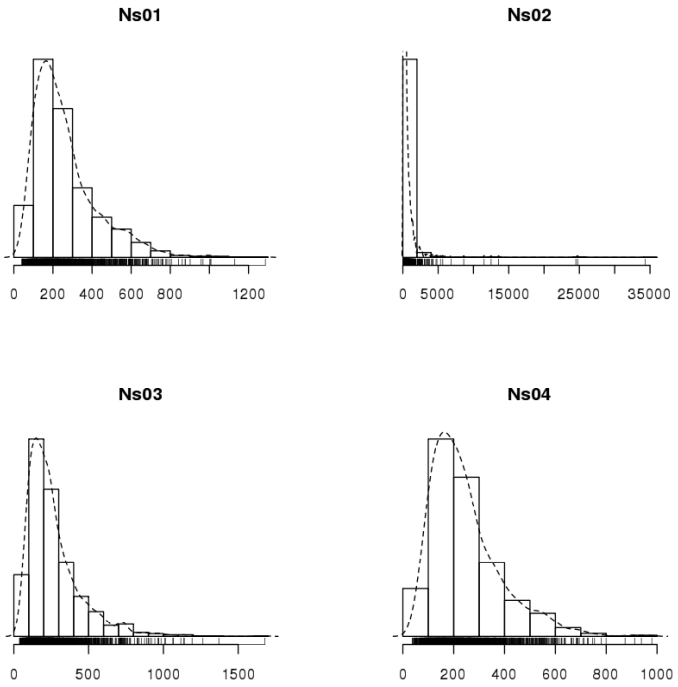


Figura 6.11: Istogrammi per il valor medio di concentrazione di radon indoor (espresso in $Bq \cdot m^{-3}$) delle stime ottenute mediante simulazioni gaussiane sequenziali; la linea tratteggiata rappresenta una stima a kernel della p.d.f. empirica; per maggiori dettagli, si faccia riferimento al testo.

id simulazione	*media*	*mediana*	*minimo*	*massimo*
Ns01	39–1286	27–1089	12–78	216–3210
Ns02	36–34309	26–1265	0–90	194–5697876
Ns03	37–1680	26–984	4–87	153–25808
Ns04	37–979	21–769	0.6–65	237–24907

Tabella 6.4: Valori dei parametri statistici per le varie simulazioni condotte e descritte in §6.5.1; i valori si riferiscono al range su tutti i valori simulati al variare del nodo della griglia di simulazione (statistica basata su 200 valori fissata la localizzazione); si tenga presente che il valore nullo per il minimo della simulazione Ns02 è stato imposto [cfr. testo per maggiori dettagli].

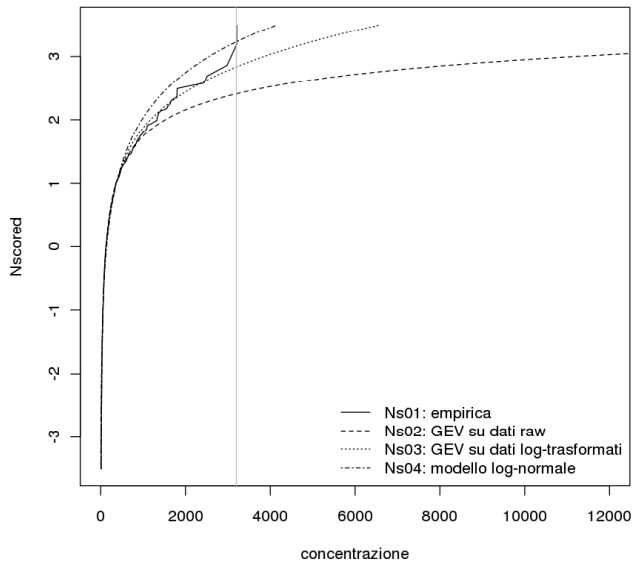


Figura 6.12: Curve utilizzate in fase di trasformazione diretta e inversa tra i dataset operativi e quelli Nscored utilizzati per le simulazioni gaussiane sequenziali.

Questi risultati trovano una spiegazione plausibile se si analizzano le curve utilizzate in fase di trasformazione (sia diretta che inversa) e riportate in figura 6.12: è evidente come le curve relative a Ns02 e Ns03, fissato un valore elevato di tipo Nscored, producano valori anche molto elevati di concentrazione, rispetto alle altre trasformazioni di riferimento; questo non deve certo stupire, visto che sono state introdotte proprio per evitare la rapida decrescita del modello log-normale. Questi effetti *eccessivi* si traducono comunque in poche situazioni facilmente individuabili sulle quali si può intervenire miratamente una volta condotta la serie di simulazioni, eventualmente correggendole nella maniera che si riterrà più opportuna.

In relazione a questi risultati, il modello più realistico sembra essere quello Ns03, in grado di popolare in modo verosimile la coda superiore della distribuzione dei valori simulati e la cui curva di trasformazione risulta, da analisi più specifiche, prossima a quella empirica nella parte centrale ma più efficace per estrarre valori elevati in fase di trasformazione inversa — come sarà chiarito tra breve. Infine, il modello log-normale sembra configurarsi come l'approccio peggiore, tagliando eccessivamente sui valori elevati — piuttosto che imporre questo modello (che peraltro non sembra essere così buono, anche se è il più diffuso in letteratura), sembra quindi preferibile ricorrere a una trasformazione di tipo *empirico*, anche se accompagnata da possibili problemi sulla modellizzazione delle code (in particolare, quella superiore).

La tabella 6.5 riporta alcuni parametri statistici relativi ai dataset simulati per un confronto con quello operativo di riferimento: globalmente, la media del fenomeno è ben riprodotta da tutte le simulazioni, con le già discusse riserve per quanto riguarda la Ns02 [cfr. nota 1]; rimane comunque un effetto di smoothing, come evidenziato dal valore della varianza, ma meno importante rispetto al caso del kriging.

	riferimento	*Ns01*	*Ns02*	*Ns02a*	*Ns02b*	*Ns03*	*Ns04*
media	239	262	534	454	444	266	246
varianza	113247	26851	2268186	441944	339469	33869	19295

Tabella 6.5: Valori (espressi in $Bq \cdot m^{-3}$) relativi ai dataset simulati per i vari tipi di simulazione; *Ns02a* identifica la simulazione basata Ns02 da cui sono stati eliminati i valori superiori a 10000 $Bq \cdot m^{-3}$(6 punti); per *Ns02b* il cut sui valori di concentrazione simulati è stato invece imposto a 6000 $Bq \cdot m^{-3}$(8 punti).

Nota 3: *Sembra interessante concludere la descrizione di questa fase dell'analisi con una nota relativa alla trasformazione Ns01 e al relativo problema di modellizzazione della coda superiore, cui si è in parte già accennato*[9]*. La distribuzione per i valori massimi di Ns01 risulta asimmetrica (coda destra più estesa), come in tutti gli altri casi, ma "stranamente"* bi-modale*, con il picco principale in corrispondenza di un valore pari a circa 3200 $Bq \cdot m^{-3}$.*

Sembra quindi che sopra un certo valore di Nscored, la trasformazione inversa restituisca il medesimo valore nella scala reale di concentrazione, andando così a creare un 'accumulo' in corrispondenza dei 3200 $Bq \cdot m^{-3}$: quest'ipotesi ha trovato conferma conducendo analisi specifiche sul software `Geostat Office`© *impiegato per questi calcoli e osservando la curva riportata in fig. 6.12.*

Tale apparente anomalia ha così permesso di verificare nel concreto possibili ricadute pratiche di implementazione della coda superiore nel caso si ricorra a una trasformazione Nscored basata su c.d.f. empirica — questo artefatto di natura computazionale risulta però trascurabile nel caso (assai più frequente) in cui le sGs vengano impiegate per ottenere le cosiddette 'probability map' [cfr. §6.5.3] *relative al superamento di un determinato valore di soglia — sempre che il valore di soglia scelto sia sufficientemente distante dal citato valore di 'accumulo'.*

Riassumendo quindi quanto discusso in relazione ai risultati ottenuti per le elaborazioni cosiddette 'E-type', si può concludere che:

> *applicare un modello log-normale ai dati di concentrazione di radon indoor in fase di trasformazione Nscored (diretta e inversa)* non *sembra essere una scelta efficace, conseguenza dell'eccessiva rapida caduta della coda superiore della distribuzione; una modellizzazione più aderente alla realtà dell'intera distribuzione si ottiene invece assumendo un modello di tipo GEV applicato ai dati log-trasformati.*

Analisi 'Probability Map'

Ho calcolato mappe di probabilità per il superamento di un determinato valore di soglia di concentrazione di radon indoor al variare del valore di soglia stesso, in una range che va da 200 a 4000 $Bq \cdot m^{-3}$: in questo caso, fintanto che i valori di soglia risultano sufficientemente bassi, non ci si aspettano grandi differenze tra le viarie tipologie di simulazione, visto che questo tipo di mappa è per sua stessa natura poco sensibile alla presenza di valori particolarmente elevati — chiaramente, in relazione al valore di soglia impostato.

In tutti i casi le *zone calde* del fenomeno vengono identificate nello spazio nelle regioni corrette, e si può affermare che:

- per valori di soglia fino a 800 $Bq \cdot m^{-3}$, le 4 tipologie di simulazione non manifestano differenze apprezzabili nelle 'probability map';
- per valori di soglia superiori, come è ragionevole aspettarsi le mappe per Ns02 Ns03 iniziano a mostrare differenze rispetto a Ns01 Ns04;

[9] Per approfondire la questione, si può far riferimento a quanto discusso da Deutsch e Journel {16}, pag. 134–138.

- Ns04 produce un fenomeno più omogeneo e meno variabile rispetto a Ns01: nelle zone calde, i valori di probabilità sono addirittura inferiori; ancora una volta, il modello log-normale rivela i suoi limiti se applicato, in ambito di prevenzione, a dati di concentrazione di radon indoor — o quantomeno, in relazione al dataset operativo di riferimento, caratterizzato dalla presenza di valori elevati piuttosto rispetto alla situazione media nazionale.

6.5.4 Simulazioni e validazione

Non potendo operare delle cross-validation sui risultati ottenuti per le simulazioni gaussiane sequenziali (a seguito delle caratteristiche proprie di questo tipo di approccio), ho invece condotto delle sGs sulle localizzazioni del dataset di validazione [cfr. quanto discusso in §6.5.1]: per ogni valore simulato, si ha così a disposizione anche il reale valore di concentrazione.

La tipologia e i parametri di simulazione, le fasi di trasformazione diretta e inversa, le correzioni per i valori negativi prodotti da Ns02 e le elaborazioni 'E-type' sono le stesse descritte in precedenza.

Una delle peculiarità delle sGs è quella di dover, almeno in linea teorica, essere in grado di riprodurre le caratteristiche *globali* del fenomeno sotto esame; per questo, inizialmente ho analizzato le proprietà statistiche dei residui — definiti come differenza tra valore reale e valore simulato — al variare del tipo di simulazione. Globalmente, si può affermare che:

- in tutti i casi, media e mediana risultano negative, rispettivamente dell'ordine di 10–20 e 70 unità, a dire che si ha una generale *sovrastima* dei valori reali: nell'ottica di una applicazione in campo di tutela della salute e/o di prevenzione, questo 'difetto' credo possa venir accettato;

- i valori di skewness [cfr. eq. (A.11)] risultano positivi (tipicamente, $skw \sim -3$), segnale di una leggera asimmetrica della distribuzione dei residui verso destra e quindi spostata verso le sottostime, a dire che se la simulazione commette errori 'gravi', questi sono per difetto — e questo, al contrario del caso precedente, risulta essere un 'difetto' meno accettabile[10];

- analisi di post-plot per sovra e sotto-stime evidenziano la preponderanza delle prime (come già posto in luce da semplici analisi statistiche descritte in precedenza) sul territorio in esame, ma per entrambe le categorie la distribuzione spaziale risulta essere omogenea e priva di cluster — questo approccio sembra quindi adatto a modellizzare le differenti situazioni locali che si manifestano in Alto Adige;

- conducendo dei fit lineari per grafici del tipo 'valori reali' *vs.* 'valori simulati', come quelli riportati in figura 6.13, si ottengono dei valori dei coefficienti di correlazione pari tipicamente a $\rho \sim 0.3$, valori non proprio incoraggianti ma comunque in linea con quanto ottenuto in lavori precedenti [cfr. ad esempio {42, capitolo 9}].

[10] Va però sottolineato come un'analisi specifica in questa direzione abbia evidenziato come siano poche le situazioni di questo tipo (dell'ordine di qualche unità per ogni simulazione), e in particolare si sono individuate due situazioni comuni a tutte le tipologie di simulazione per le quali il punto di stima è caratterizzato da un valore di concentrazione di radon indoor pari a 3787 e 3794 $Bq \cdot m^{-3}$, con un intorno popolato invece da campionamenti con valori tipici pari rispettivamente a 200 e 300-400 $Bq \cdot m^{-3}$: va da sé che situazioni di questo tipo (la cui entità di errore è chiaramente 'pesante' per la statistica) sono da considerarsi anomale, e che verosimilmente nessun algoritmo o approccio di stima potrà riprodurle in maniera corretta.

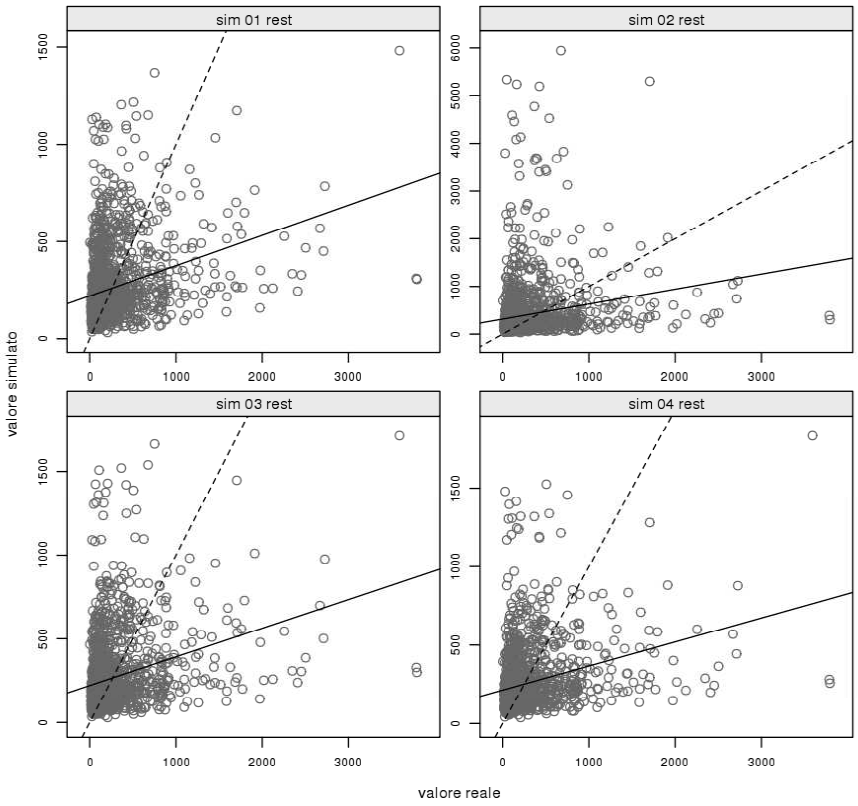


Figura 6.13: Grafici per i valori reali e i corrispondenti valori simulati al variare della tipologia di simulazione; la linea tratteggiata rappresenta la retta di pendenza unitaria, la linea continua la retta ricavata da un fit lineare condotto con i metodi dei minimi quadrati; i valori sono riportati in $Bq \cdot m^{-3}$.

6.6 Conclusioni

Identificare nella maniera corretta le zone a rischio radon (radon prone areas) significa dare la giusta importanza statistica ai valori elevati di concentrazione, ovvero disporre di un modello per la funzione densità di probabilità delle misure di radon indoor appropriato e adatto allo scopo. Attualmente, uno degli approcci più diffusi in letteratura è quello di ricorrere a modelli di tipo Normale/log-Normale, pur essendo noto che queste distribuzioni non sono le più efficaci nell'ambito dell'analisi del rischio, in quanto per loro stessa natura tendono a limitare se non addirittura trascurare il "peso" di eventuali outliers. Per questo, si è deciso di applicare la *Teoria dei Valori Estremi* (EVT) ai dati di radon indoor, per valutare la sua efficacia nella modellizzazione di questo tipo di dati e confrontare i risultati con quelli che si ottengono con l'approccio "standard".

Le analisi sono state condotte sia limitando lo studio alla coda superiore della distribuzione, sia considerando la distribuzione nella sua totalità (in linea con l'approccio Normale/log-Normale); inoltre, la validità e l'applicabilità dei modelli EVT sono stati testati sia sul dataset completo, sia su subset relativi a singoli comprensori. Infine, si è valutata la bontà dell'approccio in esame

applicando la teoria nell'ambito delle simulazioni gaussiane sequenziali (sGs).

I risultati ottenuti in questo contesto d'indagine si possono così riassumere:

- per quanto riguarda la modellizzazione della coda superiore della distribuzione [cfr. §6.3], una volta individuato nella maniera corretta il/i valore/i di soglia u per il valore di concentrazione di radon indoor oltre il quale costruire il modello di tipo GPD, questo risulta in grado di riprodurre nella maniera opportuna il comportamento della coda superiore, evitando di introdurre le pesanti sottostime che caratterizzano invece un approccio di tipo Normale/log-Normale; questi risultati restano validi anche nel caso di singoli comprensori (subset del dataset di riferimento);

- tra i vari modelli GPD disponibili, un decadimento della coda di tipo *esponenziale* risulta compatibile con il comportamento dei dati sperimentiali (sia per il dataset completo, sia per quelli relativi a singoli comprensori): questo risultato sembra importante in quanto consente una modellizzazione corretta della coda superiore con un modello che richiede la stima di un solo parametro — una caratteristica non trascurabile nel caso in cui fosse disponibile un numero esiguo di misurazioni[11];

- per quanto riguarda la modellizzazione dell'intera distribuzione empirica dei valori di concentrazione di radon indoor [cfr. §6.4], rispetto al più convenzionale e diffuso approccio di tipo Normale/log-Normale, la distribuzione GEV derivata dalla teoria dei valori estremi è in grado di descrivere in maniera più aderente alla realtà tanto l'evidente *asimmetria* che caratterizza la distribuzione empirica quanto il comportamento di entrambe le code della stessa; anche in questo caso, sembra importante sottolineare che la validità di questo approccio risulta verificata al variare della numerosità del campione e delle sue caratteristiche di tipo statistico (quali ad esempio media e valori massimi), come dimostrato dall'analisi condotta su specifici comprensori dell'Alto Adige, i quali sono stati scelti in modo da prendere in considerazione differenze evidenti nelle caratteristiche del fenomeno preso in esame;

- per valutare l'impatto di questo tipo di modellizzazioni in una applicazione diretta, si è deciso di porlo a confronto con i modelli "standard" nell'ambito delle simulazioni gaussiane sequenziali, in particolare in relazione alle trasformazioni Nscored (diretta e inversa) che caratterizzano le sGs stesse: applicare un modello log-Normale ai dati di concentrazione di radon indoor non è risultata una scelta efficace per una corretta modellizzazione del fenomeno, conseguenza dell'eccessiva rapida decrescita della coda superiore della distribuzione stessa; un modello di tipo GEV applicato ai dati log-trasformati consente invece di ottenere una rappresentazione più aderente alla realtà; si sottolinea comunque come buoni risultati si ottengano anche nel caso in cui si ricorra alla p.d.f. empirica in fase di trasformazione, soprattutto se l'output richiesto è una mappa di probabilità.

[11] Si tenga infatti presente che tutte le misurazioni inferiori al valore di soglia non andranno a far parte del dataset impiegato per analisi di questo tipo.

CAPITOLO 7

Analisi Esplorativa delle Variabili Categoriche

L'idea che ha stimolato questa analisi esplorativa è stata quella di indagare in maniera sistematica alcuni aspetti statistici che caratterizzano le numerose variabili categoriche che accompagnano la singola misura di concentrazione di radon indoor, con lo scopo di porre chiarezza sulle informazioni che tali variabili potrebbero celare, in relazione alla loro possibile influenza sulla misura di concentrazione stessa.
L'intenzione è stata quella di suddividere lo studio in tre fasi principali, in relazione al tipo *di informazione che di volta in volta di cercherà di porre in primo piano:*

i) *analisi di* frequency table *e* density plot *(per il valore di concentrazione) al variare delle singole classi per ogni variabile categorica, controllando anche eventuali significative differenze nei valori medi* ↦ come si distribuiscono le numerosità delle classi per una data variabile? Le singole classi sono sufficientemente rappresentate? Il valore di concentrazione mostra evidenti correlazioni con una data variabile/classe?

ii) *analisi della* distribuzione spaziale *delle singole classi* ↦ si evidenziano particolari pattern sul territorio di studio? La distribuzione risulta omogenea?

iii) *analisi di eventuali* correlazioni *tra coppie di variabili (utile per capire se variabili diverse portano in realtà lo stesso tipo di informazione)* ↦ quali variabili andranno incluse in un ipotetico modello di previsione e quali potranno eventualmente omesse?

La parte computazionale è stata svolta interamente in ambiente `R`{44} ricorrendo ai packages *lattice*{45}, *maptools*{35} e *vcd*{15}.

7.1 Brevi richiami teorici

Per indagare e spiegare dati categorici multidimensionali, è pratica comune ricercare strutture di indipendenza nelle variabili. Che lo scopo sia puramente esplorativo o basato su un qualche mo-

dello, tecniche quali MOSAIC PLOT e ASSOCIATION PLOT offrono senza dubbio un valido supporto: entrambe visualizzano aspetti legati all'interpretazione di tabelle di contingenza, e recentemente si sono arricchite di numerose estensioni (una delle quali sarà brevemente descritta nel paragrafo §7.1.2).

In questa breve sezione ho ritenuto utile richiamare alcuni aspetti teorici su cui le analisi successive si fondano.

7.1.1 Test

Si consideri una tabella di contingenza — utilizzata solitamente in statistica per rappresentare e analizzare le relazioni tra due o più variabili — con frequenza di cella $[n_{ij}]$ per $i = 1, \cdots, I$ e $j = 1, \cdots, J$ e somme per righe e colonne date rispettivamente da $n_{i+} = \sum_i n_{ij}$ e $n_{+j} = \sum_j n_{ij}$. Data una distribuzione per il fenomeno sotto esame con una probabilità di cella teorica data da π_{ij}, l'ipotesi nulla H_0 di *indipendenza* per le due variabili categoriche può essere formulata in questi termini:

$$H_0 \quad : \quad \pi_{ij} = \pi_{i+} + \pi_{+j} \tag{7.1}$$

Il valore atteso per la frequenza di cella in questo modello è dato da $\widehat{n}_{ij} = n_{i+}n_{+j}/n_{++}$. La più nota e usata misura della discrepanza tra valore atteso e valore osservato è il residuo di Pearson, definito come

$$r_{ij} = \frac{n_{ij} - \widehat{n}_{ij}}{\sqrt{\widehat{n}_{ij}}} \tag{7.2}$$

Segue che la maniera più conveniente per aggregare i $I \times J$ residui in una singola variabile 'test statistic' è data dalla loro somma quadratica:

$$\chi^2 = \sum_{i,j} r_{ij}^2 \tag{7.3}$$

in quanto è noto che, sotto le condizioni dell'ipotesi nulla H_0, ammette come distribuzione limite la distribuzione χ^2 con $(I-1)(J-1)$ gradi di libertà. Questo risulta essere infatti il ben noto test del χ^2, solitamente introdotto in tutti i casi in cui si fa riferimento al problema dell'indipendenza per tabelle a 2 entrate.

Tuttavia, la (7.3) non è l'unico modo plausibile per catturare le deviazioni dallo zero nei residui. Esistono infatti altri funzionali $\lambda(\cdot)$ in grado di svolgere la medesima funzione e che portano a variabili 'test statistic' $\lambda([r_{ij}])$ altrettanto ragionevoli : uno dei più indicati per l'identificazione delle celle responsabili della *dipendenza* — o, detto in altri termini, di uno scostamento *significativo* dall'indipendenza — è il massimo dei valori assoluti dei residui:

$$M = \max_{i,j} |r_{ij}| \tag{7.4}$$

Dato un valore critico pari a c_α per tale 'test statistic', tutti i residui il cui valore assoluto eccede c_α violano l'ipotesi nulla di indipendenza con un livello di significanza pari proprio a α [cfr. {39, cap. 7}]; in questo modo, le celle di interesse, cioè quelle responsabili dell'eventuale dipendenza, possono venir facilmente individuate.

7.1.2 Visualizzazioni

Le due tecniche visuali più diffuse per l'analisi dell'indipendenza nelle tabelle di contingenza a 2 entrate sono i cosiddetti *mosaic plot* e *association plot*. Entrambi risultano utili per mettere in luce differenze tra una tabella osservata $[n_{ij}]$ e quella attesa $[\widehat{n}_{ij}]$ in modo grafico; la prima tipologia si concentra sulla visualizzazione delle frequenze osservate n_{ij}, mentre la seconda sulla visualizzazione dei residui di Pearson r_{ij}. Nello specifico:

Mosaic Plot [1] — altezza e larghezza dei rettangoli rappresentati sono in relazione con le frequenze relative delle due variabili in esame; consiste semplicemente in un insieme di aree rettangolari la cui area è proporzionale alla frequenza osservata per quella determinata cella. Un rettangolo corrispondente al 100% delle osservazioni viene inizialmente diviso orizzontalmente rispetto alle frequenze della prima variabile e successivamente verticalmente rispetto alle frequenze condizionate della seconda variabile;

Association plot [2] — visualizzano una tabella dei residui di Pearson [cfr. eq. (7.2)]: ogni cella è rappresentata da un rettangolo la cui altezza (con segno) risulta proporzionale al corrispondente residuo r_{ij} e la cui larghezza risulta proporzionale alla radice quadrata dei conteggi attesi $\sqrt{\widehat{n}_{ij}}$; con questa costruzione, l'area della specifica cella risulta proporzionale al residuo grezzo $n_{ij} - \widehat{n}_{ij}$;

Interessante, come sarà discusso nel paragrafo 7.5, notare come questo insieme di approcci grafici possa venir esteso per implementare nella stessa rappresentazione anche parametri statistici relativi a test statistici sulla *significatività* dell'eventuale dipendenza riscontrata.

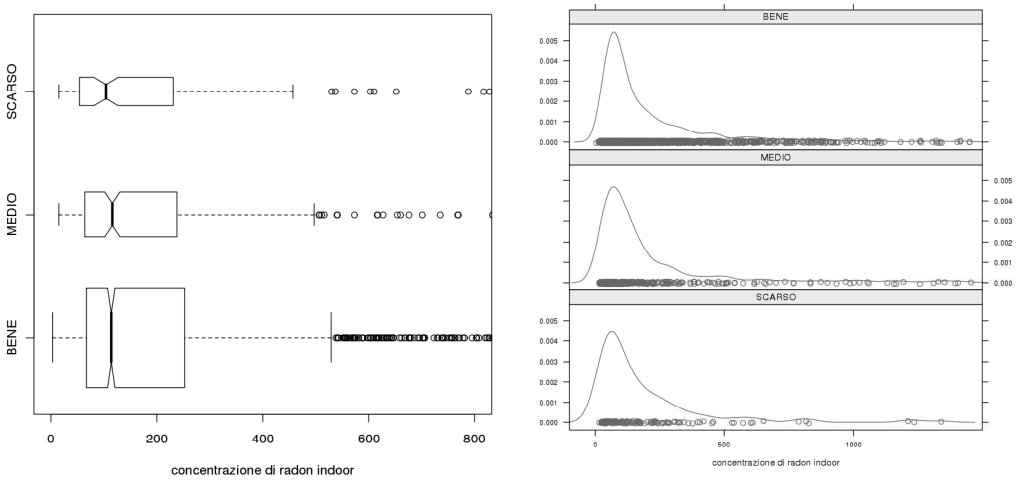
7.2 Descrizione del dataset utilizzato

Il dataset operativo su cui si basano le analisi descritte in questa fase è stato estratto da quello generale di riferimento e descritto nel paragrafo §1.3 a pagina 7; i valori di concentrazione, georeferenziati, non hanno subito alcuna correzione, e sono pertanto relativi al singolo semestre di esposizione.

Ho preventivamente estratto solo i cases per i quali fossero disponibili *tutti* i valori per ognuna delle variabili categoriche prese in esame, che nello specifico sono quelle relative alla parte per così dire *antropogenica* del fenomeno (trascurando quindi la parte *geologica*[3]), ovvero[4]:

- utilizzo;
- tipo locale;
- classe data di costruzione;
- tipo di costruzione;
- qualità degli infissi;
- contatto con il terreno;

[1] Un esempio di tale grafico è riportato nella figura 7.3a, a pagina 109.
[2] Un esempio di tale grafico è riportato nella figura 7.3b, a pagina 109.
[3] Questa scelta è stata dettata da due principali ragioni: i) il 22% circa dei valori di concentrazione non hanno informazioni relative alla parte geologica e ii) il legame/influenza tra misure di radon *indoor* e geologia non è unanimemente riconosciuto in letteratura — il dibattito su questo tema è ancora aperto e 'attivo'.
[4] Per una descrizione dettagliata delle singole variabili e del loro significato, si rimanda a quanto discusso nel paragrafo §1.3, a pagina 7.



(a) box plot

(b) density plot

Figura 7.1: Esempio delle rappresentazioni grafiche utilizzate per lo studio esplorativo delle variabili categoriche — in questo caso, i grafici si riferiscono alla variabile *qualità degli infissi*; per quanto riguarda i box plot, la larghezza del box è proporzionale alla radice quadrata del numero di osservazioni per la relativa classe; i valori di concentrazione sono riportati in $\mathrm{Bq \cdot m^{-3}}$; per chiarezza di visualizzazione, nel grafico (a) l'asse x ha subito un taglio in corrispondenza degli 800 $\mathrm{Bq \cdot m^{-3}}$, il grafico (b) in corrispondenza di 1500 $\mathrm{Bq \cdot m^{-3}}$.

- esposizione.

La citata fase di pre-processing ha ridotto la numerosità del dataset operativo a **2308** cases.

7.3 Frequency table, box plot e p.d.f.

Questa prima fase di analisi esplorativa si è concentrata sullo studio della distribuzione della numerosità delle misure di radon indoor al variare della classe di ogni variabile categorica presa in esame, con lo scopo principale di indagare nello specifico la presenza di variabili con classi particolarmente *sbilanciate*; questo, ricorrendo a FREQUENCY TABLE [cfr. tab. 7.1] e a loro rappresentazioni grafiche.

Tra le variabili prese in esame, solo *classe data di costruzione* ha una distribuzione inter-classe piuttosto uniforme; in tutti gli altri casi, ho riscontrato la presenza di almeno una classe dominante in maniera significativa sulle altre. Questo aspetto è da prendere in considerazione con la dovuta attenzione qualora si volessero 'istruire' dei modelli con le informazioni contenute nelle variabili categoriche, in quanto i risultati andranno comunque interpretati considerando che alcune classi potrebbero essere in qualche modo 'ignorate' in relazione alla loro scarsa rappresentatività.

Sfruttando rappresentazioni grafiche come quelle riportate, a titolo d'esempio, nella figura 7.1, ho quindi analizzato l'eventuale presenza di evidenti e significative differenze i) nei valori medi di concentrazione (che sono anche stati calcolati numericamente) al variare della classe e ii) nella forma della p.d.f. associata:

classe	*N*	*%*
abitazione	1951	84.5
lavoro	187	8.1
lavoro/scuola	15	1
scuola	155	6.7

(a) utilizzo

classe	*N*	*%*
<1850	364	15.7
1850–1940	264	11.4
1941–1970	596	25.8
191–1985	618	26.8
>1985	466	20.1

(b) classe data costruzione

classe	*N*	*%*
cemento	65	2.8
legno	20	0.9
mattoni	1484	64.3
prefabbricato	1	0.04
sassi	738	31.9

(c) tipo costruzione

classe	*N*	*%*
bene	1788	77.4
medio	370	16.1
scarso	150	6.5

(d) qualità degli infissi

classe	*N*	*%*
no	1460	63.3
si	848	36.7

(e) contatto con il terreno

classe	*N*	*%*
aula	143	6.2
camera da letto	511	22.1
cantina	45	1.9
corridoio	3	0.1
cucina	606	26.2
negozio	5	0.2
sala	30	1.3
salotto	742	32.2
stanza	29	1.3
ufficio	144	6.2
altro	50	2.2

(f) tipo locale

classe	*N*	*%*
east	310	13.4
nord	158	6.9
nord-est	143	6.2
nord-ovest	176	7.6
sud	465	20.2
sud-est	381	16.5
sud-ovest	638	16.7
ovest	265	11.5
flat	25	1.1

(g) esposizione

Tabella 7.1: Frequency table per le variabili categoriche impiegate nell'analisi e descritte nel paragrafo §7.2.

box plot — ho sfruttato utili e ricche rappresentazioni grafiche come quella riportata in figura 7.1a: la larghezza del box è proporzionale alla radice quadrata del numero di osservazioni per la specifica classe e attorno al valore della mediana (riga nera all'interno del box) vengono disegnati due "tagli": se questi due tagli, per due box differenti, non si sovrappongono, questo è indice di una differenza tra i valori delle mediane significativa; inoltre, si ottengono visivamente informazioni sui valori dei quartili e sulla distribuzione degli eventuali outliers. Da queste e da altre analisi condotte, si può concludere che, in relazione ai valori medi:

- *in inverno si registrano valori più elevati di concentrazione (questo è ragionevole in virtù delle differenti abitudini abitative rispetto al semestre estivo);*
- *le cantine hanno valori più elevati (anche questo in linea con le caratteristiche di questo tipo di locale: scarsa o nulla isolazione e tipicamente a diretto contatto con il terreno, anche per quanto riguarda le pareti);*
- *le abitazioni in contatto manifestano concentrazioni più elevate, come quelle di sassi (anche se quest'ultima conclusione non è così evidente come negli altri casi);*
- *a differenza di quanto ci si poteva aspettare prima di condurre l'analisi, sulla base di modelli di diffusione e in relazione alle abitudini abitative, le differenze relative alla qualità degli infissi* non *risultano particolarmente evidenti, e inoltre tra le classi "bene" e "medio" risultano praticamente nulle*[5] [cfr. fig. 7.1].

density plot — ricorrendo a rappresentazioni grafiche come quella riportata in figura 7.1b, ho confrontato le stime a kernel della p.d.f. per il valore di concentrazione di radon indoor al variare della classe di ogni singola variabile categorica, con l'idea di ricercare eventuali differenze o anomalie nella forma della p.d.f. stessa; al variare tanto della variabile che delle relative classi, la forma appare sempre come una distribuzione asimmetrica, che si estende verso destra, e che ricorda una distribuzione di tipo log-Normale[6]: questa caratteristica dei dati sembra quindi essere indipendente dalla variabile e anche dalla particolare classe della stessa, e configurarsi quindi come *propria* dei dati in esame — e per estensione, anche in relazione all'esperienza condotta su altri dataset, propria del fenomeno 'radon indoor'.

7.4 Distribuzione spaziale

Per ogni variabile categorica, ho visualizzato la distribuzione spaziale dei campionamenti in funzione della singola classe di appartenenza, con lo scopo di individuare eventuali cluster o pattern caratteristici sul territorio altoatesino; un esempio di tali rappresentazioni è riportato in figura 7.2

Questo tipo di analisi credo possa risultare particolarmente utile nel caso in cui si intenda implementare un modello di tipo geostatistico — o comunque un modello che consideri anche la componente spaziale del fenomeno — al quale fornire ulteriori informazioni (rispetto alle sole

[5] Questo risultato ha stimolato analisi ulteriori in questa direzione, con lo scopo di individuarne la causa; tuttavia, non si sono ottenuti risultati soddisfacenti e univoci. Un'ipotesi plausibile è che i valori di questa variabile non siano stati raccolti con la dovuta attenzione e/o uniformità di giudizio.

[6] Questa è infatti la distribuzione cui tipicamente si fa riferimento in letteratura; tuttavia, questo tipo di assunzione è discutibile e non sempre accettabile, come ampiamente descritto e discusso, in relazione ai dati impiegati per questo lavoro di dottorato, nel capitolo 6.

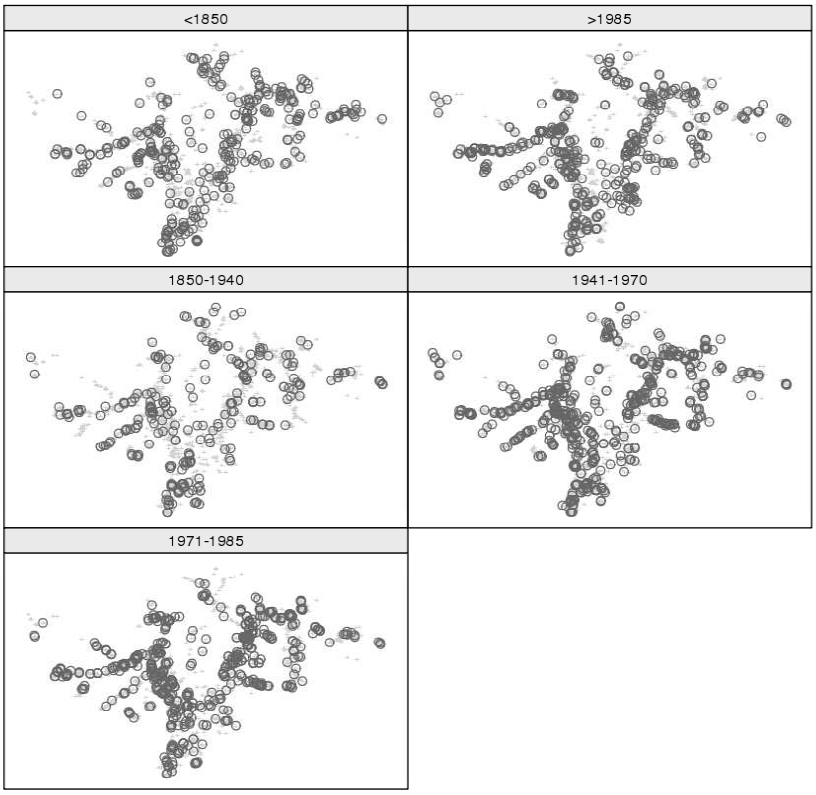


Figura 7.2: Post plot dei campionamenti al variare della classe per la variabile categorica *classe data di costruzione*; le croci in grigio rappresentano le localizzazioni di tutti i punti che costituiscono il dataset operativo, i cerchi quelli relativi alla classe indicata nella strip sopra ogni grafico.

coordinate) legate appunto alle caratteristiche proprie dell'edificio — che ritengo siano parametri la cui influenza sul valore di concentrazione misurato non possa venir trascurata, anche se la modellizzazione e l'entità di tale influenza risultano ancora distanti da una facile e chiara comprensione. Qualche idea e qualche prima implementazione in questa direzione si possono trovare nella parte III di questo lavoro.

Il risultato ottenuto in questo contesto è positivo, in quanto

> *dal punto di vista spaziale, i subset relativi alle varie classi per ogni singola variabile categorica*[7] *di tipo* antropogenico *presa in esame manifestano distribuzioni spaziali* omogenee *su tutto il territorio dell'Alto Adige.*

[7] Può essere interessante discutere un risultato emerso a proposito della variabile *tipo di costruzione*: per quanto riguarda la classe *cantina*, che coinvolge solo il 2% del campionamenti, si nota come questi ultimi si vadano a concentrare nelle zone 'calde' del fenomeno (quelle che manifestano i valori più elevati), e come siano inoltre tutte misure relative al semestre invernale. Certo, questo risultato non stupisce, ma potrebbe essere indice di una lieve *polarizzazione* verso valori elevati per queste zone — anche se sicuramente non ne costituisce la causa principale.

7.5 Correlazioni a coppie

Si è già accennato a come l'impiego di *mosaic* e *association plot* possa rendere l'analisi della presenza di eventuali correlazioni tra variabili di tipo categorico più semplice ed efficace [cfr. §7.1.2]. Per questa ultima fase dell'analisi, si è fatto ricorso a una interessante estensione delle tecniche classiche e implementata nel package *vcd*. In particolare, la colorazione di mosaic e association plot basata sui residui per i modelli di indipendenza relativi alle tabelle di contingenza è stata estesa in due direzioni:

a) utilizzo di colori HCL (Hue-Chroma-Luminace) che si adattano meglio alla percezione fisiologica dell'osservatore/utente

b) codifica dei risultati relativi ai test per la rilevanza statistica direttamente nei mosaic e association plot in base a quanto ottenuto in a).

Per quest'ultima implementazione, i valori di soglia che regolano il cambio di colore sono determinati mediante un cosiddetto *data-drive approach* basato sulle permutazioni condizionate della distribuzione di una test statistic del tipo riportato nell'equazione (7.4).

La figura 7.3 riporta un esempio di questo tipo di visualizzazioni, e in particolare un mosaic plot [cfr. fig. 7.3a] e un association plot [cfr. fig. 7.3b] per lo studio della correlazione tra le variabili categoriche *classe data di costruzione* e *tipo di costruzione*. Eventuali correlazioni statisticamente significative sono evidenziate mediante opportuni colori delle singole celle, e nello specifico ricorrendo a due differenti livelli di significatività, pari al 90 e al 99%; ancora, il colore è inoltre indice della *direzione* della deviazione rispetto alla frequenza attesa — a dire, tinte verso il blu indicano un valore osservato maggiore di quello atteso, tinte verso il rosso la situazione contraria.

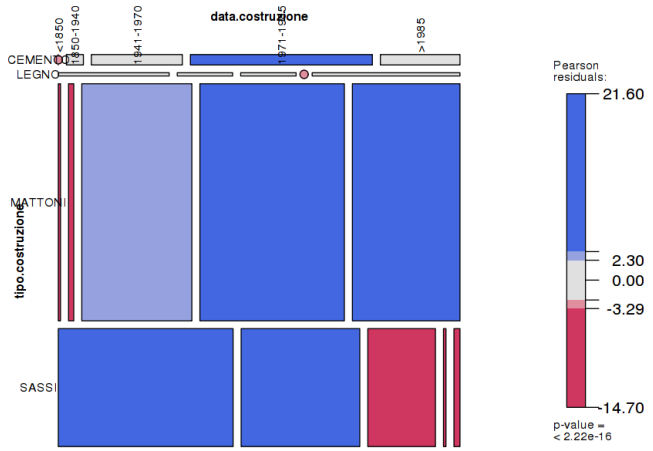
Ho quindi indagato nello specifico le singole correlazioni a coppie per le variabili categoriche antropogeniche selezionate per l'analisi descritta in questo capitolo e riportate nel paragrafo §7.2; preventivamente, per semplificare le visualizzazioni, ho escluso dal dataset l'unica misurazione per la quale il materiale di costruzione è *prefabbricato*.

I risultati ottenuti sono in linea con quello che ci può aspettare sulla base dell'esperienza acquisita e dei meccanismi più elementari che regolano le dinamiche del trasferimento del gas radon nelle abitazioni; a titolo d'esempio, si riportano alcune osservazioni che hanno trovato piena conferma nelle analisi oggettive appena descritte:

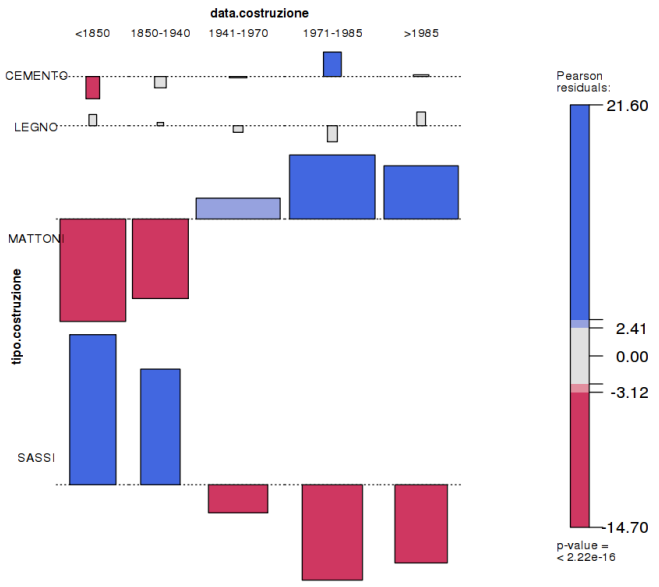
- gli edifici costruiti dopo il 1940[8] risultano correlati positivamente con la classe *mattoni* per la variabile 'tipo di costruzione', mentre correlati negativamente per la classe *sassi*; situazione complementare invece (e ragionevolmente) per gli edifici costruiti in epoche più vecchie — informazioni che si possono ricavare, in maniera simile, da quanto riportato come esempio nelle figure 7.3a e 7.3b;

- le *cantine* risultano correlate con una qualità degli infissi *scarsa*, mentre gli edifici costruiti con *mattoni* risultano significativamente correlati con una qualità degli infissi *buona* e quelli costruiti con *sassi* con una qualità degli infissi *scarsa* — certo, sono risultati che erano facilmente prevedibili, ma una conferma di tipo oggettivo va a favore delle tecniche prese in esame; inoltre, è interessante come questi stessi risultati siano confermati anche dall'analisi delle correlazioni tra *classe data costruzione* e *qualità infissi*, in quanto è stata riconosciuta anche la correlazione tra *classe data costruzione* e *tipo di costruzione*;

[8] Si tenga presente che il tipo di classificazione per la variabile *classe data di costruzione*, dettato comunque da cambiamenti più o meno riconosciuti nelle tecniche edilizie, non è certo univoco.



(a) mosaic plot



(b) association plot

Figura 7.3: Esempio delle visualizzazioni grafiche cui si è fatto ricorso per lo studio delle correlazioni tra coppie di variabili categoriche, in questo caso tra *classe data di costruzione* e *tipo di costruzione*: risulta evidente come sia statisticamente significativo il fatto che per gli edifici più "nuovi" il materiale da costruzione più diffuso sia "mattoni", mentre per quelli più "vecchii" sia "sassi"; i livelli di significatività statistica per i due livelli indicati (corrispondenti ai cambi di colore) sono rispettivamente pari al 90 e al 99%.

- come ultima curiosità, è emerso che le *scuole* sono correlate in maniera positiva con la classe per la data di costruzione *1850–1940*.

7.6 Conclusioni

Come ampiamente discusso nel paragrafo §1.3 a pagina 7, ogni singolo valore di concentrazione di attività di radon indoor è accompagnato da una numerosa serie di variabili secondarie, di natura sia quantitativa che qualitativa, variabili che caratterizzano in maniera più o meno rilevante la misura di concentrazione stessa. Si è quindi ritenuto opportuno e utile condurre delle analisi specifiche sulle varibili categoriche, con lo scopo di indagare le loro proprietà in relazione alle singole classi che le costituiscono, l'omogeneità della distribuzione spaziale delle stesse e la presenza di eventuali significative correlazioni tra coppie di variabili.

Ricorrendo a strumenti quali frequency tables, box plot, density plot, association e mosaic plot, è emerso che:

- se si eccettua la variabile 'classe data costruzione', in tutti i casi vi è la presenza di una classe dominante sulle altre per quanto riguarda la numerosità; questo aspetto dovrà ricevere la dovuta attenzione qualora si costruissero dei modelli che coinvolgano anche alcune variabili categoriche, tanto in fase di costruzione del modello stesso quanto in fase di interpretazione dei risultati;

- come ci si poteva aspettare in base a ragionamenti teorici, i valori più elevati di concentrazione sono associati a misure condotte in inverno, nelle cantine e in edifici a diretto contatto con il terreno;

- contrariamente a quanto si può prevedere in base a modelli di diffusione e in relazione alle abitudini abitative, la variabile 'qualità degli infissi' non manifesta differenze significative al variare delle classi che la caratterizzano; questa variabile non sembra quindi essere particolarmente predittiva in relazione al valore di concentrazione misurato, anche se da un punto di vista prettamente teorico ci si sarebbe aspettata un'influenza più marcata;

- per quanto riguarda la p.d.f. dei valori di concentrazione, questa appare sempre, al variare sia della variabile che delle relative classi, come una distribuzione asimmetrica che si estende verso destra: questa caratteristica dei dati sembra quindi configurarsi come propria del fenomeno radon indoor[9];

- per quanto riguarda le eventuali correlazioni a coppie, i risultati sono in linea con quanto ci si può aspettare in base all'esperienza acquisita e ai meccanismi più elementari che regolano le dinamiche di trasferimento del gas radon all'interno degli edifici — a titolo d'esempio, si sono riconosciute correlazioni significative tra le cantine e una qualità degli infissi scarsa, oppure tra edifici costruiti con mattoni e una buona qualità degli infissi; questi risultati potranno essere utili nel caso in cui si intenda costruire un modello che preveda il ricorso a variabili secondarie, in quanto sarà opportuno, nell'ottica di una economia del modello stesso, evitare di utilizzare variabili fortemente correlate tra loro, in modo da non complicare inutilmente il modello con informazioni ridondanti — situazione che tra l'altro potrebbe rendere meno affidabili e di più difficile interpretazione i risultati ottenuti.

[9] Analisi e discussioni specifiche in merito alla corretta modellizzazione di questa distribuzione si possono trovare nel capitolo 6, a pagina 75.

Inoltre, ricorrendo a visualizzazioni della distribuzione spaziale dei campionamenti relativi alle varie classi per ogni variabile categorica considerata, si evicen come questi si distribuiscano in maniera *omogenea* sull'intero territorio altoatesino, senza mostrare la presenza di cluster in alcune zone specifiche: questa si ritiene essere una caratteristica importante nel caso in cui si intenda costruire un modello che consideri la componente spaziale del fenomeno e che al contempo preveda di sfruttare l'informazione contenuta nelle variabili secondarie[10], garantendo che tale informazione risulta spazialmente rappresentativa di tutto il dominio di studio.

[10]Modelli di questo tipo verranno descritti e analizzati nel dettaglio nella parte III del presente lavoro di tesi, e nello specifico nei capitoli 10 a pagina 145, 11 a pagina 161 e 13 a pagina 185.

CAPITOLO 8

La regressione dei quantili

Sono stati fondamentalmente due gli spunti offerti dall'approccio della regressione dei quantili che hanno reso interessante questa tecnica statistica e che hanno indirizzato l'analisi presentata in questo capitolo:

i) la possibilità di valutare se le variabili secondarie (fattori) che accompagnano la misura di concentrazione di radon indoor e che caratterizzano l'edificio sede della misura abbiano una influenza costante sul valore di concentrazione in funzione del quantile della distribuzione associata — a dire, se la loro eventuale influenza si manifesti in maniera costante su tutto il range di valori di concentrazioni o se, ad esempio, si renda evidente solo per valori elevati;

ii) la possibilità di costruire dei modelli 'di riferimento' selezionando in modo opportuno le classi dei vari fattori coinvolti, in base ai quali valutare quale sia l'influenza delle altre classi/fattori in relazione a quelle che compaiono nel riferimento — a dire, la possibilità di costruire un modello per una "casa tipo", ad esempio con qualità degli infissi buona, e valutare quindi se ed eventualmente in quale misura venga influenzato il valore di concentrazione se la qualità degli infissi risulta ad esempio scarsa.

La parte computazionale è stata svolta in ambiente `R`{44}, ricorrendo ai packages *quantreg*{32}, *lattice*{45}, *vcd*{15} e *gstat*{41}.

8.1 Brevi richiami teorici

Sia y la variabile di interesse (o variabile principale), e $\mathbf{x}$ il vettore delle variabili secondarie che si crede la influenzino, essendo cioè $y = f(\mathbf{x})$. Nell'approccio classico della regressione dei minimi quadrati (LS, Least Squares), avere a disposizione la funzione per la media condizionata, ovvero la funzione che descrive come cambia il valor medio di y in relazione al vettore delle variabili

LS

secondarie $\mathbf{x}$ è solitamente tutto quello che si richiede per conoscere quale sia la relazione tra $\mathbf{x}$ e y.

Il punto chiave dell'approccio LS è che si assume che il vettore delle variabili secondarie abbia effetto solo sulla *media* della distribuzione di probabilità condizionata di $(y|\mathbf{x})$ e su nessun altro parametro, come ad esempio varianza o forma. Se la situazione è realmente questa, allora si può essere pienamente soddisfatti con la stima del modello per la media condizionata. Ma ci possono essere, e nella realtà ci sono, soprattutto in contesti quali quello ambientale o biologico in cui le interazioni tra le variabili in gioco, quando non completamente mascherate, possono essere anche molto complesse, situazioni per le quali le variabili secondarie possono influenzare la p.d.f. condizionata della variabile principale in una miriade di modi, andando ad esempio a modificare la varianza in funzione del valore della variabile principale stessa (fenomeno dell'eteroschedasticità, tipico del tra l'altro fenomeno radon indoor), allargando o contraendo la/e coda/e della distribuzione o ancora introducendo delle multi-modalità. Tutti questi aspetti possono venir efficacemente indagati ricorrendo agli strumenti offerti dalla REGRESSIONE DEI QUANTILI (QR, Quantile Regression), introdotta da Koenker e Basset {33} nel 1978 e che estende l'idea introdotta dalla regressione dei minimi quadrati alla stima della *distribuzione condizionata dei quantili* — un modello nel quale i quantili della distribuzione condizionata della variabile di interesse sono espressi in funzione del vettore delle variabili secondarie.

QR

Tra i vantaggi dell'approccio QR, ricordo quelli più utili in questo contesto:

- risulta efficace e utile in quelle situazioni per le quali la dispersione dei dati *non* è costante, ma varia con il dato stesso (situazioni di eteroschedasticità);
- rispetto all'approccio LS, è meno sensibile alla presenza di eventuali outliers o punti anomali;
- avendo a che fare con quantili (e non con la media), è possibile applicare ai dati di partenza qualsiasi tipo di trasformazione *monotona* senza per questo avere problemi in fase di trasformazione inversa.

Formalmente, come si definisce nel contesto LS la media campionaria come la soluzione di un problema di minimizzazione per la somma quadratica dei residui, è possibile definire la *mediana* come la soluzione di un problema di minimizzazione per la somma dei valori assoluti dei residui. La simmetria stessa della funzione 'valore assoluto' implica che la minimizzazione appena descritta uguagli il numero di residui positivi e negativi, assicurando così che ci sarà lo stesso numero di osservazioni sopra e sotto il valore della mediana.

E per quanto riguarda il caso generale di un altro quantile? Poiché è la simmetria stessa del valore assoluto che porta direttamente alla mediana, minimizzare una somma pesata del valore assoluto *a-simmetrico* dei residui — semplicemente dando un peso diverso ai residui positivi e a quelli negativi — potrebbe portare allora ai quantili; e questo è effettivamente quello che accade.

Risolvendo infatti la seguente equazione

$$\min_{\xi \in \mathbb{R}} \sum_i \rho_\tau (y_i - \xi) \tag{8.1}$$

dove la funzione $\rho(\cdot)$, riportata in figura 8.1, è una sorta di funzione valore assoluto "inclinata", si ottiene la stima del τ-esimo quantile campionario — infatti, per $\tau = 0.5$ quello che si ottiene è proprio la funzione valore assoluto, che guarda caso porta alla stima della mediana.

A questo punto, l'approccio LS viene in aiuto nel proporre un modello per la stima dei quantili condizionati alle variabili secondarie. Sia $(y_1, y_2, \ldots, y_N)$ un insieme di osservazioni casuali;

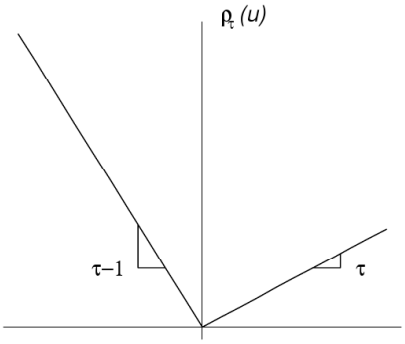


Figura 8.1: La funzione ρ che compare nell'equazione (8.1).

risolvendo l'equazione

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{N} (y_i - \mu)^2 \tag{8.2}$$

dove μ rappresenta la media, si ottiene la media campionaria, una stima della media *non*-condizionata della popolazione, $E(Y)$. Se si sostituisce lo scalare μ con una funzione parametrica del tipo $\mu(x, \beta)$ e si risolve l'equazione

$$\min_{\beta \in \mathbb{R}^+} \sum_{i=1}^{N} [y_i - \mu(x, \beta)]^2 \tag{8.3}$$

quello che si ottiene è la stima per il valore di aspettazione condizionato $E(Y|\mathbf{x})$.

Nella regressione dei quantili, si procede esattamente allo stesso modo; per ottenere una stima della mediana condizionata, ad esempio, è sufficiente sostituire lo scalare ξ dell'eq. (8.1) con la funzione parametrica $\xi(x_i, \beta)$ e imporre $\tau = 0.5$. Per ottenere stime condizionate per gli altri quantili, sarà quindi sufficiente sostituire la funzione valore assoluto con la funzione $\rho(\cdot)$ e risolvere infine l'equazione

$$\min_{\beta \in \mathbb{R}^+} \sum_{i=1}^{N} \rho_\tau \left[y_i - \xi(x, \beta) \right] \tag{8.4}$$

Se $\xi(x_i, \beta)$ viene formulata come una funzione lineare dei parametri, il problema di minimizzazione che ne risulta può facilmente essere risolto in maniera molto efficiente con metodi computazionali ormai diffusi e consolidati.

8.2 Descrizione del dataset utilizzato

Il dataset di partenza è lo stesso utilizzato per le analisi condotte e discusse nel capitolo 7, e descritto nel paragrafo §7.2, a pagina 103; i valori di concentrazione sono riferiti al semestre invernale (nessuna correzione) e sono accompagnati da tutti i valori per le variabili secondarie (fattori) considerate, che sono solo quelle di tipo *antropogenico*, ovvero che caratterizzano l'edificio sede della misura. Da questo dataset, ho eliminato alcune classi per i) facilitare le successive interpretazioni e ii) la scarsa numerosità che le caratterizza; in particolare[1]:

[1] Per una descrizione dettagliata delle variabili citate, si rimanda a quanto riportato nel paragrafo §1.3, a pagina 7.

fattore	*classe di riferimento*	*peso relativo*
contatto	no	60%
qualità infissi	bene	80%
utilizzo	abitazione	86%
tipo di costruzione	mattoni	70%
classe data costruzione	>1985	21%
tipo locale	camera da letto	21%
esposizione	sud	20%

Tabella 8.1: Descrizione delle caratteristiche della 'casa tipo': per ogni fattore, è riportata la classe di riferimento e il suo peso relativo in percentuale rispetto al totale.

- *utilizzo*: eliminata la classe 'lavoro/scuola' (0.82%);
- *tipo costruzione*: eliminate le classi 'cemento' (1.6%) e 'legno' (0.77%);
- *tipo locale*: eliminate le classi 'altro' (2.4%), 'sala' (1.3%), 'stanza' (1.2%), 'cantina' (0.33%) e 'negozio' (0.27%);

Dopo questa ulteriore selezione dei campionamenti, il dataset operativo si compone di **1671** cases. Si tenga presente infine che le analisi descritte in seguito *non* tengono conto della componente spaziale del fenomeno, anche se ogni valore di concentrazione è georeferenziato; ho comunque deciso di includere nel dataset anche le coordinate che accompagnano ogni misura, con l'intento di poter comunque controllare la distribuzione spaziale dei vari fattori considerati e condurre eventualmente analisi di tipo geostatistico [cfr. ad esempio il tentativo descritto nel paragrafo §8.3.3].

'casa tipo'

La regressione dei quantili, o quantomeno la sua implementazione mediante il package *quantreg*, prevede di definire, in fase di preparazione dei dati, le *classi di riferimento* per il modello — a dire, nel caso specifico, le caratteristiche della CASA TIPO rispetto alla quale l'influenza dei vari fattori e delle varie classi sarà successivamente valutata. La casa tipo cui si farà in seguito riferimento ha le caratteristiche riportate nella tabella 8.1.

8.3 Costruzione e analisi dei modelli

Sono stati costruiti quattro differenti modelli con lo scopo di valutare il comportamento dell'approccio QR al variare delle caratteristiche proprie dei fattori coinvolti (eventuali correlazioni, numero, ...) e anche per confrontare questi risultati con quelli ottenuti in altri contesti (come nel caso del modello 02).

Di seguito, verranno presentati i vari modelli, descrivendone le caratteristiche e le idee di base che sottendono la loro costruzione. I relativi risultati saranno discussi principalmente in relazione all'output grafico ottenuto in fase di elaborazione; si tratta di grafici che riportano, al variare del fattore considerato, l'impatto di ogni singola classe sul valore di concentrazione, in funzione del quantile τ, tenendo costanti gli altri fattori. Ho deciso di ricorrere a 19 quantili nel range $\tau \in [0.05, 0.95]$. La fascia che accompagna l'andamento dei punti in funzione di τ rappresenta l'intervallo di confidenza al 90%. Assieme a questi grafici, viene riportato anche quello relativo all'intercetta, che va interpretata come la stima della funzione densità di probabilità cumulativa per la 'casa tipo'. Ogni grafico riporta anche la stima LS dell'impatto per la classe in esame

e il relativo intervallo di confidenza al 90% (rispettivamente, linea orizzontale continua e linee orizzontali tratteggiate); va da sé che questo valore risulta costante al variare del quantile τ.

Ricordo infine che nella serie di grafici prodotti, per ogni fattore manca quello relativo alla classe di riferimento riportata nella tabella 8.1: l'influenza e l'impatto delle altre classi va infatti interpretata sempre in relazione alla classe di riferimento!

8.3.1 Modello 01

Questo primo modello è stato costruito considerando *tutti* i fattori disponibili, consapevole del fatto che alcuni sono correlati tra loro[2]. Per questo modello, per ogni fattore ho anche determinato dei box-plot per evidenziare la presenza di differenze statisticamente significative al variare delle classi: gli edifici in contatto hanno una mediana sensibilmente maggiore (188 vs. 93 $\mathrm{Bq{\cdot}m^{-3}}$), e le misure condotte in aula e ufficio sembrerebbero mostrare dei valori di concentrazione mediamente più elevati.

I grafici per questo modello sono riportati in figura 8.2. Si nota che:

- *contatto* = *sì* e *qualità degli infissi* = *scarso* mostrano degli intervalli di confidenza piuttosto stretti e degli andamenti regolari e in linea con quanto ci si può aspettare in base a considerazioni di tipo fisico: risulta infatti ragionevole che la concentrazione di radon sia maggiore per gli edifici a diretto contatto con il suolo (rispetto a quelli che non lo sono), e che sia invece minore per gli edifici che hanno una scarsa qualità degli infissi (la classe di riferimento è in questo caso 'buona'), poiché spesso è sufficiente una discreta aerazione dei locali per ridurre la concentrazione di radon negli ambienti chiusi; interessante infine come l'effetto di queste classi (considerando anche il relativo intervallo di confidenza) si stacchi dallo zero fin dai primi quantili;

- in tutti gli altri casi, invece, l'effetto delle classi inizia a manifestarsi per $\tau \in [0.4, 0.6]$, mentre per valori inferiori i coefficienti oscillano attorno allo zero; bisogna però tenere anche conto del fatto che aumentano contemporaneamente gli intervalli di confidenza, riducendo di conseguenza l'affidabilità statistica delle conclusioni che si possono trarre[3];

- per questa fase dell'analisi, l'influenza delle classi è in linea con la stima LS per la media, anche se il relativo intervallo di confidenza è spesso elevato e rende la stima compatibile con un effetto nullo;

- gli andamenti risultano in linea con le analisi più convenzionali che in precedenza sono state condotte ricorrendo a dei semplici box-plot.

In generale, mi sembra inoltre di poter affermare che la coda bassa della distribuzione per i valori di concentrazione di radon indoor *non* sembra essere influenzata sensibilmente dalle classi per le variabili secondarie di tipo antropogenico — eccettuati forse contatto e qualità degli infissi. Sembra quindi, da questa prima analisi, che

> *le caratteristiche proprie dell'edificio sede della misura — quella che è stata identificata come* parte antropogenica *del fenomeno — manifestino la loro influenza in quelle situazioni caratterizzate da* elevati *valori di concentrazione.*

[2] Ad esempio, come ampiamente discusso nel capitolo 7, è noto che *tipo di costruzione* risulta correlato con *classe data di costruzione*; in realtà, per gli altri fattori le correlazioni non risultano molto influenti.

[3] In molti casi i grafici sono piuttosto confusi negli andamenti; un'ipotesi che possa spiegare questo fenomeno è che la correlazione tra alcune classi e/o fattori possa in qualche modo 'mescolare' gli effetti o 'confondere' il modello. Indicazioni positive a sostegno di quest'ipotesi vengono dall'analisi dei modelli 03 e 04, riportate rispettivamente nei paragrafi §8.3.3 e §8.3.4.

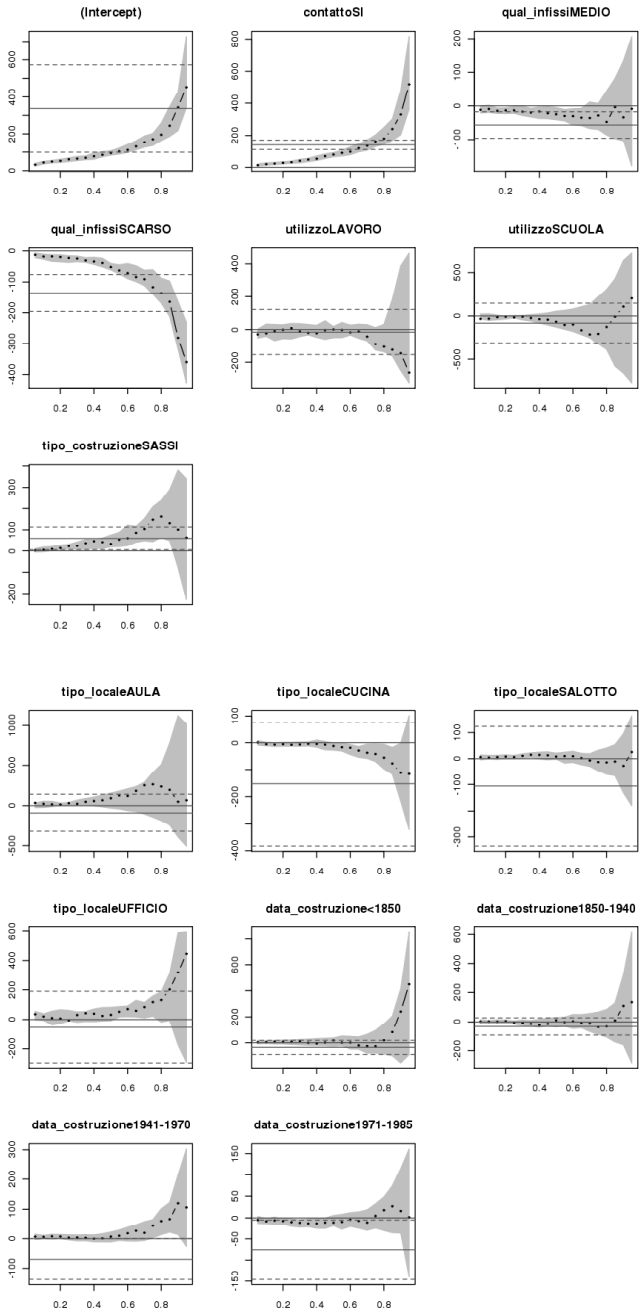


Figura 8.2: Risultati dell'analisi QR per il modello 01; per una corretta interpretazione dei grafici, si faccia riferimento a quanto discusso all'inizio del paragrafo §8.3.

8.3.2 Modello 02

Questo modello è stato costruito considerando solo i fattori che sono emersi come *significativi* dall'analisi condotta e descritta nel capitolo 9, basata sull'impiego di algoritmi di Features Selection. Le variabili secondarie considerate sono pertanto:

- tipo di costruzione;
- contatto;
- esposizione.

Questo si traduce in una 'casa tipo' di riferimento costruita con *mattoni*, *senza* contatto con il terreno e con una esposizione a *sud*.

Mi sembra interessante far notare come, in base ad analisi condotte in precedenza e in altri contesti [cfr. capp. 7 e 9], sia emerso che:

- *qualità degli infissi* sia fortemente correlato con *tipo di costruzione* e *contatto*;
- *contatto* non sia correlato con *esposizione*;
- *classe data di costruzione* e *tipo di costruzione* siano fortemente correlate;

questi risultati vanno inoltre a sostegno della scelta dei fattori operata dall'algoritmo di Features Selection — il modello è stato 'ripulito' dall'informazione ridondante portata da variabili correlate tra loro; mi aspetto quindi che i risultati che si otterranno per questo modello risultino di più facile e solida interpretazione rispetto a quanto ottenuto per il modello 01.

I grafici relativi al modello 02 sono riportati in figura 8.3, dai quali si evince che:

- *rispetto al modello 01, gli andamenti risultano globalmente più regolari e con intervalli di confidenza più stretti; probabilmente, questo si può spiegare con un minor numero di fattori coinvolti che risultano inoltre scorrelati tra loro;*
- *il fattore* esposizione *non sembra avere una influenza significativa sul valore di concentrazione, se non eventualmente per le classi W e S-E (se si tiene conto del relativo intervallo di confidenza) e solo per quantili piuttosto elevati, tipicamente* $\tau \geq 0.7$*;*
- contatto = sì e tipo di costruzione = sassi *mostrano andamenti in linea con una previsione teorica e con quanto ottenuto nel caso del modello 01, a parziale sostegno della validità del risultato ottenuto per questi fattori.*

8.3.3 Modello 03

Ho pensato di costruire questo ulteriore modello riducendo il numero di fattori, e in particolare eliminando quelli che, in base al modello 01, hanno mostrato una scarsa influenza sul valore di concentrazione — quantomeno in relazione alla coda inferiore della p.d.f. Le variabili secondarie considerate si sono ridotte a:

- contatto;
- qualità degli infissi;
- tipo di costruzione;

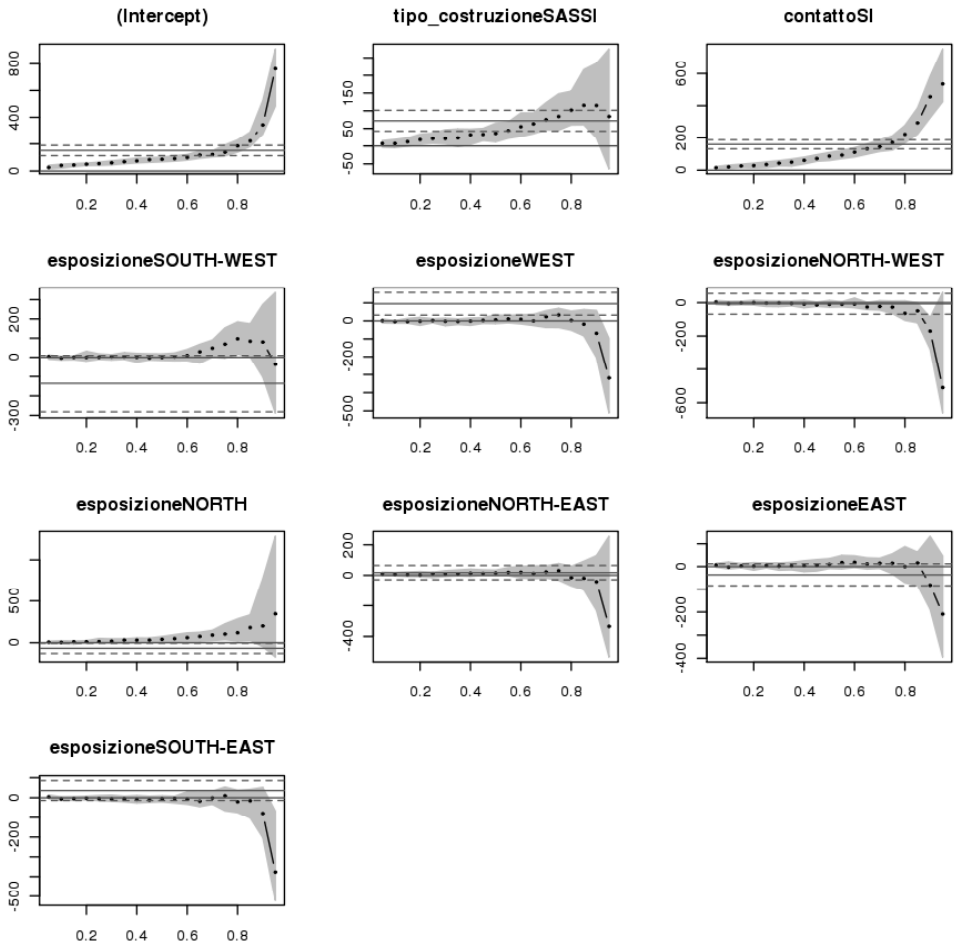


Figura 8.3: Risultati dell'analisi QR per il modello 02; per una corretta interpretazione dei grafici, si faccia riferimento a quanto discusso all'inizio del paragrafo §8.3.

- utilizzo.

Questo si traduce in una 'casa tipo' di riferimento *priva* di contatto con il terreno circostante, costruita con *mattoni*, con una *buona* qualità degli infissi; il tipo di edificio è *abitazione*.

La figura 8.4 riporta quanto ottenuto dallo studio condotto sul modello 03; si può notare come:

- la classe *contatto = sì* manifesti lo stesso andamento ottenuto per i modelli 01 e 02, con un intervallo di confidenza stretto e un andamento regolare; sembra proprio che questo risultato, peraltro in linea con le previsioni che si possono fare in base a principi fisici, sia sostenuto da una buona solidità statistica;
- la classe *qualità degli infissi = scarso* sia in linea con quanto ottenuto per il modello 01;
- la classe *tipo di costruzione = sassi* abbia un andamento più regolare rispetto a quanto

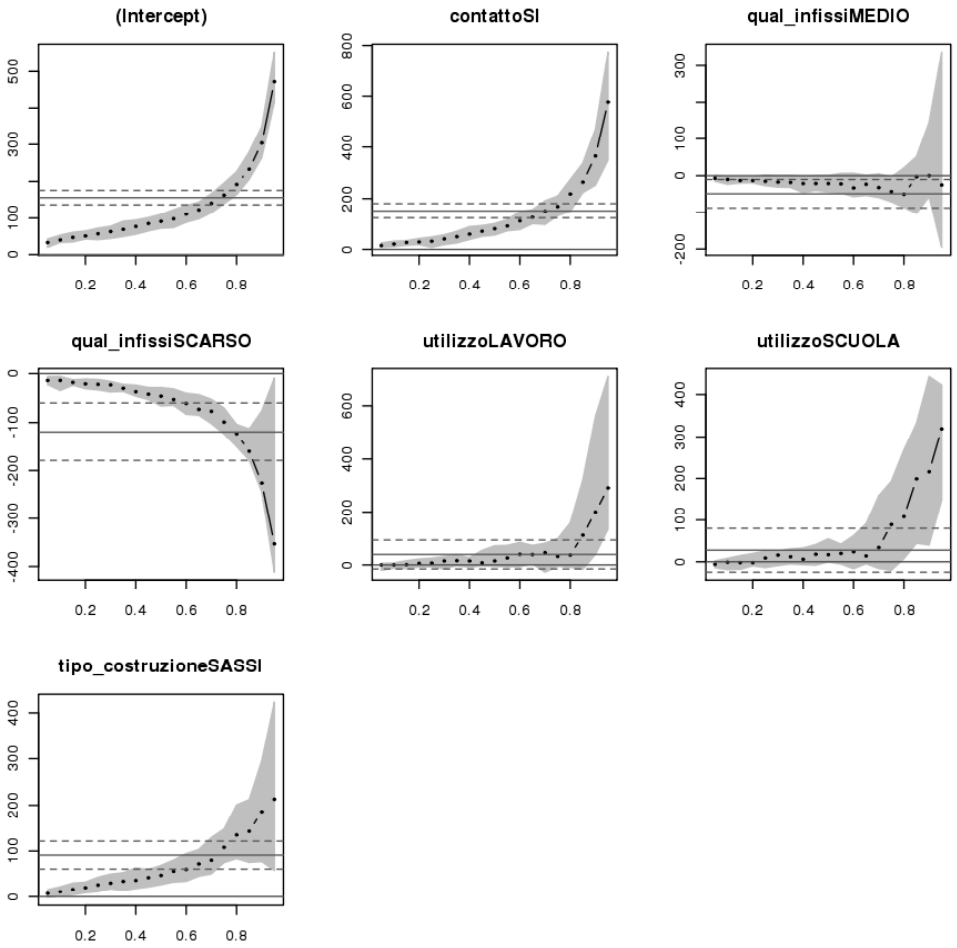


Figura 8.4: Risultati dell'analisi QR per il modello 03; per una corretta interpretazione dei grafici, si faccia riferimento a quanto discusso all'inizio del paragrafo §8.3.

ottenuto per i modelli precedenti; una possibile spiegazione potrebbe essere trovata considerando il modello 01: è nota la correlazione tra classe data di costruzione < 1850 e tipo di costruzione = sassi, in virtù della quale si può interpretare la decrescita della curva per sassi [cfr. fig. 8.2];

- il fattore *utilizzo* mostra degli andamenti molto diversi rispetto a quanto ottenuto per il modello 01 per quantili elevati (e anche il modello LS mostra differenze significative nelle stime della media); questo risultato apparentemente anomalo trova una spiegazione plausibile se si considera il fatto che vi è una forte correlazione[4] tra *utilizzo* = *lavoro* e *tipo locale* = *ufficio* e tra *utilizzo* = *scuola* e *tipo locale* = *aula*: come è chiaro da quanto riportato in figura 8.2, le coppie "giocano" in maniera opposta sul valore di concentrazione

[4]Che certo non stupisce, ma che è ampiamente verificata anche da un punto di vista statistico.

e se si combinano i due effetti, il risultato è in linea con quanto ottenuto per il fattore *utilizzo* riportato dal modello 03 — che risulta peraltro anche più leggibile.

Concludo sottolineando come il modello 03, a mio giudizio, si sia rivelato utile nel porre in luce che:

- *ricorrere a un numero "eccessivo" di fattori può portare a difficoltà di interpretazione dei risultati ottenuti, soprattutto nel caso in cui nel modello ci siano fattori fortemente correlati tra di loro — nel caso specifico esaminato,* utilizzo *e* tipo di locale*;*
- contatto = sì, qualità degli infissi = scarso *e* tipo di costruzione = sassi *manifestano la loro influenza in modo evidente e si configurano quindi come parametri significativi; interessante come queste classi siano quelle che meglio differenziano una casa "nuova" da una "vecchia", quest'ultima verosimilmente più legata alla caratteristiche geologiche del territorio: questi aspetti possono venir rivisti e paragonati con quanto discusso nel paragrafo §8.5 e nel capitolo 5.*

Analisi variografica

Alla luce della buona numerosità che caratterizza la 'casa tipo' per il modello 03 (618 cases) e dell'altrettanto buona distribuzione spaziale dei relativi campionamenti, ho provato a condurre delle analisi variografiche esplorative al fine di controllare se questo subset non desse dei buoni risultati anche dal punto di vista geostatistico. Ricordo che quello che si va ricercando è la correlazione spaziale per una *abitazione* "nuova", ovvero *priva* di contatto con il terreno circostante, costruita con *mattoni* e con una *buona* qualità degli infissi.

Ho calcolato variogrammi tradizionali omnidirezionali al variare della massima distanza tra le coppie e della risoluzione spaziale (lag-step), senza purtroppo ottenere alcun risultato soddisfacente: tutti i variogrammi sono affetti da un elevato effetto nugget e mostrano una struttura estremamente rumorosa.

Questo "parziale fallimento", da un certo punto di vista, è però in linea con quanto ottenuto dalle analisi descritte nel capitolo 5, nel quale si sostiene come si ottengano variogrammi con buone strutture e basso effetto nugget nel caso in cui si prenda in considerazione un subset che si avvicina a quello costituito da case di tipo "vecchio", teoricamente più legate agli aspetti geologici del fenomeno che, in linea di principio, dovrebbero manifestare correlazioni spaziali più forti — o, detto in altri termini, meno disturbate dai fattori antropogenici relativi alle case "nuove".

8.3.4 Modello 04

Quest'ultimo modello è stato costruito con lo stesso intento che sottende il modello 03, sostituendo il fattore *utilizzo* con *tipo di locale*, al fine di valutare l'influenza delle loro riconosciuta correlazione. In questo caso, la 'casa tipo' è pertanto *senza* contatto con il terreno, costruita con *mattoni*, la qualità degli infissi è *buona* e il dosimetro è stato esposto in una *camera da letto*.

Da quanto riportato in figura 8.5, si può affermare che:

- la classe *contatto = sì* risulta ancora in linea con quanto ottenuto per gli altri modelli studiati, a sostegno della validità e stabilità dei risultati relativi a questo fattore;
- la classe *qualità degli infissi = scarso* è anch'essa in linea con i risultati ottenuti per i modelli precedenti;

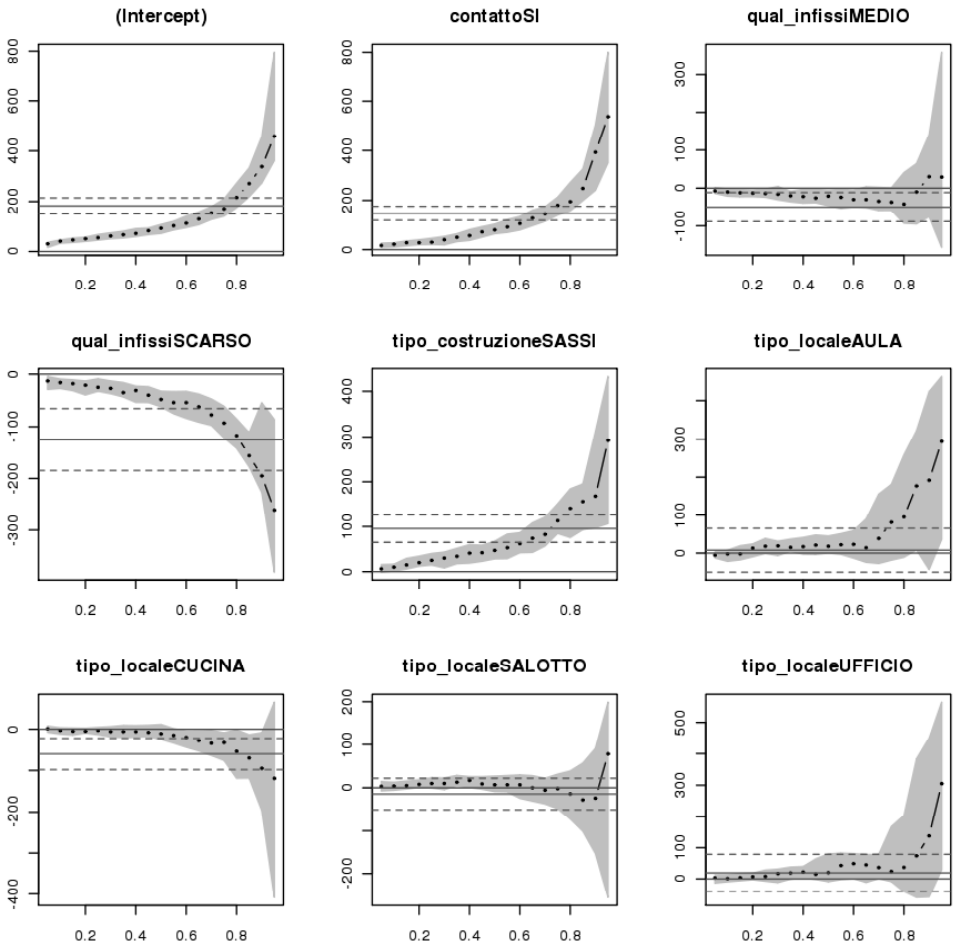


Figura 8.5: Risultati dell'analisi QR per il modello 04; per una corretta interpretazione dei grafici, si faccia riferimento a quanto discusso all'inizio del paragrafo §8.3.

- le classi *aula* e *ufficio* per il fattore *tipo di locale* risultano perfettamente in linea con quanto ottenuto per le classi *scuola* e *lavoro* relative al modello 03.

Concludo affermando che

> *dall'analisi del modello 04 si ottengono sostanzialmente le stesse informazioni che si possono ottenere dal modello 03, in virtù della forte correlazione tra le classi dei fattori* utilizzo e tipo di locale*; questo a sostegno delle ipotesi che sono state fatte in relazione ai risultati ottenuti per il modello 03.*

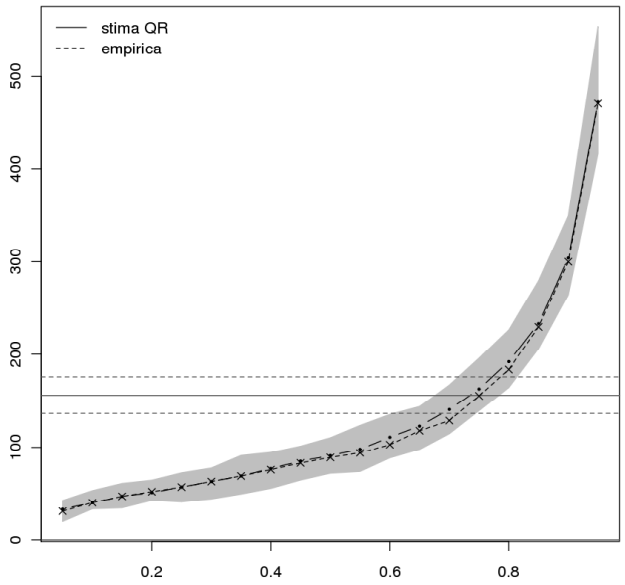


Figura 8.6: Confronto tra il parametro *intercetta* e relativo intervallo di confidenza (fascia) determinato mediante QR e stima empirica della c.c.d.f. per la 'casa tipo' relativa al modello 03; le linee orizzontali rappresentano invece la stima LS e il relativo intervallo di confidenza; per ulteriori dettagli utili alla corretta interpretazione del grafico, si faccia riferimento a quanto discusso nel paragrafo §8.4.

8.4 Verifica del significato dell'intercetta

La teoria della regressione dei quantili prevede che i coefficienti relativi al parametro identificato come *intercetta* [cfr. ad esempio quanto riportato nel primo grafico delle figure 8.2–8.5] siano una stima della funzione densità di probabilità cumulativa condizionata per la situazione di riferimento (c.c.d.f.) — quella cioè costituita dalle classi di riferimento per i vari fattori utilizzati nel modello; nel nostro caso, quella che di volta in volta è stata definita come 'casa tipo'.

Allo scopo di verificare la validità di questa interpretazione, per i quattro modelli considerati in precedenza ho pre-selezionato i cases in base alle classi di riferimento in modo da disporre di quattro subset relativi alle quattro tipologie di 'casa tipo' (quelle descritte appunto per i modelli 01–04). Su questi, ho quindi condotto semplici analisi statistiche e determinato la c.c.d.f. empirica (per gli stessi quantili utilizzati nelle precedenti analisi QR), in modo da poterla confrontare con la stima QR — ovvero, quanto ottenuto per l'intercetta QR. Un esempio visivo di questo confronto, relativo al modello 03, è riportato in figura 8.6. Si nota chiaramente come la stima empirica della c.c.d.f. sia perfettamente compatibile, entro l'intervallo di confidenza, con la stima QR; questi risultati restano validi anche per gli altri modelli esaminati.

Si può quindi concludere che si è verificato nella pratica che

> *il parametro* intercetta *ottenuto con l'approccio QR rappresenta la stima della distribuzione cumulativa densità di probabilità condizionata relativa alla situazione assunta*

> *come riferimento in fase di definizione del modello — a dire, tutti i coefficienti per le alte classi dei vari fattori che costituiscono il modello sono nulli.*

8.5 Casa tipo alternativa: una casa “vecchia”

A conclusione di questa analisi, mi è sembrato interessante provare a costruire un modello basato sugli stessi dati utilizzati nei casi precedenti (misure condotte nel semestre invernale), andando però a cambiare le classi di riferimento che definiscono la cosiddetta ‘casa tipo’; la scelta è stata dettata dai risultati ottenuti nel capitolo 5, e hanno portato alla definizione di una casa che potremmo chiamare “vecchia”, e quindi che meno risente, almeno in linea di principio, dell’influenza della parte antropogenica del fenomeno — verosimilmente, un’abitazione, o meglio, per essere più precisi, un edificio più legato alle caratteristiche geologiche. L’idea è stata quella di valutare se questa scelta portasse a risultati più leggibili o quantomeno di più facile interpretazione rispetto a quanto ottenuto finora.

La ‘casa tipo’ “vecchia” ha le seguenti caratteristiche: è una *abitazione*, *in contatto* con il terreno che la circonda, la qualità degli infissi è *scarsa* ed è costruita con *sassi*, *prima* del 1850, è esposta a *sud* e il dosimetro è stato esposto in una *camera da letto*. I risultati per questo modello — basati sul dataset relativo al modello 01 — sono riportati nei grafici di figura 8.7.

Si nota facilmente, confrontando con quanto riportato per il modello 01 in figura 8.2 a pagina 118, che:

- nei casi in cui le classi di riferimento restano le stesse, gli andamenti non subiscono variazioni significative;
- rispetto a quanto ottenuto per il modello 01, le classi relative a *qualità degli infissi* e *classe data di costruzione* mostrano degli andamenti più leggibili, e in linea con quanto ci si può aspettare da una previsione teorica;
- se si cambia la classe di riferimento, l’andamento che si ottiene risulta sensato e complementare — emblematico, in questo senso, il caso di *contatto*.

Questo tipo di confronto e analisi è stato esteso a tutti e quattro i modelli considerati in precedenza, ottenendo risultati in linea con quelli descritti per la coppia modello 01–modello 01 ‘vecchio’.

Sulla base dei risultati ottenuti, concludo la discussione affermando che

> *in alcuni casi, cambiare la classe di riferimento per un dato fattore, può avere ripercussioni sulla leggibilità dell’*influenza *delle altre classi relative al medesimo fattore sulla variabile di riferimento — in questo caso, la concentrazione di radon indoor.*

8.6 Conclusioni

Da quanto descritto nel paragrafo §1.3 a pagina 7, risulta evidente come ogni misura di concentrazione di attività di radon indoor sia accompagnata da una serie di variabili secondarie che caratterizzano tanto il contesto geologico quanto quello abitativo dell’edificio sede della misura. È sembrato interessante e utile, in questo ambito, indagare se queste variabili secondarie manifestassero la loro influenza sul valore di concentrazione realmente misurato in maniera uniforme o se piuttosto l’entità di tale influenza, se presente, fosse legata in qualche modo al valore di concentrazione stessa — ad esempio, se una variabile si mostrasse significativa solo per valori

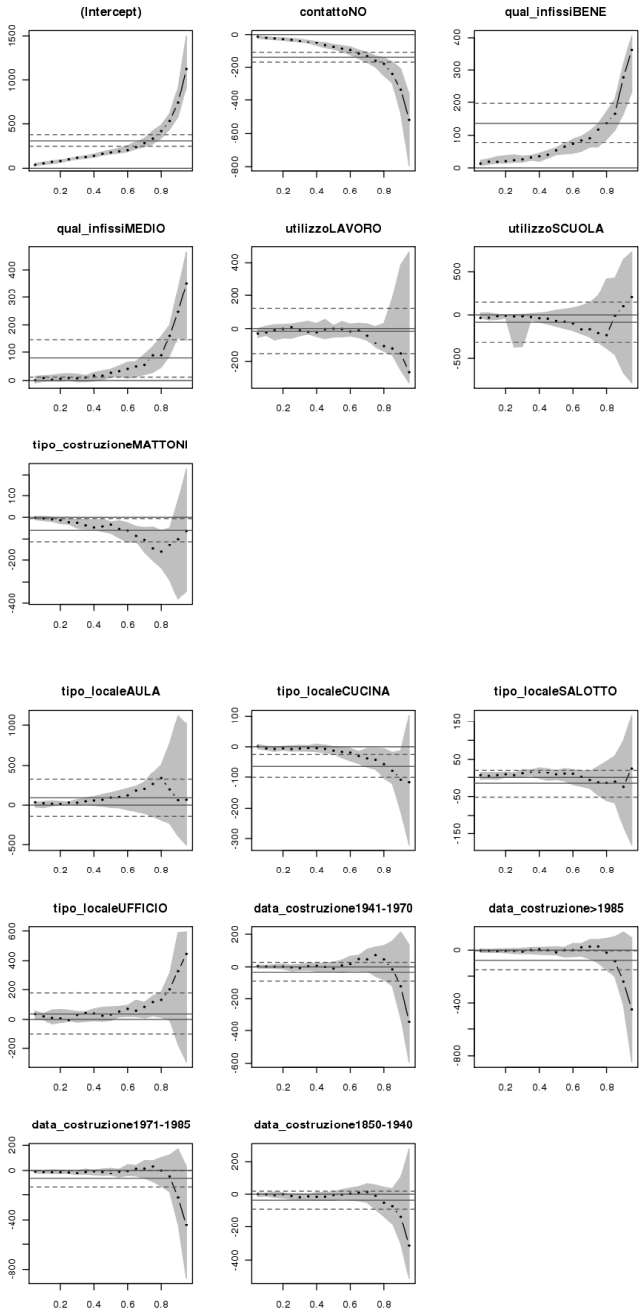


Figura 8.7: Risultati dell'analisi QR per il modello relativo alla casa "vecchia", basato sul dataset relativo al modello 01; per una corretta interpretazione dei grafici, si faccia riferimento a quanto discusso all'inizio del paragrafo §8.3, mentre per quanto riguarda questo ulteriore modello, a quanto discusso in §8.5.

elevati di concentrazione di radon indoor, e non avesse invece influenza nel caso di valori bassi. Ancora, potrebbe risultare utile costruire dei modelli di riferimento pre-selezionando il valore delle classi per una serie di variabili secondarie, in base alle quali valutare l'influenza delle altre classi — ad esempio, costruire un modello per una "casa tipo" che sia costruita con mattoni, e valutare se e in quale misura la situazione cambia se invece quella stessa casa fosse costruita di sassi.

Questi aspetti non possono essere trattati ricorrendo al diffuso metodo dei minimi quadrati, ma la teoria della *Regressione dei Quantili* (QR) fornisce invece gli strumenti teorici e pratici per affrontare entrambi i problemi descritti.

Facendo riferimento al dataset operativo descritto nel capitolo 1.3, si sono costruiti modelli differenti, al fine di valutare i risultati e il comportamento dell'approccio QR al variare di tipo e numero di variabili secondarie coinvolte, nonché per confrontare questi risultati con quelli ottenuti in altri contesti [cfr. cap. 9].

Riassumendo quanto ottenuto per i vari modelli implementati, si può affermare che:

- in generale, le variabili legate alla parte antropogenica del fenomeno radon indoor (legate alle caratteristiche dell'edificio sede della misura) manifestano in maniera evidente la loro influenza in quelle situazioni caratterizzate da *elevati* valori di concentrazione;
- ricorrere a un numero "eccessivo" di variabili secondarie per la costruzione del modello di riferimento può portare a difficoltà di interpretazione dei risultati ottenuti, soprattutto nel caso in cui alcune variabili risultino fortemente correlate tra loro;
- le classi 'contatto = sì e 'tipo di costruzione = sassi' manifestano la loro influenza in modo evidente, configurandosi quindi come parametri significativi in relazione al potere predittivo delle variabili associate; interessante notare come queste classi siano quelle che meglio differenziano una casa "nuova" rispetto a una "vecchia", quest'ultima verosimilmente più legata alla caratteristiche geologiche del territorio; ulteriori spunti di approfondimento possono essere suggeriti rivedendo questi aspetti alla luce di quanto discusso nel paragrafo §8.5 (costruzione di un modello di riferimento relativo a una "casa vecchia") e paragonandoli con le analisi condotte nel capitolo 5 (analisi variografiche relative a specifiche classi);
- è da tenere presente che cambiare la classe di riferimento che caratterizza il modello costruito può, in alcuni casi, avere ripercussioni non trascurabili sulla leggibilità dell'influenza delle altre classi in relazione alla variabile di riferimento.

Parte III

Machine Learning

CAPITOLO 9

Ricerca delle variabili predittive: Feature Selection

L'idea che ha ispirato il lavoro descritto in questo capitolo è stata quella di applicare delle tecniche di data mining (nello specifico, un algoritmo di feature selection*) con l'intento di capire se, tra le numerose variabili di tipo* antropogenico *che accompagnano il valore di concentrazione di radon indoor, risulta possibile identificare quelle che hanno una maggior influenza sul valore di concentrazione stessa — a dire, un maggior* peso predittivo.

I dati hanno subito una fase di pre-processing condotta in ambiente `R`{44}, sfruttando il quale è anche stato costruito il dataset "dummy"; la parte computazionale è stata svolta invece interamente con `Weka`{54}, ricorrendo all'algoritmo identificato come *CfsSubsetEval*.

9.1 Brevi richiami teorici

Nell'ambito delle tecniche di Machine Learning (ML), spesso risulta importante disporre di un dataset che sia il più *pulito* possibile in relazione alle variabili predittrici che lo costituiscono. Per questo, si sono sviluppati degli algoritmi con lo scopo di *identificare* ed eventualmente *rimuovere* tutta l'informazione ridondante contenuta in variabili secondarie non rilevanti; questo si traduce in un dataset operativo con un numero ridotto di FEATURES[1]. *ML*

Questo nuovo dataset verrà quindi utilizzato per le fasi successive di analisi, con il duplice vantaggio — almeno teorico! — di migliorare la qualità e la stabilità dei risultati ottenuti. Gli approcci standard al problema sono essenzialmente due:

wrapper : si opera un resample statistico del subset scelto ricorrendo all'algoritmo di ML che verrà successivamente impiegato (simile all'approccio di una cross-validation);

[1] Per consuetudine, è il termine con cui, nel gergo specifico del data mining, si fa riferimento alle variabili secondarie che costituiscono il dataset e che verranno utilizzate per predire la variabile di interesse, indicata generalmente con CLASS.

filter : le features non volute sono tolte in fase di pre-processing dei dati; tipicamente, risultano più veloci e per questo adatti a grandi dataset — altro vantaggio, è che *non* dipendono da tipo di approccio (algoritmo) di ML scelto.

L'algoritmo di feature selection dovrebbe quindi essere in grado di rispondere alla domanda: *quali sono le features iniziali da includere nel subset finale e quali quelle invece da ignorare?*

Se il dataset di partenza è costituito da n features, tutti i subset possibili saranno 2^n; per questo, si pone anche il problema di avere a disposizione una qualche *strategia di ricerca* al fine di selezionare, in base a un criterio quantitativo, il subset migliore.

Tra i più diffusi algoritmi di ricerca, si ricorda quella noto come *Best First*, cui ho fatto riferimento per le analisi che verranno discusse in seguito; essenzialmente, questo algoritmo prevede i seguenti passaggi[2]:

1. partenza da un subset vuoto;
2. generare tutte le possibili espansioni aggiungendo una singola feature per volta;
3. scegliere il migliore tra quelli costruiti al punto 2 e quindi procedere alla sua espansione;
4. se non si ottiene alcun miglioramento (sulla base di un qualche parametro quantitativo di riferimento), scegliere il "secondo migliore" e ripetere il punto 3[3].

CFS L'algoritmo impiegato nelle successive fasi di analisi è del tipo CFS, ovvero *Correlation-based Feature Selection*, e si compone sia dell'algoritmo di ricerca che della funzione atta alla valutazione della bontà di quanto ottenuto sul subset sotto esame.

L'ipotesi di base su cui si fonda è la seguente:

> *il subset di features ottimale è costituito da features fortemente correlate (predittive) con la variabile di interesse, ma allo stesso tempo scorrelate (non predittive) le une con le altre.*

L'indice su cui la scelta si basa è dato dalla seguente espressione:

$$G_s = \frac{k\overline{r_{vf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \simeq \frac{\text{predizione}}{\text{ridondanza}} \tag{9.1}$$

dove k rappresenta il numero di features del subset, $\overline{r_{vf}}$ la correlazione media delle features dello specifico subset con la variabile di interesse, $\overline{r_{ff}}$ la correlazione media tra le features. In altre parole, il numeratore dell'equazione 9.1 porta informazioni su quanto un determinato gruppo di features risulta predittivo per la variabile di interesse, mentre il denominatore porta informazioni su quanta ridondanza è presente in quello stesso subset.

[2] Questa descrizione prevede una direzione *forward*, ma l'algoritmo può essere applicato sia con direzione backward (partenza dal dataset completo e successive riduzioni del numero di features) sia con modalità bi-directional (è possibile cioè impostare quale o quali siano le feature che costituiscono il subset di partenza e quindi lasciare che l'algoritmo 'esplori' altri subset aggiungendo o togliendo features).

[3] É chiaro che dando tempo al sistema, questo potrà esplorare tutti i 2^n possibili subset: è però possibile limitare la ricerca a un determinato numero di subset che non portano a ulteriori miglioramenti nel caso in cui i tempi di calcolo risultassero eccessivi.

9.1.1 Correlazione tra le features

Al fine di disporre di un conteso comune per il calcolo delle correlazioni che compaiono nell'equazione 9.1, le features continue devono essere convertite in features di tipo categorico nella fase di pre-processing; quindi, si ricorre a una misura della correlazione tra features categoriche che si basa sulla *teoria dell'informazione*[4].

Siano X e Y due variabili casuali discrete; allora le equazioni

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y) \tag{9.2}$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log_2 p(y|x) \tag{9.3}$$

danno rispettivamente l'entropia di Y prima e dopo l'osservazione di X. La quantità per la quale l'entropia di Y diminuisce riflette l'informazione aggiuntiva su Y che viene fornita dall'osservazione di X, ed è nota come *information gain*; tale guadagno G di informazione è dato da

$$\begin{aligned} G =& H(Y) - H(Y|X) \\ =& H(X) - H(X|Y) \\ =& H(Y) + H(X) - H(X,Y) \end{aligned} \tag{9.4}$$

ed è una misura simmetrica — a dire che il guadagno di informazione ottenuto su Y dopo l'osservazione di X è uguale a quello guadagnato su X dopo l'osservazione di Y. Sfortunatamente, G risulta polarizzato verso quelle features che hanno più valori, ovvero attributi che manifestano un maggior numero di valori appariranno come portatori di maggior informazione rispetto ad attributi che ne hanno meno, anche se non sono realmente più informativi. Inoltre, le correlazioni che compaiono nell'eq. 9.1 andrebbero normalizzate per assicurare che siano tra di loro confrontabili e possano di conseguenza avere gli stessi effetti.

Per ovviare a questo tipo di problemi, limitando anche il valore di G all'intervallo $[0, 1]$, si introduce quello che è noto come *symmetrical uncertainty*:

$$S_u = 2 \times \left[\frac{G}{H(Y) + H(X)} \right] \tag{9.5}$$

9.2 Descrizione del dataset utilizzato

I valori di concentrazione di radon indoor che andranno a costituire i tre differenti dataset descritti in seguito sono stati estratti dal dataset di riferimento [cfr. §1.3] operando in fase di pre-processing le seguenti scelte:

- eliminazione di tutti i missing-values — ogni valore di concentrazione è accompagnato da *tutti* i valori corrispondenti alle variabili categoriche prese in esame;
- pre-selezione delle sole misure condotte a *piano zero*[5]

[4] Si faccia ad esempio riferimento al lavoro originale di Shannon {48}

[5] Questa variabile, anche se in parte legata alle caratteristiche dell'edificio, non farà quindi parte della successiva fase di analisi; la scelta è stata dettata da i) riconosciuta e dimostrata influenza sul valore di concentrazione e ii) numerosità esigua delle classi relative ai piani alti.

In analogia all'analisi descritta nel capitolo 7, ho deciso di focalizzare l'attenzione solo sulle variabili di tipo *antropogenico*, ossia quelle legate alle caratteristiche dell'edificio sede della misura — tutte le informazioni di tipo geologico sono in questa fase completamente ignorate, come del resto quelle relative alla parte spaziale (georeferenziazione). Le variabili prese in considerazione per questo tipo di analisi sono quindi le seguenti[6]:

- utilizzo (2);
- tipo locale (3);
- classe data di costruzione (4);

id features

- tipo di costruzione (5);
- qualità degli infissi (6);
- contatto con il terreno (7);
- esposizione (8).

Da notare che il software `Weka`{54} *non* ha modo di riconoscere un eventuale ordinamento delle classi (come ad esempio nel caso di qualità degli infissi, per la quale alto > medio > basso); del resto, non è chiaro se questo abbia un qualche effetto sul risultato finale, né se sia possibile "istruire" il software con questa informazione aggiuntiva.

Inoltre, al fine di poter valutare il comportamento e la solidità dell'algoritmo di feature selection, ho deciso di aggiungere alle variabili reali anche due variabili (fittizie) casuali — prive quindi di correlazioni con il valore di concentrazione — e identificate in questo modo:

- *rnd_ n*: numeri casuali da una p.d.f. uniforme con range di valori in $[0, 1]$;
- *rnd_ c*: serie di label casuali (A,B,C).

Infine, ho deciso di predisporre tre differenti dataset, in relazione al valore di concentrazione cui fanno riferimento:

- concentrazione semestrale (nessuna correzione al valore):
  - *inverni*: **1824** cases
  - *estati*: **89** cases
- concentrazione convertita a media annuale mediante opportuni fattori di correzione [cfr. {53}]: **1830** cases.

9.3 Test su dataset "dummy"

Prima di intraprendere l'analisi sui valori reali di concentrazione, ho ritenuto utile e opportuno costruire un dataset *ad hoc* per valutare sia le caratteristiche e il comportamento propri dell'algoritmo di feature selection, sia l'eventuale influenza della *disomogeneità* nella numerosità delle varie classi all'interno di una stessa feature — situazione che è emersa dall'analisi specifica su questo tipo di variabili, come ampiamente discusso nel paragrafo §7.3, a pagina 104.

[6] Per una descrizione dettagliata delle singole variabili e del loro significato, si rimanda a quanto discusso nel paragrafo §1.3, a pagina 7; inoltre, in parentesi, riporto anche l'identificativo usato nelle successive analisi dei risultati

id variabile	*labels*
C1	A(300), B(300), C(300)
C2	SÌ(450), NO(450)
C3	ALTO(225), MEDIO(225), BASS0(225), NULLO(225)
C1a	A(700), B(150), C(50)
C2a	SÌ(750), NO(150)
C3a	ALTO(300), MEDIO(20), BASS0(100), NULLO(480)
r	R1, R2, R3

Tabella 9.1: Descrizione delle variabili che compongono il dataset "dummy", costituito da 900 cases; in parentesi, accanto a ogni label, è riportato il numero di cases che caratterizza quella particolare classe; per quanto riguarda le label della variabile r, queste sono state ottenute a partire da numeri casuali estratti con una p.d.f. uniforme.

Il tutto è stato realizzato in ambiente `R`{44}, ricorrendo alle variabili descritte nella tabella 9.1 per la costruzione dei seguenti modelli numerici:

$$\begin{aligned} \text{M1} &= c1n + c2n - c3n \\ \text{M1a} &= c1na + c2na - c3na \\ \text{M2} &= \text{M1} + \eta_{10\%} \\ \text{M2a} &= \text{M1a} + \eta_{10\%} \\ \text{M3} &= \text{M1} + \eta_{50\%} \\ \text{M3a} &= \text{M1a} + \eta_{50\%} \end{aligned} \tag{9.6}$$

dove le variabili $c(\cdot)n(\cdot)$ sono di tipo numerico, ottenute con le seguenti sostituzioni per le variabili categoriche riportate nella tabella 9.1:

- $c1n \rightarrow$ A=1, B=0, C=-1
- $c2n \rightarrow$ SÌ=1.5, NO=-1.5
- $c3n \rightarrow$ ALTO=0.9, MEDIO=0.3, BASSO=-0.3, NULLO=-0.9

e dove $\eta_{(\cdot)\%}$ indica l'aggiunta di un rumore bianco la cui ampiezza risulta pari al valore di percentuale riportato, rispetto al range di valori del modello di riferimento M1.

9.3.1 Descrizione dei risultati ottenuti

In tutti i casi, le indagini sono state condotte ricorrendo all'algoritmo di feature selection *CfsSubsetEval* e all'algoritmo di ricerca *Best First*; per quest'ultimo, sono state prese in esame tutte le possibili strategie di ricerca, ovvero forward, backward e bi-directional.

Per ognuno dei modelli riportati nel gruppo di equazioni (9.6), i test sono stati svolti anche aggiungendo la variabile r [cfr. tab. 9.1] e successivamente, anche le altre variabili non introdotte nel modello — ad esempio, nel modello M1 introducendo come possibili predittori anche le variabili C1a, C2a, C3a.

Dalla serie di test condotti, è emerso che:

modelli senza rumore — per i modelli M1 e M1a, controllando anche l'influenza del parametro *-L* per l'algoritmo di feature selection[7]:

- la differenza di distribuzione in frequenza inter-classe *non* sembra avere un'influenza evidente;
- l'impiego dell'opzione *-L* si traduce in una stima più efficace nei casi in cui sono presenti delle variabili "spurie" (ossia, a priori non correlate con la variabile da predire).

modelli con rumore — per i modelli M2, M2a, M3, M3a, controllando anche l'influenza del parametro *-L* per l'algoritmo di feature selection:

- l'introduzione dell'opzione *-L* risulta più efficace nell'eliminare le eventuali variabili "spurie";
- al crescere della componente di rumore, sembra che l'algoritmo dia risultati migliori se le distribuzioni inter-classe sono disomogenee[8];
- in alcuni casi, può risultare utile ricorrere a una cross-validation per valutare la stabilità della scelta operata dall'algoritmo (rispetto a un generico "full training set") — opzioni offerte da `Weka`{54} in fase di calcolo;
- le distribuzioni non omogenee in frequenza inter-classe *non* sembrano, anche in questo caso, manifestare un'influenza significativa.

Concludendo, operativamente mi sembra di poter affermare, in base ai risultati dell'analisi condotta sul dataset dummy appena descritta, che:

- cautelativamente, *è meglio usare l'opzione* -L*: l'algoritmo tende a essere più selettivo, ma probabilmente è al contempo anche più sicuro;*
- non *è stata messa in luce una sensibilità evidente a classi sbilanciate (in numerosità relativa) all'interno di una stessa feature.*

9.4 Analisi dei dataset reali

Le analisi condotte sui tre differenti dataset composti da reali misurazioni di concentrazione descritti in §9.2 sono state eseguite con le stesse modalità (tipo di algoritmo e strategie di ricerca) descritte e applicate per il dataset "dummy". Inoltre, l'analisi è stata condotta sia sui valori *continui* di concentrazione, sia su opportune *classi* di concentrazione. Si ricorda inoltre che nei dataset sono presenti anche due variabili fittizie che non hanno alcuna correlazione con la variabile di riferimento.

9.4.1 Analisi sul valore continuo di concentrazione

In questa prima fase, ho deciso di prendere in considerazione tutte le classi per ciascuna feature, anche se in alcuni casi risultano molto poco popolate (come *prefabbricato* per la feature tipo

[7] Citando dall'help di Weka: *"OPTIONS: locallyPredictive – Identify locally predictive attributes. Iteratively adds attributes with the highest correlation with the class as long as there is not already an attribute in the subset that has a higher correlation with the attribute in question."*

[8] In particolare dal modello m3a: se si aggiungono variabili che non fanno parte del modello, queste vengono sempre escluse dall'algoritmo CFS; per M2a, invece, ci sono casi in cui vengono riconosciute come significative anche variabili che *non* fanno parte della definizione del modello. Benché sia stata eseguita qualche ulteriore analisi specifica, questo non ha portato a risultati concreti e convincenti. Questo aspetto meriterebbe probabilmente un approfondimento specifico.

dataset	*merit of best subset found*	*features*	*note*
inverni	.254	5–7–8	
	.254	5–7	-L
	.220	5–6–7–8	-P6 -D1
	.239	2–5–7–8	-P2
	.239	2–5–7	-P2 -L
	.254	5–7–8	-P6 -D2
estati	.383	2–3–5–7–9	
	.383	2–3–5–7–9	-L
annuali	.250	5–7–8–10	
	.250	5–7	-L
	.219	5–6–7–8	-P6 -D1
	.219	5–6–7	-P6 -D1 -L
	.250	5–7–8–10	-P6 -D2
	.250	5–7	-P6 -D2 -L

Tabella 9.2: Risultati delle analisi condotte sui dataset reali per il valore continuo di concentrazione; il numero riportato nella colonna *features* identifica le variabili selezionate dall'algoritmo in base quanto riportato in §9.2; si tenga presente che per il dataset "dummy" [cfr. §9.3], il *merit of best subset found* ha valori tipici dell'ordine di .84.

costruzione, che ha una sola misurazione); alla luce di quanto analizzato e discusso in §9.4.3, questo scelta non sembra essere determinante.

Per completezza e chiarezza di lettura dei risultati riportati nella tabella 9.2, mi sembra opportuno descrivere brevemente alcuni dettagli relativi ai parametri impiegati in fase di calcolo:

- per quanto concerne l'algoritmo di ricerca, ho utilizzato *Best First* con l'opzione *-N10*, che nella pratica consente una maggior "libertà" di ricerca sulle features da includere;
- in tutti i casi ho variato la direzione di ricerca — identificata secondo il seguente schema: D0=backward, D1=forward, D2=bi-directional —, ma senza ottenere differenze significative nei risultati;
- l'opzione *-Px* impone all'algoritmo di usare come subset di partenza quello composto dalla/e feature/s 'x'; nel caso in cui sia accoppiata all'opzione *-D1*, questo equivale a imporre che la feature 'x' sia mantenuta nel subset, anche se non dovesse risultare significativa; accoppiata con l'opzione *-D2*, non è invece garantito che la feature 'x' venga mantenuta.

Da quanto riportato nella tabella 9.2 e da ulteriori analisi condotte su questi dataset, si può concludere che:

- la direzione di ricerca *non* sembra avere alcuna influenza sul risultato finale, nemmeno imponendo la feature di partenza;
- imponendo, in base a logica e a quanto emerso da altre analisi, che la qualità degli infissi possa essere una feature rilevante, il coefficiente di merito cala — ovvero, si ottiene un subset peggiore[9];

[9] Si ricorda che il comportamento di questa variabile in relazione alla sua influenza sul valore di concentrazione è per certi aspetti dubbio; per questo, si è cercato di sfruttare le occasioni utili per mettere sotto test specifici questa variabile.

dataset	*merit of best subset found*	*features*	*note*
inverni	.254	5–7–8	
	.254	5–7	-L
	.220	5–6–7–8	-P6 -D1
estati	.362	2–3–5–7	
	.383	2–3–5–7–9	-L
	.331	2–3–5–6–7	-P6 -D1
annuali	.250	5–7–8	
	.250	5–7	-L
	.219	5–6–7–8	-P6 -D1

Tabella 9.3: Risultati delle analisi condotte sul dataset reale per il valore continuo di concentrazione dal quale sono state rimosse le features di tipo random (9–10); il numero riportato nella colonna *features* identifica le variabili selezionate dall'algoritmo in base quanto riportato in §9.2; si tenga presenta che per il dataset "dummy" [cfr. §9.3], il *merit of best subset found* ha valori tipici dell'ordine di .84.

- *le features che l'algoritmo seleziona come più predittive per il valore di concentrazione, intersecando le informazioni ricavate da tutti e tre i dataset, sono:*
  - *tipo di materiale da costruzione (5)*
  - *contatto (7)*
  - *esposizione (8)*

Potrebbe a questo punto risultare interessante un confronto tra inverni ed estati, ma la statistica su quest'ultimo dataset credo sia troppa esigua per ricavarne conclusioni affidabili (ad esempio, viene riconosciuta significativa la feature 'utilizzo', ma le abitazioni sono 82 su 89 edifici che compongono il dataset).

Infine, dalla tabella 9.2 si nota come l'algoritmo riconosca anche una feature di tipo random (nello specifico, la 9) per il dataset estati, anche ricorrendo all'opzione -L: un'indagine visiva non evidenzia correlazioni tra il valore di concentrazione e questa feature, ma il coefficiente di correlazione è pari a $\rho = -.19$, a fronte di $\rho = .027$ per il dataset inverni e $\rho = .0038$ per il dataset annuali. Probabilmente, quindi, tale risultato è frutto di un qualche artefatto di natura squisitamente computazionale.

Analisi senza variabili random

Ho inoltre condotto le stesse analisi e con gli stessi parametri operativi descritti in precedenza eliminando dal dataset operativo le variabili random (9–10); l'idea è stata quella di valutare se queste ultime fossero una ulteriore fonte di rumore che potesse compromettere l'affidabilità dell'algoritmo di selezione. I risultati ottenuti sono riportati nella tabella 9.3.

Da questo tipo di analisi, si può concludere che

> *l'algoritmo risulta essere sufficientemente robusto rispetto all'introduzione di features fittizie, ovvero* non *correlate con la variabile da predire.*

Analisi dell'influenza di eventuali correlazioni tra features

Successivamente, con l'idea di valutare se l'algoritmo non riconosca alcune features in quanto magari fortemente correlate tra di loro, ho condotto delle ulteriori analisi specifiche avendo cura

dataset	*merit of best subset found*	*features*	*note*
inverni	.145	2–5–8	
	.131	2–4–5–6–8	-P6 -D1
	.145	2–5–8	-P6 -D2
estati	.315	2–5	
	.272	2–3–5–6	-P6 -D1
annuali	.146	2–5–8	
	.133	2–4–5–6–8	-P6 -D1
	.146	5	-P6 -D2

Tabella 9.4: Risultati delle analisi condotte sul dataset reale per il valore continuo di concentrazione dal quale sono state rimosse le features di tipo random (9–10) e la feature 'contatto' (7); il numero riportato nella colonna *features* identifica le variabili selezionate dall'algoritmo in base quanto riportato in §9.2; ricorrere all'opzione -L non porta in nessun caso a risultati differenti.

di controllare preventivamente le correlazioni tra le varie features; questo per cercare una possibile spiegazione del fatto che non venga mai considerata *qualità degli infissi* come possibile variabile predittiva, poiché da un lato la logica sarebbe a favore di una sua influenza sul valore di concentrazione, dall'altro analisi differenti [cfr. capitolo 7 e capitolo 8] mostrano che l'informazione legata a questa variabile non viene significativamente riconosciuta, o quantomeno non risulta determinante sul valore di concentrazione.

Le correlazioni (condotte con gli strumenti descritti in §7.5, pag. 108) più evidenti risultano essere quelle tra:

- qualità degli infissi e tipo di costruzione (5/6): scarso-sassi, bene-mattoni;
- qualità degli infissi e contatto (7/6): medio-sì, bene-no;
- tipo di materiale da costruzione e data di costruzione (5/4);
- contatto in correlazione con tutte le altre features, tranne esposizione (7/*);

La *qualità degli infissi* risulta pertanto essere in correlazione significativa sia con *tipo di costruzione* che con *contatto*, e l'algoritmo ne dovrebbe, in linea teorica, tenere conto; inoltre, *contatto* risulta essere in correlazione con tutte le altre features: in base a queste evidenze, ho deciso di eliminare dal dataset operativo la feature *contatto* con lo scopo di valutare se l'algoritmo si trovi in qualche modo "costretto" a introdurre come significative altre features prima eliminate — l'idea è quella che contatto possa "mascherare" l'influenza di altre features con le quali è fortemente correlato.

I risultati di questa analisi sono riportati nella tabella 9.4. Come si può notare, la rimozione di *contatto* porta l'algoritmo a introdurre *utilizzo*, che però risulta avere un forte bias sulla classe abitazione; inoltre, imponendo come significativa la feature *qualità degli infissi* non si ottengono miglioramenti sul coefficiente di merito del subset. Ancora, il comportamento relativo dei tre subset è in linea con le analisi precedenti, e il possibile ruolo di portatrice di 'informazione utile' della features *qualità degli infissi* risulta ancora dubbio e non chiaramente interpretabile.

Concludendo, mi sembra di poter affermare che:

- contatto (7) *sembra essere una feature piuttosto importante, configurandosi come rappresentativa anche di altre features antropogeniche con le quali risulta essere correlata*[10]*;*
- tipo di materiale da costruzione (5)[11] *è sempre presente, a indicare che potrebbe configurarsi come un'altra feature con un buon potere predittivo.*

9.4.2 Analisi per classi di concentrazione

Come ulteriore fase dell'analisi, ho deciso di rendere categorica anche la variabile di riferimento introducendo delle *classi di concentrazione* del tipo 'sopra-sotto soglia', con l'idea di valutare se la selezione delle featrures significative risulti in qualche modo legata al valore di concentrazione stessa. Questo approccio dovrebbe, almeno nelle intenzioni, avere il duplice vantaggio di:

i) consentire un confronto sulle correlazioni con i risultati che si possono ottenere con l'analisi di variabili categoriche ricorrendo agli strumenti che si sono utilizzati nelle analisi descritte nel capitolo 7, a pagina 101;

ii) valutare se le features riconosciute come significative lo siano su tutto il range di valori di concentrazione o manifestino la loro influenza sono per determinate classi — ad esempio, una determinata features potrebbe mostrare la sua eventuale influenza solo per valori elevati di concentrazione.

> ***Problema:*** *non è stato possibile spezzare la distribuzione in più classi contigue, definendo più valori di soglia, ma solo trasformare il valore di concentrazione in una variabile di tipo indicatore (il tutto legato a delle limitazioni del software impiegato); va da sé che quanto ci si era proposti in fase di pianificazione dell'analisi avrebbe senza dubbio trovato giovamento dalla possibilità di implementare la prima idea descritta; tuttavia, questo tipo di analisi è stato successivamente sviluppato ricorrendo a un strumento statistico specifico, la* regressione dei quantili, *come ampiamente descritto nel capitolo 8 a pagina 113.*

La tabella seguente riporta i valori di soglia (espressi in $Bq \cdot m^{-3}$) impiegati nell'analisi e il corrispondente numero di cases per i vari subset che si sono venuti a costituire:

dataset	*200*		*400*		*800*	
inverni	1250	574	1562	262	1722	102
estati	75	14	84	5	89	0
annuali	1383	447	1639	191	1771	59

Dai risultati riportati in tabella 9.5, si evince che:

- *anche per questa serie di analisi,* contatto *è una features sempre presente: si ottiene quindi un ulteriore supporto al fatto che questa feature abbia un elevato peso predittivo;*

[10] Se questa ipotesi è vera, si è trovata una feature molto semplice e al contempo ricca di informazioni; la scelta di questa rispetto alle altre con le quali è correlata è inoltre in linea con la proprietà caratteristiche dell'algoritmo CfsSubsetEval di privilegiare le features con poche classi.

[11] La classe *sassi* per questa variabile ha manifestato una caratteristica interessante in relazione alle proprietà variografiche del relativo subset, come ampiamente descritto nel capitolo 5.

dataset	*merit of best subset found*	*features*	*note*
inverni (200)	.050	7–8	
	.050	7	-L
inverni (400)	.039	2–7	
	.039	2–7	-L
inverni (800)	.033	7	
	.033	7	-L
estati (200)	.136	2–7	
	.136	7	-L
estati (400)	.046	2–3–4–8	
	.046	2–3–4–8	-L
annuali (200)	.046	7–8	
	.046	7	-L
annuali (400)	.037	5–7	
	.037	7	-L
annuali (800)	.024	7–8	
	.024	7–8	-L

Tabella 9.5: Risultati delle analisi condotte sul dataset reale in funzione della classe del valore di concentrazione (il valore di soglia che identifica le due classi è riportato di volta in volta accanto al nome del dataset); il numero riportato nella colonna *features* identifica le variabili selezionate dall'algoritmo in base quanto riportato in §9.2.

- *la scelta delle features* non *sembra essere legata alla classe di concentrazione;*
- *inverni e medie annuali manifestano sempre comportamenti simili tra loro;*
- *i coefficienti di merito del subset identificato come migliore sono sensibilmente inferiori rispetto alle analisi precedenti*[12] [cfr. tab. 9.2, 9.3 e 9.4];
- *il* tipo di costruzione *risulta anche in quest'analisi significativa.*

9.4.3 Eliminazione delle classi "poco popolate"

In quest'ultima fase dello studio ho deciso di eliminare dalle varie features quelle classi per le quali la numerosità risulta troppo esigua rispetto alle quella delle altre classi della stessa feature; cosa nella pratica significhi "troppo esigua" è un criterio soggettivo, valutato caso per caso analizzando le caratteristiche delle classi della feature in esame.

Di seguito, la descrizione del tipo di interventi condotti (validi, dove non diversamente specificato, per tutti i dataset):

- *utilizzo*: ho deciso di conservare solo la classe abitazione[13];
- *tipo di locale*: tolgo le classi altro, sala, stanza, negozio, cantina, corridoio (e aula e ufficio per il dataset *estati*); restano quindi le classi camera da letto, cucina, salotto;
- *tipo di costruzione*: tolgo cemento, legno, prefabbricato; restano sassi e mattoni;

[12] Non è ancora del tutto chiaro se questo si possa legare alla bontà globale del risultato; i valori ottenuti per il dataset dummy (merit of best subset found $\simeq$ .084) sembrerebbero tuttavia supportare quest'ipotesi.

[13] Come nel caso di *qualità degli infissi*, ho preferito non eliminare dal dataset questa feature, ma conservarla anche se limitata a una sola classe: i due approcci non mi sembrano equivalenti dal punto di vista computazionale — la feature potrebbe comunque risultare significativa.

- *esposizione*: tolgo flat e no data (e nord, nord-ovest e nord-est per il dataset *estati*);
- *qualità degli infissi*: tengo solo la classe bene.

Anche in questa fase, le indagini sono state condotte lavorando anche sulle classi del valore di concentrazione (oltre al valore continuo), come introdotte e descritte in §9.4.2; la tabella seguente riporta i valori di soglia (espressi in $Bq \cdot m^{-3}$) impiegati nell'analisi e il corrispondente numero di cases per i vari subset:

dataset	*200*		*400*		*800*		*totali*
inverni	748	353	981	156	1079	58	1137
estati	48	5	51	2	53	0	53
annuali	867	270	1025	112	1102	35	1137

Il dataset *estati* non è stato preso in considerazione in questo tipo di analisi in relazione all'evidente sbilanciamento delle classi e all'esigua numerosità che lo caratterizza.

Tutte le analisi sono state eseguite in maniera analoga a quanto descritto in §9.4. Nella tabella 9.6 sono riportati i risultati ottenuti lavorando sul valore continuo di concentrazione, nella tabella 9.7 quelli ottenuti operando invece sulle classi di concentrazione. Si può concludere che:

- *come per l'analisi descritta in §9.4.2, i coefficienti di merito calano all'aumentare del valore di soglia (da tener presente che aumenta contemporaneamente lo sbilanciamento delle classi);*
- contatto *rimane comunque la feature più significativa;*
- non *si notano differenze significative legate al differente valore di soglia*

9.5 Conclusioni

Come descritto nel dettaglio nel paragrafo §1.3 a pagina 7, ogni singolo valore di concentrazione di attività di radon indoor è accompagnato da una numerosa serie di variabili secondarie; ricorrendo a un algoritmo di features selection solitamente applicato nel contesto delle analisi di data mining, si è pensato di sfruttare questo approccio per identificare in maniera per quanto possibile oggettiva le variabili secondarie di tipo antropogenico che manifestano una influenza significativa sul valore di concentrazione di radon indoor; in altri termini, quelle che mostrano avere un maggior peso predittivo.

Le analisi sono state condotte sia su dataset costituiti da misure reali, distinguendo tra misure condotte nel semestre invernale ed estivo, e tra misure convertite a medie annuali, sia su dataset costruiti ad hoc al fine di valutare l'influenza di specifici parametri (quali ad esempio la presenza di classi sbilanciate in numerosità) in maniera controllata. Ancora, l'approccio è stato applicato tanto ai valori continui di concentrazione, quanto a classi binarie di concentrazione, con lo scopo di valutare se la selezione delle variabili significative fosse in qualche modo legata al valore di concentrazione.

Riassumendo quanto emerso da questa serie di analisi, intersecando i risultati ottenuti sia per i vari dataset reali sia per quelli "artificiali", si può affermare che:

dataset	*merit of best subset found*	*features*	*note*
inverni	.284	3–5–7–8	
	.284	5–7	-L
estati	.348	3–5–7–9	
	.248	3–5–7–9	-L
annuali	.278	3–5–7–8	
	.278	5–7	-L

Tabella 9.6: Risultati delle analisi condotte sul dataset reale per il valore continuo di concentrazione dal quale sono state rimosse le classi con numerosità esigua; il numero riportato nella colonna *features* identifica le variabili selezionate dall'algoritmo in base quanto riportato in §9.2.

dataset	*merit of best subset found*	*features*	*note*
inverni (200)	.076	3–5–7	
	.076	5–7	-L
inverni (400)	.064	7–8	
	.064	7	-L
inverni (800)	.045	7–8	
	.045	7	-L
annuali (200)	.067	3–5–7–8	
	.067	5–7	-L
annuali (400)	.052	7	
	.052	7	-L
annuali (800)	.029	3–7–8	
	.029	7	-L

Tabella 9.7: Risultati delle analisi condotte sul dataset reale in funzione della classe del valore di concentrazione (il valore di soglia che identifica le due classi è riportato di volta in volta accanto al nome del dataset); dal dataset operativo sono state rimosse le classi con numerosità esigua; il numero riportato nella colonna *features* identifica le variabili selezionate dall'algoritmo in base quanto riportato in §9.2.

- le variabili secondarie che manifestano il maggior peso predittivo sul valore di concentrazione di attività di radon indoor misurata risultano essere *contatto* e *tipo di materiale da costruzione*, informazioni che risultano tra l'altro di facile reperimento in fase di installazione del dosimetro;
- la variabile *contatto*, di tipo binario, risulta in forte correlazione con altre variabili di tipo antropogenico che caratterizzano l'edificio sede della misura: questo fatto è positivo, in quanto questa "semplice" variabile sembra essere in grado di portare con sé informazioni utili contenute in altre variabili più "complesse";
- l'algoritmo preso in esame è risultato sufficientemente robusto sia rispetto alla presenza di variabili spurie, ovvero non correlate con quella di riferimento (radon indoor), sia rispetto alla presenza di variabili secondarie caratterizzate da classi sbilanciate in numerosità;
- per quanto concerne l'analisi su classi binarie di concentrazione, i risultati sono in linea con quelli ottenuti per il valore continuo di concentrazione; in particolare, la variabile con il maggior peso predittivo risulta essere anche in questo caso *contatto*; inoltre, non si notano differenze significative al variare del valore di soglia che determina le due classi di concentrazione.

CAPITOLO 10

L'approccio Weighted *k*-Nearest Neighbor

Dalle numerose analisi condotte sui dati di radon indoor, sono emersi in maniera piuttosto evidente alcuni limiti che caratterizzano gli approcci di tipo geostatistico*: sembra quindi evidente che considerare la* sola *parte spaziale del fenomeno non sia sufficiente per caratterizzarlo al meglio, o da un punto di vista diverso, non sia sufficiente per estrarre tutta l'informazione necessaria per ottenere, in una successiva fase di modellizzazione, delle stime affidabili.*
L'idea che ha guidato questo primo tentativo di applicare delle tecniche 'alternative' al fenomeno radon indoor è stata quindi quella di ricorrere all'approccio del Weighted k-Nearest Neighbor*, che pur nella sua semplicità concettuale e di implementazione, spesso si è rivelato un strumento valido ed è stato applicato in ambiti anche molto diversi tra loro.*
Inoltre, questo approccio potrebbe anche consentire un'analisi di tipo esplorativo *volta all'individuazione delle covariate che hanno maggior influenza sul valore di concentrazione di radon indoor (idea simile a quella indagata nel capitolo 9).*

La parte computazionale è stata svolta interamente in ambiente R{44} ricorrendo ai packages *kknn*{47}, *lattice*{45}, *geoR*{30} e *gstat*{41}, nonché a routine create ad hoc.

10.1 Brevi richiami teorici

Il package *kknn* per l'ambiente statistico R{44} risulta particolarmente interessante in quanto consente di implementare la tecnica che si andrà brevemente a descrivere aggiungendo delle potenzialità che sono sembrate utili in questo contesto, e in particolare i) la possibilità di utilizzare l'approccio anche per la *regressione* (e non solo per la classificazione) e ii) l'introduzione di uno *schema di pesatura* basato su delle funzioni a kernel. In questo paragrafo, ho ritenuto utile richiamare brevemente alcuni concetti di base che sottendono la tecnica utilizzata; si farà riferimento

al caso della classificazione per ragioni "storiche" e di semplicità, ma come ricordato la tecnica può essere estesa anche al caso della regressione.

10.1.1 La tecnica *k*-Nearest Neighbor (*k*-NN)

Nell'ambito della discriminazione statistica, il metodo del "vicino più prossimo" (nearest neighbor, appunto) rappresenta una delle tecniche più intuitive e semplici. È un metodo non-parametrico mediante il quale una nuova osservazione viene classificata ricorrendo alle osservazioni che costituiscono il dataset di riferimento, e in particolare a quella che le è più *vicina*: in base alle features (covariate) disponibili, alla nuova osservazione verrà quindi assegnata la classe che compete all'osservazione del dataset di riferimento che più le "assomiglia" — la determinazione del grado di somiglianza è basata sulla misura di un qualche tipo di distanza in uno spazio multidimensionale.

Formalmente, questa idea può essere descritta nel modo seguente: sia

$$L = \{(y_i, x_i), \quad i = 1, \ldots, n_L\}$$

il dataset di riferimento (noto generalmente come *training* o *learnign* set), dove $y_i \in \{1, \ldots, c\}$ rappresenta i membri di una data classe e il vettore $x_i = (x_{i1}, \ldots, x_{ip})$ rappresenta i valori relativi alle features utilizzate per la previsione. La determinazione del primo vicino sarà quindi basata su una funzione arbitraria che rappresenti una *distanza*, denotata genericamente con $d(\cdot, \cdot)$. A questo punto, per una nuova osservazione (y, x) il primo vicino $(y_{(1)}, x_{(1)})$ che appartiene al dataset di riferimento sarà determinato in base all'equazione

$$d(x, x_{(1)}) = \min_i \left(d(x, x_i)\right)$$

e la classe di appartenenza del primo vicino, indicata come $\widehat{y} = y_{(1)}$, sarà scelta come previsione per y. Si tenga presente che in questo contesto la notazione $x_{(j)}$ e $y_{(j)}$ indica rispettivamente il j-esimo primo vicino di x e la sua classe di appartenenza.

In generale, per la misura della distanza si fa ricorso alla cosiddetta distanza di Minkowski, definita come:

$$d(x_i, x_j) = \left(\sum_{s=1}^{p} |x_{is} - x_{js}|^q \right)^{\frac{1}{q}} \tag{10.1}$$

Ad esempio, se nella 10.1 si seleziona $q = 2$, quella che si ottiene è la ben nota distanza Euclidea.

Una prima estensione dell'idea appena descritta, ormai ampiamente usata nella pratica, è quella che prende il nome di k-nearest neighbor; in questo caso, non viene considerato solo il *primo* più vicino, ma i primi k vicini (osservazioni simili a quella nuova). Di conseguenza, con un sistema 'a votazione' verrà scelta la classe di appartenenza della nuova osservazione. Il parametro k deve essere definito dall'utente.

Sia k_r il numero di osservazioni appartenenti al gruppo dei primi vicini selezionati che appartengono alla classe r, in modo che:

$$\sum_{r=1}^{c} k_r = k$$

allora una nuova osservazione apparterrà alla classe l con:

$$k_l = \max_r (k_r)$$

In questo modo, si evita che sia una singola osservazione del dataset di riferimento a decidere la classe della nuova osservazione. Il grado di *località* di questa tecnica è chiaramente determinata dal parametro k: se $k = 1$ si ricade nel metodo classico descritto in precedenza, mentre per $k \to n_L$ si ottiene un voto di maggioranza sull'intero training set. Va da sé che questo porta a una previsione costante indipendentemente dai valori che caratterizzano la nuova osservazione: a questa verrà associata sempre la classe più frequente che compare nel dataset di riferimento.

10.1.2 La tecnica Weighted *k*-Nearest Neighbor (w*k*-NN)

Questa ulteriore estensione della tecnica si fonda sull'idea che le osservazioni del dataset di riferimento che sono particolarmente vicine alla nuova osservazione (y, x) dovrebbero ricevere un peso maggiore nella fase di decisione rispetto a quelle che sono più distanti da (y, x). Nel caso di k-NN, questo non accade, in quanto l'influenza dei k primi vicini è esattamente la stessa, anche se il grado di somiglianza con (y, x) potrebbe non essere tale. Per raggiungere l'obiettivo proposto, la distanza sulla quale la ricerca dei vicini si basa nella prima fase, deve essere successivamente trasformata in una misura di *somiglianza* che possa essere utilizzata come *peso* in fase di assegnazione della classe (o stima, nel caso della regressione) per la nuova osservazione.

Come nei casi precedenti, il primo passo consiste nella scelta del parametro k che determina il numero di primi vicini da considerare e che si traduce nel parametro p che compare nell'eq. (10.1). Quindi, il secondo passo consiste nel passare dalla distanza ai pesi, e questo viene svolto ricorrendo ad una arbitraria funzione a kernel. Queste sono funzioni $K(\cdot)$ della distanza d con un massimo per $d = 0$ e valori decrescenti con l'aumentare del valore della distanza d stessa. A queste funzioni sono richieste le seguenti proprietà:

- $K(d) \geq 0 \quad \forall d \in \mathbb{R}$
- $K(d)$ ammette massimo per $d = 0$
- $K(d)$ decresce in maniera monotona per $d \to \pm\infty$

Alcune tipiche funzioni a kernel ampiamente utilizzate nella pratica sono riportate in figura 10.1. In questo nuovo approccio, la scelta della $K(\cdot)$ diviene un terzo parametro che l'utente deve selezionare, ma se si eccettua il caso di un kernel rettangolare (che assegna quindi ugual peso a tutti i primi vicini, scelta non molto "furba" in questo contesto), si sottolinea come questa terza fase del processo di costruzione del modello non risulti nella pratica cruciale.

Qualsiasi funzione a kernel necessita, nella sua definizione, di un parametro che ne determini l'estensione prima che il valore della funzione raggiunga lo zero. Nell'approccio wk-NN quest'operazione viene svolta in maniera automatica sulla base del $(k+1)$-esimo primo vicino x_{k+1}, che non verrà pertanto preso in considerazione per la previsione. Una implicita standardizzazione di tutte le distanze su quella di tale elemento porta al risultato richiesto:

$$D(x, x_{(i)}) = \frac{d(x, x_{(i)})}{d(x, x_{k+1})} \qquad \text{per} \quad i = 1, \ldots, k \tag{10.2}$$

Questa distanza normalizzata D ammette sempre e solo valori nell'intervallo $[0, 1]$.

Una volta scelta la misura di somiglianza per le osservazioni che compongono il dataset di riferimento, ogni nuova osservazione (y, x) apparterrà alla classe con il maggior peso, ovvero:

$$\max_r \left(\sum_{i=1}^{k} K\left[D(x, x_{(i)})\right] I(y_{(i)} = r) \right) \tag{10.3}$$

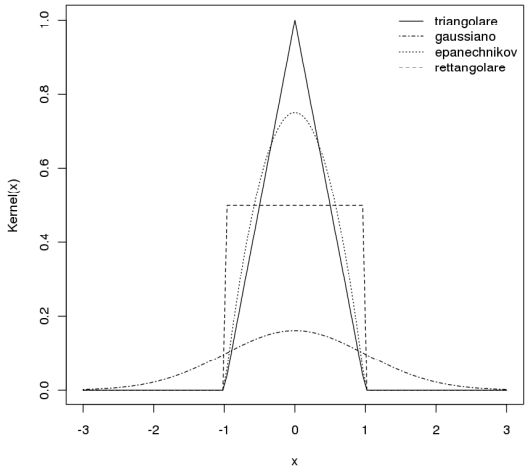


Figura 10.1: Esempi di andamento per alcune tipiche funzioni a kernel $K(x)$.

dove $I(\cdot)$ rappresenta la funzione indicatore. È facile rendersi conto che sia k-NN che NN non sono altro, a questo punto, di casi particolari di wk-NN: k-NN si ottiene imponendo un kernel di tipo rettangolare, mentre NN si ottiene per $k = 1$, indipendentemente dalla scelta di $K(\cdot)$.

Uno degli obiettivi di questo nuovo approccio è anche quello di poter disporre di un metodo che risulta, in un certo senso, indipendente da una scelta infelice del parametro k, che può portare a grandi errori di miscalssificazione; infatti, il numero di primi vicini viene implicitamente mascherato nei pesi: se k è troppo grande, viene corretto a un valore più basso in maniera automatica — se questo fosse il caso, un piccolo numero di vicini cui viene associato un peso elevato domina i rimanenti, che non avranno di conseguenza una forte influenza sulla stima in quanto verrà loro assegnato un peso inferiore.

Concludo riassumendo la struttura logico/temporale dell'algoritmo del wk-NN:

1. sia $L = \{(y_i, x_i), \quad i = 1, \ldots, n_L\}$ il dataset di riferimento costituito dalle osservazioni x_i cui è assegnata la classe y_i e sia x una nuova osservazione la cui classe di appartenenza y deve essere determinata[1];

2. si determina il $(k+1)$-esimo primo vicino a x in base a una *distanza* $d(y, x)$;

3. tale $(k+1)$-esimo primo vicino viene utilizzato per determinare la distanza normalizzata $D_{(i)}$ in base all'equazione 10.2;

4. si trasforma la distanza normalizzata $D_{(i)}$ in un peso $w_{(i)} = K(D_{(i)})$ mediante una arbitraria funzione a kernel $K(\cdot)$;

5. come previsione della classe di appartenenza y per la nuova osservazione x, si sceglie la

[1] Si ricorda che questo approccio può essere esteso sia a classi di tipo ordinale sia a variabili di tipo continuo — dal problema della classificazione a quello della regressione.

classe che mostra una maggioranza pesata sui k primi vicini:

$$\widehat{y} = \max_r \left(\sum_{i=1}^{k} w_{(i)} I(y_{(i)} = r) \right) \tag{10.4}$$

10.2 Descrizione del dataset utilizzato

Ho costruito il dataset operativo per questa analisi partendo da quello generale descritto nel paragrafo §1.3 a pagina 7, che fa riferimento a valori semestrali georeferenziati di concentrazione di attività di radon indoor (non è quindi stata apportata alcuna correzione a tali valori). La fase di preparazione dei dati ha previsto:

- la preselezione delle sole misure condotte nel *semestre invernale* e al *piano zero*, al fine di garantire una maggior uniformità dei dati e una certa coerenza con lavori precedenti;
- l'eliminazione del fattore *no data* dalla feature esposizione;
- l'eliminazione dei fattori *prefabbricato*, *cemento*, *legno* dalla feature tipo di costruzione;
- l'eliminazione dei fattori *corridoio*, *altro*, *sala*, *stanza*, *cantina* e *negozio* dalla feature tipo locale;
- l'eliminazione dell fattore *lavoro/scuola* dalla feature utilizzo.

Gran parte delle eliminazioni descritte sono state eseguite in relazione alla scarsa numerosità che caratterizza i fattori citati e anche a quanto descritto nel capitolo 8. Inoltre, ho eliminato anche alcuni punti che risultavano "isolati" da un punto di vista spaziale, andando a costituire dei piccoli cluster che avrebbero potuto creare qualche artefatto in fase di stima.

La fase di pre-processing descritta ha ridotto il dataset generale a **1638** cases.

Ogni valore di concentrazione riporta tutti i valori per tutte le features (covariate) che lo accompagnano, che nello specifico sono[2]:

- coordinate
- anno di esposizione
- utilizzo
- tipo locale
- tipo di costruzione
- classe data di costruzione
- qualità degli infissi
- contatto
- altitudine
- esposizione

[2]Per una descrizione dettagliata delle feartures citate, si faccia riferimento a quanto riportato nel paragrafo §1.3, pag. 7.

- pendenza
- curvatura

Con l'idea di ricorrere allo stesso dataset operativo anche in relazione ad altri ambiti/algoritmi di Machine Learning (ML), ho deciso di suddividerlo in tre differenti subset, come consuetudine negli approcci ML; questa procedura computazionale è stata svolta con `Geostat Office`©, in modo da preservare la distribuzione spaziale propria del dataset operativo. Ho quindi ottenuto i seguenti (sub)dataset operativi:

dataset	*N*	%
training	820	50
testing	409	25
validation	409	25

Tabella 10.1: Descrizione dei dataset operativi utilizzati per le analisi Machine Learning.

10.3 Test su dataset "dummy"

Con l'obiettivo di avere a disposizione un dataset costituito da valori la cui relazione con le varie features fosse completamente noto e sotto controllo, ho deciso di costruire ad hoc un dataset "dummy"; in questo modo, sarà possibile condurre delle analisi specifiche e mirate per valutare il comportamento dell'algoritmo di wk-NN e verificarne la sua reale efficacia prima di applicarlo ai dati reali.

10.3.1 Costruzione e descrizione del dataset

Il dataset si compone essenzialmente di due parti, una relativa alla parte spaziale del fenomeno che si intende simulare e una legata invece alla presenza di fattori (sia categorici che numerici) che contribuiscono in maniera differente nel produrre il valore della variabile di riferimento, identificata con 'val'.

parte spaziale : sono stati generati dati su una griglia regolare (50×50) con passo pari a 1, ricorrendo a una simulazione gaussiana non-condizionata e basata sul seguente modello di variogramma:

$$\gamma(\mathbf{h}) = 2 + 8 \cdot \text{sph}\left(\frac{\mathbf{h}}{20}\right) \tag{10.5}$$

ovvero un modello di tipo sferico con una varianza pari a 8 e un nugget pari a 2; per valutare l'eventuale sensibilità degli algoritmi alla presenza di anisotropie spaziali, ho ritenuto opportuno introdurne una nel modello di variogramma (10.5): il range è pari a 20 lungo la direzione di massima continuità (direzione Est-Ovest) e 12 lungo quella ortogonale (rapporto di anisotropia pari a 0.6). La bontà della simulazione è stata controllata valutando il comportamento del variogramma calcolato sui dati simulati.
A questo tipo di dati, che hanno un valor medio globale impostato in fase di simulazione pari a 11, si farà riferimento con l'etichetta 'sim1'.

'sim 1'

features : il valore della variabile di riferimento è stato costruito ricorrendo anche due features di tipo continuo e due di tipo categorico, e nello specifico:

- 'alt': feature continua con valori casuali (distribuzione uniforme) nel range $[0, 30]$;
- 'θ': feature continua con valori casuali (distribuzione uniforme) nel range $[0, 4\pi]$;
- 'cat2': feature categorica con 2 possibili valori e numerosità delle singole classi bilanciata;
- 'cat4': feature categorica con 4 possibili valori e numerosità delle singole classi sbilanciata (due classi sono numericamente più popolate rispetti alle rimanenti).

Il modello che genera i dati che compongono il dataset dummy è stato costruito in questo modo:

$$\text{val} = \text{sim1} + \text{cat2} - 0.2 \cdot \text{cat4} + 2 \cdot \text{alt} \cdot \text{sim1} + 15\cos(\theta) + \varepsilon(10) \tag{10.6}$$

facendo in modo che le varie componenti abbiano una influenza confrontabile sul valore della variabile di riferimento 'val'. Al modello è anche stato aggiunto del rumore bianco con varianza pari a 10, indicato nel modello (10.6) con $\varepsilon(10)$.

10.3.2 Descrizione dei risultati ottenuti

Inizialmente, ho costruito dei modelli parziali in modo che l'algoritmo di wk-NN non avesse a disposizione *tutta* l'informazione disponibile, ma solo una parte; sono partito fornendo solo la componente spaziale (coordinate (x, y) della griglia), per simulare la situazione che coinvolge l'approccio standard di tipo geostatistico, e aggiungendo via via ulteriori informazioni date dalle altre features che compaiono nella definizione del modello stesso [cfr. eq. (10.6)].

I vari modelli parziali sono stati valutati con una procedura di cross-validation (CV), al variare del numero k di primi vicini e del tipo di kernel impiegato (quattro tipi differenti, tra cui quello rettangolare — ricorrere a questo tipo equivale all'approccio k-NN, come descritto in §10.1.2); i parametri di confronto sono stati:

a) il minimo del valore assoluto dell'errore (min abs error);

b) il minimo dell'errore quadratico medio (min sq error);

c) il numero di primi vicini migliore (best k);

d) il tipo di kernel.

Questi risultati numerici sono riportati nella tabella 10.2.

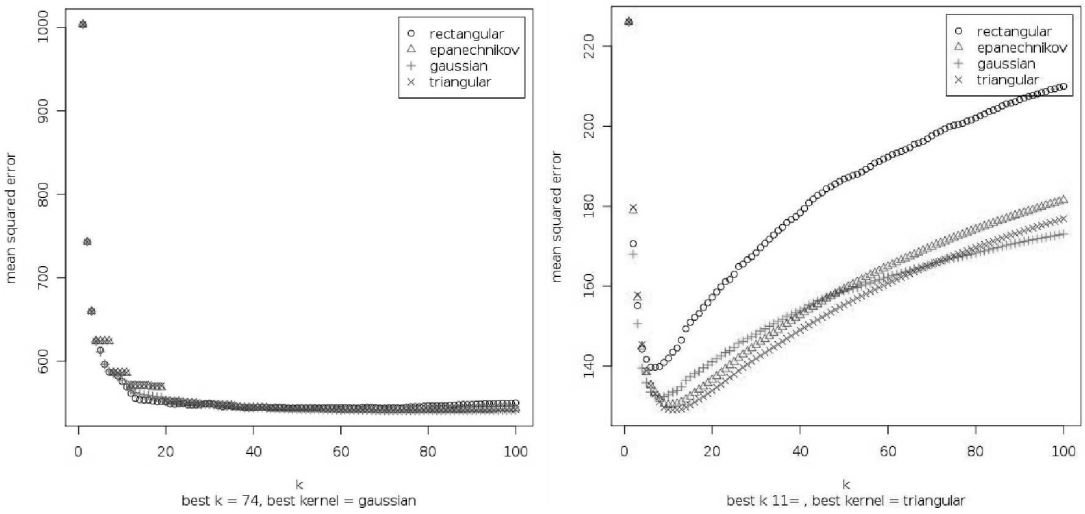
La figura 10.2 riporta invece due esempi di grafici ottenuti sempre in fase di cross-validation, che riportano l'errore quadratico medio in funzione del numero k di primi vicini, al variare del tipo di kernel; i valori numerici riportati nella tabella 10.2 sono stati ricavati da questo tipo di analisi. Tali grafici risultano inoltre utili anche come tool esplorativo, in quanto una scarsa o assente struttura nelle curve riportate è solitamente indice di una altrettanto scarsa o assente struttura (anche spaziale) nei dati stessi — un dataset completamente casuale mostrerà delle curve CV praticamente piatte.

In particolare, ho deciso di riportare i risultati relativi alle due situazioni per così dire estreme:

- la figura 10.2a si riferisce al modello più semplice, in cui compare solo la parte spaziale (in linea con l'informazione che solitamente è nota a un modello di tipo geostatistico convenzionale) [cfr. prima riga della tabella 10.2];
- la figura 10.2b si riferisce al modello completo, ovvero quello cui è stata fornita tutta l'informazione disponibile (situazione che nella pratica è raramente realizzabile) [cfr. ultima riga della tabella 10.2].

val =	*min abs error*	*min sq error*	*best kernel*	*best k*
$x+y$	18.8	540.8	gaussian	74
$+$ *cat2*	19.0	551.5	gaussian	58
$+$ *cat4*	19.6	584.2	gaussian	32
$+$ *alt*	12.9	252.8	triangular	13
$+ \theta$	11.6	220.2	triangular	9
$+$ *sim1*	9.3	129.0	triangular	11

Tabella 10.2: Alcuni parametri statistici ricavati in fase di cross-validation per i modelli descritti nella prima colonna; si noti come all'aumentare dell'informazione fornita al modello, i risultati migliorino (il modello completo è dato dall'eq. (10.6)); si tenga presente che i vari modelli sono riportati in maniera "incrementale", ovvero la nuova feature si va ad aggiungere a quelle precedenti; per ulteriori dettagli, si faccia riferimento a quanto riportato in §10.3.2.



(a) solo parte spaziale (b) modello completo

Figura 10.2: Esempi di curve di cross-validation per due modelli relativi al dataset dummy: in (a) quello relativo alla sola parte spaziale, in (b) quello completo; i relativi parametri statistici sono riportati in tab. 10.2; si noti come all'aumentare della quantità di informazione fornita al modello, le curve relative a differenti kernel mostrino differenze apprezzabili.

Nell'analizzare i risultati, si tenga comunque presente che il modello completo di riferimento (10.6) è caratterizzato anche da una parte di rumore bianco ε.

Successivamente, ho costruito ulteriori modelli variando di volta in volta il numero e il tipo di features introdotte, eliminando in alcuni casi la parte spaziale[3], focalizzando in altri l'attenzione sulla reale capacità dell'algoritmo (o meglio, della sua implementazione) di operare uno scaling opportuno dei dati — in particolare, per quanto riguarda le features di tipo continuo: non sono infatti rare le situazioni per le quali una variabile numerica caratterizzata da valori molto elevati, o range di variabilità molto esteso, possa mascherare anche completamente l'influenza di altre variabili (sia categoriche che numeriche) qualora queste assumano valori molto più piccoli.
A seguito di questa serie di test, mi sembra di poter concludere che:

- *all'aumentare dell'informazione fornita al modello (numero di features), le curve di cross-validation (come quelle riportate in fig. 10.2) mostrano strutture via via più evidenti e differenze più marcate tra i tipi di kernel indagati;*
- *il numero k di primi vicini necessari per una stima corretta risulta inversamente proporzionale alla quantità di informazione cui il modello ha accesso, a dire che sono sufficienti meno vicini se questi risultano ben caratterizzati;*
- *le variabili numeriche sembrano essere più efficaci nel predire il valore della variabile di riferimento (anch'essa numerica) rispetto a quelle categoriche; inoltre, risulta anche più evidente, nei grafici CV, la differenza tra kernel rettangolare (nessun peso) e gli altri tipi di kernel;*
- *lo scaling delle variabili è eseguito in automatico e nella maniera opportuna dall'implementazione dell'algoritmo di wk-NN fornito dal package* kknn.

Come nota finale, operativamente i) una scarsa struttura nelle curve di CV, ii) un alto valore del parametro k e iii) una non evidente differenza sul tipo di kernel sembrano essere segnali di una *mancanza di informazione* per il modello del fenomeno sotto studio.

10.4 Analisi dei dati reali

Il dataset impiegato per lo studio dei dati reali è quello cui si è fatto riferimento con l'etichetta 'training', composto di 820 valori di concentrazione di radon indoor. Per quanto riguarda questa specifica analisi, si sarebbe potuta aumentare la numerosità del dataset ricorrendo anche a quello di 'testing', non richiesto dall'algoritmo di wk-NN; tuttavia, per ragioni di uniformità e facilità di confronto con quanto sviluppato nei capitoli 11 e 13, ho preferito assumere come dataset operativo il solo dataset 'training'.

Inizialmente, ho costruito un modello del fenomeno reale estremamente semplice, basato sulla sola componente spaziale — a dire, solo sull'informazione contenuta nella localizzazione (coordinate (x, y)) del fenomeno; a questo modello si farà riferimento con l'etichetta 'sp'. Nella curva di CV, si nota una debole struttura, con un best $k = 57$; questo modello, oltre a confermare in parte che la sola componente spaziale *non* sembra essere in grado di spiegare la complessità del fenomeno radon indoor, sarà anche successivamente utilizzato come pietra di paragone rispetto a modelli più complessi, ossia con un maggior numero di informazioni disponibili.

modello 'sp'

[3] Per lo meno, quella per così dire esplicita, in quanto parte di questa informazione è comunque contenuta nella feature 'sim1'.

id modello	*features coinvolte*
01	x + y + contatto + tipo costruzione
02	x + y + contatto + tipo costruzione + utilizzo
03	x + y + altitudine + pendenza + curvatura
04	x + y + tipo costruzione + contatto + esposizione
05	tutte
06	x + y + tipo costruzione + tipo locale + classe data costruzione
07	x + y + qualità infissi
08	x + y + contatto + tipo costruzione + utilizzo + pendenza + curvatura

Tabella 10.3: Descrizione dei modelli utilizzati per lo studio condotto sui dati reali.

10.4.1 Costruzione e descrizione dei modelli

Con lo scopo di valutare tanto l'influenza del tipo di informazione fornita al modello quanto i risultati ottenuti da analisi differenti [cfr. capitoli 5 e 9], ho costruito otto differenti modelli, le cui caratteristiche di base sono riassunte nella tabella 10.3.

Uno degli aspetti che ha guidato la scelta delle features da inserire nei vari modelli è stato quello, a mio avviso non trascurabile, di tenere presente che in fase di stima, i valori di tali features devono essere accessibili; inoltre, questo avrà anche ripercussioni sul *tipo* di mappe che eventualmente si andranno a realizzare.

> ***Nota 1:*** *Se lo scopo dell'analisi è quello di produrre una mappa* generale *che fornisca le caratteristiche globali del fenomeno, nel caso più "semplice" della geostatistica la mappa prodotta sarà funzione unicamente delle coordinate del punto di stima (una volta fissati il modello di variogramma e i parametri propri dell'algoritmo di stima, tipicamente un kriging), mentre in questo caso, sarà verosimilmente necessario produrre più mappe in funzione non solo delle coordinate, ma anche dei valori propri delle features coinvolte — a dire, considerando ad esempio la feature 'contatto', si avrà una mappa per* contatto = sì *e una per* contatto = no. *Questo problema non si pone, invece, se l'obiettivo finale sarà quello di una stima* puntuale*, ovvero stimare il valore di concentrazione per un nuovo (o perché no anche esistente) edificio di cui siano noti i parametri coinvolti nella definizione del modello stesso.*

Il modello 03 è stato costruito considerando parametri caratteristici del terreno che possono facilmente essere determinati mediante strumenti GIS, una volta in possesso di un modello digitale del terreno: in questo caso, la mappa prodotta sarebbe "unica". Il modello 04 è invece stato costruito considerando le features che sono emerse come significative in base all'analisi basata su Feature Selection, descritta nel capitolo 9. Il modello 05, considerando tutta l'informazione disponibile, con lo scopo di valutare se questo porti a qualche vantaggio o aggiunga piuttosto "rumore"[4]. Il modello 07, infine, per valutare da un ulteriore punto di vista l'influenza di *qualità degli infissi*, che in linea di principio dovrebbe avere una influenza evidente e prevedibile sul valore di concentrazione, ma che, in differenti contesti di analisi, si è rivelata essere di dubbia interpretazione.

[4] In opposizione a questo modello, che potrebbe essere affetto da una *ridondanza* di informazione, ho deciso di costruirne uno che fosse il più 'economico' possibile dal punto di vista delle features coinvolte. In virtù dei risultati ottenuti nel paragrafo §7.5, a pagina 108, è noto che 'contatto' risulta essere correlato, almeno per alcune classi, sia con 'utilizzo' che con 'tipo di costruzione' — si può a ragione obiettare che il dataset operativo per le analisi condotte nel capitolo 7 e quello cui si è fatto riferimento in questa fase non siano gli stessi, tuttavia le medesime analisi di correlazione sono state eseguite nuovamente anche per il dataset di training, ottenendo gli stessi risultati ottenuti in precedenza.
In virtù di questo, ho costruito e testato un modello che ha coinvolto le features *x + y + tipo costruzione + utilizzo*, che risultano *scorrelate* tra loro; i risultati, tuttavia, non hanno portato a nulla di significativo.

modello	*min abs error*	*min sq error*	*best kernel*	*best k*
01	**167**	**111798**	**epanechnikov**	**35**
02	**166**	**111617**	**gaussian**	**56**
03	181	118686	rectangular	36
04	**168**	**113188**	**gaussian**	**69**
05	169	115519	epanechnikov	96
06	175	118376	triangular	107
07	178	116822	rectangular	55
08	166	114115	gaussian	92
sp	176	116215	rectangular	57

Tabella 10.4: Alcuni parametri statistici ricavati in fase di cross-validation per i modelli applicati ai dati reali; i vari modelli sono descritti nella tabella 10.3; in grassetto i modelli che sono stati considerati "migliori"; per maggiori dettagli, si faccia riferimento a quanto discusso nel paragrafo §10.4.1.

I vari modelli sono stati confrontati tra loro e con quello di riferimento 'sp' che coinvolge solo la parte spaziale, conducendo delle cross-validation e visualizzando le relative curve di CV; i risultati numerici sono riportati nella tabella 10.4.

Da questo tipo di analisi e confronti, si può concludere che:

- *le features relative alla configurazione del terreno (curvatura, pendenza e altitudine)* non *sembrano portare informazioni utili*[5]*, ma piuttosto aggiungere rumore al modello stesso;*
- *le features che portano informazioni utili sembrano essere* contatto, tipo costruzione e utilizzo*: questo, oltre a essere più che ragionevole da un punto di vista teorico, è anche in linea con quanto ottenuto da analisi di Feature Selection* [cfr. cap. 9] *e indagini su variogrammi categorizzati* [cfr. cap. 5]*, a ulteriore conferma dell'influenza significativa di queste features sul valore di concentrazione di radon indoor misurata negli edifici;*
- *i modelli 01 e 02 sono gli unici che mostrano una curva di CV con una struttura ben definita, a ulteriore conferma che le features* contatto, tipo costruzione e utilizzo *risultano essere significative per la descrizione del fenomeno;*
- *il modello 07 non porta a risultati soddisfacenti, a (parziale) conferma della* non *significatività della features* qualità degli infissi*;*
- *i risultati "negativi" relativi al modello 08 confermano che* pendenza e curvatura non *portano informazione utile al modello.*

10.4.2 Confronto con dataset di validazione

I vari modelli costruiti sono stati applicati successivamente al dataset di validazione (409 cases, [cfr. tab. 10.1]), col vantaggio di disporre sia del valore stimato dal modello e dall'algoritmo di wk-NN, sia del valore reale di concentrazione. Analisi e confronti sono stati eseguiti ricorrendo

[5] Questo è confermato anche da una semplice analisi visiva basata su scatter-plot 'valore di concentrazione' *vs.* 'feature', che non evidenziano alcuna dipendenza evidente — né lineare, né di altro tipo.

a grafici come quelli riportati a titolo d'esempio per il modello 01 in figura 10.3 e ai parametri statistici relativi al fit lineare basato su minimi quadrati per gli scatter plot 'valori fittati' *vs.* 'valori reali', riportati nella tabella 10.5.

Da questa ulteriore analisi, mi sembra di poter affermare che:

- *graficamente, non si notano particolari e apprezzabili differenze tra i vari modelli;*
- *in relazione ai modelli che risultano* migliori, *anche in questo caso si tratta dei modelli 01, 02 e 04 — in linea con quanto ottenuto nel paragrafo §10.4.1; sembra quindi che le features* contatto, tipo costruzione *e* utilizzo *siano utili e significative per la descrizione del fenomeno radon indoor — e fatto non trascurabile nell'ambito applicativo, sono anche facilmente accessibili;*
- *da grafici del tipo 'fitted' vs. 'conc', risulta evidente un pesante* effetto di smoothing *che caratterizza anche questo approccio — e questo vale per tutti i modelli implementati;*
- *i grafici del tipo 'res' vs. 'conc' evidenziano una netta sottostima dei valori elevati e una sovrastima di quelli bassi (un modo alternativo per rimarcare l'effetto di smoothing già descritto);*
- *nei grafici 'res' vs. 'fitted', in generale* non *si notano strutture evidenti che potrebbero essere indice di un qualche artefatto nella fase di stima; tuttavia, vi è la presenza di un leggero trend verso il basso, a testimonianza di una eventuale lieve tendenza a sovrastimare i valori bassi al crescere del valore fittato;*
- *dall'analisi dei post-plot dei residui,* non *si evidenziano particolari strutture spaziali, indice del fatto che non ci sono situazioni patologiche nelle quali i modelli vadano a concentrare la loro inefficacia (ad esempio, cluster di errori con valori particolarmente elevati), né tuttavia zone in cui i modelli funzionino particolarmente bene rispetto ad altre — a dire, la bontà dei vari modelli risulta essere grossomodo uniforme sull'intero territorio di studio.*

10.4.3 Analisi variografica dei residui

Per valutare la presenza di eventuali e significativi artefatti della procedura di stima, ho ritenuto utile calcolare i variogrammi sperimentali dei residui per i vari modelli implementati: dal punto di vista teorico, questi dovrebbero essere dei puri effetti nugget oscillanti attorno al valore di nugget del variogramma sperimentale — quest'ultimo, calcolato per il dataset di training, risulta piuttosto rumoroso e con una struttura non ben definita (alla luce delle precedenti analisi condotte, questo non stupisce più di tanto); il suo valore di nugget è pari a circa 100000.

Alcuni esempi di variogrammi relativi ai residui per il modello 01 sono riportati in figura 10.4; quello che si può osservare è che:

- *in generale,* non *si nota una struttura particolarmente evidente e, anche in relazione alla marcata rumorosità, si possono globalmente considerare degli effetti nugget oscillanti attorno al valore della varianza a priori relativa ai residui;*
- *analisi più "fini" potrebbero mettere in luce una tendenza dei variogrammi a "salire" verso il valore della relativa varianza a priori, evidenziando così la presenza di una debole struttura; tuttavia, viste le notevoli distanze in gioco (pari a circa*

id modello	*intercetta*	*pendenza*	ρ^2-*adj*	*RMSE*
01	**187±7**	**0.16±0.02**	**0.154**	**78619**
02	**193±7**	**0.14±0.02**	**0.138**	**79999**
03	207±5	0.07±0.01	0.078	85594
04	**190±7**	**0.14±0.02**	**0.196**	**74951**
05	204±5	0.09±0.01	0.104	83367
06	205±4	0.07±0.01	0.071	86475
07	204±5	0.10±0.01	0.119	81969
08	197±6	0.11±0.02	0.110	82405
sp	197±6	0.12±0.01	0.134	80700

Tabella 10.5: Alcuni parametri statistici relativi ai risultati condotti sul dataset di validazione per i modelli impiegati per l'analisi sui dati reali; i valori si riferiscono al fit lineare basato su minimi quadrati per la retta 'stima' vs. 'reale'; ρ^2-adj si riferisce al coefficiente di correlazione corretto per il numero di gradi di libertà; in grassetto, i modelli che sono stati considerati "migliori".

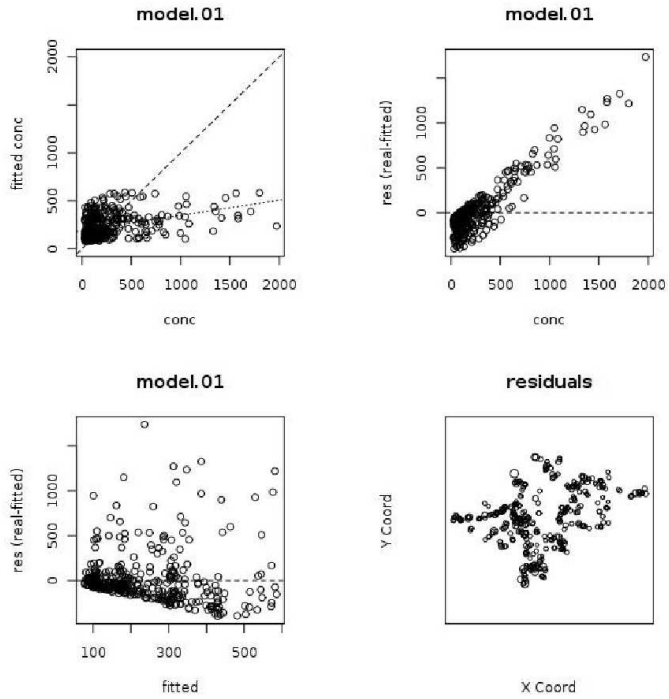


Figura 10.3: Esempio dei grafici utilizzati per valutare i modelli applicati ai dati reali [cfr. tab. 10.3] sulla base delle analisi condotte sul dataset di validazione; il grafico in basso a destra riporta un post-plot dei residui nel quale la dimensione del punto è proporzionale al valore del residuo stesso; i parametri statistici relativi al fit lineare riportato sono raccolti nella tabella 10.5.

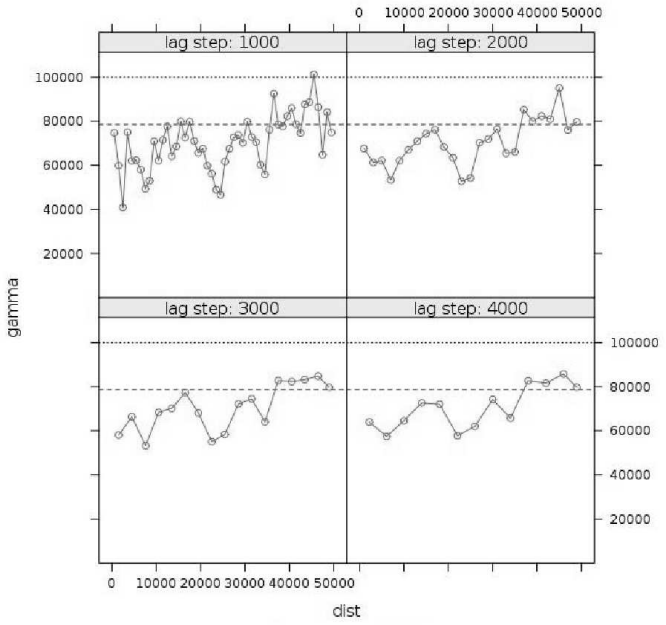


Figura 10.4: Esempio di variogramma dei residui per il modello 01, al variare del lag-step; i valori di distanza sono riportati in metri; la linea tratteggiata rappresenta la varianza a priori relativa ai residui, quella punteggiata il valore di nugget relativo al variogramma sperimentale calcolato per il dataset di training (dataset operativo).

50 km), non mi sembra che questo risultato sia da considerarsi particolarmente significativo;

- *i variogrammi dei residui si stabilizzano su un valore* inferiore *rispetto a quello del nugget che caratterizza il variogramma sperimentale di riferimento: questo potrebbe essere indice del fatto che il modello* non *è comunque in grado estrarre* tutta l'informazione disponibile.

10.5 Conclusioni

Sulla base dell'esperienza acquisita e in base a quanto ottenuto da precedenti analisi di tipo geostatistico, è emerso in maniera piuttosto evidente come il considerare unicamente la componente spaziale del fenomeno radon indoor non sia sufficiente per una trattazione esaustiva dello stesso. Nel tentativo di superare questi limiti, si è deciso di ricorrere a tecniche alternative per la costruzione di modelli in grado di prendere in considerazione anche l'eventuale informazione contenuta nelle numerose variabili secondarie (features) che accompagnano la singola misura di attività di radon indoor [cfr. anche quanto descritto nel capitolo 11]; si crede infatti che le caratteristiche dell'edificio sede della misura abbiano un'influenza significativa sul valore misurato, benché la loro interazione risulti estremamente complessa e di difficile modellizzazione.

Le analisi sono state condotte sia su un dataset di riferimento costruito ad hoc, con lo scopo di poter controllare in maniera sicura i vari aspetti di interesse, sia su un dataset reale ricavato da quello descritto nel paragrafo §1.3; si sono quindi costruiti differenti modelli da porre sotto analisi,

al variare del numero e del tipo di variabili secondarie — e quindi del tipo di informazione — accessibili ai modelli stessi.

Da quanto emerso in questa prima fase di analisi relativa all'approccio Weighted k-Nearest Neighbor, combinando i risultati ottenuti per i vari modelli costruiti, si può concludere che:

- le variabili quantitative relative alla configurazione del terreno non sembrano portare informazioni utili, ma piuttosto aggiungere rumore al modello;
- le variabili secondarie più ricche di informazione utile sono risultate essere *contatto*, *tipo di costruzione* e *utilizzo*: interessante come questo risultato, oltre a essere più che ragionevole da un punto di vista teorico, sia inoltre in linea con quanto si è ottenuto con analisi differenti, quali Feature Selection [cfr. cap. 9], analisi variografiche su specifici subset [cfr. cap. 5] e analisi General Regression Neural Network [cfr. cap. 11]; i risultati ottenuti in questa fase sembrano quindi confermare la significatività dell'influenza di queste variabili secondarie sul valore di concentrazione di radon misurato all'interno degli edifici;
- questo approccio sembra configurarsi come un buon *tool esplorativo* per indagare le proprietà dei dati di radon indoor e la loro relazione con le numerose variabili secondarie che accompagnano ogni singola misurazione; in particolare, si è rivelato utile per *identificare* le features che mostrano una qualche *influenza significativa* sul valore di concentrazione misurato negli edifici; a sostegno di questo, i risultati sono in linea con quanto ottenuto ricorrendo ad approcci differenti, come quelli affrontati e descritti nei capitoli 8 e 9;
- alla luce invece di quanto emerso mettendo a confronto le stime ottenute con wk-NN con un dataset reale di validazione, questo approccio *non* sembra essere uno strumento altrettanto valido per produrre delle mappe del fenomeno radon indoor — o quantomeno, globalmente non si ottengono risultati migliori rispetto a quanto si è ottenuto finora con gli approcci più convenzionali di tipo geostatistico[6].

[6] A tal proposito, si rimanda tuttavia a quanto discusso nel capitolo 12, a pagina 177, nel quale emerge come a livello *locale* questo tipo di approccio dia qualche risultato migliore rispetto a quanto ottenuto mediante Ordinary Kriging.

CAPITOLO 11

L'approccio General Regression Neural Network

Nella convinzione, supportata dall'esperienza acquisita e da studi condotti in ambiti differenti, che la sola parte spaziale *del fenomeno radon indoor non sia sufficiente a una trattazione e comprensione soddisfacente del fenomeno stesso, l'idea che ha ispirato il lavoro presentato in questo capitolo è stata quella di applicare gli strumenti offerti dallo stato dell'arte della statistica non-parametrica, e nello specifico l'approccio General Regression Neural Network, basato sulle Reti Neurali Artificiali, ai dati relativi alle misure di concentrazione di attività di radon indoor.*
Lo scopo è quindi quello di costruire dei modelli che coinvolgano, accanto alla parte spaziale che sicuramente ha una influenza non trascurabile, anche altri fattori legati in particolar modo alle caratteristiche proprie dell'edificio sede della misura; fattori la cui influenza è da ritenersi più che ragionevolmente significativa, ma la cui modellizzazione risulta piuttosto complessa.

La parte computazionale è stata svolta con il software `Geostat Office`© per quanto riguarda la diretta implementazione dell'algoritmo, mentre la parte di analisi statistica dei risultati è stata svolta in ambiente `R`{44}, ricorrendo a routine create ad hoc e ai packages *lattice*{45}, *geoR*{30} e *gstat*{41}.

11.1 Brevi richiami teorici

ANN

Le Reti Neurali Artificiali (Arificial Neural Network, ANN) sono sistemi analitici introdotti per affrontare problemi la cui soluzione non viene formulata (a volte, perché impossibile) in maniera esplicita; esse sono costituite di NEURONI, semplici e singole unità operative che possono essere programmate in modo da ottenere i risultati richiesti. La loro struttura si basa essenzialmente su quella biologica, riproducendo matematicamente i neuroni e la complessa struttura che li lega gli uni agli altri negli organismi viventi.

A differenza dei metodi statistici convenzionali, è ad esempio possibile programmare e testare una ANN per stimare una funzione campionata senza però conoscere la forma della funzione stessa: la ANN sono infatti in grado di stimare una funzione senza conoscere a priori — a dire, senza avere a disposizione un modello matematico — in quale modo gli output dipendano dagli input. Si dice che la ANN siano *modelli semi-parametrici*, nel senso che sono in grado di *imparare dall'esperienza*.

Le reti neurali artificiali si rendono superiori rispetto ad altri metodi nelle situazioni per le quali [cfr. {36}]:

- i dati su cui le analisi e le conclusioni saranno basate risultano confusi ('fuzzy') e/o soggetti a potenziali grandi errori;
- le relazioni tra i dati sono così nascoste che possono risultare del tutto invisibili all'operatore che ricorra ai metodi della statistica classica;
- i dati manifestano correlazioni anche molto distanti da quelle di tipo lineare;
- i dati risultano caotici dal punto di vista matematico.

Un neurone artificiale è una unità in grado di processare l'informazione che viene a esso fornita, e costituisce quindi il perno fondamentale per le operazioni di una ANN. Sono essenzialmente tre gli elementi di base di un modello di neurone:

1. un insieme di *sinapsi* o legami, ognuno dei quali è caratterizzato da un suo proprio peso (o forza del legame); il peso sarà positivo nel caso di una associazione tra sinapsi di tipo eccitatorio, negativo nel caso di una associazione di tipi inibitorio;
2. un *integratore* atto all'integrazione dei segnali in ingresso; tipicamente, si tratta di una funzione somma;
3. una *funzione di attivazione* non-lineare, per limitare l'ampiezza del segnale in uscita dal neurone (esattamente come avviene nel caso biologico).

Un neurone artificiale avrà quindi una serie di ingressi, in perfetta analogia con i dendriti del suo omologo biologico, e sarà in grado di combinare i segnali ricevuti, tipicamente mediante una somma pesata, per formare un livello di attivazione interno al neurone stesso: maggiore tale livello di attivazione, tanto più forte il segnale che esso invierà in uscita agli altri neuroni della rete artificiale ad esso collegati.

Matematicamente, una ANN ha le seguenti proprietà:

- una variabile a_i è associata a ogni nodo i;
- un peso $w_{ij} \in \mathbb{R}$ è associato a ogni legame ij tra i nodi i e j;
- un valore di bias $b_i \in \mathbb{R}$ è associato a ogni nodo i;
- una funzione di trasferimento/attivazione f_i è definita per ogni nodo i della rete; questa funzione determina lo stato del nodo in funzione del valore del bias b_i, del peso w_{ij} associato al segnale in ingresso che arriva dal legame ij e dello stato del nodo collegato a esso.

Tra le più frequentemente utilizzate in letteratura, si ricordano le seguenti funzioni di trasferimento/attivazione:

- la funzione logistica:

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{11.1}$$

- la funzione tangente iperbolica:

$$f(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{11.2}$$

11.1.1 General Regression Neural Network (GRNN)

Gli algoritmi che si rifanno all'approccio General Regression Neural Network appartengono ai ben noti modelli di regressione non-parametrici; per approfondire dal punto di vista teorico, si può fare riferimento al testo di Fan e Gijbels {20}, che riporta anche una dettagliata ed esaustiva bibliografia. *GRNN*

Tra i vantaggi di questo approccio, la possibilità di poter ottenere non solo la stima per il valor medio condizionato ai dati di riferimento in localizzazioni non campionate, ma l'intera p.d.f., mediante la quale poter stimare anche momenti di ordine superiore della stessa; inoltre, si evitano problemi di minimizzazione che possono "intrappolare" la procedura di stima di alcuni parametri del modello in un minimo locale (che può distare anche molto dal minimo assoluto), come ad esempio nel caso dell'approccio Multi Layer Perceptron (sempre basato su ANN). Tuttavia, come tutti i metodi a kernel non parametrici, i modelli GRNN hanno anche qualche svantaggio, il principale dei quali è che, essendo F la funzione che si intende stimare, tutte le approssimazioni della F saranno stime *polarizzate* della funzione non-nota di riferimento G; questo perché in generale non esiste una funzione F^* con un numero finito di parametri in grado di approssimare la funzione G.

La teoria delle GRNN si fonda sulla regressione a kernel multivariata [cfr. ad esempio {20}], il cui obiettivo è quello di approssimare la funzione densità di probabilità $F(z_1^*, \ldots, z_m^*)$ delle m variabili casuali $\mathbf{z} = (z_1, \ldots, z_m)$ ricorrendo alle n misurazioni di ogni singola variabile.

Lo stimatore a kernel multivariato $\widehat{F}$ nel caso m-dimensionale è definito come:

$$\widehat{F}(z^*) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1 \ldots h_m} K\left(\frac{z_{i1} - z_1^*}{h_1}, \ldots, \frac{z_{im} - z_m^*}{h_m} \right) \tag{11.3}$$

dove $K(\cdot)$ rappresenta la funzione a kernel multivariata e il vettore $\mathbf{h} = (h_1, \ldots, h_m)$ il parametro relativo alla larghezza di banda (bandwidth) del kernel, che gioca un ruolo fondamentale nella costruzione del modello. Tipicamente, le funzioni a kernel sono funzioni della distanza d simmetriche e decrescenti, da un valore massimo allo zero, e godono delle seguenti proprietà:

- $K(d) \geq 0 \quad \forall d \in \mathbb{R}$
- $K(d)$ ammette massimo per $d = 0$
- $K(d)$ decresce in maniera monotona per $d \to \pm\infty$

Il problema che si incontra nella fase di preparazione della rete neurale che sottende l'approccio GRNN, condotto su un dataset di training, è quello di dover trovare il parametro h *ottimale*, che nella pratica non è noto; una delle tecniche più diffuse e collaudate consiste nel ricorre a una cross-validation (CV) e all'analisi, con metodi statistici e/o geostatistici, dei residui ottenuti al variare del valore della larghezza di banda h. Successivamente, la rete dovrebbe essere testata

e validata ricorrendo a un dataset di validazione indipendente da quello di training e quindi impiegata per la fase di generalizzazione, ovvero per la stima in localizzazioni non campionate.

Il parametro h, come anticipato, è di fondamentale importanza in quanto influenza il tipo di soluzione che si può ottenere:

- se h è molto piccolo (ogni punto è indipendente dagli altri, ossia non viene riconosciuta alcuna correlazione tra loro), cioè $h \to 0$, la soluzione converge a una semplice interpolazione, ovvero $Z_m \to Z_i$ se $(x, y) \to (x_i, y_i)$ avendo indicato con (x, y) le coordinate del punto di stima;

- se h è invece molto grande (tutti i punti hanno la stessa influenza sugli altri, ossia si considera una p.d.f. uniforme), si applica un pesante effetto di smoothing e la soluzione converge a una approssimazione; nel caso limite per il quale $h \to \infty$, allora $Z_m \to 1/n \sum_i Z_i$, e quella che si ottiene non è altro che la media campionaria del dataset di riferimento.

11.2 Descrizione del dataset utilizzato

Il dataset operativo cui ho fatto riferimento per questo tipo di analisi è lo stesso descritto nel dettaglio nel paragrafo §10.2 a pagina 149, che si compone di tre differenti subset, uno di 'training' (820 cases), uno di 'testing' (409 cases) e uno di 'validation' (409 cases) costituiti da misure di concentrazione di attività di radon indoor georeferenziate e accompagnate da una nutrita serie di informazioni aggiuntive relative alle caratteristiche proprie dell'edificio sede della misura e alle caratteristiche morfologiche del terreno circostante.

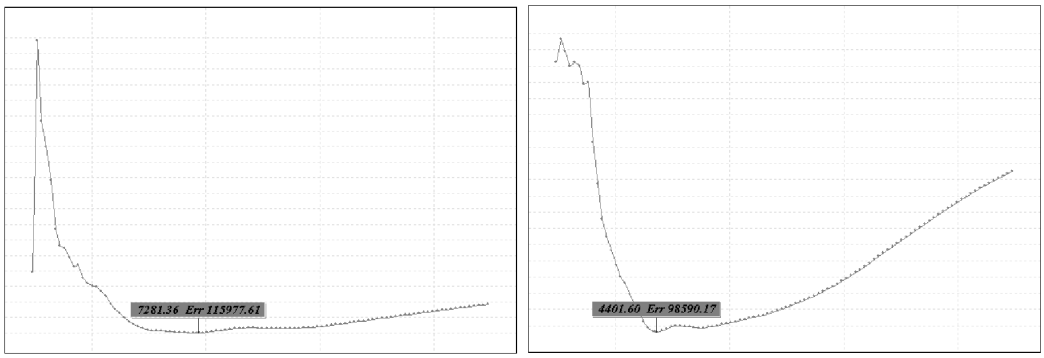
11.3 Studio dell'influenza della parte spaziale

Inizialmente, con il duplice scopo di i) testare l'algoritmo nella pratica variando alcuni parametri e ii) avere a disposizione un modello di riferimento, ho deciso di limitare l'analisi alla sola parte spaziale del fenomeno, considerando quindi unicamente le coordinate come fonte di informazione. Sono stati costruiti quattro differenti modelli, identificati come 'sp', al variare del tipo di kernel impiegato e senza operare alcuno scaling dei dati (che ricordo, in questa prima fase sono unicamente le coordinate).

dataset 'sp'

La figura 11.1a riporta un esempio di curva di CV ottenuta mediante cross-validation sul dataset operativo 'training'; questo tipo di curva è stato utilizzato per determinare il valore ottimale del parametro h_c, ovvero la larghezza di banda del kernel — il pedice "c" si riferisce al fatto che in questa prima fase non ho ritenuto opportuno introdurre anisotropie direzionali nel valore di h, a dire che tale valore risulta così *comune* a tutte le direzioni. La figura 11.1b riporta invece un esempio di curva di CV ottenuta ricorrendo al dataset 'testing': in questo caso, invece di condurre una cross-validation sul dataset operativo, si ricorre a un dataset indipendente per determinare il valore ottimale del parametro h_c; il vantaggio teorico di questa procedura risiede nel fatto che si ottiene un miglior controllo della *complessità* del modello, a dire che, almeno dal punto di vista teorico, il rischio di incappare in un overfitting dei dati dovrebbe essere inferiore rispetto a un approccio di cross-valdation.

I risultati numerici relativi ai quattro modelli costruiti, al variare del tipo di kernel e del tipo di procedura utilizzata per la determinazione del parametro h_c sono riportati nella tabella 11.1.



(a) risultati da CV (b) risultati da dataset 'testing'

Figura 11.1: Esempi di curve di cross-validation per il modello *sp03*; l'asse orizzontale riporta il valore del parametro h, quello verticale l'errore; si noti come nel caso si ricorra al un dataset 'testing', la struttura della curva di CV risulti più evidente.

id modello	h_c	*errore CV*	*kernel*	*note*
sp01	5402	98179	gaussiano	con dataset di 'testing'
sp02	8389	97361	rettangolare	con dataset di 'testing'
sp03	4401	98590	epanechnikov	con dataset di 'testing'
sp04	14252	98465	triangolare	con dataset di 'testing'
sp01	6256	116013	gaussiano	con CV
sp02	9069	115719	rettangolare	con CV
sp03	7281	115917	epanechnikov	con CV
sp04	16315	115798	triangolare	con CV

Tabella 11.1: Parametri relativi ai risultati ottenuti per i modelli che coinvolgono solo la parte spaziale, al variare del tipo di kernel e della procedura impiegata per la determinazione del valore ottimale del parametro h_c (larghezza di banda del kernel); alcuni esempi di grafici delle relative curve di CV sono riportati in figura 11.1.

Da questa analisi preliminare, si può quindi concludere che:

- *in tutti i casi, ricorrere a un dataset di 'testing' porta a una curva di CV (grafico 'errore' vs. 'h_c') con una struttura più evidente rispetto al caso di una semplice cross-validation sul solo dataset operativo 'training' — il minimo risulta più definito* [cfr. ad esempio quanto riportato in fig. 11.1];
- *con un kernel* isotropo, *in tutti i casi le curve di CV mostrano una struttura ben definita, buon indizio che i dati sono anch'essi caratterizzati da una struttura/correlazione spaziale — a conferma di risultati già noti;*
- *h_c risulta in linea con una previsione teorica in relazione al tipo di kernel: più la forma del kernel è stretta, maggiore il relativo valore di h_c per ottenere una stima affidabile*[1]*;*

[1] Una rappresentazione grafica bidimensionale dei tipi di kernel impiegati in questa analisi è riportata nella figura 10.1, a pagina 148.

- *dall'analisi degli errori — soprattutto per quanto riguarda quelli ottenuti con una procedura di semplice cross-validation — non sembra che il tipo di kernel scelto abbia una sensibile influenza sui risultati globali.*

Limitatamente al kernel di tipo gaussiano, `Geostat Office`© consente di stimare il parametro h introducendo delle asimmetrie, ovvero di stimare il valore di h_x e h_y. Questo affinamento della procedura è stato provato sia sul dataset operativo, sia sul dataset 'dummy'[2] con un modello del tipo "sim1 = x + y", dove x e y rappresentano le coordinate spaziali. Inoltre, sono anche state condotte delle analisi variografiche con lo scopo di valutare una presenza evidente di anisotropia (geometrica e/o zonale, [cfr. §B.3.1, pag. 229]) nella struttura spaziale di entrambi i dataset; se, prevedibilmente, per il dataset dummy l'anisotropia è confermata dall'analisi variografica, il dataset 'training' manifesta (e a questo punto, alla luce dell'esperienza acquisita, certo non stupisce) una struttura molto meno leggibile, se non ricorrendo a misure della correlazione spaziale non convenzionali (tipo Pairwise Relative Variogram, {16, pag. 45}).

Alla luce di quanto appena descritto e del fatto che i risultati ottenuti non sono stati del tutto riproducibili e univocamente interpretabili, ho deciso in questa fase di non introdurre anisotropie nella determinazione del parametro h.

11.3.1 Test di accuratezza

Per un confronto oggettivo dei modelli spaziali descritti, ho condotto dei test di accuratezza sugli stessi ricorrendo al dataset 'training' e al dataset 'dummy'. I risultati sono riportati nella tabella 11.2, dalla quale si evince che:

- per quanto riguarda il dataset 'dummy', i risultati sono, come è logico e lecito aspettarsi, molto buoni;
- in generale, si ottengono risultati migliori se si utilizza il dataset di 'testing' rispetto a una semplice cross-validation;
- il kernel di tipo *epanechnikov* (modello *sp03*) sembra dare i risultati migliori;
- la sola parte spaziale sembra essere in grado, nei casi migliori, di spiegare un 15% dell'intera variabilità del fenomeno in esame.

11.3.2 Test di generalizzazione

Come nel caso dei test di accuratezza, ho eseguito anche dei test di generalizzazione per valutare la capacità del modello di stimare valori di concentrazione in localizzazioni non campionate; per questo, si è rivelato utile il dataset 'validation', mediante il quale si hanno a disposizione valori della variabile di riferimento in 409 nuove localizzazioni, valori che si sottolinea non sono stati precedentemente sfruttati in nessuna fase di stima dei parametri del modelli spaziali (in questo caso, il solo parametro h). I risultati sono riportati nella tabella 11.3.

Si può facilmente notare che:

- il kernel di tipo *triangolare* sembra essere quello che porta a risultati migliori in relazione alle capacità di generalizzazione del modello;

[2] Questo dataset, costruito per avere a disposizione una serie di dati noti e sotto controllo, è descritto nel dettaglio nel paragrafo §10.3.1, a pagina 150.

id modello	*errore CV*	*RMSE*	ρ^2	*note*
sp01	98179	327.3	0.1409	con dataset di 'testing'
sp02	97361	330.5	0.1158	con dataset di 'testing'
sp03	98590	325.6	0.1446	con dataset di 'testing'
sp04	98465	328.5	0.1338	con dataset di 'testing'
sp01	111013	329.9	0.1260	con CV
sp02	115719	329.3	0.1231	con CV
sp03	115917	333.6	0.1034	con CV
sp04	115798	330.6	0.1228	con CV
dummy	2838	1.295	0.8064	con CV

Tabella 11.2: Parametri statistici relativi ai test di accuratezza condotti per i modelli spaziali sul dataset operativo 'training'; in questo caso, si riporta anche un confronto con quanto ottenuto per un dataset 'dummy' costruito ad hoc, per il quale è nota la dipendenza della variabile di interesse dalle coordinate; per maggiori dettagli, si faccia riferimento al testo.

id modello	*RMSE*	ρ^2	*note*
sp01	283.9	0.1409	con dataset di 'testing'
sp02	291.9	0.1158	con dataset di 'testing'
sp03	287.5	0.1446	con dataset di 'testing'
sp04	283.5	0.1338	con dataset di 'testing'
sp01	283.8	0.1260	con CV
sp02	288.0	0.1231	con CV
sp03	248.1	0.1034	con CV
sp04	283.1	0.1228	con CV

Tabella 11.3: Parametri statistici relativi ai test di generalizzazione condotti per i modelli spaziali sfruttando il dataset 'validation'; per maggiori dettagli, si faccia riferimento al testo.

- in opposizione a quanto ottenuto per i test di accuratezza, i risultati migliori sembra si ottengano in questo caso ricorrendo a una procedura di cross-validation per la stima del parametro h; mi sembra comunque importante sottolineare come la differenza rispetto alla procedura che prevede l'impiego di una dataset indipendente di 'testing' siano minime — e non statisticamente significative.

Nota 1: *Poiché stimare il parametro h con un dataset di 'testing' dovrebbe, dal punto di vista teorico, fornire dei risultati migliori in relazione alle capacità di* generalizzazione *del modello, ho ritenuto opportuno dividere anche il dataset 'dummy' in tre subset, perfettamente in linea con quanto fatto per il dataset operativo. Ho quindi ripetuto tutte le analisi precedenti per il modello "sim1 = x + y", determinando sia il parametro h_c che la coppia (h_x, h_y).*
Anche in questo caso, si ottengono risultati leggermente *migliori ricorrendo a una procedura di stima del/i parametro/i del modello basata su cross-validation, ma con differenze trascurabili.*
Per quanto riguarda invece l'anisotropia che caratterizza il dataset 'dummy', questa viene correttamente riconosciuta in fase di stima del/i parametro/i, ma prenderla in considerazione in fase di generalizzazione non *porta a differenze apprezzabili sui risultati ottenuti.*

Analisi dei residui

In analogia a quanto descritto nel dettaglio nei paragrafi §10.4.2 e §10.4.3, e ricorrendo alle stesse procedure di analisi condotte in ambiente R{44}, ho analizzato i residui ottenuti in fase di stima utilizzando il dataset 'validation'; anche in questo caso, è stata eseguita anche un'analisi spaziale e variografica dei residui. I risultati numerici sono riportati nella tabella 11.4, mentre la figura 11.2 riporta un esempio relativo ai variogrammi dei residui per il modello 'sp03'.

Da questo tipo di analisi, si può concludere che:

- *dall'analisi dei post-plot proporzionali dei residui,* non *si evidenziano particolari strutture spaziali, indice del fatto che non ci sono situazioni patologiche nelle quali i modelli vadano a concentrare la loro inefficacia (ad esempio, cluster di errori con valori particolarmente elevati), né zone in cui i modelli funzionino particolarmente bene rispetto ad altre — a dire, la bontà dei vari modelli risulta essere grossomodo uniforme sull'intero territorio di studio (risultato analogo a quanto ottenuto nel capitolo 10 in relazione all'approccio wk-NN);*
- *in generale, i variogrammi dei residui* non *mostrano una struttura particolarmente evidente e, anche in relazione alla marcata rumorosità, si possono globalmente considerare degli effetti nugget oscillanti attorno al valore della varianza a priori relativa ai residui — come attesa dal punto di vista teorico;*
- *i variogrammi dei residui si stabilizzano su un valore* inferiore *rispetto a quello del nugget che caratterizza il variogramma sperimentale di riferimento: questo potrebbe essere indice del fatto che il modello* non *è comunque in grado estrarre* tutta l'informazione disponibile *— questo a ulteriore supporto dell'ipotesi che la sola componente spaziale* non *sia in grado di rendere conto della complessità del fenomeno radon indoor;*
- *le differenze tra tipi di kernel e procedura di stima del parametro h risultano trascurabili in relazione alla capacità di generalizzazione del modello.*

11.4 Test su dataset 'dummy'

Con lo scopo di valutare il comportamento dell'algoritmo su un dataset del quale fossero completamente note le relazioni tra la variabile di interesse e le covariate atte alla sua descrizione, in analogia a quanto ampiamente descritto nel capitolo 10 ho ritenuto opportuno condurre le stesse analisi anche in questo contesto, ricorrendo al dataset 'dummy' descritto e commentato nel dettaglio nel paragrafo 10.3.1 a pagina 150.

I risultati ottenuti per h_c, sia in relazione ai test di accuratezza che a quelli di validazione, sono riportati nella tabella 11.5; da questi e dall'analisi delle relative curve di CV, si può concludere che[3]:

- *in tutti i casi, la curva di CV manifesta una struttura ben definita, indice del fatto che i dati sono a loro volta strutturati in relazione alla serie di covariate considerate;*

[3] Non stupisce certo che quanto ottenuto per questo approccio sia in linea con quanto ottenuto per l'approccio di Machine Learnign wk-NN, descritto nel capitolo 10.

id modello	*intercetta*	*pendenza*	ρ^2-*adj*	*RMSE*	*note*
sp01	192±6	0.13±0.02	0.1326	284	con dataset di 'testing'
sp02	196±7	0.12±0.02	0.0910	292	con dataset di 'testing'
sp03	194±7	0.14±0.02	0.1133	287	con dataset di 'testing'
sp04	193±6	0.13±0.02	0.1357	283	con dataset di 'testing'
sp01	194±6	0.13±0.02	0.1336	284	con CV
sp02	195±7	0.13±0.02	0.1093	288	con CV
sp03	196±6	0.12±0.01	0.1337	284	con CV
sp04	194±6	0.12±0.02	0.1394	283	con CV

Tabella 11.4: Alcuni parametri statistici relativi ai risultati condotti sul dataset di validazione per i modelli impiegati per l'analisi sui dati reali; i valori si riferiscono al fit lineare basato su minimi quadrati per la retta 'stima' vs. 'reale'; ρ^2-adj si riferisce al coefficiente di correlazione corretto per il numero di gradi di libertà.

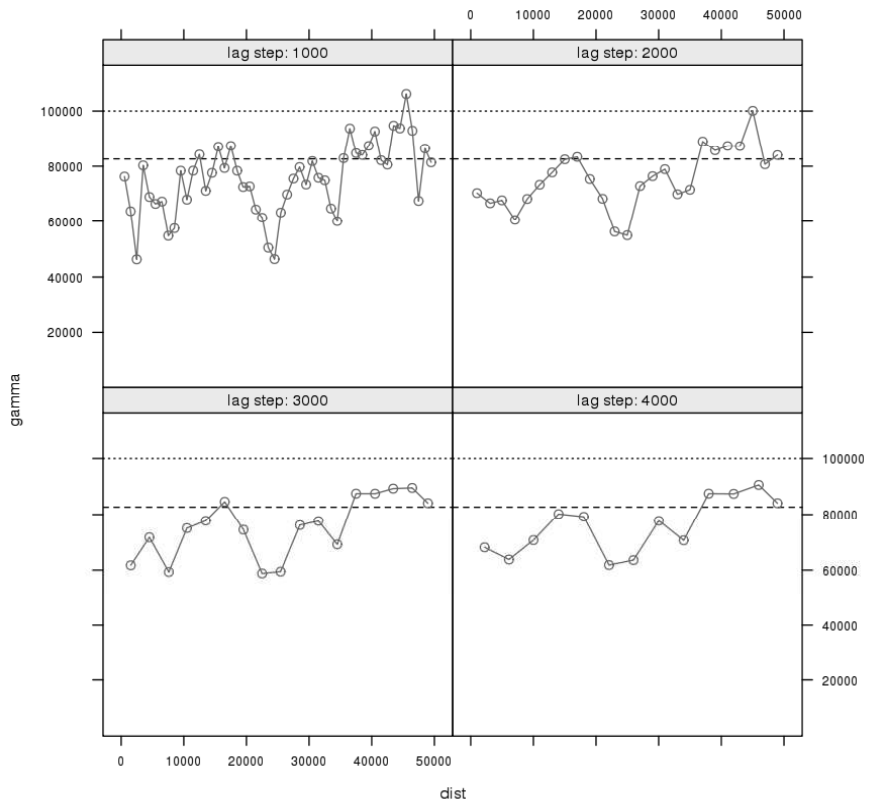


Figura 11.2: Esempio di variogramma dei residui per il modello spaziale 'sp03' al variare del lag-step; i valori di distanza sono riportati in metri; la linea tratteggiata rappresenta la varianza a priori relativa ai residui, quella punteggiata il valore di nugget relativo al variogramma sperimentale calcolato per il dataset di training (dataset operativo).

	test di accuratezza		*test di generalizzazione*		
val =	*RMSE*	ρ^2	*RMSE*	ρ^2	*errore CV*
x+y	22.81	0.2263	22.68	0.1141	0.105
+ *cat2*	22.33	0.2674	23.31	0.0754	0.109
+ *cat4*	22.31	0.2929	24.32	0.0199	0.115
+ *alt*	12.92	0.7630	15.91	0.5659	0.055
+ θ	9.06	0.8964	15.74	0.5732	0.051
+ *sim1*	5.48	0.9641	12.54	0.7382	0.031

Tabella 11.5: Alcuni parametri statistici ricavati in fase di cross validation per i modelli descritti nella prima colonna; si noti come all'aumentare dell'informazione fornita al modello, i risultati migliorino (il modello completo è infatti dato dall'eq. (10.6), pag. 151); si tenga presente che i vari modelli sono riportati in maniera "incrementale", ovvero la nuova covariata si va ad aggiungere a quelle precedenti.

- *come ci si aspetta da una previsione teorica, all'aumentare delle informazioni accessibili per il modello si ha un generale miglioramento dei risultati ottenuti.*

Problema: *La medesima procedura è stata applicata anche stimando la coppia di parametri* (h_x, h_y) *e introducendo quindi la possibilità di considerare eventuali anisotropie. In generale, anche in questo caso si ottengono risultati che evidenziano come al crescere della quantità di informazioni fornite al modello — numero di covariate — i risultati, sia per quanto riguarda i test di accuratezza che di generalizzazione, siano migliori.*
Tuttavia, si sono riscontrati dei problemi di riproducibilità *legati probabilmente all'implementazione software dell'algoritmo. Ho infatti notato che, eseguendo run successivi partendo con gli stessi dati e parametri di ingresso, i risultati si sono però rivelati lievemente differenti; per questo, pur ottenendo risultati comunque "ragionevoli", non mi sembra corretto dare a tali conclusioni un fondamento solido.*

Infine, il dataset 'dummy' si è rivelato utile anche per testare l'influenza della procedura di *scaling dei dati* da effettuare in fase di preparazione degli stessi; infatti, in questo dataset ci sono variabili con range di valori molto differenti tra loro, e quelle con valori elevati (ad esempio, le coordinate) potrebbero mascherare (anche in maniera pesante) quelle con valori inferiori (tipicamente, quelle di tipo categorico che vengono codificate con valori dell'ordine dell'unità). Operando quindi su dati con e senza uno scaling preventivo, si sono confrontati i risultati dai quali si evince che

> *nelle situazioni per le quali le covariate coinvolte nel modello abbiano range di variabilità numerica molto diversi tra loro, risulta opportuno, al fine di non perdere le informazioni portate da quelle caratterizzate da valori numerici inferiori, operare uno* scaling dei dati *in fase di pre-processing.*

11.5 Analisi dei dati reali: modelli più complessi

Il dataset impiegato per lo studio dei dati reali è quello cui si è fatto riferimento con l'etichetta 'training', composto di 820 valori di concentrazione di radon indoor georeferenziati, e descritto nel dettaglio nel paragrafo §10.2 a pagina 149.

Con lo scopo di valutare tanto l'influenza del tipo di informazione fornita al modello quanto i risultati ottenuti da analisi differenti [cfr. capitoli 5, 9 e 10], ho costruito sei differenti modelli, le cui caratteristiche di base sono riassunte nella tabella 11.6.

id modello	*covariate coinvolte*
01	x + y
02	x + y + contatto + tipo costruzione + utilizzo
03	x + y + altitudine + pendenza + curvatura
04	x + y + tipo costruzione + contatto + esposizione
05	x + y utilizzo + tipo costruzione + contatto + esposizione
06	tutte

Tabella 11.6: Descrizione dei modelli utilizzati per lo studio condotto sui dati reali.

Tra gli aspetti che hanno guidato la scelta delle covariate da inserire nei vari modelli, quello, a mio avviso non trascurabile, di tenere presente che in fase di stima, i valori di tali covariate devono essere accessibili sperimentalmente; inoltre, questo avrà anche ripercussioni sul *tipo* di mappe che eventualmente si vorranno realizzare.

> ***Nota 2:*** *Se lo scopo dell'analisi è infatti quello di produrre una mappa* generale *che fornisca le caratteristiche globali del fenomeno, nel caso più "semplice" della geostatistica, la mappa prodotta sarà funzione unicamente delle coordinate del punto di stima (una volta fissati il modello di variogramma e i parametri propri dell'algoritmo di stima, tipicamente un kriging); con questo tipo di approccio, invece, ci si troverà nella situazione di produrre più mappe in funzione non solo delle coordinate, ma anche dei valori propri delle covariate coinvolte — a dire, considerando ad esempio la covariata 'contatto', si avrà una mappa per* contatto = sì *e una per* contatto = no. *Questo problema non si pone, invece, se l'obiettivo finale sarà quello di una stima* puntuale, *ovvero stimare il valore di concentrazione per un nuovo (o perché no anche esistente) edificio di cui siano noti i parametri coinvolti nella definizione del modello stesso.*

Operativamente, le analisi sono state condotte con i seguenti parametri:

- preventivo *scaling dei dati* per evitare i già citati problemi di mascheramento delle variabili con valori numerici bassi da parte di quelle con valori elevati (in questo contesto, evidente il caso delle coordinate rispetto alla codifica delle variabili categoriche);
- per ogni modello, stima sia del singolo parametro h_c (nessuna anisotropia) che della coppia (h_x, h_y);
- utilizzo del dataset di 'testing' nella fase di determinazione dei parametri del modello;
- ricorso a un kernel di tipo *gaussiano*, l'unico per il quale il software utilizzato consenta di introdurre anisotropie.

Il modello 02 coinvolge le covariate che hanno dato i risultati migliori nel caso di wk-NN [cfr. §10.4.2, pag. 155]; il modello 03 prendere in considerazione solo le caratteristiche del terreno, facilmente accessibili mediante strumenti di tipo GIS una volta in possesso di un modello digitale del terreno; il modello 04 è stato introdotto sulla base dei risultati ottenuti nelle analisi di Feature Selection descritte nel capitolo 9.

Analisi e confronti sono stati eseguiti ricorrendo a grafici come quelli riportati a titolo d'esempio per il modello 04 in figura 11.3 e ai parametri statistici relativi al fit lineare basato su minimi quadrati per gli scatter plot ' valori fittati ' *vs.* ' valori reali ', riportati nella tabella 11.8. Inoltre, la tabella 11.7 riporta i risultati ottenuti in fase di CV eseguendo sia test di accuratezza di generalizzazione.

Si può quindi affermare che:

	test di accuratezza		*test di generalizzazione*			
id modello	*RMSE*	ρ^2	*RMSE*	ρ^2	*errore CV*	*note*
01	327.7	0.1386	283.3	0.1389	0.0273	h_c
	327.3	0.1417	283.0	0.1422	0.0273	(h_x, h_y)
02	295.8	0.3064	281.3	0.1689	0.0269	h_c
	280.4	0.3893	286.5	0.1440	0.0261	(h_x, h_y)
03	304.4	0.3000	290.6	0.0938	0.0283	h_c
	297.9	0.3210	280.2	0.1586	0.0269	(h_x, h_y)
04	307.6	0.2617	275.8	0.1824	0.0281	h_c
	273.6	0.4323	276.0	0.1998	0.0255	(h_x, h_y)
05	315.0	0.2135	281.9	0.1462	0.0281	h_c
	258.9	0.4893	280.8	0.1851	0.0254	(h_x, h_y)
06	241.0	0.6070	286.6	0.1384	0.0276	h_c
	176.0	0.7639	307.9	0.2182	0.0234	(h_x, h_y)

Tabella 11.7: Alcuni parametri statistici relativi ai risultati ottenuti in fase di cross-validation per i modelli costruiti per i dati reali e descritti nella tabella 11.6; per maggiori dettagli, si faccia riferimento al testo.

id modello	*intercetta*	*pendenza*	ρ^2*-adj*	*RMSE*	*note*
01	192±6	0.13±0.01	0.1367	283	h_c
02	176±10	0.23±0.02	0.1689	281	h_c
03	201±5	0.08±0.01	0.0916	291	h_c
04	185±8	0.20±0.02	0.1804	276	h_c
05	191±7	0.15±0.02	0.1441	282	h_c
06	184±9	0.19±0.02	0.1363	287	h_c
01	192±6	0.13±0.02	0.1401	283	(h_x, h_y)
02	179±10	0.21±0.02	0.1492	286	(h_x, h_y)
03	184±8	0.18±0.02	0.1565	280	(h_x, h_y)
04	169±10	0.26±0.03	0.1978	276	(h_x, h_y)
05	166±11	026±0.03	0.1831	281	(h_x, h_y)
06	132±16	0.44±0.04	0.2163	307	(h_x, h_y)

Tabella 11.8: Alcuni parametri statistici relativi ai risultati condotti sul dataset di validazione per i modelli impiegati per l'analisi sui dati reali; i valori si riferiscono al fit lineare basato su minimi quadrati per la retta 'stima' vs. 'reale'; ρ^2-adj si riferisce al coefficiente di correlazione corretto per il numero di gradi di libertà.

- *da un punto di vista globale, sia con* h_c *che con* (h_x, h_y) *i risultati migliori si ottengono per il modello 04;*
- *i residui si distribuiscono in maniera* uniforme *sull'intero territorio di studio, senza l'evidente presenza di cluster;*
- *per tutti i modelli indagati, con la coppia di parametri* (h_x, h_y) *l'effetto di smoothing è meno evidente rispetto all'impiego del singolo parametro* h_c *— ma comunque presente;*
- *come nel caso di wk-NN, le covariate che sembrano portare informazioni utili al modello di previsione sono* tipo di costruzione, contatto, esposizione *e* utilizzo.

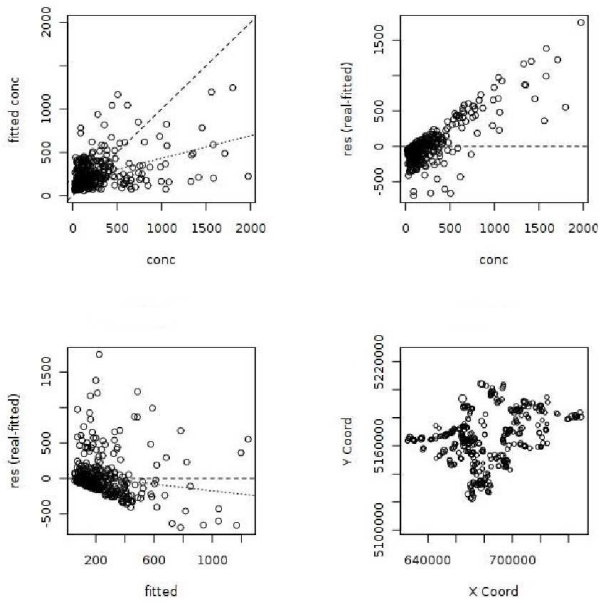


Figura 11.3: Esempio dei grafici utilizzati per valutare i modelli applicati ai dati reali [cfr. tab. 11.6] sulla base delle analisi condotte sul dataset di 'validation'; nello specifico, si tratta del modello 04 con parametri (h_x, h_y); il grafico in basso a destra riporta un post-plot dei residui nel quale la dimensione del punto è proporzionale al valore del residuo stesso; i parametri statistici relativi al fit lineare riportato sono raccolti nella tabella 11.8.

Nota 3: *Il modello 06, che comprende tutte le covariate disponibili e selezionate in fase di preparazione dei dati, appare come migliore se si limita l'analisi a quanto riportato in tabella 11.7 relativamente al test di accuratezza e alla tabella 11.8; tuttavia, questi risultati sono* falsati *dalla presenza di alcuni pesanti errori che hanno influenza non trascurabile sul fit lineare dei minimi quadrati — le performance sul test di generalizzazione non sono infatti molto buone rispetto agli altri modelli.*

11.5.1 Analisi variografica dei residui

Infine, per valutare la presenza di eventuali e significativi artefatti della procedura di stima, ho ritenuto utile calcolare i variogrammi sperimentali dei residui per i vari modelli implementati: dal punto di vista teorico, questi dovrebbero essere dei puri effetti nugget oscillanti attorno al valore di nugget del variogramma sperimentale — quest'ultimo, calcolato per il dataset di training, risulta piuttosto rumoroso e con una struttura non ben definita (alla luce delle precedenti analisi condotte, questo non stupisce più di tanto); il suo valore di nugget è pari a circa 100000.

Un esempio di tale analisi è riportato in figura 11.4 relativamente al modello 04 con parametri (h_x, h_y). I variogrammi ottenuti non mostrano particolari o evidenti differenze al variare del modello o del tipo di parametro h utilizzato, e in generale portano alle stesse conclusioni che si sono ottenute nel caso dell'approccio wk-NN descritto nel capitolo 10: struttura molto debole e assimilabile a un effetto nugget oscillante attorno al valore della varianza a priori dei residui, che risulta tuttavia inferiore rispetto al valore di nugget del variogramma sperimentale relativo

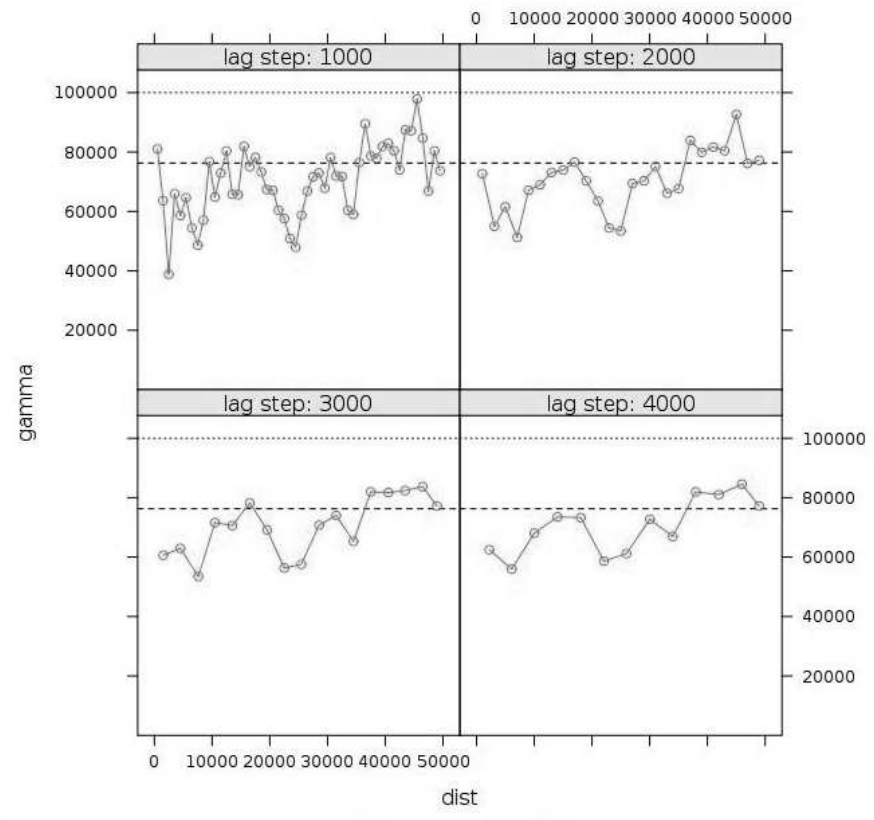


Figura 11.4: Esempio di variogramma dei residui per il modello 04 con parametri (h_x, h_y), al variare del lag-step; i valori di distanza sono riportati in metri; la linea tratteggiata rappresenta la varianza a priori relativa ai residui, quella punteggiata il valore di nugget relativo al variogramma sperimentale calcolato per il dataset di training (dataset operativo).

al dataset 'training' — anche in questo caso, ciò si può interpretare come una incapacità del modello di estrarre *tutta* l'informazione disponibile nel dataset.

11.6 Conclusioni

Sulla base dell'esperienza acquisita in ambito geostatistico relativamente al dataset operativo descritto nel paragrafo 1.3 a pagina 7, si crede che l'informazione collegata alla sola correlazione spaziale delle misure di radon indoor non sia sufficiente per una comprensione e modellizzazione soddisfacenti del fenomeno in esame. Per questo, si è rivolta l'attenzione a strumenti in grado di sfruttare, accanto a quella contenuta nella componente spaziale del fenomeno, anche l'informazione potenzialmente contenuta nelle variabili secondarie che accompagnano la singola misurazione di radon indoor, e in particolare quella relativa alla parte antropogenica del fenomeno stesso — quella cioé relativa alle caratteristiche dell'edificio sede della misura. In questo contesto, si è deciso di applicare gli strumenti offerti dall'approccio *General Regression Neural Network*, che attualmente costituisce lo stato dell'arte degli stimatori non parametrici.

Le analisi sono state condotte su vari modelli basati su un dataset operativo derivato da quello di riferimento descritto in dettaglio nel paragrafo §1.3; i vari modelli si differenziano per il numero e tipo di variabili secondarie coinvolte, consentendo così di valutare l'eventuale influenza del differente tipo di informazione fornita allo specifico modello. Inoltre, si è costruito un modello ad hoc basato su un dataset simulato in modo da poter controllare alcuni parametri coinvolti nella fase di stima dei parametri del modello.

Riassumendo i risultati ottenuti per i vari modelli implementati, si può concludere che:

- per i modelli che coinvolgono solo la parte spaziale (le sole coordinate come variabili secondarie), in tutti i casi le curve di cross-validation mostrano una struttura definita, indice del fatto che i dati in esame manifestano una qualche correlazione spaziale riconoscibile — a conferma di risultati già noti da precedenti studi di tipo geostatistico (come analisi variografiche);

- tra le variabili secondarie prese in esame, quelle che sembrano portare informazioni utili per la previsione del valore di concentrazione di radon indoor sono risultate essere *contatto*, *tipo di costruzione*, *esposizione* e *utilizzo*: interessante come queste variabili, oltre a essere ragionevoli da un punto di vista teorico, sia inoltre le stesse che sono emerse come significative in contesti differenti, quali Feature Selection [cfr. cap. 9], analisi variografiche su specifici subset [cfr. cap. 5] e analisi Weighted k-Nearest Neighbor [cfr. cap. 10]; i risultati ottenuti in questo contesto sembrano quindi confermare la significatività dell'influenza di queste variabili secondarie sul valore di concentrazione di radon misurato all'interno degli edifici;

- questo approccio sembra configurarsi come un buon *tool esplorativo* per indagare le proprietà dei dati di radon indoor e la loro relazione con le numerose variabili secondarie che accompagnano ogni singola misurazione; in particolare, si è rivelato utile per *identificare* le features che mostrano una qualche *influenza significativa* sul valore di concentrazione misurato negli edifici; a sostegno di questo, i risultati sono in linea con quanto ottenuto ricorrendo ad approcci differenti, come quelli affrontati e descritti nei capitoli 8 e 9;

- mettendo invece a confronto le stime ottenute con GRNN con un dataset reale di validazione, questo approccio *non* sembra essere uno strumento altrettanto valido per produrre delle mappe del fenomeno radon indoor — o quantomeno, globalmente non si ottengono risultati migliori rispetto a quanto si è ottenuto finora con gli approcci più convenzionali di tipo geostatistico[4].

[4] A tal proposito, si rimanda tuttavia a quanto discusso nel capitolo 12, a pagina 177, nel quale emerge come a livello *locale* questo tipo di approccio dia qualche risultato migliore rispetto a quanto ottenuto mediante Ordinary Kriging.

CAPITOLO 12

Previsioni a confronto: Machine Learnign vs. Geostatistica

L'idea che ha guidato questa parte del lavoro è stata quella di mettere a confronto i risultati ottenuti nell'ambito degli approcci di Machine Learning *testati rispetto a quanto si può ottenere invece in ambito geostatistico. Nel primo caso, i modelli sono costruiti ricorrendo anche all'informazione aggiuntiva contenuta in variabili secondarie, mentre nel secondo si prende in considerazione* unicamente *la componente spaziale del fenomeno radon indoor.*
Poiché c'è la convinzione che per una descrizione esaustiva ed efficacie del fenomeno la sola analisi spaziale non *sia sufficiente — soprattutto in relazione all'influenza non trascurabile delle caratteristiche dell'edificio sede della misura —, è sembrato quindi interessante verificare questa assunzione nella pratica, applicando appunto approcci differenti al medesimo dataset operativo costituito da misurazioni reali.*

La parte computazionale e di analisi statistica è stata svolta interamente in ambiente `R`{44} ricorrendo ai packages *geoR*{30}, *gstat*{41} e *lattice*{45}, oltre a routine create ad hoc; analisi esplorative e spaziali sulle stime ottenute sono state condotte mediante lo strumento GIS `QGis`{1}.

12.1 Descrizione del dataset utilizzato

Il dataset operativo su cui l'analisi geostatistica descritta in seguito si basa è lo stesso cui ho fatto ricorso nelle analisi di Machine Learning (ML) descritte nei capitoli 10 e 11, con l'ovvio scopo di operare i successivi confronti su modelli basati su misurazioni comuni. In particolare, si è fatto riferimento al dataset 'training' descritto nel dettaglio nel paragrafo §10.2 a pagina 149, costituito da **820** misurazioni di concentrazione di attività di radon indoor georeferenziate.

Per quanto riguarda invece le previsioni utilizzate per i confronti, si è fatto ricorso al dataset 'validation', costituito da **409** misurazioni georeferenziate, con una distribuzione spaziale sul

territorio di studio analoga a quella del dataset operativo: modelli e approcci sono stati così sottoposti a confronti basati non sulla produzione di mappe, ma su stime condotte su singole localizzazioni.

12.2 Studio di tipo geostatistico

Sulla base delle misurazioni georeferenziate che costituiscono il dataset operativo, ho condotto un'analisi geostatistica convenzionale, che prevede essenzialmente i) uno studio esplorativo variografico volto alla ricerca della presenza di una struttura spaziale nei dati (che in base ai lavori precedentemente svolti, dovrebbe essere presente!), una successiva ii) modellizzazione del variogramma sperimentale e infine il suo utilizzo per una iii) stima del valore di concentrazione in localizzazioni non campionate mediante un algoritmo di kriging — nel caso specifico, Ordinary Kriging (OK).

12.2.1 Analisi variografica

Mediante una opportuna routine creata allo scopo, ho condotto delle analisi esplorative determinando i variogrammi sperimentali al variare sia della risoluzione spaziale (lag-step) che della massima distanza tra le coppie; come ci si poteva aspettare, la leggibilità della struttura spaziale non è risultata delle migliori, e caratterizzata dalla presenza di un elevato effetto nugget. Per questo, ho deciso di ricorrere a una misura della correlazione spaziale più *robusta*, ovvero quella cui in letteratura si fa riferimento come variogramma di Cressie, dato da:

$$\gamma(\mathbf{h}) = \frac{\left[\frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}|Z(x_{i+\mathbf{h}}) - Z(x_i)|^{\frac{1}{2}}\right]^4}{0.914 + \frac{0.988}{N(\mathbf{h})}} \tag{12.1}$$

dove $N(\mathbf{h})$ rappresenta il numero di coppie per il lag $\mathbf{h}$ e $Z(\mathbf{h})$ la variabile di interesse (in questo caso, concentrazione di radon indoor).

Da analisi di questo tipo, di cui riporto un esempio nella figura 12.1, e tenendo sempre sotto controllo che il numero minimo di coppie per ogni lag fosse sufficientemente elevato, ho deciso di considerare una massima distanza pari a 35 km e un lag-step pari a 3.5 km.

Una volta identificato il variogramma sperimentale di riferimento, questo è stato fittato mediante una procedura interattiva e sfruttando il variogramma tradizionale per la stima dei parametri di varianza σ e nugget τ, mentre il variogramma di Cressie per la stima del range. Il modello di variogramma ottenuto in questo modo è un modello sferico con range pari a 15 km ca., dato dall'equazione:

$$\gamma(h) = 64997 + 46942 \cdot \text{sph}\left(\frac{h}{14777}\right) \tag{12.2}$$

Data la rumorosità del variogramma e il numero limitato di campionamenti disponibili, non ho ritenuto utile estendere e affinare l'analisi variografica con uno studio direzionale, limitandomi quindi a quella isotropa.

12.2.2 Stima mediante Ordinary Kriging

Una volta in possesso del modello di variogramma dato dall'equazione (12.2), ho condotto delle stime nelle localizzazioni del dataset 'validation' ricorrendo all'algoritmo del Ordinary Kriging [cfr. quanto discusso nel paragrafo §B.4.3, a pagina 236].

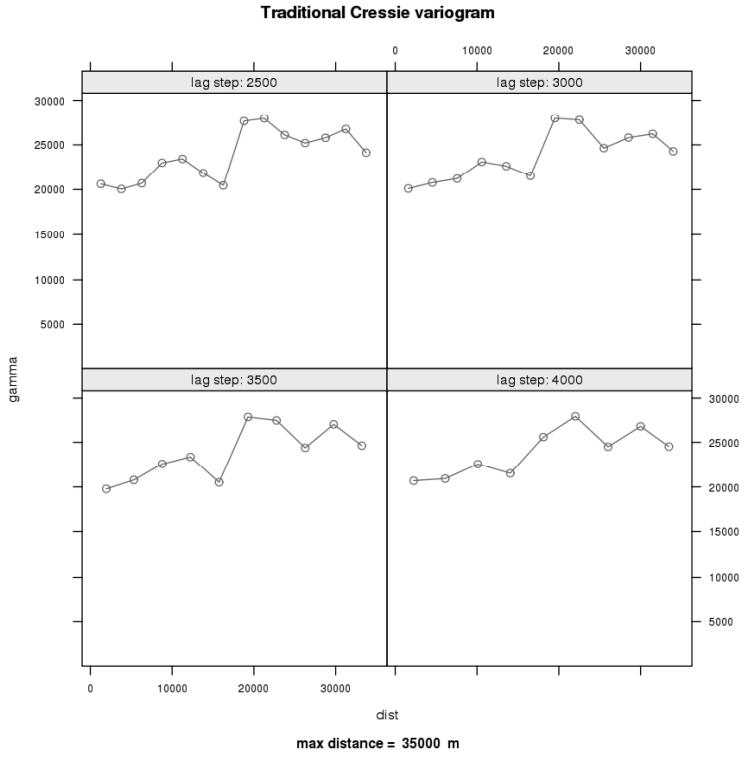


Figura 12.1: Esempio di variogrammi sperimentali utilizzati in fase esplorativa per la ricerca di una struttura spaziale riconoscibile, al variare in questo caso della risoluzione spaziale (lag-step); si tenga presente che i variogrammi riportati non sono quelli tradizionali, ma quelli definiti *robusti* basati sull'eq. (12.1).

Ho costruito quattro differenti modelli, al variare del massimo numero di punti da considerare nel vicinaggio di stima; in tutti i casi, se nel vicinaggio non sono presenti almeno tre punti, la localizzazione non viene stimata; i parametri impiegati sono riassunti nella tabella 12.1.

Il modello *OK 10* è stato costruito sulla base dei risultati ottenuti nel precedente lavoro di tesi {42}, mentre i modelli *OK 30* e *OK 60* per considerare un numero di punti da coinvolgere nella stima in linea con quanto ottenuto in relazione al parametro k dell'approccio wk-NN [cfr. tab. 10.4 a pag. 155, modelli 01 e 04].

12.3 Confronto su dataset di validazione

Le stime ottenute sulla base dei tre approcci presi in considerazione, che riassumendo sono:

- Ordinary Kriging — approccio geostatistico, che considera *solo* l'informazione contenuta nella distribuzione spaziale dei campionamenti sfruttando il modello di variogramma dato dall'eq. (12.2)
- Weighted k-Nearest Neighbor (wk-NN) — approccio Machine Learning, che oltre alla componente spaziale considera anche le informazioni contenute nelle covariate che accompagnano ogni singola misura [cfr. cap. 10]

id modello	N_{min}	N_{max}	*distanza massima*
OK all	3	∞	∞
OK 10	3	10	∞
OK 30	3	30	∞
OK 60	3	60	∞

Tabella 12.1: Descrizione dei modelli impiegati nella fase di stima dei valori di concentrazione di radon indoor mediante Ordinary Kriging; per maggiori dettagli, si faccia riferimento al testo.

id modello	*intercetta*	*pendenza*	ρ^2-*adj*	*RMSE*
OK all	182±8	0.20±0.02	0.181	276
OK 10	172±9	0.24±0.02	0.192	276
OK 30	177±9	0.22±0.02	0.187	275
OK 60	178±9	0.22±0.02	0.187	275

Tabella 12.2: Alcuni parametri statistici relativi ai risultati condotti sul dataset di validazione per i modelli basati su Ordinary Kriging impiegati per l'analisi sui dati reali; i valori si riferiscono al fit lineare basato su minimi quadrati per la retta 'stima' vs. 'reale'; ρ^2-adj si riferisce al coefficiente di correlazione corretto per il numero di gradi di libertà.

- General Regression Neural Network (GRNN) — approccio Machine Learning, che oltre alla componente spaziale considera anche le informazioni contenute nelle covariate che accompagnano ogni singola misura [cfr. cap. 11]

sono state messe a confronto sia da un punto di vista per così dire *globale*, ovvero analizzando i residui delle stime stesse, sia da un punto di vista più *specifico*, ovvero analizzando alcune localizzazioni di particolare interesse.

Sottolineo come sia stato possibile condurre questi tipi di analisi sfruttando il dataset 'validation', che ha le utili caratteristiche di:

i) avere una serie di campionamenti con una distribuzione spaziale sul territorio dell'Alto Adige che riproduce quella del dataset operativo — in questo modo, è possibile mettere a confronto i vari approcci in tutte le differenti situazioni che caratterizzano il territorio in esame;

ii) avere, per ogni localizzazione di stima, anche il valore reale di concentrazione misurato — così da poter confrontare non solo gli approcci tra di loro, ma anche con la situazione reale.

12.3.1 Analisi dei residui

Analisi e confronti sono stati eseguiti ricorrendo a grafici come quelli riportati a titolo d'esempio per il modello *OK 10* in figura 12.2 e ai parametri statistici relativi al fit lineare basato su minimi quadrati per gli scatter plot ' valori fittati ' *vs.* ' valori reali ', riportati nella tabella 12.2. Questo tipo di analisi è stato condotto in linea con quanto fatto per gli approcci di ML, descritti nei paragrafi §10.4 e §11.5.

Tenendo conto del fatto che i risultati migliori per gli approcci di ML, riportati nel dettaglio nella tabella 10.5 a pagina 157 per wk-NN e nella tabella 11.8 a pagina 172 per GRNN, sono

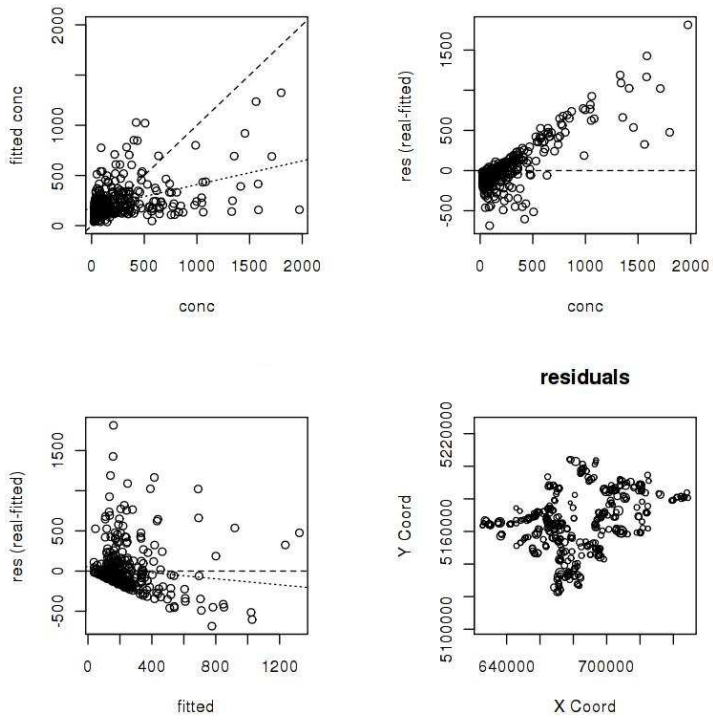


Figura 12.2: Esempio di grafici utilizzati per lo studio dei residui, relativi al modello *OK 10* [cfr. tab. 12.1]; le rette puntinate si riferiscono al fit lineare condotto mediante minimi quadrati, mentre il grafico in basso a destra riporta un post-plot dei residui in cui ogni singolo punto è proporzionale al valore del residuo stesso.

pari rispettivamente a 0.16 ± 0.02 e 0.26 ± 0.03 per il valore della pendenza e pari a 0.196 e 0.198 per il valore di ρ^2-adj, si può concludere che

> *da un punto di vista* globale, *ricorrendo a un dataset indipendente di validazione, e limitando il confronto a rappresentazioni visuali dei residui e semplici analisi statistiche sugli stessi, gli approcci di ML e quello di tipo geostatistico danno risultati paragonabili tra loro, benché i modelli di ML abbiano accesso a informazioni aggiuntive (relative alla parte* antropogenica*) rispetto al modello di OK, che prende in considerazione solo la* componente spaziale *del fenomeno radon indoor.*

12.3.2 Analisi di specifiche localizzazioni

Accanto alle analisi per così dire globali appena descritte, mi è sembrato opportuno indagare più nello specifico il comportamento dei differenti approcci controllando puntualmente alcune particolari localizzazioni; l'idea è stata quella di poter notare alcune differenze in situazioni che si potrebbero definire *difficili* per i modelli in esame.

La scelta dei punti da destinare a questo tipo di analisi è stata fatta sulla base dei punti più distanti dalla bisettrice nei grafici 'valore fittato' *vs.* 'valore reale', come quelli riportati

ad esempio in alto a sinistra nella figura 12.2 — ovvero, localizzazioni per le quali gli errori, sia in sovrastima che in sottostima, sono risultati maggiori. Interessante infine notare come tali punti siano comuni a tutti gli approcci sotto esame, indice del fatto che le situazioni per così dire anomale o difficili (piuttosto comuni per la banca dati da cui le misure sono state estratte) lo sono *indipendentemente* dall'approccio scelto, configurandosi quindi come una caratteristica intrinseca del fenomeno radon indoor.

Mediante strumenti esplorativi di tipo GIS, si sono analizzate caso per caso le situazioni appena descritte, valutando anche le caratteristiche specifiche dei campionamenti che costituiscono il vicinaggio di stima.

In tutti i casi presi in considerazione, sia per quanto riguarda le sovrastime che le sottostime:

- *in generale, gli approcci di ML portano a stime che si avvicinano di più al valore realmente misurato rispetto all'approccio dell'OK;*
- *tra gli approcci di ML indagati, wk-NN risulta essere il migliore;*
- *i casi anomali indagati si caratterizzano per intorni con edifici tipicamente in contatto con il terreno e costruiti con sassi: queste sono peculiarità che, in base ad analisi precedenti come quelle descritte nei capitoli 5, 8 e 9, hanno manifestato una influenza non trascurabile sul valore di concentrazione misurato;*
- *"istruendo" il modello con informazioni aggiuntive relative alla parte* antropogenica *del fenomeno, si ottengono delle stime che più si avvicinano alla situazione reale, a sostegno dell'ipotesi che la sola componente spaziale non sia sufficiente per una modellizzazione esaustiva del fenomeno radon indoor — questi risultati possono essere mascherati o quantomeno poco evidenti in analisi di tipo* globale, *ma risultano invece significativi se si focalizza l'attenzione sulle situazioni più complesse e anomale.*

12.4 Conclusioni

Ricorrendo a un dataset operativo comune ricavato da quello generale composto di misure di concentrazione di attività di radon indoor [cfr. §1.3], è sembrato interessante mettere a confronto i risultati che si possono ottenere ricorrendo a due differenti approcci per la stima in localizzazioni non campionate: da un lato, un modello di tipo geostatistico basato su kriging ordinario (OK), dall'altro modelli basati su tecniche di Machine Learning (ML)[1]. Nel primo caso, quindi, un modello volto alla trattazione specifica e raffinata (attraverso la modellizzazione della struttura variografica) della parte spaziale del fenomeno in esame; nel secondo, modelli in grado di aggiungere, accanto alla componente spaziale (cui non viene però dedicata una trattazione specifica), anche informazioni aggiuntive contenute in variabili secondarie che caratterizzano il contesto nel quale la misura è stata condotta.

Dalle analisi statistiche e spaziali condotte mettendo a confronto le stime ottenute su localizzazioni per le quali sono disponibili anche i reali valori di concentrazione, si può concludere che:

[1] Nello specifico, si tratta dell'approccio Weighted k-Nearest Neighbor trattato nel capitolo 10 e dell'approccio General Regression Neural Network trattato nel capitolo 11.

- limitando il confronto a rappresentazioni visuali e spaziali dei residui e a semplici analisi statistiche sugli stessi, *globalmente* gli approcci di ML e quello di tipo geostatistico (OK) danno risultati paragonabili tra loro, benché i modelli di ML abbiano accesso a informazioni aggiuntive (relative alla parte antropogenica) rispetto al modello OK;
- focalizzando invece l'attenzione su specifiche localizzazioni che hanno mostrato elevati valori del modulo del residuo (si sono quindi prese in considerazioni sia situazioni si sovrastima che di sottostima), in tutti i casi gli approcci ML portano a stime che rispecchiano meglio il valore realmente misurato, rispetto all'approccio OK: il fornire al modello informazioni aggiuntive sembra quindi manifestare la sua efficacia in situazioni "anomale" o quantomeno di più complessa modellizzazione.

Risulta infine interessante notare come i casi che sono stati oggetto di analisi specifiche siano tutti caratterizzati da intorni costituiti principalmente da edifici in contatto con il terreno e costruiti con sassi: queste sono infatti caratteristiche legate a variabili che, in base ad analisi condotte in ambiti differenti [cfr. capp. 5, 8, 9], sono risultate come le più efficaci in relazione alla loro capacità predittiva sul valore di concentrazione misurato.

CAPITOLO 13

L'approccio Support Vector Machine

L'idea di base che ha ispirato il lavoro presentato in questo capitolo è stata quella di applicare ai dati di radon indoor l'approccio del Support Vector Machine*, che nell'ambito dei classificatori ne costituisce attualmente lo stato dell'arte. Abbandonando quindi il contesto della regressione, si otterranno risultati per due classi di concentrazione, ovvero, fissato un determinato valore di soglia, si otterranno, per le localizzazioni non campionate, informazioni relative all'essere sopra o sotto la soglia — e non il valore numerico di concentrazione.*
Il poter introdurre nel modello informazioni aggiuntive disponibili in base alle features (attributi) che accompagnano la singola misura e il fatto di limitare l'analisi a due sole classi — riducendo almeno parzialmente eventuali problemi legati alla presenza di outliers e alla rumorosità dei dati stessi — si crede possa portare a risultati affidabili e solidi.

La parte computazionale è stata svolta in ambiente R{44}, ricorrendo al package *e1071*{17} e ad alcune routine create ad hoc.

13.1 Brevi richiami teorici

Nel contesto della teoria statistica dell'apprendimento, col termine APPRENDERE si fa riferimento al processo di stima di una certa funzione $y = f(\mathbf{x})$, con $\mathbf{x} \in \mathbb{R}^N$ e, in relazione al tipo di problema che viene affrontato, $y \in \mathbb{R}$ nel caso della *regressione*, $y \in [1, 2, \ldots, M]$ nel caso di una *classificazione* con M-classi, e $y \in [-1, 1]$ nel caso di una *classificazione binaria*. La stima sarà condotta sfruttando unicamente l'insieme $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_L, y_L)\}$ delle L coppie di *esempi* sperimentali che descrivono la mappatura operata dalla funzione non nota $y = f(\mathbf{x})$.

Nella pratica, un algoritmo di Machine Learning (ML) sarà chiamato a scegliere da una dato insieme di funzioni $F = \{f(\mathbf{x}, \boldsymbol{\alpha}), \boldsymbol{\alpha} \in \Lambda\}$ quella che — in base a un criterio che andrà definito — *meglio* approssima la dipendenza che si vuole modellizzare e che non è nota. Λ rappresenta un *ML*

arbitrario insieme di parametri scelti in partenza all'interno dei quali trovare quelli "ottimi" che andranno successivamente a caratterizzare l'algoritmo di ML: si tratta quindi di un problema di ottimizzazione nello spazio dei parametri $\boldsymbol{\alpha}$.

Il processo di apprendimento è definito come la stima della funzione $f(\mathbf{x})$ dall'insieme delle funzioni F definito a priori, che fornisce il valore minimo della cosiddetta *empirical risk function*, data da:

$$R_{emp} = \frac{1}{L}\sum_{i=1}^{L} Q\left(y_i, f(\mathbf{x}_i, \boldsymbol{\alpha})\right) \tag{13.1}$$

dove $Q\left(y_i, f(\mathbf{x}_i, \boldsymbol{\alpha})\right)$ è la cosiddetta loss-function, ovvero una misura della discrepanza tra la stima e il reale valore di y dato dalla funzione f per il vettore $\mathbf{x}$. In questo modo, una funzione che fornisca il minimo per l'equazione (13.1) sarà scelta come una funzione ottimale per la classificazione (o regressione); si sottolinea come questo approccio si basi esclusivamente sulla performance dell'algoritmo in relazione a un numero finito di esempi.

Tuttavia, nella pratica spesso i dati cui si ha accesso sono numericamente pochi, e la funzione determinata nel modo appena descritto potrebbe riprodurre dipendenze nei dati sperimentali che non sono però caratteristiche dell'intera popolazione da cui sono stati estratti. questo problema risulta ancora più evidente e importante nei casi in cui la dimensionalità del dataset operativo risulti elevata. La capacità del modello di descrivere la reale dipendenza funzionale a partire da una insieme finito di esempi viene chiamata in letteratura CAPACITÀ DI GENERALIZZAZIONE dell'algoritmo di ML. Tale capacità può essere controllata operando delle opportune scelte sull'insieme delle funzioni F e sulle sue dimensioni (sempre in termini delle funzioni che lo compongono).

Disporre di un insieme di funzioni in grado di operare un gran numero di separazioni (si consideri il caso della classificazione), darà origine a un basso valore per il rischio empirico, ma al contempo non sarà verosimilmente in grado di generalizzare molto — fenomeno ben noto col nome di *overfitting*. Dall'altro lato, invece, disporre di un insieme F con poche funzioni darà risultati migliori in relazione alla capacità di generalizzazione, ma le funzioni accessibili potrebbero essere così poche da non consentire una adeguata modellizzazione dei dati sperimentali.

In altre parole, quello che importa non è tanto il numero di parametri $\boldsymbol{\alpha}$ che caratterizza la funzione, quanto piuttosto la cosiddetta VC-DIMENSION, che rappresenta il numero massimo di punti distinti che in una prefissata configurazione possono essere separati dalle funzioni $f(\mathbf{x}, \boldsymbol{\alpha}) \in F$. Ad esempio, considerando le funzioni di tipo lineare[1], nel caso bi-dimensionale e per una classificazione binaria la VC-dimension è pari a 3: sul piano, infatti, è possibile costruire una configurazione di 4 punti che non possono essere separati da alcuna funzione di tipo lineare. In generale, per funzioni di tipo lineare, in $\mathbb{R}^N$ la VC-dimension è pari a $N+1$. Se la VC-dimension è troppo alta, vuol dire che si troverà verosimilmente sempre una funzione in grado di descrivere la dipendenza funzionale tra y e $\mathbf{x}$, ma questo porterà in molti casi a fenomeni di overfitting, che andrebbero evitati per ovvi motivi.

In quest'ottica, la strategia migliore e più "furba" per costruire un valido algoritmo di ML con una buona capacità di generalizzazione è quella di fittare i dati (minimizzando il rischio empirico) e al contempo mantenere la complessità del modello bassa (controllando la VC-dimension e quindi le caratteristiche di generalizzazione). Uno dei risultati fondamentali di questa teoria è che esiste un limite superiore per il *rischio strutturale* R, ovvero per la probabilità di sbagliare la previsione su punti che non sono noti. Sia $1-\delta$ la probabilità che la stima sia corretta, allora vale che:

$$R(\boldsymbol{\alpha}) \leq R_{emp}(\boldsymbol{\alpha}) + \sqrt{\frac{h\left[\log\left(\frac{2L}{h}\right)+1\right] - \log\left(\frac{\delta}{4}\right)}{L}} \tag{13.2}$$

[1] Sarà chiaro tra breve che imporre questo vincolo non è in realtà una limitazione per l'approccio Support Vector Machine, in quanto è possibile sfruttare quello che viene chiamato *kernel trick* [cfr. §13.1.2].

dove R_{emp} rappresenta il rischio empirico dato dall'eq. (13.1), h la VC-dimension e L il numero di esempi disponibili; risulta evidente dalla (13.2) che le cose vanno nella giusta direzione se h è piccolo e L è grande.

13.1.1 Classificazione mediante Support Vector Machine

L'approccio più semplice al problema della classificazione è quello di distinguere tra due classi con una superficie di separazione di tipo lineare; questa sarà quindi una retta nel caso bi-dimensionale, un piano nel caso tri-dimensionale o un iper-piano nel caso di dimensioni maggiori. Nel caso (piuttosto ideale!) che la separazione sia sempre possibile senza incappare in misclassificazioni, il dataset viene detto linearmente separabile, e l'algoritmo in grado di trovare il (iper-)piano ottimale per tale separazione lineare è noto come *large margin classifier*[2].

Poiché nella pratica sono pochi i dataset che risultano linearmente separabili, questa assunzione verrà abbandonata, per arrivare a un classificatore lineare in grado di prendere in considerazione in maniera controllata e opportuna errori di classificazione, noto come *soft margin classifier*. Infine, anche questo ulteriore passo in avanti sarà superato attraverso l'introduzione del cosiddetto *kernel trick*, che darà vita a un classificatore *non*-lineare noto come Support Vector Machine. *SMV*

Si consideri il seguente insieme di funzioni lineari di base:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{13.3}$$

dove $\mathbf{x}$ è un vettore in $\mathbb{R}^N$ (il cosiddetto *input space*), b uno scalare e $\mathbf{w}$ un vettore di parametri in $\mathbb{R}^N$ che deve essere ottimizzato. Nel caso della classificazione binaria, verrà considerata la funzione segno del secondo membro della (13.3) come output del classificatore.

Come ricordato in precedenza, la VC-dimension di tale insieme di funzioni sarà $N + 1$: il problema, in questo caso, è che tale valore è fissato e non può quindi venir controllato dalla scelta di alcun parametro che caratterizzi l'algoritmo che si vuole ottimizzare, e conseguentemente non può essere applicata l'equazione (13.2) relativa al rischio strutturale. Questa inconsistenza può essere superata introducendo il concetto di *large margin classifier*.

Large margin classifier

SVM si possono considerare in origine come una sorta di classificatori lineari. Si prenda in esame la seguente funzione di decisione, in grado di discriminare tra due classi distinte, e che dipende da $\mathbf{w}$ e b:

$$y = \begin{cases} +1 & \text{se} \quad \mathbf{w} \cdot \mathbf{x} - b \geq 1 \\ -1 & \text{se} \quad \mathbf{w} \cdot \mathbf{x} - b \leq 1 \end{cases} \tag{13.4}$$

A differenza dell'equazione (13.3), ora la decisione è presa in relazione alla posizione (distanza) del campione rispetto a un *margine* che corre lungo l'iper-piano definito dal vettore $\mathbf{w}$.

Questa proprietà risulta di particolare rilevanza in questo contesto, in quanto un risultato della teoria afferma che se l'insieme degli esempi (vettori in $\mathbb{R}^N$) che costituiscono il dataset sperimentale appartengono a una iper-sfera di raggio R, la VC-dimension h dell'insieme di funzioni

[2] Si consideri ad esempio il caso bi-dimensionale, per un problema linearmente separabile; si costruisca il classificatore binario ricorrendo alla funzione segno, in modo da trovare la retta (che per ipotesi esiste) in grado di distinguere tra i due gruppi. Si potrebbe pensare di ricorrere, per risolvere il problema della minimizzazione, al metodo dei minimi quadrati, ma il risultato non sarebbe il migliore! Quello migliore sarà quello in grado di *massimizzare* il cosiddetto *margine*, ovvero la zona che separa i due gruppi: se si dovesse tracciare la linea di separazione con un pennello, si dovrebbe optare per quello più largo possibile.

definito dalla (13.4) è sottoposto al seguente vincolo:

$$h \leq \min\left(R^2\|\mathbf{w}\|^2, N\right) + 1 \tag{13.5}$$

e se si ottiene un valore $h \leq N + 1$, allora si è fatto qualcosa di meglio rispetto a un semplice classificatore lineare. In particolare, risulta a questo punto chiaro che la VC-dimension del classificatore dato dalla (13.4) può essere controllata, minimizzando il termine $R^2\|\mathbf{w}\|^2$ per minimizzare il limite superiore di h e conseguentemente il limite superiore del rischio strutturale dato dall'equazione (13.2). Questa è la prima idea chiave alla base del classificatore SVM.

Sfruttando il fatto che, per la classificazione binaria, $y_i = \pm 1$, la (13.4) può essere riscritta in modo compatto come

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \tag{13.6}$$

A questo punto, l'idea è quella di massimizzare il margine ρ, dato dalla distanza tra l'iper-piano $f(\mathbf{x}) = 1$ e l'iper-piano $f(\mathbf{x}) = -1$. È facile ricavare che:

$$\rho = \frac{2}{\|\mathbf{w}\|} \tag{13.7}$$

Così, il classificatore SVM non solo mira a distinguere tra due classi, ma al contempo mira anche a massimizzare il margine tra queste due classi minimizzando il termine $\|\mathbf{w}\|$. L'idea intuitiva è che un iper-piano con un margine ampio dovrebbe risultare più resistente al rumore presente nei dati e possedere capacità di generalizzazione migliori rispetto a un iper-piano con un margine inferiore (o nullo).

Quello che si deve affrontare è quindi un problema di minimizzazione vincolato, risolvibile mediante moltiplicatori di Lagrange. Tralasciando i dettagli matematici, quello che si ottiene è che la funzione di decisione può essere scritta in questi termini:

$$f(\mathbf{x}) = \sum_{i=1}^{L} y_i \alpha_i \mathbf{x}_i \cdot \mathbf{x} + b \tag{13.8}$$

Se $\alpha_i = 0$, allora $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1$, mentre se $\alpha_i > 0$, allora $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1$. Gli esempi per i quali $\alpha_i > 0$ andranno a cadere esattamente sull'iper-piano $f(\mathbf{x}, \mathbf{w}, b) = \pm 1$ della superficie di decisione: questi sono chiamati Support Vector.

Essi rivestono un ruolo fondamentale, in quanto se vengono "spostati", anche la striscia subirà delle variazioni, ma se gli altri esempi venissero completamente rimossi dal dataset empirico, quest'operazione non avrebbe invece alcuna conseguenza sulla striscia stessa — cosa che ad esempio non avviene nel caso della minimizzazione con minimi quadrati. Un vantaggio non trascurabile di questo risultato è che il numero e la posizione dei support vector possono essere impiegati come criteri per la ricerca e la scelta dei parametri ottimali per il classificatore SVM.

Soft margin classifier

Tutto quanto descritto fino ad ora si applica al caso in cui il dataset empirico sia linearmente separabile, ma spesso nella pratica questo non è il caso. Le tecniche possono allora venir estese anche a queste situazioni, introducendo delle cosiddette *slack variables* $\xi_i > 0$ e riscrivendo la (13.6) in questo modo:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \tag{13.9}$$

Ancora, come nel caso precedente, ci si trova di fronte a un problema di minimizzazione vincolato, risolvibile mediante moltiplicatori di Lagrange. Il funzionale da minimizzare sarà:

$$\tau(\mathbf{w}, \xi) = \frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^{L} \xi_i \tag{13.10}$$

Il primo termine della (13.10) corrisponde a minimizzare la VC-dimension, mentre il secondo a minimizzare il numero di misclassificazioni consentite. Il parametro $C \in \mathbb{R}^+$ costituisce una sorta di peso tra i due criteri che riflette in parte quanta "fiducia" l'operatore abbia sulla bontà e affidabilità dei dati che costituiscono il dataset empirico.

13.1.2 Kernel trick

Un importante risultato della teoria, noto come teorema di Mercer, afferma che per qualsiasi funzione a kernel[3] $K(\mathbf{x}, \mathbf{x}')$ che soddisfi le condizioni del teorema stesso esiste uno spazio, detto *feature space*, sul quale la funzione $K(\mathbf{x}, \mathbf{x}')$ agisce come un prodotto scalare.

Il vantaggio fondamentale di questo teorema è che il problema della classificazione può essere trasferito dallo spazio del dataset empirico in un nuovo spazio, appunto lo spazio delle feature, sul quale è sufficiente saper calcolare un prodotto scalare; questa trasformazione può anche non essere di tipo lineare, e conseguentemente un problema che nello spazio originario non era linearmente risolvibile, lo può diventare nel nuovo spazio dopo una opportuna trasformazione e venir risolto con tecniche lineari.

Detto in altri termini, data una funzione a kernel e un algoritmo formulato unicamente in termini di prodotti scalari tra le variabili di input $\mathbf{x}$ — come nel caso dell'equazione (13.8) —, si può ottenere facilmente una forma *non*-lineare dell'algoritmo stesso sostituendo il prodotto scalare con la funzione a kernel:

$$\mathbf{x} \cdot \mathbf{x}' \mapsto K(\mathbf{x}, \mathbf{x}')$$

Accanto all'ovvio vantaggio appena descritto, vale la pena di ricordare altri aspetti estremamente positivi e interessanti legati all'introduzione delle funzioni a kernel:

- la possibilità di definire kernel su qualsiasi tipo di oggetto e applicare quindi la regressione lineare (oggetti come stringhe, immagini, oggetti tridimensionali, ...);
- la possibilità di definire in maniera esplicita lo spazio delle feature su cui si intende lavorare per la separazione/classificazione;
- la possibilità di introdurre, in maniera esplicita o implicita, conoscenze a priori sui dati nella definizione del kernel stesso (in questo contesto, la funzione a kernel può essere vista come una sorta di funzione di similarità).

> ***Nota 1:*** *Va da sé che questi vantaggi si possono anche rivelare come "trappole", qualora le funzioni a kernel non vengano scelte nel modo opportuno, oppure se ne faccia un uso non consapevole. Per questo, si sono sviluppati e diffusi alcuni kernel per così dire convenzionali, cui è buona norma fare riferimento nei primi approcci a questo tipo di algoritmi.*

13.1.3 Output di tipo probabilistico

Concludo l'introduzione teorica con un risultato che è sembrato interessante, anche perché implementato nel software utilizzato per questa analisi. La filosofia del SVM non prevede, in origine,

[3] Una definizione di tali funzioni e delle loro proprietà può essere trovata nel paragrafo §10.1.2, a pagina 147.

alcuna interpretazione di tipo probabilistico dei risultati ottenuti; tuttavia, una interpretazione di questo tipo può risultare utile in molte applicazioni pratiche.

È possibile ottenere output di questo tipo anche per SVM, ricorrendo a un post-processing legato al valore numerico della funzione di decisione utilizzata per la classificazione; infatti, questo sarà:

$$\begin{cases} -1 < f(\mathbf{x}) < +1 & \text{all'interno del margine} \\ f(\mathbf{x}) = \pm 1 & \text{ai bordi del margine — support vectors} \\ |f(\mathbf{x})| > 1 & \text{per i punti correttamente classificati} \end{cases}$$

In questo modo, il valore della funzione di decisione può essere utilizzato come una sorta di indicatore della classe di appartenenza.

Al fine di ottenere una probabilità di appartenere o meno a una data classe, è pratica comune riscalare il valore della funzione di decisione in modo che $f(\mathbf{x}) \in [0, 1]$ usando una trasformazione sigmoidale del tipo:

$$P(y = 1|f(\mathbf{x})) = \frac{1}{1 + \exp(Af(\mathbf{x}) + B)} \tag{13.11}$$

Le costanti A e B che compaiono nella (13.11) vanno determinate utilizzando i dati empirici disponibili e un opportuno criterio. Evidenze empiriche riportate in letteratura mostrano che questo tipo di interpretazione risulta appropriata nei casi di SVM lineari con bassi valori del parametro C [cfr. eq. (13.10)], mentre richiede maggior attenzione nelle situazioni di SVM non lineari.

13.2 Descrizione del dataset utilizzato

Per uniformità con gli approcci di tipo Machine Learning cui ho fatto ricorso nei capitoli 10 e 11, il dataset operativo utilizzato anche in questa analisi è lo stesso descritto nel dettaglio nel paragrafo §10.2, a pagina 149.

Il dataset di 'training' è costituito da **1229** valori di concentrazione di attività di radon indoor georeferenziati riferiti al semestre invernale (nessuna correzione per convertirli a medie annuali) e al piano zero; ogni misura è accompagnata da una ricca serie di covariate che descrivono le caratteristiche dell'edificio sede della misura (parte antropogenica del fenomeno) e alcune caratteristiche del terreno circostante. Il dataset di 'validation' si compone invece di **409** cases, la cui distribuzione spaziale rispecchia quella del dataset operativo di 'training'.

> ***Problema:*** *Rispetto a quanto descritto nel paragrafo §10.2, in questo caso sono stato costretto a unire il dataset di 'training' (820 cases) con quello di 'testing' (409 cases) causa problemi di natura computazionale: è infatti risultato impossibile ricorrere a un dataset indipendente di 'testing' per la stima dei parametri del modello. Per guadagnare quindi in numerosità e non perdere informazioni utili, ho deciso di* unire *i due dataset in uno solo, quello operativo, costituito quindi da* $820 + 409 = 1229$ *cases.*

Volendo focalizzare l'attenzione sul problema della *classificazione*, i valori di concentrazione sono stati preventivamente convertiti in due classi, fissando due differenti valori di soglia, pari rispettivamente a 200 e 400 $\text{Bq}\cdot\text{m}^{-3}$. Tali valori sono stati scelti in relazione ai limiti previsti dalla raccomandazione europea 90/143/EURATOM emanata il 21 febbraio 1990 per la tutela della popolazione contro l'esposizione al radon negli ambienti chiusi (radon indoor). Questa raccomandazione consiglia, per gli edifici residenziali, una soglia di intervento pari a 400 $\text{Bq}\cdot\text{m}^{-3}$per gli edifici esistenti, e pari a 200 $\text{Bq}\cdot\text{m}^{-3}$per quelli ancora in fase di progetto; nelle situazioni di superamento di tali valori, si raccomanda di adottare *provvedimenti semplici ma efficaci volti a ridurre il livello di radon.*

13.3 Modelli per la sola parte spaziale

Inizialmente, ho costruito un modello i cui dati in ingresso fossero costituiti unicamente dalle *coordinate* del punto di misura, sia per testare nella pratica l'algoritmo di classificazione, sia per avere un modello semplice di riferimento — in linea con quanto fatto nelle altre analisi di tipo ML.

Per questa e per tutte le analisi descritte in seguito, ho fatto riferimento a quanto suggerito da Hsu *et al.* {25}, autori della libreria `libsvm` in C++, vincitrice di numerosi riconoscimenti a livello internazionale, e di cui il package *e1071* costituisce una interfaccia per l'ambiente statistico `R`{44}.

13.3.1 Descrizione dei parametri operativi

Un opportuno *scaling dei dati* in fase di pre-processing risulta in molti casi determinante in relazione alla qualità dei risultati che si andranno a ottenere, soprattutto nelle situazioni per le quali, come quella che caratterizza il dataset operativo, i range numerici delle covariate siano molto differenti tra loro — col rischio conseguente che quelle con range maggiori dominino e mascherino l'influenza di quelle con range minori. Questa operazione ho verificato viene eseguita in automatico dal software impiegato.

Per quanto riguarda la scelta del kernel, ho optato per la funzione RBF (Radial Basis Function), data dall'equazione:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \qquad \text{con} \quad \gamma > 0 \tag{13.12}$$

in relazione al suo largo impiego con successo in applicazioni anche molto diverse tra loro; altri vantaggi sono la dipendenza da un unico parametro γ e la capacità di mappare in maniera *non*-lineare i campionamenti in uno spazio delle feature di dimensione maggiore rispetto a quello originario: a differenza di un kernel di tipo lineare, può quindi trattare anche i casi per i quali la relazione tra la classe e le covariate/attributi non sia di tipo lineare.

In relazione alle equazioni (13.10) e (13.12), il modello che si intende costruire dipenderà da soli due parametri, nello specifico C e γ. Come già sottolineato, non potendo ricorrere a un dataset indipendente di 'testing', per la loro stima ho utilizzato una *4-fold cross-validation*[4] sul dataset di 'testing', eseguendo in una prima fase una ricerca grossolana su un ampio range di valori per C e γ, e raffinando successivamente la ricerca attorno alle zone di minimo della superficie degli errori (o superficie di tuning, in quanto utilizzata per la ricerca dei valori ottimali dei parametri del modello). Una rappresentazione di una superficie di questo tipo è riportata nella figura 13.1.

13.3.2 Descrizione dei risultati ottenuti

La procedura appena descritta è stata applicata ai modello per la sola parte spaziale per entrambi i valori di soglia; tuttavia, per la soglia di 400 Bq m^{-3} si è ottenuta una superficie di tuning praticamente piatta, che quindi non consente una stima affidabile e precisa della coppia di parametri che caratterizzano il modello. Alla luce di questo, ho deciso di proseguire l'analisi limitandola alla soglia relativa ai 200 Bq m^{-3}.

Poiché anche per questo valore di soglia la superficie di tuning è comunque caratterizzata da numerose zone con minimi locali, oltre ai valori migliori per la coppia (γ, C) ho costruito un

[4] Purtroppo, non è stato possibile fare in modo che l'implicita suddivisione del dataset operativo in uno di 'training' e uno di 'testing' tenesse conto della componente spaziale dei dati, ossia costruisse i due dataset in modo da garantire una distribuzione spaziale simile sul territorio di studio.

secondo modello con una coppia alternativa di parametri, stimandone graficamente i valori; la tabella 13.1 riporta il valore dei parametri, l'errore ottenuto in fase di cross-validation, il numero di support vector utilizzati e la percentuale di misclassificazione ottenuta.

Una volta in possesso dei valori ottimali per la coppia di parametri richiesti, ho applicato il modello al dataset di 'validation', in modo da poter condurre delle analisi dei risultati avendo a disposizione, fissata la localizzazione, sia il valore stimato che quello realmente misurato. La tabella 13.2 riporta un confronto tra il numero di campionamenti stimati per una data classe e quelli che realmente vi appartengono: se l'algoritmo fosse "perfetto", una tabella di questo tipo dovrebbe essere diagonale. Infine, ho anche controllato, mediante post-plot, la distribuzione spaziale delle misclassificazioni.

Da queste prime analisi, in relazione ai modelli istruiti con la sola componente spaziale del fenomeno, si può affermare che:

- *l'influenza del valore dei parametri relativi a due differenti zone di minimo nella superficie di tuning* non *risulta evidente sui risultati globali ottenuti;*
- *la distribuzione spaziale delle misclassificazioni* non *mostra cluster evidenti (gli errori si distribuiscono in maniera piuttosto uniforme sul territorio di studio);*
- *purtroppo, per entrambi i modelli gli errori di classificazione vanno nella direzione meno accettabile in un contesto di prevenzione, a dire che quasi tutte le misclassificazioni sono tali per cui l'algoritmo di ML ha stimato un valore inferiore ai 200 Bq* m^{-3} *mentre quello misurato risulta superiore*

13.4 Costruzione di modelli più complessi

Successivamente, nella convinzione che la sola parte spaziale del fenomeno non sia in grado di catturarne tutte le caratteristiche, piuttosto complesse e interagenti tra loro in modo non sempre evidente, ho deciso di costruire altri modelli che prendessero in considerazione altre covariate, e nello specifico informazioni aggiuntive legate al contesto edilizio e geografico che accompagna il singolo valore misurato.

Le analisi sono state condotte in linea con quanto descritto nel paragrafo §13.3 e limitate al valore di soglia pari a 200 Bq m^{-3} — sia per questioni legate alla *leggibilità* della superficie di tuning, sia per ragioni legate alla *numerosità* esigua che caratterizza la classe superiore per il valore di soglia di 400 Bq m^{-3} (lo sbilanciamento delle due classi risulta pesante). In tutti i casi, le superfici di tuning sono risultate piuttosto *piatte*, con la presenza di più minimi locali: in questi casi, ho deciso di privilegiare le zone con valore del parametro C maggiore, vista la non trascurabile rumorosità dei dati. La tabella 13.3 riporta una descrizione dei vari modelli (numero e tipo di covariate coinvolte), i valori dei parametri che competono loro, l'entità dell'errore ottenuto in fase di cross-validation, il numero di support vector necessari e la percentuale di misclassificazione ottenuta in fase di validazione ricorrendo, come nel caso precedente, al dataset indipendente 'validation'.

I modelli 01, 02 e 03 sono stati costruiti in relazione ai risultati ottenuti in precedenza circa le covariate che hanno manifestato una influenza significativa sul valore di concentrazione misurato [cfr. capp. 9, 10, 11]. Il modello 04 per valutare invece l'eventuale potere predittivo delle caratteristiche morfologiche del terreno — che, a seguito delle citate analisi, non dovrebbero però portare informazioni utili alla previsione. Infine, mi è sembrato interessante "rovesciare" la prospettiva che ha guidato questa serie di analisi ed *eliminare* completamente l'informazione spaziale *esplicita*, ossia costruire i modelli 05 e 06 privandoli delle covariate relative alle coordinate

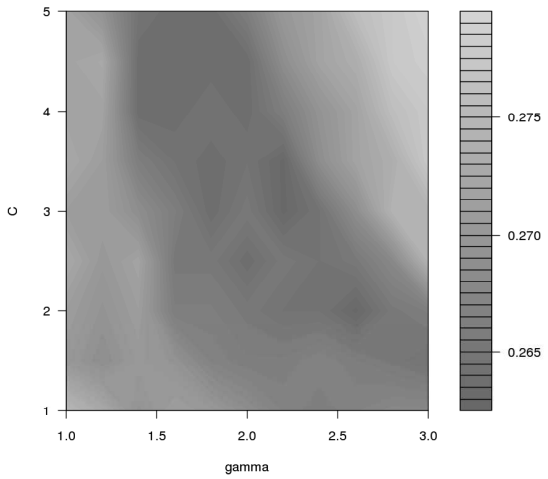


Figura 13.1: Esempio di visualizzazione grafica di una superficie di tuning per i parametri del modello C e γ; in questo modo, è possibile risalire alla coppia di parametri che fornisce il valore minimo dell'errore in fase di cross-validation.

id modello	γ	C	*errore CV*	*n SV*	*% misclass*
sp01	2.6	2.0	0.2628	727 (59%)	26
sp02	2.0	2.5	0.2636	723 (59%)	26

Tabella 13.1: Alcuni parametri relativi ai modelli di tipo spaziale; 'errore CV' indica l'errore ottenuto in fase di cross-validation, 'n SV' indica il numero (e la relativa percentuale) dei support vector utilizzati.

	reali	
stimati	$<$ *200*	$>$ *200*
$<$ *200*	268	100
$>$ *200*	8	33

(a) modello *sp01*

	reali	
stimati	$<$ *200*	$>$ *200*
$<$ *200*	267	99
$>$ *200*	8	34

(b) modello *sp02*

Tabella 13.2: Tabelle incrociate per un confronto tra stima e dati reali per le classi di concentrazione per i due modelli spaziali implementati, i cui parametri sono riportati nella tab. 13.1.

id modello	*covariate*
01	x + y + contatto + tipo costruzione
02	x + y + contatto + tipo costruzione + utilizzo
03	x + y + contatto + tipo costruzione + esposizione
04	x + y + altitudine + pendenza + curvatura
05	contatto + tipo costruzione + utilizzo
06	tipo locale + classe data costruzione + qualità infissi

id modello	γ	C	*errore CV*	*n SV*	*% misclass*
01	0.30	0.7	0.2571	714 (58%)	26
02	0.11	7	0.2538	695 (57%)	26
03	0.011	31	0.2652	749 (61%)	27
04	0.45	1	0.2799	811 (67%)	27
05	0.6	1	0.3027	749 (61%)	32
06	0.047	53	0.3320	793 (65%)	33

Tabella 13.3: Alcuni parametri relativi ai modelli più complessi che coinvolgono differenti tipi di informazione (riportati nella seconda colonna); 'err CV' indica l'errore ottenuto in fase di cross-validation, 'n SV' indica il numero (e la relativa percentuale) dei support vector utilizzati, '% misclass' la percentuale di misclassificazioni.

	reali	
stimati	< 200	> 200
< 200	264	95
> 200	12	38

(a) modello 01

	reali	
stimati	< 200	> 200
< 200	264	96
> 200	12	37

(b) modello 02

	reali	
stimati	< 200	> 200
< 200	262	98
> 200	14	35

(c) modello 03

	reali	
stimati	< 200	> 200
< 200	273	106
> 200	3	27

(d) modello 04

	reali	
stimati	< 200	> 200
< 200	246	99
> 200	30	34

(e) modello 05

	reali	
stimati	< 200	> 200
< 200	276	133
> 200	0	0

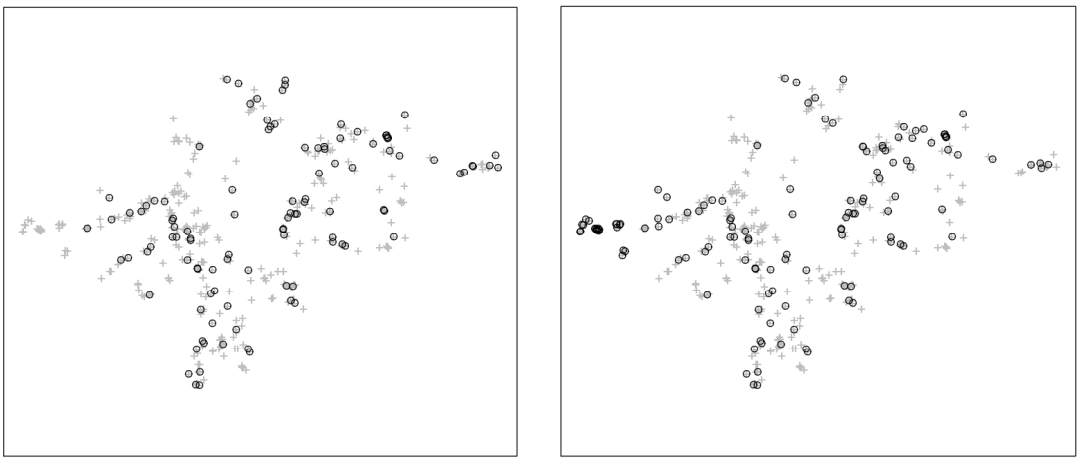
(f) modello 06

Tabella 13.4: Tabelle incrociate per un confronto tra stima e dati reali per le classi di concentrazione per i modelli descritti nella tabella 13.3, che riporta anche i relativi parametri (γ, C).

della localizzazione; inoltre, le covariate per il modello 06 non dovrebbero, in relazione a risultati ottenuti in altri contesti, essere significative nella determinazione del valore di concentrazione.

Anche in questo caso ho utilizzato il dataset indipendente 'validation' per testare e confrontare i vari modelli, sia tra loro sia in relazione ai reali valori delle classi; i risultati, in analogia a quanto discusso per i precedenti modelli spaziali, sono riportati nella tabella 13.4. Globalmente, si può concludere che:

- non *si ottengono evidenti miglioramenti sulla stima delle classi, nemmeno istruendo il modello con informazioni aggiuntive rispetto a quelle puramente spaziali;*



(a) modello 01 (b) modello 05

Figura 13.2: Post-plot delle misclassificazioni (cerchi) rispetto al dataset di 'training' (croci); si noti come per il modello 05 queste vadano a occupare intensamente la parte Ovest del territorio dell'Alto Adige; per maggiori dettagli, si faccia riferimento al testo.

- *rimane il 'difetto' di una misclassificazione polarizzata verso le sottostime già riscontrato anche nei modelli spaziali, ovvero è sempre alto il numero di campionamenti cui viene erroneamente assegnata la classe '< 200' — per i modelli 01, 02 e 03 questa misclassificazione è di qualche unità inferiore, ma certo non può essere considerato un risultato significativo;*

Questa apparente *insensibilità* dei modelli alle informazioni aggiuntive portate dalle ulteriori covariate coinvolte potrebbe forse trovare una parziale spiegazione nel fatto che, come emerso dai risultati dell'analisi dei quantili descritta nel paragrafo §8.3 a pagina 116, esse manifestano la loro influenza sul valore di concentrazione in maniera significativa al crescere del valore stesso: poiché una soglia pari a 200 $\mathrm{Bq \cdot m^{-3}}$costituisce, in relazione al dataset di riferimento, un valore non molto elevato, e la maggior parte degli esempi sono etichettati con '< 200', l'influenza delle covariate potrebbe effettivamente non essere così significativa in un contesto di questo tipo.

Ha invece dato dei risultati più interessanti e inattesi l'analisi della distribuzione spaziale delle misclassificazioni. Per i modelli 01–04, questa risulta omogenea sull'intero territorio di studio, e in linea con quanto ottenuto per i modelli spaziali; un esempio di tale distribuzione è riportato in figura 13.2a. Per quanto riguarda invece i modello 05 [cfr. fig. 13.2b] e 06, che ricordo sono stati privati dell'informazione relativa alle coordinate dei campionamenti, la distribuzione spaziale è simile per entrambi, ma *differente* rispetto agli altri modelli.

Come si nota facilmente confrontando quanto riportato nella figura 13.2, per i modelli privati della componente spaziale (esplicita), si ottiene un cluster di misclassificazioni nella zona Ovest dell'Alto Adige, che corrisponde al comune di Tubre. Caratteristiche di questa zona sono valori molto elevati di concentrazione di radon indoor ed elevata variabilità locale del fenomeno stesso, ragioni che la rendono una 'zona calda' con caratteristiche diverse rispetto alle altre zone del territorio di studio — le sue peculiarità sono pertanto legate alla sua particolare localizzazione. In relazione a questi risultati, si può verosimilmente concludere che

id modello	γ	C	*n SV*	*% misclass*
03	0.011	31	749 (61%)	27
03a	0.001	25	770 (63%)	29
03b	2	32	703 (57%)	33
03c	4	256	736 (60%)	33

Tabella 13.5: Alcuni parametri relativi ai modelli impiegati per l'analisi dell'influenza dei parametri γ e C; 'n SV' indica il numero (e la relativa percentuale) dei support vector utilizzati, '% misclass' la percentuale di misclassificazioni.

	reali	
stimati	$<$ *200*	$>$ *200*
$<$ *200*	262	98
$>$ *200*	14	35

(a) modello 03

	reali	
stimati	$<$ *200*	$>$ *200*
$<$ *200*	275	118
$>$ *200*	1	15

(b) modello 03a

	reali	
stimati	$<$ *200*	$>$ *200*
$<$ *200*	263	124
$>$ *200*	13	9

(c) modello 03b

	reali	
stimati	$<$ *200*	$>$ *200*
$<$ *200*	274	131
$>$ *200*	2	2

(d) modello 03c

Tabella 13.6: Tabelle incrociate per un confronto tra stima e dati reali per le classi di concentrazione per i modelli descritti nella tabella 13.5, che riporta anche i relativi parametri (γ, C).

la parte spaziale del fenomeno porta informazioni utili al modello: se lo si priva infatti delle coordinate dei singoli campionamenti, le misclassificazioni si vanno a concentrare in un'area che ha delle peculiarità legate alla sua specifica localizzazione sul territorio — zona che, in altre parole, si potrebbe considerare quasi un'area con caratteristiche proprie differenti rispetto a quelle generali *e medie che competono all'intero territorio altoatesino.*

13.5 Analisi dell'influenza dei parametri (γ, C)

Dato che in molti casi, soprattutto per i modelli 04 e 06, la superficie di tuning ha manifestato una struttura quasi assente, ho deciso di provare a costruire dei modelli al variare della coppia dei parametri (γ, C).

Come modello di riferimento ho scelto il modello 03, in quanto presenta una superficie di tuning con una certa struttura: invece di scegliere completamente "a caso" la coppia di parametri, ho preferito optare per dei valori che cadessero in zone caratterizzate comunque da valori di errori di CV piuttosto bassi — zone che, in altri termini, risultassero comunque papabili. Le caratteristiche dei modelli utilizzati sono riportati nella tabella 13.5, che per comodità mostra anche quelli relativi al modello 03 di riferimento.

La tabella 13.6 riporta invece i risultati più specifici relativi alle previsioni, in particolare al numero e tipo di misclassificazioni.

Dall'analisi dei dati riportati e da quelle visuali dei post-plot degli errori, si può concludere che:

- *globalmente, pur variando i parametri di qualche ordine di grandezza, non si ottengono risultati significativamente differenti (ricordo che il tipo di kernel non è stato variato, ed è del tipo RBF) — questo potrebbe essere indice di una eccessiva rumorosità dei dati o di una inadeguatezza del tipo di kernel utilizzato;*
- *la distribuzione spaziale delle misclassificazioni si differenzia invece rispetto a quella ottenuta per i modelli spaziali e i modelli 01, 02, 03 e 04: si notano dei cluster nelle 'zone calde' (come ad esempio la già citata zona di Tubre), quasi che la scelta dei valori dei parametri γ e C non si renda evidente a livello* globale, *ma piuttosto manifesti la sua influenza sulla bontà della previsione a livello* locale, *ovvero in quelle situazioni che per certi aspetti risultano più difficili da modellizzare.*

13.6 Analisi dell'output probabilistico

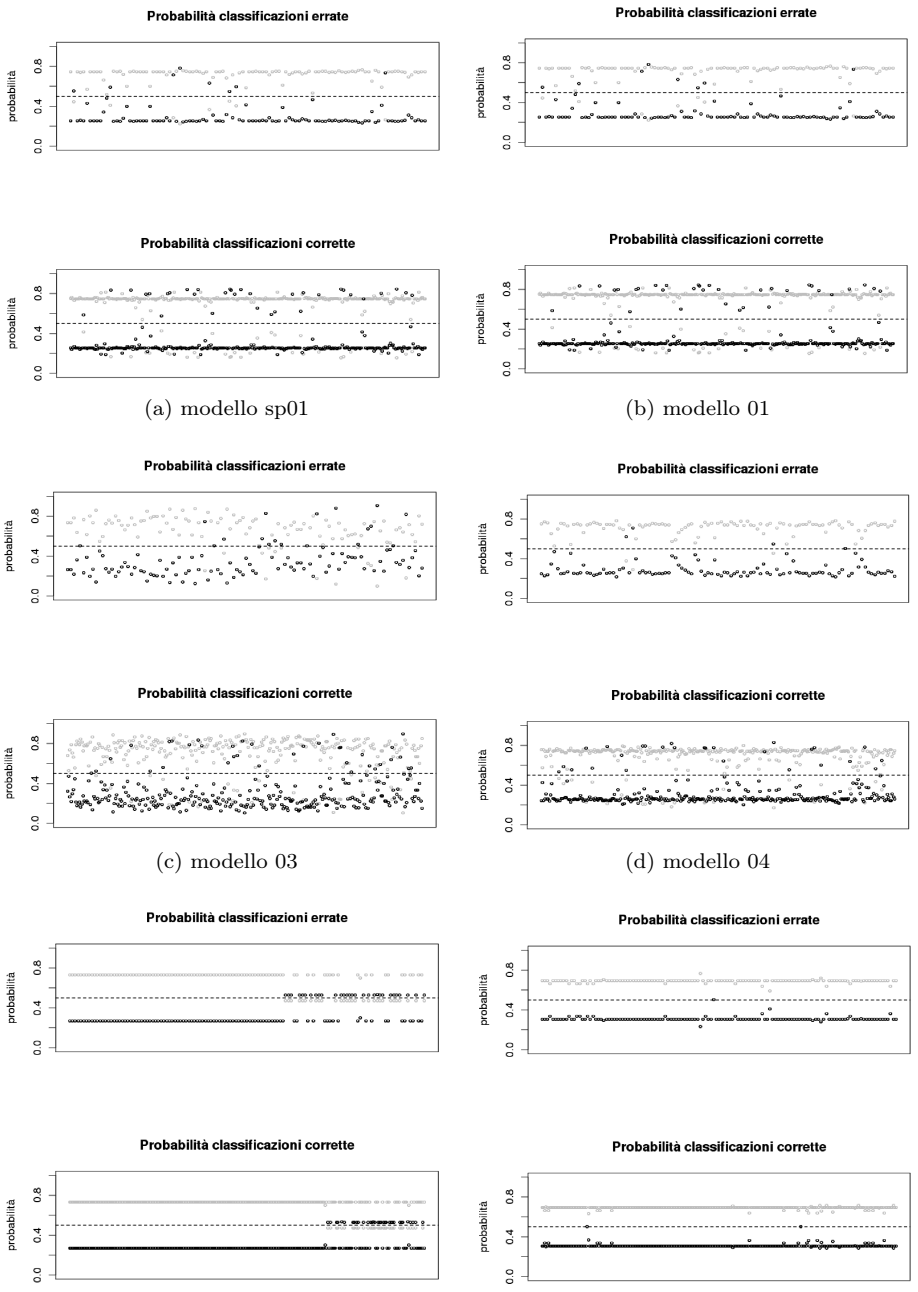
Per tutti i modelli costruiti, ho anche determinato le probabilità relative all'assegnazione della classe per ogni valore stimato, con l'idea di valutare quanto il modello in questione fosse per così dire "sicuro" dell'assegnazione fatta, sia in relazione alle classi assegnate correttamente che a quelle misclassificate: l'attesa era quella di valori di probabilità alti per le assegnazioni corrette e più bassi, attorno a 0.5–0.6, per quelle errate.

> ***Nota 2:*** *Si tenga comunque presente che le conclusioni che si possono trarre da questo tipo di analisi, come ricordato nel paragrafo §13.1.3, in virtù del fatto che si è ricorsi a un kernel RBK* non-*lineare potrebbero risultare discutibili e non supportate da solide fondamenta teoriche.*

La figura 13.3 riporta alcuni grafici utilizzati per questo tipo di analisi, in funzione del modello e del tipo di classificazione (corretta o meno). Per ogni assegnazione, il singolo grafico riporta il valore di probabilità per entrambe le classi, uno in grigio e uno di colore nero: i grafici sono pertanto simmetrici rispetto alla retta orizzontale in corrispondenza del valore 0.5, e ogni coppia verticale di punti, che si riferisce ai valori di probabilità per le due classi riferite alla singola stima, ha valori che si sommano all'unità.

In riferimento a questi grafici (per quanto riguarda il modello 02, non riportato, è graficamente analogo al modello 01), si può affermare che:

- *globalmente,* non *si notano particolari differenze tra i valori di probabilità (anche da un punto di vista dell'andamento grafico) per le classificazioni corrette o meno: la maggior parte dei valori si colloca attorno a 0.7–0.75, con poche classificazioni 'incerte' (a dire, con valori di probabilità vicini a 0.5) — eccettuato il modello 03* [cfr. fig. 13.3c]*; in altre parole, anche nelle situazioni di misclassificazione, il modello appare purtroppo comunque abbastanza sicuro del suo operato;*
- *i grafici per il modello 03, come si evince chiaramente dalla figura 13.3c, risultano i più rumorosi in relazione ai valori di probabilità e non mostrano il tipico raggruppamento attorno a un unico valore; questa incertezza sulle previsioni potrebbe essere apprezzabile nel caso delle classificazioni scorrette, ma purtroppo caratterizza anche quelle corrette;*



(a) modello sp01

(b) modello 01

(c) modello 03

(d) modello 04

(e) modello 05

(f) modello 06

Figura 13.3: Esempi di grafici che riportano i valori di probabilità per le classi assegnate in maniera corretta o meno, in relazione al dataset 'validation'; le probabilità sono state determinate in base all'eq. (13.11); per maggiori dettagli, si faccia riferimento al testo.

- *i risultati relativi ai modello 05 e 06* [cfr. figg. 13.3e e 13.3f] *fanno sospettare che ci possa essere qualche problema di natura computazionale o di implementazione dell'equazione* (13.11)*; inoltre, purtroppo non è stato possibile accedere alla procedura di stima per i parametri della* (13.11) — *che secondo la documentazione dovrebbe essere basata su una cross-validation — né ai valori degli stessi, in modo da poter eventualmente verificarne il comportamento con dei test specifici.*

Il comportamento anomalo dei modelli 05 e 06, che ricordo sono quelli privi della parte spaziale e le cui misclassificazioni si concentrano nella zona di Tubre rispetto agli altri modelli, ha fatto venire il sospetto che le misclassificazioni con bassi valori di probabilità potessero andare a collocarsi proprio nella zona di Tubre: se questa fosse la situazione, nell'errore sarebbe almeno un risultato positivo. Analisi specifiche della collocazione spaziale dei punti con una classificazione errata e un basso valore di probabilità non hanno però sostenuto questa ipotesi:

> *per tutti i modelli, le misclassificazioni che si possono definire* incerte — *a dire, con valori di probabilità prossimi a 0.5 — si collocano in maniera uniforme sull'intero territorio d'esame, senza evidenziare particolari cluster che potrebbero essere indice di un'inadeguatezza del modello su particolari sotto-domini con peculiarità caratteristiche.*

13.7 Conclusioni

Abbandonando l'ambito della regressione, è sembrato interessante valutare la possibilità di ricorrere a strumenti volti alla *classificazione* per lo studio del fenomeno radon indoor. In questo contesto, si è fatto riferimento al problema della classificazione binaria — in una localizzazione non campionata, la misurazione di concentrazione di attività di radon indoor sarà sopra o sotto un determinato valore di soglia? — ricorrendo agli strumenti messi a disposizione dall'approccio Support Vector Machine (SMV), che attualmente costituisce lo stato dell'arte per quanto concerne i classificatori.

Il dataset impiegato in questo tipo di analisi è stato costruito a partire da quello descritto in dettaglio nel paragrafo §1.3, a pagina 7, che riporta per ogni singola misura anche una serie di variabili secondarie utili alla descrizione del contesto geologico ed edilizio nel quale la misura è stata condotta. Accanto a modelli costruiti unicamente sulle coordinate, se ne sono costruiti altri al variare del numero e del tipo di variabili secondarie coinvolte.

Questa preliminare e limitata analisi ha portato alle seguenti conclusioni, basate sul confronto tra stima di appartenenza a una determinata classe (sopra o sotto soglia) e reale valore misurato (ricorrendo a un dataset indipendente di validazione):

- sia nel caso di semplici modelli che prevedano la sola componente spaziale, sia nel caso di modelli più complessi che coinvolgano ulteriori variabili secondarie, gli errori di classificazione vanno a collocarsi nella direzione meno accettabile in un contesto di prevenzione, in quanto la maggior parte delle misclassificazioni risultano tali per cui l'algoritmo di ML ha stimato un valore sotto-soglia quando quello realmente misurato è risultato invece sopra-soglia[5];

[5] Il fatto che anche i modelli più complessi siano affetti da questo infelice problema potrebbe trovare una parziale spiegazione leggendo i risultati alla luce di quanto emerso dall'analisi condotta nel capitolo 8, relativo alla regressione dei quantili: le variabili secondarie che manifestano una influenza riconosciuta sul valore di concentrazione,

- nel caso si privi il modello delle coordinate relative ai campionamenti, le misclassificazioni non risultano più spazialmente omogenee sul territorio di studio, ma si vanno a concentrare in una zona che ha delle caratteristiche proprie e legate alla sua specifica localizzazione: questo dimostra che la componente spaziale del fenomeno risulta comunque importante anche per questo tipo di modelli.

lo fanno in maniera sensibile al crescere del valore di concentrazione stessa. È quindi verosimile che una soglia pari a 200 $Bq \cdot m^{-3}$ possa essere troppo bassa, in relazione al dataset impiegato, perché l'informazione aggiuntiva contenuta nelle variabili secondarie possa manifestarsi in maniera evidente.

CAPITOLO 14

Conclusioni Generali e Possibili Prospettive...

Una buona capacità di sintesi credo presupponga una visione chiara e ben definita dell'intero materiale che si intende rivedere e presentare in maniera più omogena e strutturata. Nell'illusione (forzata) che questa sia la reale situazione in cui mi trovo, mi sembra utile, a conclusione del lavoro, cercare di fornire un quadro generale di massima che ripercorra i vari sentieri intrapresi durante il periodo di dottorato, strutturando la discussione in modo da porre in luce gli aspetti e le conclusioni che appaiono di maggior interesse, tanto in relazione ai risultati ottenuti quanto alla loro possibile ricaduta in ambito prettamente applicativo.

Si descriverà come sia stato possibile identificare le variabili secondarie che accompagnano il singolo valore misurato di concetrazione di attività di radon indoor e che sono risultate essere le più predittive per la variabile di interesse; quindi, quali sembrano essere, alla luce dell'esperienza acquisita e degli strumenti presi in considerazione, quelli più indicati per la *descrizione* dei dati e quelli più indicati per una *stima/previsione* dei valori di radon indoor in localizzazioni non campionate, dando dove necessario qualche consiglio in relazione alla loro diretta applicazione. Infine, qualche ulteriore indicazione su quanto sarebbe stato interessante fare, ma che per varie ragioni non si è riusciti a concretizzare!

Potrà forse risultare superfluo, ma mi preme sottolineare come tutto quanto verrà discusso in seguito vada comunque interpretato in relazione ai dati che sono stati oggetto di studio [cfr. §1.3], e che, come ricordato più volte, hanno delle peculiarità — come ad esempio una elevata variabilità dei dati stessi, la presenza di un evidente effetto proporzionale, una rete di monitoraggio forzatamente disomogenea e parzialmente clusterizzata (conseguenza delle caratteristiche orografiche del territorio di studio), ... — che li possono rendere ben diversi rispetto ad altri dataset relativi a differenti zone geografiche.

Infine, si tenga presente che l'attenzione è stata, nella maggior parte dei casi, rivolta alla componente *antropogenica* del fenomeno (rispetto alla pura componente spaziale) per quanto riguarda l'informazione aggiuntiva contenuta nelle covariate; questa scelta nella convinzione che la variabile protagonista delle analisi, ovvero il valore di concentrazione di attività di Radon Indoor, sia pesantemente influenzata dalla complessa e non ben definita interazione con le

caratteristiche dell'edificio, e non tanto in stretto legame con la parte geologica del fenomeno — dibattito che comunque risulta tutt'ora aperto all'interno della comunità scientifica.

14.1 Le variabili informative...

Il dataset messo a disposizione dall'APPA di Bolzano, come dettagliatamente descritto nel paragrafo §1.3 a pagina 7, riporta, per ogni singola misura condotta negli edifici, una nutrita serie di informazioni (codificate in variabili secondarie di tipo sia qualitativo che quantitativo) atte alla descrizione del contesto edilizio e geologico nei quali è stato esposto il dosimetro. In virtù di una "economicità" dei modelli che si intendono costruire e della conseguente facilità di interpretazione dei risultati, nonché di una migliore e più approfondita comprensione del 'fenomeno' radon indoor dal punto di vista fisico e quindi delle sue possibili interazioni con la parte antropogenica, ho cercato di guardare al problema da punti di vista differenti, distanti e complementari tra loro, sperando che risposte anche parzialmente sovrapponibili potessero essere indice di affidabilità delle stesse.

Combinando tra loro i risultati ottenuti i) analizzando la struttura variografica sperimentale in funzione delle covariate e delle relative classi (cap. 5), ii) ricorrendo alla Regressione dei Quantili (cap. 8), iii) utilizzando un algoritmo di Feature Selection (cap. 9) e iv) sfruttando in maniera indiretta approcci di Machine Learning (capp. 10 e 11), è emerso in maniera piuttosto evidente come le covariate che risultano essere *informative* in relazione al valore della variabile di interesse *Radon Indoor* siano **contatto con il terreno** e **tipo di materiale da costruzione**.

Inoltre, sembra interessante sottolineare come la variabile secondaria "contatto", semplice, di tipo binario e di facile reperimento, risulti in buona correlazione con altre covariate più complesse di tipo antropogenico, come discusso nelle analisi condotte nei capitoli 7 e 9: questo fatto è positivo, in quanto sembra si sia trovata una covariata semplice in grado di raccogliere informazioni utili contenute in altre covariate più complesse.

Dalle analisi citate in precedenza, sono emerse anche altre variabili secondarie con un buon potere predittivo sulla variabile di interesse, anche se non comuni a tutti i tipi di analisi. Riassumendo, la variabile radon indoor è in stretto legame con le seguenti variabili caratterizzanti il contesto in cui la misura è stata condotta:

$$\text{Radon Indoor} \rightleftharpoons \begin{cases} \textit{contatto con il terreno} \\ \textit{tipo di materiale da costruzione} \\ \text{utilizzo} \\ \text{esposizione} \end{cases}$$

Nello schema precedente, l'ordine in cui le covariate sono presentate rispecchia la frequenza con la quale sono comparse nelle varie analisi prese in considerazione.

> Operativamente, mi sembra interessante sottolineare una buona caratteristica propria del dataset cui si è fatto ricorso: conducendo analisi sulla distribuzione spaziale per le classi relative alle numerose covariate, è emerso come i relativi campionamenti si distribuiscano in maniera *uniforme* sull'intero territorio altoatesino, senza mostrare cluster evidenti in zone specifiche. Ritengo che questa prorpietà risulti particolarmente utile nei casi in cui si intendano costruire modelli "misti", ovvero che mirino a combinare tanto la parte prettamente spaziale quanto quella antropogenica (legata alle covariate): si può infatti contare sul fatto che tutto il dominio di studio risulta opportunamente rappresentato spazialmente dai campionamenti relativi alle varie classi.

14.2 Strumenti per la descrizione dei dati...

Qualunque sia il tipo di indagine, lo strumento scelto o il fine della stessa, una *conoscenza* e una *descrizione* il più possibile approfondite e accurate del dataset operativo credo siano un primo passo fondamentale e irrinunciabile cui dedicare gran parte del tempo complessivo richiesto dalla studio generale. Una buona familiarità con i dati consente infatti di operare, in fase di costruzione dei modelli, le scelte (sperabilmente) migliori fondandole su basi sufficientemente sicure, e in fase di analisi dei risultati, di interpretarli nella maniera corretta.

In quest'ottica, sono stati numerosi e diversificati gli ambiti e gli strumenti cui si è fatto ricorso, a volte in modo esplicito, altre ricavando informazioni in tale direzione in maniera indiretta. Si passeranno quindi brevemente in rassegna gli strumenti che si sono rivelati utili per una analisi esplorativa del dataset di riferimento.

Variogram Cloud (cap. 4) — questo strumento geostatistico risulta nella pratica utile ed efficace per l'identificazione di situazioni per così dire anomale del tipo "valore elevato di concentrazione con vicinaggio caratterizzato da valori bassi", mentre non è in grado di rendere altrettanto evidente la presenza di situazioni opposte.

> Operativamente, nelle situazioni in cui sono presenti anomalie il cui valore numerico si colloca all'*interno* del campo medio di variabilità che caratterizza il dataset di riferimento, queste possono essere identificate correttamente analizzando la variogram cloud su brevi distanze; nel caso siano invece presenti anomalie il cui valore numerico si colloca all'*esterno* del campo medio di variabilità che caratterizza il dataset di riferimento (anomalie che possono essere identificate come eventuali outliers), queste si rendono visibili sull'intero range della variogram cloud; visivamente, appaiono come delle *strisce orizzontali* la cui posizione lungo l'asse y risulta proporzionale all'entità dell'anomalia stessa.

Analisi Variografica su singole classi (cap. 5) — calcolare variogrammi sperimentali in funzione di singole classi relative alle differenti covariate eventualmente presenti nel dataset di riferimento ha portato a strutture di correlazione spaziale in alcuni casi di più facile interpretazione; analizzare quindi nello specifico queste classi/covariate, può rivelare informazioni utili sulla loro eventuale influenza sul valore di concentrazione, portando in luce la presenza di subset la cui analisi potrebbe fornire spunti per ulteriori analisi esplorative.

Teoria dei Valori Estremi (cap. 6) — questo strumento statistico, nato per una esplicita modellizzazione di eventi rari (o estremi, appunto), e attualmente impiegato con successo soprattutto in ambito economico e meteorologico, può venir facilmente applicato anche ai dati di radon indoor per una corretta riproduzione delle caratteristiche della sua p.d.f.; rispetto al più diffuso approccio di tipo Normale/log-Normale, infatti, le distribuzioni proposte dalla teoria dei valori estremi sono in grado di riprodurre sia l'evidente *asimmetria* quanto il comportamento di entrambe le *code* della distribuzione empirica. Mantenendo la loro validità anche se applicato a subset di dati che si differenziano sia per la numerosità che per le caratteristiche statistiche, questo strumento sembra essere sufficientemente robusto e affidabile per una corretta analisi dei dati di radon indoor.

Association e Mosaic Plot (cap. 7) — questo tipo di grafici, con la relativa estensione che rende possibile implementare nella stessa rappresentazione anche parametri statistici relativi a test sulla significatività dell'eventuale dipendenza riscontrata tra la coppia di covariate categoriche in esame, risulta essere ricco di informazioni relative tanto alla distribuzione in

frequenza delle singole classi, quanto alla eventuale correlazione tra classi appartenenti a covariate diverse.

Operativamente, questo tipo di strumenti grafici si rivela valido per confermare o smentire quanto si può prevedere da un punto di vista teorico in base alle conoscenze relative al fenomeno radon indoor, portando alla luce eventuali situazioni (apparentemente) impreviste o anomale che possono di conseguenza stimolare e suggerire ulteriori analisi in tali direzioni.

Regressione dei Quantili (cap. 8) — come strumento di tipo esplorativo, risulta particolarmente interessante e utile nell'ambito di tematiche ambientali in quanto è in grado, rispetto ai più diffusi sistemi basati sui minimi quadrati, di rivelare l'entità dell'influenza delle potenziali variabili secondarie predittive in funzione del valore della variabile di interesse: spesso infatti tale ricercata influenza può venir mascherata dalla complessa interazione tra più covariate, o manifestarsi solo per determinati range di valori (ad esempio, solo per valori bassi oppure elevati); ancora, ricorrendo a opportune visualizzazioni, è possibile valutare quanto la relazione con la variabile da predire sia significativa analizzando gli intervalli di confidenza che accompagnano la stima del valore del peso predittivo.

Operativamente, è bene tener presente che in alcune situazioni, ricorrere a un numero "eccessivo" di variabili secondarie per la costruzione del modello di riferimento può portare a difficoltà di interpretazione dei risultati ottenuti, soprattutto nel caso in cui alcune variabili risultino fortemente correlate tra loro; inoltre, si presti attenzione al fatto che cambiare la classe di riferimento che caratterizza il modello costruito può, in alcuni casi, avere ripercussioni non trascurabili sulla leggibilità dell'influenza delle altre classi in relazione alla variabile di riferimento.

Algoritmo di Feature Selection (cap. 9) — si è preso in considerazione un solo tipo di algoritmo della classe "filter", che in fase di pre-processing dei dati ha lo scopo di selezionare, tra tutte quelle disponibili, le covariate che risultano maggiormente correlate con la variabile da predire e al contempo il meno correlate possibile tra loro: in questo modo si ottiene un dataset più snello dal quale è stata, almeno in linea di principio, eliminata gran parte dell'informazione ridondante.

Operativamente, questo algoritmo si è rivelato sufficientemente robusto alla presenza di rumore (variabili spurie) e allo sbilanciamento in numerosità che può caratterizzare la distribuzione in frequenza delle classi all'interno delle singole covariate; questo metodo, di tipo "oggettivo", può inoltre trovare una valida applicazione in abbinamento ad altri tipi di tecniche esplorative di tipo visuale (e quindi più "soggettive", dal punto di vista dell'operatore), come verifica e supporto delle stesse.

Algoritmi w*k*-NN e GRNN (capp. 10 e 11) — questo tipo di algoritmi (nello specifico, rispettivamente *Weighted k-Nearest Neighbor* e *General Regression Neural Network*), possono venir con successo impiegati anche come validi strumenti per analisi di tipo esplorativo, fornendo risultati utili per indagare le proprietà dei dati di radon indoor e la loro relazione con le numerose variabili secondarie che accompagnano ogni singola misurazione; in particolare, si sono rivelati utili per *identificare* le covariate che mostrano una qualche *influenza significativa* sul valore di concentrazione misurato negli edifici.

Operativamente, va sottolineato come questi risultati si ottengano in maniera "indiretta", ovvero richiedano la costruzione di più modelli differenti e un successivo confronto tra i risultati forniti dagli stessi; i diversi modelli si dovranno distinguere

verosimilmente per numero e tipologia delle covariate in essi coinvolte, che andranno scelte in funzione del tipo di informazione che si intende ricavare.

14.3 Strumenti per la previsione/stima...

Benché il baricentro del lavoro presentato sia stato posto sugli aspetti descrittivi del fenomeno radon indoor, con lo scopo di approfondire la conoscenza di base dello stesso, qualche tentativo è stato orientato anche nella direzione della previsione del valore di concentrazione in localizzazioni non campionate — con l'idea di arrivare, alla fine, alla costruzione di un qualche tipo di *mappa*, nel senso più generale del termine.

Anche se non sempre in maniera esplicita, il metro di paragone che ho assunto come riferimento è stato quello delle *Simulazioni Gaussiane Sequenziali*[1] (sGs), che in virtù della relativa facilità di implementazione e delle diversificate e variegate elaborazioni che si possono condurre sui risultati "grezzi" ottenuti si configurano, almeno in relazione alla mia personale esperienza, come uno degli strumenti migliori in questo ambito.

Fra i vari strumenti utilizzati, vale la pena di ricordare:

Kriging log-Nornale (cap. 3) — questo approccio ha prodotto stime che sono risultate in linea con quanto ottenuto attraverso altri strumenti di tipo geostatistico, quali kriging ordinario e simulazioni gaussiane sequenziali; anche lo strumento in esame risulta affetto da un pesante effetto di smoothing sulle stime ottenute.

Operativamente, si ha il vantaggio di una variogramma più leggibile e quindi di una più facile modellizzazione dello stesso, che non si traduce però in evidenti miglioramenti in fase di stima; per contro, l'intero sistema risulta piuttosto sensibile al valore di sella e di nugget del modello di variogramma (sarà quindi necessario porre particolare attenzione nella determinazione di tali parametri) ed è necessario assumere, in maniera più o meno esplicita, un modello di tipo log-Normale perla p.d.f. dei dati di radon indoor, ipotesi che non sempre è verificata nella realtà.

Quindi, alla luce di quanto appena discusso, l'approccio del kriging log-normale non si è dimostrato un'alternativa valida ed efficace rispetto ad altri algoritmi precedentemente testati, quali appunto kriging ordinario e simulazioni gaussiane sequenziali.

Teoria dei Valori Estremi (cap. 6) — benché lo strumento non sia direttamente votato alla stima del valore della variabile di riferimento in localizzazioni non campionate, può essere applicato indirettamente a problemi di questo tipo; in particolare, può venir impiegato nella fase di trasformazione Nscored diretta e inversa come richiesto dall'implementazione delle Simulazioni Gaussiane Sequenziali (sGs). Disponendo di un modello analitico per la p.d.f. dei dati di radon indoor in grado di riprodurne le peculiarità (rispetto ad esempio a un modello di tipo log-Normale), le stime ottenute risultano più aderenti alla realtà, in particolare in relazione alla generazione di valori elevati (estremi) che sono forzatamente trascurati dagli altri modelli cui abitualmente si fa ricorso.

Operativamente, ricorrere a un modello GEV [cfr. §6.1.1] in fase di trasformazione Nscored nell'ambito delle simulazioni gaussiane sequenziali fornisce una situazione media più aderente alla realtà rispetto ad approcci di tipo Normale/log-Normale; questi risultati risultano più evidenti nel caso di vogliano ottenere delle mappe E-type (valor medio sul territorio). Nel caso in cui invece l'output richiesto sia una

[1] Informazioni a riguardo si possono trovare nel paragrafo §B.6.1, a pagina 244; applicazioni specifiche al problema della stima dei valori di radon indoor, ad esempio nel paragrafo §6.5 a pagina 89, in {42, cap. 9}, in {52} e in {3}.

mappa di probabilità, ricorrere alla p.d.f. empirica in fase di trasformazione (qualora il software lo consenta) porta a risultati altrettanto affidabili.

Bisogna però ricordare che il modello GEV *non* è limitato a valori positivi della variabile in esame, e pertanto bisognerà prestare la dovuta attenzione qualora lo si intenda applicare a dati di radon indoor, che per loro stessa natura sono limitati a valori positivi.

Operativamente, se lo scopo è quello di ricorrere a modelli GEV analitici in fase di trasformazione Nscored nel processo di una sGs, sarà necessario prestare attenzione al fatto che i valori estratti casuali dalla c.d.f. GEV risultino tutti positivi, e in caso contrario intervenire in maniera opportuna; inoltre, se il fine sarà quello di una mappa E-type, fittare il modello GEV sui dati grezzi può produrre valori eccessivamente elevati — per problemi legati al decadimento della coda superiore: in questi casi, può risultare utile fittare il modello GEV su dati log-trasformati (applicando poi la trasformazione inversa per ottenere il "vero" valore di concentrazione[2]).

Regressione dei Quantili (cap. 8) — anche se nell'implementazione che ne è stata fatta questo strumento non considera la parte spaziale del fenomeno (coordinate), consente tuttavia di poter costruire uno o più modelli di "casa tipo" con determinate caratteristiche edilizie e/o abitative fissate a priori, rispetto alle quali valutare se, in quale direzione e con quale entità la situazione cambia al variare delle caratteristiche stesse. Ad esempio, se la "casa tipo" ha una qualità degli infissi 'buona', il modello consente di valutare cosa ci si potrebbe aspettare se quella stessa casa avesse invece una qualità degli infissi 'scarsa'.

w*k*-NN e GRNN (cap. 12) — entrambi questi approcci, appartenenti alla grande famiglia di algoritmi di Machine Learning (ML), consento di trattare sia la componente spaziale che la componente antropogenica del fenomeno radon indoor, eventualmente una volta individuate in maniera opportuna le covariate realmente informative. Se da un lato la modellizzazione della parte spaziale non risulta raffinata quanto quella proposta da un approccio di tipo geostatistico, il vantaggio è quello di poter istruire il modello con l'informazione aggiuntiva codificata dalle variabili secondarie che caratterizzano ad esempio il contesto edilizio (ma potrebbe essere anche quello geologico, o entrambi).

Operativamente, da un punto di vista *globale* i risultati ottenuti mediante entrambi gli algoritmi risultano in linea con quanto si può ottenere mediante un modello geostatistico di kriging ordinario (OK) — assumendo che i vari modelli coinvolti siano stati costruiti con la medesima cura; a dire, la situazione media globale sull'intero territorio di studio viene riprodotta in maniera soddisfacente in tutti i casi.

Interessante notare però come gli approcci ML risultino più efficaci rispetto a OK nelle situazioni *locali* nelle quali si riscontrano gli errori maggiori (sia in sovra- che in sotto-stima): in questi singoli casi, infatti, le stime ML risultano più aderenti al reale valore misurato, in virtù dell'informazione aggiuntiva relativa all'intorno di stima cui hanno accesso attraverso le covariate di tipo antropogenico.

Concludendo, riporto nella tabella 14.1 uno schema di massima riassuntivo per un più diretto confronto tra le varie metodologie che sono state descritte; si tenga presente che le conclusioni tratte sono chiaramente da leggersi in relazione a quanto direttamente sperimentato nella pratica e per il dataset operativo impiegato, relativo a misure di concentrazione di attività di radon indoor raccolte in Alto Adige.

[2] Si tenga presente che questo tipo di trasformazione inversa non presenta i problemi analitici e teorici legati invece all'approccio del kriging log-Normale.

	descrizione	*previsione*	*affidabilità*	*note*
kriging log-normale	n.d.	☒	☐	👎 sensibilità ai valori di sella e nugget; 👎 assunzione del modello log-normale; 👍 leggibilità del variogramma
variogram cloud	☑	n.d.	☑	👍 possibilità di identificare punti con caratteristiche anomale rispetto al loro vicinaggio
teoria valori estremi	☑	☒	☑	👍 applicazione indiretta in sGs (trasformazioni diretta e inversa Nscored)
regressione dei quantili	☑	☐	☒	👍 informazioni sull'influenza delle covariate anche per i valori di coda della p.d.f. della variabile di riferimento; 👎 possibili difficoltà interpretative
feature selection	☑	n.d.	☑	👍 procedura "automatica" da poter mettere a confronto con tecniche visuali più "soggettive"
wk-NN	☑	☒	☑	👍 buoni risultati in previsione per situazioni locali complesse; 👎 analisi descrittiva "indiretta" (costruzione di modelli ad hoc)
GRNN	☒	☐	☒	👍 buoni risultati in previsione per situazioni locali complesse; 👎 analisi descrittiva "indiretta" (costruzione di modelli ad hoc)

Tabella 14.1: Tabella riassuntiva per un confronto di massima tra le varie tecniche analizzate in questo lavoro, limitatamente agli aspetti che sono stati direttamente implementati. I simboli riportati hanno il seguente significato: ☐ per 'sufficiente', ☒ per 'medio' e ☑ per 'buono'; nella colonna "note", 👍 indica gli aspetti positivi riscontrati, 👎 quelli negativi. La colonna "affidabilità" fornisce infine un giudizio globale relativo alla robustezza della tecnica e alla facilità di interpretazione dei risultati ottenuti rispetto alle caratteristiche dei dati esaminati.

14.4 ...e (alcune) possibili prospettive future!

Durante il periodo di dottorato, partendo da un problema comune si sono percorsi sentieri differenti, puntando di volta in volta la bussola verso orizzonti non sempre ben definiti; in alcune occasioni, la marcia si è conclusa in luoghi abbastanza stabili e sicuri, in altre, nebbia e intrico di rami hanno imposto una (temporanea?) sosta. Ma nessuno è perduto, visto che le vie ancora inesplorate sono numerose, e la prima difficoltà potrebbe proprio essere quella legata al "dove" dirigere i passi futuri...

Gli strumenti e le tecniche oggi disponibili sono indubbiamente tanti e accattivanti, e operare delle scelte risulta così legato tanto all'esperienza quanto al "gusto" personali. Senza alcuna pretesa di proporre una panoramica completa ed esaustiva, tra le metodologie per così dire "già confezionate" descriverò brevemente quelle che hanno maggiormente stimolato la mia curiosità, in relazione alla loro efficacia nel contesto dell'analisi e della modellizzazione del fenomeno/problema radon indoor.

- Nell'ambito delle Simulazioni Stocastiche, benché quelle di tipo gaussiano (sGs) si siano rivelate, almeno per quanto riguarda la mia esperienza, robuste e affidabili, è comunque noto che queste ultime, come suggerito ad esempio da Goovaerts {22}, non siano la scelta migliore nelle situazioni per le quali la correlazione tra i valori estremi sia una caratteristica importante del problema in esame; un tipo di simulazione alternativa che potrebbe venir impiegata in questo contesto è quella nota come *Simulated Annealing*, il cui algoritmo prevede di prendere come punto di partenza una "immagine" casuale del fenomeno e proporre quindi iterativamente dei cambiamenti: questi verranno accettati o meno sulla base di una funzione predefinita fino al raggiungimento di un determinato grado di accordo con l'immagine di riferimento[3]. Il problema viene così formulato in termini di ottimizzazione di una determinata funzione di riferimento, che potrebbe essere ad esempio il variogramma sperimentale, l'istogramma dei dati, una combinazione di entrambi, o quanto dettato dalla fantasia e/o dall'esperienza dell'operatore.

- Kriging di tipo *bayesiano*, di cui il ben più noto e diffuso kriging ordinario risulta essere un caso particolare. In questo contesto, si possono introdurre incertezze e conoscenze a priori tanto sui parametri che caratterizzano il variogramma, quanto su quelli che competono al fenomeno in esame; tra i principali vantaggi offerti da questo approccio:
  - la possibilità di ottenere non solo il valore per un dato parametro, ma la sua intera p.d.f.;
  - la possibilità di ottenere, anche per la singola stima, la p.d.f. completa, sulla quale poter condurre, in fase di post-processing, una gran varietà di ulteriori analisi statistiche;
  - la possibilità, attraverso la modifica delle cosiddette 'prior' e 'posterior' p.d.f., di poter "aggiornare" il modello sulla base di nuovi dati disponibili e/o nuove concoscenze acquisite.

  Tuttavia, questa tecnica presenta alcune difficoltà legate alle scelte (che possono risultare cruciali) delle 'prior' distribution che inevitabilmente dovranno essere associate a ogni parametro del modello; inoltre si fonda sull'assunzione di un modello stocastico di tipo gaussiano.

[3] Si può pensare a questo modello come a una simluazione stocastica del lento raffreddarsi di un sistema fisico — per il quale la funzione predefinita sia ad esempio l'hamiltoniana H del sistema a una certa temperatura T — fino al raggiungimento di un minimo locale per H stessa, situazione che rappresenterà quindi l'immagine stocastica di arrivo.

- Nello scorso decennio si è sviluppato un metodo che si configura come una generalizzazione estremamente interessante e promettente della teoria classica della geostatistica; quello che è noto come *Bayesian Maximum Entropy (BME) approach* [cfr. i lavori di Christakos {7} e di Christakos *et al.* {8}]. Il fondamento teorico sui cui si basa, al contempo robusto, flessibile ma anche altrettanto complesso, consente di incorporare nel modello praticamente qualsiasi tipo di informazione che possa venir codificata in maniera analitica. Il grande vantaggio è quello che le previsioni consistono in una completa p.d.f. della singola stima, così che si possano successivamente determinare svariati parametri o indicatori, a seconda del tipo di analisi e del tipo di informazione che si va ricercando (media, varianza, intervalli di confidenza, probabilità di superare un determinato valore di soglia, . . .). Inoltre, si può dimostrare come il kriging possa essere visto, in questo ben più ampio contesto, come un caso limite dell'approccio BME.

Sottolineo comunque come le tecniche di Machine Learning applicate nei capitoli 10, 11 e 13 necessitino di ulteriori approfondimenti, in quanto sono state affrontate a un livello piuttosto superficiale — soprattutto in relazione al Support Vector Machine. Accanto a quelli appena citati e brevemente descritti, anche questi approcci risultano a mio giudizio molto interessanti e promettenti nell'ambito dei problemi relativi allo studio del fenomeno/problema radon indoor.

Ci sono infine alcune idee che mi sembrerebbe interessante testare, ma che, limitatamente alle mie conoscenze in merito, non hanno ancora trovato una diretta applicazione all'interno di tecniche specifiche; in particolare:

- mi piacerebbe riuscire a "far comunicare" tra loro modelli geostatistici e di Machine Learning, potendo così sfruttare i pregi di entrambi gli approcci, che sono da un lato una raffinata e completa descrizione della correlazione spaziale dei dati, dall'altro la possibilità di ricorrere a informazioni aggiuntive senza la necessità di introdurre alcuna assunzione a priori sul modello che dovrebbe sottendere ai dati di interesse. Un'idea da sviluppare potrebbe essere quella di abbandonare, ad esempio nell'approccio Weighted k-Nearest Neighbor, la distanza di tipo euclideo per introdurne una di tipo "statistico", come quella che caratterizza appunto i modelli di tipo geostatistico;

- mi piacerebbe costrire un modello che fosse in grado di considerare, in fase di stima, non solo la pura correlazione spaziale dei dati (e l'eventuale informazione aggiuntiva propria di ulteriori possibili covariate), ma anche il valore di *incertezza sperimentale* che accompagna la singola misura di radon indoor — considerando anche questa come un "peso" caratteristico dei campionamenti presenti nel vicinaggio di stima;

- sulla base di approfondite analisi esplorative volte a una dettagliata comprensione e caratterizzazione del contesto antropogenico del fenomeno, riuscire a determinare una nuova variabile "fittizia" depurata della componente che potremmo chiamare "edilizia" e condurre su questa una mappatura (anche con tecniche convenzionali), nella speranza di ottenere una variabile con una struttura di correlazione spaziale (variogramma) di più facile modellizzazione; successivamente, in virtù della precedente analisi esplorativa condotta, in una localizzazione non campionata correggere il valore della variabile mappata in base alle caratteristiche dell'edificio presente, oppure in base a quelle di un edificio che si andrà a realizzare nella zona — avendo così la possibilità di valutare quale valore di concentrazione ci si potrebbe aspettare in base alle proprietà edilizie dell'abitazione in questione.

Parte IV

Appendici

APPENDICE A

Riferimenti Statistici

Si descrivono brevemente i principali parametri statistici e le notazioni che sono state utilizzate in questo lavoro di tesi; un buon testo di riferimento, non troppo formale e che pone l'accento sugli aspetti applicativi, è quello di Cowan {12}.

Sia S un dato insieme — definito *spazio campionario* — che contiene un certo numero di elementi, la cui interpretazione può essere la più varia; sia x una generica *variabile aleatoria* (VA), ossia una variabile che assume un valore numerico in S; x può anche essere un vettore di più VA, nel qual caso sarà indicato come $\mathbf{x} = (x_1, \ldots, x_n)$. *VA*

Con il simbolo $\widehat{(\cdot)}$ verrà indicato lo *stimatore* per la grandezza $(\cdot)$, ovvero il valore che viene assegnato alla stessa sulla base della stima che ne viene effettuata a partire da un sotto-campione dell'intera popolazione (nella pratica, i campionamenti di x disponibili nel dataset) . Un generico stimatore $\widehat{\chi}$ della grandezza χ è detto *non polarizzato* se:

$$E[\widehat{\chi}] \to \chi \text{ per } N \to \infty \tag{A.1}$$

dove $E[\widehat{\chi}]$ indica il valore atteso di χ [cfr. eq. (A.5)] e N indica la dimensione (numero di elementi) del sotto-campione su cui la stima si basa.

Diamo quindi le seguenti definizioni:

Funzione Densità di Probabilità : indicata generalmente in inglese come *Probability Density Function* (p.d.f.), rappresenta la probabilità che la variabile aleatoria x assuma un valore nell'intervallo $[x, x + dx]$: *p.d.f.*

$$f(x)dx \doteq \text{probabilità di osservare } x \text{ nell'intervallo } [x, x + dx] \tag{A.2}$$

In accordo agli assiomi della probabilità definiti da Kolmogorov nel 1933, la p.d.f. $f(x)$ deve essere normalizzata a 1:

$$\int_S f(x) = 1 \tag{A.3}$$

Funzione Densità di Probabilità Cumulativa : indicata generalmente in inglese come *Cumulative Distribution Function* (c.d.f.), rappresenta la probabilità che la variabile aleatoria x assuma un valore minore o uguale a x; la c.d.f $F(x)$ è legata alla p.d.f. $f(x)$ in questo modo:

c.d.f.

$$F(x) = \int_{-\infty}^{x} f(x')dx' \tag{A.4}$$

dove $-\infty$ indica in generale l'estremo inferiore del dominio S della VA x.

Valore Atteso : indicato anche come *Valore di Aspettazione*, per una VA x la cui p.f.d. sia $f(x)$ è un operatore lineare definito come:

$$E[x] \doteq \int_S xf(x)dx \tag{A.5}$$

Si definiscono inoltre:

- *momenti algebrici di ordine n*:

$$\alpha_n \doteq E[x^n] = \int_S x^n f(x)dx \tag{A.6}$$

- *momenti centrali di ordine n*:

$$\mu_n \doteq E[(x-\alpha_1)^n] = \int_S (x-\alpha_1)^n f(x)dx \tag{A.7}$$

Valor Medio : al valore atteso di x si fa comunemente riferimento col termine di *valore medio della popolazione*, $m = E[x]$. Lo stimatore non polarizzato per il valor medio è dato dalla media campionaria:

$$\widehat{m} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{A.8}$$

Varianza : la varianza della popolazione è definita come:

$$Var[x] = \mu_2 = E[(x - E[x])^2] = \int_S (x-m)^2 f(x)dx \equiv \sigma^2 \tag{A.9}$$

Lo stimatore non polarizzato è dato da:

$$\widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \widehat{m})^2 \tag{A.10}$$

La radice quadrata della varianza, indicata come *deviazione standard* o *scarto quadratico medio* σ (s.q.m.), risulta particolarmente utile in quanto ha le stesse dimensione della variabile x.

s.q.m.

Skewness : è un coefficiente che misura la *simmetria* della p.d.f., definito come:

$$skw \doteq \frac{\mu_3}{\sigma^3} \tag{A.11}$$

Il corrispondente stimatore non polarizzato è dato da:

$$\widehat{skw} = \frac{N\sum_{i=1}^{N}(x_i - \widehat{m})^3}{(N-1)(N-2)\widehat{\sigma}^3} \tag{A.12}$$

Si tenga presenta che per una p.d.f. Gaussiana (o Normale) il coefficiente di skewness è nullo.

Kurtosis : è un coefficiente che fornisce informazioni sul *comportamento delle code* della p.d.f. definito come:

$$krt \doteq \frac{\mu_4}{\sigma^4} - 3 \tag{A.13}$$

Il corrispondente stimatore non polarizzato è dato da:

$$\widehat{krt} = \frac{N \sum_{i=1}^{N} (x_i - \widehat{m})^4}{(N-1)(N-2)(N-3)\widehat{\sigma}^4} - 3 \tag{A.14}$$

Anche in questo caso, per una p.d.f. Gaussiana (o Normale) il coefficiente di kurtosis è nullo. Se $krt > 0$ le code risultano più "alte" rispetto a quelle di una gaussiana, e le p.d.f. di questo tipo sono dette *leptocurtiche*; se $krt < 0$, le p.d.f. sono invece dette *platicurtiche*.

APPENDICE B

La Geostatistica: Richiami Teorici

La geostatistica si presenta come una collezione di strumenti sia deterministici che probabilistici atti alla tanto alla *comprensione* quanto alla *modellizzazione* di dati che manifestino un qualche tipo di CORRELAZIONE SPAZIALE — dati cui la letteratura fa solitamente riferimento col termine di VARIABILI REGIONALIZZATE.

La nascita di tale disciplina si fa generalmente risalire agli anni sessanta, quando si è sviluppata con l'intento di risolvere alcuni problemi in campo minerario; fu allora che Matheron la definì appunto *Geostatistica*: essa si presentava come una disciplina per così dire 'ibrida', che ricorreva agli strumenti messi a disposizione dall'ingegneria mineraria, dalla geologia, dalla matematica e dalla statistica. La sua forza, rispetto alle altre discipline 'classiche', risiede nella possibilità di prendere in considerazione la variabilità spaziale sia a *corto* che a *lungo* raggio, ovvero, in termini statistici, di considerare allo stesso tempo rispettivamente la *correlazione* spaziale e l'eventuale presenza di un *trend* — andamento sistematico (aumento o diminuzione) delle misure della variabile sotto studio nello spazio, su distanze medio-lunghe in relazione alle dimensioni dell'intera area geografica esaminata.

La geostatistica "tradizionale" si basa essenzialmente su una statistica a *due-punti*, considerando la correlazione spaziale solo tra coppie punti campionati; esistono diversi strumenti analitici per determinare tale correlazione, il più diffuso e tradizionale dei quali è il VARIOGRAMMA, indicato generalmente con $\gamma(h)$: per questo l'analisi della correlazione spaziale dei dati prende comunemente il nome di ANALISI VARIOGRAFICA (o Studio Variografico) — la parte centrale di ogni studio e modellizzazione in campo geostatistico.

Recentemente, si stanno sviluppando altre tecniche alternative che consentono di considerare anche statistiche relative a più di due punti correlati spazialmente, come l'approccio basato sulla teoria bayesiana della massima entropia (BME) {8}.

Pur essendo nata in relazione ad applicazioni in campo minerario, oggi la geostatistica si configura come un insieme di tecniche matematiche e statistiche adatto a tutti quei settori in cui l'elemento determinante sia la *continuità spaziale* del fenomeno sotto esame; rispetto alla statistica classica, infatti, il maggior contributo è stato quello di implementare delle tecniche

opportune per affrontare situazioni nelle quali risulti centrale la possibilità di considerare le due caratteristiche fondamentali dei dati geostatistici, che sono:

1. la *non-ripetitività*: per ogni localizzazione campionata esiste infatti solitamente una sola osservazione;
2. la *dipendenza*: in generale, valori osservati in localizzazioni differenti manifestano una qualche correlazione.

La geostatistica è quindi in grado di mettere a disposizione una serie di strumenti, di natura probabilistica, al fine di:

- costruire un *modello* (tipicamente di natura stocastica) in grado di riprodurre e descrivere la variabilità spaziale riscontrata sperimentalmente dai campionamenti disponibili;
- utilizzare questo modello per *stimare* i valori della variabile di interesse in localizzazioni non ancora campionate, descrivendo anche l'entità dell'*incertezza* che accompagna tali stime;
- ricorrere alle informazioni così ottenute per effettuare delle decisioni — come l'individuazione di zone a rischio, la stima della quantità di materiale utile presente in un giacimento minerario, ecc.

In questa appendice si descriveranno brevemente le fasi principali che caratterizzano uno studio di tipo geostatistico, fornendo di volta in volta gli strumenti matematici di base adatti allo specifico problema trattato; per approfondire da un punto di vista prettamente teorico tali aspetti, il testo di riferimento è quello di Cressie {13}; un approccio più diretto e "applicativo" si può invece trovare nei libri di Isaaks e Srivastava {27}, Goovaerts {21}, Deutsch e Journel {16}, Kanevski e Maignan {31}, Posa {43}.

B.1 Analisi Esplorativa dei Dati Spaziali (ESDA)

Nel caso si intenda intraprendere un'analisi di tipo geostatistico (ma questo andrebbe sperabilmente applicato in generale a ogni analisi di tipo statistico), familiarizzare con i dati osservati e soprattutto con il processo fisico che li ha generati risulta essere una fase fondamentale dell'intero studio: ogni modello viene infatti costruito sulla base delle informazioni che si possono estrarre dal dataset impiegato, e questo implica inevitabilmente che *qualunque errore sistematico sulle osservazioni si rifletterà in ogni successiva fase di analisi*; inoltre, caratterizzare la prima fase di studio con una conoscenza approfondita dei dati — preferibilmente condotta con l'aiuto delle persone coinvolte nella raccolta dei dati stessi — può risultare estremamente utile tanto per *scegliere* le tecniche e i modelli più appropriati quanto per *interpretare* nel modo corretto i risultati ottenuti.

L'analisi esplorativa dei dati si configura più come un *approccio* — e non tanto come una serie di tecniche — a *come* l'analisi dovrebbe essere condotta, e si differenzia rispetto alle tecniche classiche in quanto *postpone* le assunzioni sul tipo di modello che i dati si assume seguano, lasciando che i siano i dati stessi e "rivelare" la struttura del modello:

- per l'analisi classica, la sequenza è solitamente la seguente:

  $$\text{problema} \rightarrow \text{dati} \rightarrow \text{modello} \rightarrow \text{analisi} \rightarrow \text{conclusioni}$$

- per l'analisi esplorativa dei dati, la sequenza è invece:

  $$\text{problema} \rightarrow \text{dati} \rightarrow \text{analisi} \rightarrow \text{modello} \rightarrow \text{conclusioni}$$

Una volta che i dati siano stati accuratamente ripuliti da eventuali errori di misura, di trascrizione, ecc., con particolare attenzione ai valori estremi, gli elementi basilari dell'*analisi esplorativa dei dati spaziali* (Exploratory Spatial Data Analysis) includono solitamente: *ESDA*

- un'*analisi statistica* dei dati grezzi (raw) ed eventualmente di quelli trasformati; in questa fase, l'aspetto prettamente spaziale dei dati viene tipicamente ignorato — si applicano cioè le tecniche della statistica descrittiva classica (eventualmente multivariata), ricorrendo a istogrammi, stime di media e varianza, determinazione di tabelle di frequenza, ecc.;
- una *visualizzazione grafica* dei dati, ricorrendo a grafici sia 1D che 2/3D di vario tipo, quali, tra i più diffusi:
  - semplici *postplot*: i dati vengono visualizzati nello spazio che caratterizza l'area di studio, ad esempio per indagare la distribuzione spaziale dei campionamenti;
  - *postplot indicatori*: i dati vengono suddivisi in due o più categorie, alle quali vengono assegnati simboli differenti;
  - *postplot proporzionali*: la grandezza del simbolo che rappresenta un campionamento è proporzionale al valore della misura;
  - *curve di livello* per individuare la presenza di eventuali pattern sul territorio;
  - *proiezioni* lungo particolari direzioni per l'individuazione di eventuali andamenti sistematici;
  - ...;
- un'*analisi della rete di monitoraggio* (Monitoring Network), ossia una caratterizzazione della sua risoluzione spaziale e dimensionale ricorrendo a misure geometriche e topologiche [cfr. §B.1.1]; *MN*
- un'*analisi statistica con finestre mobili* (Moving Window Statistics), particolarmente indicata per la determinazione delle presenza di un eventuale *effetto proporzionale* — legame lineare tra media locale e varianza locale [cfr. §B.1.3]; costituisce il primo passo per un'analisi *spaziale* dei dati, per il riconoscimento di possibili legami tra medie e varianze locali e per la rivelazione della presenza di *eteroschedasticità*, situazione per la quale la variabilità del fenomeno non è costante sull'area d studio[1]; *MWS*
- un'*analisi variografica* dei dati, sia grezzi che trasformati, per avere un'idea dell'eventuale presenza di direzioni privilegiate per la continuità spaziale (tutti i modelli della geostatistica classica si basano infatti sulla misura della continuità spaziale [cfr. §B.3]).

B.1.1 Descrizione della Rete di Monitoraggio (MN)

Con il termine *campionamento* si identifica solitamente una serie di metodi per selezionare e analizzare una parte di 'universo', allo scopo di effettuare di conseguenza inferenze sull'intero 'universo', assicurando economia, velocità e in certe situazioni qualità e accuratezza {9}. In particolare, si definisce CAMPIONAMENTO SPAZIALE il processo per il quale si selezionano alcuni elementi da un'area oggetto di studio per i quali sono disponibili anche alcune informazioni riguardanti l'assetto geografico di tali elementi.

In linea di principio, sarebbe auspicabile che il campionamento spaziale fosse condotto secondo una *griglia regolare* ed equispaziata su tutto il dominio di studio; nella pratica, tuttavia, in

[1] Questo fenomeno, che è piuttosto comune nella pratica, e comunque sempre presente nel caso in cui si riveli un effetto proporzionale, può avere significative influenze sulla stima della variabile regionalizzata, indipendentemente dalla tecnica impiegata per la stima stessa.

relazione a vari fattori quali ad esempio la morfologia del territorio o la necessità di avere maggiori informazioni relativamente a una sotto-area specifica[2], i punti di campionamento risultano distribuiti nello spazio in maniera *inomogenea*.

Il problema dell'omogeneità della MN è strettamente legato a due aspetti fondamentali, quali:

1. il problema del CLUSTERING, ovvero del fatto che alcune aree possono essere caratterizzate da un sovra-campionamento rispetto alle altre [cfr. §B.1.2], che si ripercuote inoltre sull'accuratezza delle stime dei parametri statistici globali, quali media e varianza;

2. la capacità della MN di *rilevare* le caratteristiche del fenomeno sotto esame in relazione alla scala scelta — risoluzione della MN;

Esistono varie tecniche di analisi atte alla descrizione sia qualitativa che quantitativa della MN, al fine di valutarne l'*omogeneità* e determinare l'eventuale presenza di clustering; tra le più utilizzate, ricordiamo:

Poligoni di Voronoi [3]: dati N punti di campionamento, attorno a ogni punto p_i presente nell'area di studio S si costruisce un *poligono di influenza* in modo che tutti i punti del piano presenti al suo interno siano più vicini a p_i rispetto a tutti gli altri $N-1$ punti; intuitivamente, un punto p_i isolato avrà quindi un poligono di influenza con area grande, mente un punto in una zona affetta da clustering avrà un poligono di influenza con area piccola; le proprietà dei poligoni di Voronoi sono descritte in {31, pagg. 31–32}; una prima analisi visiva della zona di studio con rappresentati questi poligoni può già risultare sufficiente per individuare zone sovra-campionate e/o avere informazioni sulla omogeneità della MN, ma si possono condurre anche analisi quantitative, quali:

- *istogrammi delle aree dei poligoni*, che descrivono l'irregolarità spaziale della MN: per una rete perfettamente omogenea, ci si aspettano delle funzioni δ di Dirac (ovvero tutte le aree sono uguali), mentre all'aumentare della disomogeneità, gli istogrammi tendono ad allargarsi;
- *istogrammi della distanza tra i punti*, che riportano il numero di coppie caratterizzate da una determinata distanza in funzione della distanza stessa; l'informazione che se ne ricava è simile a quella appena descritta.

Queste misure di tipo *geometrico* forniscono una descrizione quantitativa della *risoluzione spaziale* della MN.

Indice di Morishita : tra i numerosi indici statistici utilizzati per determinare il clustering dei punti [cfr. ad esempio {13}], quello di Morishita prevede di suddividere l'intera area di studio con una griglia regolare di celle (tutte con la medesima dimensione) e di determinare il valore del seguente indice:

$$I_M = K \frac{\sum_{i=1}^{K} n_i(n_i-1)}{N(N-1)} \tag{B.1}$$

dove $n_i,\ i = 1,\ldots,K$ rappresenta il numero di punti nella cella i-esima e N il numero totale di celle; il *diagramma di Morishita* riporta la dipendenza dell'indice di Morishita in funzione della dimensione delle celle ed è direttamente collegato al clustering della MN in esame; la teoria prevede infatti che:

[2] Caso tipico è ad esempio quello relativo a indagini volte alla determinazione di possibili "zone a rischio", per le quali si avranno verosimilmente maggiori campionamenti nelle aree che hanno manifestato, dopo una prima campagna di misura di carattere esplorativo, valori elevati della variabile sotto esame.

[3] Ai quali si fa riferimento anche con termini quali "area di influenza" o "celle di Dirichlet".

- per dati con distribuzione *regolare*, la curva parta da zero per raggiunge in maniera monotona un valore pari a 1 (eventualmente con qualche fluttuazione);
- per dati con distribuzione *casuale*, la curva oscilli attorno al valore 1;
- nel caso di dati affetti da clustering, la curva ammetta valori maggiori di 1 per bassi valori della dimensione della cella, mentre al crescere di quest'ultima, fino a coprire l'intera area di studio, l'indice di Morishita tenda a 1.

Questo indice risulta inoltre molto utile per distinguere e/o confrontare tra loro differenti tipi di reti di monitoraggio.

Dimensione Frattale della Rete di Monitoraggio

Nel libro di Kanevski e Maignan {31}, pag. 35 si sottolinea il fatto che per avere una MN di buona qualità non è sufficiente avere una adeguata risoluzione spaziale; anche la *dimensione frattale* della MN dovrebbe essere prossima a quella ottimale.

La dimensione frattale caratterizza in un certo senso il tipo di fenomeno che la MN è in grado di rivelare: ad esempio, se lo scopo è quello di monitorare/misurare un fenomeno in uno spazio bi-dimensionale, allora la dimensione frattale della MN dovrebbe essere prossima a 2. Sono disponibili vari metodi per determinare questa dimensione, come riportato ad esempio in {31}.

B.1.2 Il Problema del Clustering

Spesso accade che vengano condotti ulteriori e successivi campionamenti in zone di particolare interesse, per esempio zone caratterizzate da valori molto alti o molto bassi della variabile in esame; risulta intuitivo che questo fenomeno, noto appunto come CLUSTERING, abbia delle ripercussioni sulla stima dei parametri globali della variabile di interesse: come conseguenza di ottiene che

> *una rete di monitoraggio affetta da clustering raccoglie dati che* non *sono realmente rappresentativi della popolazione da cui si postula siano stati estratti/campionati.*

Si consideri ad esempio il caso frequente in cui le zone con valori elevati (mettiamo di un certo inquinante) ricevano maggior attenzione e vengano quindi condotte ulteriori misure proprio in quelle zone: la media stimata da questo nuovo dataset 'aggiornato' sarà verosimilmente più alta di quella della popolazione, in quanto si sono aggiunti valori in qualche modo "polarizzati".

L'obiettivo delle tecniche di declustering è quindi quello di *ricostruire* l'informazione corretta tenendo conto sia dell'effetto del clustering sia del campionamento preferenziale. Per fare questo, la strada più efficacie è quella di applicare degli opportuni *pesi* ai dati grezzi: ogni dato grezzo verrà quindi moltiplicato per il peso che gli compete, e saranno questi nuovi dati a costituire il dataset rappresentativo della popolazione.

Tra gli approcci più diffusi nella pratica, ricordiamo:

Desclustering Casuale : l'area di interesse viene suddivisa in una griglia regolare di celle, e per ogni cella viene selezionato in modo casuale un dato [cfr. {14, pag.3}]; tra gli svantaggi di questo metodo, la perdita di una parte dei dati, la non-ripetibilità della procedura e la definizione di un criterio per la scelta della dimensione della cella; può risultare invece utile per suddividere il dataset in più parti, ognuna delle quali rispetta la distribuzione spaziale dei campionamenti;

Declustering a Celle : anche in questo caso l'area di studio viene suddivisa mediante una griglia regolare di celle, per ognuna delle quali viene determinata la media dei valori al suo interno

(ogni valore medio riceve lo stesso peso, pari a n_i^{-1}, con n_i il numero di campionamenti nella i-esima cella); successivamente, la media globale viene calcolata come media dei valori che caratterizzano ogni singola cella, i quali sono pesati inversamente al numero di valori presenti nella cella stessa; questo metodo è veloce, efficiente e ha il vantaggio di sfruttare tutti i dati disponibili; nel caso in cui il clustering fosse rilevante, la media globale così ottenuta dipenderà però dalle dimensioni della cella: Journel {28} propone di realizzare un grafico ' media globale ' *vs.* ' dimensione cella ' e, nel caso ad esempio che si sia effettuato un campionamento preferenziale in zone con valori elevati, scegliere la dimensione della cella cui è associato il valore minimo della media;

Declustering con Poligoni di Voronoi : il peso associato a ogni campionamento sarà proporzionale all'area di influenza del punto stesso; maggiore l'area del poligono di Voronoi, maggiore la rappresentatività del punto associato.

B.1.3 Analisi Statistica con Finestre Mobili (MWS)

Questo tipo di analisi prevede di suddividere il dominio D di interesse in sotto-regioni (celle) e di condurre delle indagini statistiche (di tipo classico) all'interno di ciascuna di esse; in questo modo, è possibile avere un'idea di come si distribuiscono nello spazio i principali parametri statistici che caratterizzano il dataset in esame, quali ad esempio media, varianza, skewness [cfr. eq. (A.11)] kurtosis [cfr. eq. (A.13)], e numero di punti all'interno di ogni cella di suddivisione.

La distribuzione spaziale del valore della media costituisce inoltre un test veloce e di facile implementazione per valutare se le ipotesi di stazionarietà [cfr. §B.2.3] risultino appropriate o meno.

Le *dimensioni* delle celle (che sono spesso rettangolari) sono dettate da una scelta di compromesso: dovrebbero essere sufficientemente *grandi* da garantire una statistica affidabile (e contenere quindi un numero di punti non troppo esiguo) e al contempo sufficientemente *piccole* da evidenziare la variabilità spaziale dei parametri indagati in relazione alla scala di interesse. Nel caso in cui i campionamenti non fossero molto numerosi, si può ricorrere alla tecnica delle *MW-sovrapposte*: la griglia delle celle viene spostata sull'area di studio lungo le direzioni principali (x e y nel caso bi-dimensionale), in modo da variare, in run successivi, l'insieme dei punti che cadono all'interno di ogni cella.

Una questione importante in ambito geostatistico, per le ripercussioni che inevitabilmente comporta sulle procedure di stima, è quella del legame esistente tra *media* e *varianza* locali. Nel caso di `omoschedasticità`, questi due parametri non evidenziano alcun legame significativo tra loro— in un diagramma ' media ' *vs.* ' varianza ' i punti si distribuiscono in maniera casuale; nel caso, assai frequente, di `eteroschedasticità`, invece, la variabilità del fenomeno è legata alla media locale — quando tale legame esiste ed è di natura lineare tra media e deviazione standard, si fa solitamente riferimento a esso col termine di *Effetto Proporzionale*.

Effetto Proporzionale

Quando, in un diagramma ' media ' *vs.* ' deviazione standard ' ottenuto da un'analisi condotta mediante finestre mobili, si nota una dipendenza lineare tra i due parametri, si parla di presenza di Effetto Proporzionale, ovvero la variabilità locale del fenomeno è legata alla media locale dello stesso: più alta risulta la media, maggiore risulta la variabilità.

La presenza di questo fenomeno rende tra l'altro meno affidabile l'applicazione delle ipotesi di stazionarietà [cfr. §B.2.3], visto che la media non è costante su tutto il dominio di interesse, e porta a difficoltà nelle procedure di stima, indipendentemente dal modello impiegato — il

grado di variabilità va comunque sempre raffrontato alle dimensioni del dominio D e al grado di correlazione spaziale proprio dei dati in esame.

Nel caso in cui i dati manifestassero una distribuzione di probabilità di tipo *log-normale*, come accade spesso per dati relativi alle misure di concentrazione di Radon Indoor e/o inquinanti ambientali in generale, allora c'è da aspettarsi la presenza di un effetto proporzionale su base teorica.

Si consideri infatti una variabile y con distribuzione Gaussiana, media μ e varianza β^2; si consideri quindi una seconda variabile $x = e^y$: allora x avrà una distribuzione di probabilità *log-normale*, data da

$$f(x;\mu,\beta^2) = \frac{1}{\sqrt{2\pi\beta^2}}\frac{1}{x}\exp\left[-\frac{(\log x-\mu)^2}{2\beta^2}\right] \tag{B.2}$$

I primi due momenti della (B.2) sono dati rispettivamente da:

$$E[x] = \exp\left(\mu+\frac{1}{2}\beta^2\right) \equiv m \tag{B.3}$$

$$Var[x] = \exp(2\mu+\beta^2)[\exp(\beta^2)-1] \equiv \sigma^2 \tag{B.4}$$

È quindi facile notare come il legame tra il valor medio m e la deviazione standard σ della variabile x, log-normale, sia di tipo *lineare*, e precisamente:

$$\frac{\sigma}{m} = \left(e^{\beta^2}-1\right)^{\frac{1}{2}} \tag{B.5}$$

B.2 Trattamento Geostatistico dei Dati

Al fine di ottenere delle stime in localizzazioni non campionate, è necessario disporre di un *modello* del fenomeno sotto esame; l'approccio geostatistico alle procedure di stima rende esplicito il modello su cui le stesse si basano.

Sfortunatamente, nella pratica sono molto pochi i processi le cui caratteristiche fisico-chimico-geologiche siano conosciute con sufficiente accuratezza da rendere applicabile un modello *deterministico* del fenomeno, in quanto i fenomeni naturali sono generalmente caratterizzati da un gran numero di variabili (spesso 'nascoste') che interagiscono tra loro in maniera molto complessa; per questo

> *il processo fisico sotto esame può essere così complesso, e la nostra conoscenza di esso così limitata/parziale, che il suo comportamento può apparire* casuale*: questo non vuol certo dire che lo sia intrinsecamente, piuttosto riflette la nostra ignoranza a riguardo.*

B.2.1 La Necessità di un Modello per i Dati Spaziali

La teoria dei PROCESSI STOCASTICI riconosce proprio questa componente fondamentale di incertezza, e mediante un approccio di tipo *probabilistico* è in grado di affrontare non solo il problema della `stima` dei valori incogniti, ma anche di determinare l'`accuratezza` di tale stima.

Matheron ha definito VARIABILI REGIONALIZZATE (VR) le variabili di tipo spaziale, per enfatizzare i due aspetti fondamentali, e tra loro complementari, che le caratterizzano: *VR*

- un aspetto *casuale*, che tiene conto delle irregolarità locali del fenomeno;

- un aspetto *strutturale*, che riflette le caratteristiche globali del fenomeno (quello che è stato in precedenza chiamato "trend");

Il modello fondamentale per una variabile regionalizzata $Z(\mathbf{u})$ osservata nella localizzazione $\mathbf{u}$, dove $\mathbf{u}$ rappresenta, nel caso bi-dimensionale, il vettore di componenti spaziali (u_x, u_y), viene scritto nel modo seguente:

$$Z(\mathbf{u}) = m(\mathbf{u}) + R(\mathbf{u}) \tag{B.6}$$

questo modello probabilistico assume che la componente casuale sia descritta da $R(\mathbf{u})$ e quella strutturale da $m(\mathbf{u})$.

Si postula quindi che in ogni localizzazione $\mathbf{u}_i$, il valore osservato $z_k(\mathbf{u}_i)$ della variabile campionata sia la realizzazione di una variabile aleatoria (VA) $Z_k(\mathbf{u}_i)$, il cui valore di aspettazione (o valore atteso), indicato con $E[\cdot]$, vale proprio

VA

$$E[Z_k(\mathbf{u}_i)] = m(\mathbf{u}_i) \tag{B.7}$$

In un generico punto $\mathbf{u}$ in cui la variabile regionalizzata non è stata campionata, il valore $z(\mathbf{u})$ non è noto, tuttavia risulta ben definito; può infatti venir interpretato come una realizzazione della VA $Z(\mathbf{u})$.

L'intera famiglia delle variabili aleatorie $Z(\mathbf{u})$, con $\mathbf{u} \in D, D$ dominio/area di studio, prende il nome di FUNZIONE ALEATORIA (FA), o PROCESSO STOCASTICO; tipicamente, la definizione di FA è limitata a VA collegate a uno stesso attributo, sia z, così che un'altra FA $W(\mathbf{u}), \mathbf{u} \in D$ sarà definita per trattare un differente attributo w del fenomeno.

FA

Così come una singola VA è caratterizzata dalla propria funzione di *distribuzione di probabilità* (Probability Density Function) o funzione di *distribuzione di probabilità cumulativa* (Cumulative Density Function), una FA $Z(\mathbf{u})$ verrà caratterizzata dal rispettivo set di tutte le N-variate c.d.f. (o p.d.f.) per ogni numero N e ogni scelta delle localizzazioni $\mathbf{u}_n,\ n = 1, \ldots, N$:

p.d.f.
c.d.f.

$$F(\mathbf{u}_1, \ldots, \mathbf{u}_N; z_1, \ldots, z_N) = \text{prob}\{Z(\mathbf{u}_1) \leq z_1, \ldots, Z(\mathbf{u}_N) \leq z_N\} \tag{B.8}$$

dove $F(\mathbf{u}_1, \ldots, \mathbf{u}_N; z_1, \ldots, z_N)$ rappresenta la c.d.f. *congiunta* della FA $Z(\mathbf{u})$. Ancora, così come la c.d.f. di una singola VA $Z(\mathbf{u})$ è utilizzata per modellizzare l'incertezza associata al valore $z(\mathbf{u})$, la c.d.f. multivariata (B.8) è impiegata per modellizzare l'incertezza congiunta associata agli N valori $\{z(\mathbf{u}_1), \ldots, z(\mathbf{u}_N)\}$.

> *I dati grezzi, ovvero gli N valori sperimentali $z_k(\mathbf{u}_i)$ accessibili, sono quindi considerati come una particolare realizzazione della Funzione Aleatoria $Z(\mathbf{u})$.*

Ci sono due importanti aspetti dei dati di tipo spaziale che ne rendono complicata la trattazione statistica:

1. spesso si ha a disposizione un'*unica* realizzazione della FA[4];
2. i dati *non* sono *indipendenti* e identicamente distribuiti nello spazio, come invece è prassi assumere in un approccio statistico di tipo convenzionale;

per questo, al fine di garantirsi la possibilità di poter ottenere una qualche *inferenza statistica*, sarà necessario accettare alcune ipotesi che saranno descritte in seguito [cfr. §B.2.3].

[4] Questo impedisce tra l'altro di poter effettuare inferenza statistica su tutte le funzioni finito-dimensionali del tipo (B.8); si rendono quindi indispensabili ulteriori assunzioni, al fine di ridurre il numero di parametri da cui dipende la FA.

B.2.2 Momenti di una Funzione Aleatoria (FA)

In geostatistica lineare, della FA si richiedono solamente i primi due momenti, ovvero:

- *Valore Atteso*, o momento del primo ordine:

$$E[Z(\mathbf{u})] = m(\mathbf{u}) \tag{B.9}$$

- *Momenti del secondo ordine*; sono tre i momenti del secondo ordine più comunemente utilizzati in geostatistica:
  1. *Varianza*:

$$\begin{aligned} Var[Z(\mathbf{u})] &= E[\{Z(\mathbf{u}) - m(\mathbf{u})\}^2] \\ &= E[Z^2(\mathbf{u})] - m^2(\mathbf{u}) \end{aligned} \tag{B.10}$$

  2. *Covariogramma* o Covarianza:

$$\begin{aligned} C(\mathbf{u}, \mathbf{u}') &= E[\{Z(\mathbf{u}) - m(\mathbf{u})\}\{Z(\mathbf{u}') - m(\mathbf{u}')\}] \\ &= E[Z(\mathbf{u})Z(\mathbf{u}')] - E[Z(\mathbf{u})]E[Z(\mathbf{u}')] \end{aligned} \tag{B.11}$$

  3. *Variogramma*:

$$2\gamma(\mathbf{u}, \mathbf{u}') = Var[Z(\mathbf{u}) - Z(\mathbf{u}')] \tag{B.12}$$

     la funzione γ viene generalmente denominata come *semi-variogramma*, ma nella pratica si ricorre indifferentemente anche al termine 'variogramma' — le due funzioni differiscono infatti solo per un fattore 2.

B.2.3 Inferenza e Ipotesi di Stazionarietà

L'inferenza di qualsiasi parametro statistico che caratterizza la c.d.f. (B.8) di una FA $Z(\mathbf{u})$ richiede che siano disponibili campionamenti ripetuti della variabile $z(\mathbf{u})$: nella pratica, come già evidenziato, si ha però generalmente accesso a un solo valore di $z(\mathbf{u})$, fissato $\mathbf{u}$. Inoltre, dalle definizioni (B.11) e (B.12) si nota come sia necessario avere accesso a diverse realizzazioni anche delle coppie $\{Z(\mathbf{u}), Z(\mathbf{u}')\}$, cosa che nelle applicazioni rende l'inferenza statistica impossibile.

Il problema è aggirato con delle ipotesi aggiuntive che, è bene sottolineare, appartengono al *modello*, e non al reale processo fisico che si vuole studiare; ad esempio, se variogramma e covariogramma dipendono solamente dal vettore $\mathbf{h} = \mathbf{u} - \mathbf{u}'$, è possibile effettuare l'inferenza statistica considerando che ogni coppia di dati il cui vettore di separazione è proprio $\mathbf{h}$ (si vedrà che nella pratica si accetta una certa tolleranza su questa distanza!) può essere considerata una diversa realizzazione della coppia di VA $\{Z(\mathbf{u}), Z(\mathbf{u}')\}$.

Stazionarietà del Secondo Ordine

Un processo stocastico, o FA, è definito STAZIONARIO DEL SECONDO ORDINE se:

$\boxed{\mathcal{H}1}$

1. il momento del primo ordine (B.9) esiste e non dipende dal punto $\mathbf{u}$:

$$\exists\ E[Z(\mathbf{u})] \text{ t.c. } E[Z(\mathbf{u})] = m,\ \forall \mathbf{u} \in D \tag{B.13}$$

2. il covariogramma può essere scritto come:

$$\begin{aligned} C(\mathbf{h}) = E[Z(\mathbf{u}+\mathbf{h})Z(\mathbf{u})] - m^2, \\ \forall\ (\mathbf{u}, \mathbf{u}+\mathbf{h}) \in D \end{aligned} \tag{B.14}$$

La stazionarietà del covariogramma implica la stazionarietà della varianza e del variogramma; è facile verificare che:

$$Var[Z(\mathbf{u})] = C(0) \tag{B.15}$$

$$\gamma(\mathbf{h}) = C(0) - C(\mathbf{h}) \tag{B.16}$$

in particolare, la (B.16) implica che, sotto le ipotesi $\mathcal{H}1$, variogramma e covariogramma sono due strumenti equivalenti per la descrizione della correlazione spaziale dei dati.

Si può inoltre definire un terzo strumento, il *correlogramma*:

$$\rho(\mathbf{h}) = \frac{C(\mathbf{h})}{C(0)} = 1 - \frac{\gamma(\mathbf{h})}{C(0)} \tag{B.17}$$

L'esistenza della funzione variogramma rappresenta un'ipotesi *più debole* rispetto all'esistenza della funzione covariogramma: esistono infatti molti processi fisici che non ammettono varianza e covariogramma, ma ammettono un variogramma; l'ipotesi di stazionarietà del secondo ordine può quindi essere indebolita, assumendo la sola esistenza e stazionarietà del variogramma {43, pag. 23}.

Le Ipotesi Intrinseche

Un processo stocastico, o FA, è definito INTRINSECO se:

$\boxed{\mathcal{H}2}$

1. il momento del primo ordine (B.9) esiste e non dipende dal punto $\mathbf{u}$:

$$\exists\ E[Z(\mathbf{u})] \text{ t.c. } E[Z(\mathbf{u})] = m,\ \forall \mathbf{u} \in D \tag{B.18}$$

2. gli incrementi $\{Z(\mathbf{u}+\mathbf{h}), Z(\mathbf{u})\}$ ammettono una varianza finita che non dipende da $\mathbf{u}$:

$$Var[Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u})] = 2\gamma(\mathbf{h}), \quad \forall\ (\mathbf{u}, \mathbf{u}+\mathbf{h}) \in D \tag{B.19}$$

a dire che la stazionarietà del secondo ordine è limitata agli incrementi della FA $Z(\mathbf{u})$.

> *Sotto queste ipotesi, differenti regioni del dominio di interesse D possono essere considerate come differenti realizzazioni del processo stocastico $Z(\mathbf{u})$* {31, pag. 64}.

B.3 Analisi Variografica

L'analisi variografica, ovvero lo studio della continuità spaziale propria dei dati, costituisce un punto fondamentale in ogni studio di tipo geostatistico, in quanto è in questa fase che tipicamente l'utente sceglie il *modello* di variogramma che costituirà il parametro fondamentale impiegato nella successiva fase di stima mediante il metodo del kriging [cfr. §B.4].

La continuità spaziale può venir descritta mediante l'aiuto di differenti misure e indici, il più noto e usato dei quali è appunto il variogramma $\gamma(\mathbf{h})$. Una rappresentazione grafica che in molti casi può rivelarsi utile, soprattutto in ambito *esplorativo*, è la cosiddetta VARIOGRAM CLOUD[5].

[5] Si è fatto ampio uso di questo strumento nella serie di analisi descritte nel capitolo 4.3, a pagina 54.

Questa è costituita da un grafico del tipo ' $[Z(\mathbf{u}) - Z(\mathbf{u}')]^2$ ' *vs.* ' $\mathbf{h}$ ', e la nuvola formata da queste coppie di punti può rivelare la presenza di eventuali outliers che possono pertanto dominare la stima del variogramma sperimentale; può inoltre mettere in evidenza situazioni per le quali la distribuzione di $[Z(\mathbf{u}) - Z(\mathbf{u}')]^2$ per un certo lag $\mathbf{h_i}$ sia pesantemente a-simmetrica, caso in cui la media aritmetica per il calcolo del valore di $\gamma(\mathbf{h_i})$ potrebbe condurre a una stima non ottimale e/o affidabile.

L'analisi variografica può essere scissa in due fasi principali:

1. *stima* del variogramma e *interpretazione* della misura della continuità spaziale così ottenuta, ricorrendo ai dati grezzi o a dati opportunamente trasformati con l'intento di rendere l'interpretazione più 'leggibile';

2. *modellizzazione* della struttura spaziale identificata al punto 1 mediante modelli teorici di variogrammi — nella pratica, questo si traduce in una operazione di fitting del variogramma sperimentale con modelli teorici descritti da opportune funzioni analitiche [cfr. §B.3.3].

B.3.1 Proprietà del Variogramma

Benché la letteratura dei processi stocastici preferisca l'impiego della covarianza (B.11) rispetto al variogramma (B.12), in geostatistica la correlazione spaziale viene comunemente descritta dal variogramma[6]: la motivazione di questa scelta va ricercata sia nel lavoro svolto da Matheron {38}, sia nel fatto che, come già ricordato, esso richiede che siano soddisfatte solo le ipotesi intrinseche $\mathcal{H}2$.

Da un punto di vista fisico, il variogramma descrive la `diversità spaziale` dei dati: quando i dati sono correlati tra loro (tipicamente per piccole distanze relative), il valore del variogramma risulta piuttosto basso, mentre quando i dati sono scorrelati (e tipicamente, per grandi distanze relative), il valore di $\gamma(\mathbf{h})$ sarà piuttosto elevato.

Bisogna ricordare che se la FA soddisfa le ipotesi intrinseche, la funzione variogramma $\gamma(\mathbf{h})$ deve essere una funzione *definita positiva in modo condizionato* {6}, cioè:

$$-\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(u_i - u_j) \geq 0 \qquad \text{con} \qquad \sum_{i=1}^{n}\lambda_i = 0 \tag{B.20}$$

e questo deve valere qualunque siano i punti di supporto $(u_1, \ldots, u_n) \in D$.

Dalla (B.20) segue che:

- *ogni combinazione lineare di variogrammi con coefficienti positivi è ancora un variogramma*, ovvero:

$$\gamma(\mathbf{h}) = \sum_{i=1}^{n}\alpha_i\gamma_i(\mathbf{h}) \qquad \forall\ \alpha \geq 0 \tag{B.21}$$

- *al crescere di* $\mathbf{h}$*, il variogramma deve necessariamente crescere in modo più lento di* $|\ \mathbf{h}\ |^2$, ovvero:

$$\lim_{|\mathbf{h}|\to\infty}\frac{\gamma(\mathbf{h})}{|\ \mathbf{h}\ |^2} = 0 \tag{B.22}$$

[6] Anche se, come sarà chiaro tra breve, le equazioni del kriging impiegate nelle procedure di stima sono scritte in termini di matrici di covarianze per semplicità di calcolo.

Si passeranno ora brevemente in rassegna alcune proprietà analitiche della funzione variogramma, con l'intento di mettere in evidenza le caratteristiche che risultano importanti al fine della modellizzazione del variogramma sperimentale [cfr. §B.3.3].

Nel caso bi-dimensionale (che è quello più comune nella pratica, e comunque quello che caratterizza questo lavoro di tesi), è consuetudine, soprattutto per ragioni di praticità, *scomporre* la dipendenza spaziale di $\gamma(\mathbf{h})$ nelle due componenti di un sistema di riferimento polare, considerando quindi come coordinate spaziali il modulo h del vettore $\mathbf{h}$ (cui si fa riferimento col termine LAG) e l'angolo polare α; in questo modo

$$\gamma(\mathbf{h}) = \gamma(h_x, h_y) = \gamma(h, \alpha)$$

Fissato quindi l'angolo α, un diagramma 'variogramma' *vs.* 'lag' fornisce informazioni sulla diversità dei campionamenti al crescere della loro distanza relativa; fissato un lag, diagrammi 'variogramma' *vs.* 'α' danno informazioni sulla direzionalità del fenomeno in esame.

In generale, il grafico di $\gamma(h)$ in funzione di h presenta questo tipo di andamento:

- parte nei pressi dell'origine degli assi (da un punto di vista teorico, $\gamma(0) = 0$);
- cresce all'aumentare del lag h (i campionamenti sono in generale meno correlati);
- cresce indefinitamente (evidenziando quindi la presenza di un trend) oppure si attesta su un certo valore di plateau (noto in letteratura come valore di sella, o 'sill' in inglese).

Le informazioni principali sul fenomeno di studio che si possono estrarre dall'analisi variografica, e che sono quindi strettamente legate alla successiva fase di modellizzazione, sono essenzialmente contenute in:

andamento in prossimità dell'origine : si ottengono informazioni sulla *continuità* e sulla *regolarità* della variabile regionalizzata; si distinguono quattro diverse situazioni caratteristiche:

- comportamento *parabolico*: la VR è continua e differenziabile nel senso della media quadratica; tale comportamento è di solito associato a un fenomeno molto regolare (e raro nella pratica!);
- comportamento *lineare*: la VR è continua, ma non differenziabile; il fenomeno è quindi meno regolare di quello precedente;
- *effetto nugget* o effetto pepita[7]: la VR presenta una discontinuità nell'origine, ovvero:

$$\lim_{|\mathbf{h}|\to 0} \gamma(\mathbf{h}) = C_0 \neq 0, \quad \text{con } \gamma(0) = 0$$

  il fenomeno in questione non è continuo e molto *irregolare*; può essere legato alla presenza di `micro-strutture`, ossia componenti del fenomeno su scale inferiori a quelle caratteristiche della MN, a errori di campionamento o di localizzazione; questo effetto, spesso presente nella pratica dell'analisi, ha importanti ripercussioni anche sulla stima mediante kriging — quello che importa in realtà è il rapporto, indicato come *nugget relativo*, tra C_0 e l'eventuale valore di sella raggiunto dal variogramma;
- *curva piatta*: la VR non presenta alcuna struttura significativa, a dire che $Z(\mathbf{u})$ e $Z(\mathbf{u}+\mathbf{h})$ non sono correlate, indipendentemente dalla distanza relativa $\mathbf{h}$ della coppia di punti considerata.

[7] Storicamente, il variogramma calcolato per un deposito aurifero presentava questo fenomeno, da cui il nome stesso — in inglese, infatti, "nugget" significa appunto "pepita".

andamento per $h \to \infty$: si ottengono informazioni sulla presenza o meno di un valore di SELLA, definito in questo modo: se $\exists\ \gamma_S(h)$ t.c. $\gamma(h) \leq \gamma_S(h),\ \forall\ h \geq a$, allora γ_S è detto *valore di sella* del variogramma, mentre a, che in generale sarà un $a(\alpha)$, è detto RANGE del variogramma.
Se il variogramma ammette un valore di sella, significa che oltre un lag pari al range a, le variabili $Z(\mathbf{u})$ e $Z(\mathbf{u}+\mathbf{h})$ non sono più correlate; inoltre, sotto le ipotesi intrinseche $\mathcal{H}2$, è facile dimostrare che $\gamma_S = \sigma^2$, dove σ^2 rappresenta la varianza della FA $Z(\mathbf{u})$ — a tale varianza si fa generalmente riferimento col termine di VARIANZA A PRIORI.

Anisotropie

Se $\gamma(\mathbf{h}) = \gamma(h, \alpha)$ non varia con la direzione, allora la variabile regionalizzata è detta *isotropa*; molto più comune, nella pratica, è invece la situazione per la quale il variogramma evidenzia una dipendenza dall'angolo polare α: in questo caso, la VR è detta *anisotropa*.

Le eventuali anisotropie di $\gamma(\mathbf{h})$ sono usualmente classificate in due categorie:

- *Anisotropia Geometrica*: il variogramma manifesta la stessa sella in tutte le direzioni, ma questa è raggiunta a differenti valori del range $a(\alpha)$; questo tipo di anisotropia può essere rappresentata da un solo modello di variogramma, in quanto una trasformazione affine delle coordinate può rendere il variogramma di partenza isotropo [cfr. {43, pagg. 33–35}, o {31, pag. 76}];

- *Anisotropia Zonale*: il valore di sella $\gamma_S(\alpha)$ cambia con la direzione, ma il range rimane costante; la modellizzazione di questo tipo di anisotropia non è banale, e richiede l'utilizzo di una somma di modelli 'semplici' — la (B.21) garantisce che si otterrà ancora un variogramma — ognuno dei quali sarà caratterizzato da una propria anisotropia geometrica e dal proprio valore di sella.

Nelle applicazioni, la situazione più frequente è comunque quella in cui sono presenti entrambi i tipi di anisotropie appena descritti.

B.3.2 Stima del Variogramma Sperimentale

Per la *stima* della correlazione spaziale di variabili regionalizzate sono disponibili varie funzioni differenti; la scelta di quale di queste impiegare è dettata solitamente dal livello di accuratezza richiesto (compatibilmente con la qualità e numerosità dei dati a disposizione) e dal *tipo* di informazione che meglio viene messo in luce (e/o che si intende ricercare). Il prototipo di questa famiglia di misure è dato dal seguente *stimatore*:

$$\widehat{\chi}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} \mid z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h}) \mid^p, \quad \forall\ p > 0 \tag{B.23}$$

dove $N(\mathbf{h})$ rappresenta il numero totale di coppie di punti sperimentali separati dal vettore $\mathbf{h}$, $z(\cdot)$ il valore misurato nella localizzazione $(\cdot)$, $\chi(\cdot)$ la funzione che descrive la correlazione spaziale. Nella pratica, $p \in (0, 2]$ e se $p \to 0$, la stima risulta più robusta alla presenza di eventuali valori anomali.

Se $p = 1$, si ottiene il cosiddetto MADOGRAMMA, mentre per $p = 1/2$ si ottiene il cosiddetto RODOGRAMMA: queste due misure risultano particolarmente utili per investigare le strutture su larga scala, ma il loro impiego è consigliabile solo per indagini di tipo qualitativo.

Nel caso in cui $p = 2$, si ottiene invece il ben più noto e diffuso VARIOGRAMMA (o più precisamente, semi-variogramma), il cui stimatore è quindi dato da:

$$\widehat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})]^2 \tag{B.24}$$

Si possono individuare tre ragioni principali per le quali la geostatistica preferisce lo stimatore (B.24) rispetto ad esempio al covariogramma:

1. l'insieme dei processi spaziali che soddisfano le ipotesi intrinseche $\mathcal{H}2$ — per i quali il covariogramma non è definito — contiene l'insieme dei processi che soddisfano le ipotesi di stazionarietà del secondo ordine $\mathcal{H}1$;
2. $\widehat{\gamma}(\mathbf{h})$ non cambia se si aggiunge una costante arbitraria alla FA $Z(\mathbf{u})$ — cosa che non accade per il covariogramma;
3. $\widehat{\gamma}(\mathbf{h})$ non viene polarizzato dalla presenza di un trend costante su D — lo stimatore del covariogramma risulta invece polarizzato.

Problemi di Natura Pratica

È da sottolineare innanzitutto che la disomogeneità della rete mi monitoraggio, la presenza di zone con valori molto elevati, la presenza di outliers (punti isolati con valori particolarmente anomali), l'esistenza di una variabilità del fenomeno a differenti scale, ecc., sono fattori che possono rendere molto complessa tanto la fase di stima della continuità spaziale quanto la successiva fase di interpretazione. Inoltre, durante la fase di stima, la natura discreta del dataset deve essere presa in considerazione con i dovuti accorgimenti.

Basta infatti osservare l'espressione dello stimatore del variogramma (B.24) per rendersi conto che raramente — per non dire mai! — nella pratica si avrà accesso a un numero statisticamente significativo di coppie di campionamenti $\{z(\mathbf{u}_i), z(\mathbf{u}_j)\}$ distanti *esattamente* $\mathbf{h}$: questo implica che la stima del variogramma applicando 'alla lettera' la (B.24) risulta statisticamente inefficiente.

Per risolvere questo problema, computazionalmente si introducono delle *tolleranze* sia sul lag h che sull'angolo polare α, in modo che tutte le coppie $\{z(\mathbf{u}_i), z(\mathbf{u}_j)\}$ per le quali $h \in [h - \frac{\Delta h}{2}, h + \frac{\Delta h}{2}]$ e $\alpha \in [\alpha - \frac{\Delta\alpha}{2}, \alpha + \frac{\Delta\alpha}{2}]$ prendano parte all'insieme considerato in $N(\mathbf{h}) = N(h, \alpha)$ — in questo modo, la numerosità statistica viene recuperata[8].

È buona norma iniziare la stima del variogramma considerando il variogramma *isotropo*, ovvero una tolleranza angolare pari all'angolo giro: in questo modo, la numerosità di $N(\mathbf{h})$ è elevata, la statistica che si ottiene più affidabile, e generalmente la curva di $\widehat{\gamma}(\mathbf{h})$ abbastanza regolare. Se il variogramma isotropo non presenta una struttura interpretabile, certo non potrà andare meglio con quelli direzionali, i quali, per i motivi appena esposti, sono stimati in maniera meno efficiente — la tolleranza $\Delta\alpha$ limita il numero di punti per la statistica.

Ricordiamo che l'introduzione di queste tolleranze consente inoltre di:

- indagare il comportamento di $\widehat{\gamma}(\mathbf{h})$ al variare del numero di lag h: questo può essere utile per ridurre il rumore (eccessive fluttuazioni) che spesso caratterizza i variogrammi sperimentali;
- indagare il comportamento di $\widehat{\gamma}(\mathbf{h})$ al variare della tolleranza Δh: questo può risultare utile per individuare ad esempio la presenza di clustering — se diminuendo la tolleranza sul lag il valore del variogramma non cambia significativamente, questo è un buon indice della presenza di clustering;

[8] Journel e Huijbregts {29} sostengono che per avere una stima efficiente, il numero di punti su cui calcolare il variogramma dovrebbe essere tale che $N(\mathbf{h}) \geq 30$.

- indagare nel dettaglio l'effetto nugget, utilizzando dei lag h non uniformi, ma più fitti nei pressi dell'origine — qualora la numerosità e la risoluzione della MN lo consentano;
- indagare nel dettaglio la direzionalità del fenomeno, agendo in modo opportuno sulla tolleranza angolare $\Delta\alpha$.

Quindi, dopo aver condotto l'analisi variografica — fase fondamentale di ogni analisi geostatistica, e probabilmente anche quella che richiede il maggior tempo di studio, essendo caratterizzata tipicamente da un modo di procedere 'trial & error' — si è in possesso di un variogramma sperimentale *discreto*, costituito da un numero finito di coppie $\{\widehat{\gamma}(h), h\}$.

Se il dataset e la natura del fenomeno lo consentono, si può costruire una superficie variografica ricorrendo a più variogrammi direzionali; superfici di questo tipo risultano particolarmente utili per indagare le anisotropie del fenomeno in esame. La risoluzione angolare di tale superficie sarà legata alla tolleranza $\Delta\alpha$, il cui valore sarà una scelta di compromesso: minore la tolleranza, maggiore la risoluzione, ma maggiore anche il rumore introdotto dal minor numero di punti disponibili.

B.3.3 Modelli Teorici

Sono sostanzialmente due le ragioni principali che rendono necessaria la costruzione e il successivo impiego di un *modello* di variogramma:

1. la stima e le simulazioni di tipo geostatistico richiedono che il valore del variogramma sia noto per ogni distanza di interesse, mentre quello sperimentale lo è solo per determinati valori del lag h;
2. perché la soluzione del sistema del kriging [cfr. §B.4.2 e §B.4.3] ammetta una soluzione unica, è necessario che la funzione variogramma sia *definita positiva*.

Per questi motivi, il variogramma ottenuto sperimentalmente deve essere *fittato* con delle opportune funzioni, le cui proprietà analitiche siano tali da rendere la matrice delle covarianze impiegata nella stima mediante kriging [cfr. eq. (B.42)] *definita positiva*; tali funzioni prendono il nome di *modelli teorici di variogrammi*.

Tra i modelli più diffusi e largamente impiegati, ricordiamo:

Modello Effetto Nugget : introdotto per modellizzare eventuali discontinuità nell'origine (o meglio, nelle sue immediate vicinanze), ha la seguente forma analitica:

$$\gamma_0(h) = \begin{cases} c & \text{per } h = 0 \\ 0 & \text{per } h \neq 0 \end{cases} \tag{B.25}$$

Modello Sferico : probabilmente il più usato nella pratica, manifesta un comportamento lineare nei pressi dell'origine e assume un valore costante pari a c per distanze maggiori o uguali al valore di a, distanza per la quale si raggiunge il valore di sella; la formula analitica è data da:

$$\gamma(h) = \begin{cases} c \cdot \left[1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right] & \text{per } h \leq a \\ c & \text{per } h \geq a \end{cases} \tag{B.26}$$

In fase di fitting di un modello, può essere utile ricordare che *la tangente nell'origine raggiunge il valore di sella c a circa i 2/3 del range a.*

Modello Esponenziale : questo modello raggiunge il valore di sella c in maniera asintotica, per cui si è soliti definire il range a come un range 'effettivo', ovvero distanza per la quale il variogramma assume un valore pari al 95% di c; la formula analitica è data da:

$$\gamma(h) = c \cdot \left[1 - \exp\left(-\frac{3h}{a}\right)\right] \tag{B.27}$$

In fase di fitting di un modello, può essere utile ricordare che *la tangente nell'origine raggiunge il valore di sella c a circa 1/5 del range 'effettivo' a.*

Modello Gaussiano : utilizzato solitamente per modellizzare variabili regionalizzate estremamente continue, ha anche il vantaggio di consentire, in fase di kriging, di ottenere dei pesi negativi e quindi di ottenere dei valori stimati maggiori del valore massimo dei punti campionati e minori di quello minimo; anche questo modello raggiunge il valore di sella c in maniera asintotica — e valgono pertanto le discussioni fatte in relazione al modello esponenziale; la formula analitica è data da:

$$\gamma(h) = c \cdot \left[1 - \exp\left(-\frac{(3h)^2}{a^2}\right)\right] \tag{B.28}$$

I modelli appena descritti ammettono tutti un valore di sella, ma ci sono nella pratica situazioni per le quali il variogramma sperimentale non ammette tale valore; in questi casi si ricorre alla seguente famiglia di modelli:

Funzione Potenza : introdotta per modellizzare FA prive di varianza e covariogramma, ma che soddisfano le ipotesi intrinseche $\mathcal{H}2$; la formula analitica è:

$$\gamma(h) = c \cdot \mid h \mid^{\omega}, \quad \forall\ c > 0,\ \omega \in (0, 2) \tag{B.29}$$

È bene a questo punto notare che tutti questi modelli possono, in virtù della proprietà (B.21), essere combinati tra loro, con la sicurezza di ottenere comunque una funzione *definita positiva* — il modello così ottenuto prende il nome di MODELLO A STRUTTURE ANNIDATE; in questo modo, si possono modellizzare una gran varietà di situazioni diverse, caratterizzate da qualsiasi tipo di anisotropia.

Resta comunque una buona norma quella di optare per la soluzione più 'semplice', in quanto ogni struttura annidata dovrebbe essere, almeno in linea di principio, accompagnata da una interpretazione fisica della stessa.

B.4 Stima Puntuale Mediante Kriging

Il problema che si vuole a questo punto affrontare è quello della *stima del valore di una variabile regionalizzata in una localizzazione non campionata sulla base dei valori dei dati campionati in altre localizzazioni.*

In questo contesto, saranno esaminate solo stime di tipo *lineare*, ovvero che assumono una forma generale del tipo:

$$\widehat{Z}(\mathbf{u}) = \sum_{i=1}^{n} w_i(\mathbf{u}) Z(\mathbf{u}_i) + w_0(\mathbf{u}) \tag{B.30}$$

dove $\widehat{Z}(\mathbf{u})$ rappresenta la stima della variabile di interesse nella localizzazione $\mathbf{u}$, $w_i(\mathbf{u})$ sono degli opportuni pesi, che in generale possono dipendere dalla localizzazione in cui si effettua la stima,

e $Z(\mathbf{u}_i)$ sono gli n valori su cui la stima viene condotta — n in teoria sarà pari al numero totale N di campionamenti che costituiscono il dataset, ma nella pratica spesso si fa ricorso a un suo sottoinsieme, cioè $n < N$.

I metodi che si sono sviluppati e sono oggi disponibili sono molto numerosi, e si differenziano tra loro in base al valore che assegnano ai pesi della (B.30); tra quelli più diffusi, ricordiamo:

- metodo della media campionaria;
- metodo poligonale;
- metodo delle triangolazioni;
- la famiglia dei metodi ID (dall'inglese 'Inverse Distance'), per i quali i pesi w_i sono dati da $w_i \propto d_i^{-p}$, $p > 0$, dove d_i indica la distanza i-esima tra i punti in questione.

La geostatistica è però in grado di offrire un metodo (o meglio un'intera famiglia di metodi) di stima lineare che, rispetto a quelli appena citati e che vengono trattati nel dettaglio da Isaaks e Srivastava {27}, cap. 11, risulta essere il *migliore* nel caso di variabili regionalizzate — il significato di 'migliore' sarà chiarito in seguito.

Nessun metodo di stima, per quanto sofisticato possa essere, sarà mai in grado di fornire il valore *vero* di una VR in una localizzazione non campionata; tuttavia, un *buon* metodo dovrebbe essere in grado di manipolare nella maniera più efficiente possibile l'informazione contenuta nei dati. L'*accuratezza* della stima di una VR dipende essenzialmente da:

- numero di osservazioni e qualità dei dati;
- posizione dei campionamenti;
- distanza tra i punti (sia quella tra i punti campionati e il punto di stima, sia quella relativa tra i campionamenti stessi);
- regolarità della variabile regionalizzata.

La geostatistica, in virtù dell'introduzione di un modello di FA $Z(\mathbf{u})$ per la variabile regionalizzata, è in grado sia di fornire un metodo di stima che tiene conto di tutti i fattori appena citati, sia di accompagnare tale stima con una propria incertezza. Tale metodo prende il nome di Kriging[9], e si può dimostrare che risulta essere, nella classe definita dalla (B.30), il *migliore*, nel senso che:

1. è uno stimatore *non polarizzato*;
2. la varianza dell'errore è *minima*.

A questo metodo ci si rivolge spesso con l'acronimo **BLUE(P)**, per indicare:

- **B**est, nel senso della minimizzazione della varianza di stima;
- **L**inear, in quanto è un modello lineare;
- **U**nbiased, nel senso che in media la stima non è affetta da errori sistematici;

[9] Questa denominazione è stata attribuita a tale metodo da Matheron {37} rifacendosi al nome di D. G. Krige, un ingegnere minerario che per primo utilizzò la tecnica delle medie mobili per stimare la quantità di minerale presente in una regione di un deposito. L'idea è stata successivamente ampliata da Matheron, il quale ha fornito alla stima di una VR un solido supporto teorico.

- **E**stimator/**P**redictor.

A seconda delle caratteristiche dei dati campionati e degli obiettivi dello studio, si è sviluppata un'intera famiglia di algoritmi di kriging; tra i più diffusi, ricordiamo:

- kriging *lineari*
  - Simple Kriging (SK);
  - Ordinary Kriging (OK);
  - Universal Kriging (UK) — introdotto per trattare in modo esplicito le situazioni non stazionarie;
  - Kriging con un drift esterno — nel caso in cui il valor medio possa venir descritto da una seconda variabile che varia "lentamente" sul dominio D;
- kriging *non-lineari*
  - Kriging log-normale — introdotto per trattare dati con p.d.f. log-normale;
  - Kriging indicatore — per stimare ad esempio la c.d.f. dei campionamenti;
  - Kriging bayesiano;
  - Kriging disgiuntivo.

Una trattazione completa dal punto di vista teorico di tutti questi metodi è presentata in {13, cap. 3}; per l'approccio multivariato del problema (ossia, nel caso in cui si vogliano considerare più variabili allo stesso tempo), che prende il nome di COKRIGING, si può invece fare riferimento a {31, pagg. 111-114}, a {16, cap. 4} o a {27, cap. 17}.

In questo contesto, verranno approfonditi da un punto di vista teorico/formale solo i due algoritmi cui si è fatto ricorso in questo lavoro di tesi — SK e OK — e che risultano comunque i più diffusi.

B.4.1 Il Problema del Vicinaggio

Si è già accennato al fatto che il numero n di punti che vengono utilizzati per stimare il valore di una VR in una localizzazione non campionata [cfr. eq. (B.30)] — cui spesso si fa riferimento col termine di VICINAGGIO — può non essere pari al numero N di campionamenti che costituiscono l'intero dataset di riferimento.

Dal punto di vista del modello, la condizione migliore è quella di ricorrere a tutti gli N valori disponibili, ma è anche vero che "il modello non è la realtà"! Inoltre, sembra importante sottolineare come all'aumentare di n, aumenti anche l'effetto di smoothing che inevitabilmente accompagna qualsiasi procedura di interpolazione.

> Isaaks e Srivastava {27}, pag. 384 sostengono che *decidere quali siano i campionamenti* rilevanti *per la stima*[10] *può risultare essere un parametro più influente rispetto alla scelta dell'algoritmo impiegato per la stima stessa.*

Per questo, tutti i software che implementano il sistema del kriging o delle simulazioni stocastiche [cfr. §B.6.1] consentono di impostare un *raggio di ricerca* R_v in modo che vengano considerati, in fase di calcolo, solo i punti al suo interno — eventualmente, si può anche definire il numero massimo N_{max} e minino N_{min} di campionamenti che devono concorrere al valore della

[10] Si tenga inoltre presente che in questa fase si assume implicitamente che tali campionamenti appartengano tutti alla medesima popolazione, così come il valore che si andrà a stimare, e che non c'è modo di verificare questa ipotesi nella pratica; ancora, la scelta potrebbe essere differente per differenti regioni del dominio D.

stima, limitando così a N_{max} il numero di punti qualora all'interno dell'area definita dal raggio di ricerca, n fosse maggiore di N_{max}.

È chiaramente possibile definire anche delle aree *ellittiche* per la ricerca, qualora il fenomeno manifestasse delle evidenti anisotropie — l'asse maggiore sarà quindi orientato nella direzione di massima continuità spaziale mostrata dall'analisi variografica condotta.

B.4.2 Simple Kriging (SK)

Questo modello assume che la media m della FA stazionaria $Z(\mathbf{u})$, descritta dalla covarianza $C(\mathbf{h})$, sia nota e costante sul dominio D, ovvero:

$$Z(\mathbf{u}) = m + R(\mathbf{u}), \text{ con } \mathbf{u} \in D, m \text{ noto} \tag{B.31}$$

In questo caso, lo stimatore per la VR di interesse è dato da:

$$\widehat{Z}(\mathbf{u}) = m + \sum_{i=1}^{n} w_i(\mathbf{u}) Z(\mathbf{u}_i) \tag{B.32}$$

Minimizzando la varianza dell'errore, definita come $E[\widehat{Z}(\mathbf{u}) - Z_0]$, dove Z_0 rappresenta il valore vero sconosciuto — tale varianza σ^2_{SK} è scritta in termini dei pesi $w_i(\mathbf{u})$ e della covarianza $C(\mathbf{h})$ — si ottiene un sistema di n equazioni e n incognite (i pesi w_i, appunto) del tipo:

$$\sum_{j=1}^{n} w_{ji}(\mathbf{u}) C_{ij}(\mathbf{u}_i - \mathbf{u}_j) = C_{0i}(\mathbf{u} - \mathbf{u}_i), \quad \forall\, i = 1 \dots n \tag{B.33}$$

dove C_{ij} rappresenta la funzione covarianza — che non può essere sostituita dalla funzione variogramma a meno che la somma dei pesi non sia pari a 1.

Infine, risolvendo il sistema di equazioni (B.33) si ottengono i valori dei pesi $w_i(\mathbf{u})$ da impiegare nella (B.32) e dai quali si ottiene anche la formula per la varianza minimizzata:

$$\sigma^2_{SK}(\mathbf{u}) = C(0) - \sum_{i=1}^{n} w_i(\mathbf{u}) C_{0i}(\mathbf{u} - \mathbf{u}_i) \tag{B.34}$$

Se il modello di FA $Z(\mathbf{u})$ può essere assunto gaussiano e multivariato, allora lo stimatore dato dalla (B.32) coincide con la media condizionata $E[Z_0 \mid Z(\mathbf{u}_1), \dots, Z(\mathbf{u}_n)]$ {31, pag. 92}; inoltre, i parametri della c.c.d.f. corrispondente sono determinati dalle stime ottenute mediante SK, ovvero dalle (B.32) e (B.34): poiché tale c.c.d.f. è anch'essa gaussiana, essa risulta *completamente* determinata da questi due parametri.

$\mathcal{C}1$: Questo importante risultato è alla base dell'approccio delle Simulazioni Stocastiche Gaussiane [cfr. §B.6.1]; tale approccio è detto *parametrico* nella misura in cui è in grado di determinare in modo completo le c.c.d.f. attraverso i loro parametri (m e σ^2_{SK}); inoltre l'algoritmo associato risulta estremamente veloce e affidabile. La sua limitazione sta nell'assunzione di un modello molto specifico (quello gaussiano, appunto) che non sempre risulta appropriato.

B.4.3 Ordinary Kriging (OK)

Questo modello assume che la media m della FA $Z(\mathbf{u})$, descritta dalla covarianza $C(\mathbf{h})$ o alternativamente dal variogramma $\gamma(\mathbf{h})$, sia costante sul dominio D ma non nota, ovvero:

$$Z(\mathbf{u}) = m + R(\mathbf{u}), \text{ con } \mathbf{u} \in D, m \text{ non noto} \tag{B.35}$$

In questo caso, lo stimatore per la VR di interesse è dato da:

$$\widehat{Z}(\mathbf{u}) = \sum_{i=1}^{n} w_i(\mathbf{u}) Z(\mathbf{u}_i) \tag{B.36}$$

Imponendo che lo stimatore (B.36) non sia polarizzato, si ottiene la seguente condizione cui i pesi $w_i(\mathbf{u})$ devono sottostare:

$$\sum_{i=1}^{n} w_i(\mathbf{u}) = 1 \tag{B.37}$$

Procedendo come nel caso relativo al SK, si arriva a un problema di minimizzazione (della varianza) condizionata al vincolo imposto dalla (B.37), che viene risolto ricorrendo all'introduzione di un opportuno moltiplicatore di Lagrange, indicato con μ. Si ottiene quindi il seguente sistema di $n+1$ e equazioni e $n+1$ incognite (n pesi $w_i(\mathbf{u})$ e il parametro μ), cui si fa riferimento col termine di *sistema del kriging stazionario (o ordinario)*:

$$\begin{cases} \displaystyle\sum_{j=1}^{n} w_j(\mathbf{u}) C_{ij}(\mathbf{u}_i - \mathbf{u}_j) + \mu(\mathbf{u}) = C_{0i}(\mathbf{u} - \mathbf{u}_i), \quad \forall\ i = 1 \dots n \\ \displaystyle\sum_{i=1}^{n} w_i(\mathbf{u}) = 1 \end{cases} \tag{B.38}$$

Procedendo in maniera analoga a quanto fatto per SK, risolvendo il sistema di equazioni (B.38) si ottengono i valori dei pesi $w_i(\mathbf{u})$ e del moltiplicatore di Lagrange $\mu(\mathbf{u})$, mediante i quali risalire alla stima della VR data da (B.36) e alla varianza associata, data da:

$$\sigma^2_{OK}(\mathbf{u}) = C(0) - \sum_{i=1}^{n} w_i(\mathbf{u}) C_{0i}(\mathbf{u} - \mathbf{u}_i) - \mu(\mathbf{u}) \tag{B.39}$$

La condizione (B.37) consente di scrivere in termini della funzione *variogramma* sia il sistema del kriging stazionario :

$$\begin{cases} \displaystyle\sum_{j=1}^{n} w_j(\mathbf{u}) \gamma_{ij}(\mathbf{u}_i - \mathbf{u}_j) + \mu(\mathbf{u}) = \gamma_{0i}(\mathbf{u} - \mathbf{u}_i), \quad \forall\ i = 1 \dots n \\ \displaystyle\sum_{i=1}^{n} w_i(\mathbf{u}) = 1 \end{cases} \tag{B.40}$$

sia la corrispondente varianza:

$$\sigma^2_{OK}(\mathbf{u}) = \sum_{i=1}^{n} w_i(\mathbf{u}) \gamma_{0i}(\mathbf{u} - \mathbf{u}_i) - \mu(\mathbf{u}) \tag{B.41}$$

> *Questa proprietà consente quindi di poter applicare OK anche a quei fenomeni che non sono stazionari del secondo ordine, ma che rispettano le meno restrittive ipotesi intrinseche.*

Si può inoltre dimostrare {16, pag. 65} che il sistema di OK implicitamente *ri-stima*, ad ogni localizzazione $\mathbf{u}$, il valore della media m usata nell'espressione del SK (B.32). Poiché spesso nella pratica l'OK viene applicato ricorrendo a tecniche di ricerca di vicinaggio *mobili*, ovvero usando differenti sottoinsiemi del dataset al variare della localizzazione $\mathbf{u}$, la ri-stima implicita della media, indicata quindi con $m^*(\mathbf{u})$, dipenderà ora dalla localizzazione:

> *il kriging ordinario, se applicato con strategie di ricerca* mobili *dei punti di vicinaggio, si può considerare come un algoritmo* non stazionario*, nella misura in cui corrisponde a un modello di FA non stazionario con* $m = E[Z(\mathbf{u})]$ *variabile sul dominio* D *ma covarianza* $C(\mathbf{h})$ *stazionaria.*

È in virtù di questa importante proprietà che l'OK è risultato (e risulta) un algoritmo estremamente robusto e rimane tutt'oggi il metodo di base della geostatistica.

Uno Sguardo Intuitivo al sistema dell'OK

Il sistema del kriging stazionario dato dalla (B.38) o dalla (B.40) può essere facilmente scritto in termini matriciali (in questo caso per la funzione covariogramma):

$$\begin{pmatrix} C_{11} & \dots & C_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{n1} & \vdots & C_{nn} & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{pmatrix} = \begin{pmatrix} C_{01} \\ \vdots \\ C_{0n} \\ 1 \end{pmatrix} \tag{B.42}$$

o, con una notazione più compatta, definendo le tre matrici (il pedice ne indica la dimensione):

$$\mathbf{C}_{(n+1)\times(n+1)} \cdot \mathbf{w}_{(n+1)\times 1} = \mathbf{D}_{(n+1)\times 1} \tag{B.43}$$

Il vettore soluzione del sistema del kriging ordinario $\mathbf{w}$ è quindi legato a due differenti matrici:

$$\mathbf{w} = \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{D} \tag{B.44}$$

Presa a sé stante, la matrice $\mathbf{D}$ produce uno 'schema dei pesi' simile a quello che fornirebbe il metodo ID: infatti, anche la covarianza tra due punti generalmente decresce con l'aumento della distanza relativa; il vantaggio, in questo caso, è che non ci si limita ad andamenti del tipo $|\,\mathbf{h}\,|^{-p}$, ma si hanno a disposizione molte altre funzioni analitiche — combinazioni di quelle ammesse dai modelli teorici [cfr. §B.3.3].

> *La matrice* $\mathbf{D}$ *contiene quindi uno schema dei pesi simile a quello prodotto dai sistemi ID, nel quale però la distanza non è quella di tipo euclideo, ma una distanza di tipo* statistico*.*

La matrice $\mathbf{C}$ porta invece informazioni sulla distribuzione spaziale dei campionamenti che sono utilizzati per la stima: se due punti sono molto vicini tra loro, il relativo valore di C_{ij} sarà verosimilmente piuttosto elevato — e quindi corrisponderà a un valore basso di ${C_{ij}}^{-1}$; in questo modo

la matrice $\mathbf{C^{-1}}$*, contenendo informazioni circa le distanze relative tra tutti i campionamenti impiegati nella stima, fornisce al sistema dell'OK le informazioni sul clustering e sulla configurazione spaziale di tali campionamenti attorno al punto di stima.*

Concludendo, lo schema dei pesi dato dalla (B.44) rende il kriging ordinario sensibile ai due aspetti fondamentali che caratterizzano il problema generale della stima, ovvero la *distanza* e il *clustering*.

B.4.4 L'Effetto dei Parametri del Modello

Le stime ottenute mediante kriging richiedono che sia noto un modello per la FA $Z(\mathbf{u})$, e nello specifico i vari parametri che lo caratterizzano — unitamente, è bene non dimenticarlo, alle ipotesi che è necessario accettare, quali quelle di stazionarietà $\mathcal{H}1$ o quelle intrinseche $\mathcal{H}2$. Per questo, sembra utile analizzare brevemente quale influenza i parametri del modello hanno sulle stime ottenute con sistema il del kriging:

Effetto di Scala : nel caso in cui si impieghino variogrammi $\gamma(\mathbf{h})$ che differiscono solo per un fattore moltiplicativo:

- i pesi $w_i(\mathbf{u})$ *non* subiscono variazioni, e quindi il valore puntuale della stima non cambia;
- la varianza della stima scala invece dello stesso fattore moltiplicativo;

Effetto del range : i $\gamma(\mathbf{h})$, a parità del valore di sella raggiunto e di modello (forma analitica), possono differire tra loro per il valore del range a; anche se i pesi, raddoppiando ad esempio a, non cambiano molto, l'effetto sulla stima può essere considerevole {27, pag. 307}: raddoppiare il range è come informare il kriging che i punti sono due volte più vicini, in termini di distanza 'statistica'; tipicamente, la varianza della stima diminuisce al crescere di a

Effetto di Forma : i $\gamma(\mathbf{h})$ impiegati possono differire per il modello che è stato fittato sui dati; in questi casi, modelli diversi assegneranno pesi diversi in base principalmente al comportamento assunto nei pressi dell'origine;

Effetto dell'Anisotropia : la possibilità di modellizzare anisotropie anche molto forti manifestate dai valori campionati fornisce uno strumento molto potente per adattare la procedura di stima; è intuitivo che il sistema del kriging assegnerà pesi maggiori ai punti che si trovano lungo la direzione di massima continuità spaziale;

Effetto Nugget : nel caso in cui i $\gamma(\mathbf{h})$ differiscono solo per il valore di $\gamma(0)$:

- all'aumentare dell'effetto nugget, la stima tende a quella prodotta da una semplice media campionaria[11];
- all'aumentare dell'effetto nugget relativo (rapporto tra $C(0)$ e valore di sella) la varianza dell'errore aumenta.

[11] Questo è più che ragionevole e auspicabile, in quanto con un puro effetto nugget non si ha più alcuna correlazione spaziale, e quindi risulta logico procedere con una semplice media, trattando tutti i dati allo stesso identico modo.

B.4.5 Alcune Osservazioni Finali

Prima di concludere questa sezione, sembra utile porre l'accento su alcune questioni:

- come tutti i metodi/algoritmi di stima, anche il sistema le kriging opera una sorta di *media lineare* dei dati: questo implica che la *varianza delle stime risulta inevitabilmente inferiore rispetto a quella dei dati di partenza*; inoltre, la procedura stessa di media porta, per sua stessa natura, a una *sovrastima* dei valori bassi e a una *sottostima* dei valori elevati;

- la varianza del kriging, sia essa data dalla (B.34) (SK) o dalla (B.39) (OK), dipende *unicamente* dalla `distribuzione spaziale` dei dati impiegati nella stima, ma non dal loro attuale valore: essa misura quindi la bontà della stima in relazione a *come* i punti sono disposti attorno alla localizzazione $\mathbf{u}$, ma non tiene conto della reale varianza che questi assumono (quella cioè legata ai *valori* $z_i(\mathbf{u}_i)$); per questo, è bene considerarla sempre in maniera per così dire "relativa", e comunque non dovrebbe essere usata per la scelta del modello di variogramma o del tipo di implementazione del kriging {16, pag. 95};

- un problema che non è stato affrontato è quello relativo al SUPPORTO dei dati, definito come la forma e le dimensioni del campione su cui la misura è stata effettuata; quando tali dimensioni sono molto piccole in relazione alla scala che caratterizza lo studio, il supporto si può considerare *puntiforme* e quindi è lecito ricorrere a stime di tipo puntuale — quelle che sono state appunto descritte; il concetto di supporto e le sue implicazioni giocano un ruolo importante nella teoria della geostatistica, in quanto variandone forma e dimensione si ottiene una VR diversa da quella di partenza: i problemi principali si hanno quando l'obiettivo finale dello studio ha un supporto diverso (tipicamente, di dimensioni maggiori) da quello che caratterizza i campionamenti; si parla in questo caso di *Block Kriging*[12], per distinguerlo dal *Point Kriging* che è stato descritto in questa breve presentazione.

B.5 La Cross Validation (CV)

In ogni studio di tipo geostatistico il numero di decisioni soggettive e inter-dipendenti è spesso così elevato che, prima di intraprendere la fase finale di stima, è buona norma *validare* l'intero modello implementato e la procedura di kriging stessa.

> *L'esercizio della cross validation è analogo a quello di una "prova generale": l'idea è quella di mettere in luce quello che potrebbe andare storto, ma non assicura certo che lo "spettacolo" sarà di successo!* {16, pag. 94}

Lo scopo è quindi quello di verificare e confrontare l'influenza dei vari parametri implementati (modello di (co)-variogramma, raggio di ricerca per il vicinaggio, numero di punti da impiegare nella stima, tipo di kriging, ecc.) al fine di ottenere l'insieme di quelli 'ottimali'; in quest'ottica, la cross validation si configura come uno strumento sia *qualitativo* che *quantitativo*.

L'idea di base è la seguente: il campionamento i-esimo $z(\mathbf{u}_i)$ viene temporaneamente escluso dal dataset, e il suo valore $\widehat{z}(\mathbf{u}_i)$ stimato mediante gli altri dati sperimentali che ne costituiscono il vicinaggio; si determina quindi il residuo i-esimo come $r(\mathbf{u}_i) = z(\mathbf{u}_i) - \widehat{z}(\mathbf{u}_i)$, e il campionamento $z(\mathbf{u}_i)$ viene così reintrodotto nel dataset. L'intera procedura viene ripetuta per tutti gli N valori disponibili.

[12] Dettagliate descrizioni di questi aspetti, delle problematiche connesse e di come possono venir trattare e risolte si possono trovare in {27, capp. 13 e 19}.

Si opterà quindi per il modello/metodo che manifesta il comportamento migliore, una volta che si sia definita una opportuna *misura* per identificare il risultato migliore. È importante sottolineare infine come la cross validation non sia una tecnica basata su test-delle-ipotesi.

Indicando con l'indice k una delle K procedure di stima o varianti di implementazione sotto esame, tra i criteri cui far riferimento per l'analisi dei residui si ricordano i seguenti:

- la distribuzione degli N residui $r_k(\mathbf{u}_i)$ dovrebbe essere simmetrica, centrata attorno allo zero e con minima varianza;
- in un diagramma '$r_k(\mathbf{u}_i)$' *vs.* '$\widehat{z}_k(\mathbf{u}_i)$', i residui dovrebbero costituire una "fascia" di punti centrati attorno a una retta orizzontale passante per l'origine; inoltre, la varianza (ovvero la larghezza della fascia) dovrebbe essere costante, manifestando così un'*indipendenza* dall'attuale valore $z(\mathbf{u}_i)$ — condizione nota come *omoschedasticità*;
- gli N residui della procedura k-esima dovrebbero essere indipendenti tra loro; questo può essere controllato ricorrendo a un'analisi variografica ($\gamma(\mathbf{h})$ dovrebbe essere in questo caso un puro effetto nugget) o a una mappa che rappresenti delle isolinee per i valori dei residui (tale mappa non dovrebbe mostrare alcun trend evidente).

Bisogna comunque porre l'accento sul fatto che una CV può fornire informazioni sulla bontà del modello *solo in relazione ai campionamenti che costituiscono il dataset e alle zone del dominio D dove questi sono presenti*: nel caso in cui la rete di monitoraggio manifestasse un clustering non trascurabile o fosse particolarmente disomogenea, i risultati ottenuti da esercizi di CV andrebbero interpretati con le dovute cautele — un modello che risultasse estremamente affidabile applicato agli N dati disponibili potrebbe non esserlo altrettanto qualora venisse applicato all'intera popolazione (o qualora il dataset fosse ampliato con nuovi valori sperimentali) in quanto gli N $z(\mathbf{u}_i)$ clusterizzati non sono realmente rappresentativi della popolazione da cui sono stati estratti.

L'impiego della cross validation nella pratica è descritto nel dettaglio da Isaaks e Srivastava {27}, cap. 15.

B.5.1 Jackknife

A volte, con scopi che possono essere diversi da quelli della CV, se il dataset di partenza è sufficientemente numeroso, viene suddiviso i più parti (tipicamente, due) con lo scopo di usarne una (quella più numerosa) per la trattazione geostatistica del fenomeno, e l'altra per una successiva validazione del modello — in questo modo la validazione risulta *indipendente* dai dati usati per la costruzione del modello, nella misura in cui i due dataset non si sovrappongono.

Per la creazione dei sottoinsiemi, è utile ricorrere ad esempio a un declustering casuale [cfr. §B.1.2], col vantaggio che il sottoinsieme usato in fase di jackknife rispecchierà la distribuzione spaziale dei campionamenti di partenza.

L'utilizzo del jackknife in relazione all'impiego di reti neurali per la previsione di variabili regionalizzate è invece descritto da Kanevski e Maignan {31}.

B.6 Le Simulazioni Stocastiche

Nei paragrafi precedenti si sono approfonditi gli aspetti legati alla *previsione spaziale* di una variabile regionalizzata che si basano su modelli BLUE(P), nello specifico facendo ricorso al metodo del kriging. L'obiettivo principale è stato quello di sviluppare un modello la cui caratteristica

fosse la *migliore qualità predittiva*, ovvero un modello non-polarizzato che fornisse la minima varianza possibile tra stima e dato reale.

Questo metodo [cfr. §B.4 e §B.4.5] — come del resto tutti gli altri metodi di interpolazione —, fornisce un *unico* modello smussato della realtà, senza per altro riprodurre la reale variabilità dei dati stessi — come descritta ad esempio dal variogramma. Va però ricordato che la geostatistica, come suggerisce il nome stesso, è legata a una trattazione statistica dei dati, e assume che il valore misurato $z(\mathbf{u}_\alpha)$ sia la realizzazione di un processo stocastico $Z(\mathbf{u})$ del quale va determinata la struttura della *correlazione spaziale*.

> *L'idea che sta alla base delle simulazioni stocastiche è quella di sviluppare un* modello/generatore spaziale *di tipo* Monte Carlo[13] *che sia in grado di generare numerose e, in un senso che verrà chiarito in seguito,* equiprobabili *realizzazioni del processo stocastico* $Z(\mathbf{u})$.

Ogni realizzazione è detta IMMAGINE STOCASTICA (IS) e riflette le proprietà che sono state imposte al modello stesso di Funzione Aleatoria (FA) $Z(\mathbf{u})$; tipicamente, si richiede che le varie immagini stocastiche riproducano esattamente i valori $z(u_\alpha)$ nelle posizioni dei campionamenti, ovvero che rispettino i dati sperimentali disponibili: in questo caso, di parla di SIMULAZIONI STOCASTICHE CONDIZIONATE[14]. *IS*

A seconda di quale tipo di informazione viene impiegata nelle simulazioni, si possono identificare tre ampie classi:

- simulazioni di variabili continue, come ad esempio il valore di concentrazione di un inquinante;
- simulazione di variabili categoriche, come ad esempio la presenza o meno di un particolare tipo di vegetazione;
- simulazioni di oggetti, come ad esempio la simulazione di fratture in un corpo roccioso (si parla in questo caso di object-based simulation in contrasto con le pixel-based simulation cui appartengono le prime due classi citate).

Kanevski e Maignan {31} sostengono che le simulazioni stocastiche condizionate (CSS) possono essere usate come un modello `realistico` della variabilità spaziale del fenomeno, e che attualmente sono largamente impiegate in combinazione con differenti modellizzazioni di tipo scientifico (modelli idrologici, economici, di rischio) nei casi in cui sia importante valutare l'*incertezza* del risultato finale.

B.6.1 Principi delle Simulazioni Stocastiche

Si consideri la distribuzione, su un campo D, di una o più variabili $z(\mathbf{u})$ con $\mathbf{u} \in D$; la variabile $z(\mathbf{u})$ può essere sia continua che categorica.

[13] Si tenga presente che per sviluppare un *reale* modello spaziale di tipo Monte Carlo è necessario ricorrere a una statistica che tenga conto delle correlazioni tra le variabili — almeno quella 'a due punti' descritta dal variogramma — e in grado di considerare più punti allo stesso tempo.

[14] Le simulazioni stocastiche condizionate sono inizialmente state introdotte nel tentativo di "correggere" l'effetto di smoothing che inevitabilmente accompagna le mappe prodotte con l'algoritmo del kriging, le cui stime si possono pensare come *medie pesate mobili* dei dati disponibili: alla luce di questa visione, la loro variabilità spaziale sarà per forza di cose minore di quella posseduta dai dati di partenza. Le mappe prodotte da simulazioni stocastiche condizionate risultano invece più appropriate in studi che vogliano riprodurre la variabilità locale del fenomeno in esame.

> *Una* Simulazione Stocastica *è il processo di costruzione di realizzazioni* congiunte, equiprobabili *e* alternative *della distribuzione spaziale delle variabili aleatorie (VA) che costituiscono il modello* $Z(\mathbf{u})$.

Ciascuna di queste realizzazioni, ovvero l'insieme di tutte le realizzazioni delle variabili aleatorie $z(\mathbf{u})$, verrà indicata con l'apice l in questo modo: $\{z^l(\mathbf{u}), \mathbf{u} \in D\}$; con questo formalismo, una simulazione è detta *condizionata* se la rispettiva realizzazione rispetta i valori dei dati reali nelle loro proprie localizzazioni $\mathbf{u}_\alpha$, cioè se:

$$z^l(\mathbf{u}_\alpha) = z(\mathbf{u}_\alpha) \quad \forall\, l \tag{B.45}$$

In generale, i principi cui si fa riferimento per sviluppare delle simulazioni condizionate sono:

- le realizzazioni devono riprodurre l'istogramma rappresentativo dei dati originali (eventualmente, dopo un processo di declustering [cfr. §B.1.2], se necessario);
- le realizzazioni devono riprodurre la variabilità spaziale dei dati originali descritta dal variogramma;
- le realizzazioni devono rispettare i dati — nei punti campionati, i valori simulati devono essere uguali ai valori misurati.

Ogni immagine stocastica sarà quindi determinata:

- dai dati di condizionamento (il dataset di riferimento);
- dal modello della simulazione;
- dal tipo di algoritmo implementato.

.

Equiprobabilità delle Realizzazioni

Perché un set di L realizzazioni di uno specifico algoritmo di simulazione possa essere utilizzato per determinare probabilità legate a funzioni delle VA in esame, è necessario che sia garantita l'`equiprobabilità` delle L realizzazioni, a dire che

> *ognuna delle* L *immagini stocastiche deve avere la stessa probabilità di essere estratta di qualunque altra contenuta nell'insieme delle* L *realizzazioni.*

Questo requisito fondamentale si ottiene se ogni immagine stocastica può essere identificata con un singolo numero casuale (noto in letteratura come *realization seed number*) uniformemente distribuito nell'intervallo $[0, 1]$; in altre parole, se ciascuna realizzazione può essere riprodotta esattamente fornendo all'algoritmo di simulazione il seed number che la identifica — chiaramente, tale equiprobabilità va riferita alle specifiche del modello di simulazione, del modello di FA $Z(\mathbf{u})$, ai valori dei parametri e all'implementazione dell'algoritmo[15].

[15] Deutsch e Journel {16} fanno notare che non esiste un unico algoritmo di simulazione tanto flessibile da consentire la riproduzione dell'ampia varietà di statistiche e caratteristiche che si possono incontrare nella realtà!

Simulazioni e Algoritmi di Interpolazione

È a questo punto importante sottolineare due aspetti fondamentali per i quali le simulazioni si differenziano rispetto a qualsiasi altro tipo di algoritmo di interpolazione:

1. l'obiettivo principale degli algoritmi di interpolazione, kriging incluso, è quello di fornire la "migliore", e quindi *unica*, stima locale $z^*(\mathbf{u})$, $\mathbf{u} \in D$, per ogni valore $z(\mathbf{u})$ non campionato, senza curarsi della statistica spaziale dell'intero insieme delle stime così prodotte; quello che prevale, in questo caso, è l'`accuratezza locale`.
   Per quanto riguarda le simulazioni stocastiche, invece, l'aspetto cui viene data la precedenza è quello relativo alle `caratteristiche globali` del fenomeno e alla sua statistica — l'accento è posto sul riprodurre al meglio almeno i primi due momenti che sono accessibili sperimentalmente, quali media e variogramma, e l'istogramma dei campionamenti; detto in altri termini, le simulazioni forniscono rappresentazioni globali alternative, $z^l(\mathbf{u}), \mathbf{u} \in D$, nelle quali prevale il pattern della continuità spaziale che è stato identificato per il fenomeno in esame — e che è 'contenuto' nel modello di FA;

2. se si esclude il caso in cui si possa assumere un modello gaussiano per gli errori, il kriging fornisce una misura incompleta dell'accuratezza locale della stima [cfr. §B.4.5], e nessuna informazione circa l'accuratezza congiunta quando vengono considerate assieme differenti localizzazioni; le simulazioni sono invece concepite proprio per fornire una misura di tale accuratezza. In altre parole, se da un lato il kriging restituisce un singolo modello numerico, che può essere definito come *migliore* in senso *locale*, le simulazioni forniscono molti modelli numerici alternativi $z^l(\mathbf{u})$, ognuno dei quali può essere visto come una *buona* rappresentazione della realtà in senso *globale*; si ha quindi modo di valutare:
   a) l'**accuratezza locale** in base alle differenze riscontrate tra L alternativi valori simulati in una particolare localizzazione;
   b) l'**accuratezza globale** o **congiunta** in base alle differenze riscontrate tra L simulazioni alternative.

Gli approcci cui si ricorre nella pratica per la simulazione di variabili sia continue che categoriche sono molti e variegati[16]; tra quelli più diffusi ricordiamo:

- algoritmi Gaussian-based — modelli parametrici;
- algoritmi Indicator-based — modelli non parametrici;
- algoritmi booleani — impiegati principalmente nella simulazione di oggetti;
- algoritmi di Simulated annealing — modelli formulati in termini di un problema di ottimizzazione, senza specificare il modello per la FA: l'algoritmo parte da una immagine casuale e propone iterativamente dei cambiamenti, che possono venir accettati o meno in base a una funzione pre-definita (ad esempio, quanto ci si scosta dall'istogramma sperimentale dei campionamenti) fino al raggiungimento di un sufficiente grado di accordo con l'immagine di riferimento[17];
- impiego di reti neurali con differenti architetture [cfr. ad esempio {31, cap. 8}];

[16] Per approfondire da un punto di vista teorico modelli e algoritmi impiegati in simulazioni di tipo geostatistico, si può far riferimento al libro di Lantuejoul {34}.
[17] Si può pensare a questo modello come una simulazione stocastica del lento raffreddarsi di un sistema fisico — per il quale la funzione pre-definita sia l'hamiltoniana H del sistema a una certa temperatura T — fino al raggiungimento di un minimo locale di H stessa, che rappresenta quindi l'immagine stocastica di arrivo.

- modelli spettrali, che possono essere visti come dei duali di quelli gaussiani nei quali il variogramma viene riprodotto mediante campionamento della sua trasformata di Fourier discreta (DFT).

La questione dell'Ergodicità

Quello che si richiede a un algoritmo di simulazione è di produrre delle realizzazioni che riflettano le proprietà statistiche proprie dei dati a disposizione: il problema che si vuole affrontare in questo contesto è quello relativo a *quanto bene tali statistiche dovrebbero essere riprodotte*. Si tenga inoltre presente che una riproduzione *esatta* della statistica del modello da parte di ogni singola realizzazione simulata potrebbe non essere desiderabile, in quanto la statistica del modello è stata ricavata da un campione di dimensione finita, e non dalla reale popolazione cui il campione di misure appartiene, ed è quindi accompagnata da incertezze — e le varie immagini stocastiche dovrebbero portarne il segno.

Un modello di funzione aleatoria $Z(\mathbf{u})$ stazionario è detto ERGODICO nel parametro μ se la corrispondente realizzazione statistica μ^l, $\forall\, l$, tende a μ al crescere del campo D. Così, assumendo che $Z(\mathbf{u})$ sia stazionario ed ergodico e che il campo di simulazione sia sufficientemente grande[18], ci si dovrebbe aspettare che la statistica del modello venga riprodotta in maniera esatta da ogni realizzazione l-esima.

Nella pratica, però, come ad esempio riportato in {16, pag. 130}, anche utilizzando un campo di simulazione quattro volte maggiore del range del variogramma modello e una griglia di simulazione di 10 000 punti, le fluttuazioni ergodiche dei parametri ricavati da differenti simulazioni sono notevoli.

In molte applicazioni per le quali non si ha la certezza assoluta che la statistica del campione rispecchi in maniera ottimale quella della popolazione, tali fluttuazioni possono risultare particolarmente *utili*, in quanto da esse è possibile ricavare informazioni circa l'incertezza che accompagna i risultati ottenuti dall'elaborazione delle simulazioni. Così, qualora il dataset cui si facesse riferimento fosse ad esempio costituito da campionamenti fortemente addensati in zone particolari (situazione che potrebbe portare a non essere troppo confidenti nella statistica da questi ricavati), una scelta ragionevole e di tipo conservativo potrebbe essere quella del modello che manifesta le "peggiori" proprietà ergodiche!

> *La sola corrispondenza che la teoria delle simulazioni garantisce è quella ricavata da una media, nel senso del* valore di aspettazione*, di un gran numero di immagini stocastiche: meno ergodico il modello di FA scelto, maggiore il numero di realizzazioni richiesto per avvicinare tale valore di aspettazione.*

L'Approccio delle Simulazioni Sequenziali

L'idea che sta alla base delle simulazioni sequenziali è ormai ben nota ed è stata introdotta nell'ambito geostatistico da Alabert e Massonat nel 1990; questo tipo di simulazione può essere considerato come il solo vero e generale algoritmo di simulazione {31, pag. 131}.

c.p.d.f.

L'approccio delle simulazioni sequenziali si basa essenzialmente sulla possibilità di ricavare il valore di una variabile $Z(\mathbf{u})$ dalla *funzione distribuzione di probabilità condizionata* (c.p.d.f.) dato il valore di una seconda variabile correlata alla prima, nella stessa localizzazione $\mathbf{u}$ [cfr. {16, pagg. 123–127}]: l'idea è quella di *estendere il condizionamento in modo da includere* tutti *i dati*

[18] Quello che ha reale rilevanza è la dimensione del campo D rispetto al range del variogramma che caratterizza il modello $Z(\mathbf{u})$ e non la densità di discretizzazione di tale campo.

disponibili in un definito intorno della posizione $\mathbf{u}$*, includendo non solo i dati di partenza, ma anche quelli relativi a tutte le precedenti simulazioni condotte.*

Si consideri la funzione distribuzione di probabilità congiunta di N variabili aleatorie Z_i, con N molto grande; le N variabili aleatorie Z_i possono rappresentare il valore di uno stesso attributo di un processo stocastico $Z(\mathbf{u})$ in N differenti localizzazioni, o N differenti attributi nella stessa localizzazione $\mathbf{u}$, o ancora una combinazione di K differenti attributi in M nodi di una griglia, con $N = K \cdot M$. Si consideri quindi il *condizionamento* di queste N VA da parte di un set di informazioni di n dati di *qualsiasi* tipo; la corrispondente N-variata *funzione distribuzione di probabilità cumulativa condizionata* (c.c.d.f.) sarà data da: *c.c.d.f.*

$$F_N(z_1, \ldots, z_N \mid (n)) = \text{Prob}\{Z_i \leq z_i,\ i = 1 \ldots, N \mid (n)\} \tag{B.46}$$

Si noti che l'equazione (B.46) ha validità del tutto generale e che alcune o tutte le Z_i possono essere di tipo categorico.

A questo punto, si può mostrare come sia possibile estrarre un campione N-variato dalla c.c.d.f. (B.46) ricorrendo a N passaggi, ognuno dei quali richiede una c.c.d.f. univariata con un livello di condizionamento crescente:

1. estrarre un valore z_1^l dalla c.c.d.f. univariata di Z_1, dati gli (n) valori sperimentali di partenza (dataset — spesso nella pratica ci si limita a una parte di esso, in base alla scelta del vicinaggio); il valore z_1^l viene quindi considerato come un nuovo dato *condizionante*, in modo che il set di informazioni venga aggiornato a $(n+1) = (n) \cup \{Z_1 = z_1^l\}$;

2. estrarre un valore z_2^l dalla c.c.d.f. univariata di Z_2, dati gli $(n+1)$ valori di condizionamento; anche in questo caso, il set di informazioni va aggiornato a $(n+2) = (n+1) \cup \{Z_2 = z_2^l\}$;

3. ripetere in maniera sequenziale per tutte le N variabili aleatorie Z_i.

Il set dato da $\{z_i^l,\ i = 1 \ldots, N\}$ rappresenta quindi una immagine stocastica simulata e congiunta delle N variabili aleatorie dipendenti Z_i.

Questa procedura di simulazioni sequenziali prevede quindi la determinazione di N c.c.d.f. univariate, e precisamente

$$\begin{aligned} &\text{Prob}\{Z_1 \leq z_1 \mid (n)\} \\ &\text{Prob}\{Z_2 \leq z_2 \mid (n+1)\} \\ &\text{Prob}\{Z_3 \leq z_3 \mid (n+2)\} \\ &\vdots \\ &\text{Prob}\{Z_N \leq z_N \mid (n+N-1)\} \end{aligned} \tag{B.47}$$

Il principio su cui si basano le simulazioni sequenziali è *indipendente* dall'algoritmo o dal modello impiegati per determinare la sequenza (B.47) di c.c.d.f. univariate; inoltre, la sequenza di decomposizione appena descritta ha una validità del tutto generale [cfr. {31, pag. 132}].

A questo punto, un problema potrebbe essere dato dal dove reperire le c.c.d.f. e le loro proprietà statistiche; uno degli algoritmi più utilizzati nella pratica assume che il modello di FA $Z(\mathbf{u})$ — e conseguentemente anche tutte le c.c.d.f. — siano di tipo Gaussiano: in questa caso è possibile ricavare i primi due momenti di tali distribuzioni ricorrendo a N sistemi di simple kriging (SK) [cfr. $\mathcal{C}1$].

Simulazioni Gaussiane Sequenziali (SGS) Il modello di processo stocastico gaussiano è unico in statistica, sia per la sua semplicità analitica sia perché risulta essere la distribuzione limite di molti teoremi analitici, noti globalmente come *teorema limite centrale* [cfr. {12, pagg. 147-149}].

Sia $\{z(\mathbf{u}),\ \mathbf{u} \in D\}$ un fenomeno spaziale continuo[19] generato dalla somma di differenti sotto-fenomeni *indipendenti* $\{j_k(\mathbf{u}),\ \mathbf{u} \in D\},\ k = 1, \ldots, K$, che ammettano distribuzioni spaziali simili: allora la distribuzione spaziale di $z(\mathbf{u})$ può essere modellizzata da un processo stocastico gaussiano multivariato, ovvero

$$Z(\mathbf{u}) = \sum_{k=1}^{K} Y_k(\mathbf{u}) \approx \text{Gaussiano} \tag{B.48}$$

Si presti attenzione al fatto che la condizione più stringente sulla (B.48) è data non tanto dal valore di K o dal fatto che le componenti $Y_k(\mathbf{u})$ siano equamente distribuite, quanto piuttosto dall'ipotesi di *indipendenza* delle $Y_k(\mathbf{u})$ stesse.

Se da un lato, infatti, gli errori di misura possono spesso essere considerati come eventi indipendenti, nell'ambito delle scienze naturali i differenti processi geologici, fisici e biologici che hanno generato il fenomeno osservato sono difficilmente indipendenti tra di loro, e in alcuni casi potrebbero non risultare nemmeno additivi. Nonostante questo, l'approccio basato su un modello parametrico di tipo gaussiano multivariato è largamente utilizzato, soprattutto in relazione a:

- la sua semplicità analitica (è infatti l'unico modello per il quale la c.d.f. è completamente e analiticamente nota);
- la dettagliata conoscenza teorica che lo caratterizza;
- il suo impiego, con successo, in molti campi, anche diversi tra loro.

L'algoritmo e la procedura delle Simulazioni Gaussiane Sequenziali (SGS) sono ben noti fin dalla prima pubblicazione, nel 1992, della libreria GSLib {16}; una tipica SGS di una variabile *continua* $z(\mathbf{u})$ modellizzata da un processo stocastico stazionario e gaussiano multivariato $Z(\mathbf{u})$, prevede quindi i seguenti passaggi:

1. determinare la funzione di distribuzione cumulativa (c.d.f.) $F_Z(z)$ rappresentativa dell'intera area di studio, e non solo dei campionamenti-z disponibili; nel caso in cui il dataset fosse affetto da clustering, poiché in questo contesto lo scopo è quello di una stima *globale*, sarà opportuno ricorrere a una c.d.f. relativa al dataset declusterizzato, che dovrebbe essere utilizzata in entrambe le fasi di trasformazione delle variabili [cfr. punti 2 e 6 della procedura];
2. ricorrendo alla c.d.f. $F_Z(z)$, si procede a una *trasformazione normale standardizzata* dei dati-z in dati-y, i quali avranno così una c.d.f. gaussiana standardizzata — media nulla e varianza unitaria[20]; tuttavia la normalità della c.d.f. univariata non è sufficiente, in quanto il modello che si vuole sviluppare è sì normale, ma multivariato: per questo risulta necessario verificare la normalità bi- e n-variata dei dati[21];

[19] "Continuo" a significare che $z(\mathbf{u})$ *non* è caratterizzato da variabili categoriche discrete o da una sovrapposizione di differenti popolazioni.

[20] Questo può sempre essere fatto, in quanto una trasformazione non lineare può sempre trasformare una qualsiasi c.d.f. continua in un'altra c.d.f. [cfr. ad esempio {27, pagg. 469-476}].

[21] Con una sola realizzazione, è impossibile testare la normalità di ordine superiore a due, tuttavia questo non è necessario, in quanto l'intero modello si basa su una statistica al massimo a due punti, ovvero il covariogramma $C_Y(\mathbf{h})$; come sostengono Deutsch e Journel {16}, pag. 144, nella pratica, *se dalla statistica campionaria non si può dimostrare che la normalità bivariata è violata, la scelta di un modello gaussiano multivariato dovrebbe essere la prima cui rivolgersi per la simulazione di una variabile continua.*

3. testare la validità dell'ipotesi di normalità bi-variata dei dati-y, ovvero che la c.d.f. bi-variata di ogni coppia di valori $Y(\mathbf{u})$ e $Y(\mathbf{u}+\mathbf{h})$ sia normale $\forall \mathbf{u}, \forall \mathbf{h}$; esistono vari modi per testare la normalità bi-variata di dati che assumono un istogramma normale; uno dei più rilevanti sfrutta il legame analitico che lega la covarianza $C_Y(\mathbf{u})$ a un qualsiasi valore della c.d.f. gaussiana bi-variata standardizzata {16, pag. 142}:

$$\begin{gathered}\text{Prob}\{Y(x) \leq y_p, Y(x+h) \leq y_p\} = \\ p^2 + \frac{1}{2\pi}\int_0^{\arcsin C_Y(\mathbf{h})} \exp\left[-\frac{y_p^2}{1+\sin(\theta)}\right] d\theta\end{gathered} \tag{B.49}$$

dove $y_p = G^{-1}(p)$ è il quantile p normale standardizzato, G la c.d.f. normale standardizzata e $C_Y(\mathbf{h})$ il correlogramma del modello gaussiano standardizzato $Y(\mathbf{u})$.

A questo punto, la probabilità bivariata (B.49) può essere legata alla covarianza indicatore per la soglia y_p:

$$\begin{gathered}\text{Prob}\{Y(x) \leq y_p, Y(x+h) \leq y_p\} = \\ E[I(\mathbf{u};p) \cdot I(\mathbf{u}+\mathbf{h};p)] = p - \gamma_I(\mathbf{h};p)\end{gathered} \tag{B.50}$$

dove $I(\mathbf{u};p) = 1$ se $Y(\mathbf{u}) \leq y_p$, $I(\mathbf{u};p) = 0$ altrimenti, e $\gamma_I(\mathbf{h};p)$ rappresenta il variogramma indicatore per il quantile p alla soglia y_p. Il test consiste quindi nel confrontare il variogramma indicatore $\gamma_I(\mathbf{h};p)$ con il valore teorico dato dalla (B.49)[22].

Esiste anche un altro test empirico {31, pag. 133}, e più semplice da applicare, che consiste nel confrontare i due termini di questa equazione:

$$\frac{\sqrt{\text{variogramma}(\mathbf{h})}}{\text{madogramma}(\mathbf{h})} \stackrel{?}{=} \sqrt{\pi} \tag{B.51}$$

nel caso di una variabile aleatoria con distribuzione normale bi-variata, i due termini che compaiono nella (B.51) dovrebbero essere uguali a differenti lag $\mathbf{h}$;

4. se il modello gaussiano multivariato può venir adottato con sufficiente sicurezza per le variabili y, allora la c.d.f. locale è normale con media e varianza ottenute dalle equazioni di Simple Kriging; l'ipotesi di stazionarietà impone di usare il SK con media nulla in questo passaggio: tuttavia, nel caso in cui si disponesse di un numero sufficiente di dati in modo da poter considerare un modello non-stazionario per la media $E[Y(\mathbf{u}]$ di $Y(\mathbf{u})$, ricorrendo alle equazioni di Ordinary Kriging la media sarebbe implicitamente ri-estimata per ogni intorno di $\mathbf{u}$; in ogni caso, per la c.c.d.f. gaussiana andrebbe sempre usata la varianza ottenuta da SK;

5. iniziare finalmente la simulazione gaussiana sequenziale, che prevede i seguenti passi:
   a) definire un percorso *casuale*[23] che visiti tutti i nodi della griglia (non necessariamente regolare) una sola volta; ad ogni nodo $\mathbf{u}$, selezionare un certo numero di punti (definendo un opportuno vicinaggio di ricerca) che costituiranno l'insieme di condizionamento (includendo sia i dati-y originali, sia quelli delle precedenti realizzazioni condotte);

[22] Si noti che la variabile indicatore $I(\mathbf{u};p)$ è la stessa sia che venga definita sui dati di partenza z che su quelli trasformati y, fin tanto che le soglie z_p e y_p siano entrambe i p-quantili delle rispettive c.d.f.; in altre parole, i variogrammi indicatori sono invarianti per qualsiasi trasformazione lineare o non-lineare monotona crescente {16, pagg. 142-144}.

[23] La teoria non specifica la sequenza con la quale gli N nodi della griglia dovrebbero essere simulati; tuttavia, la pratica ha mostrato che la scelta migliore, per evitare possibili artefatti, è quella di considerare un percorso casuale {26}.

b) ricorrere a SK con il modello di variogramma ricavato dai dati-y per determinare i parametri — media e varianza — della c.c.d.f. del processo stocastico $Y(\mathbf{u})$ nella posizione $\mathbf{u}$;
c) estrarre un valore simulato $y^l(\mathbf{u})$ dalla c.c.d.f. appena determinata;
d) aggiungere il valore simulato $y^l(\mathbf{u})$ al dataset dei valori;
e) procedere al nodo successivo e ripetere i punti descritti fino al completamento di tutti i nodi della griglia;

6. applicare una *trasformazione inversa* ai valori gaussiani simulati $\{y^l(\mathbf{u}),\ \mathbf{u} \in A\}$ in modo da ottenere i valori simulati per la variabile di partenza $\{z^l(\mathbf{u}),\ \mathbf{u} \in A\}$.

Le fasi di *trasformazione* dei dati prevedono l'impiego di una c.d.f. che viene ricavata dagli N dati sperimentali accessibili mediante la costruzione di un istogramma cumulativo; nel caso migliore, ovvero se tutti gli N valori $z(\mathbf{u}_\alpha)$ risultano differenti, tale istogramma sarà costituito da N classi.

Essendo le simulazioni stocastiche una tecnica detta di 'espansione', nel senso che genera molti più dati di quelli di partenza, se N non è molto elevato, la risoluzione che caratterizza la c.d.f. campionaria potrebbe non essere sufficiente, in particolare per quanto riguarda le due code della distribuzione stessa — che sono relative a probabilità piuttosto basse, e quindi verosimilmente pochi valori appartenenti alle stesse saranno disponibili nel dataset.

Per questo, soprattutto in fase di *trasformazione inversa*, risulta necessario introdurre dei `modelli di interpolazione` per avere informazioni relative alle zone della c.d.f. che vanno oltre il minimo e massimo valore ricavati sperimentalmente — ovvero, avere a disposizione dei modelli per le due code della distribuzione; la simulazione è infatti in gradi di generare valori della variabile in esame anche oltre il range che caratterizza il dataset. La scelta della modellizzazione della coda superiore della c.d.f. può risultare particolarmente importante per le situazioni ambientali nelle quali z rappresenta ad esempio la concentrazione di un inquinante, in quanto le scelte operate in questa fase possono ripercuotersi in maniera significativa sui valori estremi generati dalle simulazioni.

Per approfondire le questioni e i modelli legati al problema appena esposto, si può fare riferimento a {16, pagg. 134-138}.

Bibliografia

[1] 'Quantum GIS (QGIS), a user friendly Open Source Geographic Information System'. GPL (2008). Version 0.11.0–Metis
URL http://qgis.org/

[2] AA.VV. *Radon and Its Decay Products in Indoor Air*. Environmental Science and Technology. W. Nazaroff and A. V. Nero, Jr (1988). ISBN 0-471-62810-7

[3] A. Bertolo, C. Bigliotto, C. Giovani, M. Garavaglia, M. Spinella, L. Verdi, e S. Pegoretti. 'Spatial distribution of indoor radon in Triveneto (northern Italy): a geostatistical approach'. In 'VIII International Workshop: Geological Aspects of Radon Risk Mapping', (26-30 settembre 2006)

[4] F. Bochicchio, G. C. Venuti, C. Nuccetelli, S. Piermattei, S. Risica, L. Tommasino, e G. Torri. 'Result of the Representative Italian National Survey on Radon Indoors'. *Health Physics*, 71(5):741–748 (1996)

[5] J. P. Chiles e P. Delfià. *Geostatistics, Modeling Spatial Uncertainty*. Wiley series in probability and statistics. John Wiley & Sons (1999)

[6] G. Christakos. 'On the Problem of Permissible Covariance and Variograms Models'. *Water Resources Research*, 20(2):251–265 (1984)

[7] G. Christakos. *Modern Spatiotemporal Geostatistics*. Oxford University Press (2000)

[8] G. Christakos, P. Bogaert, e M. Serre. *Temporal GIS. Advanced Function for Field-Based Application*. Springer Verlag, Heidelberg (2002)

[9] W. G. Cochran. *Samplign Techniques*. Wiley, NY (1978)

[10] S. G. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London (2001)

[11] J. Conrady, K. Martin, J. Lembcke, e H. Martin. 'The True Size of Lung Cancer Risk from Indoor Radon: Hidden behind a Smoke Screen?' In 'PreCura Institute for Preventive Medicine, International Congress Series', volume 1225, pagine 253–258 (2002)

[12] G. Cowan. *Statistical Data Analysis*. Oxford University Press (1998)

[13] N. A. C. Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Mathematicl Statistics. John Wiley & Sons, inc. (1993). ISBN 0-471-00255-0

[14] M. David. *Handbook of Applied Advanced Geostatistical Ore Reserve Estimation*. Elsevier Science Publisher B. V. (1988)

[15] A. Z. David Meyer e K. Hornik. *vcd: Visualizing Categorical Datas* (2008). R package version 1.0-8
URL http://www.jstatsoft.org/v17/i03/

[16] C. V. Deutsch e A. G. Journel. *GSLIB – Geostatistical Software Library and User's Giude*. Applied Geostatistics Series. Oxford University Press, seconda edizione (1998). ISBN 0-19-510015-8

[17] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, e A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien* (2008). R package version 1.5-18

[18] G. Dubois. 'An Overview of Radon Surveys in Europe'. EC. Office for Official Publication of the European Communities, Luxembourg (2005). EUR 21892 EN

[19] G. Dubois e P. Bossew. 'The radon "noise" and its geostatistical implications: risk mapping or mapping at risk?' In 'XI International Congress for Mathematical Geology', IAMG2006, Liege, Belgium (3–8 settembre 2006)

[20] J. Fan e I. Gijbels. *Local Polinomyal Modelling and its Application*, volume 66 di *Monograph on Statistical and Applied Probability*. Chapman and Hall, London (1997)

[21] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Applied Geostatistics Series. Oxford University Press (1978). ISBN 0-19-511538-4

[22] P. Goovaerts. 'Impact of the Simulation Algorithm, Magnitude of Ergodic Fluctuations and Number of Realizations on the Spaces of Uncertainty of Flow Properties'. *Stochastic Environmental Research and Risk Assessment*, 13(3):161–182 (1999)

[23] E. J. Gumbel. 'Les Valeurs extremes des distributions statistiques'. In 'Annales de l'I.H.P.', pagine 115–158. tomo 5 (1935)

[24] F. E. Harrell e con il contributo di molti altri autori. *Hmisc: Harrell Miscellaneous* (2007). R package version 3.4-3
URL http://biostat.mc.vanderbilt.edu/s/Hmisc

[25] C.-W. Hsu, C.-C. Chang, e C.-J. Lin. 'A Practical Guide to Support Vector Classification'. Rapporto tecnico, National Taiwan University, Dept of Computer Science (maggio 2008)
URL http://www.csie.ntu.edu.tw/~cjlin

[26] E. H. Isaaks. *The Application of Monte Carlo Methods to the Analysis of Spatially Correlated Data*. Tesi di dottorato, Stanford University, Stanford, CA (1990)

[27] E. H. Isaaks e R. M. Srivastava. *An Introduction to Applied Geostatistics*. Applied Geostatistics Series. Oxford University Press (1989)

[28] A. G. Journel. 'Non-parametric Estimation of Spatial Distributions'. *Mathematical Geology*, 15:445–468 (1983)

[29] A. G. Journel e C. J. Huijbregts. *Mining Geostatistics*. Academic Press, London (1981)

[30] P. J. R. Jr e P. J. Diggle. 'geoR: a package for geostatistical analysis'. *R-NEWS*, 1(2):14–18 (June 2001). ISSN 1609-3631
URL http://CRAN.R-project.org/doc/Rnews/

[31] M. Kanevski e M. Maignan. *Analisys and Modelling of Spatial Environmental Data*. EPFL Press (2004). ISBN 2-940222-02-9

[32] R. Koenker. *quantreg: Quantile Regression* (2008). R package version 4.17
URL http://www.r-project.org

[33] R. Koenker e G. Basset. 'Regression Quantile'. *Econometrica*, 46(1):35–50 (1978)

[34] C. Lantuejoul. *Geostatistical Simulation. Models and Algorithms*. Springer, Berlin Heidelberg (2002)

[35] N. J. Lewin-Koh, R. Bivand, contributions by Edzer J. Pebesma, E. Archer, S. Dray, D. Forrest, P. Giraudoux, D. Golicher, V. G. Rubio, P. Hausmann, T. Jagger, S. P. Luque, D. MacQueen, A. Niccolai, e T. Short. *maptools: Tools for reading and handling spatial objects* (2008). R package version 0.7-13

[36] T. Masters. *Practical Neural Network Recipies in C++*. Academic Press (1993)

[37] G. Matheron. 'Principles of Geostatistics'. *Economic Geology*, 58:1246–1266 (1963)

[38] G. Matheron. *La Theorie des Variables Ragionalisee et ses Applications*. Masson, Paris (1965)

[39] J. A. Mazanec e H. Strasser. *A Nonparametric Approach to Perceptions-based Market Segmentation: Foundations*. Springer, Berlin (2000)

[40] J. P. Mc Laughlin. 'Approaches to the Assessment of Long Term Exposure to Radon and its Progeny'. *The Science of the Total Environment*, 272:53–60 (2001)

[41] E. J. Pebesma. 'Multivariable geostatistics in s: the gstat package'. *Computers & Geosciences*, 30:683–691 (2004)

[42] S. Pegoretti. *Mappatura del radon in Alto Adige: un approccio geostatistico*. Tesi di laurea, Univeristà degli Studi di Trento (2005)

[43] D. Posa. *Introduzione alla Geostatistica*. Adriatica Editrice Salentina, Lecce (1995)

[44] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2008). ISBN 3-900051-07-0
URL http://www.R-project.org

[45] D. Sarkar. *lattice: Lattice Graphics* (2008). R package version 0.17-13

[46] M. Schlather. *RandomFields: Simulation and Analysis of Random Fields* (). R package version 1.3.30
URL http://www.stochastik.math.uni-goettingen.de/institute/index.php

[47] K. Schliep e K. Hechenbichler. *kknn: Weighted k-Nearest Neighbors* (2007). R package version 1.0-5

[48] C. E. Shannon. 'A mathematical theory of communication'. *Bell System Technical Journal*, 27:379–423,623–656 (1948)

[49] A. G. Stephenson. 'evd: Extreme value distributions'. *R News*, 2(2):31–32 (June 2002)
URL http://CRAN.R-project.org/doc/Rnews/

[50] M. Theus. 'Interactive data visualization using mondrian'. *Journal of Statistical Software*, 7(11):1–9 (11 2002). ISSN 1548-7660
URL http://www.jstatsoft.org/v07/i11

[51] C. Thomas-Agnan, Y. Aragon, A. Ruiz-Gazen, T. Laurent, e L. Robidou. *GeoXp: Interactive exploratory spatial data analysis* (2008). R package version 1.2
URL http://w3.univ-tlse1.fr/GREMAQ/Statistique/SP_etheme3.pdf

[52] L. Verdi e S. Pegoretti. 'Radon mapping in South-Tyrol: a geostatistical analysis'. In '4th Dresden Symposium "Survey of Geo-Hazards"', (26-30 settembre 2005)

[53] L. Verdi e S. Pegoretti. 'Mappatura del radon in Alto Adige: un'analisi di tipo geostatistico'. In 'III Convegno nazionale "Controllo ambientale degli agenti fisici: dal monitoraggio alle azioni di risanamento e bonifica"', Biella (7–9 giugno 2006)

[54] I. H. Witten e E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, seconda edizione (2005)
URL http://www.cs.waikato.ac.nz/ml/weka/