SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di

STATISTICA E METODI QUANTITATIVI

Dottorato di Ricerca in STATISTICA E MATEMATICA PER LA FINANZA Ciclo XXIX

Curriculum in STATISTICA METODOLOGICA

TITOLO TESI

Una specificazione semiparametrica del modello di regressione M-Quantile ad effetti casuali con applicazioni a dati ambientali georeferenziati

Cognome CARCAGNI' Nome ANTONELLA

Matricola 787792

Tutore : RICCARDO BORGONI

Coordinatore: GIORGIO VITTADINI

ANNO ACCADEMICO 2015/2016

INDICE

1. INTRODUZIONE 8

Obiettivo e contenuto della tesi 8

2. I MODELLI DI REGRESSIONE QUANTILE ED M-QUANTILE PER DATI GERARCHICI 13

2.1 Il modello di regressione quantile 16

2.2 Il modello di regressione quantile per dati raggruppati 18

2.3 Il modello di regressione M-Quantile ed M-Quantile ad effetti casuali 23

3. SPLINE 28

3.1 Smoothing 28

3.2 Le spline come strumento matematico per l'approssimazione di funzioni 29

3.3 Smoother basati sulle spline 31

3.4 La regressione con spline penalizzate 33

3.5 Selezione del parametro di smoothing λ e gradi di libertà 37

3.6 Spline come modelli misti 38

3.7 Smoothing spline bivariato 41

3.8 Low-Rank Radial Smoothers 45

4. MODELLO SEMIPARAMETRICO M-QUANTILE AD EFFETTI CASUALI 47

4.1 Modello di regressione quantile non parametrico 47

4.2 Dal modello M-Quantile non parametrico al modello semi-parametrico M-Quantile 49

4.3 Modello semiparametrico M-Quantile ad effetti casuali 51

5. DISEGNO DI SIMULAZIONE BASATO SUL MODELLO 56

5.1 Primo disegno di simulazione 56

5.2 Risultati 59

5.3 Secondo Disegno di simulazione 62

5.4 Risultati 64

5.5 Terzo Disegno di simulazione 67

5.6 Risultati 68

6. ANALISI DELLE CONCENTRAZIONI DEL RADON INDOOR IN LOMBARDIA TRAMITE IL MODELLO M-QUANTILE 70

6.1 Introduzione 70

6.2 Il radon indoor in Lombardia 72

6.3 Dataset e statistiche descrittive della concentrazione di radon indoor in Lombardia 73

6.4 Modelli M-Quantile per IRC 81

6.5 Analisi preliminari 82

6.6 Effetto della distanza dalla faglia: modello con spline univariata 84

6.7 Analisi dei trend spaziali 86

6.7.1 Mappe di IRC stimato 867

6.8 Identificazione delle Radon Prone Area (RPA) 89

6.9 Le determinanti della concertazione di IRC: modello additivo con covariate di edificio e effetto random di litologia e spline bivariata 92

6.9.1 IRC in base ai Profili abitativi 94

7. DISCUSSIONE E CONCLUSIONI 99

BIBLIOGRAFIA 103

APPENDICE 116

INDICE TABELLE

Tabella 1 Valore degli indici RMSE E MRB% per ogni valore fissato delle componenti di varianza e M-quantili **60**
Tabella 2 Valori del MASE per ciascun scenario e M-quantile (0.25, 0.50, 0.75). **64**
Tabella 3 Valore medio del coefficiente di determinazione R^2 per ciascun M-quantile **65**
Tabella 4 Rapporto tra MASE ciascuno scenario e per ciascun M-quantile **69**
Tabella 5 Caratteristiche fisiche del Radon (da ARPA) **72**
Tabella 6 Statistiche descrittive per la variabile di risposta IRC **75**
Tabella 7 Statistiche descrittive per la variabile di risposta IRC per ogni caratteristica costruttiva **76**
Tabella 8 Valori di IRC per le 11 classi litologiche **79**
Tabella 9 Stima dei parametri del modello semiparametrico M-quantile con spline univariata per gli M-quantili 0.25,0.50 e 0.75 **85**
Tabella 10 Stima dei parametri per gli M-quantili 0.25, 0.50 e 0.75 **87**
Tabella 11 Parametri stimati per il modello Radon Prone Areas **90**
Tabella 12 Stima dei parametri del modello per gli M-quantili 0.25, 0.50 e 0.75 **93**
Tabella 13 Profili abitativi a minor rischio per l'M-quantile 0.25 **95**
Tabella 14 Profili abitativi a maggior rischio per l'M-quantile 0.25 **95**
Tabella 15 Profili abitativi a minor rischio per l'M-quantile 0.50 **98**
Tabella 16 Profili abitativi ad elevato rischio per l'M-quantile 0.50 **96**
Tabella 17 Profili abitativi a basso rischio per l'M-quantile 0.75 **97**
Tabella 18 Profili abitativi ad elevato rischio per l'M-quantile 0.75 **97**

INDICE DELLE FIGURE

Figura 1 Confronto tra i due modelli: modello ad effetti fissi ed modello misto. La linea tratteggiata è la funzione vera mentre la linea solida è la funzione stimata (tratta da Ruppert et al., 2003 pag.110). 40
Figura 2 I cerchi neri (o) rappresentano le coordinate spaziali (lat=latitudine e lon=longitudine) dei siti ottenuti alla prima delle 1000 iterazioni del dataset 56
Figura 3. Posizione dei nodi della funzione spline radiale bivariata 58
Figura 4 Grafico delle relazioni funzionali 63
Figura 5 Boxplot dei coefficienti di determinazione R^2 per ciascun scenario e M-quantile 66
Figura 6 Grafici delle relazioni funzionali univariate: a) a salti, b) esponenziale e c) ciclica 67
Figura 7 Locazione spaziale dei punti di misura (a) e locazione spaziale dei punti di misura il cui valore supera i 300 Bq/m^3 (b) 74
Figura 8 Boxplot della concentrazione di Radon indoor (IRC) in ciascuna delle province lombarde 75
Figura 9 Grafico della IRC per caratteristiche abitative e materiali da costruzione 77
Figura 10 Classificazione delle 11 classi litologiche (tratta da Borgoni et al., 2010) 79
Figura 11 IRC per le classi litologiche Depositi Fini, Detriti, Morene, Calcari 80
Figura 12 Lineamenti di faglia della Regione Lombardia 82
Figura 13 Istogramma della IRC (y) 83
Figura 14 a e b) Normal Q-Q Plot dei residui di primo e secondo livello 85
Figura 15 Distanza di Cook per le osservazioni a e per le classi litologiche b 84
Figura 16 Boxplot di IRC per classe litologica 84
Figura 17 Grafico del IRC (y) vs distanza dalla faglia: linea verde per l'M-quantile 0.25; linea rossa per l'M-quantile 0.50 e linea blu per l' M-quantile 0.75 86
Figura 18 I 1000 punti di predizione: i cerchi neri ricadono nel territorio regionale mentre i cerchi rossi si collocano al di fuori della Regione 88
Figura 19 Mappe per gli M-quantili 0.25, 0.50 e 0.75 89
Figura 20 Superficie dell' 85-esimo M-quantile di IRC 91
Figura 21 Mappe Radon Prone Areas per i valori di riferimento 200 Bq/m3 (a) e 300 Bq/m3 (b) 91

A Temistocle

RINGRAZIAMENTI

Un ringraziamento particolare al prof. Riccardo Borgoni che mi ha seguito costantemente in questo lavoro e durante l'esperienza del dottorato di ricerca. Inoltre, un sentito ringraziamento al dott. Davide Guido per l'aiuto materiale datomi in questi anni di dottorato; dottorato che non avrei mai iniziato e frequentato senza la sua decisiva presenza e supporto. Inoltre ringrazio me stessa per la determinazione dimostrata in questi tre anni di esami e ricerca.

1. INTRODUZIONE

Obiettivo e contenuto della tesi

Questo lavoro di tesi ha come finalità lo sviluppo e l'implementazione di un modello semiparametrico M-quantile ad effetti random che sia in grado di cogliere l'eventuale presenza di un trend spaziale nei dati ambientali in particolare quelli relativi alla concentrazione di radon indoor. Il modello proposto è una estensione del modello M-quantile ad effetti casuali di base in cui è stata inclusa una componente spaziale.

La componente spaziale è modellata combinando insieme un'intercetta random (Chambers e Tzavidis, 2006) che coglie l'effetto del gruppo e un termine semiparametrico per catturare la regolarità residua nello spazio (Pratesi et al. 2009). Quest'ultima componente è trattata mediante una spline bivariata delle coordinate geografiche dei siti di campionamento. Come proposto da Rupert et al. (2003), i coefficienti dei nodi della spline bivariata sono trattati come effetti random.

L'approccio di massima verosimiglianza robusta (Richardson and Welsh, 1995) e un metodo sequenziale a due stadi è stato adottato per ottenere la stima dei parametri del modello (Tzavidis et al., 2016).

Il modello sviluppato è stato applicato ai dati di concentrazione del radon in locali chiusi raccolti in Lombardia con lo scopo di individuare le determinanti della concentrazione del radon e per identificare le aree geografiche più esposte ad elevati valori di tali concentrazioni.

Il radon è un gas nobile radioattivo naturalmente presente nelle rocce, dalle quali diffonde in atmosfera raggiungendo valori di concentrazione solitamente molto bassi (Vaupotic et al., 2010). Il tasso di rilascio di radon dalle rocce e dal suolo è fortemente condizionato dalla concentrazione di uranio e radio (entrambi, sono elementi radioattivi progenitori del radon) contenuto in esse, dalla granulometria, dal tipo di minerali nel quale l'uranio si trova (Appleton, 2013). Diversa invece la situazione in cui questo gas si accumula in ambienti confinati quali gli edifici, in tali situazioni le concentrazioni possono raggiungere valori anche superiori ai 4000 $Bq \bullet m^{-3}$ e gli effetti sulla salute (supportati da evidenze epidemiologiche (Kreienbrock et al., 2001; Darby et al., 2005; Krewski et al., 2005)) possono risultare importanti. Il radon penetra negli edifici, ad esempio, attraverso aperture o microfratture presenti nelle fondamenta, l'abitazione stessa intrappola il gas radioattivo limitandone la dispersione in atmosfera e per questo motivo la sua concentrazione aumenta. Fonti secondarie di radon, riconosciute, sono i materiali da costruzione degli edifici queste con un effetto additivo o esclusivo rilasciano radon nell'ambiente confinato. Il Radon che si accumula nell'edificio può

decadere nei suoi sottoprodotti radioattivi, Polonio 218 e Polonio 214, solidi e dotati di carica elettrica che gli rende capaci di legarsi alle componenti dell'aerosol presenti nell'aria, le quali, se inalate possono raggiungere i polmoni (Field, 2015; US-EPA, 2003; WHO, 2010). Una volta raggiunti i polmoni, i due isotopi instabili del Polonio iniziano a loro volta a decadere rilasciando particelle alfa, la cui energia fornisce una dose radiologicamente significativa idonea a danneggiare il DNA dei tessuti epiteliali, che se non riparato dai meccanismi cellulari potrebbe innescare la patologia oncologica (ICRP (Recommendations of the International Commission on Radiological Protection),1991).

L'Agenzia Internazionale per la Ricerca sul Cancro (IARC) ha inserito il Radon nelle categorie di cancerogenicità di Gruppo1 ("evidenza sufficiente di cancerogenicità per l'uomo").

Anche, recenti studi epidemiologici hanno concluso che esiste una associazione tra l'esposizione al Radon indoor e l'insorgenza della patologia oncologica. Inoltre, in Europa, sono imputabili al Radon circa 9% delle morti per tumore al polmone e il 2% di tutte le morti per cancro (Darby et al., 2005). Queste percentuali ovviamente, possono variare a seconda della nazione presa in considerazione (Kim et al., 2016). Ad esempio, in Italia è stato stimato dall'Istituto Superiore di Sanità che il radon provoca dal 5 al 20% dei casi di tumore al polmone, pari cioè a da 1.500 a 5.500 casi ogni anno. In sostanza dopo il fumo di tabacco è la seconda causa di carcinoma polmonare (McColl et al., 2015).

Per questo, la comprensione del fenomeno e la sua modellizzazione hanno assunto negli ultimi anni una forte valenza socio-sanitaria, soprattutto in relazione agli effetti cancerogeni di questo gas sulla salute umana.

Negli ultimi due decenni, la presa di coscienza del rischio indotto dal Radon indoor ha portato le istituzioni governative e intergovernative a promulgare delle normative in materia di radioprotezione dall'esposizione al Radon.

In Italia una normativa per la protezione dall'esposizione al Radon nelle abitazioni non è stata ancora emanata, ma una di tipo cogente, entrata in vigore nel 2001 con il D.Lgs 241/00, (che ha modificato il D.Lgs 230/95) esiste solo per i luoghi di lavoro. Tale decreto prevede, tra l'altro, l'obbligo da parte dell'esercente di misurare la concentrazione di radon in tutti i locali sotterranei e, nel caso questa superi i 500 Bq/m^3 (livello d'azione), di valutare in maniera più approfondita la situazione e di intraprendere azioni di bonifica, nel caso i cui i locali siano sufficientemente frequentati da lavoratori. Tuttavia la protezione dal radon indoor nelle abitazioni è prevista nella nuova direttiva europea 2013/59/Euratom in materia di protezione dalle radiazioni ionizzanti, approvata il 5 dicembre 2013 (che abroga le direttive 89/618/Euratom, 90/641/Euratom, 96/29/Euratom, 97/43/Euratom e 2003/122/Euratom), che dovrà quindi essere obbligatoriamente recepita nella normativa italiana entro febbraio 2018. Tale direttiva prevede che gli Stati Membri dell'Unione Europea adottino un livello

di riferimento di concentrazione di radon non superiore a 300 Bq/m^3 fermo restando che gli stati comunitari sono liberi di mantenere il livello di riferimento adottato nella loro normativa se prevede un valore più restrittivo.

Con questa Direttiva si richiede anche di definire un piano d'azione nazionale che affronta i rischi di lungo termine dovuti alle esposizioni al radon nelle abitazioni, negli edifici pubblici e nei luoghi di lavoro per qualsiasi fonte di radon, sia essa il suolo, i materiali da costruzione o l'acqua. Gli Stati Membri dovranno anche promuovere interventi volti a individuare le abitazioni che presentano concentrazioni di radon (media annua) superiori al livello di riferimento e, nel caso, incoraggiano, con strumenti tecnici o di altro tipo, misure di riduzione della concentrazione di radon. Nella direttiva 96/29/Euratom e nel D.lgvo 241/00 di recepimento di tale direttiva, queste aree erano dette Radon Prone Areas anche se non esisteva (e non esiste ancora adesso) un criterio di identificazione univoco atto a definirle e di individuarle. L'obiettivo dell'individuazione delle Radon Prone Areas è quello di aiutare le autorità locali e nazionali a sviluppare una strategia adeguata per ridurre al minimo l'esposizione a questo cancerogeno, per migliorare la qualità della vita e la salute della popolazione.

Il rischio di esposizione può essere mitigato ricorrendo alle migliori pratiche e tecniche edilizie esistenti (e.g. Long et al., 2013; US EPA, 2001) per tanto una volta mappate le aree del territorio, a cui è associata una alta probabilità di superare il valore di riferimento stabilito per legge, questo strumento potrà (dovrà) essere consultato dai proprietari o dalle imprese di costruzione affinchè possano mettere in atto le "azioni di rimedio" più adeguate. Nel caso si sia pianificato di costruire nuovi edifici, queste mappe possono essere anche utilizzate per identificare le aree su cui è opportuno intraprendere misure preventive (McColl et al., 2015). Perciò, l'elaborazione di mappe ad alta risoluzione, accurate e statisticamente robuste hanno un grande risvolto sia sociale che economico oltre ad essere essenziali per sviluppare e attuare gli approcci più efficienti per ridurre l'esposizione della popolazione.

In Italia dal 1989 al 1997 è stata realizzata una campagna di misura nazionale per valutare l'esposizione al radon della popolazione, organizzata dall'ISPRA (ex APAT) e dall'istituto superiore di sanità (ISS) in collaborazione con le strutture regionali competenti. L'indagine sui i livelli di concentrazione ha permesso di rilevare le concentrazioni medie annuali di radon nelle abitazioni e di stimare il numero di abitazioni in cui la concentrazione supera determinati livelli. Le misure sono state condotte, in alcuni comuni di ogni regione, in un totale di circa 5361 abitazioni.

La media annuale nazionale della concentrazione di radon è risultata pari a 70 Bq/m^3, superiore a quella mondiale che è stata stimata intorno a 40 Bq/m^3. Nel 4,1 % delle abitazioni si è misurata una concentrazione superiore a 200 Bq/m^3, e nello 0.9% una concentrazione superiore a 400 Bq/m^3

(rispetto i livelli d'azione individuati e stabiliti nella raccomandazione dell'Unione Europea 90/143/Euratom).

In Lombardia, così come nel Lazio, sono state riscontrate le più elevate concentrazioni di radon. Da un punto di vista legislativo le Regioni hanno il compito di definire le zone con probabilità alta di concentrazione di radon (radon prone area), sulla base di dati già disponibili e dei risultati di apposite campagne di monitoraggio.

In ottemperanza alla normativa, nel 2003 in Lombardia è stata svolta una prima campagna di misura su scala regionale (svolta con una collaborazione tra ARPA Lombardia e i Dipartimenti di Prevenzione delle AASSLL), allo scopo di individuare le aree del territorio lombardo con la maggiore probabilità di avere alte concentrazioni di radon indoor, a cui ha fatto seguito, nel 2009-2010 la realizzata di una seconda campagna regionale.

Lo studio della concentrazione di Radon indoor è complesso in quanto contribuiscono al fenomeno molte variabili di natura diversa ma fortemente interagenti, le quali possono essere raggruppate in tre categorie: spaziali e geografiche (geologiche, litologiche e pedologiche) (Bossew et al., 2013; Cinelli et al., 2010; Friedmann e Bossew, 2010; Ielsch et al., 2010; Kemski et al., 2009; Miles e Appleton, 2005,2010; Tapia et al., 2006), temporali (meteorologiche e influenze antropogeniche) (Bossew e Lettner, 2007; Burke et al., 2010; Denman et al., 2007; Groves-Kirkby et al., 2006; Miles, 2001), e caratteristiche architettoniche degli edifici (età della costruzione, materiali edili, tipo di fondamenta, tipo di edificio ecc) (Friedmann, 2005; Friedmann e Groeller, 2010; Girault e Perrier, 2012a,b; Kemski et al., 2009). Il dataset impiegato si compone di un sottocampione di 900 valori georeferenziati di concentrazione di radon indoor (media annuale) misurati all'interno di abitazioni, della regione Lombardia durante una delle prime campagna di monitoraggio.

Il presente lavoro è strutturato nel seguente modo: nel capitolo 2 viene proposta una review del modello quantile e M-quantile sia per dati indipendenti che per dati gerarchici; nel capitolo 3 vengono presentate le spline univariate e bivariate; nel capitolo 4 viene descritto il modello Semiparametrico M-quantile ad effetti random, sviluppato in questo lavoro di tesi e in sezione 5 vengono riportati i risultati dello studio di simulazione mirato a verificare la prestazioni di stima e predittive del modello proposto e a confrontarne la performance con il modello non-parametrico M-Quantile P-spline discusso da Pratesi et al. (2009); nel capitolo 6 viene riportata l'applicazione su dati di concentrazione di radon indoor (IRC); nel capitolo 7 sono riportate le considerazioni conclusive su questo lavoro di tesi.

Le appendici riguardano: 1) alcune definizioni matematiche e concetti di base relativi al modello di regressione quantile e M-quantile; 2) descrizione dell'algoritmo di stima dei parametri del modello; 3) tabelle complete dei profili di rischio abitativo individuati per i diversi M-quantili considerati.

2. I MODELLI DI REGRESSIONE QUANTILE ED M-QUANTILE PER DATI GERARCHICI

Nello studio dei fenomeni ambientali accade spesso di disporre di dati organizzati in una struttura di tipo gerarchico, essendo gli stessi classificati, in via naturale e/o in modo funzionale all'analisi, in classi o gruppi. Il tipo di dati raccolti riflettono il disegno di campionamento attuato per raccogliere le informazioni sui parametri della matrice ambientale o sul contaminante di interesse. Spesso, qualunque sia il fenomeno ambientale in esame non si può prescindere della componente spaziale e/o temporale per avere una corretta interpretazione.

I modelli multilevel, quali metodologie statistiche più adatte a dedurre le informazioni presenti all'interno delle strutture gerarchiche, tengono conto in maniera esaustiva sia della presenza di relazioni tra le variabili appartenenti ad ogni livello sia delle relazioni tra i livelli differenti, considerando in tal modo l'effetto netto sulle unità e le interazioni in esse presenti (Kreft e De Leeuw, 1998; Snijders e Bosker, 1999).

L'obiettivo di questi modelli è quello di modellare il valore atteso condizionato di una variabile di risposta *y* dato un insieme di covariate **x**, tenendo al contempo conto delle struttura di dipendenza insita nei dati.

Tuttavia, limitarsi al valore atteso come unico parametro di locazione potrebbe non garantire una completa descrizione della distribuzione condizionata. Inoltre, alcune covariate potrebbero esercitare un'influenza diversa sui quantili della distribuzione condizionata. Non di rado, specialmente in un contesto ambientale, l'interesse cade sulla code della distribuzione, ad esempio, là dove si vanno a posizionare specifici valori soglia (ad es. contaminanti in matrici ambientali il cui valore limite è regolamentato dalla legislazione (Radon, PCB e diossine, metalli pesanti ecc.) oppure nello studio della distribuzione di specie in un habitat in funzione di fattori limitanti (temperatura, pH, risorse, ecc).

Un altro aspetto fondamentale riguarda la presenza di outliers nei dati che potrebbe inficiare la stima dei parametri del modello, in questo caso sarebbe più opportuno applicare un approccio robusto.

Il modello di regressione quantilica (QR), sviluppato per dati indipendenti da Koenker e Bassett (1978), fornisce uno strumento analitico alternativo che nel corso degli anni è divenuto un approccio molto popolare e affermato nella letteratura statistica.

Data la crescente popolarità, il modello di regressione quantilica ben presto è stato esteso ai dati gerarchici in particolare longitudinali in cui i cluster sono i soggetti su cui vengono rilevate più misure ripetute.

Una delle prime proposte è stato il modello di regressione sulla mediana di Jung (1996), che include una struttura di correlazione tra misure ripetute e si basa sull'approccio di quasi-verosimiglianza. Koenker (2004) ha sviluppato un modello ad intercetta casuale per dati longitudinali e ha proposto un metodo di stima conosciuto come ℓ_1 penalizzato, che risulta meno rigido nelle assunzioni rispetto ad altre alternative. Negli anni successivi, tra gli altri, Geraci e Bottai (2007; 2014) e Liu e Bottai (2009), hanno presentato il modello di regressione quantilica ad effetti random basato sulla distribuzione asimmetrica di Laplace.

Circa dieci anni dopo lo sviluppo e la presentazione del metodo di regressione quantilica, Breckling e Chambers (1988) hanno proposto il modello di regressione M-quantile che è sostanzialmente una generalizzazione di tipo quantilico del modello di regressione-M (M si riferisce alla tipologia di stimatore, si veda allegato B), basato sull'utilizzo della funzione di influenza. Questo approccio può essere definito come una combinazione del modello di regressione quantilico ed *expectile* (Newey e Powell, 1987) che integra in un unico modello le proprietà di robustezza ed efficienza che caratterizzano questi due ultimi modelli. Infatti, specificando una determinata funzione di perdita si possono ottenere come casi particolari il modello di regressione expectile, M-Quantile e quantile.

Il modello di regressione M-Quantile come sviluppato da Breckling e Chambers (1988), non tiene conto dell'effetto del raggruppamento nell'analisi dei dati gerarchici. Questo aspetto è stato sviluppato successivamente da Tzavidis et al. (2010, 2016) estendendo il modello lineare M-Quantile per i quantili di una distribuzione condizionata mediante l'inclusione del termine di intercetta casuale. Il modello di regressione ad effetti random M-quantile per dati gerarchici permette di stimare i parametri di un modello ad intercetta random a due livelli Tzavidis et al. (2016) e a tre livelli (Borgoni et al., 2016). Quando la forma della relazione funzionale tra la variabile di risposta e le covariate non è nota oppure manifesta degli andamenti complessi adottare un approccio semiparametrico o non parametrico si dimostra più vantaggioso. In questa tipologia di modelli il ruolo molto importante è giocato dalle spline. La spline può essere inclusa all'interno del modello di regressione lineare con lo scopo di rendere il predittore lineare più flessibile e capace di cogliere andamenti locali della variabile di risposta del modello al variare della o delle variabili esplicative. Il modello di regressione diviene quindi capace di trattare effetti non lineari delle esplicative sulla risposta, modellizzare effetti che si modificano al variare dell'esplicativa sul suo supporto e stimare eventuali punti di cambio della relazione. La regressione spline basata sulla penalizzazione è un metodo comunemente usato per stimare complesse relazioni non-lineari tra due variabili. La bontà dell'adattamento del modello ai dati dipende dal numero e dalla posizione dei nodi e dal parametro di smoothing.

In letteratura sono stati proposti molti modelli che combinano il modello ad effetti misti e i modelli di regressione non parametrica rappresentata dalla spline (Parise et al. (2001), Coull et al. (2001), Coull et al. (2001a) e Opsomer (2008)).

Il termine di spline come suggerito in Ruppert et al. (2003) può essere considerato con un ulteriore effetto casuale e per tanto gestibile all'interno del modello misto stesso.

Di recente tali modelli sono stati estesi anche all'approccio quantilico ed M-quantile sia per dati continui che di conteggio (Pratesi et al. (2008), Pratesi et al. (2009) e Dreassi et al. (2014) di cui viene riportata una breve trattazione nel capitolo 4.

In questo capitolo vengono descritte le principali caratteristiche del modello di regressione quantile e M-quantile e le relative estensioni per dati gerarchici. Nel caso dei dati gerarchici, l'attenzione si è soffermata ai modelli proposti da Koenker (2004), Geraci e Bottai (2007), Tzavidis (2012; 2016).

2.1 Il modello di regressione quantile

Come già detto, l'idea di modellare i quantili della distribuzione condizionata $f(y|\mathbf{x})$ di una variabile di risposta *y* dato l'insieme di covariate **x** è stata sviluppata ed introdotta da Koenker e Bassett nel 1978.

Il modello di regressione quantile in specifici ambiti si è rivelato più appropriato rispetto al modello di regressione ordinaria. Come ricordato in precedenza infatti, in alcuni casi, l'interesse dei ricercatori è indirizzato verso le code della distribuzione e le covariate potrebbero esercitare differenti effetti sui diversi quantili della distribuzione condizionata. Tra le maggiori proprietà che hanno reso questo approccio molto consolidato e applicato in diversi ambiti di ricerca, si annoverano la robustezza rispetto all'outliers, l'efficienza per un ampio range delle distribuzioni degli errori e l'equivarianza per trasformazioni monotone.

La formulazione del modello Koenker e Basset (1978) si basa su un criterio di ottimizzazione per definire il quantile di una variabile casuale. Più precisamente, in analogia a quanto succede per la media campionaria, che può essere definita come la soluzione del problema di minimizzazione della somma di scarti al quadrato, possiamo definire ogni singolo quantile come la soluzione del seguente problema di minimo (per maggiori dettagli si veda appendice A)

$$Min_{\varepsilon_q \in R} \left\{ \sum_{i \in \{i | X_i \geq \varepsilon_q\}} q|X_i - \varepsilon_q| + \sum_{i \in \{i | X_i < \varepsilon_q\}} (1-q)|X_i - \varepsilon_q| \right\}$$

Questa formulazione risulta utile per passare direttamente all'ambito regressivo. Con riferimento alla *i*-esima unità campionaria, si consideri il modello lineare nella forma

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_q + e_i \qquad \text{con } i = 1, \dots, n$$

dove *n* è la numerosità campionaria, y_i è la variabile risposta rilevata sull'unità *i*-esima, $\mathbf{x}_i$ è un vettore di p variabili esplicative con prima componente pari a **1** per garantire la presenza dell'intercetta, $\boldsymbol{\beta}$ è il vettore dei coefficienti fissi del modello e e_i è il termine d'errore. L'assunzione $Q_q(e_i|y_i, \mathbf{x}_i) = 0$ è necessaria per garantire che la distribuzione degli errori sia centrata sul *q*-esimo

quantile $(0 \leq q \leq 1)$. Un modello lineare per il q-esimo quantile della variabile risposta y_i condizionatamente alle variabili esplicative $\mathbf{x}_i$ può essere quindi scritto come

$$Q_q(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_q.$$

Il vettore dei parametri $\boldsymbol{\beta}_q$ viene stimato minimizzando

$$\sum_{i=1}^{n} \rho_q(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_q)$$

dove $\rho_q(u)$ è la funzione di perdita

$$\rho_q(u) = \begin{cases} qu & u > c \\ (1-q)u & u < c \end{cases}$$

Ovvero

$$\widehat{\boldsymbol{\beta}}_q = \underset{\beta_i}{\operatorname{argmin}} \sum_{i:y_i \geq \mathbf{x}_i^T \boldsymbol{\beta}_q} q|\, y_i - \mathbf{x}_i^T \boldsymbol{\beta}_q| + \sum_{i:y_i \leq \mathbf{x}_i^T \boldsymbol{\beta}_q} (1-q)|y_i - \mathbf{x}_i^T \boldsymbol{\beta}_q|$$

risolvendo un problema di minimo che è espresso come un problema di programmazione lineare (LP) (Buchinsky, 1998). L'algoritmo del simplesso (Barrodale e Roberts, 1974) è usato per risolvere il problema LP in ambito della QR di moderate dimensioni. Questo algoritmo purtroppo si è rivelato molto lento quando il numero di osservazioni *n* è elevata (*n* >100000) (Chen and Wei 2005) pertanto Portnoy and Koenker (1997) hanno proposto il metodo del punto interno che si è mostrato più efficiente. Di recente un approccio euristico (*finite smoothing algorithm*) è stato suggerito da Chen (2004, 2007) che in presenza di un elevato numero di covariate si è rivelato più veloce e accurato.
Alcuni autori tra i quali Koenker e Machado (1999) e Yu e Moyeed (2001) teorizzarono il legame che intercorre tra la minimizzazione della funzione di perdita per i quantili e la teoria di verosimiglianza attraverso la distribuzione asimmetrica di Laplace (ALD). Per meglio comprendere questo aspetto e anche la trattazione che segue sui modelli di regressione per dati raggruppati (Geraci e Bottai 2007;2014) definiamo la ALD.
Una variabile casuale *Y*~ALD (μ, σ, τ) ha la funzione di densità di probabilità della forma

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}$$

dove $\rho_\tau(u)$ è la funzione di perdita, $0 \leq \tau \leq 1$ è il parametro di asimmetria, $\sigma > 0$ è il parametro di scala e $-\infty \leq \mu \leq +\infty$ è il parametro di locazione.

Yu et al. (2001) hanno presentato una definizione alternativa del modello di regressione quantilica proprio utilizzando la funzione di distribuzione asimmetrica di Laplace. La funzione di perdita assegna un peso τ e $1-\tau$ rispettivamente alle osservazioni maggiori e/o minori di μ e inoltre si può notare che $P(y \leq \mu) = \tau$.

Se adesso definiamo $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ assumendo che $y_i \sim ALD(\mu_i, \sigma, \tau)$ la verosimiglianza per n osservazioni indipendenti è

$$L(\boldsymbol{\beta}, \sigma; \boldsymbol{y}, \tau) \propto \sigma^{-1} \exp\left\{-\sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \mu_i}{\sigma}\right)\right\}$$

Ottimizzando rispetto a β le stime che si ottengono per tali parametri sono equivalenti a quelle che si ottengono minimizzando la funzione obiettivo vista all'inizio della trattazione.

Dopo la sua apparizione, il modello di regressione quantilico è stato oggetto di un continuo e sempre crescente sviluppo, sia a livello teorico che empirico. Attualmente, il suo ambito di applicazione copre tutte le aree delle scienze applicate (Yu et al., 2003). Una buona parte della letteratura sulla QR si concentra sui problemi di stima e sull'estensione di questo modello ai modelli già adottati in ambito classico ad esempio Powell (1986) ha adattato la QR per i dati censurati, Machado e Silva (2005) hanno proposto una sua generalizzazione per i dati di conteggio, Koenker e Bilias (2001), Fitzenberger e Wilke (2006) la applicano ai dati di durata. Diversi autori, dei cui lavori sarà discusso nel paragrafo successivo, l'hanno estesa ai modelli per dati longitudinali.

2.2 Il modello di regressione quantile per dati raggruppati

Come spesso accade in molti ambiti di ricerca (epidemiologico, ambientale, sociale ed economico) i dati derivanti da determinati disegni di campionamento presentano una struttura di dipendenza. In questo caso la struttura di dipendenza deve essere presa in considerazione per evitare di ottenere stime dei parametri del modello affette da bias.

Prima di introdurre il modello di regressione quantilico ad effetti casuali, si fornisce di seguito una breve descrizione dei modelli lineari ad effetti misti nella loro versione standard per modellare i dati gerarchicamente strutturati. Nel seguito si suppone di disporre di *n* osservazioni della variabile di risposta *y* raggruppate in *m* gruppi. Si indichi la dimensione del gruppo *j* (*j*=1,….,*m)* con n_j.

Un modello misto ad effetti casuali è specificato dalla seguente equazione:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \boldsymbol{u}_j + e_{ij}$$

dove

y_{ij} è il vettore dei valori assunti dalla *i*-esima unità nel *j*-esimo gruppo, con *i*=1,…, n_j e *j*=1,….,*m*. $\mathbf{x}_{ij}^T$ è un vettore di covariate, di dimensione *p*, associate all'unità *i* nel gruppo *j*, $\boldsymbol{\beta}$ è il vettore $p \times 1$ dei coefficienti di regressione e $\boldsymbol{z}_{ij}^T$ è il vettore $m \times 1$ dell'indicatori del gruppo, usato per definire la parte random del modello. Inoltre, $\boldsymbol{u}$ è il vettore $m \times 1$ di effetti random ascrivibili al gruppo e e_{ij} è l'effetto random individuale. Si assuma anche che $\boldsymbol{u} \sim N(0, \sigma_u^2 \boldsymbol{I}_m)$ e $e_{ij} \sim N(0, \sigma_e^2)$ e tra loro indipendenti.

Per stimare i parametri incogniti, $\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2$, del modello si utilizza la stima di massima verosimiglianza basata sulla distribuzione marginale di *y* (Harville 1979) la cui funzione di log-verosimiglianza è:

$$l(\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$$

dove $\mathbf{V} = \boldsymbol{\Sigma}_e + \mathbf{Z}\boldsymbol{\Sigma}_u\mathbf{Z}^{\mathbf{T}}$, $\boldsymbol{\Sigma}_e = \sigma_e^2\mathbf{I}_n$, $\boldsymbol{\Sigma}_u = \sigma_u^2\mathbf{I}_m$ e $\mathbf{Z}$ è una matrice $n \times m$ di costanti note e positive. La funzione di perdita quadratica utilizzata presume che le assunzioni di normalità siano rispettate ma spesso la presenza di outliers rendono le stime dei parametri del modello affette da bias ed inefficienti (Richardson e Welsh, 1995).

Nei modelli lineari multilivello, si stima il valore atteso condizionato della variabile di risposta *y* tenendo conto della struttura gerarchica dei dati ma, come ogni modello di regressione, esso non caratterizza l'intera distribuzione condizionata di una variabile dipendente. Il modello di regressione quantilica nella versione sopra presentata, pur rispondendo a quest'ultima necessità, lascia irrisolto l'aspetto connesso alla struttura di dipendenza.

Negli ultimi anni, il bisogno di estendere le potenzialità del modello di regressione quantilica per osservazioni indipendenti ai dati dipendenti ha portato allo sviluppo di diversi approcci tra i quali ricordiamo: il modello ad effetti fissi (Canay, 2011), modello con variabili strumentali (Harding e Lamarche, 2009), modelli lineari misti (Geraci e Bottai, 2014), i modelli ad effetti fissi penalizzati (Koenker, 2004; Lamarche, 2010), modelli ad effetti random (Geraci e Bottai, 2007; Barry et al., 2016) .

Come menzionato, uno dei primi lavori presenti in letteratura che riguarda i modelli quantilici per dati longitudinali è dovuto a Koenker (2004), che introduce un modello di regressione quantilica ad effetti fissi e una classe di stimatori basati su un approccio l_1 penalizzato mediante il quale cerca di risolvere le difficoltà emerse nello stimare più parametri nell'ambito della programmazione lineare (Koenker e Hallock, 2000; Abrevaya and Dahl, 2008).

Ricordiamo che stiamo parlando di dati longitudinali in cui le misure ripetute sul singolo soggetto vanno a costituire un determinato cluster. Ora, assumiamo di avere i dati (misura ripetute) $\left(\mathbf{x}_{ij}^{T} y_{ij}\right)$ con $i = 1, \dots, m$ e $j = 1, \dots n_i$. Dove $\mathbf{x}_{ij}^{T}$ sono le righe dei vettori di dimensione *p* della matrice $\mathbf{X}_i$, e $\boldsymbol{y}_{ij}$ è il *j*-esimo scalare della misura (di una variabile casuale continua) effettuata sull' *i*-esimo soggetto.

La stima dei parametri viene ricondotta al seguente problema di ottimo:

$$\min_{\alpha,\beta} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \omega \rho_q \left(y_{ij} - \mathbf{x}_{ij}^{T} \beta - \alpha_i\right) + \lambda \sum_{i=1}^{m} |\alpha_i|$$

in cui ω è il peso che controlla l'influenza dell'*q*-esimo quantile sulla stima dell'effetto individuale e λ è il parametro di penalizzazione. La penalità serve a portare la stima degli effetti individuali α verso zero in modo tale da migliorare le prestazioni della stima di β. Oltre a stimare l'effetto delle covariate, la stima di α usando il termine di penalità potrebbe riflettere le conoscenze che si hanno a priori sugli effetti random (Koenker, 2004, Geraci e Bottai,2007).

I risultati che scaturiscono adottando questo metodo dipendono dalla scelta del parametro λ da qui la necessità di selezionare un valore idoneo. A tal proposito Lamarche (2006, 2010) propone di selezionarlo minimizzando la varianza asintotica stimata. Canay (2011) suggerisce di stimare i coefficienti β mediante uno stimatore a due stadi, consistente e asintoticamente normale, in grado di gestire nel primo stadio la presenza dei parametri α_i.

Il metodo di Koenker potrebbe essere non applicabile ai modelli più complessi, in quanto tratta gli effetti random come parametri fissi e un aumento della dimensionalità potrebbe creare dei problemi. Di recente Geraci e Bottai (2007), hanno proposto l'approccio basato sulla distribuzione asimmetrica di Laplace (ALD) degli errori ed effetti random gaussiani (o distribuiti come una Laplace per pervenire alla robustezza) e l'applicazione dell'algoritmo E-M per la stima dei parametri.

Assumendo una ALD comune il loro metodo vincola gli errori ad essere omoschedastici e avere una tendenza alla mediana. Gli autori aggirano il problema legato al modello di Koenker, specificando l'effetto di *location-shift* come random e considerando i parametri della loro distribuzione dipendente dal quantile *q.* In questo modo il livello ottimale di penalità viene scelto automaticamente.

Quindi Geraci e Bottai (2007) partono definendo la funzione quantilica della variabile di risposta y_{ij} tramite un modello misto come

$$G_{y_{ij}|u_j}\left(q \middle| \mathbf{x}_{ij}, u_i\right) = \mathbf{x}_{ij}^{T} \beta + u_i, \quad j = 1, \dots., n_i, i = 1, \dots., m$$

dove $G_{y_{ij|u_j}}(.) \equiv F^{-1}_{y_{ij|u_j}}(.)$ è l'inversa della funzione di distribuzione cumulata della risposta condizionata sull'effetto random *location-shift* u_i

Assumendo che y_{ij}, condizionatamente a u_i, per $j = 1, \ldots, n_i$ e $i = 1, \ldots, m$ sono indipendenti e distribuiti come un ALD

$$f(y_{ij}|\beta, u_i, \sigma) = \frac{q(1-q)}{\sigma} exp\left\{-\rho_q\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right)\right\}$$

dove $\mu_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i$ è il predittore lineare del q-esimo quantile (Geraci e Bottai, 2007).

L'effetto random induce una struttura di correlazione tra le osservazioni misurate sullo stesso soggetto. Si assume anche, che gli effetti random u_i siano mutualmente indipendenti e identicamente distribuiti in accordo ad una data funzione di densità f_u caratterizzata da un parametro φ che dipende dai quantili φ_q. Allo stesso modo si assume che e_{ij} siano indipendenti tra loro e da u_i.

Dato $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$ e $f(\boldsymbol{y}_i|\beta, u_i, \sigma) = \prod_{j=1}^{n_i} f(y_{ij}|\beta, u_i, \sigma)$ la distribuzione di densità del i-simo soggetto condizionata sull'intercetta random u_i , la funzione di densità completa di $(\mathbf{y_i}, u_i)$ per $i = 1, \ldots, m$ è:

$$f(\mathbf{y}_i, u_i|\beta, \sigma, \varphi) = f(\mathbf{y}_i|\beta, u_i, \sigma) f(u_i|\varphi)$$

dove $f(u_i|\varphi)$ è la densità di u_i e β, σ, φ sono i parametri di interesse.

Allora se $\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_m})$ e $\mathbf{u} = (u_1, \ldots, u_m)$, la densità congiunta di $(\mathbf{y}, \mathbf{u})$ basata su *m* soggetti è:

$$f(\mathbf{y}, \mathbf{u}|\beta, \sigma, \varphi) = \prod_{i=1}^{m} f(\mathbf{y}_i|\beta, u_i, \sigma)\, f(u_i|\varphi)$$

Per quanto attiene la distribuzione degli effetti casuali, se si considerano indipendenti e identicamente distribuiti come una normale $u_i \sim N(0, \varphi)$ per $= 1, \ldots, m$ la densità congiunta di $(\mathbf{y_i}, u_i)$ per i-esimo soggetto è data da:

$$f(\mathbf{y}_i, u_i|\beta, \sigma, \varphi) = f(u_i|\varphi) \prod_{j=1}^{n_i} f\left(\mathrm{y}_{ij}|\beta, u_i, \sigma\right) \left\{\frac{q(1-q)}{\sigma}\right\}^{n_i} \frac{1}{\sqrt{2\pi\varphi}} \exp\left[-\sum_{j=1}^{n_i}\left\{\rho_q\left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right)\right\} - \frac{1}{2\varphi}u_i^2\right]$$

Mentre se $u_i \sim ALD(0, \varphi, 0.5)$ la densità congiunta di $(\mathbf{y_i}, u_i)$ per i-esimo soggetto sarà data da:

$$f(\mathbf{y}_i, u_i|\beta, \sigma, \varphi) = f(u_i|\varphi) \prod_{j=1}^{n_i} f\left(\mathrm{y}_{ij}|\beta, u_{i'}, \sigma\right) \left\{\frac{q(1-q)}{\sigma}\right\}^{n_i} \frac{1}{4\varphi} \exp\left[-\sum_{j=1}^{n_i} \left\{\rho_q\left(\frac{y_{ij}-\mu_{ij}}{\sigma}\right)\right\} - \frac{1}{2\varphi} u_i^2\right]$$

La densità marginale di **y** si ottiene integrando rispetto all'effetto casuale **u**

$$f(\mathbf{y}|\beta, \sigma, \varphi) = \int_{R^N} f(\mathbf{y}, \mathbf{u}|\beta, \sigma, \varphi) d\mathbf{u}$$

Quindi l'inferenza per i parametri β, σ, φ dovrebbe basarsi sulla verosimiglianza marginale *L*

$$L(\beta, \sigma, \varphi; \mathbf{y}) = \sum_{i}^{m} L(\beta, \sigma, \varphi; \mathbf{y}_i)$$

Per stimare i parametri del modello viene suggerito l'algoritmo E-M basato sul metodo Montecarlo che consente di evitare la risoluzione di un integrale multidimensionale (Bottai Geraci, 2007) come quello definito in precedenza.

Liu e Bottai (2009) e Geraci e Bottai (2014), hanno generalizzato questo metodo per includere più effetti casuali nel modello (modello misto quantilico). Per la stima dei parametri gli autori hanno costruito un pacchetto di `R` in cui è implementata la funzione `lqmm`. La procedura `lqmm` permette di ottenere la stima dei parametri a partire dalla verosimiglianza marginale ottenuta dall'integrazione numerica della funzione di verosimiglianza rispetto all'effetto casuale.

Il metodo di integrazione suggerito è la quadratura di Gauss-Hermite se si assume che *u* sia normalmente distribuito o il metodo di quadratura di Gauss-Laguerre se invece *u* è distribuito secondo un ALD.

A questo punto la funzione di log-verosimiglianza marginale è

$$l_j(\beta, \sigma, \varphi|y_j) = logL_j(\beta, \sigma, \varphi|y_j)$$

I metodi di ottimizzazione quali l'algoritmo di Nelder-Mead o metodo del gradiente possono essere usati per ottenere le stime dei coefficienti fissi e dei componenti di varianza del modello.

Successivamente anche Smith et al. (2015) hanno presentato un'ulteriore estensione dei modelli misti in ambito quantilico. Barry et al. (2016) hanno implementato sia il modello quantile ad effetti random

che il modello expectile ad effetti random per i dati longitudinali, fornendo le proprietà asintotiche degli stimatori dei parametri del modello e suggerendo un appropriato stimatore della matrice di varianza-covarianza.
In alternativa, Farcomeni (2012) ha specificato un modello Markoviano latente per i quantili condizionati in cui gli effetti casuali possono variare nel tempo in modo tale da modellare anche la sorgente di eterogenità non osservata che evolve nel tempo. Di fatto, i termini di intercetta casuale variano nel tempo in accordo ad una catena di Markov di primo ordine omogenea.
Seguendo questa linea Marino et al. (2015) hanno sviluppato un modello misto Markoviano che considera contemporaneamente entrambe le sorgenti di eterogeneità non osservata (costante e variabile nel tempo).

2.3 Il modello di regressione M-Quantile ed M-Quantile ad effetti casuali

Ricordiamo che la regressione quantile può essere pensata come la generalizzazione di un modello di regressione sulla mediana e la regressione expectile è una generalizzazione di tipo quantile della regressione sulla media (Newey e Powell, 1987). Questi due aspetti vengono integrati in uno stesso framework che definisce una generalizzazione di tipo quantile dei modelli di regressione robusti basati sulla funzione di influenza (M-regression). La regressione M è un metodo, basato sulla funzione di influenza, introdotto da Huber (1973) per garantire la robustezza delle stime dei parametri in presenza di dati anomali. La regressione M controlla l'effetto degli outliers trattando la parte residuale la cui grandezza è maggiore di un dato cutoff c, ponendola pari a c.
La regressione M-Quantile (MQ) (Brecking e Chambers, 1988) è la generalizzazione di un modello di tipo quantile basato su una funzione di influenza.
L'M-Quantile di ordine q della densità condizionata di y dato un insieme di covariate $\mathbf{x}$, $f(y|\mathbf{x})$, è definita come la soluzione $MQ_y(q|\mathbf{x};\psi)$ dell'equazione

$$\int \psi_q\{y - MQ_y(q|\mathbf{x};\psi)\} f(y|\mathbf{x})\, dy = 0$$

ψ_q è una funzione di influenza asimmetrica, che è la derivata prima della funzione di perdita asimmetrica ρ_q (si veda appendice B).
Nel modello di regressione lineare M-Quantile per la variabile di risposta y dato un insieme di covariate $\mathbf{x}$ si assume che:

$$MQ_{y_i}(q|\mathbf{x}_i,\psi) = \mathbf{x}_i^T \boldsymbol{\beta}_{\psi q}$$

Specificando diversamente la funzione perdita asimmetrica ρ_q è possibile ottenere il modello di regressione lineare expectile (ρ_q quadratica e $q \neq 0.5$)(Newey and Powell, 1987), il modello OLS (ρ_q quadratica e $q = 0.5$), invece, con la funzione di perdita di Koenker e Bassett (1978) si ottiene il modello di regressione quantilica.

Per quanto riguarda il modello MQ si assume che la funzione di perdita asimmetrica sia quella ρ_q di Huber (Breckling and Chambers, 1988):

$$\rho_q(u) = \begin{cases} 2c|u| - c^2\{qI(y > 0) + (1-q)I(y \leq 0)\} & |u| > c \\ u^2\{qI(y > 0) + (1-q)I(y \leq 0)\} & |u| \leq c \end{cases}$$

dove *c* è una costante di tuning e *I* la funzione indicatrice. La costante *c* può essere usata per scambiare la robustezza per l'efficienza nell'adattamento del modello di regressione MQ. Con *c* positivo ma che tende a zero aumenta la robustezza ma diminuisce l'efficienza (ci si muove verso la regressione quantile), mentre con *c* grande e positivo decresce la robustezza e aumenta l'efficienza (ci si muove verso la regressione *expectile*).

Le stime di $\boldsymbol{\beta}_{\psi q}$ si ottengono minimizzando

$$\sum_{i=1}^{n} \rho_q\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{\psi q}\right)$$

Calcolando il gradiente e annullandolo si ottengono le equazioni di stima:

$$\sum\nolimits_{i=1}^{n} \psi_q(r_{iq})\mathbf{x}_i = 0$$

dove $r_{iq} = y_i - \mathbf{x}_i^T \boldsymbol{\beta}_{\psi q}$,

$$\psi_q\left(r_{iq}\right) = 2\psi\left(s^{-1} r_{iq}\right)\left\{qI\left(r_{iq} > 0\right) + (1-q)I\left(r_{iq} \leq 0\right)\right\}$$

e

$$\psi(u) = uI(-c \leq u \leq c) + c\ \mathrm{sgn}(u)I(|u| > c),$$

con $s > 0$, stima del parametro di scala. L'applicazione dell'algoritmo minimi quadrati pesati iterati porta alla soluzione dell'equazioni di stima (Kokic et al, 1987).

Il modello di regressione M-Quantile risulta un modello flessibile in quanto si possono utilizzare una vasta gamma di funzioni di influenza come ad esempio la funzione Huber, proposta Huber 2 o la funzione di Hampel .

Il modello di regressione M-quantile è stato esteso da Tzavidis et al. (2010, 2016) nel modello di regressione M-quantile ad effetti random per tener conto della struttura gerarchica dei dati quando si modellano i quantili della distribuzione condizionata $f(y|\mathbf{x})$.

Un metodo di stima robusto è stato proposto da Richardson e Welsh (1995). Questo approccio consiste nel sostituire, nella funzione di log-verosimiglianza, la funzione di perdita quadratica con una nuova funzione che cresce insieme ai residui ma ad un tasso più lento.

La funzione di log-verosimiglianza (proposta I), per i modelli ad effetti random diventa:

$$l(\boldsymbol{\beta}, \sigma_u^2, \sigma_\varepsilon^2) = -\frac{K_1}{2}\log|\mathbf{V}| - \rho(\boldsymbol{r})$$

dove $\boldsymbol{r} = \mathbf{V}^{-\frac{1}{2}}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$, ρ è la funzione di perdita di cui ψ è la derivata e $K_1 = E[e\psi(e)]$ è il fattore di correzione per la consistenza con $e \sim \mathrm{N}(0, \boldsymbol{I_n})$ e $\psi(\mathbf{r})$ e $\boldsymbol{r}^T\psi(\boldsymbol{r})$ limitate.

Un'alternativa consiste nel risolvere l'equazione di stima (log verosimiglianza) per σ_u^2 , σ_e^2 nell'ambito della stima di massima verosimiglianza robusta (Huber proposta II):

$$\frac{1}{2}\psi(\boldsymbol{r}^T)\mathbf{V}^{-\frac{1}{2}}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}\mathbf{V}^{-\frac{1}{2}}\psi(\boldsymbol{r}) - \frac{K_2}{2}tr(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}) = 0$$

dove $K_2 = E[\psi(e)\psi(e)^T]$ con $e \sim N(0, \boldsymbol{I}_n)$ (Richardson e Welsh, 1995).

Il modello di regressione M-quantile ad effetti random (MQRE) è stato proposto da Tzavidis et.al., (2010, 2015) .

Nel caso più semplice, nel modello lineare M-quantile è inclusa una intercetta random per il gruppo ed è definito come:

$$MQ_{y_{ij}}\left(q\middle|\mathbf{x}_{ij}, \boldsymbol{u}, j; \psi\right) = \mathbf{x}_{ij}{}^T\boldsymbol{\beta}_{\psi q} + \mathbf{Z}_{ij}^T\boldsymbol{u}$$

dove $\boldsymbol{u}$ è il vettore $m \times 1$ di effetti casuale di gruppo.

Uno dei principali vantaggi della M – stima è che permette di ottenere la stima robusta sia degli effetti fissi che degli effetti random.

Per ottenere le equazioni di stima si applica la stima di massima verosimiglianza robusta (Huber proposta II) pesando in modo asimmetrico i residui (Sinha e Rao, 2009) ottenendo :

$$\mathbf{X}^{\mathrm{T}}\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\boldsymbol{r_q}) = \mathbf{0}$$

$$\frac{1}{2}\psi_q(\boldsymbol{r}_q)^T\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\boldsymbol{r}_q) - \frac{K_{2q}}{2}tr\left(\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right) = 0$$

$$\frac{1}{2}\psi_q(\boldsymbol{r}_q)^T\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\boldsymbol{r}_q) - \frac{K_{2q}}{2}tr\left(\mathbf{V}_q^{-1}\right) = 0$$

$\boldsymbol{r}_q = \mathbf{U}_q^{-\frac{1}{2}}(y - \mathbf{X}\boldsymbol{\beta}_{\psi q})$ è il vettore di residui scalati le cui componenti sono r_{ijq}, $\mathbf{U}_q$ è una matrice diagonale, sulla cui diagonale principale vi sono gli elementi u_{ijq} uguali ai componenti posti sulla diagonale della matrice di covarianza $\mathbf{V}_q = \mathbf{\Sigma}_{eq} + \mathbf{Z}\mathbf{\Sigma}_{uq}\mathbf{Z}^T$ con $\mathbf{\Sigma}_{eq} = \sigma_{eq}^2\mathbf{I}_n$ e $\mathbf{\Sigma}_{uq} = \sigma_{\gamma q}^2\mathbf{I}_m$ dove $\sigma_{eq}^2, \sigma_{uq}^2$ sono i parametri di varianza MQ-specifici, $K_{2q} = E\left[\psi_q(e)\psi_q(e)^T\right]$ con $e \sim N(0, \boldsymbol{I}_n)$.

Per risolvere le equazioni è stata proposta una procedura (Tzavidis et al., 2016) basata sull'uso congiunto dell'algoritmo di Newton-Raphson e del metodo iterativo del punto fisso (Anderson, 1973). In particolare gli effetti fissi sono stimati mediante l'algoritmo di Newton- Raphson mentre i parametri di varianza sono stimati con il metodo del punto fisso. Questo approccio permette di prevenire problemi di convergenza durante la stima delle componenti di varianza riscontrabili con ricorso all'algoritmo di Newton-Raphson.

Quest'ultimo modello rappresenta una alternativa sia al modello quantilico ad effetti casuali proposto da Geraci e Bottai (2007, 2014) che al modello di regressione quantilica penalizzata di Koenker (2004).

Una ulteriore proposta è dovuta ad Alfò et al. (2017), i quali hanno esteso i modelli quantilici ed M-quantilici per dati longitudinali nel framework dei modelli a mistura finita a partire dall'approccio semiparametrico basato sui modelli a mistura finita proposto da Aitkin (1996, 1999). I componenti della mistura finita sono i cluster di individui che condividono un valore omogeneo dei parametri casuali del modello di cui la distribuzione è ignota ma approssimata mediante una distribuzione discreta. Considerare una distribuzione non parametrica per gli effetti casuali ha lo scopo di rendere il modello di regressione M-quantile robusto anche alla errata specificazione del modello (Alfò et al., 2017).

Un ulteriore sviluppo dei modelli di regressione M-quantile è il modello per dati multivariati longitudinali e continui. In questo contesto i parametri casuali discreti, individuo-specifici, in particolare sono utilizzati per tener conto sia della dipendenza entro la stessa risposta rilevata in tempi diversi che dell'associazione tra differenti risposte osservate sulla stessa unità di campionamento in un dato momento. Allo stesso tempo, è stata apportata una riparametrizzazione a livello dei predittori lineari per controllare una eventuale dipendenza tra i parametri casuali individuo-specifici e il vettore delle covariate osservate (Alfò et al., 2017). Come nel caso del modello a mistura finita dei modelli di regressione quantilica e M-quantilica, anche in questo caso la stima dei parametri del modello è stata ottenuta mediante l'applicazione dell'algoritmo E-M.

Negli ultimi anni, molti sforzi sono stati compiuti per sviluppare ulteriormente il modello M-quantilico sia a livello teorico che empirico, nonostante ciò rimangono ancora molti aspetti interessanti da approfondire e altri da potenziare in modo tale da renderla applicativa in tutti quei casi che richiedono l'adozione di un approccio robusto.

3. SPLINE

Questo capitolo introduce la tecnica di smoothing spline e in particolare le spline penalizzate. Negli ultimi decenni tale approccio è diventato uno dei più applicati strumenti di smoothing. Questo metodo non parametrico può essere considerato una generalizzazione dello smoothing spline le cui proprietà risiedono nella flessibilità di scelta delle basi e della tipologia di penalità da applicare. La principale attrattività del metodo risiede nel fatto di essere una diretta estensione del modello di regressione lineare (O'Sullivan, 1986; Kelly e Rice 1990; Gray 1992, 1994; Eilers e Marx, 1996) e nel legame con i modelli misti, che permette di utilizzare la metodologia propria di tali modelli per la stima dei parametri.

3.1 Smoothing

Uno *smoother*, secondo la definizione data da Hastie e Tibshirani (1990), è uno strumento che permette di riassumere l'andamento di una variabile risposta *y* in funzione di uno o più predittori *x*. La funzione stimata si presenta meno variabile rispetto a *y* stesso e per questo motivo prende il nome di *smoother*. Una proprietà importante di uno *smoother* è la sua natura non parametrica in quanto non viene fatta alcuna assunzione sulla forma funzionale tra *y* e *x*. Gli *smoothers* hanno principalmente due utilizzi. Il primo è descrittivo, in quanto possono essere utilizzati per facilitare l'interpretazione del plot di *y vs x*. Il secondo utilizzo, invece, è la stima della dipendenza della media di *y* rispetto ai suoi predittori. Esistono vari tipi di smoothers, ad esempio le medie mobili, le regressioni *kernel*, le funzioni *spline*. In questo studio verrà utilizzato quest'ultimo tipo di *smoother*.

Il termine spline nasce nell'ambito dell'ingegneria dove sorge la necessità di progettare lamine flessibili in grado di connettere punti in posizioni prefissate ma libere di assumere andamenti curvilinei nei tratti intermedi senza punti angolosi e discontinuità

Da un punto di vista strettamente matematico per spline (Azzalini e Scarpa, 2004) si intende una funzione *f(x)* vincolata a passare esattamente per *k* punti detti nodi in corrispondenza dei quali si forza la funzione ad una continuità che di norma arriva fino alla derivata seconda. Ciò permette di approssimare la funzione *f(x)*, nota su un numero di punti prefissati, in modo che la curva approssimante interpoli la funzione sui punti (ovvero passi per quei punti) e sia regolare nei tratti

intermedi. Un'approssimazione di tale funzione può essere ottenuta tramite combinazioni lineari di curve polinomiali a tratti dette appunto spline.

3.2 Le spline come strumento matematico per l'approssimazione di funzioni

Più precisamente sia $f: [a, b] \rightarrow \mathbb{R}$ e si voglia approssimare *f(x)* tramite una curva *S(x)* regolare a tratti, ovvero tale che, considerata la restrizione di *S* su un sottointervallo di una partizione di [a,b], questa abbia predefiniti requisiti in termini di derivabilità e continuità. Al fine di definire la spline si fissano *k* punti in [a,b], indicati nel seguito con $x_1, x_2, \dots, x_k$, tali che $a = x_1 < x_2 < \cdots < x_k = b$. Questi punti sono denominati nodi della spline. I nodi possono essere allocati in modo equidistante, cioè distanziati uniformemente, o posizionati sui quantili della variabile indipendente. Se i nodi sono equidistanti si parla anche di spline cardinale. I due nodi coincidenti con gli estremi *a* e *b* del dominio della funzione sono detti nodi esterni gli altri vengono denominati nodi interni. I nodi partizionano l'intervallo in *k-1* sottointervalli contigui. Si assuma di conoscere il valore, $f(x_i)$, della funzione *f(x)* su ciascun nodo k_i.

Su ciascun intervallo della partizione di [a,b], si considera un polinomio $S_i(x),\ i = 1, 2, \dots, \mathrm{k} - 1$,e si definisce la funzione *S(x)* polinomiale a tratti su [a,b] come

$$S(x) = \begin{cases} S_1(x) & a = k_1 \le x \le k_2 \\ S_2(x) & k_2 \le x \le k_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ S_i(x) & k_i \le x \le k_{i+1} \\ \cdot & \cdot \\ \cdot & \cdot \\ S_{k-1}(x) & k_{k-1} \le x \le k_k = b \end{cases}$$

La funzione *S(x)* coincide quindi con il polinomio $S_i(x)$ sul *i*-esimo sottointervallo di [a, b]. Al fine di garantire a *S(x)* una "sufficiente regolarità" sui nodi, si richiede che essa soddisfi a dei vincoli di continuità. Per rispettare, inoltre, l'informazione che si ha sulla funzione da interpolare, ovvero il fatto che assuma sui nodi valori noti, si richiede alla spline di interpolare tali punti imponendo dei vincoli di interpolazione.

La funzione polinomiale a tratti *S(x)* che soddisfa i vincoli di continuità e interpolazione costituisce una spline.

Supponendo che ciascun polinomio $S_i(x_i)$ sia di grado *p*=1, si ottiene una spline lineare. In questo caso

$$S_i(x) = a_i(x - k_i) + b_i$$

per $k_i \leq x \leq k_{i+1}$

la spline avrà queste proprietà:

1. $S(x) \in C^0[a, b]$
2. $S_i(x)$ è lineare su ogni $[k_i, k_{i+1}]$

La spline dipende da $(k-1) + (k-1) = 2(k-1)$ parametri detti gradi di libertà, rappresentati dall'intercetta e dal coefficiente angolare di ciascuna retta relativa a ciascuno dei $k - 1$ segmenti identificati dai k nodi.

La funzione *S(x)* sarà quindi completamente specificata una volta fissati i valori di questi parametri. *k* di tali parametri sono identificabili in base ai vincoli di interpolazione che impongono che i valori della funzione interpolata e del polinomio coincidano su ognuno dei k nodi prescelti. I rimanenti k − 2 parametri sono identificati in base ai vincoli di continuità posti sui $k - 2$ nodi interni.

Spesso è conveniente, anche alla luce delle applicazioni statistiche considerate nel seguito, utilizzare una rappresentazione alternativa delle spline in cui la spline *S*(*x*) approssimante la funzione *f*(*x*) è ottenuta come una combinazione lineare di una base di funzioni $(B_1(x), B_2(x), \ldots, B_k(x))$ ovvero:

$$\mathrm{S}(x) = \sum_{\mathrm{j}=1}^{k} \beta_j \mathrm{B}_j(\mathrm{x})$$

dove $\mathrm{B}_j(\mathrm{x})$ rappresenta la *j*-esima componente della base lineare relativa all'intervallo *j*-esimo identificato dai nodi disposti sul campo di variazione di *x* e $\beta_1, \ldots, \beta_k$ un insieme di coefficienti opportuni.

3.3 Smoother basati sulle spline

Con l'obiettivo di descrivere le caratteristiche degli smoother basati sulle spline consideriamo il seguente modello

$$y_i = f(x_i) + e_i$$

in cui f è la funzione di smooth nella covariata x_i, che necessita di essere stimata dall'insieme di osservazioni (y_i, x_i) (*scatterplot smoothing*); y_i è la variabile di risposta e e_i sono variabili casuali i.i.d distribuite come una normale $N{\sim}(0, \sigma_e)$.

Al fine di stimare la funzione di lisciamento viene richiesto che venga rappresentata in modo tale da essere ricondotta ad un modello lineare. Questo può essere fatto scegliendo una opportuna base, ovvero definendo lo spazio delle funzioni di cui *f* (o una sua approssimazione) è un elemento.

Per tanto è possibile esprimere la spline lineare approssimante la funzione *f(x)* come una combinazione lineare di una base di funzioni $(B_1(x), B_2(x), \dots, B_k(x))$ ovvero:

$$f(x) = \sum_{j=1}^{k} \beta_j B_j(x)$$

dove $B_j(x)$ rappresenta la *j*-esima componente della base. Le componenti $B_j(x)$ sono funzioni definite su sottointervalli contigui della partizione di [a,b] determinata dalla scelta dei nodi.

Sostituendo questa definizione di *f(x)* nella formulazione del modello lineare si ottiene:

$$y_i = \sum_{j=1}^{K} \beta_j B_j(x) + e_i$$

che permette di utilizzare i metodi di stima classici per stimare i coefficienti beta della combinazione lineare.

Ad esempio, nel caso in cui la scelta verte sulle base spline polinomiali troncate di grado p, le funzioni base sono

$$B_0(x) = 1, B_1(x) = x, B_2(x) = x^2, \dots, B_p(x) = x^p,$$

$$B_{p+1}(x) = (x - \kappa_1)_+^p, \dots, B_{p+K}(x) = (x - \kappa_K)_+^p$$

con $(x - \kappa)_+$:

$$(x_i - \kappa_k)_+ = \begin{cases} 0 & x \leq \kappa_{\mathrm{k}} \\ x - \kappa_k & x > \kappa_{\mathrm{k}} \end{cases}$$

In generale per $p = 1$, una funzione di questo tipo $(x - \kappa_{\mathrm{k}})_+^1$ è conosciuta come funzione base spline lineare ed un insieme di tali funzioni è detta base spline lineare. Qualsiasi combinazione lineare di funzioni-base spline lineari $1,\ x, (x - \kappa_1)_+, \ldots., (x - \kappa_K)_+$ è una funzione lineare a tratti con i nodi a $\kappa_1, \ldots, \kappa_{\mathrm{K}}$.

La funzione spline $f(x)$con questa base diviene:

$$f(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \beta_{pk} (x - \kappa_k)_+^p$$

Esistono diverse tipologie di base spline quali spline naturali, spline naturali cubiche, B-spline, spline radiali e combinazioni di esse che trovano un ampia applicazione in ambito statistico e che verranno considerate nel seguito di questo lavoro. Per una più completa e dettagliata descrizione si veda Ruppert et al. (2003, 2009) e Wood (2006).

Dopo aver selezionato la base, il passo successivo è quello di scegliere la posizione e il numero di nodi. I nodi dovrebbero essere uniformemente spaziati nel range di valori di *x* o posizionati ai quantili della distribuzione della covariata.

La scelta del numero dei nodi è molto importante in quanto non solo definisce il numero di funzioni base ma soprattutto influenza il livello di lisciamento del modello spline. Considerando i casi estremi: con K=0 ci si riconduce ad un modello di regressione polinomiale mentre se il numero di nodi uguaglia il numero di osservazioni *n* avremmo che il modello spline corrisponde esattamente all'interpolazione dei dati. In generale, siamo tenuti a scegliere un numero di nodi intermedio ma compreso nell'intervallo $0 \leq \mathrm{K} \leq n$, facendo attenzione al fatto che se optiamo per un numero, comunque, troppo piccolo o troppo grande potremmo avere rispettivamente un problema di un eccessivo lisciamento che porta ad ignorare il reale pattern dei dati e, dall'altro lato, di una eccessiva flessibilità che induce un overfit. Quando la dimensione del campione è elevata, l'utilizzo di $\mathrm{K} << n$ nodi produce modelli più parsimoniosi e riduce il costo computazionale.

Gli smoother che utilizzano un numero di funzione base minore di *n* sono detti *low-rank* mentre quelli che fanno uso di un numero di funzioni base approssimativamente uguale alla dimensione del campione sono detti *full-rank* (Hastie, 1996). L'utilizzo dei lisciatori *low-rank* (ad esempio P-spline)

può essere fatto risalire a Parker e Rice (1985), O'Sullivan (1986, 1988) e Kelly e Rice (1990), ma hanno raggiunto la massima popolarità dopo la pubblicazione di Eilers e Marx (1996) e Hastie (1996).

3.4 La regressione con spline penalizzate

Nel paragrafo precedente è stato messo in evidenza l'importanza della scelta del numero (e quindi della dimensione della base) e della posizione dei nodi nel determinare un auspicabile livello di fitting.

La letteratura relativa alla tecniche di spline è ricca di metodi che prevedono l'utilizzo di una penalità per controllare il lisciamento del fit e che costituiscono un'alternativa ai complessi algoritmi di ricerca del numero e della posizione ottimale di nodi (si veda ad esempio, O'Sullivan, 1986; Eubank,1988; Green e Silverman, 1994; Wahba, 1990; Eilers, e Marx, (1996) e Ruppert et al. (2003,2009). L'idea di base prevede di mantenere fisso il numero di nodi ad una dimensione un po' più grande rispetto a quello che riteniamo necessaria ma smussare la loro influenza mediante un termine di penalità da aggiungere alla funzione obiettivo da minimizzare.

Due approcci, in particolare, hanno avuto una crescente popolarità nella letteratura statistica e nelle scienze applicate: 1) il primo usa le funzioni base B-spline (de Boor (1978) e Dierckx (1993)), i nodi uniformemente spaziati e un termine di penalità discreta che controlla il livello di lisciamento (introdotto da Eilers e Marx, (1996) e notoriamente conosciuto come P-spline); 2) il secondo fa uso delle funzioni base potenza troncate, dei nodi posizionati sui quantili della variabile indipendente e di una penalità di tipo *rigde* sulle funzioni-base potenza troncate, qualunque sia il suo grado (Ruppert et al., 2003; 2009). Nel seguito vengono descritti entrambi i metodi, a partire dal quello proposto da Ruppert et al. (2003), con l'obiettivo di delinearne le principali caratteristiche (per maggiori dettagli si consulti Eilers e Marx, (1986; 2015).

Consideriamo il modello spline, generale, in forma matriciale, con un numero di nodi abbastanza elevato K (modello penalizzato di Ruppert et al.2003):

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{e}$$

con $\mathbf{y} = (y_1, y_2, \dots, y_n)$, vettore delle variabile di risposta, $\mathbf{X}$ matrice delle basi spline (potenze troncate), $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_{11}, \dots, \beta_{pk})$ vettore dei coefficienti dei nodi e $\boldsymbol{e} = (e_1, e_2, \dots, e_n)$ vettore degli errori. Come già ribadito precedentemente, non vincolare la stima del coefficiente relativo al *k*-esimo nodo β_{pk} potrebbe portare ad un adattamento ondulato. Per porre rimedio a questa situazione, in base alla selezione delle basi spline e all'assunzioni del modello, è possibile scegliere tra diverse tipologie di vincolo:

- $max\left|\beta_{pk}\right| < C$
- $\sum\left|\beta_{pk}\right| < C$
- $\sum\beta_{pk}^2 < C$

Ad esempio, se viene adottata la base spline polinomiale troncata, la penalità più semplice (Wand, 1999) consiste nel vincolare la somma al quadrato dei coefficienti dei nodi β_{pk} dato per scontato che alla base ci sia una ragionevole scelta del valore della costante C.

Quindi facendo uso del vincolo $\sum\beta_{pk}^2 < C$ e della matrice **D**

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{(p+1)\times(p+1)} & \mathbf{0}_{(p+1)\times K} \\ \mathbf{0}_{K\times(p+1)} & \boldsymbol{I}_{K\times K} \end{bmatrix}$$

il problema di minimizzazione vincolata è

$$\min\|\mathbf{y} - \mathbf{X\beta}\|^2 \text{ soggetta a } \boldsymbol{\beta}^T\mathbf{D}\,\boldsymbol{\beta} \leq C$$

Utilizzando i moltiplicatori di Langrange la soluzione del problema di minimo equivale a scegliere $\boldsymbol{\beta}$ per minimizzare:

$$\min\|\mathbf{y} - \mathbf{X\beta}\|^2 + \lambda^2\boldsymbol{\beta}^T\mathbf{D}\,\boldsymbol{\beta}$$

ovvero

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X^TX} + \lambda^2\mathbf{D})^{-1}\mathbf{X}^T$$

I valori stimati della risposta saranno ottenuti come:

$$\hat{\mathbf{y}} = \mathbf{X}\,(\mathbf{X^TX} + \lambda^2\mathbf{D})^{-1}\boldsymbol{X}^T\mathbf{y}$$

per qualsiasi valore $\lambda \geq 0$.

Il termine di penalità $\lambda^2\,\boldsymbol{\beta}^T\mathbf{D}\,\boldsymbol{\beta}$ è detto *roughness penalty* perché penalizza fit che sono troppo grossolani, generando così un risultato più smussato. La quantità di smoothing è controllato dal

parametro λ . La stima penalizzata comprime tutti i coefficienti della funzioni base spline verso zero diversamente dalla metodologia di selezione dei nodi, che contrae alcuni coefficienti a zero e lascia i restanti inalterati. Il parametro di smoothing λ gestisce il trade-off tra il fit e il lisciamento del modello: per $\lambda \to \infty$ si ha il fit di un polinomio di grado *p*-esimo, mentre per $\lambda = 0$ il termine di penalità non ha più alcuna importanza e pertanto corrisponde alla stima non vincolata.

Eilers e Marx (1996) generalizzano l'approccio di O'Sullivan (1986,1988) il quale ha suggerito di prendere una base costituita da molte B-spline e di utilizzare una penalità discreta ricavata dall'integrale della derivate seconde della curva al quadrato. Tale integrale misura l'irregolarità (R) di una funzione $f(x)$:

$$R = \int_l^u [f''(x)]^2 \, dx$$

dove *l* e *u* sono gli estremi del dominio di *x*. Allora, se $f(x) = \sum_j \beta_j B_j(x)$, si può determinare una matrice **P** tale che $\mathbf{R} = \boldsymbol{\beta}^T \mathbf{P} \boldsymbol{\beta}$ **,** i cui elementi sono dati dall'integrale del prodotto delle derivate seconde delle B-spline adiacenti. Introducendo la penalità sulle derivate seconde l'autore formula la funzione obiettivo come:

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{k} \beta_j B_j(x_i) \right\}^2 + \lambda \int_{x_{min}}^{x_{max}} \left\{ \sum_{j=1}^{k} \beta_j B_j''(x_i) \right\}^2 dx$$

Nel contesto del lisciamento basato sulle spline, la derivata prima porta ad una equazione semplice e al fit di una funzione lineare a tratti mentre le derivate di ordine superiore conducono ad un complesso sistema di equazioni e ad un lisciamento molto marcato.

La determinazione di **P** non è banale e diventa alquanto gravosa quando si considerano le derivate di terzo e quarto ordine e più. Recentemente, anche Wand e Ormerod (2008) hanno esteso l'idea di O'Sullivan per considerare le derivate di ordine più elevato e inoltre, avvalendosi di un sistema algebrico computerizzato hanno costruito un formulario.

Le P-spline, così come sviluppate da Eilers e Marx, 1996, eludono il problema tralasciando completamente le derivate e gli integrali.

Seguendo questo approccio, il termine di penalità **P** è discreto ed è, come già menzionato sopra, specificato come operatore differenza sui coefficienti della funzione B-spline. Quindi la penalità **P** è una matrice di dimensioni $c \times c$ definita come $\mathbf{P} = \lambda(\Delta^d)^T \Delta^d$ dove Δ^d è l'operatore differenza di ordine *d*. Per il vettore dei coefficienti β l'operatore differenza è definito ricorsivamente da:

$$\Delta^1\beta_i = \beta_i - \beta_{i-1}$$
$$\Delta^2\beta_i = \Delta^1(\Delta^1\beta_i) = \beta_i - 2\beta_{i-1} + \beta_{i-2}$$
$$\vdots$$
$$\Delta^d\beta_i = \Delta^1(\Delta^{d-1}\beta_i)$$

L'ordine della penalità *d* controlla i cambiamenti che avvengono tra nodi adiacenti. Una differenza di primo ordine ($d = 1$), penalizza i salti tra coefficienti successivi mentre una differenza di secondo ordine penalizza le deviazioni da un trend lineare (cioè da $2\beta_{i-1} - \beta_{i-2}$). Quindi il fit della P-spline ha un interessante proprietà: il grado delle B-spline e l'ordine della penalità possono essere scelte in modo totalmente indipendente.

Introducendo la penalità sulle differenze finite dei coefficienti adiacenti delle B-spline la funzione obiettivo diventa:

$$\sum_{i=1}^{n}\left\{y_i - \sum_{j=1}^{k}\beta_j B_j(x_i)\right\}^2 + \lambda \sum_{j=d+1}^{k}(\Delta^d\beta_j)^2$$

Data la matrice $\boldsymbol{D}_d$ tale che $\boldsymbol{D}_d\boldsymbol{\beta} = \Delta^d\boldsymbol{\beta}$, se si sostituisce la penalità con $\lambda\|\boldsymbol{D}_d\boldsymbol{\beta}\|^2 = \lambda\boldsymbol{\beta}^T\boldsymbol{D}_d^T\boldsymbol{D}\boldsymbol{\beta} = \lambda\,\boldsymbol{\beta}^T\boldsymbol{P}\,\boldsymbol{\beta}$ si ottiene esattamente il termine di penalità espresso da O'Sullivan. E'stato dimostrato che entrambi gli approcci sono simili quando, nel termine di penalità, si tratta la divergenza del secondo ordine, tuttavia nel metodo P-spline è più semplice includere penalità di qualsiasi ordine nell'equazione di regressione.

3.5 Selezione del parametro di smoothing λ e gradi di libertà

Selezionare un valore ottimale del parametro di smoothing λ è di notevole importanza poiché ha una forte influenza sul fit. A questo scopo si può utilizzare uno dei criteri di selezione del modello come AIC e *Mallows Cp* oppure ricorrere ad uno dei metodi tra i più comuni in letteratura quali il *cross-validation* (CV) o il *generalized cross validation* (GCV) (Wahba,1990). Tutte queste tecniche devono rispondere all'obiettivo di selezionare un valore di λ tale che la funzione stimata $\hat{f}$ sia il più vicina possibile ad f , ma differiscono nel modo di misurare la prossimità tra le due.
Il criterio di cross-valitation seleziona il valore di λ che minimizza la funzione obiettivo:

$$CV(\lambda) = \sum_{i=1}^{n} \left[y_i - \hat{f}_{-i}(x_i; \lambda) \right]^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(1 - s_{ii})^2}$$

dove s_{ii} sono gli elememti della diagonale della matrice hat $\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X^T X} + \lambda \mathbf{D})^{-1} \mathbf{X^T}$.
Per il criterio *generalized cross validation* seleziona il valore di λ che minimizza

$$GCV(\lambda) = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(n - \sum_{i=1}^{n} s_{ii})^2}$$

Per entrambe si seleziona il valore di λ che corrisponde ai minori valori di CV e GCV. I due criteri sono abbastanza simili ma il GCV è dotato della proprietà di invarianza (Wood, 2006; Wahba, 1990).
Il parametro λ non ha una diretta interpretazione della quantità di 'struttura' che è stata imposta sul fit, per fare ciò possiamo ricorrere alla definizione di gradi di libertà di un modello.
Generalizzando il concetto di gradi di libertà del modello lineare, si possono definire i gradi di libertà del fit corrispondente al parametro di smoothing λ come:

$$df_{fit} = tr\ (\mathbf{S}_\lambda)$$

Questa quantità può essere interpretata come il numero di parametri equivalente di cui necessitiamo per ottenere lo stesso fit da un modello parametrico.
Consideriamo la matrice hat S_λ allora

$$df_{fit} = tr\ (\mathbf{X}(\mathbf{X^T X} + \lambda \boldsymbol{D})^{-1} \mathbf{X^T}) = tr((\mathbf{X^T X} + \lambda \mathbf{D})^{-1} \mathbf{X^T X})$$

Per un lisciatore spline penalizzato con *k* nodi e grado *p* è facile dimostrare che

$$tr(\mathbf{S}_0) = p + 1 + k$$
$$tr(\mathbf{S}_\lambda) = p + 1 \text{ come } \lambda \to \infty$$

Così, valori positivi di λ corrispondono a

$$p + 1 < df_{fit} < p + 1 + k$$

3.6 Spline come modelli misti

La specificazione del termine spline tramite effetti casuali è giustificato essenzialmente da due motivi: fornisce un quadro analitico per comprendere perché lo smoothing spline è uno smoother ottimo e la rappresentazione della non linearità attraverso l'eterogeneità non osservata aiuta a chiarire il bisogno di utilizzare tecniche non parametriche.

La relazione tra i modelli di regressione non parametrici e i modelli misti è stata stabilita per la prima volta da Green (1987) and Speed (1991), ma solo quando sono stati sviluppati i software in grado di manipolare i modelli misti, questo aspetto è diventato un argomento di ricerca molto indagato (Wang, 1998; Zhang et al., 1998; Verbyla et al., 1999). I primi modelli proposti in tal senso si basavano sull'uso della tecnica di smoothing spline ma ben presto sono stati estesi al contesto delle spline penalizzate che faceva uso di funzioni-base potenza troncate per la costruzione della matrice di regressione (Brumback et al., 1999; Coull et al., 2001; Wand, 2003). Anche Eilers (1999) propose l'interpretazione delle P-spline come modello misto e successivamente Currie e Durban (2002) utilizzarono il suo approccio per estenderlo ai casi in cui esiste eteroschedasticità o autocorrelazione degli errori.

Nel paragrafo 3.4 è stato ricordato come stimare la funzione *f* (*x*) che esprime l'effetto non lineare di *x* sulla risposta mediante le tecniche di spline penalizzata, ora vedremo come la stima può essere trattata all'interno di un modello misto. A tal fine, consideriamo ancora una volta il modello di regressione spline penalizzato con base spline polinomiale troncata. Per semplificare l'illustrazione occorre apportare alcune modifiche alla notazione e riscrivere il modello

$$y_i = \sum_{j=1}^{K} \beta_j B_j(x) + e_i$$

nel seguente modo

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \sum_{k=1}^{K} u_{pk}(x_i - \kappa_k)_+^p + e_i$$

con $\boldsymbol{\beta}$ i coefficienti del polinomio, e $\boldsymbol{u}$ il vettore dei coefficienti delle funzioni lineari troncate. In corrispondenza dei due vettori si definiscono le matrici **X** dei polinomi e **Z** ottenuta dalle funzioni base del tipo polinomi troncati. La funzione di stima ai minimi quadrati penalizzati (vedi equazione di stima nel paragrafo 3.4) divisa per la varianza di errore $\boldsymbol{\sigma_e^2}$ diventa

$$\frac{1}{\boldsymbol{\sigma_e^2}}||\mathbf{y} - \mathbf{X\beta} - \mathbf{Zu}||^2 + \frac{\boldsymbol{\lambda^2}}{\boldsymbol{\sigma_e^2}}||\boldsymbol{u}||^2.$$

Da questa espressione si evince che la funzione obiettivo usata nella stima è analoga al criterio BLUP (best linear unbiased prediction) (vedi ad esempio McCulloch e Searle, 2001) dei modelli misti.
Con le quantità sopra definite, si perviene alla rappresentazione del modello di regressione spline penalizzata come un modello misto che in forma matriciale è

$$\boldsymbol{y} = \mathbf{X\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e}$$

con ciascun elemento di $\boldsymbol{u}$ distribuito secondo una distribuzione normale $N(0, \sigma_u^2)$. Allo stesso modo il termine di errore è gaussiano $e \sim N(0, \sigma_e^2 \mathbf{I})$, gli errori sono incorrelati tra loro e indipendenti dagli effetti random $\boldsymbol{u}$ da cui:

$$Cov\begin{bmatrix}\boldsymbol{u} \\ e\end{bmatrix} = \begin{bmatrix}\sigma_u^2 \boldsymbol{I} & 0 \\ 0 & \sigma_e^2 \boldsymbol{I}\end{bmatrix}$$

Il valore fittato di *f* sarà

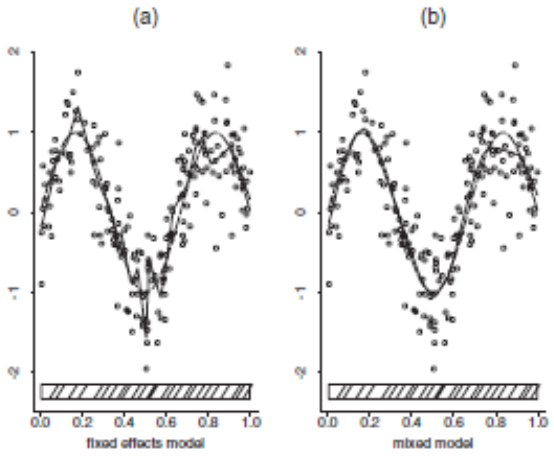
$$\tilde{\boldsymbol{f}} = \boldsymbol{C}\left(\boldsymbol{C^T C} + \frac{\sigma_e^2}{\sigma_\gamma^2}\boldsymbol{D}\right)^{-1}\boldsymbol{C^T y}$$

con **D** = *diag*(0,0,1,1,….,1) e $\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}]$ e $\lambda^2 = \frac{\sigma_e^2}{\sigma_u^2}$. Quindi, nel framework dei modelli misti, il rapporto di varianze $\frac{\sigma_e^2}{\sigma_u^2}$ gioca il ruolo del parametro di *smoothing* λ.

La differenza tra i modelli è dovuta al fatto che adesso l'effetto casuale è associato a ciascun locazione del nodo.

I due grafici riportati in figura1(tratta da Ruppert et al., 2003 pag.110) sono stati ottenuti implementando il modello ad effetti fissi, che tratta i coefficienti dei nodi come un effetto fisso, e il modello ad effetti misti in cui i coefficienti dei nodi sono trattati come effetti casuali, usando dati simulati da una funzione $f(x) = \sin(3\pi x)$ con $0 < x < 1$. Si può vedere come nel caso del modello misto, figura 1-(a), si ottenga un lisciamento e un adattamento alla funzione $f(x)$ migliore, mentre con il modello ad effetti fissi i dati sono overfittati.

Figura 1. Confronto tra i due modelli: modello ad effetti fissi ed modello misto. La linea tratteggiata è la funzione vera mentre la linea solida è la funzione stimata (tratta da Ruppert et al., 2003 pag.110).



Analogamente a quanto visto prima, per riformulare la P-spline come un modello ad effetti misti si divide il primo termine della funzione da minimizzare per $\boldsymbol{\sigma_e^2}$ e il termine di penalità per $\boldsymbol{\sigma_a^2}$

$$\frac{1}{\boldsymbol{\sigma_e^2}}||\mathbf{y} - \mathbf{Ba}||^2 + \frac{\boldsymbol{a^T P a}}{\boldsymbol{\sigma_a^2}}$$

dove $\boldsymbol{\sigma_e^2}$ e $\boldsymbol{\sigma_a^2}$ sono rispettivamente la varianza degli errori e la varianza dell'effetto random.

Il modello ad effetti casuali a cui corrisponde questa funzione ha la forma

$$\boldsymbol{y} = \mathbf{B}\boldsymbol{a} + \boldsymbol{e}$$

Con il coefficiente random che si distribuisce come una normale $a \sim N(0, \sigma_u^2 \mathbf{P}^{-1})$ e il termine di errore anch'esso normale $e \sim N(0, \sigma_e^2 \mathbf{I})$.

Le penalità di ordine *d* non penalizzano le potenze di *x* fino al grado $d-1$. Per questo motivo la matrice **P** è singolare e di conseguenza $\boldsymbol{u}$ avrà una distribuzione degenere. Per risolvere il problema si ri-parametrizza (Eirles, 1999) il modello scomponendo $\mathbf{B}\boldsymbol{a}$ in due parti additive

$$\mathbf{B}\boldsymbol{a} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u}$$

dove $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{D}^{\boldsymbol{T}}(\boldsymbol{D}\boldsymbol{D}^{\boldsymbol{T}})^{-\mathbf{1}}$ con **D** matrice delle differenze . Operando in questo modo le *d* colonne di **X** generano lo spazio nullo dei polinomi di **P** e le $n-d$ colonne di **Z** generano il suo complemento. Con questa rappresentazione **u** ha una distribuzione non degenere e il modello misto corrispondente è

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e}$$

con $u \sim N(0, \sigma_u^2\, \boldsymbol{I})$ ed $e \sim N(0, \sigma_e^2 \mathbf{I})$

I vantaggi derivanti dall'uso del modello ad effetti misti sono legati alla relativa facilità di implementazione grazie alla possibilità di utilizzare software esistenti. Inoltre la stima del parametro di lisciamento viene sostituita dalla stima dei due parametri di varianza e quindi, non richiede metodi quali il cross-validation. La stima dei parametri si può ottenere applicando il metodo di massima verosimiglianza ML o REML (Restricted Maximum Likelihood), (Patterson and Thompson (1971).

3.7 Smoothing spline bivariato

Come visto nella sezione precedente le tecniche di smoothing spline penalizzate necessitano di un insieme di funzioni base che permette di cogliere la struttura non lineare presente nei dati. Similmente lo smoothing bivariato necessita di funzioni base bivariate. L'estensione da una base unidimensionale a basi bivariate può essere attuata almeno in due modi diversi: uno basato sul prodotto e l'altro sulla rotazione (Ruppert et al., 2003).

Supponiamo di avere due covariate continue *s* e *t* predittive della variabile di risposta y. Il modello di regressione non lineare bivariato generale è:

$$y_i = f(s_i, t_i) + e_i$$

dove *f* è una funzioni bivariata a variabili reali.

Nell'ambito di un modello di regressione è possibile modellare un termine di interazione (il prodotto delle due variabili *s* e *t*) per tener conto dell'effetto congiunto di queste due covariate sulla risposta *y*, aggiungendo il termine di regressione $\gamma_s\, s_i\, t_i$ nel modello additivo lineare:

$$y_i = \beta_0 + \beta_s s_i + \beta_t t_i + \gamma_s\, s_i\, t_i + e_i$$

La diretta estensione alla base lineare troncata porta al modello:

$$y_i = \beta_0 + \beta_s s_i + \sum_{k=1}^{K^s} u_k^s (s_i - k_k^s)_+ + \beta_t t_i + \sum_{k=1}^{K^t} u_k^t (t_i - k_k^t)_+ + \gamma\, s_i\, t_i + \sum_{k=1}^{K^s} v_k^s s_i (t_i - k_k^t)_+ + \sum_{k=1}^{K^t} v_k^t t_i (s_i - k_k^s)_+ + \sum_{k=1}^{K^s} \sum_{k'=1}^{K^t} v_{kk'}^{st} (s_i - k_k^s)(t_i - k_k^t)_+ + e_i$$

dove K^s e K^t sono il numero di nodi riferibili rispettivamente alla covariate *s* e *t*, $u_k^s \sim N(0, \sigma_s^2)$ e $u_k^t \sim N(0, \sigma_t^2)$ inoltre il modello è specificato mediante le seguenti funzioni base

$$1, s, (s - k_1^s)_+, \ldots, (s - k_{K^s}^s)_+$$
$$1, t, (t - k_1^t)_+, \ldots, (t - k_{K^t}^t)_+$$

formando tutte le coppie di prodotti tra un elemento del primo insieme di funzioni e uno del secondo. Una caratteristica di queste spline è la dipendenza delle coordinate dall'orientazione dagli assi. Lo smoothing potrebbe cambiare se, ad esempio, le locazioni spaziali sono misurate su assi con diversa orientazione. Se il fenomeno non dipende dalla orientazione spaziale, l'invarianza per rotazione può essere raggiunta mediante l'utilizzo delle funzioni spline radiali che sono funzioni base della forma

$$C(||(s,t) - (k^s, k^t)||)$$

per qualsiasi funzione univariata *C*. Siccome il valore della funzione a (s, t) dipende solo dalla distanza dal nodo (k^s, k^t) la funzione è radialmente simmetrica attorno a questo punto.
Per agevolare e snellire l'illustrazione, viene presentata la classe degli smoother radiali unidimensionali.

Facendo uso delle funzioni base radiali lo smoothing di un scatterplot può essere ottenuto come segue. Innanzitutto definiamo le matrici **X** e **Z** come:

$$\mathbf{X} = [1, x_i]_{1 \le i \le n}$$

$$\mathbf{Z} = \left[\left(x_i - x_j \right)_+ \right]_{1 \le i,j \le n}$$

come abbiamo già visto nel paragrafo 3.4, i valori fittati dal modello sono

$$\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\gamma}}$$

dove $\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \underset{\beta,\gamma}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X\beta} - \mathbf{Z\gamma}\|^2 + \lambda^2 \begin{bmatrix} \beta \\ \boldsymbol{\gamma} \end{bmatrix}^T \mathbf{D} \begin{bmatrix} \beta \\ \boldsymbol{\gamma} \end{bmatrix}$ e $\mathbf{D} = diag\ (0,0,1, \ldots\ldots, 1)$.

Le stime dei parametri sono EBLUP (Empirical Best Linear Unbiased Predictor ovvero miglior predittore empirico lineare non distorto) nel ambito dei modelli misti:

$$\boldsymbol{y} = \mathbf{X\beta} + \mathbf{Z}\boldsymbol{\gamma} + e$$

$$Cov \begin{bmatrix} \boldsymbol{\gamma} \\ e \end{bmatrix} = \begin{bmatrix} \sigma_\gamma^2 \boldsymbol{I} & 0 \\ 0 & \sigma_e^2 \boldsymbol{I} \end{bmatrix}$$

$$\lambda^2 = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_\gamma^2}$$

A questo punto si effettua una trasformazione delle basi lineari troncate mediante una combinazione lineare delle colonne di **X** e **Z** in modo tale che **X** rimanga immutata e **Z** divenga la matrice radialmente simmetrica

$$\mathbf{Z}_R = \left[\left| x_i - x_j \right| \right]_{1 \le i,j \le n}$$

Questa trasformazione è espressa in termini di una matrice $\mathbf{L}_{(n+2) \times (n+2)}$ per la quale

$$[\mathbf{X}\ \mathbf{Z_R}] = [\mathbf{X}\ \mathbf{Z}]\mathbf{L}$$

Il vettore dei valori fittati per la nuova base sono

$$\widehat{\boldsymbol{y}}_R = \mathbf{X}\hat{\beta}_R + \mathbf{Z_R}\widehat{\boldsymbol{\gamma}}_R$$

dove $\hat{\beta}_R$ e $\widehat{\boldsymbol{\gamma}}_R$ sono dati da

$$\begin{bmatrix}\hat{\beta}_R \\ \hat{\boldsymbol{\gamma}}_R\end{bmatrix} = \underset{\beta,\gamma}{\operatorname{argmin}} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_R\boldsymbol{\gamma}\|^2 + \lambda^2 \begin{bmatrix}\beta \\ \boldsymbol{\gamma}\end{bmatrix}^T \mathbf{L^T D\ L} \begin{bmatrix}\beta \\ \boldsymbol{\gamma}\end{bmatrix}$$

Se la matrice $\mathbf{Z}_R$ è radialmente simmetrica questo non vale per il fattore di penalità $\lambda^2 \begin{bmatrix}\beta \\ \boldsymbol{\gamma}\end{bmatrix}^T \mathbf{L^T D\ L} \begin{bmatrix}\beta \\ \boldsymbol{\gamma}\end{bmatrix}$.

Inoltre il termine di penalità così come è espresso non può essere usato nel contesto multivariato pertanto si può effettuare scelta alternativa dello stesso che consiste nel termine $\lambda\boldsymbol{\gamma}^T\ \mathbf{Z}_R\boldsymbol{\gamma}$ che porta alla famiglia di *smoother* del tipo *thin plate spline* (Green e Silverman, 1994). In questo modo la stima dei parametri è

$$\begin{bmatrix}\hat{\beta} \\ \hat{\boldsymbol{\gamma}}\end{bmatrix} = \underset{\beta,\gamma}{\operatorname{argmin}}\ ||\boldsymbol{y}\ -\ \mathbf{X}\beta\ -\ \mathbf{Z}_R\boldsymbol{\gamma}||^2\ +\ \lambda\boldsymbol{\gamma}^T\ Z_R\boldsymbol{\gamma}$$

$\mathbf{Z}_R$ non è propriamente una matrice di covarianza poiché non gode sempre della proprietà di essere una matrice definita positiva per cui il modello considerato fin ora non è un modello misto ad effetti random, tuttavia si possono effettuare delle operazioni per renderlo tale.

Un primo metodo per rettificare la mancanza di questa proprietà è sostituire $\mathbf{Z}_R$ con la matrice

$$\mathbf{Z}_P = \left[-\left|x_i - x_j\right|_{1\le i,j\le n} + |x_i| + \left|x_j\right|\right]$$

che è semi-definita o definita positiva quindi il modello diventa

$$y = \mathbf{X}\beta_R + \mathbf{Z}_P\boldsymbol{\gamma}_R + e$$

$$Cov\begin{bmatrix}\boldsymbol{\gamma} \\ e\end{bmatrix} = \begin{bmatrix}\sigma_\gamma^2\mathbf{Z}_P^{-1} & 0 \\ 0 & \sigma_e^2\boldsymbol{I}\end{bmatrix}$$

Il modello è un modello ad effetti misti effettivamente valido e i parametri del modello non sono influenzati dalla sostituzione di $\mathbf{Z}_R$ con $\mathbf{Z}_P$ (French, Kammann e Wand 2001).

La funzione $C(r) = -|r|$ utilizzata nella definizione di $\mathbf{Z}_P$ è conosciuta come funzione di covarianza generalizzata (Kitanidis 1997).

Un ulteriore metodo di correzione prevede di utilizzare la funzione di covarianza propria $e^{-|x_i - x_j|/\rho}$ con $\rho > 0$ per ottenere

$$\mathbf{Z}_E \equiv \begin{bmatrix} e^{-|x_i - x_j|/\rho} \\ 1 \le i,j \le n\end{bmatrix}$$

3.8 Low-Rank Radial Smoothers

Ruppert et al. (2003) raccomandano l'utilizzo delle funzioni base radiali per pervenire alla *low-rank thin plate splines* in virtù dei vantaggi computazionali che ne derivano. In questo paragrafo si descrive brevemente l'estensione low-rank per le spline radiali.

Dato l'insieme dei nodi $\kappa_1, \dots, \kappa_{\mathrm{K}}$, un'approssimazione basata su un insieme ridotto di funzioni base

$$C(|x_i - \kappa_k|) \text{ con } 1 \leq k \leq \mathrm{K}$$

proviene dall'implementazione del modello misto

$$\boldsymbol{y} = \mathbf{X}\beta + \mathbf{Z}_{\mathrm{K}}\boldsymbol{\gamma} + e\ , \quad Cov(\boldsymbol{\gamma}) = \sigma_\gamma^2(\Omega_{\mathrm{K}}^{\mathbf{-1/2}})(\Omega_{\mathrm{K}}^{\mathbf{-1/2}})^{\boldsymbol{T}}$$

in cui X, Z_{K} , Ω_{K} sono così definite:

$$X = [1, x_i]_{1 \leq i \leq n}$$

$$Z_K \equiv [C(|x_i - \kappa_k|)]_{1 \leq i \leq n, 1 \leq k \leq K} \text{ e } \Omega_{\mathrm{K}} \equiv [C(|\kappa_k - \kappa_{k'}|)]_{1 \leq k, k' \leq K}$$

e γ è un vettore casuale $K \times 1$. Usando la trasformazione $\mathbf{Z} = \mathbf{Z}_K(\Omega_K^{-\frac{1}{2}})$ il modello finale può essere riscritto come

$$y = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathrm{e}$$

$$Cov\begin{bmatrix}\boldsymbol{\gamma} \\ e\end{bmatrix} = \begin{bmatrix}\sigma_\gamma^2\boldsymbol{I} & 0 \\ 0 & \sigma_e^2\boldsymbol{I}\end{bmatrix}$$

Fin qui sono state discusse le funzioni base radiali ad una dimensione ora andremo a considerarle in più dimensioni.

Come visto, con le funzioni base radiali, la dipendenza nei dati è dovuta esclusivamente alla distanza tra coppie di punti unidimensionali del tipo:

$$|x_i - \kappa_k| \text{ con } 1 \leq i \leq n\ , 1 \leq k \leq K.$$

Quindi l'estensione al caso multidimensionale $\mathbf{x}_i \in R^d$ si ottiene in modo immediato ricorrendo alla distanza tra punti d-dimensionali:

$$||\mathbf{x}_i - \kappa_k|| \quad 1 \leq i \leq n\ , 1 \leq k \leq K$$

Per $\mathbf{x}_i \in R^d$ $1 \leq i \leq n$ e $\kappa_k \in R^d$ $1 \leq k \leq K$ la *thin plate spline* può essere ottenuta utilizzando la matrice del disegno

$$\mathbf{X} = [1, \mathbf{x}_i{}^T]_{1 \leq i \leq n}$$

e

$$\mathbf{Z} = \left[C(||\mathbf{x}_i - \kappa_K||)_{1 \leq k \leq K}\right]_{1 \leq i \leq n} \left[C(||\kappa_k - \kappa_{k'}||)_{1 \leq k, k' \leq K}\right]^{-1/2}$$

$$\text{dove } C(\boldsymbol{r}) = \begin{cases} ||\boldsymbol{r}||^{2m-d} & \text{per } d \text{ dispari} \\ ||\boldsymbol{r}||^{2m-d} \log||\boldsymbol{r}|| & \text{per } d \text{ pari} \end{cases}$$

dove *m* è un intero che soddisfa la proprietà $2m - d > 0$ e controlla il lisciamento. La funzione C(.) è applicata in modo tale che in caso di rango pieno (quando il numero di nodi è uguale al numero di osservazioni) il modello per il classico smoothing bivariato conduca a una splines di tipo *thin plate*. Nella definizione di $\mathbf{Z}$ il secondo termine a destra è una trasformazione usata per semplificare le procedure di stima.

In alternativa si potrebbero utilizzare le funzioni radiali corrispondenti alla funzione di covarianza propria descritta in precedenza.

4. MODELLO SEMIPARAMETRICO M-QUANTILE AD EFFETTI CASUALI

In questa sezione, dopo una preliminare introduzione all'approccio non parametrico proposto in letteratura, viene presentata e descritta una estensione semi-parametrica del modello di regressione M-quantile ad effetti casuali per dati georeferiti che è il fulcro di questo lavoro di ricerca. Tale modello combina le caratteristiche di robustezza e flessibilità propri dei modelli M-quantile e dei modelli di regressione spline.

4.1 Modello di regressione quantile non parametrico

Molti fenomeni, soprattutto quelli ambientali, sono molto difficili da modellare poiché non è semplice trovare a priori un forma funzionale adatta alla loro descrizione. Un modo di risolvere questo problema è di utilizzare un approccio non parametrico. Invece di adottare una funzione specifica, questo approccio che fa uso di procedure automatiche, con il vantaggio di ridurre la specificazione del predittore del modello alla scelta di pochi parametri che regolano l'andamento e il lisciamento della curva. Purtroppo uno dei svantaggi dei modelli non parametrici è quello che l'interpretazione dei risultati diventa meno agevole.

Come delineato nel capitolo 2, il modello di regressione quantilica permette di stimare i quantili della distribuzione condizionata della variabile di risposta y dati i predittori $\mathbf{x}$. Come nel caso della regressione standard, una funzione lineare potrebbe non esser adatta per rappresentare questa relazione.

I metodi non parametrici nell'ambito del modello di regressione quantilica sono stati studiati estensivamente. Molti autori hanno esplorato l'argomento in relazione al metodo kernel (Chaudhuri, 1991; Fan et al., 1994; Yu e Jones, 1998; Takeuchi et al., 2006; Kai et al., 2011; mentre Hendricks e Koenker (1992) e Koenker et al. (1994) hanno utilizzato, rispettivamente, il metodo di regressione spline a basso rango e il metodo di smoothing splines. Ng and Maechler (2007) hanno implementato un modello B-spline vincolato (COBS). Un ulteriore approccio, presente in letteratura, è il un modello quantilico basato sulla P-spline (Bollaerts et al., 2006). Yoshida (2012) si è occupato del modello di regressione quantilica con Spline Penalizzata e del confronto tra questo e gli altri modelli (spline non penalizzata e smoothing spline) e delle sue proprietà asintotiche. Andriyana et al. (2014) implementano un modello quantilico a coefficienti variabili basato sulle P-spline.

Con l'approccio non parametrico il modello di regressione quantilico può essere riformulato come segue:

$$Q_q(Y|x_1) = s(x_1)$$

in cui $s(\,)$ è una funzione non nota e possibilmente non lineare. Qualsiasi metodo proposto in letteratura (menzionata sopra) può essere utilizzato per ottenere la stima della funzione *s( ).*

Il punto cruciale di un modello di regressione quantilica non parametrica è la selezione del parametro di smoothing λ. Anche in questo contesto, sono stati presentati molti approcci: Koenker et al. (1994) hanno usato una versione modificata dello Schwarz Information Criterion; Oh et al. (2004) hanno proposto una versione robusta dell'approccio cross-validation e Nychka et al. (1995) hanno fornito un approssimazione dell'Cross Validation noto come ACV (Approximate Cross Validation) per ridurre il costo computazionale. In seguito Yuan (2006) ha introdotto un approccio generalizzato all'Cross-Validation approssimato (Generalized Approximate Cross Validation (GACV)). Andriyana et al. (2014), selezionano il parametro di smoothing mediante la L-curva (Frasso and Eilers, 2015) ovvero una funzione che rappresenta il trade-off tra fedeltà e penalizzazione. Infine, Reiss and Huang (2012) sfruttando la relazione tra penalità e modelli misti, hanno presentato un metodo di selezione basato sulla verosimiglianza. Tutti questi metodi di selezione si basano sull'utilizzo di una griglia di ricerca (*grid search*), su cui calcolare il modello per poi scegliere λ, pertanto hanno un tempo computazionale molto elevato. A questo si deve aggiungere l'influenza delle covariate, infatti maggiore è il numero di covariate maggiori saranno le dimensioni della griglia e di conseguenza più elevati i tempi di calcolo.

4.2 Dal modello M-quantile non parametrico al modello semi-parametrico M-quantile

Come detto nel paragrafo precedente, quando la forma funzionale tra la variabile di risposta e le covariate non è nota oppure è particolarmente complessa il modello di regressione non parametrico basato sulle spline penalizzate può offrire una maggiore utilità. Esprimendo i coefficienti delle spline del modello (vedi capitolo 3) come un effetto casuale questo può essere riformulato e implementato come un modello ad effetti misti (Ruppert et al., 2003).

Sfruttando questa proprietà Opsomer et al. (2008), nell'ambito della stima per piccole aree, hanno proposto un modello semiparametrico ottenuto combinando il modello ad effetti casuali e il modello di regressione P-spline. Seguendo questa linea Urgate et al. (2009), con lo scopo di prevedere i valori futuri della variabile di risposta in ciascuna delle piccole aree, hanno presentato un modello longitudinale semiparametrico. Questo modello deriva dalla combinazione di un termine non-parametrico che gestisce il trend temporale e di un effetto casuale specifico per ciascuna area.

A questo si aggiunge il recente lavoro di Torabi e Shokoohi (2015) mediante il quale hanno proposto un modello GLM basato sulla P-spline nel caso la variabile di risposta sia discreta.

Ad oggi alcuni tentativi sono stati fatti per estendere il modello di regressione M-quantile in ambito non parametrico (Pratesi et al., 2006, 2008, 2009; Salvati et al., 2011; Dreassi et al., 2015).

Pratesi et al. (2008) hanno esteso questo approccio al metodo M-quantilico per la stima dei parametri di piccole aree utilizzando una specificazione non parametrica degli M-quantili della variabile di risposta y condizionati ad un insieme di covariate. L'utilizzo della spline bivariata e del modello di regressione M-quantilico permettono di controllare la variabilità spaziale presente nei dati per poi pervenire alla stima delle piccole aree.

Come già delineato, i modelli semi-parametrici sono una estensione dei modelli di regressione nei quali una o più variabili predittive entrano nel modello senza specificare a priori la forma funzionale che definisce la relazione tra variabile di risposta e le covariate (Ruppert et al., 2003). Il modello semi-parametrico M-quantile si ottiene da una diretta estensione del modello M-quantile standard.

Nel caso univariato, data una funzione di influenza ψ, un modello non parametrico per il q-esimo quantile può essere formulato come segue

$$M_q(x_1, \psi) = \tilde{f}_{\psi q}(x_1)$$

Dove la funzione $\tilde{f}_{\psi q}(x)$ non è nota e nel contesto di smoothing si assume continua e differenziabile. Si suppone quindi che questa funzione sia approssimata sufficientemente bene dalla funzione:

$$f_{\psi q}[x_1; \boldsymbol{\beta}_\psi(q), \boldsymbol{\gamma}_\psi(q)] = \beta_{0\psi}(q) + \beta_{1\psi}(q)x_1 + \ldots + \beta_{p\psi}(q)x_1^p + \sum_{k=1}^{K} \boldsymbol{\gamma}_{k\psi}(q)\,(x - \kappa_j)_+^p$$

dove p è il grado della spline, K è l'insieme dei nodi *(k=1,…K)*, $\boldsymbol{\beta}_\psi(q) = (\beta_{0\psi}(q), \beta_{1\psi}(q), \ldots., \beta_{p\psi}(q))^T$ è il vettore dei coefficienti che entrano nella parte parametrica e $\boldsymbol{\gamma}_\psi(q) = (\gamma_{1\psi}(q), \ldots., \gamma_{k\psi}(q))^T$ è il vettore dei coefficienti della spline che ci permette di cogliere la non linearità presente nella relazione. Come ricordato in precedenza, molte tipologie di basi possono essere utilizzate, in particolare per lo smooting bivariato è possibile usare una base radiale (Ruppert et al., 2003).

Poiché il numero di nodi influenza il risultato dello smoothing, un numero eccessivo per esempio potrebbe produrre un andamento irregolare, quindi ne viene limitato l'impatto imponendo un vincolo sulla dimensione dei coefficienti delle spline richiedendo che $\sum_{k=1}^{K} \boldsymbol{\gamma}_{k\psi}{}^2(q)$ sia limitato da una costante mentre i parametri $\boldsymbol{\beta}_\psi(q)$ rimangono non vincolati.

La stima dei parametri è effettuata risolvendo le *(1+p+K)* equazioni di stima:

$$\sum_{i=1}^{n} \psi_q \left(y_i - \boldsymbol{x_i}\boldsymbol{\beta}_\psi(q) - \boldsymbol{z_i}\boldsymbol{\gamma_\psi}(q)\right)(\boldsymbol{x_i}, \boldsymbol{z}_i)^T + \lambda \begin{bmatrix} \mathbf{0}_{(1+p)} \\ \boldsymbol{\gamma}_\psi(q) \end{bmatrix} = \mathbf{0}_{(1+p+K)}$$

dove $\boldsymbol{x}_i$ è *i*-sima riga dalla matrice $\boldsymbol{X}$ di dimensioni ($n \times (1 + p)$) e $\boldsymbol{z}_i$ è la *i*-sima riga dalla matrice $\boldsymbol{Z}$ ($n \times$ K), λ è il moltiplicatore di Lagrange che controlla il livello di lisciamento.

Un algoritmo iterativo è stato impiegato per ottenere la stima di $\boldsymbol{\beta}_\psi(q)$ e di $\boldsymbol{\gamma}_\psi$.

4.3 Modello semiparametrico M-quantile ad effetti casuali

Nel modello ad effetti misti così come nel modello M-quantile ad effetti casuali l'assunzione di indipendenza è implicita, tuttavia nelle applicazioni ambientali o epidemiologiche questa è generalmente violata. Infatti, è noto che esiste un certo grado di dipendenza tra osservazioni vicine nello spazio rispetto a quelle lontane. Inoltre la parte fissa del modello potrebbe essere poco flessibile in tutti quei casi in cui la relazione funzionale tra variabili di risposta e covariate non è lineare.

Partendo da quanto proposto dai lavori di Opsomer et al. (2008) e di Pratesi et al. (2006,2008,2009), in questa tesi viene presentata un'ulteriore estensione del modello M-quantile. Tale modello è ottenuto considerando il termine di spline, che implementa una struttura spaziale, come se fosse un effetto random e includendo un'ulteriore componente random nel modello M-quantile per tener conto della struttura gerarchica dei dati. Un modello così formulato consiste, quindi, di una combinazione di un effetto casuale di gruppo e di un termine di spline atto a cogliere la dipendenza spaziale in modo sufficientemente flessibile.
Nella fattispecie, il modello semi-parametrico M-quantile ad effetti random è espresso come:

$$MQ_q(y) = \mathbf{X}\boldsymbol{\beta}_{q\psi} + \mathbf{Z}\boldsymbol{u}_q + \mathbf{Z}_{sp}\boldsymbol{\gamma}_q$$

dove y e $\mathbf{X}$ sono l'usuale vettore dei valori della variabile di risposta e la matrice degli effetti fissi mentre:
$\mathbf{Z}$ è la matrice di dimensione $n \times p$ di variabili binarie che identificano i gruppi
$\boldsymbol{u}_q$ è il vettore degli effetti casuali del gruppo.
$\mathbf{Z}_{sp_{(n \times K)}}$ è la matrice della spline radiale bivariata (*thin plate spline)* costruita, nella fattispecie, utilizzando le coordinate cartografiche $\mathbf{x}_i = (x_{1i}, x_{i2i})$ dei siti di rilevazione:

$$\mathbf{Z}_{sp} = [C(\boldsymbol{x}_i - k_j)]_{\substack{1 \le i \le n, \\ 1 \le j \le K}} [C(k_j - k_{j\prime})]_{\substack{1 \le j \le K, \\ 1 \le j\prime \le K}}^{-1/2}$$

dove $C(\boldsymbol{r}) = ||\boldsymbol{r}||^2 log\, ||\boldsymbol{r}||$ è la funzione di covarianza e k_j ($j=1,\ldots,K$) sono i nodi della spline. Analoga specificazione si ha quando la componente semiparametrica è costituita da una spline univariata, in tale caso $\mathbf{Z}_{sp_{(n \times K)}}$ risulta ad esempio essere

$$\mathbf{Z}_{sp} = \begin{bmatrix} (x_1 - k_1)_+ & \cdots & (x_n - k_K)_+ \\ \vdots & \ddots & \vdots \\ (x_1 - k_1)_+ & \cdots & (x_n - k_K)_+ \end{bmatrix}$$

Tornando al caso bivariato, $\boldsymbol{x}_i = (x_{1i}, x_{2i})$ e K numero di nodi selezionati.

$\boldsymbol{\gamma}_q$ è il vettore degli effetti random associati al termine di spline che ci permette di rappresentare la variabilità spaziale di larga scala.

Nel caso nelle spline bivariate la locazione ideale dei nodi da selezionare è complessa, tuttavia si può ricorrere ad esempio a due algoritmi *space filling designs* (Nychka & Saltzman, 1998) e l'algoritmo clara (Clustering LARge Applications) (Kaufman & Rousseeuw, 1990) che consentono di ottenere un sotto insieme di osservazioni adeguato a coprire lo spazio. Clara è un algoritmo di partizionamento k-medoids per dataset di grandi dimensioni. L'idea dell'algoritmo è la seguente: invece di prendere in considerazione l'intero insieme dei dati, viene scelta una piccola porzione di questi supponendo che essa sia rappresentativa di tutti i dati. I medoidi (il medoid è il punto localizzato più centralmente in un cluster) vengono, quindi, scelti da questo campione usando l'algoritmo PAM (*Partitioning Around Medoids*). Se i campioni vengono selezionati in maniera casuale, dovrebbero rappresentare abbastanza fedelmente l'insieme dei dati originario e i medoidi rappresentativi individuati dovrebbero essere simili a quelli che si sarebbero costruiti utilizzando l'intero insieme dei dati.

La complessità di ciascuna iterazione è $O(ks^2 + k(n - k))$, dove *s* è la dimensione del campione, *k* è il numero di cluster ed *n* è il numero totale di oggetti.

L'efficacia di CLARA dipende dalla dimensione del campione infatti Clara cerca i migliori *k* medoid nel campione selezionato dai dati pertanto non può trovare il miglior clustering se ciascun medoid del campione non è tra i migliori *k* medoid. Per esempio, se un oggetto o_i è uno dei migliori *k* medoidi ma non è selezionato durante il campionamento, CLARA non troverà mai il miglior clustering. Esso, pertanto, è disposto ad accettare una diminuzione della qualità dei risultati finali a vantaggio, però, dell'efficienza.

Il vantaggio nel rappresentare il trend non parametrico mediante la spline penalizzata deriva dalla flessibilità e dalle proprietà di questo strumento come: l'assenza di effetti di bordo che la rende utile nel caso di previsioni, sono lisciatori a basso rango (*low rank smoothers*), cioè la dimensione delle basi è minore rispetto alla dimensione dei dati e infine la scelta dei nodi non è così tanto importante poiché è presente il termine di penalità.

Ai fini della specificazione del modello si assume, infine, che gli effetti random si distribuiscono secondo una distribuzione normale e sono tra di loro indipendenti in particolare:

$$\boldsymbol{u}_q \sim N(0, \sigma^2_{uq}\boldsymbol{I}_m) \text{ effetto gruppo}$$

$$\boldsymbol{\gamma}_q \sim N(0, \sigma^2_{\gamma q}\boldsymbol{I}_K) \text{ effetto random legato al termine di spline}$$

$$e_q \sim N(0, \sigma^2_{eq}\boldsymbol{I}_n) \text{ termine di errore}$$

Da qui, la matrice di varianza e covarianza della variabile risposta diventa

$$\mathbf{V}_q = \boldsymbol{\Sigma}_{eq} + \mathbf{Z}\boldsymbol{\Sigma}_{uq}\mathbf{Z}^{\mathrm{T}} + \mathbf{Z}_{sp}\boldsymbol{\Sigma}_{\gamma q}{\mathbf{Z}_{sp}}^{\mathrm{T}}$$

con

$$\boldsymbol{\Sigma}_{uq} = \sigma^2_{uq}\boldsymbol{I}_m$$

$$\boldsymbol{\Sigma}_{\gamma q} = \sigma^2_{\gamma q}\boldsymbol{I}_K$$

$$\boldsymbol{\Sigma}_{eq} = \sigma^2_{eq}\boldsymbol{I}_n,$$

Utilizzando l'approccio di massima verosimiglianza robusta e il metodo proposto da Tzavidis (2016) (vedi capitolo 1, paragrafo 2), ovvero ottimizzando la funzione di log-verosimiglianza robusta

$$l\left(\beta, \sigma^2_{uq}, \sigma^2_{\gamma q}, \sigma^2_{eq}\right) = = \frac{K_2}{2} - \log|\mathbf{V}| + \rho(\mathbf{U}^{-\frac{1}{2}}(y - \mathbf{X}\boldsymbol{\beta}))$$

si ottengono le stime dei parametri del modello.

Come riportato in precedenza, per risolvere le equazioni, si adotta l'algoritmo di Newton- Raphson per la stima dei coefficienti fissi e il metodo iterativo del punto fisso per le componenti di varianza (Anderson, 1973) secondo la procedura avanzata da Tzavidis et al. (2016). Per la descrizione dell'algoritmo di stima si veda appendice C.

La terza equazione è relativa alla stima dei coefficienti della spline, che, come suggerito da Ruppert et al. (2003), vengono considerati e quindi stimati come se fossero degli effetti random. Questo permette di evitare metodi di stima e tecniche esterne come la tecnica di cross-validation per la scelta del parametro di lisciamento.

Dalla funzione di log-verosimiglianza robusta (proposta II in Richardson and Welsh (1995)) *l(***θ***)*, ($\boldsymbol{\theta} = \boldsymbol{\beta}_q, \sigma^2_{uq}, \sigma^2_{\gamma q}, \sigma^2_{eq}$) e adottando l'idea di stima dei residui asimmetrici pesati definiamo le equazioni di stima dei parametri del modello:

$$\mathbf{X}^{\boldsymbol{T}}\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\mathbf{r}_q) = \mathbf{0}$$ (cofficienti fissi)

$$\frac{1}{2}\psi_q(\mathbf{r}_q)^T\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\mathbf{r}_q) - \frac{K_{2q}}{2}tr\left(\boldsymbol{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\right) = 0$$

(componente di varianza del gruppo)

$$\frac{1}{2}\psi_q(\mathbf{r}_q)^T\mathbf{U}_{\mathrm{q}}^{\frac{1}{2}}\mathbf{V}_{\mathrm{q}}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}\mathbf{V}_{\mathrm{q}}^{-1}\mathbf{U}_{\mathrm{q}}^{\frac{1}{2}}\psi_q(\mathbf{r}_q) - \frac{K_{2q}}{2}tr\left(\mathbf{V}_{\mathrm{q}}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}\right) = 0$$

(componente di varianza della spline)

$$\frac{1}{2}\psi_q(\mathbf{r}_q)^T\mathbf{U}_{\mathrm{q}}^{\frac{1}{2}}\mathbf{V}_{\mathrm{q}}^{-1}\mathbf{V}_{\mathrm{q}}^{-1}\mathbf{U}_{\mathrm{q}}^{\frac{1}{2}}\psi_q(\mathbf{r}_{\mathbf{q}}) - \frac{K_{2q}}{2}tr\left(\mathbf{V}_{\mathrm{q}}^{-1}\right) = 0$$

(componente di varianza residua)

Le derivate seconde pure e miste, calcolate per le tre componenti di varianza, sono rispettivamente:

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial(\sigma_u^2)^2} = -\frac{1}{2}[\mathbf{U}^{-1}r\,\psi'(\mathrm{r})^T]\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1} + \frac{1}{2}\psi(\mathrm{r})^T\mathbf{U}^{-\frac{1}{2}}\mathbf{V}^{-1} - \psi(\mathrm{r})^T\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}(\psi(r)\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1} + 0.5tr((\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}K_2)))$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\sigma_e^2\partial\sigma_u^2} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\sigma_u^2\partial\sigma_e^2} = -\frac{1}{2}[\mathbf{U}^{-1}r\,\psi'(\mathrm{r})^T]\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1} + \frac{1}{2}\psi(\mathrm{r})^T\mathbf{U}^{-1/2}\mathbf{V}^{-1} - \psi(\mathrm{r})^T\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}(\psi(\mathrm{r})\mathbf{U}^{\frac{1}{2}}\boldsymbol{V}^{-1} + 0.5tr((\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^{\mathbf{T}}K_2)))$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial(\sigma_g^2)^2} = -\frac{1}{2}[\mathbf{U}^{-1}r\,\psi'(\mathrm{r})^T]\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}diag(\mathbf{Z}_{\mathbf{sp}}\mathbf{Z}_{\mathbf{sp}}^{\mathbf{T}}) + \frac{1}{2}\psi(\mathrm{r})^T\mathbf{U}^{-1/2}\mathbf{V}^{-1}diag(\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}) - \psi(\mathrm{r})^T\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}\psi(\boldsymbol{r})\,\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1} + 0.5tr\big(\mathbf{V}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}K_2\big)$$

$$\frac{\partial^2 l(\mathbf{\theta})}{\partial\sigma_g^2\partial\sigma_u^2} = \frac{\partial^2 l(\mathbf{\theta})}{\partial\sigma_u^2\partial\sigma_g^2} = -\frac{1}{2}[\boldsymbol{U}^{-1}r\,\psi'(\mathrm{r})^T]\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}diag\big(Z_{sp}Z_{sp}^T\big) + \frac{1}{2}\psi(\mathrm{r})^T\mathbf{U}^{-\frac{1}{2}}\mathbf{V}^{-1}diag\big(\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}\big) - \psi(\mathrm{r})^T\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}\psi(\mathrm{r})^T\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1} + 0.5tr\big(\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}\,\mathbf{V}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}K_2\big)$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial(\sigma_e^2)^2} = -\frac{1}{2}[\mathbf{U}^{-1}r\,\psi'(\mathrm{r})^T]\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1} + \frac{1}{2}\psi(\mathrm{r})^T\mathbf{U}^{-1/2}\mathbf{V}^{-1} - \psi(\mathrm{r})^T\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\mathbf{V}^{-1}\psi(\mathrm{r})\,\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1} + 0.5tr((\mathbf{V}^{-}\mathbf{V}^{-}K_2)$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\sigma_e^2\partial\sigma_g^2} = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\sigma_g^2\partial\sigma_e^2} = -\frac{1}{2}[\mathbf{U}^{-1}r\,\psi'(\mathrm{r})^T]\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1} + \frac{1}{2}\psi(\mathrm{r})^T\mathbf{U}^{-\frac{1}{2}}\mathbf{V}^{-1} - \psi(\mathrm{r})^T\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}\psi(r)\mathbf{U}^{\frac{1}{2}}\mathbf{V}^{-1} + 0.5tr(\big(\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{Z}_{\mathrm{sp}}\mathbf{Z}_{\mathrm{sp}}^{\mathrm{T}}K_2\big)$$

Per valutare la precisione della stima calcoliamo gli errori standard analiticamente. Invertendo la matrice Hessiana $\mathbf{H}_{(3\times3)}$ otteniamo la matrice osservata di informazione di Fisher $\mathbf{I}(\theta)$ (è definita come l'opposto del valore atteso dell'Hessiana della log-verosimiglianza), quindi risulta:

$$\mathbf{I}_{(3\times3)}(\theta) = -\mathbf{H}_{(\mathbf{3}\times\mathbf{3})}^{-\mathbf{1}}(\boldsymbol{\theta})$$

La radice quadrata degli elementi di posizione $I_{[1,1]}$, $I_{[2,2]}$ e $I_{[3,3]}$ restituiscono gli errori standard della varianza dell'effetto random di gruppo σ_u, della varianza del termine di spline σ_g e della varianza residua σ_e:

$$\sigma_u = \sqrt{(I_{[1,1]})}$$
$$\sigma_g = \sqrt{(I_{[2,2]})}$$
$$\sigma_e = \sqrt{(I_{[3,3]})}$$

Per quanto riguarda la stima degli errori standard dei coefficienti fissi del modello calcoliamo le derivate prime e seconde rispetto ai $\boldsymbol{\beta}$ e successivamente, tramite lo stimatore sandwich, otteniamo gli errori standard. La funzione di score è:

$$\mathbf{S} = \frac{\partial l}{\partial \beta_i} = \frac{\mathbf{X^T V^{-1} U^{\frac{1}{2}} \psi(r)^T \psi(r) U^{\frac{1}{2}} V^{-1} X}}{n-p}$$

mentre la derivata seconda è

$$\mathbf{G} = \frac{\partial^2 l}{\partial \beta_i^2} = \mathbf{X^T V^{-1}}\, \psi'(\mathbf{r})\mathbf{X}$$

Da qui si ottiene la matrice osservata di informazione di Fisher tramite lo stimatore sandwich

$$\mathbf{B} = \mathbf{G^{-1} S\, G^{-1}}$$

Prendendo la radice quadrata degli elementi posti sulla diagonale di B si ottengono gli errori standard dei coefficienti fissi.

In questa sezione abbiamo introdotto il modello semi-parametrico M-quantile ad effetti random in cui il temine semi-parametrico è una spline bivariata. Per il modello che include un termine di spline univariata valgono esattamente gli stessi metodi di stima.

Nella prossima sezione saranno descritti tre studi di simulazione *model-based* condotti per valutare la performance complessiva del modello in termini di stima e capacità predittiva, sia in assoluto che relativamente a un modello alternativo proposto da Pratesi et al. nel 2009.

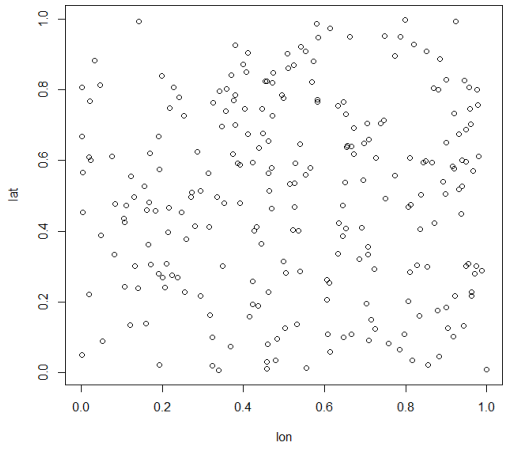
5. DISEGNO DI SIMULAZIONE BASATO SUL MODELLO

Sono stati predisposti ed effettuati tre distinti disegni di simulazione con obiettivi differenti: con il primo disegno si vuole valutare la performance di stima del modello, con il secondo si indaga la prestazione a livello di predizione e con l'ultimo l'efficacia del modello proposto nel caso univariato, comparandolo con il modello Non-parametric M-quantile P-spline proposto da Pratesi et al. (2009).

5.1 Primo disegno di simulazione

Fissato il numero di osservazioni pari a n = 500, su una regione quadrata di dimensioni [0,1]×[0,1], generiamo da una uniforme n punti $\mathbf{x} = (x_{1i}, x_{2i})\ i = 1, \dots n$ corrispondenti alle coordinate geografiche dei punti di misura, dislocati nello spazio (Fig. 1). Con riferimento all'analisi del radon indoor in Lombardia, proposta nel capitolo 6, si suppone, ad esempio, che ciascun punto di misura identifichi un edificio, ognuno dei quali posizionato su una precisa classe litologica e che i dati simulati, abbiano una struttura gerarchica. La variabile che determina il raggruppamento, rappresentata dalla classe litologica nel capitolo relativo all'applicazione ambientale, è modellizzata tramite un effetto random a 11 livelli. Ciascun livello compare nel dataset simulato con una diversa frequenza di punti di misura. Al fine di mantenere un parallelismo tra lo studio di simulazione e l'analisi, la frequenza, nelle varie classi della variabile di gruppo, è proporzionale alle frequenze osservate nella situazione reale.

Figura 2 I cerchi neri (o) rappresentano le coordinate spaziali (lat=latitudine e lon=longitudine) dei siti ottenuti alla prima delle 1000 iterazioni del dataset



Le due variabili spaziali andranno a costituire la parte non stocastica del modello. Il vettore dei relativi coefficienti non stocastici sarà indicato con $\boldsymbol{\beta}$.

Il disegno di simulazione consiste nei seguenti passi:

a) Simulazione delle tre componenti casuali relative al quantile q-esimo: l'effetto random relativo al termine di spline γ, la componente di variabilità di gruppo u e la componente di variabilità residua e da distribuzioni normali indipendenti

$$\boldsymbol{\gamma}_q \sim N(0, \sigma^2_{gq})$$

$$\boldsymbol{u}_q \sim N\left(0, \sigma^2_{uq}\right)$$

$$e_q \sim N(0, \sigma^2_{eq})$$

con valori di σ^2_{gq} σ^2_{uq} e σ^2_{eq}, corrispondenti rispettivamente alle componenti di varianza del termine di spline, dell'effetto random legato al gruppo e all'errore, fissati e scelti tra diverse combinazioni di valori.

b) Costruzione della matrice delle funzioni di base spline radiale bivariata ($\mathbf{Z}_{sp}$) utilizzando come covariate il vettore $\tilde{\mathbf{x}}$ delle locazioni $\tilde{\mathbf{x}}_i = (x_{1i}, x_{2i})$ identificate dalle coordinate di ciascun punto i incluso nel campione. Per allocare i nodi κ_j j=1,…, K nello spazio, utilizziamo l'algoritmo clara implementato nel pacchetto `cluster` (Maechler et al., 2016) del software `R` (`R` core team, 2015), il quale ci consente di avere una copertura adeguata nello spazio delle osservazioni. Il numero di nodi K può essere modificato di volta in volta per studiare l'influenza di questi sulla stima dei parametri del modello.

Si definiscono di seguito le seguenti matrici:

$$\mathbf{Z}_R = [||x_i - k_j|^{2r-2} log||x_i - \kappa_j|||]_{1\leq i\leq n, 1\leq j\leq K}$$

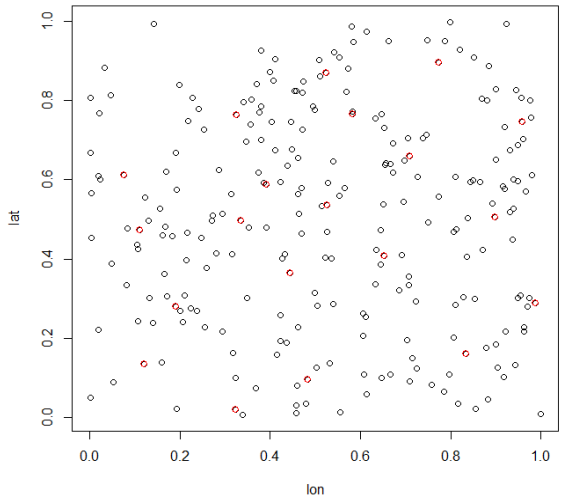
$$\boldsymbol{\Omega} = [||k_s - k_t|^{2r-2} log||k_s - k_t||]_{1\leq s,t\leq K}$$

$$\boldsymbol{Z}_{sp} = \boldsymbol{Z}_R \boldsymbol{\Omega}^{-1/2}$$

La matrice $\boldsymbol{Z}_{sp}$ andrà a pre-moltiplicare $\boldsymbol{\gamma_q}$ ossia l'effetto random legato al termine di spline (Opsomer et al., 2008, Ruppert et al., 2003).

Nel grafico in figura 3 i cerchi rossi (°) sono le posizione dei K =50 knots (nodi) utilizzati nella simulazione, mentre i restanti cerchi (°) sono le coordinate spaziale delle locazioni dei punti di misura. La matrice $\boldsymbol{Z}$ è la matrice del disegno connessa al gruppo litologico.

Figura 3. Posizione dei nodi della funzione spline radiale bivariata



c) Calcolo dei valori della variabile di risposta y una volta scelti i coefficienti fissi β attraverso l'equazione del modello:

$$y = \mathbf{X}\boldsymbol{\beta}_q + \mathbf{Z}\boldsymbol{u_q} + \mathbf{Z}_{\boldsymbol{sp}}\boldsymbol{\gamma_q} + e$$

I valori delle varianze vengono fissati all'inizio di ogni ciclo di simulazioni.

d) Il processo di generazione del dataset viene ripetuto T=1000 volte stimando, su ciascun set di dati simulato, le tre componenti di varianza e i coefficienti fissi del modello.

Infine, per valutare la performance e stabilità della procedura di stima, per ciascun vettore dei parametri stimati, calcoliamo il *Mean Relative Bias* (MRB) percentuale e il *Root Mean Square Error* relativo (RMSE). Le formule degli indici vengono riportate di seguito:

$$\text{MRB\%} = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\theta}^{(t)} - \theta}{\theta} \times 100$$

$$\text{RMSE} = \sqrt{\left[\frac{\sum \frac{(\hat{\theta}^{(t)} - \theta)^2}{\theta^2}}{T}\right]}$$

dove per $\hat{\theta}$ si intende il valore stimato e per θ il valore vero fissato a priori combinando tre valori della componente di varianza del termine di spline (basso (25), medio(100), alto (400)), e due valori per ciascuna delle rimanenti componenti di varianza (basso pari a 25 sia per σ_u^2 che σ_e^2, e alto pari a 400 per σ_e^2 e 100 per σ_u^2). Per la componente spline si sono considerati più livelli di valori essendosi questa dimostrata più difficile da stimare e si è quindi ritenuto opportuno valutare il comportamento della stima su una griglia più fitta di possibili valori. Sono state fatte numerose ulteriori simulazioni a questo proposito (non riportate esplicitamente nella tesi per motivi di spazio) che comunque hanno fornito risultati in linea con quanto riportato nel seguito. La procedura è ripetuta per gli M-quantili di ordine 0.25, 0.50 e 0.75.

5.2 Risultati

Per varie combinazioni di valori delle componenti di varianza e per i diversi M-quantili sono riportati in tabella (Tabella 1) gli indici RMSE e MRB%. Per i coefficienti fissi otteniamo valori ottimali per entrambi gli indici per tutti gli M-quantili considerati.

Mentre per quanto riguarda le componenti di varianza, in generale, si ha un peggioramento nei valori di RMSE e MRB% andando verso i quantili più estremi.

Ad esempio, le variazione in percentuale rispetto al valore vero della componente di varianza spline, più alte in valore assoluto si hanno quando il valore fissato per le componenti di varianza del termine di spline, di gruppo e d'errore è basso (σ_g^2=25, σ_u^2=25, σ_e^2=25). Lo stesso risultato si ha quando i valori veri della componente spline e di errore sono elevati e la varianza del gruppo è bassa (σ_g^2=400, σ_e^2=400 e σ_u^2=25**)**, quando le componenti di varianza di gruppo e del termine di errore sono bassi e la componente di varianza spline è fissata ad un valore intermedio (σ_g^2=100, σ_u^2=25, σ_e^2=25**).**

Valori di MRB% migliori per tutti gli M-quantili si ottengono nel caso in cui il valore per la componente di varianza del gruppo è alto (σ_u^2=100**)**, quello della componente d'errore (σ_e^2=25) e quello della spline è basso (σ_g^2=25**).**

In tutte le combinazioni dei valori fissati, si può notare una sottostima, più o meno marcata (evidente in particolar modo agli M-quantili estremi) della componente di varianza spline.

La combinazione contrassegnata in tabella con due asterischi (**) indica che il valore della componente di varianza d'errore è fissata a σ_e^2=36 invece del consueto 25. L'innalzamento di questo valore si è reso necessario per portare a termine tutte le 1000 simulazioni in tutti gli M-quantili, in quanto dopo alcune simulazioni veniva generato un dataset sul quale l'algoritmo di stima non raggiungeva la convergenza bloccando in questo modo tutto il processo di simulazione.

Tabella 1. Valore degli indici RMSE E MRB% per ogni valore fissato delle componenti di varianza e M-quantili

Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=25**	$\boldsymbol{\sigma_g^2}$**=25**	$\boldsymbol{\sigma_e^2}$ **=25**
M-quantile=25						
MRB%	-3.10	-0.34	-0.097	-9.03	-20.72	-13.93
RMSE	0,06	0,14	0,18	0,53	0,72	0,29
M-quantile=50						
MRB%	0,13	-0,68	0,17	-3,97	-13,48	5,63
RMSE	0,05	0,10	0,15	0,51	0,61	0,11
M-quantile=75						
MRB%	3.29	-0.53	0.21	-10,46	-20,46	-10,38
RMSE	0,06	0.33	0.17	0,52	0,79	0,28
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=25**	$\boldsymbol{\sigma_g^2}$**=100**	$\boldsymbol{\sigma_e^2}$ **=25**
M-quantile=25						
MRB%	-2.45	-2.93	0.26	-22.39	-32.78	-8.57
RMSE	0.75	0.34	0.61	0.61	0.64	0.24
M-quantile=50						
MRB%	0.57	-0.60	-2.63	1.80	-10.94	9.87
RMSE	0.09	0.27	0.56	0.53	0.53	0.14
M-quantile=75						
MRB%	2.30	-2.67	4.19	-17.41	-18.60	-13.22
RMSE	0.23	0.36	0.61	0.62	0.57	0.23
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=25**	$\boldsymbol{\sigma_g^2}$**=400**	$\boldsymbol{\sigma_e^2}$ **=25**
M-quantile=25						
MRB%	-2.58	-0.70	-5.01	-5.3	-10.74	3.6
RSME	0,34	0,65	0,91	0.34	0.88	0.17
M-quantile=50						
MRB%	-3.40	2.22	-0.13	5.15	-8,14	6,89
RMSE	0,47	0,55	0,89	0.28	0.68	0.12
M-quantile=75						
MRB%	6.02	-2.66	-2.88	-3.17	-12.00	5.57
RMSE	0,37	0,67	0,92	0.31	0.82	0.12
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=25**	$\boldsymbol{\sigma_g^2}$**=25**	$\boldsymbol{\sigma_e^2}$ **=400**
M-quantile=25						
MRB%	-4,10	-2,40	-6,68	-5,87	-8,43	-3,58
RMSE	0,19	0,92	3,51	0,35	0,36	0,09
M-quantile=50						
MRB%	-1,04	1,38	2,745	-2,63	-1,57	6,65
RMSE	0,18	1,04	3,57	0,30	0,21	0,015
M-quantile=75						
MRB%	4,02	3,32	1,26	-3,21	-3,11	-8,30
RMSE	0,20	1,23	3,58	0,44	0,29	0,27
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**= 25**	$\boldsymbol{\sigma_g^2}$**=100**	$\boldsymbol{\sigma_e^2}$ **=400**
M-quantile=25						
MRB%	-6.41	1.36	-5.26	-6,31	-13,09	-18,58
RMSE	0.15	0.57	1.83	0.56	0.69	0.32
M-quantile=50						
MRB%	0.24	-2.99	0.03	-3,18	-13,91	-3,18
RMSE	0.11	0.54	1.74	0.51	0.53	0.12
M-quantile=75						

MRB%	5.28	-0.29	-3.13	-5,24	-11,77	-18,49
RMSE	0.19	0.55	1.75	0.55	0.73	0.33
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=25**	$\boldsymbol{\sigma_g^2}$**=400**	$\boldsymbol{\sigma_e^2}$ **=400**
M-quantile=25						
MRB%	-1.32	-2.21	5.37	-11.00	-30.02	-7.31
RMSE	0.40	1.12	0.51	0.84	1.02	0.56
M-quantile=50						
MRB%	0.63	-3.33	-0.74	12.78	-29.99	4.42
RMSE	0.36	1.03	0.48	0.80	0.99	0.52
M-quantile=75						
MRB%	4.79	0.54	4.75	14.80	-30.14	-6.18
RMSE	0.40	1.09	0.51	0.83	1.01	0.55
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=100**	$\boldsymbol{\sigma_g^2}$**=25**	$\boldsymbol{\sigma_e^2}$ **=25**
M-quantile=25						
MRB%	-5.53	1.19	0.17	-3.39	-8.00	-6.31
RMSE	0.08	0.16	0.45	0.54	0.89	0.27
M-quantile=50						
MRB%	-0.37	1.01	0.23	-1.28	-3.95	2.06
RMSE	0,06	0,15	0,17	0.52	0.63	0.15
M-quantile=75						
MRB%	4.68	0.90	0.28	-3.01	-7.76	-5.66
RMSE	0.04	0.19	0.38	0.59	0.77	0.15
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=100**	$\boldsymbol{\sigma_g^2}$**=100**	$\boldsymbol{\sigma_e^2 = 25}$
M-quantile=25						
MRB%	-4.32	1.55	-0.19	-26.2	-33.15	-1.41
RMSE	0.10	0.29	0.33	0.52	0.64	0.23
M-quantile=50						
MRB%	-0.38	1.41	0.43	-2.63	-19.69	15.37
RMSE	0.09	0.28	0.33	0.53	0.54	0.18
M-quantile=75						
MRB%	3.48	1.50	0.99	-26.04	-33.81	1.01
RMSE	0.10	0.29	0.34	0.49	0.62	0.25
Valore vero**	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=100**	$\boldsymbol{\sigma_g^2}$**=400**	$\boldsymbol{\sigma_e^2}$ **=36**
M-quantile=25						
MRB%	-2.07	-10.05	-6.93	-22.81	-19.98	18.37
RMSE	0.10	0.27	0.35	0.38	0.10	0.22
M-quantile=50						
MRB%	0.17	-1.17	-0.04	2.91	-9.11	19.84
RMSE	0.03	0.23	0.30	0.35	0.09	0.20
M-quantile=75						
MRB%	2.77	-0.73	-3.94	-17.73	11.54	11.39
RMSE	0.06	0.27	0.36	0.37	0.11	0.22
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=100**	$\boldsymbol{\sigma_g^2}$**=25**	$\boldsymbol{\sigma_e^2}$ **=400**
M-quantile=25						
MRB%	3,35	-0,67	4,32	7,39	5,99	5,20
RMSE	0,08	0,36	1,23	0,92	1,04	0,15
M-quantile=50						
MRB%	0,35	-0,002	-0,06	-8,76	-4,49	6,28
RMSE	0,06	0,31	1,10	0,49	0,95	0,10
M-quantile=75						
MRB%	-9,54	1,37	-1,66	-9,63	6,86	5,80
RMSE	1,1	0,74	1,29	0,51	1,00	0,20
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=100**	$\boldsymbol{\sigma_g^2}$**=100**	$\boldsymbol{\sigma_e^2}$ **=400**
M-quantile=25						
MRB%	-7.48	-0.76	-1.34	-0.32	-0.38	-0.19
RMSE	0.13	0.33	0.38	0.53	0.71	0.22
M-quantile=50						
MRB%	0.97	-2.70	-3.78	-7.08	-6.29	3.17
RMSE	0.11	0.30	0.36	0.51	0.70	0.21
M-quantile=75						
MRB%	10.88	-1.23	-1.72	-4.32	-16.32	-9.40
RMSE	0.12	0.31	0.38	0.55	0.78	0.31
Valore vero	$\boldsymbol{\beta_0}$**=100**	$\boldsymbol{\beta_1}$**=22,8**	$\boldsymbol{\beta_2}$**=7,0**	$\boldsymbol{\sigma_u^2}$**=100**	$\boldsymbol{\sigma_g^2}$**=400**	$\boldsymbol{\sigma_e^2}$ **=400**
M-quantile=25						

MRB%	-11.65	1.22	10.58	-6.24	-14.27	-5.27
RMSE	0.23	0.62	0.70	0.65	0.88	0.30
M-quantile=50						
MRB%	-1.23	-0.44	8.74	-4.42	-11.36	2,99
RMSE	0.18	0.60	0.65	0.53	0.70	0,10
M-quantile=75						
MRB%	9.49	-0.64	5.88	-8.10	-17.46	4.38
RMSE	0,21	0,61	0,66	0,56	0,78	0,31

5.3 Secondo Disegno di simulazione

Con il secondo disegno di simulazione si vuole valutare la capacità predittiva del modello. A tal fine è stata costruita una griglia regolare di $n = 500$ siti utilizzata come reticolato di punti di previsione. I siti della griglia sono stati classificati in $g = 11$ gruppi. Analogamente al caso precedente le coordinate dei punti di misura sono state ottenute simulando le due variabili x_{i1}, x_{i2} da due distribuzioni uniformi nell'intervallo [0,1].

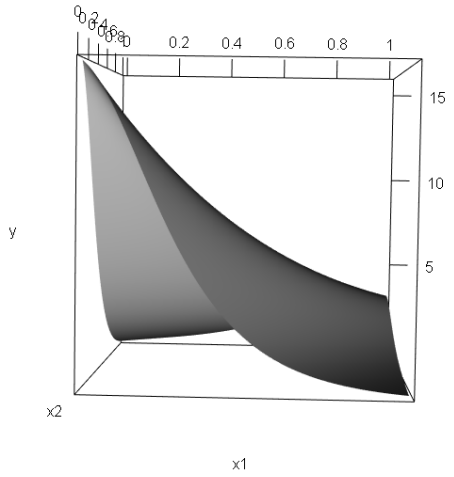
La relazione funzionale tra le covariate x_{i1}, x_{i2} e la variabile di risposta y_i con $i = 1, \dots\dots, n$ (Figura 4a e 4b) ha le seguenti forme:

1. $f(x_1, x_2) = \left(-(4x_1^2 + 4x_2^2 - 5)\right)$

2. $f(x_1, x_2) = \frac{40 \times exp\left[8\{(x_1-0.5)^2+(x_2-0.5)^2\}\right]}{exp[8\{(x_1-0.2)^2+(x_2-0.7)^2\}]+exp[8\{(x_1-0.7)^2+(x_2-0.2)^2\}]}$

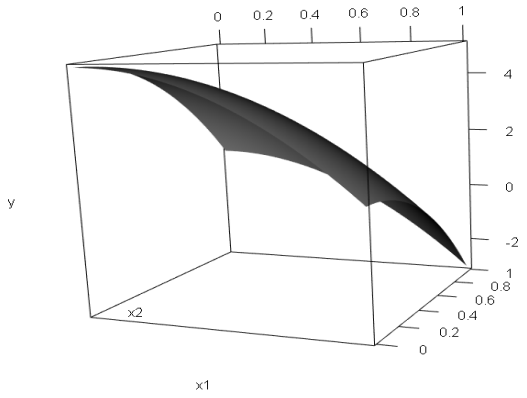
Figura 4 Grafico delle relazioni funzionali

a



$$f(x_1, x_2) = \frac{40 \times exp[8\{(x_1 - 0.5)^2 + (x_2 - 0.5)^2\}]}{exp[8\{(x_1 - 0.2)^2 + (x_2 - 0.7)^2\}] + exp[8\{(x_1 - 0.7)^2 + (x_2 - 0.2)^2\}]}$$

b



$$f(x_1, x_2) = \left(-(4x_1^2 + 4x_2^2 - 5)\right)$$

Ad ogni iterazione sono simulati gli effetti casuali *u,* cioè l'effetto gruppo, e la componente di errore *e.* I valori della variabile di risposta y_{ij} relativi a ciascun punto x_{i1}, x_{i2} si ottengono attraverso l'equazione

$$y_{ji} = f(x_{i1},\ x_{i2})\ + u_j + e_{ji}$$

dove *j* è l'indice del gruppo *j*=1,….,11 e u_j e e_{ji} si distribuiscono come due normali incorrelate

$$e_{ji} \sim N(0, \sigma_e^2)$$

$$u_j \sim N(0, \sigma_u^2)$$

Volendo valutare anche le proprietà di robustezza del modello, sono stati indagati quattro scenari di simulazione con riferimento alla presenza di outlier nella parte stocastica del modello:

1) effetto random di gruppo e componente d'errore prive di outliers, entrambe si distribuiscono come una normale di media 0 e deviazione standard 2;

2) errore generato da una mistura di gaussiane, di cui il 10% è una gaussiana di media 0 e deviazione standard 9 e il rimanente 90% è una normale di media zero e deviazione standard 2;

$$e_{ji} \sim 0{,}9 \times N(0; 2) + 0{,}1 \times N(0; 9)$$

mentre l'effetto random è privo di outliers e si distribuisce come al punto 1;

3) effetto random di gruppo con outliers ed componente casuale come il punto 1. Gli effetti random per le classi che vanno da 1 a 7 sono generati da una normale di media 0 e deviazione standard pari a 3 e i rimanti (da 8 a 11) da una distribuzione normale di media 0 e deviazione standard 9 (outliers);
4) entrambi le componenti sono affette da outliers (vedere punti 2 e 3).
Come nel caso precedente sono stati considerati gli M-quantili di ordine q = 0.25, q = 0.5 e q = 0.75. Per ciascun M-quantile sono state calcolate le seguenti quantità quali indici di performance della previsione: MASE (mean average squared errors) e il coefficiente di determinazione R^2.

$$MASE = \frac{1}{T \times N} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\hat{m}_{\psi,q}^{t}[x_{i1}, x_{i2}] - m_{\psi,q}[x_{i1}, x_{i2}] \right)^2$$

$$R^2 = cor^2(\hat{m}_{\psi,q}^{t}[x_{i1}, x_{i2}]; m_{\psi,q}[x_{i1}, x_{i2}])$$

dove $m_{\psi,q}(x_{i1}, x_{i2})$ è la vera distribuzione ottenuta a ciascun quantile e $\hat{m}_{\psi,q}^{r}$ è la stima ottenuta.

T=1000 indica il numero di repliche effettuate.

5.4 Risultati

I risultati scaturiti dal disegno di simulazione che valuta la bontà di predizione sono presentati in tabella 2 e 3 e in grafico in figura 5.
In tabella 2 sono riportati valori del MASE per tutti e quattro gli scenari investigati. In figura 5 vi sono i boxplot dei valori di R^2 i cui valori medi per ciascun quantile sono riportati in tabella 3.

Tabella 2. Valori del MASE per ciascun scenario e M-quantile (0.25, 0.50, 0.75).

$f(x_1, x_2)$	M-Quantile=0.25	M-Quantile=0.50	M-Quantile=0.75	Contaminazione
Exp	5,3	10,0	7, 7	no
Exp	7,3	5,7	3,1	outliers in *e*
Exp	10,0	10,4	8,5	outliers in *u*
Exp	8,5	6,2	3,7	outliers in *e* e *u*
Par	6,3	9,3	6,8	no
Par	3,1	5,7	3,2	outliers in *e*
Par	9,5	9,7	7,9	outliers in *u*
Par	8,4	5,8	3,7	outliers in *e* e *u*

'Exp' è la relazione esponenziale tra y e le covariate x_1,e x_2; 'Par' è la relazione funzionale paraboloide tra y e le covariate.
E' anche specificato se i dati simulati sono contaminati, i.e., se prevedono outliers nella condizione sperimentale.

In generale, i valori più elevati del MASE si hanno quando consideriamo la relazione esponenziale con outliers nell'effetto random di gruppo (i cui valori sono 10, 10.4, 8.5 rispettivamente per M-quantile 0.25, 0.50 e 0.75) mentre i valori più bassi si manifestano quando è considerata una relazione

di tipo paraboloide e contaminazione a livello di errori (i cui valori sono 3.1, 5.7, 3.2 rispettivamente per M-quantile 0.25, 0.50 e 0.75).

Per la relazione esponenziale il valore più basso dell'indice è 3.1 in corrispondenza del M-quantile più estremo (e con la presenza di outliers nel termine di errore) mentre il più elevato è 10.4 per il M-quantile centrale (presenza di outliers nel termine di gruppo).

Per la relazione funzionale di tipo paraboloide (con outliers nel termine di errore) i valori più bassi del MASE sono 3.1 e 3.2 in corrispondenza dei quantili più estremi mentre il più alto valore è di 9.7 per il quantile centrale e outliers nel termine di gruppo.

Tabella 3. Valore medio del coefficiente di determinazione R^2 per ciascun quantile

$f(x_1, x_2)$	M-Quantile=0,25	M-Quantile=0, 50	M-Quantile=0,75	Contaminazione
Exp	0,94	0,96	0,95	no
Exp	0,97	0,96	0,96	outliers in *e*
Exp	0,96	0,97	0,96	outliers in *u*
Exp	0,97	0,98	0,98	outliers in *e* e *u*
Par	0,95	0,95	0,94	no
Par	0,97	0,96	0,96	outliers in *e*
Par	0,72	0,98	0,97	outliers in *u*
Par	0,97	0,98	0,98	outliers in *e* e *u*

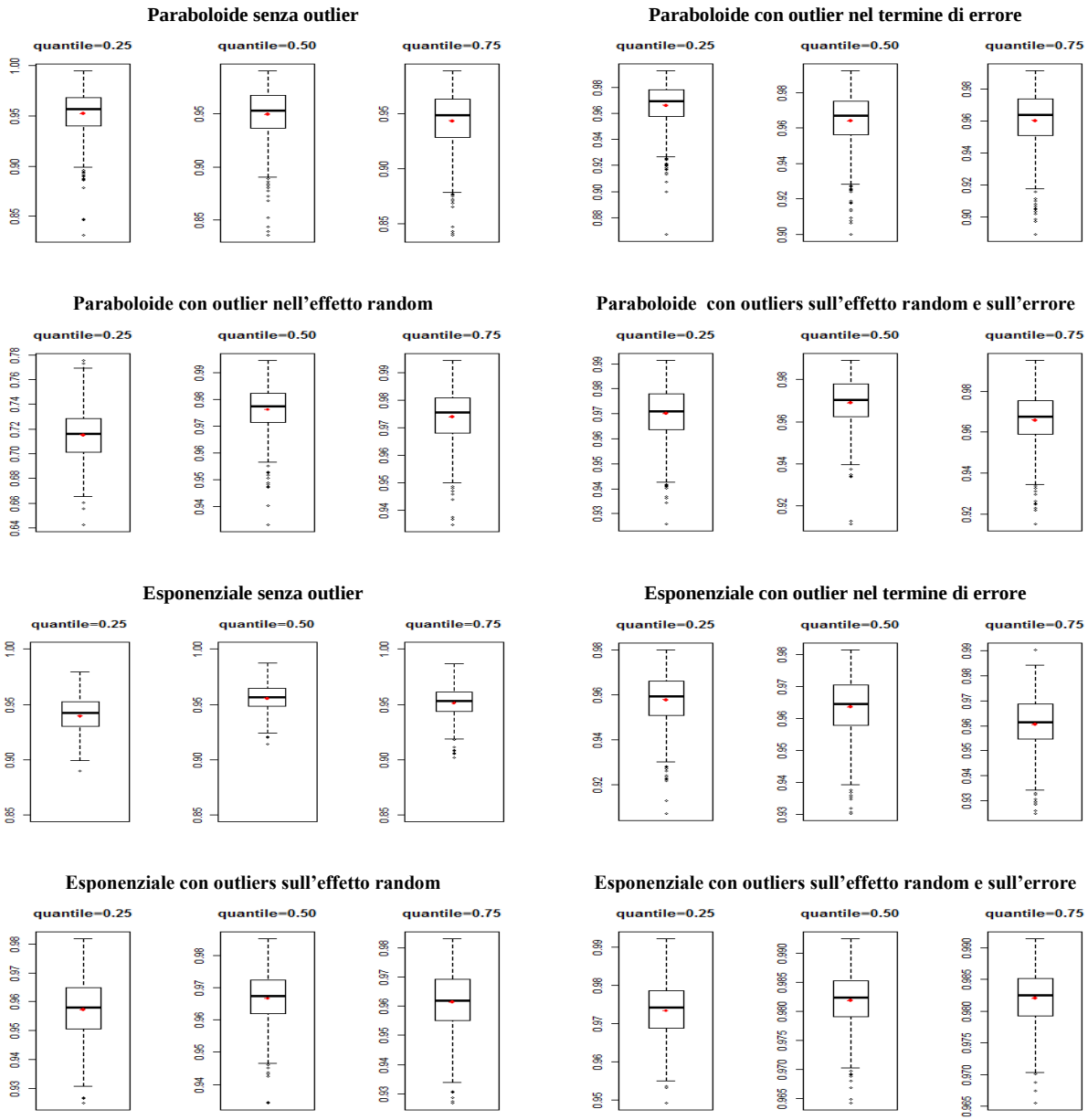
'Exp' è la relazione esponenziale tra y e le covariate x_1,e x_2; 'Par' è la relazione funzionale paraboloide tra y e le covariate.
E' anche specificato se i dati simulati sono contaminati, i.e., se prevedono outliers nella condizione sperimentale.

I valori medi del coefficienti di determinazione sono prossimi alla unità per tutte le tipologie di relazioni funzionali considerate e variano in un intervallo che va da un minimo di 0.72 ad un massimo di 0.98.

In generale, sia per la relazione funzionale esponenziale che per quella paraboloide i valori più alti si hanno nel caso in cui c'è presenza di outliers in entrambe le componenti stocastiche del modello.

Si hanno comunque performance comparabili del modello proposto per le due specificazione alternative. L'approccio considerato sembra quindi ragionevolmente valido qualunque sia la relazione funzionale esistente fra le diverse variabili.

Figura 5 Boxplot dei coefficienti di determinazione R^2 per ciascun scenario e M-quantile



Nei boxplot riportati in figura 5 sono rappresentati i valori di R^2 calcolati per ciascun scenario investigato. Il punto rosso indica il valore medio di R^2. In generale lo scostamento del valore medio dal valore mediano è irrilevante pertanto si può concludere che la distribuzione è approssimativamente simmetrica. In tutti si manifesta la presenza di outliers verso il basso.

5.5 Terzo Disegno di simulazione

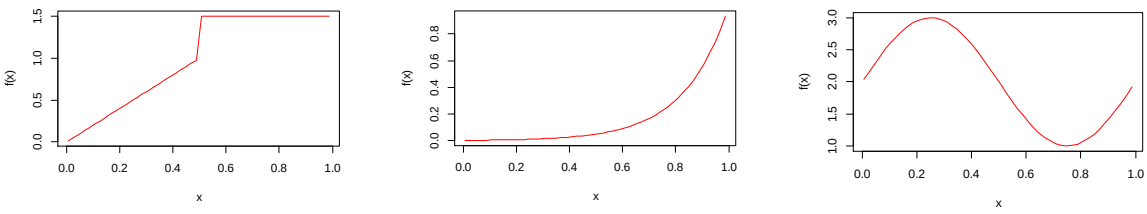
Questo ulteriore studio di simulazione è stato condotto per confrontare il modello proposto con il modello non-parametrico M-quantile P-spline proposto da Pratesi et al. 2009.
Per rendere i risultati comparabili è stato replicato il disegno di simulazione utilizzato dagli autori.
Per generare la vera relazione tra la variabile di predizione univariata *x* e la variabile di risposta *y* sono stati usati i tre modelli riportati qui di seguito (grafici in Figura 6).

$$f(x) = \{1 + 2 \times (x - 0.5) \times (x <= 0.5) + 0.5 \times (x > 0.5)\}, (a\ salti)$$

$$f(x) = \left\{\frac{exp(6 \times x)}{400}\right\}, (esponenziale)$$

$$f(x) = \{2 + sin(2 \times \pi \times x)\}, (ciclica)$$

Figura 6 Grafici delle relazioni funzionali univariate: a) a salti, b) esponenziale e c) ciclica



a) Funzione a salti **b) Funzione esponenziale** **c) Funzione ciclica**

In questo caso fissiamo il numero di osservazioni n = 200 e generiamo i valori della variabile *x* da una distribuzione uniforme in [0, 1]; i valori di *y* sono generati a ciascuna replica addizionando il termine di errore alla relazione funzionale definita sopra.

$$y = f(x) + e$$

Per quanto concerne la distribuzione della componente di errore sono stati considerati due distinti scenari:

1. distribuzione normale di media 0 e deviazione standard 0,4;

2. errore normale con outliers: mistura di gaussiane in cui il 10% delle realizzazioni (outliers) si distribuisce come una normale di media 0 e deviazione standard 2 e il rimanente 90% è estratto come una normale di media 0 e deviazione standard 0,4.

$$e_i \sim 0{,}9 \times N(0;\ 0{,}4) + 0{,}1 \times N(0;\ 2)$$

Per il modello semiparametrico M-Quantile ad effetti random (nel seguito SMQRE) è stata usata una base lineare troncata e numero di nodi K = 39 per ciascun quantile considerato. La selezione dei nodi è avvenuta mediante la funzione `default.knots()` del pacchetto `SemiPar` (Wand, 2014) del sofware `R`.
Al fine di comparare le performances dei due modelli è stato calcolato a ciascun M-quantile (q=0.20, 0.50, 0.70) il Mean Average Squared Errors (MASE) definito come:

$$MASE = \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\hat{m}_q{}^t(x_i) - m_q(x_i) \right)^2 / T$$

dove T=1000 è il numero di repliche

In tabella 4 viene riportato il rapporto tra i MASE dei due modelli per ciascun scenario (dati non contaminati e contaminati), ciascuna relazione funzionale e per ciascun M-quantile. Il MASE di riferimento è quello del modello non parametrico M-Quantile P-spline il cui rapporto è di 1.
Un valore maggiore di 1 è una evidenza contro il modello proposto mentre un valore minore di 1 è a vantaggio di quest'ultimo.

5.6 Risultati

Ad esclusione della relazione ciclica su dati contaminati, dove il rapporto tra MASE è inferiore (M-quantile 0.20) o leggermente superiore all'unità, il modello proposto ha una prestazione peggiore rispetto al modello di riferimento.

Per quanto attiene la relazione funzionale esponenziale (dati senza contaminazione), si sono ottenuti valori molto elevati ai due M-quantili estremi (contrassegnati in tabella con due asterischi **).

Tabella 4. Rapporto tra MASE ciascuno scenario e per ciascun M-quantile

$f(x_1, x_2)$	**M-Quantile=0,20**	**M-Quantile=0,50**	**M-Quantile=0,70**	**Contaminazione**
a salti	5,6	12	8,15	no
esponenziale	**	0,1	**	no
ciclica	7,3	0,8	4,3	no
a salti	2,3	1,8	2,5	si
esponenziale	11,04	14,4	6,9	si
ciclica	0,9	1,3	1,3	si

**valore elevato

In assoluto la prestazione peggiore si riscontra quando si considera la relazione esponenziale i cui dati sono affetti da contaminazione.

Si evidenzia anche che, solo per l'M-quantile centrale e per la funzione relazionale esponenziale e ciclica in assenza di outliers il modello proposto sembra rispondere meglio.

6. ANALISI DELLE CONCENTRAZIONI DEL RADON INDOOR IN LOMBARDIA TRAMITE IL MODELLO M-QUANTILE

In questo capitolo è presentata l'applicazione del modello, descritto nel capitolo 4, sui dati di concentrazione di Radon indoor. Dopo la descrizione del dataset e il calcolo delle statistiche descrittive, viene applicato il modello semiparametrico M-quantile ad effetti random con l'obiettivo di ottenere le mappe di IRC e relativamente al M-quantile 0.85 la mappa Radon Prone Area. Inoltre, facendo uso dei valori stimati dal modello è stato possibile individuare i profili abitativi più a rischio in termini di IRC elevati.

6.1 Introduzione

Negli ultimi decenni il riconoscimento del danno indotto dall'esposizione al gas Radon sulla salute umana ha portato molte istituzioni nazionali ed internazionali a emanare normative ed elaborare/attuare piani d'azione (e.g. Piano Nazionale Radon in Italia) per la mitigazione del rischio di esposizione della popolazione e l'identificazione, mediante campagne di monitoraggio, di quelle aree geografiche che hanno maggiore probabilità di avere un'alta concentrazione di radon (*radon prone areas*).

Come già detto nel capitolo introduttivo (capitoli 1), la principale sorgente naturale di radon indoor è il radon presente nella frazione gassosa dei pori del suolo, la cui concentrazione è fortemente connessa alle caratteristiche geologiche del territorio. La concentrazione di radon indoor è influenzata dalla permeabilità del suolo e delle rocce e dal contenuto di radio e altre miscele di gas che possono rilasciare radionuclidi. Sorgenti secondarie di radon sono i materiali da costruzione degli edifici.

Molti modelli statistici sono stati utilizzati per studiare l'influenza di diverse variabili (sia geologico-strutturali che costruttive) sulla concentrazione di radon indoor (IRC), tra cui i modelli di regressione classici. Tuttavia le tecniche di regressione sulla media condizionata presentano dei limiti, qualora si è interessati alle code della distribuzione condizionata della variabile risposta, in presenza di eterogeneità spaziale, outliers e quando la distribuzione è asimmetrica. Inoltre, se si prendono in considerazione i valori limite individuati dalla legislazione in materia di protezione contro le esposizione di radon, modellare la media condizionata della variabile di risposta in funzione di predeterminati valori dei predittori porta a informazioni non esaurienti sul fenomeno in esame.

In letteratura sono stati proposti alcuni approcci robusti, che consentono di ottenere una maggiore comprensione di come la concentrazione di radon indoor sia influenzata da determinate covariate. Tra questi vi è il modello di regressione quantilica utilizzato nel lavoro di Borgoni (2011) e di Fontanella et al. (2015). In entrambi i lavori viene adottato un modello di regressione quantilica che tiene anche conto della dipendenza spaziale insita nei dati. Infatti, nel primo lavoro viene proposto un modello quantile autoregressivo spaziale mentre nel secondo una regressione quantilica spaziale in un framework di un modello gerarchico bayesiano, nel quale la matrice delle covariate indipendenti è definita da fattori spaziali latenti e la struttura spaziale è incorporata nel termine di errore del modello, attraverso la specificazione di un processo spaziale di Laplace asimmetrico. Recentemente Sarra et al. (2016) hanno proposto un modello di regressione spaziale bayesiano in cui sono specificati dei parametri che controllano la dipendenza spaziale nei dati.

I dati raccolti durante le campagne di monitoraggio, oltre ad essere georeferiti, presentano una struttura gerarchica naturale che non può essere ignorata. Non tener conto della struttura dei dati aumenta il rischio di ottenere risultati inferenziali imprecisi o non validi. I modelli multilivello (Goldstein, 2011; Snijders and Bosker, 1999) sono adatti per l'analisi di dati che presentano pattern complessi di variabilità.

I modelli multilivello sono stati applicati in vari studi per modellare la concentrazione di radon indoor (Price et al., 1996; Apte et al., 1999; Gelman, 2006; Bochicchio et al., 2013; Almasri et al., 2009). Si tratta di modelli che presentano due livelli di gerarchia con cui si è cercato di spiegare la variabilità spaziale della IRC e ottenere le mappe dei valori della concentrazione di radon indoor per identificare le aree ad alto rischio. In Borgoni et al. (2014) viene adottato un modello gerarchico a tre livelli che spiega come IRC dipenda da un insieme di variabili secondarie, che sono riferite a ciascun livello della gerarchia (rispettivamente piani, edificio e litologia) e fornisce una stima della quantità di variabilità non spiegata.

In questo lavoro di tesi viene proposto un approccio M-quantile e, in particolare, viene esteso il modello M-Quantile per includere la componente spaziale. La componente spaziale è modellata combinando una intercetta random che cattura l'effetto della litologia sull'IRC e un termine semiparametrico che coglie la regolarità residua nello spazio (Pratesi et al., 2009). Quest'ultimo termine è modellato mediante una spline bivariata (*thin plate spline*) delle coordinate geografiche dei siti di campionamento.

Seguendo quanto proposto da Rupert et al., 2003 (vedi capitolo 3, paragrafo sulla relazione tra modelli di regressione spline e modelli ad effeti misti), trattiamo i coefficienti dei nodi della spline bivariata come un effetto random. Per la stima dei parametri del modello è stato utilizzato l'approccio di

massima verosimiglianza robusta (Richardson and Welsh, 1995) e l'algoritmo a due stadi (Tzavidis et al., 2015).

Il modello semi-parametrico M-Quantile ad effetti random è stato applicato ai dati di concentrazione di Radon indoor in Lombardia ottenuti durante la prima campagna di misurazione del 2003-04.

L'implementazione dei modelli riportati in questa sezione è stata attuata con il software statistico open source `R` (`R` core team, 2015).

6.2 Il radon indoor in Lombardia

Il radon (Rn) è un gas nobile, chimicamente inerte, che può diffondere liberamente senza cambiare le proprie caratteristiche fisiche (vedi Tabella 5). Proviene dal decadimento del radioisotopo Ra-226 originato, per decadimenti successivi, dal 'capostipite' U-238 diffusamente presente nella crosta terrestre in concentrazione variabile in funzione della particolare conformazione geologica. Il radon ha un tempo di emivita di circa 3.8 giorni, che può essere considerato breve. Essendo un gas nobile non interagisce con gli altri elementi presenti nel suolo riuscendo a raggiungere l'interfase suolo/aria agevolmente da cui viene esalato dalla crosta terrestre in atmosfera. Rimane prevalentemente intrappolato nella matrice solida nella quale avviene il decadimento del Ra-226 e solo una piccola frazione, quella emessa dal Ra-226 posto alla periferia dei singoli elementi solidi (superfici e zone di fratture delle rocce, grani di terreno o di sabbia ecc), emerge dal suolo o si discioglie nelle acque.

Tabella 5 Caratteristiche fisiche del Radon (da ARPA)

Numero atomico	86
Peso atomico	222
Densità(*g/l*)	9.73
Pressione critica (*atm*)	62
Temperatura critica (^{0}C)	105
Punto di ebollizione a 1 *atm*	-62
Coefficiente di dispersione in aria (cm^2/s)	0.1

Il gas Radon, emergente dal suolo o portato in superficie dalle acque, diffonde rapidamente nell'atmosfera, producendo concentrazioni molto basse negli ambienti outdoor.

Diversa è la situazione degli ambienti indoor (edifici) o sotterranei (grotte, caverne, ...) penetrati dal gas Radon e dai quali difficilmente può liberarsi per diffondere in atmosfera, in questo caso si possono avere concentrazioni anche molto elevate. In particolare, concentrazioni elevate si hanno nei locali

interrati degli edifici che, se da una parte costituiscono la più diretta via di penetrazione del Radon emergente dal suolo, dall'altra sono generalmente anche i locali meno aerati. Anche i materiali usati nelle costruzioni, che contengono percentuali variabili del 'genitore' Ra-226 possono contribuire in modo significativo alla concentrazione di attività del Radon negli edifici.
I fattori dai cui dipende la concentrazione di Radon indoor sono: la temperatura, la pressione, il riscaldamento artificiale (per effetto camino ovvero la differenza di temperatura tra interno ed esterno e il vento, provocano una differenza di pressione tra l'interno dell'edificio (pressione inferiore) e l'ambiente circostante, in particolare il terreno sottostante, che induce un risucchio di aria più fredda contente Radon) e la diversa modalità di uso degli infissi esterni e/o impianti di aerazione, dal vento (la concentrazione diminuisce in ambienti ben areati), contatto con suolo, materiale utilizzato per costruire la casa.

6.3 Dataset e statistiche descrittive della concentrazione di radon indoor in Lombardia

I dati utilizzati in questa analisi sono stati raccolti da ARPA (Agenzia Regionale Protezione Ambiente) durante la prima campagna di misurazione del radon indoor in Lombardia nel 2003-05. Come emerge dalla prima indagine nazionale condotta dal 1989 al 1994 (Bochicchio et al., 2005) questa regione (insieme a Lazio e Campania) è esposta ad un'alta concentrazione di radon indoor, registrando un valore medio di 116 Bq/m^3 contro i 70 Bq/m^3 della media nazionale. La differenza tra le medie delle regioni è da mettere in relazione alla naturale variabilità spaziale del fenomeno, dovuta principalmente al diverso contenuto di uranio nelle rocce e nei suoli e alla loro differente permeabilità. Un basso livello medio non esclude l'esistenza di aree limitate ad alta concentrazione di radon.
A questo dato si deve aggiungere che la Lombardia è una delle regioni italiane più popolose d'Italia pertanto la valutazione dell'esposizione della popolazione residente al Radon indoor e come l'IRC varia nello spazio è un problema, sia di carattere ambientale che di protezione della salute umana, estremamente rilevante in questa regione.
In questa tesi per l'analisi delle problematiche connesse alle concentrazioni di radon indoor negli edifici è stato considerato un campione di 900 punti di misura di IRC raccolti sul tutto il territorio regionale per i quali erano disponibili tutte le variabili costruttive degli edifici.
A tale scopo è stato utilizzato un rilevatore a tracce CR-39 collocato in sito per un anno. Il rilevatore è stato cambiato dopo sei mesi circa, di conseguenza sono state ottenute due misure semestrali. In fine l'IRC che consideriamo in questo lavoro deriva dalla media annuale tra le misure ottenute nei due semestri pesata per il tempo di esposizione del dosimetro.

Contestualmente alla misurazione della IRC sono state rilevate mediante un questionario le caratteristiche delle abitazioni in cui sono state effettuate le misure quali: materiale da costruzione delle pareti, tipo di connessione con il suolo, anno di costruzione e ristrutturazione dell'edificio, tipologia di edificio, dotazione di impianti di condizionamento e aerazione misurata in ore medie giornaliere, materiale della pavimentazione. Queste variabili sono state riconosciute in più studi come rilevanti nell'influenzare la IRC (Price et al.1996, Levesque et al.,1997, Brauner et al., 2013). Le variabili sono state ricodificate come binarie o dicotomiche; alcune presentano già una natura binaria (ad es. edificio in contatto con il suolo vs non in contatto con il suolo, edificio singolo vs non singolo) mentre altre sono state dicotomizzate (aerazione e anno di costruzione/ristrutturazione). Le misure di IRC e le variabili costruttive dell'edificio sono state combinate per ottenere un unico dataset.

Per 900 punti di misura rilevati a piano terra e dislocati su tutto il territorio lombardo (Figura 7a), riportiamo la media annuale di IRC e gli altri quantili (Tabella 2). In figura 7 b sono rappresentate le locazioni spaziali delle misure di IRC che superano il valore di attenzione 300 Bq/m^3 secondo la recente direttiva europea EUROTOM 2013/59. Queste sono principalmente localizzate nella provincie di Bergamo. Nel grafico in Figura 8 sono riportati i boxplot dei valori di IRC per ciascuna provincia. In ogni provincia la distribuzione di IRC è asimmetrica come si evince dal discostamento dei punti rossi che indicano la media dalla mediana, inoltre le province di Sondrio, Bergamo, Brescia, Lecco, Monza-Brianza e Varese presentano dei valori di IRC elevati e dispersi intorno alla mediana. Si riscontrano molti valori di IRC anomali, ad esempio a Sondrio sono stati misurati valori anomali che vanno da circa 500 a 1700 Bq/mq^3.

Figura 7 Locazione spaziale dei punti di misura (a) e locazione spaziale dei punti di misura il cui valore supera i 300 Bq/m^3 (b)

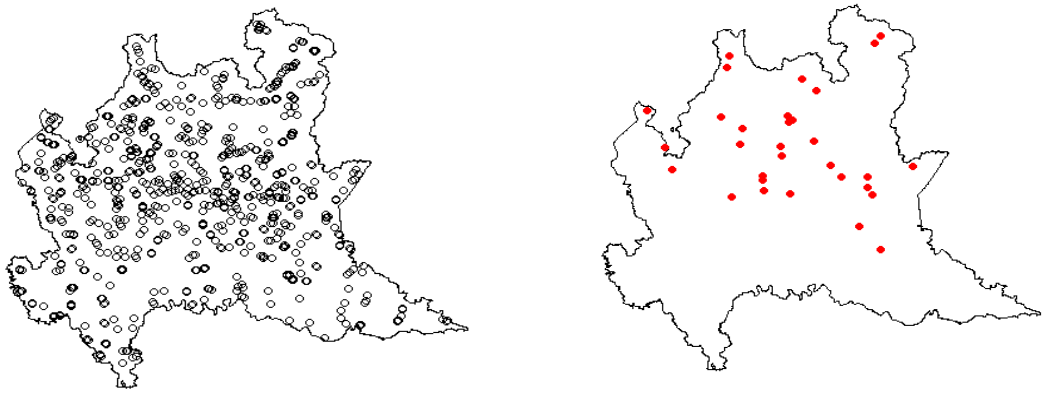
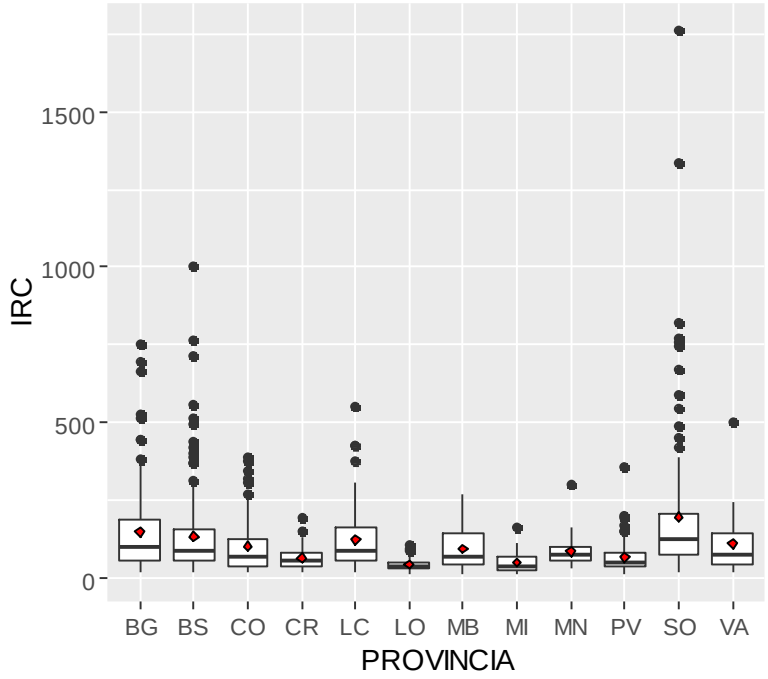


Figura 8 Boxplot della concentrazione di Radon indoor (IRC) in ciascuna delle province lombarde



In tabella 6 sono presentate alcune statistiche di sintesi della IRC media annuale: il valore medio registrato per il campione di osservazioni N=900 è di 117.92 Bq/m^3, il massimo valore è di 1762.5 Bq/m^3 mentre il minimo 12.5 Bq/m^3. 59 sono i valori superano o eguagliano il valore di riferimento legislativo di 300 Bq/m^3. In percentuale il 13.88% della IRC è maggiore dei 200 Bq/m^3 mentre il 6.5% delle misure supera i 300 Bq/m^3.

Tabella 6 Statistiche descrittive per la variabile di risposta IRC

Min	1°quartile	Mediana	Media	3°quartile	Max	Numerosità	sd	IRC≥200 Bq/m3	IRC≥300 Bq/m^3
12.5	46.06	75.98	117.92	140.15	1762.5	900	136.2	125 (13.88%)	59 (6.5%)

In tabella 7 sono riportate le statistiche descrittive condizionatamente a ciascuna delle caratteristiche abitative che consideriamo in questo studio. Si evince che differenti variabili costruttive agiscono in modo differente sui diversi livelli di IRC e anche che gli effetti variano al cambiare dei fattori. Per esempio, considerando la variabile pavimentazione le differenze tra i 20-esimi e l'80-esimi percentili delle IRC misurate in abitazioni con pavimenti in marmo/granito rispetto a quelle dotate di pavimento in altro materiale è 4.9 Bq/m^3 (20-simo percentile) e 14.2 Bq/m^3 (80-simo percentile). Per quanto riguarda la variabile pareti, l'IRC delle case con pareti in pietra rispetto a quelle con materiale diverso le differenze tra i percentili 20 e 80 è di 10.9 Bq/m^3 (20-simo percentile) e di 61.3 Bq/m^3 (80-simo

percentile). Andando a considerare se la casa è dotata di sistema di areazione o meno, la differenza nelle misure di IRC è 1.3 Bq/m^3 (20-simo percentile) e 69 Bq/m^3 (80-simo percentile). Confrontando la IRC misurate in abitazioni edificate in contatto con il suolo e quelle separate da esso, sempre per medesimi percentili, abbiamo una differenza di 5.9 Bq/m^3 (20-simo percentile) e 49.8 Bq/m^3 (80-simo percentile). In fine, confrontando la IRC misurate in abitazioni singole con quelle in connessione la differenza è di 10.6 Bq/m^3 (20-simo percentile) e 57.8 Bq/m^3 (80-simo percentile).
E' evidente che le differenze tra le misure di IRC sono più accentuate per i percentili più elevati.

Tabella 7 Statistiche descrittive per la variabile di risposta IRC per ogni caratteristica costruttiva

Caratteristiche abitazione	N°siti	Min	20%	Media	Mediana	80%	Max	Sd
Pavimentazione:								
Marmo/Granito =1	47	19.9	37.0	100.8	63.3	147.0	742.9	114.6
Altro =0	853	12.5	41.9	118.9	76.6	161.2	1762.5	137.3
Pareti:								
Pietra =1	112	16.0	51.2	171.1	93.8	212.0	1762.5	233.2
Altro =0	788	12.5	40.3	110.4	73.2	150.7	1003.9	114.3
Aria condizionata								
Si =1	60	19.9	40.6	85.4	65.7	95.7	555.2	84.5
No =0	840	12.5	41.9	120.2	76.7	164.7	1762.5	138.9
Anno di edificazione								
Prima anni '90=0	525	12.5	41.0	117.0	74.4	161.4	1336.2	134.6
Dopo anni '90=1	375	16.0	42.6	119.2	79.0	160.9	1762.5	138.6
Connessione suolo								
Contatto =0	352	13.8	45.0	135.9	84.6	195.3	1003.9	143.0
Non contatto=1	548	12.5	39.1	106.4	70.9	145.5	1762.5	130.5
Tipo di edificio								
Non Singolo =0	323	13.8	35.4	95.4	64.7	130.0	1762.5	130.3
Singolo =1	577	12.5	46.0	130.5	84.2	187.8	1336.2	137.9

Ancora, in base a quanto riportato in tabella 7 possiamo riassumere che:
Gli edifici con pareti in pietra presentano concentrazioni mediamente (171.1Bq/m^3) superiori rispetto a locali con pareti di altro materiale (110.4 Bq/m^3). Le costruzioni con pavimenti in marmo /granito hanno una concentrazione di radon indoor in media inferiore (100.8 Bq/m^3) rispetto a costruzioni con pavimenti in altro materiale (118.9 Bq/m^3).
Gli edifici privi di impianto di aerazione presentano in media una concentrazione di radon indoor superiore (120.2 Bq/m^3) rispetto a quelle dotate (85.4Bq/m^3).

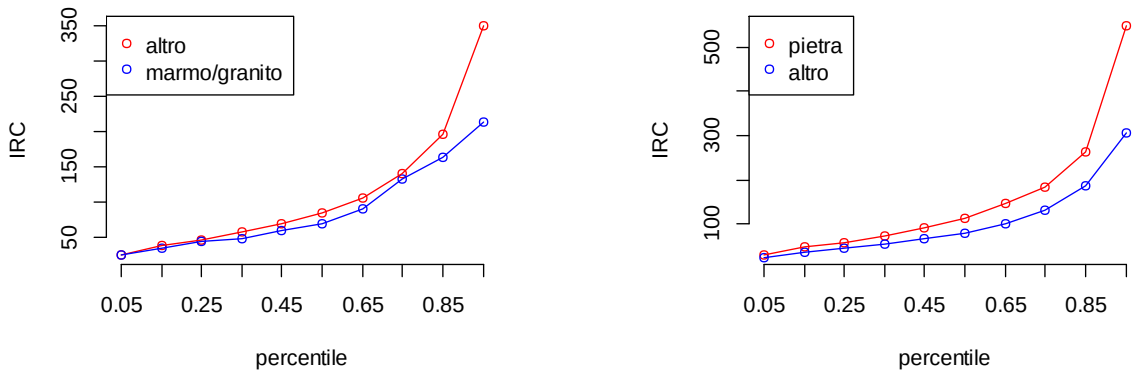
Le costruzioni edificate prima degli anni novanta hanno una concentrazione di radon indoor in media in valore inferiore (117.0 Bq/m^3) a quelle costruite/ristrutturate successivamente a tale anno (119.2 Bq/m^3)

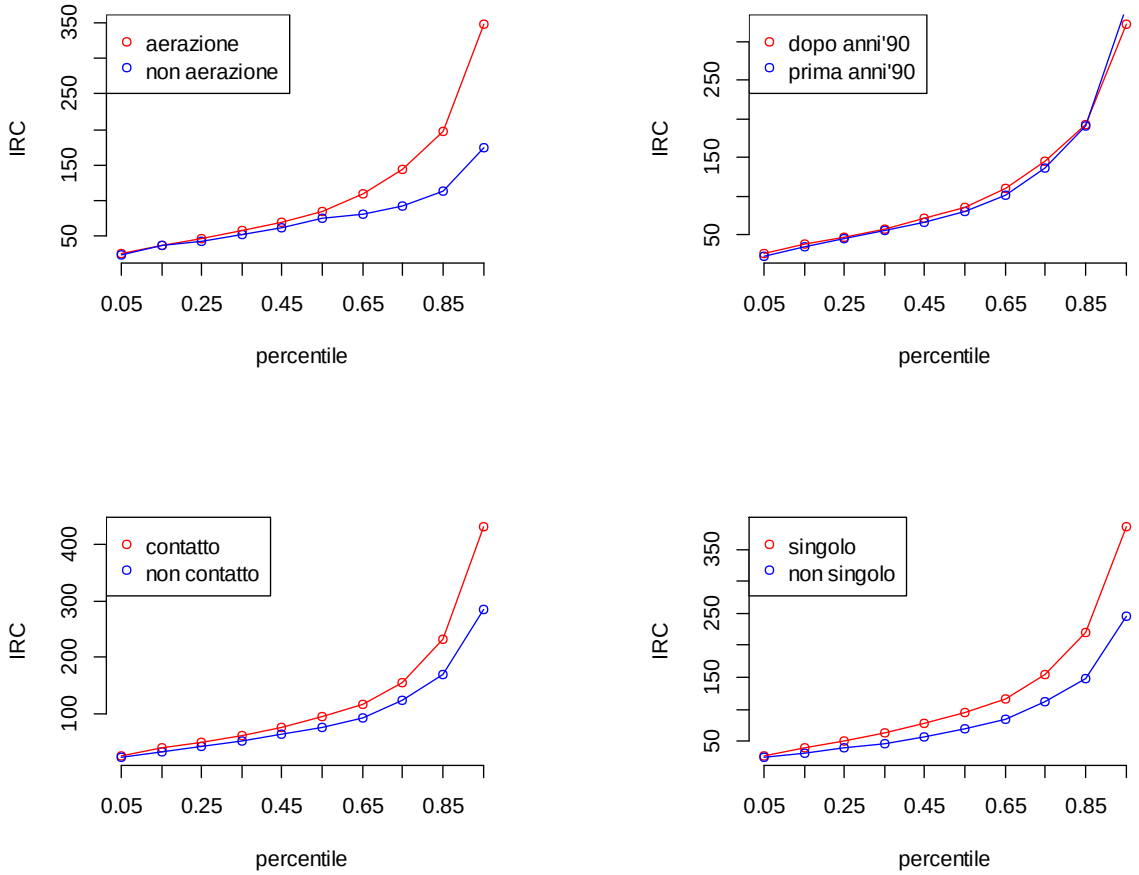
Le abitazioni isolate dal terreno (presenza di un vespaio o di un sotterraneo) presentano in media una concentrazione di radon indoor inferiore (106.4 Bq/m^3) rispetto ai locali a diretto contatto (130.5Bq/m^3). Gli edifici isolati nel loro immediato intorno (singoli) presentano concentrazioni di radon indoor mediamente superiori (132.4 Bq/m^3) rispetto agli edifici non singoli, come condomini o ville a schiera (95.4 Bq/m^3).

In figura 9 la IRC in funzione delle variabili antropiche/edili considerate in ogni quantile $0 \leq q \leq 1$. Per le costruzioni con pavimenti in marmo/granito e altro materiale si inizia a delineare una certa differenza nella concentrazione di radon indoor oltre il percentile 0.75, arrivando ad una netta separazione delle due linee di grafico al percentile più estremo. Per quanto riguarda le due classi in cui è diviso l'anno di costruzione per tutti i percentili considerati i due grafici rimangono accostati e non si manifesta mai una netta separazione.

Per tutte le restanti variabili la differenze si manifestano già ai percentili più bassi ma con un diverso grado di separazione tra le due linee.

Figura 9 Grafico della IRC per caratteristiche abitative e materiali da costruzione





Come già menzionato le caratteristiche di porosità e di permeabilità del suolo e delle rocce sottostanti così come la presenza di formazioni rocciose molto fratturate o con faglie, sono fattori che influenzano il rilascio, la diffusione e il trasporto del Radon nell'ambiente. A questo si deve aggiungere il fatto che differenti tipologie di roccia presentano una concentrazione di progenitori del radon diverso (troviamo un maggior contenuto nei graniti, nei tufi e in rocce metamorfiche acide e vulcaniche in genere).

Data l'importanza del fattore litologico nell'influenzare la variabilità di IRC, questo è stato inserito nel nostro modello come fattore random di gruppo, in quanto ci si aspetta che abitazioni edificate sulla stessa tipologia litologica abbiano valori simili di IRC.

La regione Lombardia ha una composizione geologica molto eterogenea poiché presenta molti litotipi, per motivi legati all'analisi sono state individuate e selezionate 11 classi.

Per poter ottenere l'informazione i dati sono stati associati alla mappa geo-litologica (scala 1:250000) che partizionano il territorio della regione Lombardia nelle 11 classi litologiche.

Per maggior dettagli e approfondimenti su come è stata ottenuta la mappa litologica con le 11 classi in figura 10 si veda Borgoni et el., 2010, 2011).

Figura 10 Classificazione delle 11 classi litologiche (tratta da Borgoni et al., 2010)

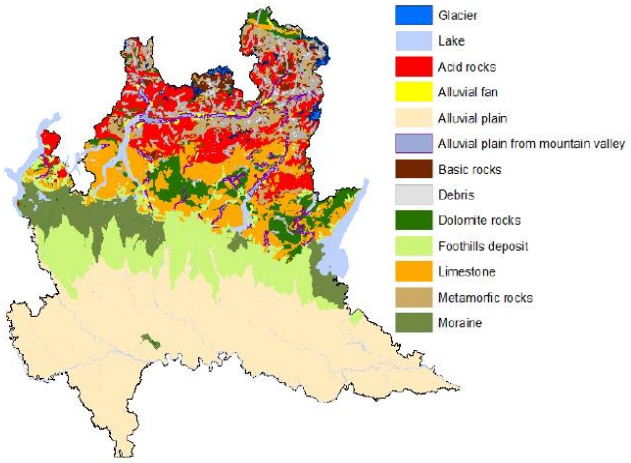


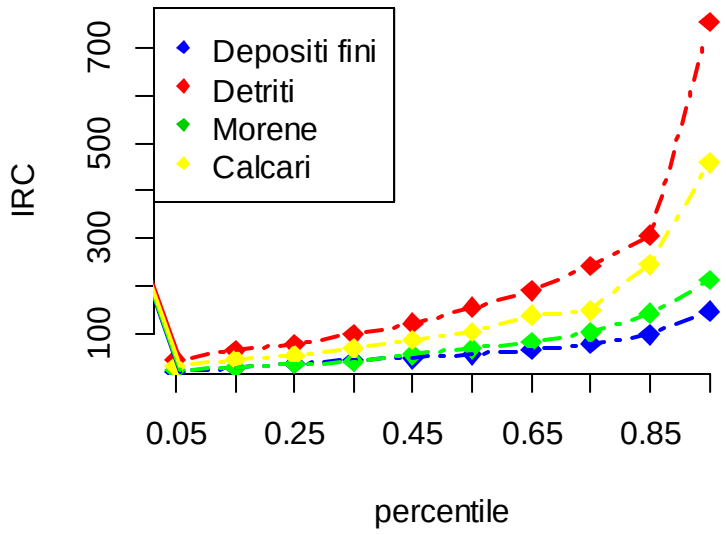
Tabella 8 Valori di IRC per le 11 classi litologiche

Classe litologica	Numerosità delle classi	Concentrazione media IRC Bq/m^3	IRC ≥ 300 Bq/m^3
1 Depositi fini	263	64.65	2
2 Alta Pianura	121	100.14	4
3 Calcari	94	143.08	10
4 Conoidi	28	143.79	3
5 Detriti	78	231.12	13
6 Dolomie del Triassico	68	167.61	11
7 Magmatiche Metamorfiche acide	51	147.44	6
8 Magmatiche Metamorfiche Basiche	10	104.3	1
9 Metamorfiche	56	136.54	4
10 Alluvionale di Montagna	43	153.16	3
11 Morene	88	83.10	2

In tabella 8 è riportata la numerosità delle 11 classi litologiche, la concentrazione media di radon indoor in ciascuna di esse e il numero di valori che eguagliano e superano il valore di 300 Bq/m^3.
Le IRC medie maggiori sono presenti nelle abitazioni localizzate sulle classi litologiche: detriti (231.12 Bq/m^3), dolomie (167.61 Bq/m^3), alluvionali di montagna (153.16 Bq/m^3), Magmatiche Metamorfiche acide (159.67 Bq/m3), Conoidi (143.79 Bq/m^3), Calcari (143.08 Bq/m^3), metamorfiche (136.54 Bq/m^3), Magmatiche Metamorfiche Basiche (104.3 Bq/m^3), Alta pianura (100.14 Bq/m^3).
La IRC minore è stata misurata negli edifici ubicate sui depositi fini (64.65 Bq/m^3).
Per quanto riguarda i valori maggiori o uguali al valore di riferimento limite 300 Bq/m^3, la classe litologica che ha registrato il maggior numero è la classe dei detriti (13 valori), dolomie del triassico (11 valori), i calcari (10 valori) e magmatiche metamorfiche acide (6 valori). Un solo valore di IRC ≥ 300 Bq/m^3 e misurato sulla classe delle rocce Magmatiche Metamorfiche Basiche.

In figura 11 la IRC in funzione delle classi litologiche che presentano in media un valore di IRC elevato (Detriti e Calcari) e IRC basso (Depositi fini e Morene) in ogni quantile $0 \leq q \leq 1$.

Figura 11 IRC per le classi litologiche Depositi Fini, Detriti, Morene, Calcari



Per le classi litologiche Detriti e Calcari si inizia a delineare una certa differenza nella concentrazione di radon indoor oltre il percentile 0.85, arrivando ad una netta separazione delle due linee di grafico al percentile più estremo. Per quanto riguarda le classi Morene e Depositi fini per tutti i percentili investigati i due grafici rimangono accostati e non si manifesta mai una netta separazione.

La maggior distanza si ha tra i Detriti e i Depositi fini, che diventa molto evidente già dai percentili più bassi suggerendo che esiste un differente effetto della litologia detritica e dei depositi fini sui diversi M-quantili della distribuzione di IRC.

Le caratteristiche geologiche possono agire a piccola scala principalmente attraverso le strutture tettoniche, alte concentrazioni di IRC sono state misurate vicino faglie, fratture o forme carsiche, le quali agevolano il flusso verso la superficie del gas radon. Per considerare la componente geologica locale sono stati, dapprima, sovrapposti i siti di campionamento alla mappa dei lineamenti tettonici della regione Lombardia (Figura 12) e, successivamente, è stata calcolata la distanza di ciascuna locazione *s* del sito di campionamento al lineamento tettonico più vicino (per maggiori dettagli vedi Borgoni et al., 2010. Indicando con A la linea tettonica, la distanza è ottenuta attraverso

$$d(s, A) = \inf_{x \in A} \|s - x\|$$

La distanza dalla faglia andrà a costituire una variabile che in seguito sarà utilizzata sia come covariata che entrerà nella parte deterministica che nella costruzione della matrice spline univariata dei modelli implementati.

Figura 12 Lineamenti di faglia della Regione Lombardia

6.4 Modelli M-Quantile per IRC

In questa sezione sono riportati i risultati ottenuti implementando il modello semiparametrico M-Quantile ad effetti casuali (si veda il capitolo 4 per la trattazione metodologica) sui dati di IRC. Le covariate che entrano nella parte fissa del modello generale sono relative alle caratteristiche costruttive dell'edificio. Come delineato in precedenza, queste variabili sono riconosciute essere quelle che influenzano maggiormente l'accumulo di concentrazione di radon all'interno dell'edificio. La concentrazione di radon indoor tende a variare nello spazio mostrando un pattern di valori regolare in funzione di un dato numero di fattori geo-ambientali e antropici. Le coordinate geografiche dei punti di misura possono essere considerate delle variabili proxy di tutti quei fattori che risultano non misurate o che è impossibile misurare. In questo lavoro includiamo questa componente in tre modi.

Siccome nella regione in esame IRC mostra una chiara tendenza che va da Sud a Nord (Borgoni et al., 2011), un *trend surface model* (Cade et al. (2005), Koenker e Mizera (2004)) è stato specificato semiparametricamente mediante una *thin plate spline* bivariata delle coordinate geografiche. In secondo luogo, ci si aspetta che la concentrazione di Radon indoor misurata nelle abitazioni che sono costruite sulla stessa tipologia di suolo o roccia siano più simili rispetto a quelle ubicate su classi litologiche diverse. Di conseguenza i dati mostrano una struttura gerarchica e perciò è utile esplicitare anche questo aspetto nella specificazione del modello. Per tener conto della naturale struttura gerarchica dei dati è inclusa nel modello un effetto random che catturi la variabilità caratterizza le aree geologiche.

Inoltre per tener conto dell'influenza dell'elementi tettonici, la distanza dal punto di misura dalla faglia ad esso più prossima, è stata inserita nel modello come predittore lineare poiché l'analisi esplorativa suggerisce che una relazione di questo tipo è ragionevole.

In tutti i modelli che saranno presentati e discussi il termine di spline è specificato tramite un effetto casuale (Ruppert. et al., 2003).

La stima dei parametri è ottenuta applicando il metodo a due stadi sequenziali già descritto da Tzavidis et al. (2016) (vedi capitolo 2), per dati longitudinali e gerarchici. Tale metodo prevede come primo step la stima dei coefficienti fissi del modello mediante l'algoritmo di Newton Raphson e successivamente la stima delle componenti di varianza mediante metodo del punto fisso. Gli errori standard sono ottenuti mediante le formule riportate nel capitolo 4.

6.5 Analisi preliminari

Prima di presentare i risultati ottenuti applicando il modello semiparametrico M-Quantile ad effetti casuali riportiamo quanto ottenuto dalle analisi preliminari di tipo diagnostico.

La distribuzione della concentrazione di radon indoor è fortemente asimmetrica e discostante dalla distribuzione normale (Figura 13). Questa è una situazione emersa e riconosciuta in tutti gli studi che hanno interessato la IRC (Nero et al., 1984; Borgoni et al., 2010, 2011, 2014)). La figura 14 a, b riporta invece un qq-plot per i residui di primo e secondo livello ottenuti fittando un modello misto semiparametrico che include quindi il termine di spline bivariata, i cui quantili sono confrontati con quelli di una distribuzione normale. Per l'analisi è stata utilizzata la funzione `lme` del pacchetto `nlme` di `R`. E' evidente la presenza di numerosi outlier collocati nella coda di destra.

Figura 13 Istogramma della IRC (y)

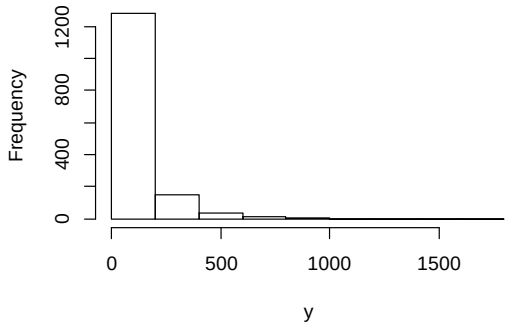
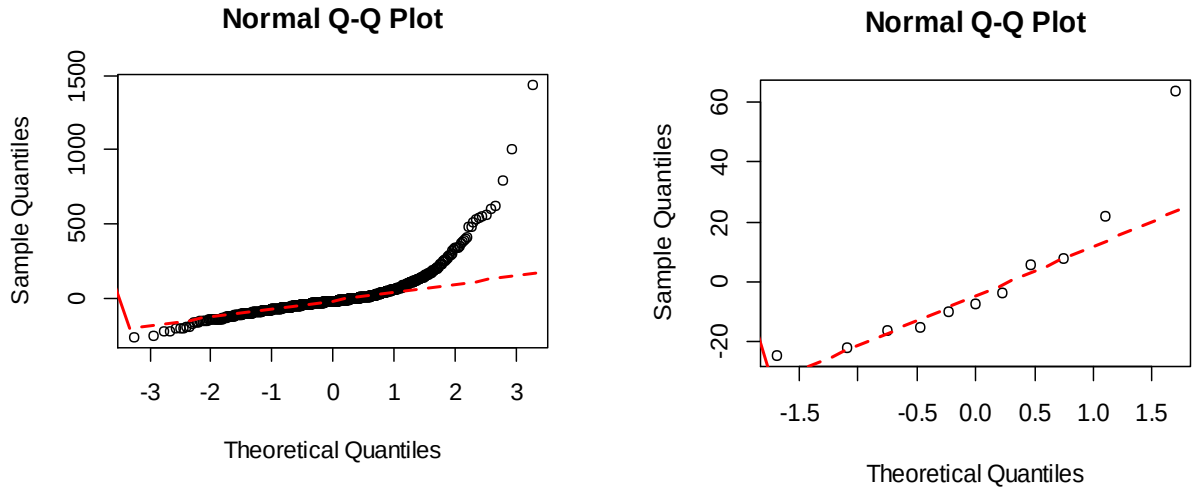


Figura 14 a e b) Normal Q-Q Plot dei residui di primo e secondo livello

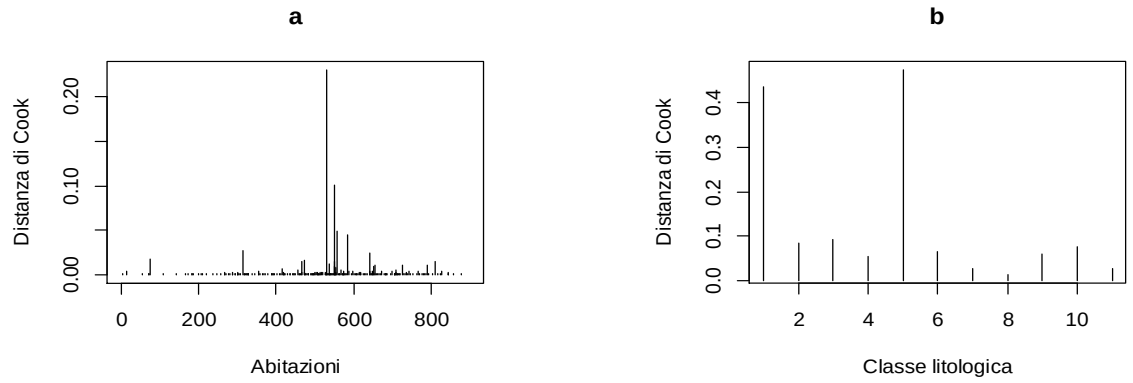
a **b**



A questo punto, appare opportuna una analisi più dettagliata per verificare che il modello non sia sensibile a delle particolari osservazioni. Per fare ciò utilizziamo la distanza di Cooks per modelli misti ottenuta con la funzione `cooks.distance` del pacchetto `HLMdiag` (Loy A. e Hofmann H., 2014) di `R`.

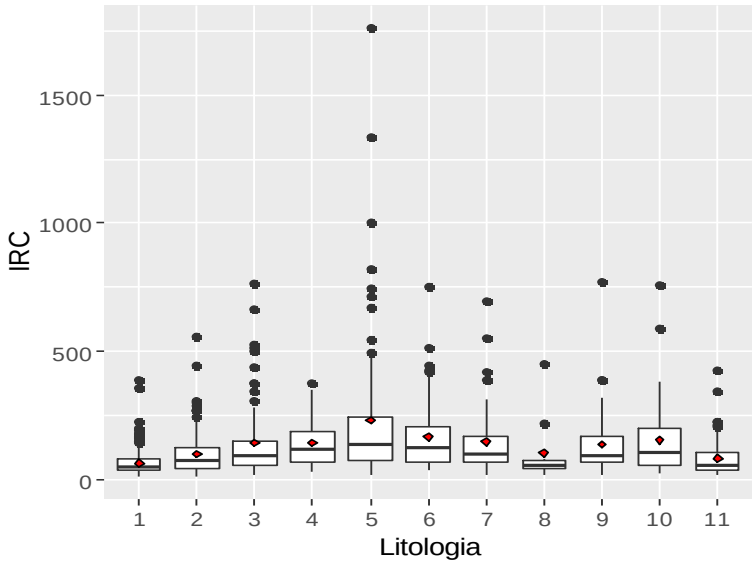
In figura 15 (a) e (b) sono riportati i grafici della Distanza di Cook rispettivamente per le osservazioni (15 a) (abitazioni) e per le classi litologiche (15 b). Entrambi suggeriscono la presenza di alcune osservazioni che rendono sensibile il modello e giustificano l'adozione di un modello robusto quale il modello semiparametrico M-Quantile ad effetti casuali.

Figura 15 Distanza di Cook per le osservazioni a e per le classi litologiche b



Come si evince dalla figura 16 la presenza di outliers nelle misure di radon indoor si manifesta in tutte le classi litologiche, in particolare nelle rocce detritiche si sono misurate IRC che raggiungono e addirittura superano di 1000 Bq/m^3.

Figura 16 Boxplot di IRC per classe litologica



6.6 Effetto della distanza dalla faglia: modello con spline univariata

L'obiettivo di questa analisi è considerare l'influenza delle caratteristiche geologico-strutturali del territorio sulla concentrazione di radon indoor. La variabile, distanza dalla faglia è stata inserita nel modello tramite una trasformazione spline univariata come descritta nel capitolo 4 ed è stata

preliminarmente normalizzata tra 0 e 1 per avere una maggiore stabilità numerica nell'algoritmo di stima.

Il modello utilizzato è specificato in forma matriciale come segue:

$$MQ_q(y|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta_q} + \mathbf{L}_{sp}\boldsymbol{s_q}$$

dove

$\boldsymbol{y}_{(900\times1)}$ è il vettore della variabile di risposta concentrazione di radon indoor;

$\mathbf{X}_{(900\times2)}$ matrice delle covariate (distanza dalla faglia) la cui prima colonna è un vettore di elementi unità che consente di includere il termine d'intercetta

$\boldsymbol{\beta_q}_{(\mathbf{1}\times\mathbf{2})}$ è il vettore dei parametri fissi del modello .

$\mathbf{L}_{\boldsymbol{sp}_{(900\times20)}}$ è la matrice della spline univariata costruita utilizzando la covariata distanza dalla faglia.

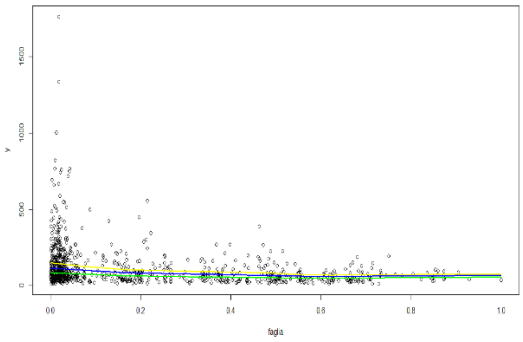
$\boldsymbol{s_q}_{(\mathbf{20}\times\mathbf{1})}$ é il vettore degli effetti random del modello

La matrice spline univariata $\boldsymbol{L_{sp}}$è ottenuta utilizzando una funzione spline lineare troncata della variabile distanza faglia. Per la costruzione della spline sono stati utilizzati K=20 nodi di cui 19 posizionati ai quantili della distanza e il primo nodo è stato fissato intorno a 1200 m (0.062), distanza a cui si nota una variazione di andamento (Borgoni et al., 2010). La stima della componente di varianza legata alla spline univariata $(\hat{\sigma}^2_{fq})$ è di 11524.51per la mediana, di 1436.46 per il 1° quantile e di 49365.28 per il 3°quantile. La stima del coefficiente fisso (faglia) ha valori che aumentano in valore assoluto andando dal 1°quantile al 3°quantile. Coerentemente con quanto ci si aspetta, i valori del termine di intercetta per ciascun quantile mostrano un andamento crescente. La stima dei parametri del modello con solo termine di spline univariata sono riportati in Tabella 9.

Tabella 9 Stima dei parametri del modello semiparametrico M-quantile con spline univariata per gli M-quantili 0.25,0.50 e 0.75

	M-Quantile=0.25		**M-Quantile=0. 50**		**M-Quantile=0.75**	
Parametri	**Stima**	**Errore standard**	**Stima**	**Errore standard**	**Stima**	**Errore standard**
$\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$ ***(intercetta)***	76.52	2.24	113.88	3.58	148.15	5.13
$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ ***(faglia)***	-84.00	51.65	-161.71	106.60	-284.88	205.34
$\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\boldsymbol{e}}$ ***(residuo)***	999.24	59.58	3778.06	288.04	5758.24	429.52
$\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\boldsymbol{f}}$ ***(spline)***	1436.46	1193.85	11524.51	8910.52	49365.28	22950.95

Figura 17 Grafico del IRC (y) vs distanza dalla faglia: linea verde per l'M-quantile 0.25; linea blu per l'M-quantile 0.50 e linea gialla per l' M-quantile 0.75



Come si evince dalla figura 17, per tutti i quantili investigati, se si esclude un modesto cambiamento nell'andamento della retta nelle immediate vicinanze del primo nodo (posizionato a 0.062), in generale l'effetto della distanza dalla faglia sull'IRC è lineare.
Data la linearità mostrata la variabile distanza dalla faglia viene inserita nella parte deterministica dei modelli presentati alla fine di questa sezione in modo lineare.

6.7 Analisi dei trend spaziali

Come abbiamo affermato più volte in questo capitolo, l'esalazione di radon dal suolo dipende, non solo dal tipo di roccia, ma anche da altri fattori, ad esempio la fratturazione della roccia che favorisce la migrazione dal suolo. Per tanto la distanza di un punto di misura dalla faglia più vicina viene inserita nel modello.
Per l'analisi del trend spaziale è stato costruito ed implementato un modello in cui i coefficienti della spline bivariata sono considerati un effetto random γ

$$MQ_q(y|\mathbf{x}) = \mathbf{X}\boldsymbol{\beta_q} + \mathbf{Z}\boldsymbol{u_q} + \mathbf{Z}_{sp}\boldsymbol{\gamma_q}$$

Le coordinate cartografiche in proiezione Gauss-Boaga sono state normalizzate prima di essere utilizzate nella costruzione della matrice spline $\boldsymbol{Z_{sp}}$ (vedi capitolo 4 Metodi). Mediante l'algoritmo clara (pacchetto `cluster` di `R` (Maechler et al., 2016)) sono stati individuati e localizzati sul territorio lombardo K=50 nodi. La costante di tuning è stata fissata a c =1.345. Inoltre, i valori iniziali necessari ad inizializzare la stima dei parametri del modello sono stati ottenuti dal modello semiparametrico misto con spline.
$\boldsymbol{y}_{(900\times1)}$ è il vettore della variabile di risposta della concentrazione di radon indoor;

$\mathbf{X}_{(900 \times 3)}$ matrice delle covariate (longitudine, latitudine e distanza dalla faglia) compresa il vettore $1_{(900 \times 1)}$ per includere il termine d'intercetta;

$\mathbf{Z}_{sp_{(900 \times 50)}}$ è la matrice della spline radiale bivariata (*thin plate spline)* costruita utilizzando le coordinate geografiche $\mathbf{x_i} = (x_{1i}, x_{i2i})$ dei siti di rilevazione presenti sul territorio.

$\mathbf{Z}_{(900 \times 11)}$ è la matrice del disegno

$\boldsymbol{u_{q(11 \times 1)}}$ vettore degli effetti random di gruppo

$\boldsymbol{\gamma_{q(50 \times 1)}}$ vettore degli effetti random dei coefficienti della spline bivariata

In Tabella 10 sono riportate le stime delle componenti di varianza rispettivamente del termine random di spline ($\hat{\sigma}^2_{gq}$), gruppo ($\hat{\sigma}^2_{uq}$), della componente erratica ($\hat{\sigma}^2_{eq}$) e della e le stime dei coefficienti fissi del modello privo di covariate specifiche di edificio ma relative alle caratteristiche spaziali valutate agli M-quantili 0.25, 0.50, 0.75.

I valori stimati delle tre componenti di varianza più alti sono associati rispettivamente al M-quantile 0.75. Per quanto attiene il valore stimato $\hat{\beta}_0$ si evidenzia un trend crescente andando dal quantile 0.25 al quantile 0.75.

Tabella 10 Stima dei parametri per gli M-quantili 0.25, 0.50 e 0.75

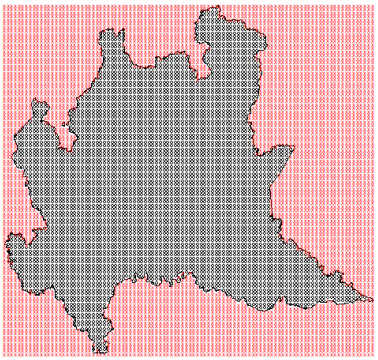
	M-Quantile=0.25		***M-Quantile=0.50***		***M-Quantile=0.75***	
Parametri	**Stima**	**Errore Standard**	**Stima**	**Errore Standard**	**Stima**	**Errore Standard**
$\widehat{\boldsymbol{\beta}}_{0\,(intercetta)}$	29.68	18.89	53.80	30.42	99.24	77.41
$\widehat{\boldsymbol{\beta}}_{1\,(faglia)}$	-9.36	8.31	-19.11	13.44	-51.58	27.58
$\widehat{\boldsymbol{\beta}}_{2\,(longitudine)}$	40.57	23.66	47.60	38.20	91.06	101.50
$\widehat{\boldsymbol{\beta}}_{3\,(latitudine)}$	31.43	26.77	39.10	43.13	31.54	110.12
$\widehat{\boldsymbol{\sigma}}^2_{e\,(residuo)}$	745.78	120.81	3213.71	178.43	6468.04	199.36
$\widehat{\boldsymbol{\sigma}}^2_{u\,(litologia)}$	48.30	44.41	194.02	238.23	308.89	429.64
$\widehat{\boldsymbol{\sigma}}^2_{g\,(spline)}$	211.83	33.57	905.70	122.54	3651.11	145.49

6.7.1 Mappe di IRC stimato

Una volta ottenute la stima dei parametri il modello semiparametrico M-Quantile ad effetti random è stato possibile costruire le mappe di IRC per tutti gli M-quantili considerati.

Per ottenere le mappe di IRC stimato, preliminarmente costruiamo un grigliato $[100 \times 100]$ di 10000 punti di predizione (di cui 4651 cadono interamente nel territorio regionale), da sovrapporre alla mappa della Lombardia (Figura 18).

Figura 18 I 1000 punti di predizione: i cerchi neri ricadono nel territorio regionale mentre i cerchi rossi si collocano al di fuori della Regione

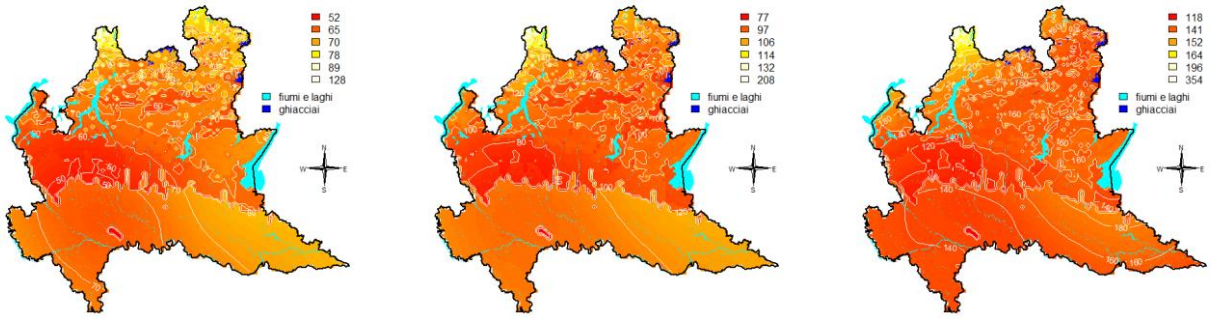


Ogni punto griglia, presente nell'area regionale, è stato associato ad una determinata classe litologica. Un'operazione di overlay tra lo shapefile delle 11 formazioni litologiche e le coordinate dei punti griglia, attuata mediante il comando `over` del pacchetto `sp` di `R` (Pebesma, 2012, 2016), ha permesso di individuare l'esatta associazione. La distanza dalla faglia è stata ricalcolata: è stata costruita una regione quadrata di dimensione pari a quelle del grigliato quindi, facendo uso dello shapefile dei lineamenti tettonici sono state sovrapposte le linee di faglia su tale regione ed è stata calcolata la distanza della faglia dal punto di misura interno alla griglia. Una nuova matrice delle spline $\mathbf{Z}_{sp}$ è stata costruita con le coordinate cartografiche dei punti griglia.
I valori predetti per gli M-quantili q = 0.25, 0.50, 0.75 sono stati ottenuti da:

$$MQ_q(y|\mathbf{x}) = \hat{\beta}_0 + \hat{\beta}_1\, dist.faglia_{\,ij} + \hat{\beta}_2 latitudine_i + \hat{\beta}_3 longitudine_i + \boldsymbol{litologia}_{\mathrm{ij}}^{\mathrm{T}}\hat{\boldsymbol{u}} + \boldsymbol{z_{sp.i}}\hat{\boldsymbol{\gamma}}$$

In Figura 19 sono riportate le mappe degli M-quantili di IRC così ottenuti. Per tutti gli M-quantili i valori più bassi di IRC sono localizzati nella regione sud-ovest (Provincia di Pavia, nella provincie di Mantova e Lodi). I valori più alti nella parte nord-est della regione in particolare Sondrio. In generale i risultati mostrano come nell'area di pianura, dove il substrato alluvionale, poco permeabile al gas, presenta uno spessore maggiore, la presenza di radon sia poco rilevante; nelle aree montane e pedemontane, in provincia di Sondrio, Varese, Bergamo, Brescia e Lecco, le concentrazioni sono risultate invece più elevate (Borgoni et al., 2010).

Figura 19 Mappe per gli M-quantili 0.25, 0.50 e 0.75



M-quantile 0.25 ***M-quantile 0.50*** ***M-quantile 0.75***

6.8 Identificazione delle Radon Prone Area (RPA)

La necessità di proteggere e tutelare la salute della popolazione esposta al Radon ha portato, in tutto il mondo, molte istituzioni governative ad emanare delle leggi che prevedono l'esecuzione di programmi di monitoraggio, al fine di individuare le aree geografiche dove è più probabile trovare concentrazioni di Radon indoor che eccedono il livello di riferimento (Gue, 2015). Queste aree ad alto rischio sono dette Radon Prone Area. Generalmente, all'interno di una nazione la RPA è una regione geografica in cui si stima che più di una data percentuale di case possieda una concentrazione di Radon Indoor che eccede il livello di riferimento ma in realtà, esistono molte definizioni delle RPA (World Health Organization (WHO), 2009). La legislazione italiana conformandosi a quanto stabilito dal WHO definisce la RPA come delle aree geografiche in cui c'è un'alta probabilità di trovare una elevata concentrazione di Radon (art 10-ter,comma 2, D. l.vo 241/2000).

Negli ultimi anni molte tecniche statistiche sono state utilizzate per identificare le RPA. Tra queste tecniche vanno menzionate innanzitutto quelle geostatistiche (Zhu et al., 2001, Chaouch et al., 2003, Dubois et al., 2007, Borgoni et al., 2010; 2013), i modelli gerarchici (Apte et al., 1999; Price et al., 1996), il modello spaziale multiprocesso in ambito bayesiano (Smith e Cowles, 2007), il modello di regressione lineare multivariata (Hodgson et al., 2014), il modello di regressione logistica (Elio et al., 2017) e i vari approcci basati sui metodi di cluster detection adottati per definire i cluster spaziali e per comprendere il fenomeno a livello geografico attraverso l'individuazione delle regioni dello spazio che sono anomale, inattese o interessanti (Sarra et al., 2016).

Borgoni et al. (2010) hanno anche suggerito un approccio basato sui quantili che accoppia la tecnica di kriging con la simulazioni sequenziali (gaussiane) Monte Carlo per approssimare, in ogni punto dello spazio, la distribuzione condizionata della IRC. Come visto sono molti i modelli statistici

impiegati per individuare le Radon Prone Areas, tuttavia ricorrere al modello M-quantile sembra il modo più diretto e naturale di operare per identificarle.

Supponiamo che il livello di riferimento stabilito dalla legge sia φ, allora, per la definizione appena data avremmo che la RPA è l'area dove la probabilità che la IRC superi il valore φ sia alta

$$P(IRC > \varphi) = alta$$

dove, per alta, si intende un valore ">q" $(0 \leq q \leq 1)$ con q prefissato. Questo implica che il quantile $\xi_{(1-q)}$ di ordine $(1-q)$ di IRC soddisfa $\xi_{(1-q)} > \varphi$. Quindi, con riferimento ai quantili della distribuzione dell'IRC, la RPA si può definire come quell'area dove un quantile $\xi_{(1-q)}$ di alto ordine di IRC cade sopra il valore di riferimento.

Con l'obiettivo di identificare le RPA stimiamo il modello descritto nel paragrafo 6.8 all'M-quantile 0.85. In Tabella 11 sono riportate le stime dei parametri del modello e in Figura 20 la distribuzione della IRC. Ricordiamo che, sia per costruire le mappe RPA che la mappa di distribuzione della concentrazione del Radon indoor è stato utilizzato il grigliato descritto in precedenza.

Tabella 11 Parametri stimati per il modello Radon Prone Areas

	M-Quantile=0.85	
Parametri	**Stima**	**Errore Standard**
$\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$ ***(intercetta)***	178.75	275.62
$\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$ ***(faglia)***	-76.13	44.00
$\widehat{\boldsymbol{\beta}}_{\mathbf{2}}$ ***(longitudine)***	237.45	385.68
$\widehat{\boldsymbol{\beta}}_{\mathbf{3}}$ ***(latitudine)***	-154.31	387.98
$\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\boldsymbol{e}}$ ***(residuo)***	7758.44	220.07
$\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\boldsymbol{u}}$ ***(litologia)***	302.36	402.06
$\widehat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\boldsymbol{g}}$ ***(spline)***	75250.16	28397.95

Figura 20 Superficie dell'85-esimo M-quantile di IRC

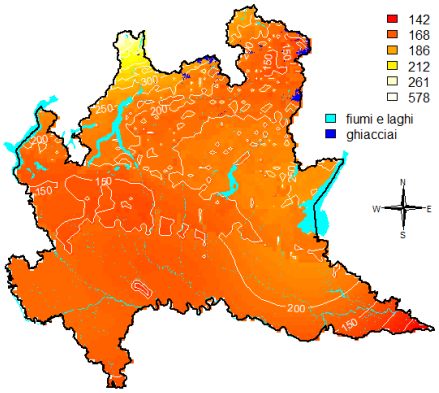
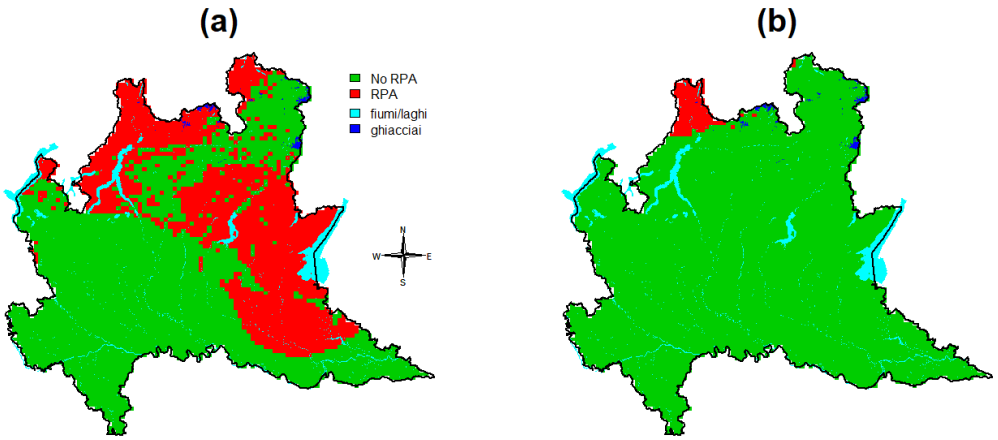


Figura 21 Mappe Radon Prone Areas per i valori di riferimento 200 Bq/m3 (a) e 300 Bq/m3 (b)



Ad ogni cella del grigliato viene associato un valore di IRC che supera il valore limite considerato. La figura 21 (a) e 22 (b) mostra, rispettivamente, i risultati ottenuti considerando i valori che superano i 200 Bq/m^3 (valore limite stabilito nella Direttiva Europea EURATOM 1990/43) e i 300 Bq/m^3 (valore limite suggerito dalla Direttiva Europea più recente EURATOM 2013/59). In rosso vi sono le aree dove l'85-esimo M-quantile stimato è al di sopra dei valori di riferimento 200 Bq/m^3 e 300

Bq/m^3 mentre in verde le aree che sono al di sotto di tali valori. Per quanto riguarda il valore di riferimento 200 Bq/m^3 le Radon Prone Area sono localizzate nella parte estrema della provincia di Varese, nella provincia di Sondrio, parte della provincia di Bergamo e Brescia e nelle provincie di Como e Lecco mentre per il valore 300 Bq/m^3 si trovano solo nella porzione orientale della Provincia di Sondrio.

6.9 Le determinanti della concertazione di IRC: modello additivo con covariate di edificio e effetto random di litologia e spline bivariata

La concentrazione di Radon in una abitazione non dipende solamente dalla morfologia e tipologia litologica sulla quale è costruito ma anche da determinanti antropiche quali: tipologia di edificio, età dell'edificio, attacco a terra (contatto con il suolo), materiali da costruzione (di pareti e di pavimenti), impianto di condizionamento.

Dato ciò, il modello implementato e descritto nel paragrafo precedente è stato esteso per considerare le variabili predittive.

Per tanto il modello ha la forma

$$MQ_q(y|\mathbf{X}) = \boldsymbol{X\beta_q} + \boldsymbol{Zu_q} + \boldsymbol{Z}_{sp}\boldsymbol{\gamma_q}$$

La variabile dipendente y rappresenta la concentrazione media annua di radon indoor.

Tra le variabili indipendenti abbiamo sia fattori ambientali che fattori antropici connessi alle caratteristiche delle edificio.

Fattori ambientali: Distanza di un punto dalla faglia più vicina (variabile normalizzata) e le classi litologiche che verranno utilizzate come fattore di raggruppamento

Fattori antropici/edili:

Pavimentazione: variabile dicotomica (marmo/granito vs altro)

Tipo di edificio: variabile dicotomica (singoli vs non singoli)

Età edificio: variabile dicotomica (edificato prima degli anni '90 vs edificato/ristrutturato dopo gli anni '90

Impianto di aereazione/condizionamento: variabile dicotomica (presente vs assente)

Materiale delle pareti: variabile dicotomica (pietra vs altro materiale)

Contatto con il suolo: variabile dicotomica (in contatto con il terreno vs non in contatto)

Nell'espressione matriciale del modello, tutte queste variabili andranno a costituire la matrice del disegno $\mathbf{X}_{(900\times10)}$ che include il vettore $\mathbf{1}_{(900\times1)}$ per l'intercetta e le coordinate spaziali longitudine e latitudine. Inoltre:

$\boldsymbol{\beta}_{q\,(1\times10)}$ è il vettore dei coefficienti fissi del modello.

$\boldsymbol{u}_{q\,(1\times11)}$ vettore degli effetti random delle classi litologiche

$\mathbf{Z}_{(900\times11)}$ matrice di incidenza

$\mathbf{Z}_{sp\,(900\times50)}$matrice dei coefficienti delle spline bivariata costruita utilizzando le coordinate spaziali delle abitazioni

$\boldsymbol{\gamma}_{\boldsymbol{q}\,(50\times1)}$ effetto random dei coefficienti dei nodi

Per tanto la parte deterministica del modello è legata all'influenza delle covariate di edificio ma anche a quella spaziale mediante latitudine e longitudine e geologico-strutturale (distanza dalla faglia).

Ai fini della stima il numero di nodi è stato posto pari a 50, il parametro c =1.345 e i valori iniziali sono ottenuti dall'implementazione del modello misto P-spline standard.

L'intercetta associata a ciascuno degli M-quantili condizionati rappresenta il profilo abitativo baseline ad una data locazione geografica e classe litologica ovvero di una ipotetica abitazione, non singola costruita prima degli anni novanta priva di impianto di aerazione e di vespaio o di qualsiasi sistema che la isola dal suolo.

In Tabella 12 sono riportati i risultati della stima dei parametri per il modello con tre effetti random: spline bivariata, litologia e componente erratica, per i quantili 0.25, 0.50, 0.75.

Tabella 12 Stima dei parametri del modello per gli M-quantili 0.25, 0.50 e 0.75

	M-Quantile=0.25			**M-Quantile=0.50**			**M-Quantile=0.75**		
Parametri	**Stima**	**Errore standard**	**p-value**	**Stima**	**Errore Standard**	**p-value**	**Stima**	**Errore Standard**	**p-value**
$\hat{\beta}_{0(intercetta)}$	27.24	13.53	0.044	45.59	25.12	0.069	84.93	74.86	0.256
$\hat{\beta}_{1(distanza\ faglia)}$	-3.21	8.27	0.697	-10.25	13.47	0.446	-35.19	28.08	0.21
$\hat{\beta}_{2(pavimenti\ marmo)}$	-5.41	5.54	0.328	-8.88	8.87	0.316	-22.43	17.32	0.195
$\hat{\beta}_{3(pareti\ pietra)}$	3.77	3.94	0.338	6.53	6.31	0.301	16.86	12.38	0.173
$\hat{\beta}_{4(età\ dopo\ anni\ '90)}$	7.99	2.55	0.001	11.25	4.09	0.005	14.11	8.003	0.077
$\hat{\beta}_{5(singolo)}$	6.35	2.69	0.018	10.36	4.32	0.016	24.98	8.49	0.003
$\hat{\beta}_{6(non\ contatto\ con\ il\ suolo)}$	-5.73	2.67	0.032	-9.74	4.28	0.022	-18.24	8.35	0.029
$\hat{\beta}_{7(no\ condizionatore)}$	1.09	5.16	0.832	-1.48	8.29	0.858	-7.59	16.24	0.64
$\hat{\beta}_{8(longitudine)}$	33.40	15.44	0.030	42.43	30.23	0.160	87.73	97.85	0.369
$\hat{\beta}_{9(latitudine)}$	36.62	18.77	0.05	48.81	35.19	0.165	38.58	106.14	0.716
$\hat{\sigma}^2_{e(residuo)}$	819.88	48.91		3172.99	241.93		6283.10	421.41	

$\hat{\sigma}^2_{u(gruppo)}$	42.21	26.41		149.70	85.48		220.13	92.08	
$\hat{\sigma}^2_{g(spline)}$	75.96	-		491.78	-		3140.52	-	

Per tutti i modelli (M-quantili) la componente di varianza di gruppo $\hat{\sigma}^2_{uq}$ aumenta andando dal M-quantile 0.25 al M-quantile 0.75. Allo stesso modo, le componenti di varianza legata al termine di spline e di errore, così come i valori (in valore assoluto) dei coefficienti fissi hanno valori che aumentano al crescere del M-quantile.

Per gli M-quantili 0.25 e 0.50 le variabili predittive statisticamente significative sono età di costruzione dell'edificio (edificato/ristrutturato dopo gli anni 90), il tipo di edificio (singolo) e piano campagna (non contatto con il suolo). Per il M-quantile 0.75 le variabili significative sono solo due ovvero il tipo di edificio (singolo) e piano campagna (non contatto con il suolo).

6.9.1 IRC in base ai Profili abitativi

A partire dalla stima dei parametri forniti dal modello è stato possibile ottenere gli M-quantili di concentrazione di radon indoor per diversi profili abitativi.

Relativamente agli M-quantili 0.25 e 0.50, costruiamo un grigliato $2^3 \times 11$ combinando le 3 variabili indipendenti dicotomiche che sono risultate statisticamente significative (età edificio, tipo di edificio e contatto con il suolo), le 11 classi litologiche e un solo valore delle coordinate spaziali (latitudine=5056487, longitudine=1564925). La coordinata spaziale è stata preventivamente normalizzata.

Le tabelle con tutti gli 88 profili individuati per gli M-quantili 0.25 e 0.50 sono riportati in appendice D.

Il valore di IRC per ciascuna delle 88 combinazioni e per l'M-quantile di ordine q=0.25 e 0.50 è ottenuta da:

$$MQ_q(IRC) = \hat{\beta}_0 + \hat{\beta}_1 età + \hat{\beta}_2 tipo + \hat{\beta}_3 contatto + \hat{\beta}_4 latitudine + \hat{\beta}_5 longitudine + \mathbf{Z}\hat{u} + \boldsymbol{Z}_{sp}\hat{\boldsymbol{\gamma}}$$

Nelle tabelle sottostanti sono riportate i primi dieci profili abitativi che hanno valori di IRC minori (Tabella 13) e i primi 10 profili abitativi con i valori di IRC più elevati (Tabella 14) per l'M-quantile 0.25.

Tabella 13 Profili abitativi a minor rischio per l'M-quantile 0.25

Numero profilo	Età	Tipo	Connessione suolo	Litologia	IRC
61	Prima anni ‘90	Non singolo	No	Magmatiche Metamorfiche Basiche	40.07
85	Prima anni ‘90	Non singolo	No	Morene	43.52
5	Prima anni ‘90	Non singolo	No	Depositi fini	43.77
57	Prima anni ‘90	Non singolo	Si	Magmatiche Metamorfiche Basiche	45.81
63	Prima anni ‘90	Singolo	No	Magmatiche Metamorfiche Basiche	46.43
69	Prima anni ‘90	Non singolo	No	Metamorfiche	47.63
62	Dopo anni ‘90	Non singolo	No	Magmatiche Metamorfiche Basiche	48.07
13	Prima anni ‘90	Non singolo	No	Alta Pianura	48.29
81	Prima anni ‘90	Non singolo	Si	Morene	49.25
29	Prima anni ‘90	Non singolo	No	Conoidi	49.29

I primi tre profili meno a rischio sono quelli che corrispondono ad edifici non singoli costruiti su rocce magmatiche metamorfiche basiche, morene e depositi fini, non in connessione con il suolo ed costruite prima degli anni novanta.

Tabella 14 Profili abitativi a maggior rischio per l'M-quantile 0.25

Numero profilo	Età	Tipo	Connessione suolo	Litologia	IRC
36	Dopo anni ‘90	Singola	Si	Detriti	84.88
44	Dopo anni ‘90	Singola	Si	Dolomie del Triassico	81.40
40	Dopo anni ‘90	Singola	Si	Detriti	79.15
34	Dopo anni ‘90	Non singola	No	Detriti	78.52
35	Prima anni ‘90	Singola	No	Detriti	76.88
48	Dopo anni ‘90	Singola	No	Dolomie del Triassico	75.67
42	Dopo anni ‘90	Non singola	Si	Dolomie del Triassico	75.05
76	Dopo anni ‘90	Singola	Si	Alluvionale di Montagna	73.95
43	Dopo anni ‘90	Singola	Si	Dolomie del Triassico	73.41
38	Dopo anni ‘90	Non singola	No	Detriti	72.79

Le abitazioni più a rischio sono quelle che sono costruite in diretto contatto con il substrato roccioso detritico e dolomitico, isolate nelle loro immediate vicinanze dalle altre costruzioni ed edificate dopo gli anni novanta (numero profilo 36 e 44). Minore è il rischio del profilo numero 40 che vede l’edificio non direttamente in connessione con la roccia detritica ne con altre abitazioni e costruito dopo gli anni 90. Leggermente inferiore a quest’ultimo è il rischio del profilo 34 in cui le costruzioni sono sia in contatto tra loro che con il substrato detritico ma costruite dopo gli anni novanta.

Il quinto profilo (35) più critico è quello per le abitazioni costruite su substrato detritico, singole e edificate prima degli anni ottanta.

Nelle tabelle sottostanti sono riportate i primi dieci profili abitativi che hanno valori di IRC minori (tabella 15) e i primi 10 profili abitativi con i valori di IRC più elevati (tabella 16) per l’M-quantile 0.50.

Tabella 15 Profili abitativi a minor rischio per l'M-quantile 0.50

Numero profilo	Età	Tipo	Connessione suolo	Litologia	IRC
61	Prima anni ‘90	Non singolo	No	Magmatiche Metamorfiche Basiche	60.36
5	Prima anni ‘90	Non singolo	No	Depositi fini	62.63
85	Prima anni ‘90	Non singolo	No	Morene	62.94
57	Prima anni ‘90	Non singolo	Si	Magmatiche Metamorfiche Basiche	70.10
69	Prima anni ‘90	Non singolo	No	Metamorfiche	70.42
63	Prima anni ‘90	Singolo	No	Magmatiche Metamorfiche acide	70.72
53	Prima anni ‘90	Non singolo	No	Magmatiche Metamorfiche acide	70.82
62	Dopo anni ‘90	Non singolo	No	Magmatiche Metamorfiche acide	71.61
1	Prima anni ‘90	Non singolo	Si	Depositi fini	72.38
81	Prima anni ‘90	Non singolo	Si	Morene	72.69

I profili abitativi caratterizzati da un rischio minore sono i primi tre, si tratta di abitazioni non in contatto con il substrato roccioso di tipo magmatico-basico, depositi fini e morene, costruite prima degli anni novanta e in connessione con altre abitazioni. Per quanto riguarda il profilo 57 e 69 non ci sono differenze sostanziali nei valori di radon ottenuti ma nel quarto abbiamo le abitazioni in contatto con il substrato magmatico basico, in connessione tra loro e costruite prima degli anni ottanta mentre nel quinto le case sono in contatto con la roccia metamorfica.

Tabella 16 Profili abitativi ad elevato rischio per l'M-quantile 0.50

Numero profilo	Età	Tipo	Connessione suolo	Litologia	IRC
36	Dopo anni ‘90	Singolo	Si	Detriti	134.71
40	Dopo anni ‘90	Singolo	No	Detriti	124.96
34	Dopo anni ‘90	Non singolo	Si	Detriti	124.35
35	Prima anni ‘90	Singolo	Si	Detriti	123.46
44	Dopo anni ‘90	Singolo	Si	Dolomie del Triassico	121.53
38	Dopo anni ‘90	Non singolo	No	Detriti	114.61
39	Prima anni ‘90	Singolo	No	Detriti	113.71
33	Prima anni ‘90	Non singolo	Si	Detriti	113.10
48	Dopo anni ‘90	Singolo	No	Dolomie del Triassico	111.80
42	Dopo anni ‘90	Non singolo	Si	Dolomie del Triassico	111.18

Il profilo abitativo 36 è quello potenzialmente più a rischio, e rappresenta una abitazione in contatto con substrato roccioso detritico che non ha connessione alcuna con le altre abitazioni circostanti ed edificata dopo gli anni novanta.

Il secondo profilo (40) riguarda abitazioni non in contatto con il substrato roccioso detritico ne con altri edifici costruiti dopo gli anni novanta. Il terzo profilo (34) pur presentando un valore di radon molto simile al precedente si distingue da questo per le abitazioni in contatto con la roccia detritica e con le altre abitazioni.

Il quarto profilo (35) riguarda abitazioni edificate prima degli anni novanta, isolate nelle loro immediate vicinanze e in contatto con la roccia detritica.

Le variabili risultate statisticamente significative per l’M-quantile 0.75 sono solo due (tipo edificio e contatto con il suolo) pertanto si è costruita un grigliato $2^2 \times 11$ per un totale di 44 punti griglia.

$$MQ_{0.75}(IRC) = \hat{\beta}_0 + \hat{\beta}_1 tipo + \hat{\beta}_2 contatto + \hat{\beta}_3 latitudine + \hat{\beta}_4 longitudine + \mathbf{Z}\hat{u} + \boldsymbol{Z}_{sp}\hat{\boldsymbol{\gamma}}$$

Nelle tabelle sottostanti sono riportate i primi dieci profili abitativi che hanno valori di IRC minori (Tabella 17) e i primi 10 profili abitativi con i valori di IRC più elevati (Tabella 18) per l'M-quantile 0.75. In questo caso i profili che presentano a minor rischio sono quelli corrispondenti ad abitazioni connesse tra loro e non in contatto con il substrato roccioso della tipologia depositi fini.

Tabella 17 Profili abitativi a basso rischio per L'M-quantile 0.75

Numero profilo	Tipo	Connessione suolo	Litologia	IRC
3	Non singolo	No	Depositi fini	92.22
43	Non singolo	No	Morene	92.44
35	Non singolo	No	Metamorfiche	92.95
31	Non singolo	No	Magmatiche Metamorfiche Basiche	103.02
27	Non singolo	No	Magmatiche Metamorfiche Acide	107.19
39	Non singolo	No	Alluvionale di Montagna	109.02
1	Non singolo	Si	Depositi fini	110.46
41	Non singolo	Si	Morene	110.69
33	Non singolo	Si	Metamorfiche	111.20
7	Non singolo	No	Alta Pianura	113.06

Tabella 18 Profili abitativi ad elevato rischio per L'M-quantile 0.75

Numero profilo	Tipo	Connessione suolo	Litologia	IRC
18	Singolo	Si	Detriti	179.39
22	Singolo	Si	Dolomie del Triassico	175.58
14	Singolo	Si	Conoidi	161.54
20	Singolo	No	Detriti	161.15
10	Singolo	Si	Calcari	159.18
24	Singolo	No	Dolomie del Triassico	157.34
6	Singolo	Si	Alta Pianura	156.29
17	Non singolo	Si	Detriti	154.41
38	Singolo	Si	Alluvionale di Montagna	152.25
21	Non singolo	Si	Dolomie del Triassico	150.60

Riassumendo, uno dei profili abitativi più rischiosi in quanto potrebbe presentare una concentrazione di Radon indoor elevata con conseguenti effetti sanitari sulla popolazione esposta corrispondono ad abitazioni singole, costruite dopo gli anni novanta su litologie detritiche che non sono dotate di sistemi che lo isolano dal suolo (vespaio).

Presumibilmente si tratta di abitazioni che hanno subito di recente degli interventi di efficientamento energetico che prevede, ad esempio, l'uso di materiali isolanti atti a ridurre gli scambi di energia con l'ambiente esterno e per tanto imprigionano il Radon all'interno del locale non favorendone la sua diluizione e dispersione, inoltre i condomini o altre tipologie di abitazioni che prevedono le pareti in contatto sembrerebbero più difficili da penetrare rispetto a quelle singole.

Tenendo presente che per l'identificazione dei profili si è scelto di includere solo le variabili statisticamente significative e che per agevolarne l'individuazione, un solo punto di coordinate cartografiche (in particolare latitudine e longitudine di Begamo) possiamo cercare di riportare quanto ritrovato alla situazione della realtà della regione Lombardia. A tale scopo facendo un cross-check tra il dataset (filtrando per le variabili) e i profili ottenuti è emerso che nella provincia di Sondrio, Bergamo, Brescia e Como esistono abitazioni corrispondenti al profilo più rischioso e che presentano IRC elevati (Tabella 15).

I profili meno rischiosi realmente presenti in Lombardia si trovano nella provincia di Lodi, Pavia, e Cremona, Como, Varese.

7. DISCUSSIONE E CONCLUSIONI

In questo lavoro di tesi è stato proposto un modello semi-parametrico M-quantile ad effetti casuali in grado di cogliere il trend spaziale nei dati ambientali.

Nel modello la componente spaziale è stata descritta combinando il termine casuale di intercetta con quello semi-parametrico. L'inclusione del termine di intercetta consente di catturare l'effetto del gruppo sulla variabile di risposta mentre la presenza del termine semi-parametrico coglie la variabilità regolare nello spazio geografico. Il modello è stato trattato come un modello ad effetti misti in cui i coefficienti dei nodi spline entravano come un ulteriore effetto random.

La combinazione dell'approccio M-quantile (Breckling J. and Chamber R., 1988) con quello di smoothing (mediante spline bivariata), in un framework di modelli ad effetti misti, ha permesso di ottenere un modello facilmente trattabile, che coniugasse le caratteristiche di robustezza e flessibilità proprie degli approcci citati. Inoltre, in questo modo, è stato possibile stimare direttamente il parametro di smoothing, λ, senza ricorrere a metodi esterni, quale ad esempio la cross-validation.

Per quanto attiene la stima robusta dei parametri del modello, si è optato per un metodo a due stadi in cui gli effetti fissi e le componenti di varianza sono stati stimati sequenzialmente (Tzavidis et al., 2010, 2014, 2016; Richardson and Welsh, 1995).

Le prestazioni del modello, in termini di stima e di predizione, sono state esaminate predisponendo due distinti studi di simulazione Monte Carlo. Un terzo disegno è stato predisposto per valutare l'efficacia del modello proposto nel caso univariato, comparandolo con il modello di riferimento Non-parametric M-quantile P-spline.

Nella prima simulazione Monte Carlo, per ogni combinazione della componente di varianza del termine di spline (basso (25), medio(100), alto (400)), e due valori per ciascuna delle rimanenti componenti di varianza (basso pari a 25 sia per σ_u^2 che σ_e^2, e alto pari a 400 per σ_e^2 e 100 per σ_u^2), sono stati generati 1000 dataset sui quali sono stati stimati sia i coefficienti fissi che le componenti di varianza. Ricordiamo che abbiamo scelto tre distinti valori per la componente spline poiché si sono riscontrate le maggiori criticità. Per valutare concretamente la performance e stabilità della procedura di stima, per ciascun vettore dei parametri stimati, è stato calcolato il *Mean Relative Bias* (MRB) percentuale. In generale la componente di varianza spline in tutti gli M-quantili risulta sottostimato più o meno fortemente, inoltre valori di MRB più elevati in valore assoluto si sono ottenuti in diverse combinazioni, in particolare, dove la componente di varianza della spline e dell'errore erano

rispettivamente alto e basso. A contrario le prestazioni migliori si sono avute quando la componente di varianza di gruppo era intermedio e le altre due avevano valori bassi.

I risultati ottenuti dal secondo disegno indicano che questo approccio funziona discretamente in termini di predizione, anche in presenza di dati anomali e con relazioni complesse tra covariate e variabile di risposta.

Il terzo disegno ha però evidenziato dei deficit quando comparato con un modello alternativo proposto precedentemente in letteratura (Pratesi et al., 2009).

Successivamente, il modello è stato applicato a dati ambientali reali quali le concentrazione di radon indoor negli edifici. Il Radon è un gas chimicamente inerte, inodore, incolore ma radioattivo i cui effetti cancerogeni sono ormai riconosciuti. L'esposizione al gas Radon avviene principalmente in ambienti confinati quali residenze private e posti di lavoro dove può raggiungere concentrazioni anche molto elevate. Il Radon può penetrare negli ambienti confinati ad esempio attraverso crepe e fenditure, giunti di connessione, canalizzazioni, ecc. presenti nell'attacco a terra delle costruzioni (Barros-Dios, et al., 2007) ma può anche essere rilasciato in quantità più o meno elevate dai materiali da costruzione che talvolta presentano forti esalazioni. Nel primo caso gioca un ruolo importante la natura del substrato roccioso su cui sono costruiti gli edifici specialmente se questi presentano una qualsiasi forma di connessione con il suolo (tipologia e tecnologia costruttiva dell'attacco a terra); infatti ogni classe litologica è caratterizzata da un contenuto naturale di Uranio, che dipende dalle caratteristiche mineralogiche e strutturali delle rocce stesse, di cui il Radon è un prodotto di decadimento. Molti studi hanno evidenziato che le abitazioni che presentavano alte concentrazioni di Radon Indoor erano costruite su formazioni rocciose di tipo granito, gneiss, sedimentarie e sedimentarie altamente fratturate (Farah, et al., 2012; Minda, et al., 2009; Park, et al., 2011). Le rocce carbonatiche (calcari e dolomie) sono soggette a fenomeni carsici, che modellando fratture formano cave e sistemi reticolari all'interno della formazione, contribuendo in questo modo al trasporto in superficie di alte quantità di Radon (Buttafuoco, et al., 2010; Kropat, et al., 2014). Altri fattori rilevanti sono le caratteristiche stesse dell'edificio (forma, dimensione, disposizione delle bucature, livello rispetto al suolo dei locali abitati, ecc.) le modalità di uso dell'edificio/abitazione e abitudini di chi ci vive (ad esempio quanto fanno aerare la casa, ecc.).

Tutti questi fattori, con effetto additivo o esclusivo, contribuiscono all'aumento della concentrazione di Radon indoor e rendono l'IRC soggetta ad una grande variabilità sia spaziale che temporale. In questo lavoro ci siamo concentrati unicamente alla modellizzazione della variabilità spaziale oltre che agli effetti differenziali che i fattori costruttivi hanno sui diversi punti della distribuzione condizionata di IRC.

L’applicazione del modello sui dati provenienti dalla prima campagna di monitoraggio condotta da ARPA in Lombardia nel 2003-2004, ha permesso di mappare le concentrazioni di Radon indoor, di ottenere la mappa delle Radon Prone area per l’M-quantile 0.85, di individuare le variabili che maggiormente contribuiscono alla spiegazione del fenomeno e quindi di ricostruire i profili di rischio abitativo.

In particolare dall’implementazione del modello semiparametrico M-quantilico che considera solo le componenti geologico-strutturali e geografiche è stato possibile ottenere la mappa di Radon Prone Area. E’ emerso che le aree geografiche che hanno una più alta probabilità di eccedere il valore di 200 $Bq \cdot m^{-3}$ sono localizzate nella estrema provincia di Varese, Sondrio, Bergamo e Brescia, Lecco e Como, mentre quelle che eccedono i 300 $Bq \cdot m^{-3}$ si trovano nella parte più orientale della Provincia di Sondrio.

Successivamente, nella parte deterministica di questo modello sono state incluse i fattori costruttivi ed antropici. Le variabili che influenzano gli M-quantili condizionati 0.25, 0.50 sono il tipo di edificio, l’anno di costruzione o ristrutturazione e il contatto con il suolo, confermando quanto trovato da Borgoni (2011) mentre i fattori che agiscono sul M-quantile condizionato 0.75 sono il tipo di edificio e il contatto con il suolo. In particolare, per una generica abitazione ad una data localizzazione geografica ed edificata su una determinata classe litologica, non essere in connessione con il substrato roccioso ha un effetto protettivo rispetto all’IRC tanto più intenso tanto più si va verso M-quantili estremi. Allo stesso tempo, il passaggio da una costruzione non singola a una singola e dall’anno di edificazione/ristrutturazione prima degli anni novanta a dopo gli anni novanta, al netto delle altre variabili, hanno un effetto positivo sull’IRC (nel senso che contribuisce ad un suo aumento). Inoltre, diversi tipi di fondamenta (piano interrato, vespaio, basamento in cemento) hanno un rilevante impatto sulla variabilità di IRC (Brauner et al., 2013). Molti studi riportano che abitazioni in diretto contatto con il suolo (basamento) presentano valori di IRC più elevati rispetto a quelle dotate di piani interrati o semi-interrati (Borgoni et al., 2011, Andersen et al.., 2007) e che l’essere in contatto diretto con il suolo è uno dei fattori che determina elevate contrazioni di radon indoor (Demoury et al., 2013). Contrariamente a quanto affermato in questi studi, altri autori hanno messo in evidenza che case con piani interrati o semi-interrato hanno un elevato IRC rispetto a quelle con basamento in cemento (Arvela et al.,2012; Alghamdi e Aleissa, 2014; Kitto & Green, 2008) questo perché sono scarsamente ventilate e consentono al gas di raggiungere livelli di concentrazioni pericolose (Alghamdi e Aleissa, 2014, Harnapp et al., 1997). Per quanto attiene la tipologia di costruzione, anche Sundal et al. (2004) indicano che le abitazioni non singole, come i condomini, mostrano un livello di IRC più basso rispetto alle case singole, questo è in parte dovuto al fatto che le case indipendenti sono generalmente costruite su un singolo piano in prossimità con il suolo (Borgoni et al., 2011; Demoury et al., 2013).

In definitiva, i profili abitativi maggiormente a rischio sono edificati in contatto con classi litologiche di tipo detritico e dolomitico, dopo gli anni novanta e sono degli edifici singoli mentre le abitazioni a minor rischio sono quelle multifamiliari edificate su depositi fini e morene prima degli anni novanta. Circa l'applicazione ai dati reali, il modello ha risposto bene riuscendo a cogliere la componente spaziale (trend) presente nei dati, come riportato nella letteratura pregressa (e.g. Borgoni et al., 2010). Un limite di questo modello, oltre ad aver manifestato oneri computazionali (elevato tempo-macchina), è quello di peccare a livello di stabilità dell'algoritmo di stima che si è manifestato sia durante esecuzione del primo disegno di simulazione sia durante l'analisi dei trend spaziali. Come si è visto i valori della componente di varianza di gruppo (sia nell'analisi dei trend spaziali che nell'analisi congiunta tra fattori edilizi e spaziali) erano più basse rispetto a quelle delle altre due componenti. Presumibilmente ciò è dovuto alla fatto che si sia adottata una classificazione delle litologia più grossolana su di una scala 1:25.000 riducendo la variabilità rispetto a quella che è la realtà.

La costruzione della mappa di Radon Prone Area e le identificazioni dei profili di rischio possono costituire uno strumento utile alle istituzioni preposte alla pianificazione urbanistica, nel caso si dovessero costruire nuove abitazioni, e alla protezione sanitaria della popolazione (individuazione di residenze esistenti che posseggono le caratteristiche rilevate potrebbero essere sottoposte ad azioni di mitigazione).

La complessità dei fenomeni spaziali/ambientali richiedono modelli altrettanto elaborati per poter essere descritti ed interpretati in modo adeguato. Di conseguenza una prima linea di sviluppo potrebbe focalizzarsi sulla costruzione di un modello semiparametrico M-quantile a tre livelli gerarchici in modo da tener conto di eventuali ulteriori sorgenti di variabilità nello spazio.

Affinchè il modello diventi operativo a tutti gli effetti è necessario sviluppare maggiormente l'aspetto connesso alla derivazione degli errori standard e del suo stimatore sia a livello teorico che empirico. Molto interessante infatti, sarebbe disporre di uno stimatore dello standard error per la stima di ciascun M-quantile nelle diverse locazioni spaziali e dell'intervallo di confidenza per tale parametro.

BIBLIOGRAFIA

- Aitkin, M., (1999). A general maximum likelihood analysis of variance components in generalized linear models. Biometrics, 55,117-128.

- Aitkin, M., (1996).A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and Computing, 6,251-262.

- Alfò, M., Marino, M.F., Ranalli, M.G., Salvati, N., (2017). Multivariate M-quantile regression for longitudinal data: analysis of the Millennium Cohort Study data.

- Alfò, M., Ranalli, M.G., Salvati, N. (2017).Finite mixtures of quantiles and M-quantile models. Statistics and Computing, 27(2), 547-570.

- Alghamdi, A. S., Aleissa, K. A, (2014). Influences on indoor radon concentrations in Riyadh, Saudi Arabia; Radiation Measurements62:35-40. Doi 10.1016/j.radmeas.2014.01.010.

- Almasri, A., Andersson, E. M., Barregård, L., (2009). A study of residential radon in Sweden using multi-level analysis. Health Physics, 96, 442-449.
- Anderson, T.W., (1973) Asymptotically efficient estimation of covariance matrices with linear covariance structure. Ann. Statist., 1, 135–141.
- Andersen, C. E., Raaschou-Nielsen, O., Andersen, H.P., Linda, M., Gravesen, P., Thomsen, B.L., Ulbak. K., (2007). Prediction of Rn-222 in Danish dwellings using geology and house construction information from central databases; Radiation Protection Dosimetry 123 (1):83-94. doi: 10.1093/rpd/ncl082.
- Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., Tukey, J.W., (1972). Robust Estimates of Location: Survey and Advances. Princeton: Princeton University Press.
- Antoch, J., Ekblom, H., Visek, J.A., (1998). Robust Estimation in Linear Model. XploRe Macros: http://www.quantlet.de/codes/rob/ROB.html
- Apte, M.G., Price, P.N., Nero, A.V., Revzan, K.L., (1999). Predicting New Hampshire indoor radon concentrations from geologic information and other covariates. Environ. Geol , 37, 181–194.
- Appleton, J.D., Miles, J.H.C., (2010). A statistical evaluation of the geogenic controls on indoor radon concentrations and radon risk. J. Environ. Radioact., 101(10), 799-803.

- Arvela, H., Holmgren, O., Reisbacka, H., (2012). Radon prevention in new construction in Finland: a nationwide sample survey in 2009. Radiation Protection Dosimetry 148 (4):465-474. doi: 10.1093/rpd/ncr192.
- Azzalini, A., Scarpa, B., (2004). Analisi dei dati e data mining. Collana di statistica e probabilità applicata. Springer.
- Barry, A. D., Charpentier, A., Oualkacha, K., (2016). Quantile and Expectile Regression for random effects model. <hal-01421752>

  Barros-Dios, J. M., Ruano-Ravina, A., Gastelu-Iturri, J., Figueiras, A.,(2007). Factors underlying residential radon concentration: Results from Galicia, Spain; Environmental Research 103 (2):185-190. doi: 10.1016/j.envres.2006.04.008.
- Bochicchio, F., Zunic, Z.S., Carpentieri, C., Antignani, S., Venoso, G., Carelli, V., Cordedda, C., Veselinovic, N., Tollefsen, T., Bossew, P., (2013). Radon in indoor air of primary schools: a systematic survey to evaluate factors affecting radon concentration levels and their variability. Indoor Air. http://dx.doi.org/10.1111/ina.12073.
- Borgoni, R., (2011). A quantile regression approach to evaluate factors influencing residential indoor radon concentration. Environmental Modelling and Assessment, 16, 3, 239-250
- Borgoni, R., Quatto, P., Somà, G., De Bartolo, D. A., (2010). Geostatistical approach to define guidelines for radon prone area identification. Statical Methods Application, 19, 255–276.
- Borgoni, R., Tritto, V., de Bartolo, D., (2013). Identifying radon-prone building typologies by marginal modelling. J. Appl. Stat. 40 (9), 2069-2086.
- Borgoni, R., De Francesco, D., De Bartolo, D. A., Tzavidis, N., (2014). Hierarchical modeling of indoor radon concentration: how much do geology and building factors matter?. Journal of Environmental Radioactivity, 138, 227-237.
- Borgoni, R., Del Bianco, P., Salvati, N., Schmid, T., Tzavidis, N., (2016).Modelling the distribution of health related quality of life of advanced melanoma patients in a longitudinal multi-centre clinical trial using M-quantile random effects regression. Statistical Methods in Medical Research, 1-18.
- Bossew, P., Stojanovska, Z., Zunic, Z. S., & Ristova, M. (2013). Prediction of Indoor Radon Risk from Radium concentration in soil: Republic of Macedonia Case Study. Romanian Journal of Physics, 58, S29–S43, Bucharest.
- Bossew, P. and Lettner, H., (2007) Investigations on indoor radon in Austria, part 1: seasonality of indoor radon concentration, J. Environ. Radioactiv., 98, 329–345.
- Brauner, E.V., Rasmussen, T.V., Gunnarsen, L., (2013). Variation in residential radon levels in new Danish homes. Indoor Air, 23, 311-317.

- Brauner, E. V., Andersen, C. E., Sorensen, M., Andersen, Z. J., Gravesen, P., Ulbak, K., Hertel, O., Pedersen, C., Overvad, K., Tjonneland, A., Raaschou-Nielsen, O. (2012). Residential radon and lung cancer incidence in a Danish cohort; Environmental Research 118:130-136. doi: 10.1016/j.envres.2012.05.012.
- Breckling, J., Chambers, R., (1988) M-quantiles. Biometrika (75), 761-771.
- Bryk, A.S., Raudenbush, S. W., (1992) Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage Publications.
- Buchinsky, M., (1998). Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research. Journal of Human Resources, 33(1), 88-126.
- Bunke, H., Bunke, O., (1986). Statistical Inference in Linear Models. John Wiley & Sons, Chichester, U.K
- Burke, Ó., Long, S., Murphy, P., Organo, C., Fenton, D., Colgan, P.A., (2010). Estimation of seasonal correction factors through Fourier decomposition analysis-a new model for indoor radon levels in Irish homes J. Radiol. Prot. 30, 433–43.
- Buttafuoco, G., Tallarico, A., Falcone, G., Guagliardi, I., (2010)A geostatistical approach for mapping and uncertainty assessment of geogenic radon gas in soil in an area of southern Italy; Environmental Earth Sciences 61 (3):491-505. doi: 10.1007/s12665-009-0360-6.
- Canay, I.A., (2011). A simple approach to quantile regression for panel data. The Econometrics Journal, 14(3),368-386.
- Chambers, J.M., Hastie, T.J., (1991). Statistical Models in S, Chapman & Hall Computer Science Series.
- Chambers, R., Tzavidis, N., (2006). M-quantile models for small area estimation. Biometrika 93 (2), 255–268.
- Chambers, R.; Tzavidis, N.; and Salvati, N., (2009). Borrowing strength over space in small area estimation: Comparing parametric, semiparametric and non-parametric random effects and M-quantile small area models, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper. http://ro.uow.edu.au/cssmwp/32
- Chen, C., (2007). A finite smoothing algorithm for quantile regression. Journal of Computational and Graphical Statistics ,16(1), 136–164.
- Chen, C., (2004). An adaptive algorithm for quantile regression. In Theory and Applications of Recent Robust Methods (Hubert M, Pison G, Struyf A and Van Aelst S eds), 39–48.

- Cinelli G, Tondeur F, Dehandschutter B., (2010) Development of an indoor risk map of the Wallon region of Belgium, integrating geological information. Environmental Earth Sciences, 62, 809–819. doi: 10.1007/s12665-010-0568-5.
- Comincioli, V., (2005). Analisi numerica: metodi, modelli, applicazioni Apogeo.
- Darby, S., Hill, D., Auvinen, A., Barros-Dios, J.M., Baysson, H., Bochicchio, F., Doll, R. (2005). Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case–control studies. BrMedJ.;330:223–227.
- de Boor, C., (2001). A Practical Guide to Splines, Revised edition, Applied Mathematical Sciences 27, Springer–Verlag, New York.
- Demoury, C., Ielsch, G., Hemon, D., Laurent, O., Laurier, D., Clavel, J., Guillevic, J.,(2013). A statistical evaluation of the influence of housing characteristics and geogenic radon potential on indoor radon concentrations in France; Journal of Environmental Radioactivity 126:216-225. doi: 10.1016/j.jenvrad.2013.08.006.
- Denman, A., Crockett, R., Groves-Kirkby, C., Phillips, P., Gillmore, G., Woolridge, A., (2007). The value of seasonal correction factors in assessing the health risk from domestic radon—a case study in Northamptonshire, UK Environ. Int. 33 34–44
- Dierckx, P. (1993). Curve and Surface Fitting with Splines. Clarendon Press, Oxford.
- DIRETTIVA 2013/59/EURATOM: "Norme fondamentali di sicurezza relative alla protezione contro i pericoli derivanti dall'esposizione alle radiazioni ionizzanti, Retrieved from:http://eurlex.europa.eu/legalcontent/IT/TXT/?uri=uriserv%3AOJ.L_.2014.013.01.0001.01.ITA.
- Dreassi, E., Ranalli, M. G., Salvati, N., (2015). Semiparametric M-quantile Regression for Count data. Statistical methods in medical research 23 (6), 591-610.
- EC (European Commission). (1996). Council Directive 96/29/Euratom of 13 May 1996 laying down basic safety standards for the protection of the health of workers and the general public against the dangers arising from ionizing radiation. Official Journal L-159 of 27/06/96. European Commission, Bruxelles.
- EC (European Commission). (1997). Radiation Protection 88 Recommendations for the implementation of title VII of the European Basic Safety standards Directive (BSS) concerning significant increase in exposure due to natural radiation sources.
- EC (European Commission). (2013). Council Directive 2013/59/Euratom of 5 December 2013 laying down Basic Safety Standards for Protection against the Dangers Arising from Exposure to Ionising Radiation. Official Journal L13 of 17/01/2014 European Commission, Bruxelles.

- Eilers, P.H.C., Marx, B.D., (1996). Flexible smoothing with B-splines and penalties (with discussion). Statistical Science, 11, 89–121.
- Eilers, P.H.C., Marx, B.D., (2010). Splines, knots, and penalties. WIREs Comp Stat, 2: 637–653. doi:10.1002/wics.125.
- Elìo, J., Crowley, Q., Scanlon, R., Hodgson, J., Long, S., (2017). Logistic regression model for detecting radon prone areas in Ireland. Science of The Total Environment, 599, 1317-1329.
- US-EPA, (2003). EPA assessment of risks from radon in homes. Office of Radiation and Indoor Air, Washington, DC.
- Farah, C., Beard, K., Hess, C. T., Hock, J. M., (2012). Analyzing spatial and temporal (Rn)-R-222 trends in maine. Health Physics102 (2):115-123. doi: 10.1097/HP.0b013e318231aa9a.
- Farcomeni, A., (2012). Quantile Regression for longitudinal data based on latent Markov subject-specific parameters Statistics and Computing, 22, 141-152 (the routine described in appendix to compute sums on the log scale is available in the package snipEM, on CRAN).
- Farcomeni, A.,Viviani, S., (2015). Longitudinal quantile regression in presence of informative drop-out through longitudinal-survival joint modeling, Statistics in Medicine, 34, 1199-1213.

  Field, W., (2015). Radon: An Overview of Health Effects. Encyclopedia of Environmental Health, 745-753.
- Fitzenberger, B., Wilke, R., (2006). Using Quantile Regression for Duration Analysis. Allgemeines Statistisches Archiv, 90(1), 103-118.
- Fontanella, L., Ippoliti, L., Sarra, A., Valentini, P., Palermi, S., (2015). Hierarchical generalised latent spatial quantile regression models with applications to indoor radon concentration. Stochastic Environmental Research and Risk Assessment, 29 (2), 357–367
- French, J.L., Kammann, E.E., Wand, M.P., (2001). Comment on paper by Ke and Wang. Journal of the American Statistical Association 96, 1285–1288.

  Friedmann, H., (2005). Final results of the Austrian radon project. Health Phys, 89 (4): 339-348.
- Friedmann, H., Bossew, P., (2010). Selected statistical problems in spatial evaluation of Rn related variables. Nukleonika -Original Edition- 55(4), 429-432.
- Friedmann, H., Gröller, J., (2010) An approach to improve the Austrian radon potential map by Bayesian statistics. J Environ Radioact.
- Fellner, W.H., (1986). Robust Estimation of Variance Components. Technometrics, 28, 51-60.

- Kaufman, L., Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley.
- Kitto, M. E., Green, J. G., (2008). Mapping the indoor radon potential in New York at the town ship level; Atmospheric Environment 42 (34):8007-8014. doi: 10.1016/j.atmosenv.2008.06.039.
- Kropat, G., Bochud, F., Jaboyedoff, M., Laedermann, J. P., Murith, C., Palacios, M., Baechler S.,(2014). Major influencing factors of indoor radon concentrations in Switzerland; Journal of Environmental Radioactivity129:7-22. doi:10.1016/j.jenvrad.2013.11.010.
- Geraci, M., Bottai, M., (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. Biostatistics, 8 (1),140-154.
- Geraci, M., Bottai, M., (2014). Linear quantile mixed models. Statistics and Computing, 24, 461- 474.
- Gilchrist, W., (2000). Statistical Modelling with Quantile Functions. Boca Raton (FL): Chapman & Hall, CRC.
- Girault, F., Perrier, F., (2012 a). Estimating the importance of factors influencing the radon-222 flux from building walls, Sci. Total Environ., 433, 247–263.
- Girault, F., Perrier, F., (2012b) . Measuring effective radium concentration with large numbers of samples. Part 2—general properties and representativity, J. Environ. Radioact., 113, 189–202.
- Gray, R. J., (1992). Flexible models for analyzing survival data using splines, with applications to breast cancer progno- sis. Journal of the American Statistical Association 87, 942-951.
- Gray, R. J., (1994). Spline-based test in survival analysis. Biometrics 50, 640-652.
- Green, P. J., Silverman, B.W., (1994). Nonparametric Regression and Generalized Linear Models. London: Chapman & Hall.
- Groves-Kirkby, C.J., Denman, A.R., Phillips, P.S.,(2009). Lorenz Curve and Gini Coefficient: novel tools for analysing seasonal variation of environmental radon gas, J. Environ. Manage., 90, 2480–2487.
- Gue, L., (2015). David Suzuki Foundation Report: Revisiting Canada's Radon Guideline(ISBN digital: 978–1-897375-91-4 print: 978–1-897375-90-7).
- Hampel, F.R., (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69, 383-393.

- Hampel, F.R., Ronchetti, E., Rousseeuw, P. J., Stahel, W., (1986) Robust Statistics: The Approach Based on Influence Functions, New York: John Wiley.
- Harding, M., Lamarche, C.,(2009). A quantile regression approach for estimating panel data models using instrumental variables. Economics Letters, 104(3), 133-135.
- Hartley, H. O., Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. Biometrika, 54, 93-108.
- Harnapp, V. R., Dollwet, H.A., Rong, S.Y., (1997). Airborne radon in homes in Summit County, Ohio: A geographic analysis; Ohio Journal of Science 97 (1):17-23.
- Hastie, T.J., Tibshirani, R.J., (1990). Generalized Additive Models, Chapman and Hall, London.
- Hastie, T.J., (1996). Pseudosplines. Journal of the Royal Statistical Society, Series B, 58, 379–396.
- Hastie, T.J., Tibshirani, R.J., (1990a). Exploring the nature of covariate effects in the proportional hazards model. Biometrics 46, 1005-1016.
- Hastie, T. J., Tibshirani, R. J., (1990b). Generalized Addi-tive Models. London: Chapman and Hall.
- Harville A.D., (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association, 72, 320-338.
- He, X., (1997). Quantile curves without crossing, Amer. Statist. 51, 186–192.
- Hodgson, J., Carey, S., Scanlon, R., (2014). Developing a New National Radon Risk Map, (Dublin, Ireland)
- Huber, P.J., (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35 (1), 73-101.
- Huber, P.J., (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. The Annals of Statistics, 1(5), 799-821.
- Huber, P.J., (1981). Robust Statistics. John Wiley & Sons, NewYork.
- Huggins, R.M., (1993). A Robust Approach to the Analysis of Repeated Measures. Biometrics, 49, 255-268.
- Huggins, R.M., Loesch, D.Z., (1998). On the Analysis of Mixed Longitudinal Growth Data. Biometrics. 54(2),583-95.
- Ielsch, G., Cushing, M.E., Combes, P., Cuney, M., (2010) Mapping of the geogenic radon potential in France to improve radon risk management: methodology and first application to region Bourgogne. J Environ Radioact., 101, 813–820.

- Jung, S.H., 1996. Quasi-likelihood for median regression models. Journal of the American Statistical Association 91, 251–257.
- Jureckova, J., Sen, P.K., (1996). Robust statistical procedures: asymptotics and interrelations. John Wiley & Sons
- Kang, S. J. (1993). Root Multiplicities of the Hyperbolic Kac-Moody Lie Algebra HA(1)1. Journal of Algebra. 160 (2), 492-523.
- Kelly, C. Rice , J., (1990). Monotone smoothing with application to dose response curves and the assessment of synergism. Biometrics 46, 1071-1085.
- Kemski, J., Klinger, R., Siehl, A., Valdivia-Manchelo, M., 2009. From radon hazard to risk prediction-based on geological maps, soil gas and indoor measurements in Germany. Environ. Geol. 56, 1269-1279.
- Kitanidis, P.K., (1997). Introduction to Geostatistics: Applications in Hydrogeology. Cambridge University Press.
- Koenker, R. (2004). Quantile regression for longitudinal data. Journal of Multivariate Analysis, 91, 74-89.
- Koenker, R., Bassett, G., (1978). Regression quantiles. Econometrica, 46, 33-55.
- Koenker, R., D'Orey, V., (1987). Computing regression quantiles. Applied Statistics,36, 383-393.
- Koenker, R., Bilias, Y., (2001). Quantile regression for duration data: A reappraisal of the pennsylvania reemployment bonus experiments. Empirical Economics, 26(1),199-220.
- Koenker, R., Hallock, K.F., (2001). Quantile Regression. Journal of Economic Perspectives,15(4),143-156.
- Koenker, R., (2004). Quantile regression for longitudinal data. Journal of Multivariate Analysis, 91(1), 74-89.
- Kokic, P., Chambers, R., Breckling, J., Beare, S., (1997). A measure of production performance. Journal of Business and Economic Statistics, 15, 445-451.
- Kreft Ita G.G., De Leeuw J., (1998). Introducing Multilevel Modeling, Sage, London.
- Lamarche, C., (2010). Robust penalized quantile regression estimation for panel data. Journal of Econometrics, 157(2), 396-408.
- Lecoutre, J. P., Tassi, P., (1987). Robust and nonparametric statistics. Economica, Paris, 455.
- Lévesque, B., Gauvin, D., McGregor, R.G., Martel, R., Gingras, S., Dontigny, A., Walker, W.B., Lajoie, P., Létourneau, E., (1997). Radon in residences: Influences of geological and housing characteristics. Health Phys, 72, 907–914.

- Liu, Y., Bottai, M., (2009). Mixed-Effects Models for Conditional Quantiles with Longitudinal Data. The International Journal of Biostatistics, 5(1), articolo 28.
- Loy A., Hofmann, H., (2014). HLMdiag: Diagnostic Tools for Hierarchical (Multilevel) Linear Models in R. Journal of Statistical Software, 56 (5),1-28. URL http://www.jstatsoft.org/v56/i05/.
- Longford, N. T., (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. ETS Research Report Series, 1987 (1), i–26.
- Longford, N. T., (1993). Random coeficient models. Oxford: Clarendon Press.
- Machado J. A. F., Santos Silva J. M. C., (2005). Quantiles for counts. Journal of the American Statistical Association, 100(472),1226-1237.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., P. Roudier. Package 'cluster', (2016).
- Marino, M. F., Farcomeni, A., (2015). Linear quantile regression models for longitudinal experiments: an overview, METRON, 73, 229-247.
- Miles, J. C. H., (2001). Temporal variation of radon levels in houses and implications for radon measurement strategies Radiat. Prot. Dosim. 93 369–75
- Miles, J.C.H., Appleton, J.D., (2005). Mapping variation in radon potential both between
- and within geological units. J. Radiolog. Prot. 5, 256-276.
- Minda, M., Tóth, G., Horváth, I., Barnet, I., Hámori, K., Tóth, E., (2009). Indoor radon mapping and its relation to geology in Hungary. Environmental Geology 57(3), 601-609.
- Mosteller, F. and Tukey, J. (1977). Data Analysis and Regression. Addison-Wesley. of Statistics, 37, 381-399.
- Newey, W., Powell, J., (1987). Asymmetric Least Squares Estimation and Testing. Econometrica, 55(4), 819-47.
- Nychka, D., Saltzman, N., (1998). Design of air quality monitoring networks. In Nychka, Douglas, Piegorsch, Walter W. and Cox, Lawrence H. (eds), Case studies in environmental statistics
- Opsomer J.D., Claeskens G., Ranalli M.G., Kauermann G., Breidt F. J., (2008). Nonparametric small area estimation using penalized spline regression, Journal of the Royal Statistical Society: Series B, 70, 265-286.
- O'Sullivan, F., (1986). A Statistical Perspective on Ill-Posed Inverse Problems (with Discussion). Statistical Science 1, 505-527.

- O'Sullivan, F., (1988). Nonparametric estimation of relative risk using splines and cross validation. SIAM Journal of Science and Statistical Computation 9, 531-542.
- Park, C.H., Jang, S.Y., Kim, S.J., Moon, J.H., (2011). Effects of bedrock type on the indoor radon concentrations at the office buildings in gyeongju, korea. Nuclear Technology & Radiation Protection 26 (3):226-232. doi: 10.2298/ntrp1103226p.
- Parker, R. L., and Rice, J. A., (1985), Discussion of "Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting" by B. W. Silverman, Journal of the Royal Statistical Society, Series B, 47, 40–42.
- Patterson, H. D., Thompson, R., (1971). Recovery of inter-block information when block sizes are unequal. Biometrika. 58 (3), 545.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. e R Core Team (2017). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131, https://CRAN.R project.org/package=nlme.
- Portnoy, S., Koenker, R., (1997). The Gaussian hare and the Laplacian tortoise: computation of squared-error vs. absolute-error estimators. Statistical Science 12, 279–300.
- Pratesi, M, Ranalli, MG, Salvati N., (2009). Nonparametric M-quantile regression using penalised splines. Journal of Nonparametric Statistics 21 (3), 287-304.
- Pratesi, M., Ranalli, M.G., Salvati, N., (2006). Nonparametric M-quantile Regression via Penalized Splines ASA Proceedings on Survey Research Methods, Alexandria, VA, pp. 3596–3603.
- Pratesi, M., Ranalli, M.G., Salvati, N., (2008). Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. Environmetrics 19, 687–701.
- Price, P.N., Nero, A.V., Gelman, A., (1996). Bayesian prediction of mean indoor radon concentrations for Minnesota counties. Health Phys, 71, 922–936.
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Requiaa, W. J., Roigb, H.L., Adamsc, M. D., Zanobettid, A., Koutrakisd, P., (2016). Mapping distance-decay of cardiorespiratory disease risk related to neighborhood environments. Environmental Research, 151, 203–215.
- Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D. (2017) Package 'MASS': Support Functions and Datasets for Venables and Ripley's MASS
- Richardson, A.M., Welsh, A.H., (1995). Robust Estimation in the Mixed Linear Model. Biometrics, 51, 1429-1439.

- Rieder, H., (1994). Robust Asymptotic Statistics. Springer, New York.
- Rousseeuw, P.J., Leroy, A.M., (1987). Robust Regression and Outlier Detection. John Wiley and Sons, New York.
- Ruppert, D., Wand, M.P., Carrol, R.J., (2003). *Semiparametric Regression*; Cambridge Series in Statistical and Probabilistic Mathematics.
- Ruppert, D., Wand, M.P., Carroll, R.J., (2009). Semiparametric regression during 2003–2007. The Electronic Journal of Statistics. 3, 1193–1256.
- Yu, K., Moyeed, R. A., (2001). Bayesian quantile regression. Statistics and Probability Letters 54, 437–47.
- Yu, K., Zhang, J., (2005). A three-parameter asymmetric Laplace distribution and its extension. Communications in Statistics-Theory and Methods 34, 1867–79.
- Yu, K., Lu, Z., Stander, J.,(2003). Quantile regression: Applications and current research areas. Journal of the Royal Statistical Society Series D:The Statistician, 52(3):331350.
- Salvati, N., Ranalli, M. G., Pratesi, M., (2011). Small area estimation of the mean using non-parametric M-quantile regression: A comparison when a linear mixed model does not hold. Journal of Statistical Computation and Simulation, 81(8), 945–964.
- Sarra, A., Fontanella, L., Valentini, P., Palermi, S., (2016). Quantile regression and Bayesian cluster detection to identify radon prone areas. Journal of Environmental Radioactivity,164, 354–364.
- Searle, S.R., Casella, G., McCulloch, C.E., (1992). Variance components. New York: Wiley.
- Simonoff, J. (1996), Smoothing Methods in Statistics, Springer, New York.
- Sinha, S.K., Rao, J.N.K., (2009). Robust Small Area Estimation. The Canadian Journal of Statistics 37 (3), 381–399.
- Smith, B. J., Cowles, M. K., (2007). Correlating point referenced radon and areal uranium data arising from a common spatial process. Journal of the Royal Statistical Society Series C, 56, 313–326.
- Staudte, R.G., Sheather, S.J., (1990). Robust Estimation and Testing. Wiley, New York.
- Stone, C.J., (1986). The dimensionality reduction principle for generalized additive models. Ann. Statist.,14, 590–606.
- Street, J.O., Carroll, R.J., Ruppert, D., (1988). A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares. The American Statistician, 42, (2),152-154.

- Sundal, A.V., Henriksen, H., Soldal, O., Strand, T., (2004). The influence of geological factors on indoor radon concentrations in Norway;Science of the Total Environment 328 (1-3):41-53. doi: 10.1016/j.scitotenv.2004.02.011.
- Takeuchi, I., Le, Q.V.T., Sears, D., Smola, A.J., (2005). Nonparametric quantile regression, J. Mach. Learn. Res. 7,1001–1032.
- Tapia, R., Kanevski, M., Maignan, M., Gruson, M., (2006): Comprehensive multivariate analysis of indoor radon data in Switzerland, 8th International Workshop "Geological aspects of radon risk mapping", Prague, 26–30 September.
- Torabi, M., Shokoohi, F., (2015). Non-parametric generalized linear mixed models in small area estimation. Can. J. Statistics, 43: 82–96. doi:10.1002/cjs.11236
- Tzavidis N, Salvati N, Schmid T, Flouri E, Midouhas E. (2015). Longitudinal analysis of the Strengths and Difficulties Questionnaire scores of the Millennium Cohort Study children in England using M-quantile random effects regression. Journal of Royal Statistical Society Series A, 179, 427-452.
- Tzavidis, N., Marchetti, S., Chambers R., (2010). Robust estimation of small area means and quantiles. Australian and New Zealand Journal of Statistics 52 (2), 167–186.
- Ugarte, M.D., Goicoa, T., Militino, A.F., Durban, M., (2009). Spline smoothing in small area trend estimation and forecasting. Computational Statistics and Data Analysis 53, 3616–3629
- Vaupotic, J., Kobal, I., Krizman, M.J., (2010). Background outdoor radon levels inSlovenia. Nukleonika 55, 579-582
- Wahba, G., (1990). Spline Models for Observational Data, Philadelphia: SIAM.
- Wand M (2014). SemiPar: Semiparametic Regression. R package version 1.0-4.1, URL http: //CRAN.R-project.org/package=SemiPar.
- Wand, M.P., Ormerod, J.T., (2008). On semiparametric regression with OSullivan penalized splines. Australian and New Zealand Journal of Statistics 50, 179-198.
- Wasserman, L. (2006), All of Nonparametric Statistics, Springer, New York.
- Wand, M.P., 2003. Smoothing and mixed models. Computational Statistics 18, 223–249.
- Wood, S., 2003. Thin plate splines regression. Journal of the Royal Statistical Society. Series B 65, 95–114.
- Wood, S., 2006. On confidence intervals for generalized additive models based on penalized regression splines. Australian and New Zealand Journal of Statistics 48, 445–464.
- WHO (2009). WHO handbook on indoor radon: a public health perspective. http://apps.who.int/iris/bitstream/10665/44149/1/9789241547673_eng.pdf

WHO (2010). WHO handbook on indoor radon: a public health perspective. World Health Organization, France (2010)

Zhua, H.C., Charleta, J.M., Poffijnb, A., (2001). Radon risk mapping in southern Belgium: an application of geostatistical and GIS techniques. Science of The Total Environment, 272, (1–3), 203-210.

APPENDICE

Appendice A

I quantili univariati sono definiti come particolari indici di posizione della distribuzione, cioè il θ-th quantile è il valore di y tale che $P(Y \leq y) = \theta$. Partendo dalla funzione di distribuzione cumulata (CDF)

$$F_Y(y) = F(y) = P(Y \leq y)$$

la funzione quantile è definita come la sua inversa

$$Q_Y(\theta) = Q(\theta) = F_Y^{-1}(\theta) = inf\{y: F(y) > \theta\}$$

per $\theta \in (0,1)$. Se $F(.)$è una funzione strettamente crescente e continua, allora $F^{-1}(\theta)$ è l'unico numero reale y tale che $F(y) = \theta$ (Gilchrist 2000).

Meno comune è la presentazione dei quantili come particolari centri di distribuzione c, ottenuti minimizzando la somma ponderata delle deviazioni assolute (Hao and Naiman 2007). In una tale prospettiva il quantile θ-esimo è quindi:

$$q_\theta = argmin_c E[\rho_\theta(Y - c)]$$

dove $\rho_\theta(.)$ è la seguente funzione di perdita:

$$\rho_\theta(y) = [\theta - I(y < 0)]y = [(1 - \theta)I(y \leq 0) + \theta I(y > 0)]|y|$$

Tale funzione di perdita è quindi una funzione di perdita assoluta asimmetrica, la quale è una somma ponderata delle deviazioni assolute, in cui un peso (1 - θ) viene assegnato a deviazioni negative e un peso θ è utilizzato per le deviazioni positive.

Nel caso Y sia una variabile discreta con distribuzione di probabilità f(y)= $P(Y = y)$ il problema di minimo diviene:

$$q_\theta = argmin_c E[\rho_\theta(Y - c)] = argmin_c \left\{(1 - \theta)\sum_{y \leq c}|y - c|f(y) + \theta \sum_{y > c}|y - c|f(y)\right\}.$$

Lo stesso criterio è adottato nel caso di variabili casuali continue salvo sostituire l'operazione di sommatoria con quella di integrale:

$$q_\theta = argmin_c E[\rho_\theta(Y - c)]$$
$$= argmin_c \left\{(1 - \theta)\int_{-\infty}^{c}|y - c|f(y)\,d(y) + \theta \int_{c}^{\infty}|y - c|f(y)\,d(y)\right\}$$

Al fine di mostrare la formulazione dei quantili univariati come soluzioni di un problema di minimo (Koenker 2005), assumiamo, senza perdita di generalità, che Y è una variabile casuale continua e dividiamo in due termini il valore atteso della somma della deviazione assoluta da un dato centro c

$$E|Y-c| = \int_{y\in R} |y-c| f(y)\, d(y) =$$

$$\int_{y<c} |y-c| f(y)\, dy + \int_{y>c} |y-c| f(y)\, dy =$$

$$\int_{y<c} (c-y) f(y)\, dy + \int_{y>c} (y-c) f(y)\, dy$$

Poiché il valore assoluto è una funzione convessa , differenziando $E|Y-c|$ rispetto a c e annullando le derivate parziali otteniamo la soluzione ottima:

$$\frac{\partial}{\partial c} E|Y-c| = 0.$$

La soluzione può, allora, essere ottenuta derivando e integrando per parti.

La soluzione non cambia pre-moltiplicando le due componenti di $E|Y-c|$ rispettivamente per la costante θ e $(1-\theta)$. Questo ci permette di formulare lo stesso problema per il quantile θ generico e con la stessa strategia ottenere:

$$\frac{\partial}{\partial c} E[\rho_\theta(Y-c)] = \frac{\partial}{\partial c}\left\{(1-\theta)\int_{-\infty}^{c} |y-c| f(y)\, d(y) + \theta \int_{c}^{\infty} |y-c| f(y)\, d(y)\right\}$$

quindi

$$\frac{\partial}{\partial c} E[\rho_\theta(Y-c)] = (1-\theta)F(c) - \theta\big(1-F(c)\big) = 0$$

e poi q_θ come la soluzione del problema di minimizzazione:

$$F(c) - \theta F(c) - \theta + \theta F(c) = 0 \rightarrow F(c) = \theta \rightarrow c = q_\theta$$

L'approccio utilizzato può essere generalizzato nell'ambito dei modelli di regressione quantilica come segue. Infatti interpretando Y come variabile di risposta e X come un insieme di variabili predittive,

$$\widehat{q_Y}(\theta, \boldsymbol{X}) = argmin_{Q_Y(\theta,\boldsymbol{X})} E[\rho_\theta(Y - Q_Y(\theta, \boldsymbol{X}))]$$

dove $Q_Y(\theta, \boldsymbol{X}) = Q_\theta[Y|X=x]$ è la funzione del quantile condizionato. Analogamente al modello lineare l'equazione diventa:

$$\hat{\beta}(\theta) = argmin_\beta E[\rho_\theta(Y - \boldsymbol{X}\beta)]$$

Per un dato valore di θ, il modello corrispondente mostra come il θ -th quantile della distribuzione condizionata di y varia al variare di x. Ad esempio per θ =0.1, l'iperpiano separa il 10% della distribuzione condizionata dal rimanente 90%.

La stima del θ^{th} β_θ è ottenuta minimizzando la funzione obiettivo

$$Q(\beta) = \sum_{i:y_i \geq x_i^T \beta_q}^{N} q\left|y_i - x_i^T \beta_q\right| + \sum_{i:y_i < x_i^T \beta_q}^{N} (1-q)\left|y_i - x_i^T \beta_q\right|$$

con metodi di programmazione lineare (PL) quali ad esempio, il metodo del simplesso e il metodo dei punti interni.

Appendice B

B.1 Robustezza

I fondatori di una teoria della robustezza completa formalizzata ed applicabile sono Huber e Hampel. Gli approcci seguiti dai due autori non sono i medesimi: Huber cerca di ottimizzare la situazione peggiore in cui ci si può trovare (approccio minimax), Hampel è invece il padre dell'approccio infinitesimale, basato sulla funzione d'influenza.

L'approccio infinitesimale si basa su tre concetti fondamentali: robustezza qualitativa, funzione d'influenza e punto di rottura. Poiché molte statistiche dipendono solamente dalla funzione di ripartizione empirica (fre) dei dati (o dalla fre e dalla numerosità campionaria *n*), queste possono essere viste come funzionali definiti nello spazio delle distribuzioni di probabilità (o sostituite da funzionali per ogni *n*), rendendo possibile l'applicazione di concetti quali continuità e derivazione.

Lo strumento dell'approccio infinitesimale è la funzione d'influenza e le varie quantità da esse derivate. Essa descrive l'effetto (approssimato e standardizzato) di un'osservazione aggiuntiva di valore x su una statistica T, dato un ampio campione estratto dalla distribuzione F. Da un punto di vista matematico, la funzione d'influenza IF(x; T,F) non è altro che la derivata direzionale in F della statistica T nella direzione data dalla distribuzione δx di Dirac.

B.2 Gli stimatori M

Gli stimatori M sono una generalizzazione degli stimatori di massima verosimiglianza proposta da Huber nel 1964. Successivamente diversi autori, quali Andrews et al. (1972), Hampel et al. (1986), Burke e Burke (1986), Lecoutre e Tassi(1987), Robusseeuw and Leroy (1987), Staudte and Sheather (1990), Rieder (1994), Jureckova and Sen (1996), Antoch et al. (1998), svilupparono e ampliarono la classe di stimatori M.

Si supponga di voler determinare lo stimatore del parametro θ nel modello parametrico $P=\{P_\vartheta, \theta \in \Theta\}$. Sia $\rho : (\Omega, \Theta) \rightarrow R$, una funzione dotata di derivata $\psi(x, \theta)=(\partial/\partial\theta)\rho(x, \theta)$. L' M-stimatore T_n si ottiene risolvendo il seguente problema di ottimizzazione:

$$\min_{\theta\in\Theta} \sum_{i=1}^{n} \rho(X_i, \theta)$$

o

$$\min_{\theta\in\Theta} E_{P_n}[\rho(X, \theta)] \quad (1)$$

in cui $\rho(.,.)$ un'idonea funzione di perdita che attribuisce un determinato peso alle *n* osservazioni.

Si osservi che, se *f*(x,θ) è la funzione di densità di P_ϑ, allora la stima di massima verosimiglianza è la soluzione di minimo di

$$\min_{\theta\in\Theta}\sum_{i=1}^{n}(-logf(X_i,\theta)).$$

Dunque, se nella (1) si pone $\rho(X_i,\theta) = -logf(X_i,\theta)$ si ottiene lo stimatore di massima verosimiglianza. Se la funzione è differenziabile in θ ed è dotata di derivata continua $\psi(\cdot,\theta) = \frac{\partial}{\partial\theta}\rho(\cdot,\theta)$, allora T_n è la radice o (sono le radici) dell'equazione:

$$\sum_{i=1}^{n}\psi(X_i,\theta) = 0$$

quindi

$$\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,T_n) = E_{P_n}[\rho(X,T_n)] = 0 \quad T_n \in \Theta \qquad (2)$$

Dalla (1) e dalla (2) discende che l'M-funzionale corrispondente a T_n è la soluzione che si ottiene minimizzando:

$$\min_{T(P)\in\Theta}\int_x \rho(x,T(P))dP(x) = E_P[\rho(X,T(P))] \qquad (3)$$

o come soluzione della seguente equazione:

$$\min_{T(P)\in\Theta}\int_x \psi(x,T(P)dP(x) = E_P[\psi(X,T(P))] \qquad (4)$$

Dove la funzione T(P) è Fisher consistente se la soluzione delle equazioni in (3) o in (4) sono univocamente determinate.

B.3 M-stimatore del parametro di posizione

Sia θ il parametro di posizione del modello da stimare e $X_1, X_2, \ldots, X_n$ osservazioni indipendenti ed identicamente distribuite $F(x-\theta)$ con $\theta \in$ R. La funzione di distribuzione F è in generale incognita; L'M-stimatore del parametro di posizione T_n è ottenuto come soluzione di minimo dell'equazione:

$$\min_{\theta\in\Theta}\sum_{i=1}^{n}\rho(X_i-\theta)$$

Se $\rho(.,.)$ è differnziabile con derivata continua $\psi(\cdot)$, allora T_n risolve l'equazione:

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0.$$

Il corrispondente M –funzionale T(F) è consistente secondo Fisher e fornisce

$$\int_x \rho(x - \theta) dP(x) = min$$

Avremmo un'unica soluzione $\theta = 0$, cioè la soluzione dell'equazione è:

$$\int_x \psi(X - \theta) dP(x) = 0$$

B.4 M-stimatori decrescenti

Gli M-stimatori decrescenti più popolari sono gli M-stimatori del tipo ψ, in cui la funzione ψ risulta non decrescente vicino all'origine ma decresce fino ad annullarsi lontano da essa.

Le loro funzioni ψ sono scelte in modo tale che ridiscendano verso zero agevolmente, in modo tale da soddisfare $\psi(x) = 0$ per tutte le *x* per cui |X| > k, dove k è il punto di minimo.

La funzione ψ non deve scendere troppo ripidamente, perché potrebbe avere una cattiva influenza sul denominatore della varianza asintotica:

$$\frac{\int \psi^2 dF}{(\int \psi' dF)^2}$$

dove F è la distribuzione di un modello mistura. L'effetto è particolarmente dannoso quando un valore negativo grande di $\psi'(x)$ si combina con un alto valore positivo di $\psi^2(x)$ ed inoltre esiste un cluster di outliers vicino ad *x*.

Un esempio di M-stimatore decrescente è quello di Hampel dove la funzione ψ è dispari e definita per ogni valore di *x* come:

$$\psi(x) = \begin{cases} x, & 0 \le |x| \le a \ \ (segmento\ centrale) \\ asign(x), & a \le |x| \le b \ \ (alto\ e\ basso\ segmento \\ \dfrac{a(k-|x|)}{k-b} sign(x), & b \le |x| \le k \quad (fine\ pendenza) \\ 0 & k \le |x| \ (coda\ di\ destra\ e\ di\ sinistra \end{cases}$$

Lo stimatore di Tukey *biweight* o biquadrato è definito a partire da funzioni ψ date, per ogni k positivo, da

$$\psi(x) = \begin{cases} x\left[\left(1-\frac{x}{k}\right)^2\right]^2, & |x| \le k \\ 0, & |x| > k \end{cases}$$

L'M-stimatore di Huber (1964) è

$$\psi(x) = \begin{cases} x, & |x| \le k \\ ksign(x), & |x| > k \end{cases}$$

B.5 La funzione d'influenza

Uno degli strumenti dell'approccio infinitesimale è la funzione d'influenza (IF), introdotta da Hampel (1974) inizialmente col nome di curva d'influenza. La IF è essenzialmente uno strumento euristico che gode di una importante interpretazione intuitiva.

Sia T una statistica (funzionale) definita su un sottoinsieme convesso di F(Ω), si dirà che T è Gateaux differenziabile in F, se esiste una funzione reale a_1 tale che per ogni G nel dominio di T vale

$$\lim_{t\to 0} \frac{T\big((1-t)F + tG\big) - T(F)}{t} = \int a_1(x)G(dx).$$

In modo equivalente ponendo F=G si può scrivere

$$\frac{\partial}{\partial t}\big[T\big((1-t)F + tG\big) - T(F)\big]_{t=0} = \int a_1(x)G(dx)$$

Ora ponendo F=G, ne discende che

$$\int a_1(x)F(dx) = 0$$

$$\frac{\partial}{\partial t}\big[T\big((1-t)F + tG\big) - T(F)\big]_{t=0} = \int a_1(x)dG(dx) = \int a_1(x)(G-F)(dx)$$

Essendo $a_1(x)$ definita solo implicitamente, non è ancora chiaro il suo significato. Se si sostituisce G con il δx di Dirac, e quest'ultimo è nel dominio di T, si può dare una formulazione più esplicita.
Formalmente possiamo dare la definizione di funzione di influenza IF:
La funzione d'influenza di T in F è data da:

$$IF(x,T,F) = \lim_{t\to 0} \frac{T\big((1-t)F + t\delta_x\big) - T(F)}{t}$$

per tutti i valori di $x \in \Omega$ per cui il limite esiste.
Si noti che la IF può essere equivalentemente definita con

$$IF(x,T,F) = \frac{\partial}{\partial t}\big[T\big((1-t)F + t\delta_x\big)\big]_{t=0}$$

Le condizioni d'esistenza della funzione d'influenza sussistono praticamente in tutte le applicazioni statistiche comuni, per cui non è in genere necessario andarle a verificare.
Esistono versioni finite della IF, quale la *sensistivity curve* (SC) di Tukey che misura l'effetto di un'osservazione aggiuntiva *x* su una statistica T_n in funzione del valore di *x*.

$$SCn-1(x) := n[T_n(x_1,\ldots,x_n-1,x) - T_n-1(x_1,\ldots,x_n-1)]$$

Se la statistica T_n è un funzionale, cioè se $T_n(x_1,\ldots,x_n) = T(F_n)$ per ogni *n*, allora

$$\mathrm{SC}_{n-1}(x) = \frac{\left[T\left(\left(1-\frac{1}{n}\right)F_{n-1} + \frac{1}{n}\delta_x\right) - T(F_{n-1})\right]}{\frac{1}{n}}$$

dove F_{n-1}è un'approssimazione di F e *t = 1/n*. In molte situazioni $\mathrm{SC}_{n-1}(x)$ covergerà a IF(x; T,F).
Vi sono alcune quantità derivate dalla IF, che ne riassumono gli aspetti rilevanti per quanto riguarda lo studio della robustezza di una statistica, e che originano altrettante nozioni di robustezza. La più importante è la *gross-error sensitivity*, che misura la sensibilità di una statistica alla presenza di outliers.
Alla *gross-error sensitivity* è legata la nozione di B-robustezza (dove B sta per *bias*), che può essere vista come robustezza rispetto agli outliers. La seconda quantità derivata dalla IF è la *local-shift sensistivity*, che misura la sensibilità di una statistica ad approssimazioni dovute agli arrotondamenti dei valori.
La funzione di influenza per uno stimatore di tipo M è esprimibile come:

$$IF(x,T,F) = M(\theta)^{-1}\psi(X_i,\theta)$$

$$M(\theta) = -\int \frac{\partial}{\partial\theta}\psi(X_i,\theta)$$

Pertanto la IF di uno stimatore di tipo M è proporzionale alla funzione di stima che lo definisce ($\psi(.,\theta)$). L'indice di sensibilità rispetto ai grandi errori è finito solo se $\psi(.,\theta)$ è limitata (nell'ipotesi che $M(\theta) \neq 0$) Allora, ne consegue che uno stimatore di tipo M è B-robusto se la funzione che lo definisce è limitata, cosa che per gli stimatori di massima verosimiglianza accade raramente.
La IF di un funzionale T(F) è direttamente collegata alla sua varianza asintotica tramite la seguente relazione:

$$V(T,F) = E_\theta[IF(x,T,F)^2]$$

Prendendo come riferimento la IF di uno stimatore di tipo M, si ottiene la sua varianza asintotica:
se dim(θ)=1

$$V(\tilde{\theta}) = \frac{Q(\theta)}{M(\theta)^2}$$

con $Q(\theta) = \int \psi(X_i,\theta)dF$
se dim(θ)>1

$$V(\tilde{\theta}) = M(\theta)^{-1}Q(\theta)M(\theta)^{-T}, \quad \text{con } Q(\theta) = E_\theta[\psi(X_i,\theta)\psi(X_i,\theta)^T]$$

Sotto condizioni di regolarità si ha inoltre che $(\tilde{\theta})$è consistente e asintoticamente normale:
Riassumendo, gli stimatori di tipo M sono uno strumento utile per ottenere stime robuste dei parametri del modello oggetto di studio. Essi vengono usati anche nelle regressioni robuste.

Appendice C

1.C Algoritmo di stima dei parametri

1° step Inizializzazione dell'algoritmo

Assegniamo i valori di inizializzazione alle tre componenti di varianza, i quali sono ottenuti applicando il modello misto P-spline (funzione `lme` del pacchetto `nlme` di `R`).

2°step

Dati i parametri di varianza costruiamo la matrice di covarianza $\mathbf{V}_q$ e stimiamo $\boldsymbol{\beta}_{\psi q}$ risolvendo iterativamente l'equazione

$$\boldsymbol{\beta}_{\psi q}{}^{t+1} = \boldsymbol{\beta}_{\psi q}{}^{t} + \left\{\mathbf{X}^T\mathbf{U}_q^{-\frac{1}{2}}\mathbf{D}_q(\boldsymbol{\beta}_{\psi q}{}^{t})\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{X}\right\}^{-1}\mathbf{X}^T\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\mathbf{r}_q)$$

dove $\mathbf{D}_q(\boldsymbol{\beta}_{\psi q}{}^{t})$ è una matrice diagonale con il *j*-esimo elemento diagonale $D_{ijq} = \frac{\partial\psi_{ijq}(r_{ijq})}{\partial(r_{ijq})}$

3°step

Utilizziamo la stima dei coefficienti per ottenere quella delle componenti di varianza. L'utilizzo del metodo di stima del punto fisso iterativo richiede che le equazioni di stima siano riformulate come segue

$$\psi_q(\mathbf{r}_q)^T\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^T\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\mathbf{r}_q) - K_{2q}tr\left\{\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^T\mathbf{V}_q^{-1}\left(\mathbf{Z}\mathbf{Z}^T\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{I}_n\right)\begin{pmatrix}\sigma_{uq}^2\\ \sigma_{\gamma q}^2\\ \sigma_{eq}^2\end{pmatrix}\right\} = 0$$

$$\psi_q(\mathbf{r}_q)^T\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\mathbf{r}_q) - K_{2q}tr\left\{\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{V}_q^{-1}\left(\mathbf{Z}\mathbf{Z}^T\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{I}_n\right)\begin{pmatrix}\sigma_{uq}^2\\ \sigma_{\gamma q}^2\\ \sigma_{eq}^2\end{pmatrix}\right\} = 0$$

$$\psi_q(\mathbf{r}_q)^T\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{I}_n\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\mathbf{r}_q) - K_{2q}tr\left\{\mathbf{V}_q^{-1}\mathbf{I}_n\mathbf{V}_q^{-1}\left(\mathbf{Z}\mathbf{Z}^T\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{I}_n\right)\begin{pmatrix}\sigma_{uq}^2\\ \sigma_{\gamma q}^2\\ \sigma_{eq}^2\end{pmatrix}\right\} = 0$$

L'algoritmo del punto fisso per l'equazioni di stima alla *t*-esima iterazione è

$$\begin{pmatrix}\sigma_{uq}^{2(t+1)}\\ \sigma_{\gamma q}^{2(t+1)}\\ \sigma_{eq}^{2(t+1)}\end{pmatrix} = \left[A\begin{pmatrix}\sigma_{uq}^{2(t)}\\ \sigma_{\gamma q}^{2(t)}\\ \sigma_{eq}^{2(t)}\end{pmatrix}\right]^{-1} a\begin{pmatrix}\sigma_{uq}^{2(t)}\\ \sigma_{\gamma q}^{2(t)}\\ \sigma_{eq}^{2(t)}\end{pmatrix}$$

dove A e a sono definite come segue:

$$A\begin{pmatrix}\sigma^2_{uq}\\ \sigma^2_{\gamma q}\\ \sigma^2_{eq}\end{pmatrix}=\begin{pmatrix} K_{2q}tr[\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\boldsymbol{T}}\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\boldsymbol{T}}] & K_{2q}tr[\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\boldsymbol{T}}\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T] & K_{2q}tr[\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\boldsymbol{T}}\mathbf{V}_q^{-1}\mathbf{I}_n] \\ K_{2q}tr[\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\boldsymbol{T}}]K_{2q}tr[\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T]K_{2q}tr[\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{V}_q^{-1}\mathbf{I}_n] \\ K_{2q}tr[\mathbf{V}_q^{-1}\mathbf{I}_n\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\boldsymbol{T}}] & K_{2q}tr[\mathbf{V}_q^{-1}\mathbf{I}_n\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T] & K_{2q}tr[\mathbf{V}_q^{-1}\mathbf{I}_n\mathbf{V}_q^{-1}\mathbf{I}_n] \end{pmatrix}$$

$$a\begin{pmatrix}\sigma^2_{uq}\\ \sigma^2_{\gamma q}\\ \sigma^2_{eq}\end{pmatrix}=\begin{pmatrix} \frac{1}{2}\psi_q(\mathbf{r}_q)^T\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{Z}\mathbf{Z}^{\boldsymbol{T}}\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\mathbf{r}_q) \\ \frac{1}{2}\psi_q(\mathbf{r}_q)^T\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{Z}_{sp}\mathbf{Z}_{sp}^T\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\mathbf{r}_q) \\ \frac{1}{2}\psi_q(\mathbf{r}_q)^T\mathbf{U}_q^{\frac{1}{2}}\mathbf{V}_q^{-1}\mathbf{I}_n\mathbf{V}_q^{-1}\mathbf{U}_q^{\frac{1}{2}}\psi_q(\mathbf{r}_q) \end{pmatrix}$$

4°step

Si itera lo step 2 e 3 fino a convergenza.

Una volta ottenute la stima dei coefficenti fissi e delle componenti di varianza questi saranno utilizzati per ottenere le stime dei coefficienti random del modello $\boldsymbol{u}_q$ e $\boldsymbol{\gamma}_q$

$$\mathbf{Z}^{\mathrm{T}}\boldsymbol{\Sigma}_{eq}^{-1/2}\psi_q\{\boldsymbol{\Sigma}_{eq}^{-1/2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_q-\boldsymbol{Z}\boldsymbol{u}_q-\boldsymbol{Z}_{sp}\boldsymbol{\gamma}_q)\}-\boldsymbol{\Sigma}_{uq}^{-\frac{1}{2}}\psi_q\left\{\boldsymbol{\Sigma}_{uq}^{-\frac{1}{2}}\boldsymbol{u}_q\right\}=0$$

$$\boldsymbol{Z}_{sp}^T\boldsymbol{\Sigma}_{eq}^{-1/2}\psi_q\{\boldsymbol{\Sigma}_{eq}^{-1/2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_q-\boldsymbol{Z}\boldsymbol{u}_q-\boldsymbol{Z}_{sp}\boldsymbol{\gamma}_q)\}-\boldsymbol{\Sigma}_{\gamma q}^{-\frac{1}{2}}\psi_q\left\{\boldsymbol{\Sigma}_{\gamma q}^{-\frac{1}{2}}\boldsymbol{\gamma}_q\right\}=0$$

Appendice D

1. Gli 88 profili abitativi per l'M-quantile centrale ordinati in ordine decrescente di IRC

Numero profilo	Età	Tipo	Contatto suolo	Litologia	Latitudine	Longitudine	IRC
36	1	1	0	5	0,5056487	0,1564925	134,7067
40	1	1	1	5	0,5056487	0,1564925	124,9629
34	1	0	0	5	0,5056487	0,1564925	124,3498
35	0	1	0	5	0,5056487	0,1564925	123,457
44	1	1	0	6	0,5056487	0,1564925	121,5316
38	1	0	1	5	0,5056487	0,1564925	114,6061
39	0	1	1	5	0,5056487	0,1564925	113,7133
33	0	0	0	5	0,5056487	0,1564925	113,1002
48	1	1	1	6	0,5056487	0,1564925	111,7879
42	1	0	0	6	0,5056487	0,1564925	111,1748
43	0	1	0	6	0,5056487	0,1564925	110,282
76	1	1	0	10	0,5056487	0,1564925	109,7909
20	1	1	0	3	0,5056487	0,1564925	105,7349
28	1	1	0	4	0,5056487	0,1564925	105,1228
12	1	1	0	2	0,5056487	0,1564925	104,4354
37	0	0	1	5	0,5056487	0,1564925	103,3565
52	1	1	0	7	0,5056487	0,1564925	102,1759
68	1	1	0	9	0,5056487	0,1564925	101,7682
46	1	0	1	6	0,5056487	0,1564925	101,4311
47	0	1	1	6	0,5056487	0,1564925	100,5382
80	1	1	1	10	0,5056487	0,1564925	100,0471
41	0	0	0	6	0,5056487	0,1564925	99,92513
74	1	0	0	10	0,5056487	0,1564925	99,43404
75	0	1	0	10	0,5056487	0,1564925	98,5412
24	1	1	1	3	0,5056487	0,1564925	95,99122
32	1	1	1	4	0,5056487	0,1564925	95,37912
18	1	0	0	3	0,5056487	0,1564925	95,37812
26	1	0	0	4	0,5056487	0,1564925	94,76602
16	1	1	1	2	0,5056487	0,1564925	94,69163
19	0	1	0	3	0,5056487	0,1564925	94,48528
84	1	1	0	11	0,5056487	0,1564925	94,29737
10	1	0	0	2	0,5056487	0,1564925	94,07853
4	1	1	0	1	0,5056487	0,1564925	93,98512
27	0	1	0	4	0,5056487	0,1564925	93,87318
11	0	1	0	2	0,5056487	0,1564925	93,18568
56	1	1	1	7	0,5056487	0,1564925	92,43213
72	1	1	1	9	0,5056487	0,1564925	92,0245
50	1	0	0	7	0,5056487	0,1564925	91,81903
60	1	1	0	8	0,5056487	0,1564925	91,71168

66	1	0	0	9	0,5056487	0,1564925	91,4114
51	0	1	0	7	0,5056487	0,1564925	90,92619
67	0	1	0	9	0,5056487	0,1564925	90,51856
45	0	0	1	6	0,5056487	0,1564925	90,18141
78	1	0	1	10	0,5056487	0,1564925	89,69032
79	0	1	1	10	0,5056487	0,1564925	88,79748
73	0	0	0	10	0,5056487	0,1564925	88,18438
22	1	0	1	3	0,5056487	0,1564925	85,6344
30	1	0	1	4	0,5056487	0,1564925	85,02231
23	0	1	1	3	0,5056487	0,1564925	84,74156
88	1	1	1	11	0,5056487	0,1564925	84,55365
14	1	0	1	2	0,5056487	0,1564925	84,33481
8	1	1	1	1	0,5056487	0,1564925	84,24141
31	0	1	1	4	0,5056487	0,1564925	84,12946
17	0	0	0	3	0,5056487	0,1564925	84,12846
82	1	0	0	11	0,5056487	0,1564925	83,94055
2	1	0	0	1	0,5056487	0,1564925	83,62831
25	0	0	0	4	0,5056487	0,1564925	83,51636
15	0	1	1	2	0,5056487	0,1564925	83,44197
83	0	1	0	11	0,5056487	0,1564925	83,0477
9	0	0	0	2	0,5056487	0,1564925	82,82887
3	0	1	0	1	0,5056487	0,1564925	82,73546
54	1	0	1	7	0,5056487	0,1564925	82,07531
64	1	1	1	8	0,5056487	0,1564925	81,96796
70	1	0	1	9	0,5056487	0,1564925	81,66768
58	1	0	0	8	0,5056487	0,1564925	81,35486
55	0	1	1	7	0,5056487	0,1564925	81,18247
71	0	1	1	9	0,5056487	0,1564925	80,77484
49	0	0	0	7	0,5056487	0,1564925	80,56937
59	0	1	0	8	0,5056487	0,1564925	80,46202
65	0	0	0	9	0,5056487	0,1564925	80,16174
77	0	0	1	10	0,5056487	0,1564925	78,44066
21	0	0	1	3	0,5056487	0,1564925	74,38474
86	1	0	1	11	0,5056487	0,1564925	74,19683
6	1	0	1	1	0,5056487	0,1564925	73,88459
29	0	0	1	4	0,5056487	0,1564925	73,77264
87	0	1	1	11	0,5056487	0,1564925	73,30398
13	0	0	1	2	0,5056487	0,1564925	73,08515
7	0	1	1	1	0,5056487	0,1564925	72,99174
81	0	0	0	11	0,5056487	0,1564925	72,69088
1	0	0	0	1	0,5056487	0,1564925	72,37864
62	1	0	1	8	0,5056487	0,1564925	71,61115
53	0	0	1	7	0,5056487	0,1564925	70,82565
63	0	1	1	8	0,5056487	0,1564925	70,7183
69	0	0	1	9	0,5056487	0,1564925	70,41802

57	0	0	0	8	0,5056487	0,1564925	70,1052
85	0	0	1	11	0,5056487	0,1564925	62,94716
5	0	0	1	1	0,5056487	0,1564925	62,63492
61	0	0	1	8	0,5056487	0,1564925	60,36148

2. Gli 88 profili abitativi per l'M-quantile 0.25 ordinati in ordine decrescente di IRC

Numero profilo	Età	Tipo	Contatto suolo	Litologia	Latitudine	Longitudine	IRC
36	1	1	0	5	0,5056487	0,1564925	84,88199
44	1	1	0	6	0,5056487	0,1564925	81,4091
40	1	1	1	5	0,5056487	0,1564925	79,15168
34	1	0	0	5	0,5056487	0,1564925	78,52376
35	0	1	0	5	0,5056487	0,1564925	76,88534
48	1	1	1	6	0,5056487	0,1564925	75,67878
42	1	0	0	6	0,5056487	0,1564925	75,05086
76	1	1	0	10	0,5056487	0,1564925	73,95135
43	0	1	0	6	0,5056487	0,1564925	73,41244
38	1	0	1	5	0,5056487	0,1564925	72,79344
52	1	1	0	7	0,5056487	0,1564925	71,81604
39	0	1	1	5	0,5056487	0,1564925	71,15503
33	0	0	0	5	0,5056487	0,1564925	70,5271
20	1	1	0	3	0,5056487	0,1564925	70,03827
28	1	1	0	4	0,5056487	0,1564925	69,3764
46	1	0	1	6	0,5056487	0,1564925	69,32055
12	1	1	0	2	0,5056487	0,1564925	68,38311
80	1	1	1	10	0,5056487	0,1564925	68,22104
68	1	1	0	9	0,5056487	0,1564925	67,71896
47	0	1	1	6	0,5056487	0,1564925	67,68213
74	1	0	0	10	0,5056487	0,1564925	67,59312
41	0	0	0	6	0,5056487	0,1564925	67,05421
56	1	1	1	7	0,5056487	0,1564925	66,08572
75	0	1	0	10	0,5056487	0,1564925	65,9547
50	1	0	0	7	0,5056487	0,1564925	65,4578
37	0	0	1	5	0,5056487	0,1564925	64,79679
24	1	1	1	3	0,5056487	0,1564925	64,30796
4	1	1	0	1	0,5056487	0,1564925	63,85973
51	0	1	0	7	0,5056487	0,1564925	63,81938
18	1	0	0	3	0,5056487	0,1564925	63,68004
32	1	1	1	4	0,5056487	0,1564925	63,64609
84	1	1	0	11	0,5056487	0,1564925	63,60733

26	1	0	0	4	0,5056487	0,1564925	63,01817
16	1	1	1	2	0,5056487	0,1564925	62,6528
19	0	1	0	3	0,5056487	0,1564925	62,04162
10	1	0	0	2	0,5056487	0,1564925	62,02487
72	1	1	1	9	0,5056487	0,1564925	61,98865
78	1	0	1	10	0,5056487	0,1564925	61,8628
27	0	1	0	4	0,5056487	0,1564925	61,37975
66	1	0	0	9	0,5056487	0,1564925	61,36073
45	0	0	1	6	0,5056487	0,1564925	61,3239
11	0	1	0	2	0,5056487	0,1564925	60,38646
79	0	1	1	10	0,5056487	0,1564925	60,22439
60	1	1	0	8	0,5056487	0,1564925	60,16149
54	1	0	1	7	0,5056487	0,1564925	59,72749
67	0	1	0	9	0,5056487	0,1564925	59,72231
73	0	0	0	10	0,5056487	0,1564925	59,59646
8	1	1	1	1	0,5056487	0,1564925	58,12942
55	0	1	1	7	0,5056487	0,1564925	58,08907
22	1	0	1	3	0,5056487	0,1564925	57,94972
88	1	1	1	11	0,5056487	0,1564925	57,87702
2	1	0	0	1	0,5056487	0,1564925	57,50149
49	0	0	0	7	0,5056487	0,1564925	57,46115
30	1	0	1	4	0,5056487	0,1564925	57,28786
82	1	0	0	11	0,5056487	0,1564925	57,24909
23	0	1	1	3	0,5056487	0,1564925	56,31131
14	1	0	1	2	0,5056487	0,1564925	56,29456
3	0	1	0	1	0,5056487	0,1564925	55,86308
17	0	0	0	3	0,5056487	0,1564925	55,68338
31	0	1	1	4	0,5056487	0,1564925	55,64944
70	1	0	1	9	0,5056487	0,1564925	55,63042
83	0	1	0	11	0,5056487	0,1564925	55,61068
25	0	0	0	4	0,5056487	0,1564925	55,02152
15	0	1	1	2	0,5056487	0,1564925	54,65614
64	1	1	1	8	0,5056487	0,1564925	54,43117
9	0	0	0	2	0,5056487	0,1564925	54,02822
71	0	1	1	9	0,5056487	0,1564925	53,992
77	0	0	1	10	0,5056487	0,1564925	53,86615
58	1	0	0	8	0,5056487	0,1564925	53,80325
65	0	0	0	9	0,5056487	0,1564925	53,36408
59	0	1	0	8	0,5056487	0,1564925	52,16483
6	1	0	1	1	0,5056487	0,1564925	51,77118
53	0	0	1	7	0,5056487	0,1564925	51,73084
86	1	0	1	11	0,5056487	0,1564925	51,51878
7	0	1	1	1	0,5056487	0,1564925	50,13277
21	0	0	1	3	0,5056487	0,1564925	49,95307
87	0	1	1	11	0,5056487	0,1564925	49,88037

1	0	0	0	1	0,5056487	0,1564925	49,50484
29	0	0	1	4	0,5056487	0,1564925	49,29121
81	0	0	0	11	0,5056487	0,1564925	49,25244
13	0	0	1	2	0,5056487	0,1564925	48,29791
62	1	0	1	8	0,5056487	0,1564925	48,07294
69	0	0	1	9	0,5056487	0,1564925	47,63377
63	0	1	1	8	0,5056487	0,1564925	46,43452
57	0	0	0	8	0,5056487	0,1564925	45,8066
5	0	0	1	1	0,5056487	0,1564925	43,77453
85	0	0	1	11	0,5056487	0,1564925	43,52213
61	0	0	1	8	0,5056487	0,1564925	40,07629

2. I 44 profili abitativi per l'M-quantile 0.75 ordinati in ordine decrescente di IRC

Numero profilo	**Tipo**	**Contatto Suolo**	**Litologia**	**Latitudine**	**Longitudine**	**IRC**
18	1	0	5	0,5056487	0,1564925	179,3941
22	1	0	6	0,5056487	0,1564925	175,5862
14	1	0	4	0,5056487	0,1564925	161,544
20	1	1	5	0,5056487	0,1564925	161,1527
10	1	0	3	0,5056487	0,1564925	159,1842
24	1	1	6	0,5056487	0,1564925	157,3448
6	1	0	2	0,5056487	0,1564925	156,2881
17	0	0	5	0,5056487	0,1564925	154,4072
38	1	0	10	0,5056487	0,1564925	152,255
21	0	0	6	0,5056487	0,1564925	150,5994
26	1	0	7	0,5056487	0,1564925	150,4202
30	1	0	8	0,5056487	0,1564925	146,2567
16	1	1	4	0,5056487	0,1564925	143,3026
12	1	1	3	0,5056487	0,1564925	140,9428
8	1	1	2	0,5056487	0,1564925	138,0467
13	0	0	4	0,5056487	0,1564925	136,5572
34	1	0	9	0,5056487	0,1564925	136,1877
19	0	1	5	0,5056487	0,1564925	136,1658
42	1	0	11	0,5056487	0,1564925	135,6779
2	1	0	1	0,5056487	0,1564925	135,4512
9	0	0	3	0,5056487	0,1564925	134,1973
40	1	1	10	0,5056487	0,1564925	134,0136
23	0	1	6	0,5056487	0,1564925	132,358
28	1	1	7	0,5056487	0,1564925	132,1788
5	0	0	2	0,5056487	0,1564925	131,3013
32	1	1	8	0,5056487	0,1564925	128,0153
37	0	0	10	0,5056487	0,1564925	127,2682
25	0	0	7	0,5056487	0,1564925	125,4333

29	0	0	8	0,5056487	0,1564925	121,2698
15	0	1	4	0,5056487	0,1564925	118,3158
36	1	1	9	0,5056487	0,1564925	117,9463
44	1	1	11	0,5056487	0,1564925	117,4365
4	1	1	1	0,5056487	0,1564925	117,2098
11	0	1	3	0,5056487	0,1564925	115,9559
7	0	1	2	0,5056487	0,1564925	113,0599
33	0	0	9	0,5056487	0,1564925	111,2009
41	0	0	11	0,5056487	0,1564925	110,6911
1	0	0	1	0,5056487	0,1564925	110,4644
39	0	1	10	0,5056487	0,1564925	109,0268
27	0	1	7	0,5056487	0,1564925	107,1919
31	0	1	8	0,5056487	0,1564925	103,0284
35	0	1	9	0,5056487	0,1564925	92,95949
43	0	1	11	0,5056487	0,1564925	92,44969
3	0	1	1	0,5056487	0,1564925	92,22301