Università degli Studi di Firenze
Dipartimento di Statistica "G. Parenti"
Dottorato in Statistica Applicata XI ciclo

Metodi di rappresentazione cartografica in epidemiologia geografica: l'effetto confine

Emanuela Dreassi

Relatore:
Annibale Biggeri

Correlatori:
Corrado Lagazio
Pietro Rigo

Coordinatore:
Giovanni Maria Marchetti

Indice

Introduzione **1**

1 Il problema dell'effetto confine in statistica spaziale **6**
Premessa 6
1.1 Processi spaziali definiti su reticoli 6
1.2 Modelli autoregressivi spaziali 7
1.3 L'effetto confine e la sua correzione nella stima dei modelli autoregressivi spaziali 11
Considerazioni conclusive 14

2 L'effetto confine nei metodi di stima del rischio relativo **16**
Premessa 16
2.1 Il rapporto standardizzato di mortalità 17
2.2 Il modello bayesiano empirico 18
2.3 Il modello bayesiano gerarchico 20
2.3.1 Modello scambiabile per i rischi relativi 21
2.3.2 Modello spaziale per i rischi relativi 22
2.3.3 Modello di convoluzione gaussiana per i rischi relativi . 23
2.4 Implementazione del modello bayesiano gerarchico 26
2.5 Il problema dell'effetto confine nei metodi di stima per il rischio relativo 28
Considerazioni conclusive 30

3 Metodi di correzione per l'effetto confine nel modello bayesiano gerarchico **31**
Premessa 31
3.1 Metodo dei dati pesati 33
3.2 Metodi dei dati mancanti 34
3.2.1 Algoritmo EM doppiamente stocastico 35

3.2.2 Algoritmo EM con passo M stocastico 39
3.2.3 Algoritmo Chained Data Augmentation a tre vettori . 40
3.3 Valutazione dei metodi di correzione tramite simulazioni . . . 41
Considerazioni conclusive 47

4 L'atlante di mortalità toscano **49**
Premessa 49
4.1 Fonti e materiali 49
4.2 Risultati per il tumore allo stomaco maschile 52
Considerazioni conclusive 69

Conclusioni **71**

A Correzione per l'effetto confine nei metodi che usano distanze tra eventi **74**
A.1 Processi spaziali di punto e metodi che usano distanze tra e da eventi 74
A.2 L'effetto confine e la sua correzione nei metodi che usano distanze tra e da eventi 77
A.2.1 La mappa di forma toroidale 78
A.2.2 Eliminazione di una corona interna 79
A.2.3 Estensione dell'area di studio 80
A.2.4 Schema di pesi 80

B Correzione per l'effetto confine nei modelli autoregressivi spaziali **83**
B.1 Stima di massima verosimiglianza per modelli autoregressivi spaziali gaussiani di primo ordine 83
B.2 Metodi proposti per la correzione dell'effetto confine nei modelli autoregressivi spaziali 87
B.2.1 La mappa di forma toroidale 87
B.2.2 Eliminazione di una corona interna 87
B.2.3 Costruzione di un'area esterna artificiale 88
B.2.4 Costruzione di più aree esterne artificiali 88
B.2.5 Costruzione di aree esterne usando un modello trend surface 89
B.2.6 Metodo dei dati mancanti 90
B.2.7 Modello di regressione con una variabile dummy 92
B.2.8 Metodo dei minimi quadrati generalizzati 93

C Algoritmi per il trattamento dei dati mancanti **95**
C.1 La condizione di ignorabilità 95
C.2 L'algoritmo EM 97
C.2.1 Massimizzazione della funzione di verosimiglianza . . . 98
C.2.2 Massimizzazione della a posteriori 99
C.3 L'algoritmo Monte Carlo EM, EM stocastico ed EM-tipo . . . 100
C.4 L'algoritmo poor man's data augmentation, data augmentation e chained data augmentation 101

D Gibbs sampler, adaptive rejection sampling e test di convergenza **104**
D.1 Gibbs sampler 104
D.2 Rejection sampling e adaptive rejection sampling 105
D.3 Test di convergenza di Geweke 108

E Materiale relativo all'atlante di mortalità toscano **109**

Bibliografia **120**

Indice delle tabelle

3.1 Valori degli indici TMSE, BMSE, TVAR e BVAR per i vari metodi e per i due diversi schemi spaziali 47

4.1 Stima dei coefficienti del modello CAR 57
4.2 Confronto tra le distribuzioni empiriche del rischio relativo nel comune di Firenzuola ottenute adottando il modello nelle varie versioni 66
4.3 Confronto tra le stime degli iperparametri per i diversi metodi 68
4.4 Confronto tra i diversi metodi per le aree lungo il bordo nord-est della Toscana 68
4.5 Differenze tra le diverse stime per le aree lungo il bordo nord-est della Toscana rispetto alle stime ottenute con informazione completa CO 69

E.1 Comuni della Toscana e confinanti, codici ISTAT e adiacenze . 111
E.2 Cause di morte analizzate nell'atlante di mortalità toscano . . 119

Indice delle figure

1.1 Dipendenze in uno schema autoregressivo del primo ordine definito su un reticolo regolare 10
1.2 La mappa di forma toroidale 13
1.3 La mappa espansa con una (a) o più (b) aree esterne artificiali 14

2.1 Modello scambiabile per i rischi relativi 22
2.2 Modello autoregressivo intrinseco per i rischi relativi 23
2.3 Modello di convoluzione gaussiana per i rischi relativi 26

3.1 Modello bayesiano gerarchico esteso alle aree esterne alla regione 36
3.2 Schema spaziale a *cluster* (a) ed a *trend* (b) 42
3.3 Mappe medie e mappe degli errori per lo schema spaziale a *cluster* metodo IN (a), CO (b), PE (c), MS (d) e CH (e) . . . 44
3.4 Mappe medie e mappe degli errori per lo schema spaziale a *trend* metodo IN (a), CO (b), PE (c), MS (d) e CH (e) 45

4.1 Mappa a livello comunale della Toscana 51
4.2 Focolaio del tumore allo stomaco maschile sulla mappa degli SMR, individuato in studi precedenti 53
4.3 Stima *kernel* della distribuzione del rischio relativo 54
4.4 Stime SMR (a) e EB (b) 55
4.5 Stime SE (a) e IN (b) 56
4.6 Stime CO 58
4.7 Differenze assolute tra SE e IN (a) e tra IN e CO (b) 59
4.8 Stime PE (a), MS (b), DS (c) e CH (d) 61
4.9 *Box-plot* delle distribuzioni relative alle 20 iterazioni, dell'algoritmo EM doppiamente stocastico, per il parametro di *heterogeneity* (a) e di *clustering* (b), il rischio relativo di Firenzuola (c) e di Castel del Rio (d) 62

4.10 Valore R nell'analisi delle componenti di varianza per la convergenza dell'algoritmo EM doppiamente stocastico 63
4.11 Differenze assolute tra CO e PE (a), MS (b), DS (c) e CH (d) 65
4.12 Rischio relativo nel comune di Firenzuola, metodo IN (a), CO (b), PE (c), MS (c), DS (d) e CH (e) 67

A.1 Un'interpretazione della mappa di forma toroidale 78
A.2 Determinazione dei pesi da assegnare alle distanze 81

B.1 Assegnazione dei valori per le aree esterne artificiali mediante estrapolazione di una superficie stimata usando un modello *trend surface* 89

D.1 *Upper* (- - -) e *lower* (···) *hull* 108

E.1 Mappa a livello comunale della Toscana con codifica ISTAT . 110

Introduzione

Una delle applicazioni più interessanti dell'epidemiologia geografica è la descrizione della variazione del rischio per una malattia, o causa di morte, su una data regione. I risultati ottenuti da tali studi sono utilizzati a fini descrittivi e per fornire indicazioni eziologiche, cioè per suggerire associazioni tra fattori di rischio e malattia; in questo senso tali applicazioni costituiscono il punto di partenza per la pianificazione di ulteriori analisi mirate ad investigare le cause di una bassa o elevata mortalità in alcune aree della regione esaminata. Queste ricerche hanno avuto un grande sviluppo negli ultimi anni, trovando applicazione soprattutto nello studio delle relazioni fra ambiente e morbilità (o mortalità), in particolare nell'analisi degli effetti sulla salute umana delle alterazioni dell'ambiente legate a fattori inquinanti.

Oggetto di studio sono i casi di malattia (o di mortalità) presenti in una data regione. Solitamente i dati sono disponibili in forma aggregata, per aree più o meno estese in cui viene ripartita la regione di studio; si noti che di solito la suddivisione della regione è fatta, convenientemente, in base a criteri di carattere amministrativo, dato che questi sono anche i criteri che generalmente determinano l'aggregazione dei dati. In alcuni casi, per malattie rare, si possono avere dati individuali, ossia la posizione dei singoli eventi di malattia nella regione. Sia per l'analisi di dati in forma aggregata, che per quelli individuali, è interessante l'uso di modelli propri della statistica spaziale, che permettono di considerare un'importante informazione contenuta nei dati: quella relativa alla distribuzione nello spazio del fenomeno.

Negli ultimi anni si è assistito ad un grande sviluppo di tecniche per l'analisi di dati spaziali (si vedano per esempio Cox e Isham 1980, Cliff e Ord 1981, Ripley 1981, Diggle 1983, Upton e Fingleton 1985, Haining 1990, Cressie 1991) ma poca attenzione è stata rivolta al problema del cosiddetto "effetto-confine" (*edge-effect* o *boundary problem*), ossia al fatto che l'applicazione di tali tecniche può portare a risultati affetti da gravi errori; infatti, queste metodologie, essendo state sviluppate nella teoria in "spazi infiniti",

non tengono conto dei "confini", che invece nella realtà sono sempre presenti. In particolare, nel caso di dati aggregati (*lattice data*), quando si utilizzano modelli che definiscono delle interazioni sul piano infinito, questi risultano non correttamente specificati, in quanto i dati disponibili riguardano soltanto una parte di esso; analogamente, nel caso di dati individuali (*spatial point patterns*), l'analisi può essere seriamente affetta dalla prossimità degli eventi al confine della parte di piano analizzato, dato che si può avere un'errata determinazione delle distanze, tra e da eventi, utilizzate per l'inferenza.

Per la descrizione della variabilità del rischio relativo di mortalità in una certa regione, spesso si dispone soltanto di dati in forma aggregata; in tal caso la descrizione è fatta attraverso la rappresentazione della stima del rischio relativo, per ogni area in cui è ripartita la regione, mediante cartogrammi a colori graduati raccolti in atlanti di mortalità. Questi forniscono alcune informazioni supplementari rispetto al semplice elenco dei valori stimati, riguardo alla contiguità e vicinanza delle aree ad elevato o basso rischio; l'obiettivo, infatti, è quello di ricercare andamenti spaziali particolari del rischio relativo, cioè indagare sulla causa della malattia individuando aree contigue a rischio omogeneo.

L'ampiezza delle aree in cui la regione risulta suddivisa, determina il grado di dettaglio della mappa. Naturalmente, se l'obiettivo è solo quello di dare una descrizione del fenomeno, il dettaglio può essere meno fine, quando invece si conduce una ricerca eziologica, è necessario disporre di un maggiore dettaglio, onde individuare aree sottoposte ad una particolare esposizione. I vantaggi di operare su piccole aree sono evidenti: quanto più dettagliata è la scala geografica su cui viene eseguita l'analisi, tanto maggiori sono le possibilità di svelare fenomeni interessanti, come la presenza di sorgenti di esposizione di piccola dimensione. Purtroppo, i risultati forniti da un'analisi condotta su aree di piccola dimensione sono di difficile interpretazione, questo perché i metodi classici utilizzati per la stima del rischio relativo forniscono valori fortemente instabili. Si noti, inoltre, che tali metodi non considerano la struttura spaziale dei rischi relativi, ossia non considerano un'importante informazione, cioè che individui, o gruppi di individui, residenti in aree vicine della regione, presentano, generalmente, rischi di mortalità simili, dato che sperimentano stessi fattori ambientali, possiedono medesime abitudini alimentari e più in generale hanno stili di vita simili.

L'uso della cartografia per la rappresentazione della variazione della mortalità su una regione non è una novità e per molti anni ha rappresentato il maggiore strumento per la pianificazione di interventi in ambito socio-

sanitario. Recentemente si è assistito ad un ritorno di interesse per tale disciplina, grazie anche allo sviluppo di metodi statistici, quali i modelli bayesiani empirici e il modello bayesiano gerarchico di Besag, York e Mollié (1991), che permettono sia di ottenere valide stime del rischio relativo anche per analisi condotte su aree di piccola dimensione, sia di considerare la struttura spaziale dei rischi relativi.

L'idea di questo lavoro nasce all'interno di un progetto bilaterale CNR (Consiglio Nazionale delle Ricerche) e BIOMED 2 (*European Union Biomed 2 — Disease Mapping and Risk Assessment*) ed inoltre integra una ricerca sanitaria, a cura della Regione Toscana, finalizzata alla redazione dell'atlante di mortalità toscano a livello comunale per gli anni 1971–1994.

Il presente lavoro si propone come obiettivo di analizzare in che misura le analisi che non tengono conto dell'effetto confine forniscono un'immagine fuorviante della reale distribuzione del rischio relativo sulla regione e quindi la ricerca e la valutazione di metodologie originali per operare una sua correzione. Infatti, in analisi geografiche quali la "mappatura" del rischio relativo di malattia, qualora si analizzino dati aggregati e si definisca una struttura autoregressiva spaziale per i rischi relativi (una dipendenza fra aree adiacenti), si possono avere delle distorsioni nelle stime, dato che l'informazione circa le aree adiacenti è spesso incompleta per alcune aree della regione, segnatamente quelle al confine. Se tale distorsione in alcuni casi può essere trascurabile, in altri deve essere necessariamente considerata, dato che i risultati delle analisi effettuate possono portare a formulare ipotesi eziologiche errate.

Scopo della tesi, in particolare, è proporre alcune metodologie originali per correggere per l'effetto confine il modello bayesiano gerarchico di Besag, York e Mollié, usato per la stima dei rischi relativi riportati nell'atlante di mortalità toscano.

Nella tesi viene descritto, preliminarmente, come in statistica spaziale si presenti il problema dell'effetto confine e vengono analizzate le metodologie, presenti in letteratura, per operare la sua correzione. Da queste sono state sviluppate le metodologie proposte nella tesi per effettuare la correzione nel modello bayesiano gerarchico; i metodi proposti sono poi valutati sulla base dei risultati forniti da esperimenti di simulazione e dei risultati ottenuti applicando i metodi ad un esempio reale.

In particolare, nel capitolo 1 è presentato il problema dell'effetto confine per quanto riguarda l'analisi di dati aggregati mediante modelli autoregressivi spaziali; sono inoltre illustrate e valutate le tecniche presenti in letteratura

per correggere tale effetto. Per una trattazione del problema nell'analisi di dati individuali, che non è di primario interesse ai fini del lavoro, si rimanda all'appendice A. Nel capitolo 2 sono analizzati gli usuali metodi adottati per la stima del rischio relativo di mortalità per l'analisi di dati aggregati e come in essi si presenti il problema dell'effetto confine; si noti che nessuna considerazione a riguardo è presente in letteratura. In particolare è analizzato il modello gerarchico bayesiano di Besag, York e Mollié per il quale sono proposte delle metodologie originali per operare la correzione per l'effetto confine; esse sono presentate e valutate tramite un esperimento di simulazione nel capitolo 3. Uno dei metodi proposti riguarda l'inserimento del modello in un algoritmo iterativo di imputazione di dati; in particolare sono proposti tre algoritmi originali alternativi. Per mettere in luce quale sia la genesi degli algoritmi proposti, nell'appendice C sono analizzati gli algoritmi di imputazione che sono presenti in letteratura. Le tecniche di correzione proposte sono state utilizzate per la redazione dell'atlante di mortalità toscano. Nel capitolo 4 sono presentati i risultati relativi all'applicazione dei metodi di correzione suggeriti ad un caso particolare già analizzato in passato in letteratura: la stima del rischio relativo di mortalità per tumore allo stomaco della popolazione maschile della regione Toscana nel periodo 1981–1988.

La scelta di studiare tale causa di morte è stata fatta per meglio analizzare l'effetto confine. La mortalità per tumore allo stomaco, infatti, ha rischi elevati lungo il "bordo" nord-est della regione, al confine con l'Emilia Romagna e Marche; è evidente l'interesse per una potenziale distorsione dovuta all'effetto confine nelle stime dei rischi relativi, quando si ha un aumento del rischio proprio nelle aree poste lungo il confine. Inoltre, nonostante la forte diminuzione del rischio di mortalità per tumore allo stomaco in entrambi i sessi, questa rimane la causa di mortalità che in Toscana presenta un rischio più elevato rispetto alle altre regioni italiane; è quindi interessante individuare, nel modo più preciso possibile, quale siano le zone che presentano i rischi più elevati, in modo da indagare le possibili cause.

Colgo l'occasione per ringraziare Annibale Biggeri per avere guidato e stimolato l'intero lavoro, Corrado Lagazio, Marco Marchi, Mariangela Vigotti ed Alessandra Petrucci per i suggerimenti e la disponibilità a discutere i problemi inerenti alle mappature di rischio, Geoffrey McLachlan per la competenza riguardo gli algoritmi per il trattamento di dati mancanti e Andrew Gelman, Sonia Petrone, Steve Brooks e Rossella Berni per gli utili consigli relativi alla convergenza degli algoritmi proposti. Un infinito grazie va a

Pietro Rigo e Giorgio Calzolari per i loro consigli ed il loro costante incoraggiamento. Un doveroso grazie a tutto il Dipartimento di Statistica "G. Parenti" dell'Università di Firenze ed al Dipartimento di Matematica della *Queensland University* di Brisbane (Australia) che mi hanno ospitato durante il lavoro. Grazie, infine, a Bruno Chiandotto e Giovanni Maria Marchetti che hanno coordinato il corso di dottorato ed a Laura Neri, Anna Gottard, Angela D'Elia, Antonella Zanobetti, Antonella D'Agostino ed Emilia Rocco che hanno vissuto, insieme a me, questa esperienza.

Capitolo 1

Il problema dell'effetto confine in statistica spaziale

Premessa

In questo capitolo viene analizzata la letteratura esistente sul problema dell'effetto confine nelle tecniche adottate per l'analisi di dati spaziali. L'attenzione è rivolta ai processi autoregressivi spaziali, è infatti in tale ambito che sono state sviluppate le ricerche e le proposte per effettuare la correzione per l'effetto confine nel caso di dati aggregati.

In particolare, nel paragrafo 1.1 è definito un processo spaziale stazionario e isotropico, assunzione di base per le tecniche analizzate, e vengono fornite delle nozioni sui processi definiti su reticoli. Nel paragrafo 1.2 sono presentati i modelli autoregressivi spaziali di primo ordine; il problema dell'effetto confine in tali modelli e le proposte fatte in letteratura per la sua correzione sono analizzati nel paragrafo 1.3.

Poiché tali proposte derivano dai metodi suggeriti per operare la stessa correzione nei metodi che usano distanze tra e da eventi usati per l'analisi di dati individuali, il problema dell'effetto confine cui sono soggetti, e quindi le correzioni proposte, sono presentate, per completezza, nel paragrafo A.1 e nel paragrafo A.2 dell'appendice A.

1.1 Processi spaziali definiti su reticoli

Si definisce processo stocastico un insieme di variabili casuali indicizzato da un insieme S

$$\{X(\boldsymbol{s}) \mid \boldsymbol{s} \in S\}.$$

Generalmente $S \subseteq \mathbb{R}^d$, in particolare, quando si è interessati ai processi sul piano, si considera $d = 2$, cioè processi spaziali in due dimensioni con $S \subseteq \mathbb{R}^2$; si veda per esempio, per tali definizioni, Ripley (1981) e Cressie (1991). Un processo si dice stazionario (od omogeneo) se invariante rispetto a traslazioni, cioè la sua distribuzione è inalterata quando l'origine dell'insieme degli indici viene traslato; se l'invarianza vale anche rispetto a rotazioni intorno all'origine, si parla di processo isotropico.

Se l'insieme degli indici S è un insieme numerabile, cioè $S \subseteq \mathbb{N}$ (ogni indice rappresenta ad esempio un'area in cui i dati sono raccolti in forma aggregata) l'insieme S è detto reticolo (*lattice*). Si può definire un reticolo infinito come

$$S \equiv \{i : i = 1, \ldots\}. \tag{1.1}$$

Esso è corredato da un'informazione sulle adiacenze, determinate in base a confini geografici, distanze o altre funzioni; per ogni $i = 1, \ldots$, si ha

$$W_i \equiv \{j : j \text{ è adiacente ad } i, \text{ per } j = 1, \ldots\}. \tag{1.2}$$

L'insieme costituito dagli indici i (nel caso analizzato aree) e dalle loro adiacenze W_i (le aree che confinano con ciascuna area considerata) è detto reticolo spaziale (*spatial lattice*) ed è definito come

$$SW \equiv \{(i, W_i) : i = 1, \ldots\}.$$

Esso può essere rappresentato, teoricamente, mediante un grafo, dove gli indici (cioè le aree dove vengono rilevate le osservazioni in forma aggregata) sono rappresentate da nodi e la struttura delle adiacenze da archi (l'esistenza di un arco che congiunge due nodi indica l'adiacenza tra le rispettive aree). I dati aggregati possono essere considerati come una realizzazione di un processo spaziale definito su un reticolo finito. Si noti che un reticolo può essere regolare o irregolare, ed il processo definito sul reticolo può essere discreto o continuo.

1.2 Modelli autoregressivi spaziali

I processi spaziali definiti su un reticolo regolare a variabile continua sono la diretta estensione, a due dimensioni, dei processi temporali, usati quale modello per le serie temporali; per essi, quindi, si utilizzano strumenti che derivano da quelli adottati per lo studio di queste. Quando si analizza la struttura autoregressiva del processo, si parla di modelli autoregressivi

spaziali; l'estensione dai processi temporali a quelli spaziali, quindi il passaggio da una a due dimensioni, non è però immediata, dato che si tratta di modellare dipendenze non più unidirezionali ma multidirezionali.

Quando si considerano dei modelli autoregressivi di primo ordine per una serie temporale X_t, con $t = \ldots, -1, 0, 1, \ldots$, il meccanismo autoregressivo può essere espresso mediante tre diversi modelli (sempre mantenendo il coefficiente di autocorrelazione $\mid \rho \mid < 1$ affinché sia soddisfatta la condizione di stazionarietà):

- il modello autoregressivo simultaneo

$$X_t = \rho X_{t-1} + \epsilon_t,$$

  con media $\mathrm{E}(X_t) = 0$ e $\mathrm{E}(\epsilon_t) = 0$, varianza $\mathrm{var}(\epsilon_t) = \sigma^2$, e covarianza $\mathrm{cov}(\epsilon_t, \epsilon_{t-s}) = 0$ e $\mathrm{cov}(X_{t-s}, \epsilon_t) = 0$ per $s > 0$;

- il modello autoregressivo condizionale

$$\mathrm{E}(X_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots) = \rho x_{t-1}$$

  e

$$\mathrm{var}(X_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots) = \sigma^2;$$

- il modello covarianza

$$\mathrm{E}(X_t) = 0 \qquad \text{e} \qquad \mathrm{cov}(X_t, X_{t-s}) = \frac{\sigma^2}{1-\rho^2}\rho^{|s|}.$$

Anche nel caso di processi spaziali si possono dare diverse formulazioni del meccanismo autoregressivo; in questo paragrafo sono illustrati i modelli autoregressivi spaziali simultanei e condizionali, sempre limitatamente a quelli del primo ordine.

Whittle (1954) propone il seguente modello autoregressivo simultaneo (SAR, *Simultaneous Autoregressive*)

$$X_i = \sum_{j \neq i} g_{ij} X_j + \epsilon_i,$$

per $i = 1, \ldots, n$, dove gli ϵ_i rappresentano dei termini di errore incorrelati con $\mathrm{E}(\epsilon_i) = 0$ e $\mathrm{var}(\epsilon_i) = \sigma_i^2$ e g_{ij} definisce la struttura autoregressiva del modello.

Bartlett (1971) e Besag (1974) definiscono il seguente modello autoregressivo condizionale (CAR, *Conditional Autoregressive*)

$$\mathrm{E}(X_i \mid \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) = \sum_{j \neq i} g_{ij} x_j \quad \text{e} \quad \mathrm{var}(X_i \mid \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) = \sigma_i^2,$$

per $i = 1, \ldots, n$, con $\boldsymbol{X}_{-i}$ e $\boldsymbol{x}_{-i}$ che indicano, rispettivamente, il vettore di variabili casuali $\boldsymbol{X}$ e il vettore dei valori osservati $\boldsymbol{x}$ senza considerare la componente i-esima; il termine g_{ij} definisce ancora la struttura autoregressiva del modello.

In entrambi i modelli, il termine g_{ij} può essere definito come ρw_{ij}, con ρ il coefficiente di autocorrelazione spaziale e w_{ij} l'elemento generico della matrice di adiacenze $\boldsymbol{W}$ usata per descrive l'insieme delle adiacenze delle aree del reticolo; ogni elemento w_{ij} della matrice è unitario se le aree i e j sono adiacenti, nullo in caso contrario.

La stima di massima verosimiglianza del parametro di autocorrelazione spaziale ρ, per un modello autoregressivo spaziale gaussiano di primo ordine, si ottiene massimizzando il profilo di verosimiglianza; per il modello nella formulazione condizionale assume la forma

$$\ell(\rho \mid \boldsymbol{W}, \boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\hat{\sigma}^2 + \frac{1}{2}\log \mid \boldsymbol{I} - \rho\boldsymbol{W} \mid,$$

con

$$\hat{\sigma}^2 = \frac{\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}}{n}, \tag{1.3}$$

per il modello nella formulazione simultanea

$$\ell(\rho \mid \boldsymbol{W}, \boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\hat{\sigma}^2 + \log \mid \boldsymbol{I} - \rho\boldsymbol{W} \mid,$$

con

$$\hat{\sigma}^2 = \frac{\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W}^T)(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}}{n}, \tag{1.4}$$

per la dimostrazione e approfondimenti si veda il paragrafo B.1 dell'appendice B.

Per analizzare le dipendenze in un modello autoregressivo spaziale del primo ordine, si consideri un reticolo regolare di $n = p \times q$ aree, indicando con $X_{l,k}$ la generica area, per $l = 1, \ldots, p$ e $k = 1, \ldots, q$, si veda la figura 1.1.

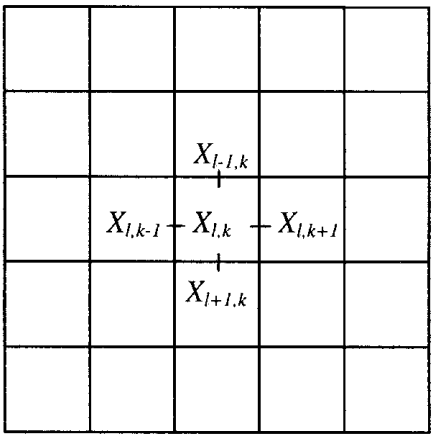


Figura 1.1: Dipendenze in uno schema autoregressivo del primo ordine definito su un reticolo regolare

Il modello autoregressivo simultaneo di primo ordine (si veda per esempio Cliff e Ord, 1981) è definito come

$$X_{l,k} = \rho(X_{l-1,k} + X_{l+1,k} + X_{l,k-1} + X_{l,k+1}) + \epsilon_{l,k},$$

con la condizione di stazionarietà $\mid \rho \mid < 1/4$ e le dovute modifiche nel caso di aree al confine. Infatti, per esse, vengono limitate le aree adiacenti ponendo uguale a zero il valore delle variabili che rappresentano il fenomeno fuori del confine, dove il processo non è osservato, cioè $X_{l,k} = 0$ per $l > p$ o $k > q$. Per lo stesso reticolo il modello autoregressivo condizionale di primo ordine è dato da

$$\mathrm{E}(X_{l,k} \mid X_{-l,-k} = x_{-l,-k}) = \rho(x_{l-1,k} + x_{l+1,k} + x_{l,k-1} + x_{l,k+1}),$$

dove $X_{-l,-k}$ indica che dalla matrice $\boldsymbol{X}$ è stata tolta la componente $X_{l,k}$ e $x_{l,k}$ indica l'elemento generico della matrice dei valori osservati sul reticolo; la condizione di stazionarietà è ancora $\rho < 1/4$, vengono inoltre operate le stesse modifiche per le aree al confine.

Si noti che il valore $X_{l,k}$ relativo all'area (l,k) del reticolo, in entrambi i modelli, dipende solo dai valori assunti dal processo nelle aree confinati; queste sono quattro nel caso si consideri un sistema di vicinato "a torre", otto nel caso se ne consideri uno "a regina" (in questo caso sono considerate adiacenti le aree aventi un lato o un vertice in comune con l'area in esame).

Per le $2p + 2q - 4$ aree che si trovano al confine della regione, invece, il numero di aree adiacenti è ridotto a tre o due nel caso si consideri un sistema di vicinato a torre, ed a cinque o tre quando sia adottato uno a regina.

Considerando ancora un reticolo regolare $p \times q$ di n aree, la stima del coefficiente di autocorrelazione (si veda Whittle, 1954) può essere ottenuta quale soluzione di un equazione che coinvolge la stima del coefficiente di autocovarianza

$$\gamma(r, s) = \mathrm{E}(X_{l,k}, X_{l+r,k+s}),$$

e la stima della varianza σ^2; tale metodo prende il nome di approssimazione della verosimiglianza di Whittle. Le stime usate sono date rispettivamente da

$$\hat{\gamma}(r, s) = \frac{1}{n} \sum_{l=1}^{p-r} \sum_{k=1}^{q-s} x_{l,k} x_{l+r,k+s}, \tag{1.5}$$

per il coefficiente di autocovarianza e, per la varianza, dall'espressione (1.3) per il modello autoregressivo condizionale e dalla (1.4) per il modello autoregressivo simultaneo.

1.3 L'effetto confine e la sua correzione nella stima dei modelli autoregressivi spaziali

Nel modello autoregressivo condizionale spaziale di primo ordine viene fatta un'assunzione di markovianità rispetto ad una struttura di adiacenze, cioè la dipendenza tra le variabili del processo definito sul reticolo è spiegata da una opportuna serie di relazioni di indipendenza condizionale. Viene definita una struttura di adiacenze del tipo di quella dell'espressione (1.2); la distribuzione condizionale è definita per tutte le adiacenze, ma i dati a disposizione sono noti soltanto per alcune. Il modello è definito sul reticolo infinito (1.1), ma è applicato a quello finito

$$S \equiv \{i : i = 1, \ldots, n\}$$

e le adiacenze considerate sono quindi

$$W_i \equiv \{j : j \text{ è adiacente ad } i, \text{ per } j = 1, \ldots, n\}.$$

Così, quando si considera un modello autoregressivo del primo ordine, l'insieme delle variabili a cui doversi condizionare è più ampio di quanto in effetti venga fatto.

In entrambi i modelli autoregressivi, considerando un reticolo di $i = 1, \ldots, n$ aree, il problema dell'effetto confine si presenta quando X_i è influenzato da X_j e l'area j è esterna alla regione di studio ($j \neq 1, \ldots, n$); in tal caso si ha un problema di non corretta specificazione del modello in cui il vero processo che governa X_i non è accuratamente rappresentato. In particolare, considerando un reticolo regolare $p \times q$ di n aree, il problema della mancanza di una corretta specificazione del modello riguarda le $2p + 2q - 4$ aree al confine.

Se nelle serie temporali l'effetto confine può essere trascurato, passando da $d = 1$ a $d = 2$ le sue conseguenze divengono più gravi, dato che aumentano i nodi del grafo con cui è rappresentato l'insieme che indicizza il processo, per i quali la distribuzione condizionale è mal specificata; in generale il problema diventa ancora più serio per $d > 2$.

In letteratura è stata analizzata la distorsione per l'effetto confine nella stima di massima verosimiglianza del parametro di autocorrelazione ρ nel caso di un modello autoregressivo simultaneo o condizionale gaussiano di primo ordine su un reticolo regolare, poco è stato fatto per schemi autoregressivi più complicati o per reticoli irregolari.

Uno dei primi contributi alla ricerca sull'effetto confine si deve a Guyon (1982). Egli propone la seguente correzione per l'effetto confine della stima del coefficiente di autocorrelazione spaziale: questa si ottiene inserendo quale stima dell'autocovarianza nell'equazione di approssimazione di Whittle l'espressione

$$\hat{\gamma}'(r,s) = \frac{n\hat{\gamma}(r,s)}{(p - \mid r \mid)(q - \mid s \mid)}.$$

Per approfondimenti si vedano Cressie (1991) a pag. 476–479, Ripley (1984) e per altri sviluppi Dahlhaus e Künsch (1987).

Haining (1977), Griffith (1983) e Griffith e Amrhein (1983) hanno mostrato che la stima di ρ, ottenuta quale massimizzazione del profilo di verosimiglianza, senza correzioni per l'effetto confine, tende ad essere una sottostima del vero parametro; fortunatamente quando n aumenta la distorsione diminuisce.

Sono state proposte varie metodologie per correggere per l'effetto confine la stima del coefficiente di autocorrelazione nei modelli autoregressivi spaziali.

Una soluzione è marginalizzare rispetto alle variabili che descrivono il fenomeno all'esterno della regione, ossia integrare ogni distribuzione condizionale rispetto alle variabili non osservabili (si veda Cressie, 1991 pag. 422), ma questo non è sempre possibile, almeno analiticamente.

Un'altra soluzione è considerare la regione infinita immaginandola di forma toroidale, in questo modo anche le aree che si trovano al confine posseggono un insieme completo di adiacenze, si veda la figura 1.2.

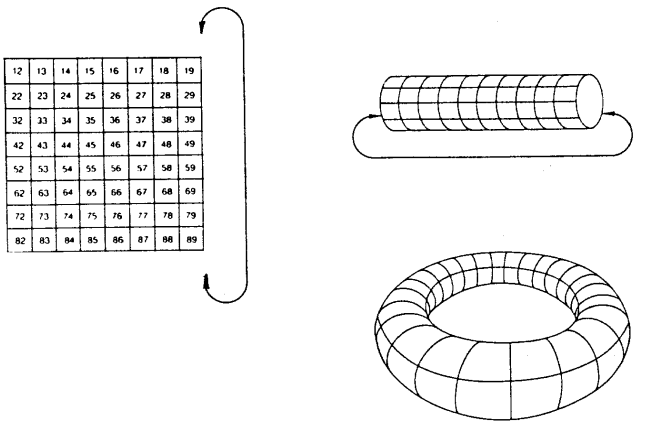


Figura 1.2: La mappa di forma toroidale

Una soluzione è quella di considerare, nel modello autoregressivo, il condizionamento ai valori assunti dal processo nelle aree che si trovano al confine della regione (area di guardia, o *empirical buffer zone* o corona interna), ma non considerare queste aree nell'analisi oltre che come aree adiacenti a quelle interne alla regione; non viene cioè considerata nel modello la loro equazione autoregressiva. Un inconveniente di tale approccio è quello di provocare una notevole perdita di informazione riguardo al fenomeno analizzato.

Un modo per evitare tale perdita di informazioni è quella di estendere la regione sotto studio, costruendo una corona esterna empirica, in modo che le aree escluse siano aree che non sono di interesse.

Quando non è possibile reperire i dati relativi a tali aree, si può costruire una corona esterna artificiale, composta da una (si veda la figura 1.3.a) o più aree (figura 1.3.b); la variabile casuale relativa a ciascuna area della corona esterna assume, quale modalità, un certo valore: esso può essere la media aritmetica dei valori assunti nelle aree della regione, o il valore estrapolato da una superficie stimata attraverso un modello *trend surface*, o un valore imputato secondo le metodologie applicate quando si svolgono delle analisi in condizioni di dati mancanti.

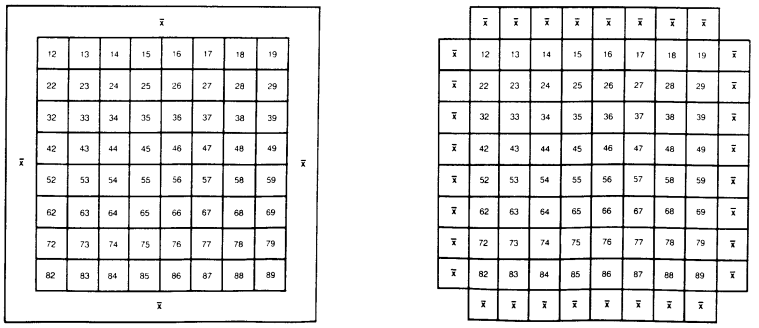


Figura 1.3: La mappa espansa con una (a) o più (b) aree esterne artificiali

Seguendo il metodo proposto da Ripley (1981) per la correzione per l'effetto confine nel caso di dati individuali (si veda paragrafo A.2.4), Griffith (1983) suggerisce l'adozione di uno schema di pesi; tale metodo non è però ulteriormente specificato, ad esempio non è chiaro in quale modo siano definiti i pesi.

Le metodologie sopra descritte, insieme con altre (una prevede l'introduzione di una variabile *dummy* che discrimina le aree al confine dalle altre, l'altra riguarda la teoria delle stime dei minimi quadrati generalizzati), sono descritte, con i corrispondenti riferimenti bibliografici, nel paragrafo B.2 dell'appendice B.

Considerazioni conclusive

In questo primo capitolo sono state presentate alcune soluzioni presenti in letteratura riguardo alla correzione per l'effetto confine della stima del coefficiente di autocorrelazione nei modelli autoregressivi spaziali. Molti dei metodi derivano da quelli suggeriti per la correzione per l'effetto confine nelle analisi di dati individuali (si veda il paragrafo A.2 dell'appendice A). Malgrado tali metodi di correzione siano usati per due tipologie di modelli e dati diversi, tali metodi di correzione sono sostanzialmente unici: l'abolizione dei confini (considerando la superficie a forma toroidale), la costruzione di un'area di guardia interna (*empirical buffer zone*) o esterna (*artificial buffer zone*), o l'introduzione di un sistema di pesi.

In questo capitolo sono state presentate tali metodologie quali proposte possibili per una loro applicazione in un altro contesto: la correzione per

l'effetto confine del modello bayesiano gerarchico usato per la stima del rischio di mortalità in una data regione geografica quando si dispone di dati in forma aggregata. Quando si studia il rischio di mortalità su una regione, il numero di decessi osservati, che sono i dati in forma aggregata disponibili in tali studi, sono considerati come una realizzazione di un processo spaziale definito su un reticolo irregolare: usualmente si assume che il valore osservato su ogni area rappresenti la modalità di una variabile casuale Poisson, indicizzata attraverso un elemento del reticolo e parametrizzata tramite il rischio relativo. In tal senso occorre stimare i parametri del processo spaziale; in ottica bayesiana, si considera un ulteriore processo, definito sullo stesso reticolo: associata ad ogni area, in questo caso, si considera una variabile casuale continua, che descrive il rischio della stessa. Assumendo un modello di interazione spaziale per il rischio relativo è possibile considerare la natura spaziale dei dati ed ottenere delle mappe da cui sia possibile trarre importanti informazioni epidemiologiche. Per fare ciò si può estendere la teoria dei modelli autoregressivi spaziali condizionali esaminati in questo capitolo al caso di reticoli irregolari (si veda Besag 1975 e Künsch 1987). Poiché si considera ancora un modello autoregressivo spaziale, i metodi di correzione per l'effetto confine analizzati nel paragrafo precedente rappresentano dei validi suggerimenti per operare la correzione nei modelli bayesiani cui siamo interessati. Si noti che in questo caso l'interesse non sarà sul parametro di autocorrelazione spaziale del modello autoregressivo condizionale, ma sui valori teorici che si ottengono per il processo.

Nel prossimo capitolo vengono descritti alcuni dei modelli maggiormante utilizzati per la stima del rischio relativo di mortalità su una regione quando si dispone di dati in forma aggregata; l'attenzione è rivolta al modello bayesiano gerarchico. Viene poi evidenziato come tali modelli, ed in particolare quello di interesse, siano influenzati dall'effetto confine. Da tale capitolo si sviluppano i seguenti, con le proposte per operare la correzione per l'effetto confine e la loro applicazione.

Capitolo 2

L'effetto confine nei metodi di stima del rischio relativo

Premessa

La descrizione della variazione del rischio relativo su una regione, quando si dispone di dati in forma aggregata, è fatta mediante la sua mappatura, generando carte di patologia che sono generalmente raccolte in atlanti di mortalità. Per la costruzione di queste si considera una partizione della regione di interesse in n aree, individuate in base all'aggregazione presente nei dati disponibili (usualmente secondo criteri di carattere amministrativo).

Generalmente si opera su una serie di frequenze di eventi di malattia (o di decessi) O_i, con $i = 1, \ldots, n$ l'area cui si riferiscono, ed una serie di frequenze attese E_i, solitamente ottenute con il metodo della standardizzazione indiretta, applicando alla popolazione dell'area, suddivisa per età e sesso, una serie di tassi specifici calcolati sul totale delle aree o derivati da una popolazione esterna presa come riferimento. È necessario considerare nell'analisi i casi attesi poiché i tassi specifici di mortalità variano notevolmente con l'età, quindi occorre standardizzare per tale fattore; in tal modo si evita che le stime del rischio relativo siano influenzate da differenze nella struttura della popolazione tra le aree (si veda Breslow e Day, 1975).

Lo scopo dell'analisi è stimare il rischio relativo θ_i corrispondente a ciascuna area $i = 1, \ldots, n$; essa è ottenuta usando l'informazione data dai casi osservati e attesi su ognuna di esse.

Sono stati sviluppati vari metodi di stima del rischio relativo: da quello di massima verosimiglianza, a quelli che si basano sulla quasi-verosimiglianza penalizzata e sulla pseudo-verosimiglianza (si vedano Green, 1984 e Besag, 1986), a modelli bayesiani empirici e modelli bayesiani gerarchici.

In questo capitolo sono analizzati i metodi generalmente usati per la stima del rischio relativo: il metodo della massima verosimiglianza (il rapporto standardizzato di mortalità), il modello bayesiano empirico (con particolare riferimento a quello definito da Clayton e Kaldor, 1987) e il modello bayesiano gerarchico di Besag, York e Mollié (1991). Nel paragrafo 2.5 viene illustrato in quale modo sia presente il problema dell'effetto confine per i modelli analizzati.

2.1 Il rapporto standardizzato di mortalità

Una semplice stima del rischio relativo, per la generica area i-esima, è rappresentata dal rapporto standardizzato di mortalità (o di incidenza) (SMR, *Standardized Mortality Ratio*) dato da

$$\mathrm{SMR}_i = \frac{O_i}{E_i}.$$

Generalmente si assume che il numero degli eventi osservati O_i in ogni area i, per $i = 1, \ldots, n$, dato il rischio relativo θ_i nella stessa, si distribuisca come una variabile casuale di Poisson di media $E_i\theta_i$: sulla regione analizzata, i casi osservati si distribuiscono, dunque, come un processo di Poisson non omogeneo. Sotto tale assunzione, $\hat{\theta}_i = O_i/E_i = \mathrm{SMR}_i$ è lo stimatore di massima verosimiglianza di θ_i.

Tali stime, quando l'analisi è condotta su aree scarsamente popolate, sono fortemente instabili: se il numero di eventi attesi è basso, un caso osservato in più o in meno può dar origine a stime dei rischi relativi molto diverse. In tal senso, quando l'analisi è condotta su aree di piccole dimensioni, quindi su aree anche scarsamente popolate, si ha in genere un'elevata variabilità delle stime a causa di una eterogeneità nei livelli individuali di rischio. Le mappe che riportano le stime dei rischi relativi ottenute tramite l'indice SMR tendono ad essere dominate da quelle più instabili; esse sono pertanto difficilmente interpretabili da un punto di vista epidemiologico. È quindi necessario adottare "tecniche di lisciamento" (*smoothing*), si veda Breslow (1984); in letteratura sono stati suggeriti, tra gli altri, vari approcci, quali il modello bayesiano empirico e quello bayesiano gerarchico.

Per il rischio relativo viene considerato un modello di Poisson log-lineare, il cui predittore lineare, nell'approccio classico, contiene un termine η_i che rappresenta la variazione del rischio nello spazio (l'effetto area), cioè

$$\log(\theta_i) = \mu + \eta_i.$$

Nel predittore lineare può essere considerato un termine costante μ che rappresenta il livello di riferimento per i logaritmi dei rischi relativi (generalmente la loro media), ogni effetto area rappresenta lo scostamento del logaritmo del rischio relativo di ogni area da quest'ultimo; per semplicità si può assumere un valore $\mu = 0$.

Il modello bayesiano gerarchico e quello bayesiano empirico permettono di "filtrare il rumore" dovuto alla variabilità poissoniana e quindi forniscono stime che danno un'immagine più chiara della variazione geografica; tutto ciò modellando separatamente la variabilità poissoniana da quella non poissoniana che rappresenta la reale variazione del rischio. Tali modelli, oltre all'informazione data dai casi osservati, sfruttano un'informazione a priori sulla variabilità del rischio relativo sulla mappa, specificando una distribuzione sugli effetti area, definendoli come casuali. In questo caso si considera un modello lineare generalizzato, il cui predittore lineare contiene degli effetti casuali, cioè un modello lineare generalizzato misto (GLMM, *Generalized Linear Mixed Model*, si veda Breslow e Clayton, 1993).

Sugli effetti casuali può essere definito un modello scambiabile: i rischi relativi delle aree della regione non hanno una struttura spaziale, sono quindi indipendenti, in genere tendono a non differire troppo tra di loro e sono così prossimi ad un loro valore medio (la media generale sulla mappa). Oppure, inserendo un'informazione sulla spazialità del fenomeno, cioè la posizione di ogni area sulla mappa, i rischi relativi hanno una struttura spaziale e sono simili per aree adiacenti: definendo una dipendenza markoviana, la stima del rischio relativo in ogni area è prossima ad una media locale.

Considerando gli effetti area come casuali si ottengono delle stime, per i rischi relativi, che risultano essere un compromesso tra l'indice SMR e la media generale o locale dei rischi stessi.

2.2 Il modello bayesiano empirico

Clayton e Kaldor (1987) suggeriscono degli stimatori bayesiani empirici per il rischio relativo. Questi stimatori derivano da quelli che James e Stein (1961) hanno proposto, nell'ambito della teoria delle decisioni, per l'analogo problema della stima di p medie di altrettante normali. Tali stimatori hanno una perdita errore quadratico minore, per $p > 2$, rispetto a quella degli stimatori di massima verosimiglianza; si deve ad Efron e Morris (1973) la giustificazione di tali stimatori in termini di un approccio bayesiano empirico.

Clayton e Kaldor (1987) definiscono un modello bayesiano empirico assumendo che i casi osservati si distribuiscano, condizionatamente ai rischi relativi, come una variabile di Poisson di valore atteso $E_i\theta_i$, per $i = 1, \ldots, n$. Si assume che i rischi relativi siano delle variabili indipendenti e identicamente distribuite secondo una distribuzione gamma, con parametro inverso di scala α e parametro di forma β, quindi con media β/α e varianza β/α^2. La stima del rischio relativo è data dal valore atteso della distribuzione a posteriori dei rischi relativi. La distribuzione per i rischi relativi, condizionata alle osservazioni, è ancora una distribuzione gamma con parametro inverso di scala $E_i+\alpha$ e il parametro di forma $O_i+\beta$. Il valore atteso della distribuzione è dato da

$$\mathrm{E}(\theta_i \mid O_i, \alpha, \beta) = \frac{O_i + \beta}{E_i + \alpha}$$

e quindi la stima bayesiana empirica dei rischi relativi è data dall'espressione precedente, in cui α e β sono sostituiti con le loro stime $\hat{\alpha}$ e $\hat{\beta}$. Queste possono essere ottenute dai dati, che seguono una distribuzione binomiale negativa. Infatti, le osservazioni hanno una distribuzione di Poisson condizionatamente ai rischi relativi, ed una distribuzione marginale binomiale negativa con media $(E_i\beta)/\alpha$ e varianza $(E_i\beta)/\alpha + (E_i^2\beta)/\alpha^2$.

Le stime bayesiane empiriche hanno una dispersione minore rispetto a quelle ottenute tramite l'indice SMR; ogni stima è un compromesso tra quest'ultimo e la stima della media della distribuzione assunta per i rischi relativi, cioè tra O_i/E_i e $\hat{\beta}/\hat{\alpha}$:

$$\hat{\theta_i}' = z_i \frac{O_i}{E_i} + (1 - z_i)\frac{\hat{\beta}}{\hat{\alpha}}.$$

La formula evidenzia come lo stimatore sia una media pesata tra l'indice SMR e la media della distribuzione a priori sul rischio relativo, il peso $z_i = E_i/(E_i + \hat{\alpha})$ dipende dal numero dei casi attesi. Nel caso le stime siano basate su un gran numero di eventi attesi, esse saranno vicine alla stima SMR, dato che il peso sarà prossimo ad uno; per malattie rare o aree poco popolate, un peso maggiore sarà dato alla media della distribuzione a priori, così da ottenere stime lisciate (si ha un'attrazione dei rapporti standardizzati verso la media generale). In questo modo le stime stabili, cioè ottenute da molte osservazioni, sono preservate, mentre quelle instabili, ottenute basandosi su poche osservazioni, sono "lisciate", ovvero sono attratte verso la media generale. Louis (1984) e Gosh (1992) propongono stimatori bayesiani empirici vincolati per evitare una diminuizione troppo drastica della variabilità sulla mappa (*oversmoothing*). Il modello bayesiano empirico fornisce

valide stime puntuali del rischio relativo, ma non della variabilità dello stesso, dato che non considera la variabilità di α e β; per ovviare a tale problema Morris (1983) propone l'uso del metodo delta, Laird e Louis (1987) e Biggeri, Braga e Marchi (1993) propongono ed adottano procedure *bootstrap*.

2.3 Il modello bayesiano gerarchico

Besag, York e Mollié (1991) suggeriscono un approccio bayesiano basato su di un modello gerarchico a tre stadi. Le stime sono ottenute combinando due tipi di informazione: quella relativa ai casi osservati descritta tramite la verosimiglianza poissoniana e l'informazione data dalla distribuzione a priori sui rischi relativi.

L'inferenza bayesiana circa il rischio relativo ignoto di ogni area è basata ancora sulla sua distribuzione a posteriori

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}) \propto p(\boldsymbol{O} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

La funzione di verosimiglianza del rischio relativo $\boldsymbol{\theta}$ dati i casi osservati $\boldsymbol{O}$, poiché le osservazioni, condizionatamente al rischio relativo, sono assunte indipendenti, è data da

$$p(\boldsymbol{O} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(O_i \mid \theta_i).$$

La distribuzione a priori $p(\boldsymbol{\theta} \mid \lambda)$ riflette l'informazione circa la variazione del rischio relativo sulla mappa; generalmente è parametrizzata tramite l'iperparametro λ, che può essere uno scalare, o un vettore, a seconda del modello specificato sui rischi relativi. A differenza del modello bayesiano empirico in cui i parametri della distribuzione sul rischio relativo sono stimati dai dati, nel modello bayesiano viene definita su di essi una distribuzione $p(\lambda)$. La distribuzione a posteriori congiunta dei parametri $\boldsymbol{\theta}$ e λ è data da

$$p(\boldsymbol{\theta}, \lambda \mid \boldsymbol{O}) \propto p(\boldsymbol{O} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \lambda) p(\lambda)$$

e la distribuzione marginale a posteriori del rischio relativo è dunque

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}) = \int p(\boldsymbol{\theta}, \lambda \mid \boldsymbol{O}) d\lambda. \tag{2.1}$$

Una stima puntuale dell'insieme degli n rischi relativi è data da una misura di locazione della distribuzione (2.1), generalmente il suo valore atteso.

Nella definizione del modello bayesiano gerarchico, per il numero dei casi osservati dato il rischio relativo (primo stadio del modello bayesiano gerarchico), si assume una distribuzione

$$p(O_i \mid \theta_i) = \exp(-E_i\theta_i)\frac{(E_i\theta_i)^{O_i}}{O_i!},$$

dove θ_i rappresenta il rischio relativo della i-esima area. La distribuzione a priori sul rischio relativo (secondo stadio del modello bayesiano gerarchico) può essere specificata in vari modi.

2.3.1 Modello scambiabile per i rischi relativi

Il rischio relativo di un'area è indipendente da quelli delle altre aree condizionatamente al valore dell'iperparametro λ, conseguentemente si ha una eterogeneità dei rischi relativi non strutturata spazialmente. Le stime dei rischi relativi sono prossime al loro valore di riferimento (la media generale sulla mappa).

La distribuzione a priori sui rischi relativi è data da

$$p(\boldsymbol{\theta} \mid \lambda) = \prod_{i=1}^{n} p(\theta_i \mid \lambda),$$

dove la a priori $p(\theta_i \mid \lambda)$ è la stessa per ciascuna area i. Generalmente sono usate o la distribuzione gamma (coniugata della Poisson) o la distribuzione lognormale. Se $\mu = 0$, si può assumere per il logaritmo del rischio relativo $x_i = \log \theta_i$ una distribuzione normale con media nulla e varianza λ. Indicando con $\boldsymbol{x}$ il vettore dei termini x_i si definisce la distribuzione

$$p(\boldsymbol{x} \mid \lambda) \propto \frac{1}{\lambda^{\frac{n}{2}}} \exp\left(-\frac{1}{2\lambda}\sum_{i=1}^{n} x_i^2\right). \tag{2.2}$$

Quando si opera la scelta di distribuzioni non coniugate, come in questo caso, non si dispone di espressioni analitiche per la distribuzione a posteriori e quindi si deve ricorrere a soluzioni approssimate con metodi, cosiddetti, Monte Carlo *Markov Chain*: mediante il *Gibbs sampler*, si può simulare la distribuzione a posteriori marginale del rischio relativo e quindi produrre stime puntuali e intervallari; il funzionamento di tale metodo è brevemente descritto nel paragrafo D.1 dell'appendice D.

La rappresentazione del modello bayesiano gerarchico con un modello scambiabile per i rischi relativi è rappresentata dal grafo nella figura 2.1.

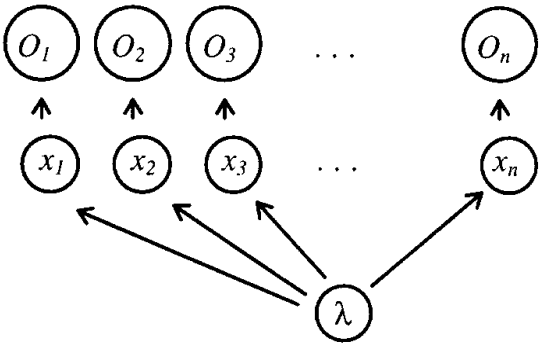


Figura 2.1: Modello scambiabile per i rischi relativi

Esso rappresenta un insieme di assunzioni di indipendenza: le osservazioni O_i sono indipendenti da λ dato x_i per $i = 1, \ldots, n$ e sono inoltre indipendenti tra loro, dato $\boldsymbol{x}$; viene poi assunta l'indipendenza tra O_i e $\boldsymbol{x}_{-i}$ (con $\boldsymbol{x}_{-i}$ il vettore $\boldsymbol{x}$ senza la componente i-esima) e la mutua indipendenza tra i logaritmi dei rischi relativi $\boldsymbol{x}$, dato il valore dell'iperparametro λ.

2.3.2 Modello spaziale per i rischi relativi

Un altro modo di definire una distribuzione a priori è quella di assumere una eterogeneità dei rischi relativi strutturata spazialmente: la distribuzione del rischio relativo di una certa area, condizionata ai valori dei rischi relativi delle altre, dipende soltanto dal rischio relativo delle aree adiacenti. In questo caso la stima del rischio relativo di un area è vicina al valore di una media locale. Per descrivere tale struttura spaziale può essere utilizzato un modello autoregressivo condizionale spaziale (come quello descritto nel paragrafo 1.2) che specifica, per il logaritmo del rischio relativo, una distribuzione condizionale normale con la media che dipende dalla media dei logaritmi dei rischi relativi per le aree adiacenti. La forma usuale del modello autoregressivo condizionale gaussiano assume che la varianza condizionale sia costante, ma per mappe irregolari, dove il numero di aree confinanti può variare, sembra più valido un modello gaussiano autoregressivo intrinseco (si veda Künsch, 1987), dove la varianza condizionale del logaritmo del rischio relativo ($x_i = \log \theta_i$), dati tutti gli altri logaritmi dei rischi relativi della mappa, è inversamente

proporzionale al numero di aree adiacenti. In questo caso si ha

$$p(\boldsymbol{x} \mid \lambda) \propto \frac{1}{\lambda^{\frac{n}{2}}} \exp\left[-\frac{1}{2\lambda}\sum_{i=1}^{n}\sum_{j<i} w_{ij}(x_i - x_j)^2\right], \tag{2.3}$$

con w_{ij} l'elemento generico della matrice di adiacenze. La rappresentazione del modello bayesiano gerarchico con una distribuzione sul logaritmo del rischio relativo autoregressiva intrinseca gaussiana, è rappresentata mediante il grafo della figura 2.2.

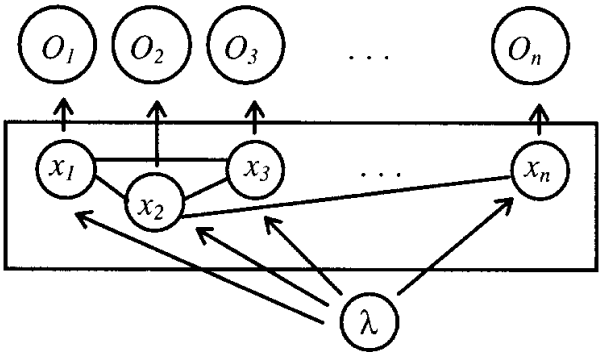


Figura 2.2: Modello autoregressivo intrinseco per i rischi relativi

Si noti, in questo caso, la dipendenza condizionale (legame senza direzione) fra i logaritmi dei rischi relativi delle aree confinanti. L'assenza di tale legame indica che le aree non sono confinanti e che i logaritmi dei rischi relativi sono condizionatamente indipendenti dato λ.

2.3.3 Modello di convoluzione gaussiana per i rischi relativi

La scelta della a priori da adottare può essere guidata da una conoscenza epidemiologica riguardo alla regione studiata: quando le aree sono piccole ed i casi osservati poco numerosi, si preferisce in genere assumere una eterogeneità strutturata spazialmente, per aree vaste sembra invece più appropriato assumere un modello che considera una eterogeneità dei rischi non strutturata spazialmente. Nel caso intermedio si può assumere un modello che spiega l'eterogeneità dei rischi come solo parzialmente dovuta ad una interdipendenza spaziale; Besag, York e Mollié (1991) suggeriscono un modello misto di a priori: per ogni area il logaritmo del rischio relativo è dato dalla somma di

due componenti casuali indipendenti e gaussiane (convoluzione gaussiana), una che rappresenta la parte di eterogeneità non strutturata (*heterogeneity*) e l'altra quella strutturata spazialmente (*clustering*). Si noti che solo definendo tale modello per i rischi relativi è possibile considerare entrambe le componenti. Spesso è adottata una distribuzione normale per il logaritmo del rischio relativo, in particolare si ricorre alla teoria dei modelli lineari generalizzati misti per la specificazione della distribuzione del rischio relativo θ_i dell'area i-esima

$$x_i = \log \theta_i = v_i + u_i, \tag{2.4}$$

dove il logaritmo del rischio relativo è una funzione lineare di v_i e u_i; nel predittore lineare può essere inserito anche un termine μ quale livello comune di riferimento, per semplicità si può ancora assumere $\mu = 0$.

Il termine casuale *heterogeneity* v_i si distribuisce come una variabile casuale normale di media nulla e varianza λ_v, cioè in modo analogo alla (2.2). La distribuzione di ogni termine è data da

$$p(v_i \mid \lambda_v) \propto \frac{1}{\sqrt{\lambda_v}} \exp\left(-\frac{v_i^2}{2\lambda_v}\right).$$

La componente di *clustering* u_i si distribuisce seguendo un modello autoregressivo gaussiano intrinseco come quello dell'equazione (2.3), con la seguente distribuzione condizionale per il generico termine u_i

$$p(u_i \mid \boldsymbol{u}_{-i}, \lambda_u) \propto \sqrt{\frac{w_{i.}}{\lambda_u}} \exp\left[-\frac{w_{i.}(u_i - \bar{u}_i)^2}{2\lambda_u}\right]; \tag{2.5}$$

cioè una distribuzione normale con media

$$\bar{u}_i = \frac{1}{w_{i.}} \sum_{j \neq i} w_{ij} u_j,$$

dove $w_{i.}$ indica il numero di aree adiacenti all'area i-esima, e varianza $\lambda_u / w_{i.}$. Si noti che la media è data dalla media dei valori assunti dalle altre componenti nelle aree adiacenti la i-esima ed anche che sia la media che la varianza sono, in questo caso, condizionate alla struttura spaziale, cioè alla matrice $\boldsymbol{W}$.

Si noti che la distribuzione sulle osservazioni, riparametrizzata secondo il modello di convoluzione gaussiana per il rischio relativo, diviene

$$p(O_i \mid u_i, v_i) = \frac{1}{O_i!} \exp[-E_i \exp(u_i + v_i) + O_i \log(E_i) + O_i(u_i + v_i)].$$

Per gli iperparametri λ_v e λ_u (terzo stadio del modello bayesiano gerarchico) viene assunta una distribuzione a priori gamma inversa (coniugata con la normale), in modo da ottenere, grazie all'indipendenza degli iperparametri dalle osservazioni dati i parametri $\boldsymbol{v}$ e $\boldsymbol{u}$, distribuzioni a posteriori in forma chiusa. Le distribuzioni a priori sono date da

$$p(\lambda_v \mid a_v, b_v) \propto \lambda_v^{-(a_v+1)} \exp\left(-\frac{b_v}{\lambda_v}\right) \quad \text{e} \quad p(\lambda_u \mid a_u, b_u) \propto \lambda_u^{-(a_u+1)} \exp\left(-\frac{b_u}{\lambda_u}\right)$$

in cui a_u e a_v rappresentano i parametri di forma e b_v e b_u i parametri di scala. Si possono adottare distribuzioni non informative improprie, ponendo $a_v = a_u = -1$ e $b_v = b_u = 0$; per motivi di convergenza sarà necessario porre $b_v = b_u = \epsilon$ (per esempio $\epsilon = 0.005$) per evitare problemi tecnici nell'algoritmo *Gibbs sampler* cui è necessario ricorrere per la procedura di stima (si veda Besag, York e Mollié, 1991 pag. 9). In questo caso le distribuzioni a priori sono date da

$$p(\lambda_v) \propto \exp\left(-\frac{\epsilon}{\lambda_v}\right) \qquad \text{e} \qquad p(\lambda_u) \propto \exp\left(-\frac{\epsilon}{\lambda_u}\right).$$

Alternativamente, possono essere specificate distribuzioni a priori informative, si vedano Mollié (1996) a pag. 372 e Bernardinelli *et al.* (1995). Infatti, quando si analizzano i rischi di mortalità per malattie rare (ad esempio il tumore alla mammella maschile), è necessario ricorrere ad una informazione a priori sulla distribuzione dei rischi relativi onde ottenere dei risultati interpretabili da un punto di vista epidemiologico.

Si noti che i parametri λ_u e λ_v controllano l'importanza di ciascuna componente nella convoluzione: se λ_u/λ_v è piccolo si ha una prevalenza di eterogeneità non strutturata spazialmente, se è grande si ha una eterogeneità strutturata spazialmente, se è pari a $\bar{w}$ (il numero medio di aree adiacenti) le due componenti hanno la stessa importanza nella convoluzione.

La rappresentazione tramite un grafo del modello bayesiano gerarchico con una a priori sul logaritmo del rischio relativo data dalla convoluzione gaussiana della (2.4) è rappresentata nella figura 2.3; in questo caso, la dipendenza condizionale fra i logaritmi dei rischi relativi è rappresentata dal legame senza direzione fra i termini di *clustering* $u_1, \ldots, u_n$ delle aree adiacenti. Si noti che i termini $v_1, \ldots, v_n$, condizionati all'iperparametro λ_v, sono tra loro indipendenti.

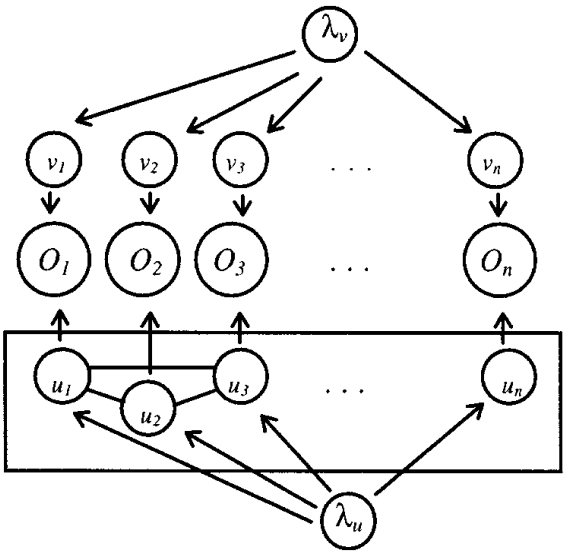


Figura 2.3: Modello di convoluzione gaussiana per i rischi relativi

2.4 Implementazione del modello bayesiano gerarchico

Per il modello gerarchico bayesiano definito nel paragrafo precedente, la distribuzione a posteriori congiunta del logaritmo del rischio relativo e degli iperparametri è data da

$$p(\boldsymbol{x}, \boldsymbol{\lambda} \mid \boldsymbol{O}) = \frac{p(\boldsymbol{O}, \boldsymbol{\lambda}, \boldsymbol{x})}{\int\int p(\boldsymbol{O}, \boldsymbol{\lambda}, \boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{\lambda}} \propto p(\boldsymbol{O} \mid \boldsymbol{x}) p(\boldsymbol{x} \mid \boldsymbol{\lambda}) p(\boldsymbol{\lambda}).$$

Da questa possono essere determinate, a meno di un fattore di scala, le distribuzioni a posteriori condizionali di ciascun parametro del modello rispetto a tutti gli altri (distribuzioni condizionali complete) necessarie per il *Gibbs sampler*. Infatti, la distribuzione a posteriori di $\boldsymbol{x}$ condizionatamente a $\boldsymbol{\lambda}$ sarà data da

$$p(\boldsymbol{x} \mid \boldsymbol{\lambda}, \boldsymbol{O}) = \frac{p(\boldsymbol{x}, \boldsymbol{\lambda} \mid \boldsymbol{O})}{p(\boldsymbol{\lambda} \mid \boldsymbol{O})},$$

e poiché $p(\boldsymbol{\lambda} \mid \boldsymbol{O}) = p(\boldsymbol{\lambda})$ non dipende da $\boldsymbol{x}$ si ha semplicemente

$$p(\boldsymbol{x} \mid \boldsymbol{\lambda}, \boldsymbol{O}) \propto p(\boldsymbol{x}, \boldsymbol{\lambda} \mid \boldsymbol{O}).$$

Quindi per derivare la distribuzione condizionale completa di $\boldsymbol{x}$, basta considerare i termini della congiunta a posteriori che dipendono da $\boldsymbol{x}$.

Assumendo un modello di convoluzione gaussiana per i rischi relativi, poiché $\boldsymbol{u}$ e $\boldsymbol{v}$ sono tra loro indipendenti e lo sono anche λ_u e λ_v, si ha la seguente

$$p(\boldsymbol{x} \mid \boldsymbol{\lambda}) = p(\boldsymbol{v}, \boldsymbol{u} \mid \lambda_v, \lambda_u) \propto p(\boldsymbol{v} \mid \lambda_v) p(\boldsymbol{u} \mid \lambda_u)$$

da cui

$$\begin{aligned}
p(\boldsymbol{u}, \boldsymbol{v}, \lambda_u, \lambda_v \mid \boldsymbol{O}) &\propto \prod_{i=1}^{n} p(O_i \mid v_i, u_i) p(\boldsymbol{v} \mid \lambda_v) p(\boldsymbol{u} \mid \lambda_u) p(\lambda_v) p(\lambda_u) \\
&\propto \prod_{i=1}^{n} \frac{\exp[-E_i \exp(u_i + v_i) + O_i \log E_i + O_i(u_i + v_i)]}{O_i!} \times \\
&\times \frac{1}{\lambda_v^{\frac{n}{2}}} \exp\left(-\frac{1}{2\lambda_v} \sum_{i=1}^{n} v_i^2 \right) \times \\
&\times \frac{1}{\lambda_u^{\frac{n}{2}}} \exp\left[-\frac{1}{2\lambda_u} \sum_{i=1}^{n} \sum_{j<i} w_{ij} (u_i - u_j)^2 \right] \times \\
&\times \lambda_v^{-(a_v+1)} \exp\left(-\frac{b_v}{\lambda_v} \right) \times \lambda_u^{-(a_u+1)} \exp\left(-\frac{b_u}{\lambda_u} \right).
\end{aligned}$$

Definire la distribuzione condizionale completa di ogni singolo termine del vettore $\boldsymbol{v}$ è immediato; per determinare quella di ogni termine di *clustering* u_i, è necessario considerare che, poiché

$$p(\boldsymbol{u} \mid \lambda_u) = p(u_i \mid u_{-i}, \lambda_u) p(u_{-i} \mid \lambda_u)$$

si ottiene

$$p(u_i \mid u_{-i}, \lambda_u, \boldsymbol{O}) \propto p(O_i \mid u_i) p(u_i \mid u_{-i}, \lambda_u).$$

In particolare la distribuzione condizionale completa per il termine v_i è data da

$$p(v_i \mid u_i, \lambda_v, O_i) \propto \exp\left[O_i v_i - E_i \exp(u_i + v_i) - \frac{v_i^2}{2\lambda_v} \right],$$

mentre la distribuzione condizionale completa per il termine u_i sarà data da

$$p(u_i \mid \boldsymbol{u}_{-i}, v_i, \lambda_u, O_i) \propto \exp\left[O_i u_i - E_i \exp(u_i + v_i) - \frac{w_{i.}(u_i - \bar{u}_i)^2}{2\lambda_u} \right],$$

con $\boldsymbol{u}_{-i}$ che indica il vettore dei valori aggiornati dei parametri di *clustering* escluso quello i-esimo e $\bar{u}_i$ la media dei termini di *clustering* relativi alle aree adiacenti alla i-esima. Si noti che i parametri v_i e u_i non sono direttamente

"collegati" nel grafo in figura 2.3; la dipendenza nelle condizionali complete deriva dalla definizione di $p(O_i \mid u_i, v_i)$ nell'espressione della distribuzione congiunta a posteriori. Poiché le distribuzioni condizionali complete per i termini $\boldsymbol{u}$ e $\boldsymbol{v}$ sono non standard, ma comunque log-concave, si può campionare da esse tramite l'*adaptive rejection sampling* (ARS), il cui funzionamento è descritto nel paragrafo D.2 dell'appendice D.

Le condizionali complete per i termini λ_v e λ_u sono date rispettivamente da

$$p(\lambda_v \mid \boldsymbol{v}) \propto p(\boldsymbol{v} \mid \lambda_v)p(\lambda_v) \qquad \text{e} \qquad p(\lambda_u \mid \boldsymbol{u}) \propto p(\boldsymbol{u} \mid \lambda_u)p(\lambda_u).$$

Quando per l'iperparametro λ_v si assume una distribuzione a priori gamma inversa di parametri a_v e b_v, la distribuzione a posteriori condizionale completa è una gamma inversa con parametri $a_v + n$ e $b_v + \sum_{i=1}^{n} v_i^2/2$; se la distribuzione a priori sull'iperparametro λ_u è una gamma inversa di parametri a_u e b_u, la distribuzione a posteriori condizionale completa è una gamma inversa con parametri $a_u + n/2$ e $b_u + \sum_{i=1}^{n} \sum_{j<i} w_{ij}(u_i - u_j)^2/2$. Nel caso di distribuzioni a priori non informative improprie si ottengono le seguenti condizionali complete

$$p(\lambda_v \mid \boldsymbol{v}) \propto \lambda_v^{-\frac{n}{2}} \exp\left[-\frac{1}{\lambda_v}\left(\epsilon + \frac{1}{2}\sum_{i=1}^{n} v_i^2\right)\right]$$

e

$$p(\lambda_u \mid \boldsymbol{u}) \propto \lambda_u^{-\frac{n}{2}} \exp\left[-\frac{1}{\lambda_u}\left(\epsilon + \frac{1}{2}\sum_{i=1}^{n}\sum_{j<i} w_{ij}(u_i - u_j)^2\right)\right],$$

cioè distribuzioni gamma inversa proprie da cui il campionamento può essere fatto usando gli algoritmi standard. Si noti che nell'algoritmo *Gibbs sampler* saranno campionate, ad ogni iterazione, $2n + 2$ distribuzioni: i due iperparametri λ_u e λ_v ed i termini v_i e u_i, per $i = 1, \ldots, n$.

2.5 Il problema dell'effetto confine nei metodi di stima per il rischio relativo

Quando si analizza la variazione del rischio relativo su una regione, il modello utilizzato è definito su una superficie infinita, ma i dati a disposizione e l'interesse sono limitati ad una parte di questa, quella entro i confini. Tali confini possono essere naturali (la regione è isolata e confina totalmente o in

parte con il mare) o arbitrari (stabiliti in base a ragioni di carattere politico-amministrativo). In quest'ultimo caso, il processo continua oltre i confini e quindi considerare solo una parte di esso può provocare delle distorsioni che devono essere considerate; le aree al confine "perdono dei vicini", ossia è limitato l'insieme delle loro aree adiacenti e quindi viene considerata una matrice di adiacenze limitata alle n aree considerate. In questo modo si ha una mancanza di informazioni relativamente al fenomeno analizzato.

La stima attraverso il rapporto standardizzato di mortalità, non tenendo conto della spazialità del fenomeno, cioè non considerando le interazioni spaziali fra le aree, banalmente non risente del problema dell'effetto confine.

Nel caso di un modello bayesiano gerarchico, la distorsione dovuta all'effetto confine può essere meno elevata quando si assume un modello scambiabile per i rischi relativi: anche se non si considerano le interazioni spaziali tra le aree, conseguentemente all'operazione di *smoothing* viene introdotta una distorsione per l'effetto confine. Infatti, il rischio relativo di ogni area è attratto verso la media generale che non è però valutata esattamente, a causa della limitata informazione riguardo le aree oltre il confine. Comunque la perdita di informazione relativa alle aree esterne al confine è trascurabile quando la proporzione di aree note è elevata. Si noti che lo stesso tipo di problema può presentarsi nel modello bayesiano empirico.

Nel caso si consideri la spazialità dei rischi relativi nel modello bayesiano gerarchico, il metodo può risentire seriamente del problema dell'effetto confine, in quanto le stime del rischio relativo per le aree al confine sono fatte considerando solo una parte dell'informazione circa la spazialità del fenomeno. Tali stime saranno fortemente influenzate dai rischi relativi delle aree a loro adiacenti interne alla regione di studio e non saranno affatto influenzate da quelli delle aree confinanti esterne alla regione. Per le aree al confine, considerando un modello autoregressivo intrinseco per i rischi relativi della regione, si ha un'errata specificazione della distribuzione per il logaritmo del loro rischio relativo condizionata agli altri logaritmi dei rischi relativi, ossia la determinazione della media locale è influenzata da una mancanza di informazioni relative al fenomeno. Tutto ciò può comportare una distorsione nelle stime per i rischi relativi della regione; la distorsione introdotta per le aree al confine è in parte estesa anche alle aree interne, i cui rischi relativi sono stimati considerando una distribuzione condizionata a valori distorti.

Assumendo un modello di convoluzione gaussiana per i logaritmi dei rischi relativi, si uniscono le due fonti di distorsione sopra descritte, ossia l'errata determinazione della media generale dei rischi relativi e l'inesatta

specificazione delle distribuzioni condizionali per le aree al confine.

Considerazioni conclusive

Sono stati descritti i metodi di stima che sono generalmente utilizzati per stimare il rischio relativo di mortalità in una data regione geografica. È stata poi discussa la lacuna che tutti i modelli presentano: essi non considerano i confini, e quindi, le stime da essi fornite, possono risultare distorte per l'effetto confine a causa di una mancanza di informazione sul processo.

Il modello cui si è interessati è quello bayesiano gerarchico a tre stadi, in particolare quello che adotta per i rischi relativi un modello di convoluzione gaussiana. In tale contesto, il problema dovuto alla perdita di informazione per le aree esterne al confine della regione risulta scarsamente considerato e non esistono in letteratura suggerimenti per correggere gli effetti di tale perdita. Nel prossimo capitolo, pertanto, vengono proposte e valutate alcune metodologie atte a risolvere tale problema.

Capitolo 3

Metodi di correzione per l'effetto confine nel modello bayesiano gerarchico

Premessa

Nel capitolo 2 è stato descritto il modello bayesiano gerarchico di Besag, York e Mollié (1991) ed inoltre è stato sottolineato come esso possa fornire delle stime dei rischi relativi distorte per l'effetto confine. In tale senso è quindi necessario considerare delle metodologie che compensino per tale effetto il modello. In questo capitolo sono presentate delle proposte originali che sono poi confrontate e valutate mediante un esperimento di simulazione.

I metodi suggeriti, prendono in parte spunto da quelli già presenti in letteratura per la correzione della stima del coefficiente di autocorrelazione nei modelli autoregressivi spaziali, si veda il capitolo 1.

Per correggere la distorsione introdotta per l'effetto confine nel modello di Besag, York e Mollié, le soluzioni tradizionali, quali eliminare una corona interna dai risultati o considerare la mappa a forma toroidale, o provocano una perdita di informazione eccessiva o sono poco realistiche. Altre soluzioni, quali costruire una corona esterna artificiale, oppure introdurre un sistema di pesi nel modello, sembrano, invece, suggerimenti validi.

Quando, come avviene in alcuni casi, i dati relativi alle aree esterne, ma adiacenti alla regione, sono disponibili, questi possono essere usati per correggere le stime dei rischi relativi delle aree interne; in tale situazione si costruisce una corona esterna empirica. Le aree appartenenti a tale corona sono impiegate per ottenere le stime di quelle interne alla regione, ma la stima che si ottiene su esse non viene considerata, dato che questa è influenzata dall'effetto confine. Si ha così una perdita di informazioni relativa alle aree esterne che in ogni caso non sono di interesse. Le stime dei rischi relativi si

ottengono semplicemente estendendo il modello alle aree della corona esterna, quindi considerando una matrice di adiacenze ed un insieme di dati (casi osservati e attesi) più ampi, cioè un'informazione completa; tale metodo sarà indicato con la sigla CO.

Si noti che la costruzione di una corona esterna empirica rientra nelle proposte, presenti in letteratura, per la correzione per l'effetto confine della stima del coefficiente di autocorrelazione spaziale nei modelli autoregressivi spaziali. In realtà, considerarlo come tale, è forse fuorviante, sembra più giusto considerarlo come un metodo che permette l'uso di un'informazione completa. I risultati che si ottengono da tale metodo sono usati come termine di paragone, quindi per dare, in qualche modo, una misura dell'effetto confine nella stima del rischio relativo di ciascuna area, altrimenti non determinabile, quando si conducano analisi empiriche.

Se non si dispone di informazioni per la corona esterna si può correggere la distorsione per l'effetto confine indotta sulle stime considerando una corona esterna artificiale usata per la corretta specificazione delle distribuzioni condizionali delle aree al confine. Anche in questo caso il modello viene esteso ad una regione più ampia, ma per le aree al confine della regione di interesse si lavora in condizioni di mancanza di informazioni; così la costruzione di tale corona esterna può essere effettuata nell'ambito delle tecniche usate per l'analisi di problemi di dati mancanti.

La distorsione per l'effetto confine si presenta soprattutto nella stima del rischio relativo delle aree poste al confine della regione analizzata. Infatti, esse sono fortemente influenzate dalle stime delle aree a loro confinanti ed interne alla regione e non lo sono affatto dalle stime di quelle esterne. Per operare la correzione per tale effetto si può quindi pensare di introdurre uno schema di pesi che riducano almeno l'influenza delle aree al confine sulla stima dei rischi della regione. Attribuendo un peso minore alle aree che si trovano al confine della regione analizzata (la corona interna della regione), si riduce la loro influenza sulle stime delle aree interne alla regione e delle altre aree poste al confine.

Si noti che l'applicazione dei metodi proposti per la correzione dell'effetto confine dipende dal grado di informazione disponibile circa le aree confinanti. Si possono infatti conoscere o meno, sia il numero di aree adiacenti ed esterne alla regione di interesse, che i casi attesi in tali aree (quindi la struttura della loro popolazione).

3.1 Metodo dei dati pesati

Un semplice approccio per la correzione dell'effetto confine consiste nel pesare i dati in modo da ridurre l'influenza delle aree al confine sulle stime dei parametri del modello.

Tale correzione può essere applicata anche quando si ignorino il numero delle aree fuori confine e la struttura della loro popolazione (quindi i casi attesi). Una volta che il modello sia correttamente specificato, per ogni area della regione viene definito un peso, proporzionale all'informazione disponibile sulle aree a lei confinanti. Quando per un'area si dispone di un'informazione limitata, la stima del rischio relativo è prossima alla stima SMR.

La stima può essere ottenuta usando il modello gerarchico bayesiano ma introducendo, in questo caso, un peso a priori, sia sui casi osservati che su quelli attesi, in modo che l'indice SMR rimanga inalterato; denotiamo tale metodo con la sigla PE.

Come detto il peso deve essere funzione della quantità di informazione mancante e può dipendere dalla natura e dallo scopo dell'analisi. Esso, per ogni area, può essere definito come

$$\text{peso}_i = m(d_i),$$

con $m(\cdot)$ una funzione applicata ad una grandezza d_i. Quale valore d_i può essere assunta la distanza tra il centroide dell'area ed il confine, o il rapporto tra la lunghezza del perimetro dell'area comune al confine esterno della regione l_i^* e la lunghezza del perimetro dell'area l_i; in quest'ultimo caso una semplice scelta della funzione $m(\cdot)$ è quella che definisce il peso come

$$\text{peso}_i = 1 - \frac{l_i^*}{l_i},$$

si noti che, usando tale peso, si può correggere il modello per l'effetto confine anche senza conoscere il numero di aree confinanti ed esterne alla regione.

Un altro schema di pesi può essere la proporzione del numero di aree confinanti non note $w_{i.}^* - w_{i.}$ sul totale delle confinanti $w_{i.}^*$, cioè

$$\text{peso}_i = 1 - \frac{w_{i.}}{w_{i.}^*},$$

ovviamente quando il numero delle aree adiacenti ed esterne alla regione sia noto.

I pesi possono essere anche specificati come funzione della densità della popolazione nelle aree con informazione mancante. Ad esempio, se sono disponibili i casi attesi per le aree oltre il confine, è possibile definire una struttura di pesi che dipenda da questi: ad esempio, per ogni area, il peso adottato è dato dal rapporto tra la somma dei casi attesi nelle aree confinanti note e quella dei casi attesi in tutte le aree confinanti, cioè

$$\text{peso}_i = \sum_{j=1}^{n} (w_{ij} E_j) \,/\, \sum_{j=1}^{n+m} \left(w^*_{ij} E^*_j \right), \tag{3.1}$$

con $\boldsymbol{W}^*$ che indica la matrice delle adiacenze ampliata ad $n+m$ aree ed $\boldsymbol{E}^*$ il vettore dei casi attesi per le $n+m$ aree. Questo può essere considerato come "peso ottimale", in quanto tiene conto di tutta l'informazione disponibile, anche relativamente alla popolazione delle aree esterne.

3.2 Metodi dei dati mancanti

L'altro metodo proposto consiste nell'estensione delle metodologie usate per l'analisi di problemi in cui vi siano dati mancanti; per una trattazione sommaria di tali tecniche si veda l'appendice C.

Gli algoritmi iterativi analizzati in tale appendice sono generalmente adottati al fine di semplificare la trattazione di un problema di dati incompleti, oppure in ambito di imputazione di dati. In questo lavoro vengono presentati degli algoritmi, per vari aspetti originali, con una diversa finalità: ottenere delle stime corrette per l'effetto confine tramite la corretta specificazione del modello.

Innanzitutto il modello è esteso anche alle aree esterne al confine, i cui dati, in questa situazione, risultano mancanti. Ovviamente è necessario che il modello sia correttamente specificato anche per le aree al confine e che il meccanismo di generazione dei dati mancanti possa essere considerato ignorabile secondo la definizione di Rubin (1976), (si veda il paragrafo C.1).

Mediante gli algoritmi che proponiamo la stima dei rischi relativi è fatta attraverso l'imputazione dell'informazione mancante per le aree esterne alla regione: tramite questa si costruisce una regione di guardia esterna sfruttando l'informazione derivante dalla struttura spaziale dei rischi relativi; le quantità ignote delle aree esterne al confine sono trattate come nodi stocastici di un modello grafico bayesiano.

Le variabili ausiliarie inserite nel modello sono convenientemente trattate quali mancanti e su esse si definisce una certa distribuzione condizionale

completa da cui campionare. Si noti che l'interesse non è quello di determinare la distribuzione marginale delle variabili ausiliarie, bensì quella delle variabili del modello originario, una volta che sia stata operata la correzione per l'effetto confine tramite l'esatta specificazione del modello.

Considerando il caso fortunato in cui si conosca la popolazione o il numero dei casi attesi per le aree esterne al confine, l'informazione mancante per tali aree è relativa ai casi osservati. La ricostruzione dell'informazione mancante è fatta sfruttando anche la conoscenza dei casi attesi nelle aree esterne alla regione. Si estende il modello e si applica il metodo *Gibbs sampler*, usando le condizionali complete "estese", cioè condizionate anche ai parametri (le nuove variabili) inseriti nel modello e le condizionali complete di tali parametri.

L'approccio che prevede l'estensione del modello può essere considerato come una derivazione del concetto di *auxiliary variables* di Besag e Green (1993); mentre l'uso di metodi Monte Carlo sul modello esteso prende le mosse dall'algoritmo *data augmentation* proposto da Tanner e Wong (1987).

In questo paragrafo sono proposti tre algoritmi originali che permettono la ricostruzione dell'informazione mancante. Essi sono tra loro alternativi, in questo senso è interessante effettuare una loro analisi comparativa, oltre che la semplice valutazione della loro efficacia.

L'originalità di tali algoritmi sta nel fatto che essi, poichè applicati ad un modello in cui la stima dei parametri è fatta tramite campionamento (*Gibbs sampler*), hanno al loro interno una componente stocastica che si va ad unire a quella introdotta per operare l'imputazione dei dati mancanti.

Gli algoritmi sono presentati prima per il modello gerarchico bayesiano a tre stadi di Besag, York e Mollié considerando una distribuzione per i rischi relativi non definita, poi considerando per essi un modello di convoluzione gaussiana (si veda paragrafo 2.3.3).

Nel seguito, useremo l'indice *obs* per indicare le quantità osservate, *mis* per quelle mancanti e *com* per l'insieme delle due.

3.2.1 Algoritmo EM doppiamente stocastico

Il primo algoritmo proposto deriva dall'algoritmo EM stocastico, differendo da questo per il fatto che oltre ad avere il passo E eseguito tramite campionamento dalla distribuzione predittiva, ha il passo M eseguito mediante campionamento dalle distribuzioni condizionali complete (l'algoritmo *Gibbs sampler*). Per questo motivo abbiamo deciso di denominare tale algoritmo come EM doppiamento stocastico (con la sigla DS).

In particolare l'algoritmo prevede un passo E in cui sono imputati i dati mancanti campionando dalla distribuzione predittiva $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}_{com})$ con i parametri $\boldsymbol{\theta}_{com}$ pari alla loro stima corrente ottenuta al passo M nell'iterazione precedente; la stima dei parametri è fatta mediante l'algoritmo *Gibbs sampler* sul modello esteso, quindi considerando sia i dati osservati che quelli imputati e campionando anche i parametri $\boldsymbol{\theta}_{mis}$, così da ottenere $\boldsymbol{\theta}_{com}$.

Si noti che la distribuzione predittiva $p(O_{mis} \mid O_{obs}, \boldsymbol{\theta}_{com})$ utilizzata per l'imputazione dei dati mancanti, per come è definito il modello, diviene in questo caso $p(\boldsymbol{O}_{mis} \mid \boldsymbol{\theta}_{mis})$, dato che $\boldsymbol{O}_{mis}$ è condizionatamente indipendente da $\boldsymbol{O}_{obs}$ dato $\boldsymbol{\theta}_{mis}$ e indipendente da $\boldsymbol{\theta}_{obs}$.

Il generico passo dell'algoritmo è:

passo E:

estrarre $\boldsymbol{O}_{mis}$ da $p(\boldsymbol{O}_{mis} \mid \boldsymbol{\theta}_{mis})$;

passo M:

estrarre $\boldsymbol{\theta}_{com}$ da $p(\boldsymbol{\theta}_{com} \mid \boldsymbol{O}_{com}, \lambda)$;
estrarre λ da $p(\lambda \mid \boldsymbol{\theta}_{com})$.

In particolare, nel caso si assuma un modello di convoluzione gaussiana per i rischi relativi, il modello esteso è rappresentato, mediante un grafo, nella figura 3.1.

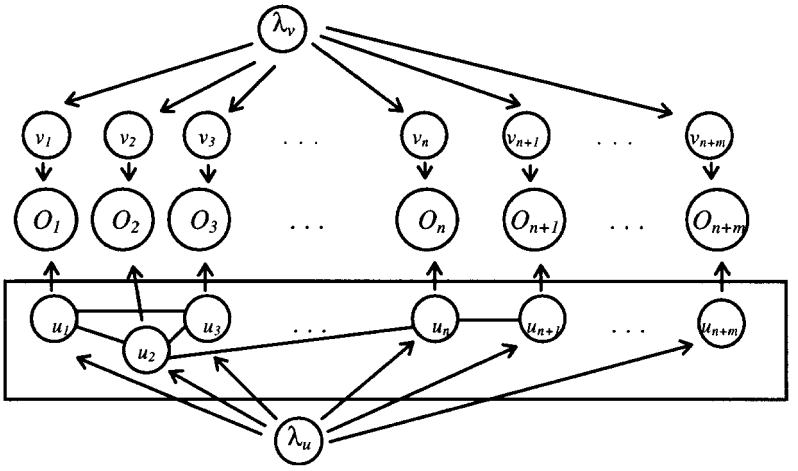


Figura 3.1: Modello bayesiano gerarchico esteso alle aree esterne alla regione

In questo caso, indicando con t la generica iterazione dell'algoritmo di imputazione dei dati, $j = 1, \ldots, J$ la generica iterazione dell'algoritmo di campionamento *Gibbs sampler*, con $i = 1, \ldots n$ le aree note e $i = n+1, \ldots, n+m$ quelle ignote, l'algoritmo può essere formulato come segue. Siano dati i valori di E_i per $i = 1, \ldots, n+m$, O_i per $i = 1, \ldots, n$. La distribuzione predittiva da cui campionare O_i è una distribuzione di Poisson di media $E_i\theta_i$, con $i = n+1, \ldots, n+m$. I valori $u_i^{(t-1)}$ e $v_i^{(t-1)}$ nel passo E sono dati dalla media della distribuzione corrispondente campionata al passo M dell'iterazione precedente. Alla prima iterazione $u_i^{(0)}$ e $v_i^{(0)}$ possono essere considerati nulli, quindi assumendo dei rischi relativi di partenza unitari; questo non influenza il risultato finale di stima dell'algoritmo.

Per il passo M della prima iterazione è necessario fissare i valori iniziali per i vettori di dimensione $n+m$ $\boldsymbol{v}^{0(1)}$, $\boldsymbol{u}^{0(1)}$ e gli scalari $\lambda_v^{0(1)}$ e $\lambda_u^{0(1)}$; per le iterazioni successive, il valore di partenza per ogni sequenza campionata può essere posto uguale all'ultimo campionato o alla media di tutti i valori campionati, nell'iterazione precedente. Le distribuzioni da cui campionare sono definite nel paragrafo 2.4. Con $\boldsymbol{u}_{-i}^{j,j-1(t)}$ si considera il vettore dei termini di *clustering* correnti relativi alle aree adiacenti con l'area i-esima, alcuni aggiornati nella iterazione j dell'algoritmo *Gibbs sampler* altri da aggiornare (quindi campionati all'iterazione $j-1$).

La generica iterazione t-esima è data da

passo E:

$$\begin{array}{lll} \text{estrarre } O_i^{(t)} & \text{da } p(O_i^{(t)} \mid u_i^{(t-1)}, v_i^{(t-1)}), & \\ & \text{per } i = n+1, \ldots, n+m; & \end{array}$$

passo M:

$$\begin{array}{lll} \text{estrarre } v_i^{j(t)} & \text{da } p(v_i^{j(t)} \mid u_i^{j-1(t)}, \lambda_v^{j-1(t)}, O_i^{(t)}), & \\ \text{estrarre } u_i^{j(t)} & \text{da } p(u_i^{j(t)} \mid \boldsymbol{u}_{-i}^{j,j-1(t)}, v_i^{j(t)}, \lambda_u^{j-1(t)}, O_i^{(t)}), & \\ & \text{per } i = 1, \ldots, n+m; & \\ \text{estrarre } \lambda_v^{j(t)} & \text{da } p(\lambda_v^{j(t)} \mid \boldsymbol{v}^{j(t)}); & \\ \text{estrarre } \lambda_u^{j(t)} & \text{da } p(\lambda_u^{j(t)} \mid \boldsymbol{u}^{j(t)}). & \end{array}$$

Per stabilire la convergenza dell'algoritmo, sono confrontate, per ogni parametro del modello, le sequenze *Gibbs* ottenute al passo M di un certo numero T di iterazioni consecutive. Il confronto è stato fatto sulla base dell'analisi delle componenti di varianza, in modo analogo a quanto fatto, in un altro contesto, da Gelman e Rubin (1992a, 1992b). L'adozione di un metodo ori-

ginale per stabilire la convergenza dell'algoritmo è resa necessaria dal fatto che non è possibile ottenere la convergenza ad un valore in due iterazioni successive né per i dati imputati né per le stime dei parametri; questo a causa della natura doppiamente stocastica dell'algoritmo. Si noti, peraltro, che la convergenza ad un valore delle stime di tutti i parametri, non è necessariamente ottenuta nel caso di convergenza dell'algoritmo, dato che l'algoritmo di campionamento potrebbe non aver esplorato l'intera distribuzione.

Gelman e Rubin (1992a, 1992b), Gelman *et al.* (1995, pag. 331–332), Gelman (1996) pag. 136–139, utilizzano un'analisi delle componenti di varianza per stabilire la convergenza ad una stessa distribuzione a posteriori di più sequenze parallele, cioè campionate (per esempio mediante un algoritmo *Gibbs sampler*) da una stessa distribuzione ma considerando, nell'algoritmo di campionamento, dei valori iniziali diversi. Si denoti il generico parametro del modello bayesiano con θ; per esso si possiedono T sequenze campionate di lunghezza J, cioè i valori θ_{tj} con $t = 1, \ldots, T$ e $j = 1, \ldots, J$. Per ogni parametro vengono determinate la varianza tra le T medie delle sequenze di lunghezza J

$$B = \frac{J}{T-1} \sum_{t=1}^{T} (\bar{\theta}_{t.} - \bar{\theta}_{..})^2$$

con

$$\bar{\theta}_{t.} = \frac{1}{J} \sum_{j=1}^{J} \theta_{tj} \qquad \text{e} \qquad \bar{\theta}_{..} = \frac{1}{T} \sum_{t=1}^{T} \bar{\theta}_{t.}$$

e la media delle varianze entro le sequenze

$$W = \frac{1}{T} \sum_{t=1}^{T} s_t^2 \qquad \text{con} \qquad s_t^2 = \frac{1}{J-1} \sum_{j=1}^{J} (\theta_{tj} - \bar{\theta}_{t.})^2.$$

Per analizzare le sequenze si può considerare il valore dell'indice

$$\hat{R} = \left(\frac{B}{J} + \frac{(J-1)}{J} W \right) / W.$$

Tale indice è dato dal rapporto

$$R = \frac{\sigma_B^2 + \sigma_W^2}{\sigma_W^2}$$

dove σ_B^2 rappresenta la varianza tra le sequenze e σ_W^2 la varianza dovuta alla componente di errore casuale. Una stima di questi è data da

$$\hat{\sigma_B^2} = \frac{B - W}{J} \qquad \text{e} \qquad \hat{\sigma_W^2} = W$$

poiché

$$\mathrm{E}(B) = \sigma_W^2 + J\sigma_B^2 \qquad \text{e} \qquad \mathrm{E}(W) = \sigma_W^2,$$

con B e W (la media dei quadrati MS, *Mean Squares*) corrispondenti alle due componenti di varianza (si veda ad esempio Winer, 1971). L'indice R è dato dal rapporto tra due stime della varianza: al numeratore una sovrastima della varianza nell'ipotesi che i punti di partenza siano campionati da una distribuzione sovradispersa rispetto alla distribuzione a posteriori di interesse (una stima corretta se è ugualmente dispersa o la lunghezza di ogni singola sequenza tende all'infinito) e al denominatore la stima della varianza entro le sequenze che rappresenta una sottostima della varianza, dato che la sequenza può non aver "esplorato" completamente la distribuzione a posteriori. Le sequenze sono campionate da una stessa distribuzione, quindi è stata raggiunta la convergenza dell'algoritmo di campionamento (*Gibbs sampler*), quando la statistica R assume valori prossimi ad 1; nella pratica viene utilizzato quale valore di soglia 1.2, esso, infatti, si è rivelato in molti casi un criterio diagnostico adeguato.

In questo caso il test è applicato per individuare se per ogni parametro, da un certo numero di iterazioni in poi, le sequenze *Gibbs* siano o meno campionate da una stessa a posteriori. Quando ciò avviene il test di Gelman e Rubin deve necessariamente segnalarlo, in quanto sotto l'ipotesi di convergenza dell'algoritmo di imputazione $\sigma_B^2 = 0$ e quindi si ottengono valori unitari per l'indice R. In pratica, per ogni parametro del modello esteso, si applica il test prima su tutte le T sequenze, poi scartando ad una ad una quelle iniziali, sino a quando non si ottiene un valore per l'indice R inferiore a 1.2.

3.2.2 Algoritmo EM con passo M stocastico

Il secondo algoritmo è una semplificazione del precedente; in questo caso l'imputazione dei dati mancanti avviene usando una misura di locazione (ad esempio la media) della distribuzione predittiva, dati i valori dei parametri fissati alla loro stima corrente. In questo caso l'algoritmo, che chiameremo EM stocastico nel passo M (con la sigla MS), ha una componente stocastica nel passo M di massimizzazione. Quindi differisce dall'algoritmo EM stocastico per il fatto che la procedura di campionamento è nel passo M anziché nel passo E e dall'EM-tipo per il fatto che il passo M è eseguito mediante campionamento.

Questa volta, ad ogni passo E, sono imputati i valori per i dati mancanti $\boldsymbol{O}_{mis} = \text{int}(\boldsymbol{E}_{mis}\boldsymbol{\theta}_{mis})$, con $\text{int}(\cdot)$ che indica l'operatore valore intero. Nel caso si assuma un modello di convoluzione gaussiana per i rischi relativi, nel passo E sono considerati i vettori $\boldsymbol{v}^{(t-1)}$ e $\boldsymbol{u}^{(t-1)}$ dati ancora dalla media dei valori campionati nel passo M dell'iterazione precedente; fissando valori nulli per i vettori $\boldsymbol{v}^{(0)}$ e $\boldsymbol{u}^{(0)}$, alla prima iterazione dell'algoritmo, sarà imputata la quantità $\boldsymbol{O}_i^{(1)} = \text{int}(\boldsymbol{E}_i)$ per $i = n+1, \ldots, n+m$.

Il passo generico t dell'algoritmo, analizzato nella formulazione relativa al modello di convoluzione gaussiana per i rischi relativi, è in questo caso:

passo E:

$$\begin{array}{lll} \text{porre } O_i^{(t)} = & \text{int}(E_i \exp(u_i^{(t-1)} + v_i^{(t-1)})), & \\ & \text{per } i = n+1, \ldots, n+m; & \end{array}$$

passo M:

$$\begin{array}{lll} \text{estrarre } v_i^{j(t)} & \text{da } p(v_i^{j(t)} \mid u_i^{j-1(t)}, \lambda_v^{j-1(t)}, O_i^{(t)}), & \\ \text{estrarre } u_i^{j(t)} & \text{da } p(u_i^{j(t)} \mid \boldsymbol{u}_{-i}^{j,j-1(t)}, v_i^{j(t)}, \lambda_u^{j-1(t)}, O_i^{(t)}), & \\ & \text{per } i = 1, \ldots, n+m; & \\ \text{estrarre } \lambda_v^{j(t)} & \text{da } p(\lambda_v^{j(t)} \mid \boldsymbol{v}^{j(t)}); & \\ \text{estrarre } \lambda_u^{j(t)} & \text{da } p(\lambda_u^{j(t)} \mid \boldsymbol{u}^{j(t)}). & \end{array}$$

La regola di arresto di questo algoritmo è data dall'ugualianza tra le osservazioni imputate in due iterazioni consecutive. In tal caso le distribuzioni condizionali complete a posteriori da cui si campiona nelle due iterazioni sono identiche.

3.2.3 Algoritmo Chained Data Augmentation a tre vettori

Il terzo algoritmo prevede l'imputazione delle osservazioni all'interno del *Gibbs sampler*: le osservazioni per le aree esterne sono viste come parametri del modello anziché come dati. In tale senso l'algoritmo può essere implementato semplicemente estendendo la procedura di campionamento dell'algoritmo *Gibbs sampler* anche a tali parametri; la distribuzione condizionale completa per i parametri $\boldsymbol{O}_{mis}$,è del tutto analoga alla distribuzione predittiva.

Tale algoritmo deriva dell'algoritmo *chained data augmentation*; mentre originariamente l'algoritmo è formulato per due vettori, cioè un vettore di dati mancanti ed un vettore di parametri, nella versione proposta è esteso a tre vettori, per questo motivo abbiamo denotato tale algoritmo come *chained data augmentation* a tre vettori (con la sigla CH). Il generico passo dell'algo-

ritmo, quando non sia stato specificato il modello sul rischio relativo, è dato da

$$\begin{array}{ll} \text{estrarre } \boldsymbol{O}_{mis} & \text{da } p(\boldsymbol{O}_{mis} \mid \boldsymbol{\theta}_{mis}); \\ \text{estrarre } \boldsymbol{\theta}_{com} & \text{da } p(\boldsymbol{\theta}_{com} \mid \boldsymbol{O}_{com}, \lambda); \\ \text{estrarre } \lambda & \text{da } p(\lambda \mid \boldsymbol{\theta}_{com}). \end{array}$$

Nel caso si assuma un modello di convoluzione gaussiana per i rischi relativi, l'algoritmo di imputazione, con $j = 1, \ldots, J$ le iterazioni dell'algoritmo *Gibbs sampler*, diviene

$$\begin{array}{ll} \text{estrarre } O_i^j & \text{da } p(O_i^j \mid v_i^{j-1}, u_i^{j-1}), \\ & \qquad \text{per } i = n+1, \ldots, n+m; \\ \text{estrarre } v_i^j & \text{da } p(v_i^j \mid u_i^{j-1}, \lambda_v^{j-1}, O_i^j), \\ \text{estrarre } u_i^j & \text{da } p(u_i^j \mid \boldsymbol{u}_{-i}^{j,j-1}, v_i^j, \lambda_u^{j-1}, O_i^j), \\ & \qquad \text{per } i = 1, \ldots, n+m; \\ \text{estrarre } \lambda_v^j & \text{da } p(\lambda_v^j \mid \boldsymbol{v}^j); \\ \text{estrarre } \lambda_u^j & \text{da } p(\lambda_u^j \mid \boldsymbol{u}^j). \end{array}$$

Per tale algoritmo sarà necessario fissare i valori iniziali di $\boldsymbol{v}^0$, $\boldsymbol{u}^0$, λ_v^0 e λ_u^0. La convergenza di tale algoritmo è stata verificata applicando il test di Geweke (si veda il paragrafo D.3) a tutti i parametri del modello esteso, osservazioni mancanti comprese. Infatti, poiché l'algoritmo *chained data augmentation* a tre vettori è traducibile in un algoritmo *Gibbs sampler*, è sembrato naturale adottare, quale test di convergenza, uno dei test maggiomente usati per verificare la convergenza di quest'ultimo.

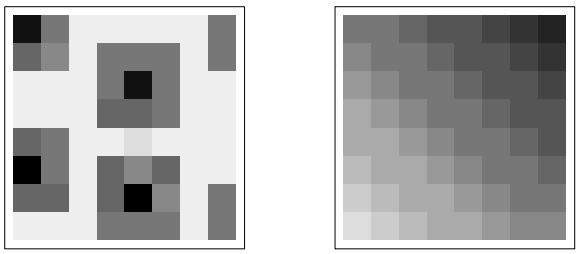
3.3 Valutazione dei metodi di correzione tramite simulazioni

Sono considerate 100 mappe quadrate 10×10 (la regione di interesse è il quadrato 8×8 con 64 aree, le altre 36 rappresentano la corona esterna), ogni area $i = 1, \ldots, 100$ ha un numero di casi osservati pari ad un valore simulato da una variabile Poisson$(10\, \theta_i^{ve})$, usando per i rischi relativi (rappresentati dai parametri θ_i^{ve}) sia uno schema spaziale a *cluster* che a *trend* lineare, con θ_i^{ve} a vari livelli: da un minimo di 0.65 ad un massimo di 2.10. Le mappe dei rischi relativi θ_i^{ve} relative ai due schemi spaziali sono rappresentate nella figura 3.2.a (schema a *cluster*) e nella figura 3.2.b (schema a *trend*); i livelli di grigio rappresentano i valori del rischio relativo da un valore minimo di 0.65 ad un massimo di 2.10; l'utilizzo di più categorie di grigio rispetto a quanto viene in

genere fatto per i cartogrammi, permette di ottenere maggiore informazione dalle mappe che sono analizzate.

Per le adiacenze è stato considerato uno schema "a regina": ogni area del reticolo regolare confina quindi con otto aree, eccetto le aree poste al confine che confinano con tre o cinque aree. Per le simulazioni sono stati scelti questi due schemi, in quanto essi rappresentano due situazioni limite; infatti, le strutture spaziali con cui si manifesta generalmente il rischio relativo di mortalità, possono essere considerate intermedie a questi; ad esempio, dai risultati dell'atlante toscano di mortalità, si vede che il tumore allo stomaco o alla laringe si manifesta seguendo uno schema del rischio relativo sulla Toscana prossimo a quello *trend*, mentre per altre cause di morte, ad esempio il tumore al polmone femminile, si ha uno schema più vicino a quello *cluster*.

Le stime sono ottenute per:
- le 100 mappe di 64 aree, con informazione incompleta, ottenendo θ_{ij}^{in} per ogni area i e simulazione j;
- le 100 mappe di 100 aree, ossia il caso in cui si ha un'informazione completa con l'introduzione nell'analisi una corona esterna empirica, ottenendo θ_{ij}^{co} per ogni area i e simulazione j;
- le 100 mappe di 64 aree, con la corona interna pesata, i pesi sono 1, 3/8 e 5/8 dati dalla proporzione delle aree adiacenti note sul totale delle adiacenti, ottenendo θ_{ij}^{pe} per ogni area i e simulazione j;
- le 100 mappe di 100 aree usando il metodo che prevede l'imputazione dei dati tramite il valore atteso della distribuzione predittiva (algoritmo EM stocastico nel passo M), ottenendo θ_{ij}^{ms} per ogni area i e simulazione j;
- le 100 mappe di 100 aree usando il metodo che prevede l'imputazione dei dati tramite l'algoritmo *chained data augmentation* a tre vettori, ottenendo θ_{ij}^{ch} per ogni area i e simulazione j.



(a) (b)

Figura 3.2: Schema spaziale a *cluster* (a) ed a *trend* (b)

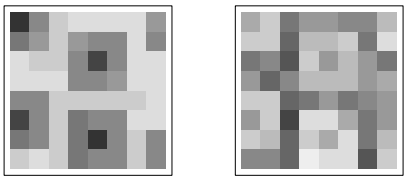
La figura 3.3 e la figura 3.4, rispettivamente per il caso di schema spaziale a *cluster* e a *trend*, mostrano le mappe delle medie delle stime (la media è fatta sulle 100 mappe, quindi sull'indice j) ottenute con i diversi metodi: $\bar{\theta}_i^{in}$ con dati incompleti (a), $\bar{\theta}_i^{co}$ con dati completi (b), $\bar{\theta}_i^{pe}$ con dati pesati (c), $\bar{\theta}_i^{ms}$ con dati imputati tramite la media della distribuzione predittiva (usando l'algoritmo EM stocastico nel passo M) (d) e $\bar{\theta}_i^{ch}$ usando l'algoritmo *chained data augmentation* a tre vettori (e), per $i = 1, \ldots, 64$. Le mappe medie sono 8×8, cioè riguardano la regione di interesse, e sono mostrate seguendo una scala di grigio della stessa gradazione di quella usata precedentemente.

Per il metodo DS che prevede l'imputazione dell'informazione tramite campionamento (algoritmo EM doppiamente stocastico), non è stato effettuato l'esperimento di simulazione data l'impossibilità di stabilire la convergenza dell'algoritmo per ciscuna delle 100 mappe in modo automatico, occorre infatti analizzare attentamente i risultati dell'analisi delle componenti di varianza; inoltre, teoricamente, il metodo fornisce gli stessi risultati del metodo *chained data augmentation* a tre vettori.

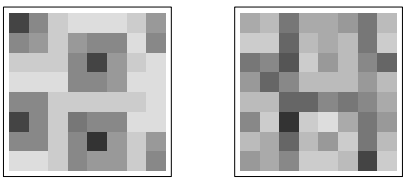
Nella figura 3.3 e nella figura 3.4 sono mostrate, a fianco delle mappe medie, anche le relative mappe degli errori nel caso di uno schema spaziale *cluster* e a *trend*. Esse sono relative ai risultati medi ottenuti applicando i diversi metodi di correzione; il confronto è fra la mappa media ottenuta per ogni metodo di correzione proposto e la mappa vera; denotando con $\mathrm{abs}(\cdot)$ l'operatore valore assoluto, si ha

$$\frac{\mathrm{abs}(\bar{\theta}_i^{*} - \theta_i^{ve})}{\theta_i^{ve}},$$

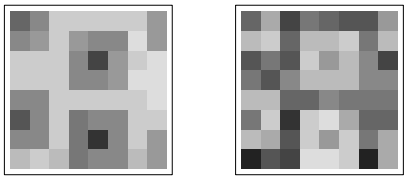
con $\bar{\theta}_i^{*}$ il valore del rischio relativo attribuito all'area i-esima determinato come media dei 100 valori ottenuti per le 100 mappe, stimati con il generico metodo, indicato con il simbolo $*$. Le mappe degli errori seguono una gradazione di grigio che va da un minimo di 0 ad un massimo di 0.40. Sono state mantenute delle scale di grigio fisse per operare un diretto confronto tra le mappe vere e quelle medie, per lo stesso motivo sono usati livelli di grigio fissi, se pur diversi da quelli delle mappe medie, per tutte le mappe degli errori. Dall'analisi delle mappe medie si nota come tutti i metodi ricostruiscano, in media, la vera mappa (quella della figura 3.2). I metodi di correzione che prevedono l'imputazione dell'informazione mancante sembrano dare migliori risultati rispetto a quello che prevede l'introduzione di uno schema di pesi a priori nel modello, questo è soprattutto evidente nel caso di uno schema spaziale a *trend*.



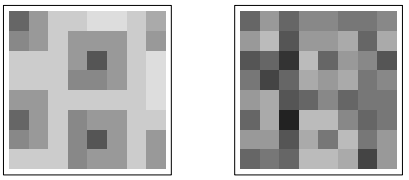
(a)



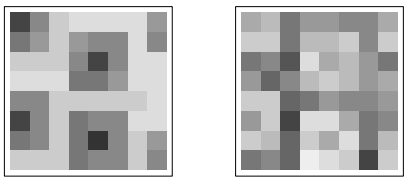
(b)



(c)

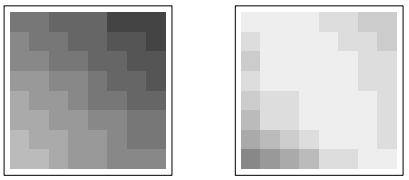


(d)

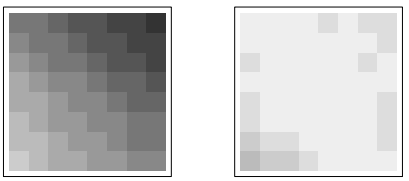


(e)

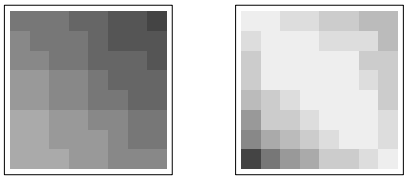
Figura 3.3: Mappe medie e mappe degli errori per lo schema spaziale a *cluster*, metodo IN (a), CO (b), PE (c), MS (d) e CH (e)



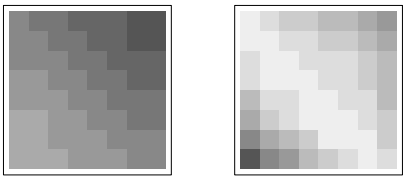
(a)



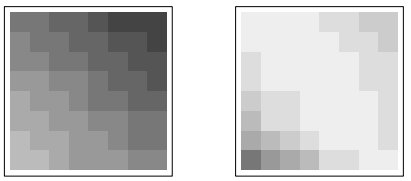
(b)



(c)



(d)



(e)

Figura 3.4: Mappe medie e mappe degli errori per lo schema spaziale a *trend*, metodo IN (a), CO (b), PE (c), MS (d) e CH (e)

Dei due metodi di imputazione analizzati, quello che opera tramite un algoritmo *chained data augmentation* a tre vettori sembra comportarsi meglio, in entrambi gli schemi.

È interessante osservare come questo metodo influenzi i risultati solo limitatamente alle aree poste al confine del reticolo; lo stesso non avviene utilizzando il metodo di pesatura dei dati, né tantomeno imputando l'informazione mancante tramite la media della distribuzione predittiva.

Si noti, inoltre, analizzando la mappa relativa agli errori nella figura 3.4.a, che quando si applica il modello bayesiano gerarchico ai dati incompleti nel caso di uno schema spaziale a *trend*, l'effetto confine segue un andamento lineare ai confini della regione analizzata.

Le mappe medie analizzate aiutano soltanto a dare un'idea di come funzionino ed in quali aree intervengano i metodi correzione, senza considerare cosa in realtà succeda per ciascuna delle mappe analizzate. Per meglio valutare questi metodi, è necessario adottare delle misure che considerino i risultati ottenuti su tutte le 100 mappe, e non solo l'immagine della media di tali risultati; sono stati quindi usati i seguenti quattro indici

$$\text{TMSE} = \sum_{i=1}^{64}\sum_{j=1}^{100}\frac{(\theta^*_{ij} - \theta^{ve}_i)^2}{100} \qquad \text{BMSE} = \sum_{i=1}^{28}\sum_{j=1}^{100}\frac{(\theta^*_{ij} - \theta^{ve}_i)^2}{100}$$

$$\text{TVAR} = \sum_{i=1}^{64}\sum_{j=1}^{100}\frac{(\theta^*_{ij} - \bar{\theta}^*_i)^2}{99} \qquad \text{BVAR} = \sum_{i=1}^{28}\sum_{j=1}^{100}\frac{(\theta^*_{ij} - \bar{\theta}^*_i)^2}{99}$$

dove la sommatoria per i è fatta sulle 64 aree della regione di interesse (per gli indici TMSE e TVAR) o sulle 28 aree della corona interna (per gli indici BMSE e BVAR); θ^*_{ij} indica la stima del rischio relativo, ottenuta adottando il metodo * sulla i-esima area nella j-esima mappa simulata e $\bar{\theta}^*_i$ il loro valore medio. I risultati sono mostrati nella tabella 3.1, riportata sotto.

Dai valori dell'indice TMSE si nota che il metodo di correzione che prevede l'imputazione dell'informazione mancante fornisce una mappa che si discosta dalla mappa reale, in misura quasi uguale a quella ottenuta usando il modello gerarchico bayesiano con la completa informazione. Lo stesso si ha considerando l'indice BMSE dato dalla somma degli scarti quadratici medi solo per il bordo della regione analizzata; il metodo di correzione che ricostruisce l'informazione mancante diminuisce notevolmente gli errori commessi ignorandola; la ricostruzione è fatta esattamente. Il metodo che prevede l'inserimento di uno schema di pesi permette una limitata correzione della distorsione che si ha usando un'informazione incompleta.

Tabella 3.1: Valori degli indici TMSE, BMSE, TVAR e BVAR per i vari metodi e per i due diversi schemi spaziali

Schema spaziale	Metodi	TMSE	BMSE	TVAR	BVAR
cluster	IN	21.709	6.756	3.268	1.535
	CO	4.365	1.851	2.782	1.218
	PE	20.133	5.700	2.563	0.947
	MS	5.136	2.330	2.119	0.948
	CH	4.626	2.073	3.257	1.463
trend	IN	6.173	2.950	1.347	0.721
	CO	1.083	0.525	1.023	0.476
	PE	5.675	2.542	0.992	0.400
	MS	1.368	0.878	0.624	0.312
	CH	1.592	0.856	1.336	0.650

Tramite gli indici TVAR e BVAR si può considerare cosa succede per le 100 mappe, simulate e stimate con i 4 diversi metodi. Si noti che il metodo che imputa l'informazione usando la media della distribuzione predittiva provoca un abbassamento degli indici TVAR e BVAR; ciò indica che il metodo tende a dare stessi risultati per tutte le 100 mappe stimate, a causa del fatto che, adottando tale algoritmo, non viene considerata la variabilità della distribuzione predittiva. Si noti, quindi, che tale metodo fornisce delle eccellenti stime, su cui però è impossibile costruire metodi inferenziali, dato che la loro variabilità è sottostimata.

Considerazioni conclusive

I metodi proposti danno buoni risultati sia nel caso di uno schema spaziale a *cluster* che a *trend*. Si noti che per lo schema a *cluster*, si ha un notevole aumento degli indici TMSE e BMSE, dovuta alla pessima qualità della ricostruzione della mappa vera che si ottiene mediante il modello bayesiano (con e senza correzioni). Per contro, per lo schema a *trend*, la qualità dei risultati è senz'altro accettabile. Alla luce dei risultati ottenuti dall'esperimento di simulazione, il metodo che fornisce i miglior risultati è l'algoritmo *chained data augmentation* a tre vettori. Esso, infatti, ricostruisce, in media, la mappa usata per le simulazioni in maniera analoga a quanto fatto dal modello quando sia considerata l'informazione completa ed inoltre influenza soltanto le aree al confine della regione, senza apportare modifiche per la stima relativa alla aree interne. Teoricamente i risultati forniti dall'algorit-

mo *chained data augmentation* a tre vettori sono gli stessi di quelli ottenuti applicando l'algoritmo EM doppiamente stocastico; si noti però che il primo è computazionalmente meno dispendioso.

Nel prossimo capitolo i metodi sono applicati ad un caso reale: la stima del rischio di mortalità per tumore allo stomaco per la regione Toscana nel periodo 1981–1988. I metodi di correzione, in base ai risultati ottenuti sugli esperimenti di simulazione presentati in questo capitolo, sembrano adatti per tale esempio, dato che, come detto, da studi precedenti il rischio per tumore allo stomaco nella Toscana si presenta con uno schema spaziale prossimo a quello a *trend* lineare.

Si noti che i risultati desunti dalle simulazioni proposte sono validi quando si consideri un reticolo regolare e una struttura della popolazione fissa per ogni area. Alla luce di tali considerazioni, applicando i metodi di correzione ad un caso reale definito su un reticolo irregolare e con aree aventi una diversa struttura ed entità di popolazione tra loro, si possono ottenere risultati leggermente diversi. In tale caso, infatti, giocano un ruolo importante sia una struttura di adiacenze più complessa, sia la struttura e la densità della popolazione su cui si basa sia il metodo che prevede la determinazione di un sistema di pesi da applicare ai dati che il metodo che prevede la ricostruzione dell'informazione mancante.

Tutti i metodi proposti sono stati implementati mediante programmi in linguaggio FORTRAN, usati sia per l'esperimento di simulazione che per l'applicazione ai dati reali relativi alla mortalità nella regione Toscana.

Capitolo 4

L'atlante di mortalità toscano

Premessa

Per stimare i rischi relativi di morte per la popolazione residente nella regione Toscana riportati nell'atlante di mortalità toscano 1971–1994, è stato usato il modello bayesiano gerarchico di Besag, York e Mollié (1991) (si veda paragrafo 2.3), nella versione originale e in quella corretta per l'effetto confine seguendo i metodi proposti nel capitolo 3.

Dopo la descrizione, nel primo paragrafo, dei dati disponibili e dei materiali utilizzati per le analisi relative all'atlante, nel secondo paragrafo sono analizzati i risultati ottenuti per il caso particolare del tumore allo stomaco per la popolazione maschile nel periodo 1981–1988. La validità dei metodi di correzione applicati è valutata attraverso il reperimento dell'informazione sulle aree circostanti le zone di confine per il periodo considerato: ossia i dati relativi ai comuni della Liguria, dell'Emilia Romagna, delle Marche, dell'Umbria e del Lazio confinanti con quelli della Toscana. Tramite tale informazione si ottengono delle stime del rischio relativo (metodo indicato nel capitolo 3 con la sigla CO) utili per dare una misura dell'effetto confine; i risultati ottenuti dal metodo con o senza correzione per l'effetto confine, sono infatti confrontati con questi.

4.1 Fonti e materiali

Per ogni comune della Toscana è disponibile il numero di decessi, distinti per sesso e per causa, avvenuti dal 1971 al 1994. I dati utilizzati derivano da quelli raccolti dall'Istituto Centrale di Statistica (ISTAT) in occasione della rilevazione annuale della mortalità; ossia da quelli ricavati dai certificati di

morte raccolti correntemente dall'ISTAT. Soltanto alcune delle informazioni presenti nelle schede ISTAT, utilizzate per la raccolta dei dati relativi al defunto, sono però disponibili: l'età, il sesso, la causa di morte, il comune in cui è avvenuto il decesso e, per quanto riguarda la residenza, sino al 1980, solo la provincia quando il decesso è avvenuto in un comune diverso da quello di residenza. Usando tali dati, la mortalità per ogni singolo comune di residenza non è immediatamente ottenibile. È possibile, infatti, individuare esattamente soltanto i residenti deceduti nel proprio comune di residenza e non coloro che sono deceduti in un comune diverso, dato che per essi è riportata soltanto l'informazione relativa alla provincia di residenza e non del comune. D'altra parte presso i comuni non sono disponibili notizie sulla causa di morte di coloro che sono morti in un comune diverso da quello di residenza, poiché la causa di morte è registrata nel comune di decesso ma non è trasmessa a quello di residenza. Poiché l'ISTAT fornisce i dati a livello comunale solo a partire dall'anno 1980, per il periodo 1971–1979 è stato necessario operare una disaggregazione dei dati, basandosi sulle informazioni disponibili presso le anagrafi comunali. Per riattribuire i deceduti in un comune diverso da quello di residenza, al comune di residenza stesso, sono stati rilevati, a livello comunale, alcuni dati anagrafici su tali individui, che grazie a tali informazioni, vengono individuati tra i deceduti i cui dati sono forniti dall'ISTAT. Grazie a tale operazione di disaggregazione, l'unità areale, cioè l'area geografica che viene utilizzata quale unità elementare nell'analisi, è quella comunale; essa rappresenta la più piccola area amministrativa per cui sono disponibili i dati di mortalità. Poiché tale operazione è stata fatta, in ambito italiano, solo per la redazione di quest'atlante, esso è l'unico che ha dei risultati ad un tale livello di definizione.

Nella figura 4.1 è rappresentata la suddivisione territoriale della regione Toscana nelle 9 province e i 287 comuni. Si noti che questa è relativa alla situazione antecedente alla costituzione della provincia di Prato; per maggiori dettagli su tale suddivisione e sulle adiacenze considerate tra i comuni, si rimanda alla lettura dell'appendice E.

Le cause di morte sono classificate in base alla Classificazione Internazionale delle Malattie (*International Classification of Diseases*, ICD); in particolare, in base all'ottava revisione della classificazione (ICDVIII) per gli anni 1971–1978 e in base alla nona (ICDIX) per gli anni 1979–1994. Dato che, nel corso del periodo cui sono relativi i dati analizzati interviene il cambiamento nella codifica della patologia, è stato necessario un raccordo per rendere omogenea la classificazione. Peraltro la lunghezza del periodo analizzato,

pur presentando problemi per quanto riguarda la coerenza delle informazioni utilizzate, rappresenta sicuramente un pregio ed una innovazione rispetto ad esperienze precedenti che considerano periodi più limitati. Le cause di morte analizzate per entrambi i sessi e la relativa codifica ICDVIII e ICDIX è riportata nella tabella E.2.

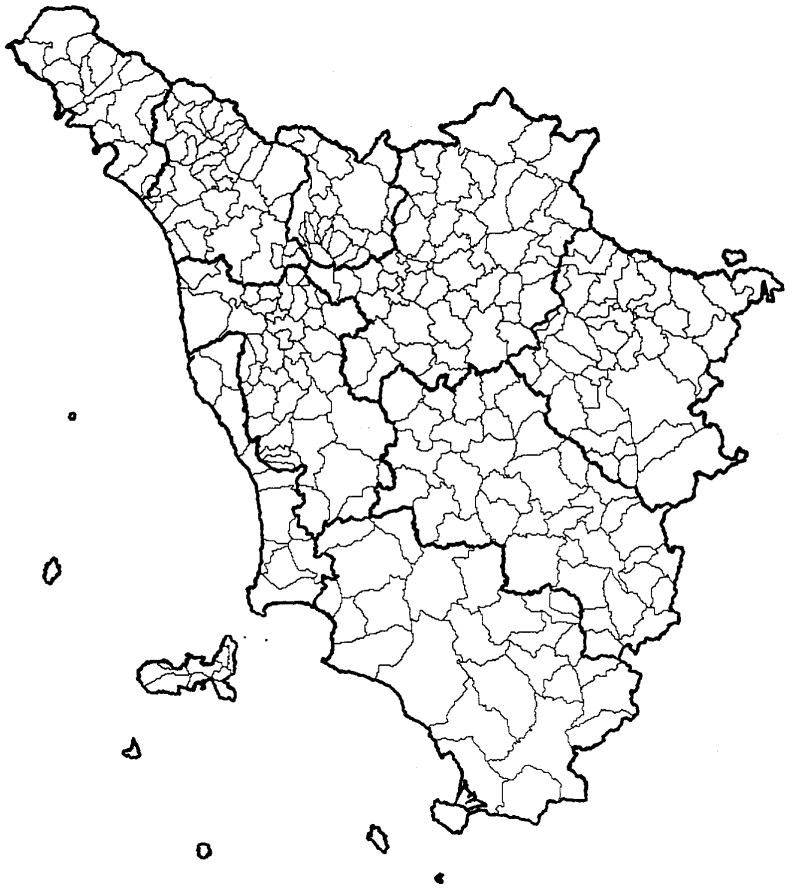


Figura 4.1: Mappa a livello comunale della Toscana

La determinazione dei casi attesi è stata effettuata in modo indiretto: per ogni comune, il numero di casi attesi per sesso e classe di età è stato ottenuto moltiplicando i tassi specifici osservati nella popolazione presa come standard (in questo caso la Toscana nel periodo 1971–1994 distinguendo i tassi per causa, sesso e classe di età) per la popolazione del comune stimata nelle varie fasce di età per entrambi i sessi. Questi valori sono stati sommati per

tutte le età ottenendo così, per maschi e femmine, il numero di decessi per area che ci si sarebbe attesi se la popolazione in esame avesse sperimentato per ogni classe di età la stessa mortalità verificatasi in quella di riferimento. D'interesse sarà il discostarsi da tale valore, quindi l'individuazione di zone che hanno al loro interno aree a basso o elevato rischio.

I dati relativi alla struttura per età e sesso della popolazione sono rilevati dall'ISTAT solo in occasione dei censimenti. Utilizzando quest'informazione e quella delle risultanze anagrafiche (il totale dei residenti e le componenti dinamiche della popolazione, cioè movimento naturale e migratorio), l'ISTAT calcola e pubblica, per gli anni intercensuali, una stima della struttura delle popolazioni regionali ma non di quelle provinciali e comunali. A colmare questa lacuna sono intervenuti diversi istituti di ricerca, ad esempio l'Istituto Superiore di Sanità. Strutture stimate o calcolate diversamente comportano, come ovvio, valori diversi negli indicatori, in pratica però tali differenze sono minime. Il limite inferiore di età considerato è posto a 35 anni, questo perché si sono riscontrati alcuni errori nella determinazione della struttura della popolazione per età inferiori; indagini preliminari consentono comunque di affermare che qualora si considerino anche le classi di età inferiori ai 35 anni, le differenze sono trascurabili, anche perchè l'incidenza delle patolgie considerate è relativamente bassa per queste classi di età. La metodologia di mappatura, ossia la costruzione dei cartogrammi utilizzati per la descrizione della variazione dei rischi relativi, richiede che questi siano espressi come variabili categoriche. Ciò si ottiene dividendo la scala dei valori che essi assumono in intervalli disgiunti ed esaustivi, attraverso la specifica di valori di *cut-off*; ad ognuno degli intervalli specificati viene associato un dato colore sulla mappa, in particolare una diversa gradazione nella scala dei grigi, da bianco per rischi di mortalità bassi al nero per quelli elevati.

4.2 Risultati per il tumore allo stomaco maschile

A titolo esemplificativo, si analizzano i risultati ottenuti applicando i diversi metodi ai dati di mortalità per tumore allo stomaco nella popolazione maschile.

Per tale esempio sono stati considerati i dati relativi al periodo 1981–1988 a livello comunale, per la regione Toscana (e i comuni confinanti con essa per considerare un caso di informazione completa). I tassi specifici per sesso e classe di età, utilizzati per la determinazione dei casi attesi, sono relativi all'Italia di quel periodo.

La scelta del periodo analizzato è stata fatta per effettuare un confronto con risultati ottenuti da studi passati, sempre a livello comunale, per lo stesso periodo nella zona tra la Romagna, l'Aretino ed il Pesarese (si veda Cislaghi *et al.*, 1995a, 1995b). Mediante tali studi è stato individuato un "focolaio" per il rischio di tumore allo stomaco fra le provincie di Arezzo e Forlì; esso è rappresentato dalla circonferenza nera riportata sulla mappa originale degli SMR nella figura 4.2. Dall'analisi della sola area della regione Toscana, vedremo, tale andamento viene contraddetto, in quanto i rischi sono elevati anche nella parte superiore della provincia di Firenze, così da far pensare ad un focolaio di tumore centrato in un'area spostata a nord-ovest rispetto a quella reale.

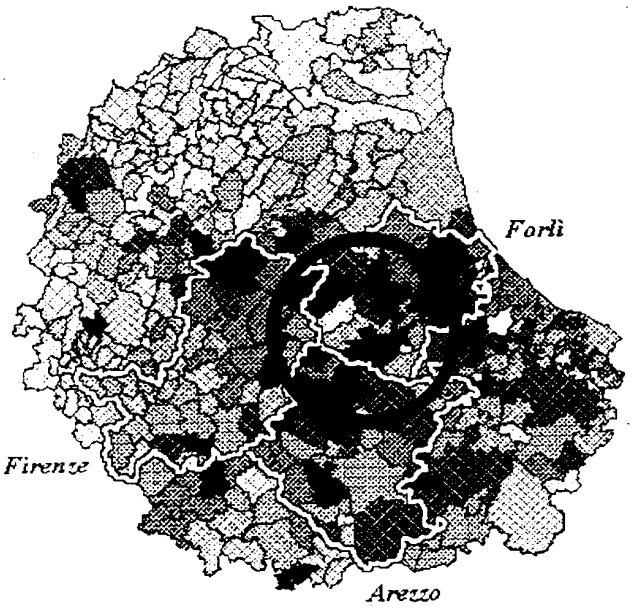


Figura 4.2: Focolaio del tumore allo stomaco maschile sulla mappa degli SMR, individuato in studi precedenti

Le mappe, che descrivono le stime dei rischi relativi ottenute utilizzando i vari metodi, sono elaborate usando una categorizzazione dei rischi relativi; i livelli di grigio utilizzati rappresentano diversi livelli del rischio relativo, le cui soglie sono scelte in base alla distribuzione dello stesso sulla regione Toscana. Le categorie sono state individuate sulla densità del rischio relativo nella regione Toscana, ossia analizzando la stima *kernel* di tale densità, ottenuta dai valori dei 287 rischi relativi stimati sui comuni toscani nel caso

di informazione completa; si veda la figura 4.3.

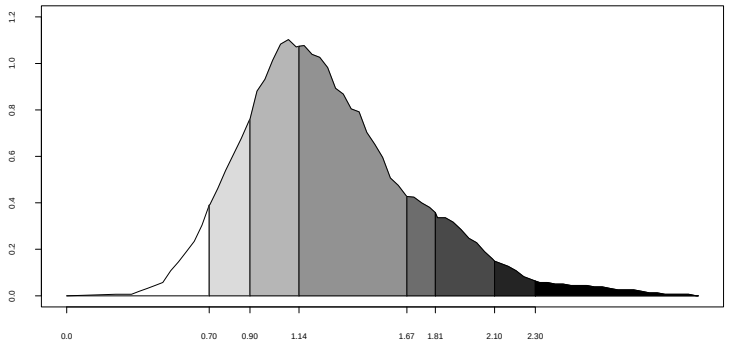
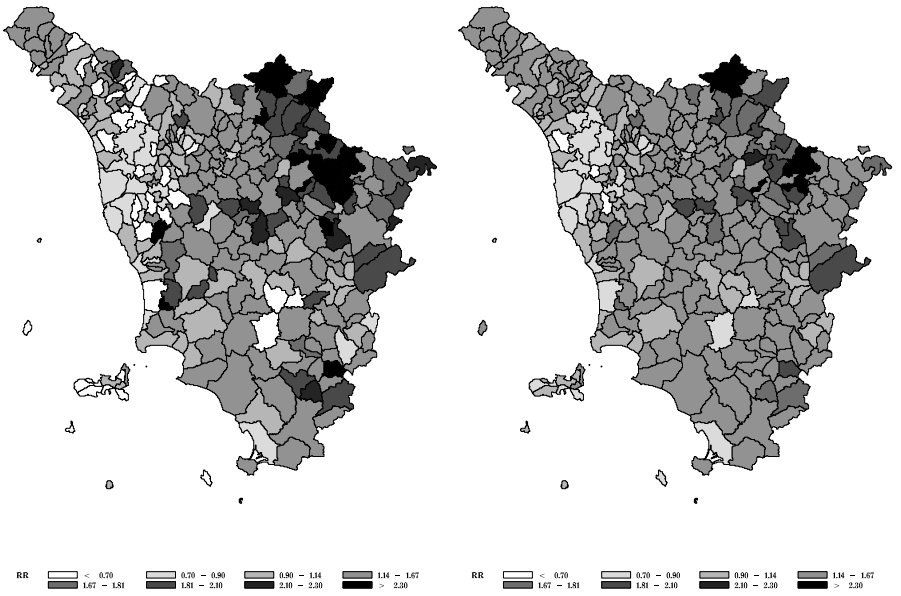


Figura 4.3: Stima *kernel* della distribuzione del rischio relativo

La lettura delle mappe può essere agevolata dalla consultazione della figura E.1 e della tabella E.1 nell'appendice E; esse permettono di associare ad ogni area, rispettivamente, il codice ISTAT corrispondente e il nome del comune relativo a tale codice. Si noti che i comuni dell Toscana sono i primi 287 della tabella e quelli il cui nome è scritto con un carattere diverso sono i comuni che si trovano al confine della Toscana.

Nella figura 4.4.a sono mostrate le mappe relative alle stime di massima verosimiglianza, ossia gli SMR (si veda il paragrafo 2.1); nella figura 4.4.b le stime bayesiane empiriche (EB) secondo l'approccio proposto da Clayton e Kaldor (1987) descritto nel paragrafo 2.2, usando come distribuzione per il rischio relativo di ciascun comune una distribuzione gamma con parametro di forma β e parametro inverso di scala α; i rischi relativi sono assunti indipendenti da comune a comune.

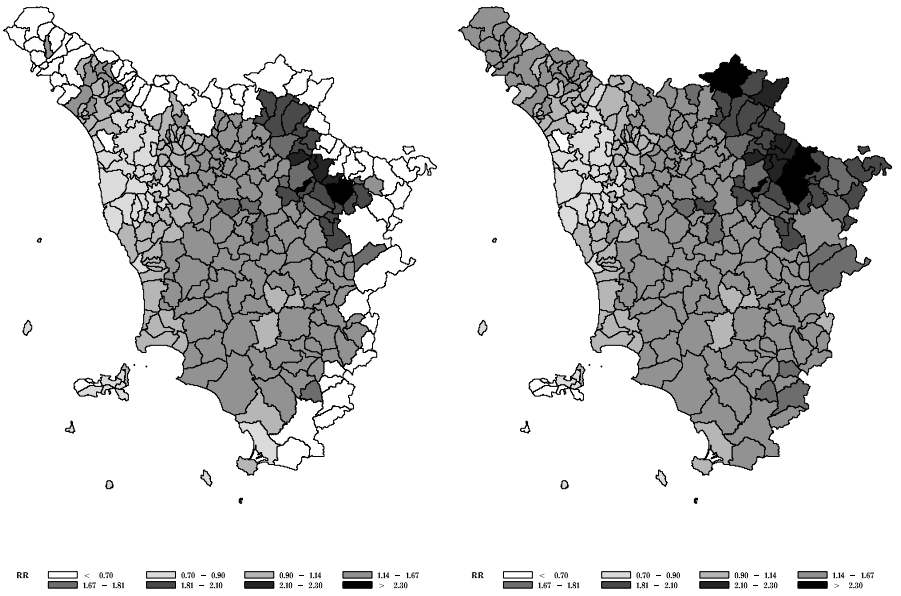
La mappa che mostra le stime del rischio relativo (RR) date dall'indice SMR, a causa della instabilità delle stime, fornisce un'immagine non chiara della variabilità geografica del rischio di mortalità (come visto nel paragrafo 2.1). Le stime ottenute per i parametri della distribuzione gamma nell'approccio bayesiano empirico sono $\hat{\beta} = 9.102$ e $\hat{\alpha} = 6.923$, quindi la media della distribuzione del rischio relativo è 1.315 e la varianza 0.189. Si noti che la mappa delle stime bayesiane empiriche risulta "smussata" rispetto a quella degli SMR; l'assunzione di distribuzioni gamma indipendenti sui rischi relativi permette però soltanto uno smussamento generale della mappa, non quello locale, data la natura non spaziale del modello.



(a) (b)

Figura 4.4: Stime SMR (a) e EB (b)

Per considerare una dipendenza locale tra i rischi relativi di comuni adiacenti ed ottenere delle stime che non risentano dell'*oversmoothing* cui sono affette quelle bayesiane empiriche, si considera il modello bayesiano gerarchico descritto nel paragrafo 2.3. Lo stesso modello viene applicato sia su una regione ristretta, cioè la regione Toscana senza considerare i comuni al confine (la corona interna), sia sulla regione nella sua completezza, ossia su dati che sono incompleti in quanto mancanti dell'informazione sui comuni confinanti con la regione stessa. Quando si applica il modello alla regione senza la corona interna, metodo indicato con SE, esso viene limitato considerando solo 231 comuni, sui quali viene definita una matrice di adiacenze ridotta. La mappa delle stime dei rischi relativi ottenute in questo modo è mostrata nella figura 4.5.a. Nella figura 4.5.b è invece rappresentata la mappa dei rischi relativi stimati mediante il modello bayesiano sulla base dell'insieme informativo costituito dai soli comuni della regione Toscana (metodo denominato IN).



(a) (b)

Figura 4.5: Stime SE (a) ed IN (b)

Dall'analisi della figura 4.5.b, sembra, in effetti, presente una distorsione nelle stime dei rischi relativi dovuta all'"effetto confine": infatti, la mappa stimata senza tenerne conto fornisce un'immagine distorta rispetto a ciò che conoscenze, derivanti da studi precedenti su un'area più estesa, fanno pensare. Se non si opera, infatti, una correzione per l'effetto confine, considerando i dati della sola Toscana, si è portati a ipotizzare l'esistenza di un "focolaio" di tumore centrato in un comune più a nord, di quello che, invece, viene supposto (si veda la figura 4.2).

Una possibile misura dell'effetto confine può essere data, non considerando la distorsione della stima dei rischi relativi ma la distorsione nella stima del coefficiente di autocorrelazione spaziale del modello autoregressivo condizionale definito sui rischi relativi; si noti però che tale informazione è ai nostri fini di poca utilità. Usando il modello autoregressivo condizionale (CAR) gaussiano di primo ordine analizzato nel paragrafo 1.2 (esteso a reticoli irregolari, si veda Besag 1975) per il logaritmo degli SMR per tumore allo stomaco, qualora non si consideri la completa informazione (come visto nel paragrafo 1.3), si ottiene una sottostima del parametro di autocorrelazione

ρ. Le stime dei parametri del modello CAR (σ^2 ed il coefficiente di autocorrelazione ρ) sono riportate nella tabella 4.1, rispettivamente per la Toscana senza considerare i comuni della corona interna, per la Toscana e per essa più i comuni della corona esterna (quindi operando una correzione per l'effetto confine).

Tabella 4.1: Stima dei coefficienti del modello CAR

Dati considerati	σ^2	ρ
Toscana senza corona interna	0.4463	0.1314
Toscana	0.3756	0.1418
Toscana più corona esterna	0.3351	0.1554

Le stime ottenute risentono dell'effetto confine, infatti in presenza di una minore informazione, il coefficiente di autocorrelazione spaziale è sottostimato. È, quindi, utile operare una correzione con i metodi proposti nel capitolo 3.

Per la valutazione dei metodi di correzione, i risultati ottenuti sono confrontati con quelli relativi all'applicazione del modello nella sua formulazione originale usando la completa informazione relativa alla mortalità per tumore allo stomaco in un'area più estesa che comprende, oltre alla regione Toscana, anche le aree (i comuni) confinanti della Liguria, Emilia Romagna, Marche, Umbria e Lazio (corona esterna), ossia 69 comuni, per un totale di 356 aree (riportate nella tabella E.1). L'introduzione di una corona esterna permette di confrontare le stime relative alle aree interne alla regione considerando una completa informazione per quelle al confine; le stime per le aree esterne saranno a loro volta affette dal problema dell'effetto confine e quindi non verranno considerate, anche perché non sono di nostro interesse.

Le stime dei rischi relativi per i comuni toscani, sono ottenute usando il modello bayesiano esteso alla zona comprendente la Toscana ed i comuni non toscani confinanti con essa. Per la definizione esatta delle distribuzioni condizionali sui rischi relativi si è fatto uso di una matrice di adiacenze estesa. La mappa dei rischi relativi ottenuta usando il modello bayesiano con dati completi, detto metodo CO, è mostrata nella figura 4.6, mantenendo ancora la stessa gradazione di grigio usata precedentemente. Si noti come in tal caso i risultati confortino quelli forniti dagli studi precedenti ad opera di Cislaghi *et al.* (1995a, 1995b).

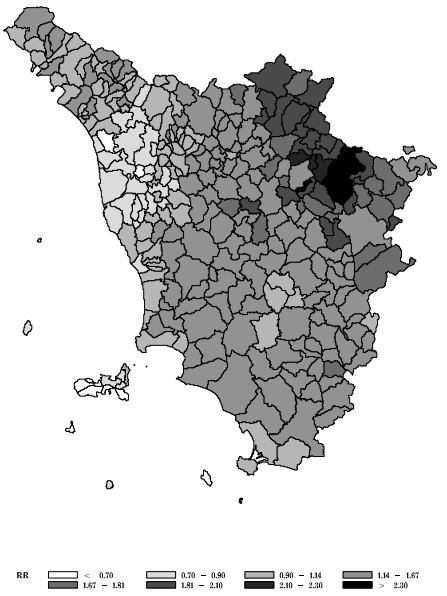
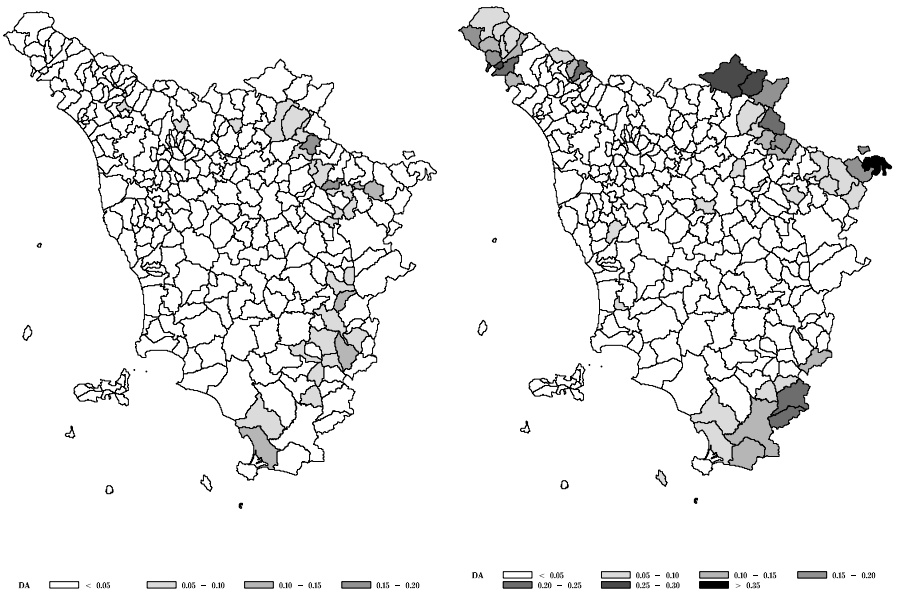


Figura 4.6: Stime CO

Vengono mostrate nella figura 4.7 le differenze in valore assoluto tra le mappe considerate; in particolare, nella figura 4.7.a la differenza in valore assoluto tra le stime ottenute considerando i comuni della Toscana (IN) e quelle ottenute non considerando i comuni della prima corona interna, ossia i comuni al confine (SE); si noti che in questo caso naturalmente la prima corona interna non è considerata nei risultati.

Nella figura 4.7.b è rappresentata la differenza assoluta tra le stime dei rischi ottenuti usando i dati incompleti della sola Toscana (IN) e quelli completi con la corona esterna dei comuni confinanti con la Toscana (CO).

Le differenze si riferiscono alle stime ottenute considerando l'informazione completa (quindi le stime ottenute applicando il metodo di correzione CO), poiché esse si possono considerare più prossime al reale valore del rischio relativo. Le mappe delle differenze assolute sono determinate considerando valori *cut-off* fissi per otto classi, in modo da poter ottenere un buon dettaglio.



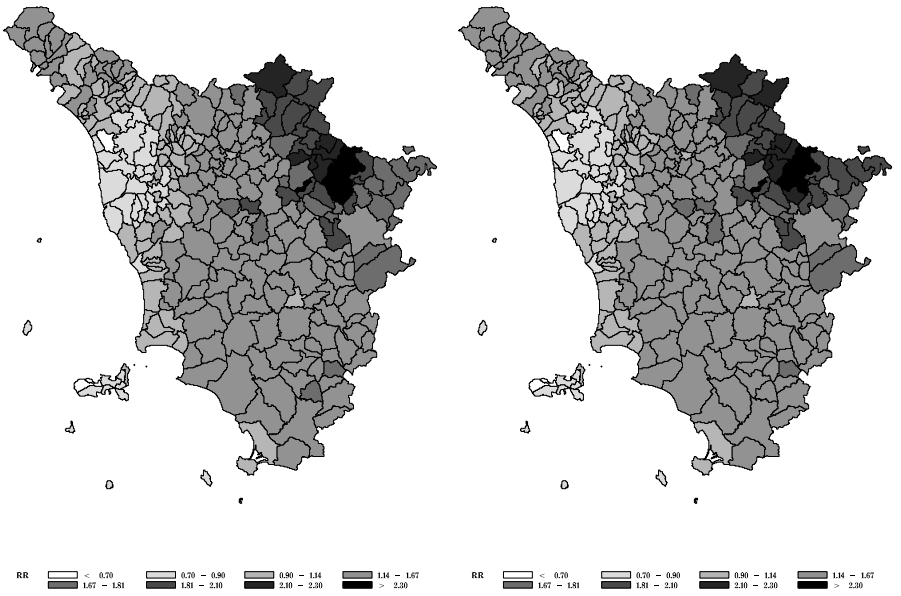
(a) (b)

Figura 4.7: Differenze assolute tra SE e IN (a) e tra IN e CO (b)

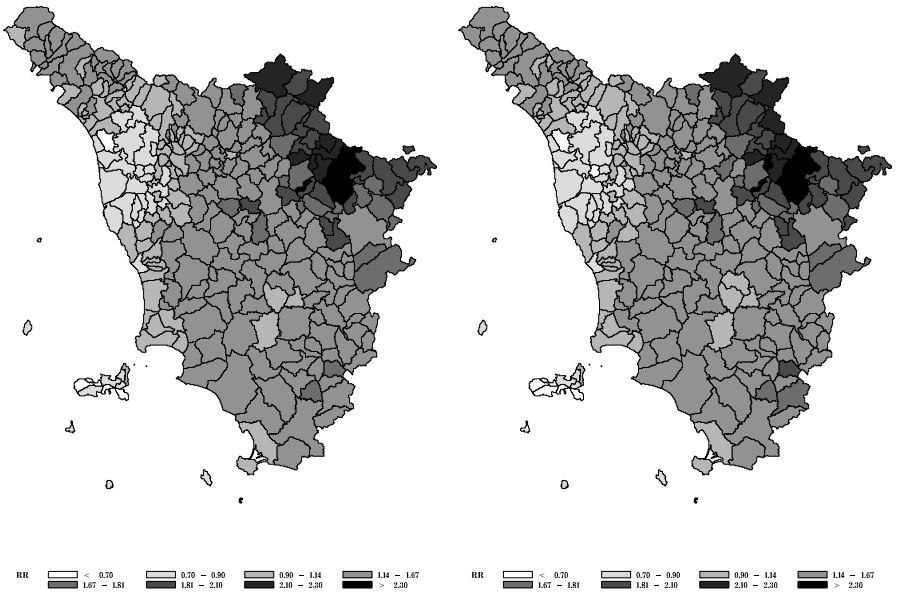
Dall'analisi delle mappe si nota l'assenza di linearità dell'effetto confine (questo avviene soltanto nello schema limite a *trend* e per reticoli regolari, come visto nella figura 3.4.a). Tale proprietà non consente, così, di effettuare una correzione per le aree della prima corona interna, basandosi sulla distorsione, sempre determinabile, che si ha sulla seconda corona interna eliminando, dal procedimento di stima, le aree della prima corona interna. Ma soprattutto si nota, confrontando i risultati che si ottengono applicando il modello bayesiano gerarchico ai dati relativi alla sola regione Toscana (figura 4.5.b) e a quelli completati mediante la considerazione della corona esterna (figura 4.6), che le maggiori differenze sono soprattutto relative ai comuni al confine, mentre piccole variazioni si hanno nelle aree interne. In modo particolare risultano distorte quelle nel "bordo" nord-est della Toscana. A causa dell'effetto confine i rischi relativi dei comuni di Firenzuola in modo particolare, ma anche di Palazzuolo sul Senio, Marradi, S. Godenzio e Sestino risultano sovrastimati. L'effetto confine sul bordo nord-ovest e sud-est, è di minore interesse in quanto di entità minore e riferito a comuni con un rischio di mortalità molto vicino al valore di riferimento.

Viene quindi adottato il modello bayesiano gerarchico con le modifiche per la correzione per l'effetto confine proposte nel capitolo 3. Applicando i metodi di correzione al modello bayesiano, nel caso del metodo che prevede la limitazione del peso delle aree al confine (PE) è stato usato quale peso, per ogni area, il rapporto tra la somma dei casi attesi nelle aree a lei confinanti note e quelli di tutte le aree confinanti (come dall'espressione 3.1). La mappa relativa alle stime ottenute pesando i dati è mostrata nella figura 4.8.a. La mappa relativa alle stime ottenute imputando i dati mancanti con il metodo MS, qundi tramite la media della distribuzione predittiva, è riportata nella figura 4.8.b; quando l'imputazione è fatta tramite campionamento dalla distribuzione predittiva (metodo DS) si ottiene una mappa come quella mostrata in figura 4.8.c. Nel caso si adotti il metodo CH, la mappa che si ottiene è mostrata nella figura 4.8.d. Le mappe corrette per l'effetto confine con i metodi proposti, evidenziano un andamento del rischio relativo più conforme a quello mostrato nella mappa ottenuta considerando una maggiore informazione (metodo di correzione CO).

La convergenza per le sequenze *Gibbs* ottenute per ogni parametro e per ogni metodo utilizzato, è stata verificata tramite il test di convergenza di Geweke (si veda paragrafo D.3). Dopo aver scartato 1000 campioni quale *burn-in*, sono stati considerati 3500 valori campionati; per il metodo CH il *burn-in* considerato è di 2000 valori. Per il metodo DS, sono state effettuate 20 iterazioni (sempre considerando per l'algoritmo *Gibbs sampler* 3500 valori dopo averne scartati, ogni volta, i primi 1000 quale *burn-in*). Nella figura 4.9 sono mostrati i *box-plot* relativi alle distribuzioni empiriche per le 20 iterazioni relative all'iperparametro di *heterogeneity* (figura 4.9.a), a quello di *clustering* (figura 4.9.b) e ai rischi relativi dei comuni di Firenzuola (figura 4.9.c) e di Castel del Rio (figura 4.9.d). La sequenza considerata per simulare la distribuzione a posteriori dei parametri del modello è quella relativa all'ultima iterazione; l'avvenuta convergenza dell'algoritmo è provata tramite l'analisi di componenti di varianza utilizzando il test di Gelman e Rubin (1992a, 1992b), come descritto nel paragrafo 3.2.1.

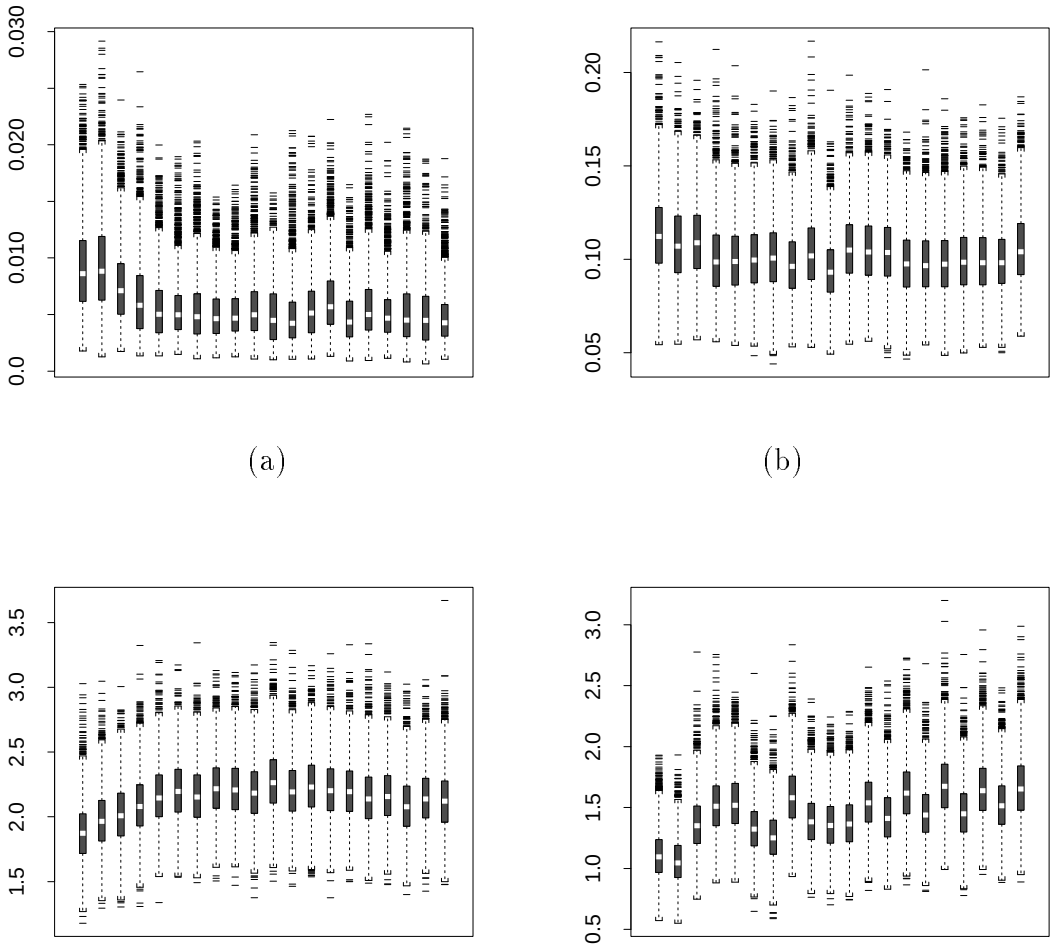


(a) (b)



(c) (d)

Figura 4.8: Stime PE (a), MS (b), DS (c) e CH (d)



(a) (b)

(c) (d)

Figura 4.9: *Box-plot* delle distribuzioni relative alle 20 iterazioni, dell'algoritmo EM doppiamente stocastico, per il parametro di *heterogeneity* (a) e di *clustering* (b), il rischio relativo di Firenzuola (c) e di Castel del Rio (d)

I valori assunti dall'indice R, considerando tutte le 20 iterazioni e poi scartando le sequenze relative alle prime iterazioni dall'analisi (dalla prima alle prime diciotto), sono mostrati in figura 4.10, relativamente agli iperparametri λ_u e λ_v e al rischio relativo di due comuni presi come esempio. Il primo, Firenzuola, si trova al confine della Toscana; il secondo, Castel del Rio, confinante con la Toscana ma esterno ad essa (in provincia di Bologna) è stato scelto in quanto rappresentante la situazione "limite" di area della corona est che confina con una sola area della regione di interesse. In tale situazione l'informazione sulla quale si basa l'imputazione del valore osservato, è minima. Dal grafico si evince che l'algoritmo può essere ritenuto a convergenza dall'undicesima iterazione.

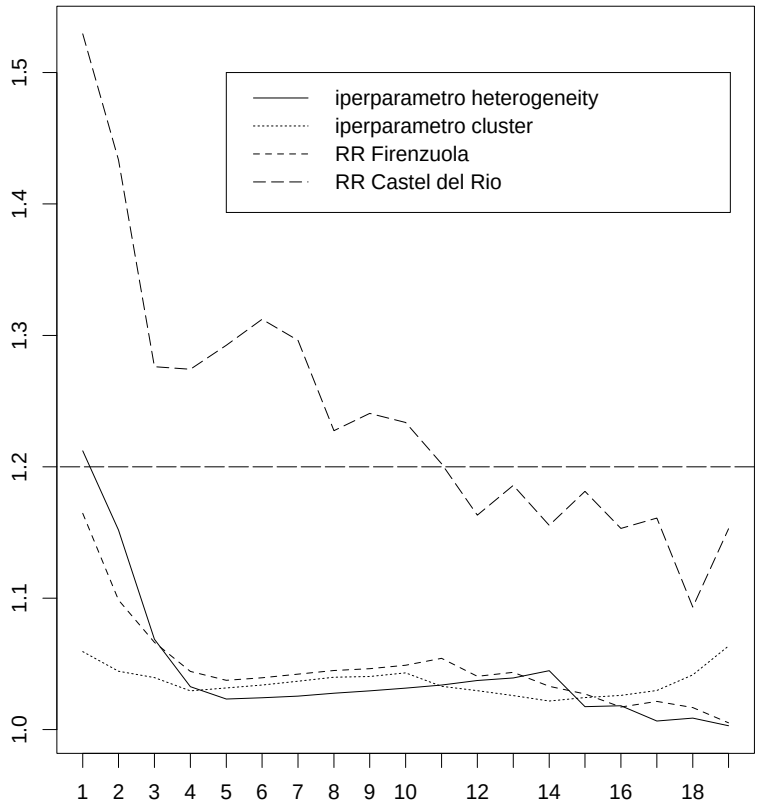
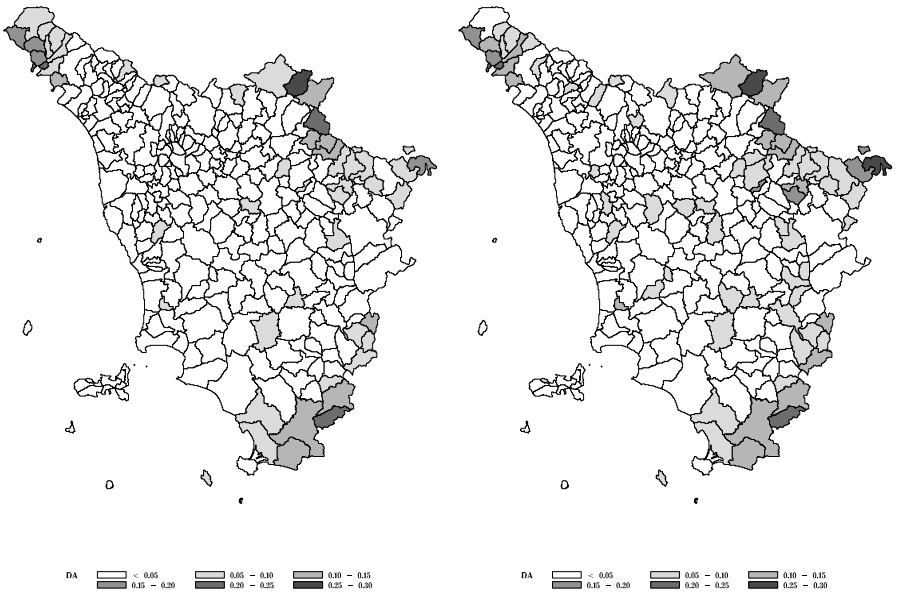


Figura 4.10: Valore R nell'analisi delle componenti di varianza per la convergenza dell'algoritmo EM doppiamente stocastico

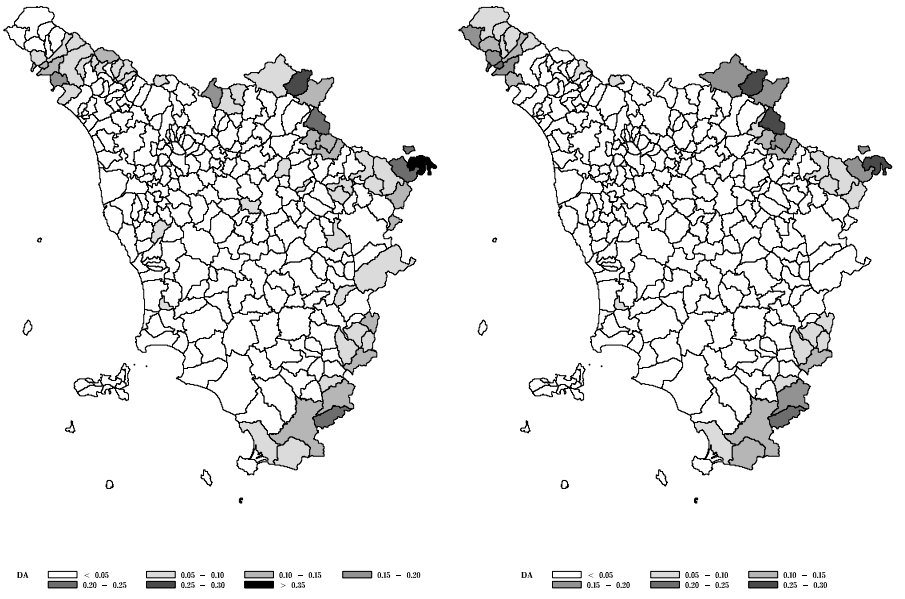
Le differenze in valore assoluto tra le mappe corrette con i metodi proposti e quella ottenuta con dati completi (metodo di correzione CO) sono mostrate nella figura 4.11. In particolare, le differenze per la mappa corretta pesando i dati (metodo PE) in figura 4.11.a, quelle per la mappa corretta tramite l'imputazione dei dati con il metodo MS in figura 4.11.b, con quello DS in figura 4.11.c e con il CH in figura 4.11.d.

Nell'analisi dei risultati ottenuti nell'esperimento di simulazione è stata considerata la differenza relativizzata al "vero" valore del rischio relativo di ogni area. In questo caso sono considerate le differenze assolute, sia perché non si dispone del valore vero dei rischi relativi, sia perché le differenze relative non sono di interesse: queste darebbero, infatti, troppa importanza a differenze piccole in aree a rischi bassi, quando l'interesse, per come si presenta il problema in questo esempio specifico, è soprattutto nelle aree a rischio elevato. Le mappe possono essere utilmente confrontate con quella mostrata in figura 4.7.b, che evidenzia le differenze assolute tra mappa con informazione incompleta (IN) e mappa con informazione completa (CO). Da queste si vede come il metodi di correzione applicati al modello forniscono delle stime per i rischi relativi che in valore assoluto si discostano in misura minore da quelle ottenute usando un informazione maggiore sul fenomeno, relativa alle aree esterne al confine. In modo analogo a quanto visto nell'esperimento di simulazione, l'algoritmo di imputazione dell'informazione mancante CH agisce soprattutto al confine della regione, dove maggiore è la distorsione che l'effetto confine induce. Il metodo PE e soprattutto quello MS influenzano invece anche le stime dei rischi relativi per i comuni interni alla Toscana. Dall'analisi di queste mappe si nota come il maggiore intervento del metodo di correzione interviene per il comune di Firenzuola, al confine nord della provincia di Firenze. Nella tabella 4.2 sono riportati i valori medi, lo scarto quadratico medio, i valori minimi e massimi ed i quantili al 2.5%, 50% e 97.5% delle distribuzioni campionate per il rischio relativo di questo comune usando i diversi metodi.

Nella figura 4.12 sono raffigurate le sequenze campionate dalla distribuzione a posteriori del rischio relativo del comune e la stima *kernel* della densità, prima nel caso di mancanza di informazioni circa i comuni confinanti delle province di Bologna e Ravenna, poi nel caso di completa informazione e nel caso in cui si corregga per l'effetto confine il metodo di stima bayesiano gerarchico.



(a) (b)



(c) (d)

Figura 4.11: Differenze assolute tra CO e PE (a), MS (b), DS (c) e CH (d)

In particolare la figura 4.12.a è relativa alla distribuzione ottenuta nel caso di informazione incompleta (metodo IN), la 4.12.b quando si ha completa informazione (metodo CO), la figura 4.12.c correggendo tramite l'introduzione di uno schema di pesi (metodo PE), la figura 4.12.d quando sono imputati i dati mancanti quale valore atteso della distribuzione predittiva (metodo MS), la figura 4.12.e quando l'imputazione è fatta campionando da essa (metodo DS) e infine la figura 4.12.f adottando l'algoritmo *chained data augmentation* (metodo CH). Si noti che la linea tratteggiata, riportata nelle figure, rappresenta il valore medio della distribuzione campionata, quindi la stima del rischio relativo per il comune di Firenzuola adottando i diversi metodi.

Tabella 4.2: Confronto tra le distribuzioni empiriche del rischio relativo nel comune di Firenzuola ottenute adottando il modello nelle varie versioni

	IN	CO	PE	MS	DS	CH
media	2.3059	2.0466	2.1467	2.1516	2.1286	2.2312
s.q.m.	0.2968	0.2521	0.3225	0.2261	0.2401	0.2750
min	1.4464	1.3144	1.1961	1.4580	1.4776	1.4538
max	3.4151	3.1749	3.6105	3.0260	3.6720	3.4545
2.5%	1.7700	1.5900	1.6000	1.7500	1.6900	1.7300
50%	2.2900	2.0300	2.1200	2.1400	2.1200	2.2200
97.5%	2.9200	2.5600	2.8400	2.6200	2.6500	2.8100

Dai valori riportati nella tabella 4.2, si nota che conformemente a quanto visto dai risultati sull'esperimento di simulazione, il metodo MS riduce, anche se in quantità minore, la varianza della distribuzione del rischio relativo per ciascuna area.

Nella tabella 4.3 sono riportate le stime degli iperparametri, ossia la varianza dei termini di *heterogeneity* e quella dei termini di *clustering* (a meno di un fattore moltiplicativo dato dall'inverso del numero di aree adiacenti). Le stime sono date dalla media della distribuzione campionata tramite *Gibbs sampler*; sono inoltri riportati gli scarti quadratici medi (s.q.m.) delle distribuzioni campionate.

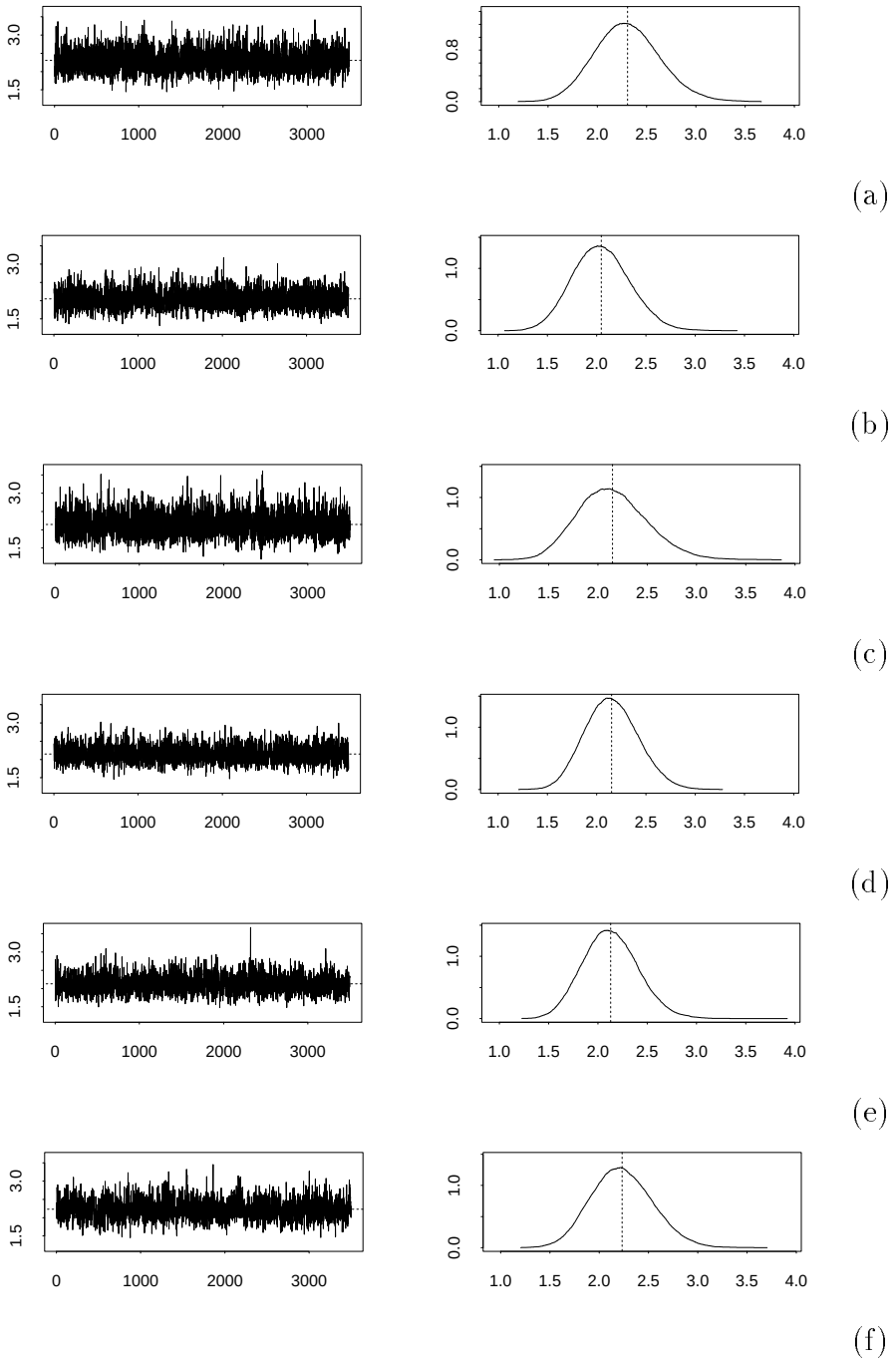


Figura 4.12: Rischio relativo nel comune di Firenzuola, metodo IN (a), CO (b), PE (c), MS (d), DS (e) e CH (f)

Tabella 4.3: Confronto tra le stime degli iperparametri per i diversi metodi

parametro	SE	IN	CO	PE	MS	DS	CH
$\hat{\lambda}_v$	0.0058	0.0060	0.0086	0.0055	0.0048	0.0047	0.0065
s.q.m.($\hat{\lambda}_v$)	0.0031	0.0028	0.0039	0.0026	0.0025	0.0023	0.0034
$\hat{\lambda}_u$	0.0912	0.1034	0.1302	0.1005	0.0813	0.1063	0.1116
s.q.m.($\hat{\lambda}_u$)	0.0217	0.0213	0.0275	0.0215	0.0154	0.0202	0.0232

Nella tabella 4.4 sono riportati i valori ottenuti adottando i diversi stimatori per le aree lungo il bordo nord-est della Toscana (ordinate da nord verso sud). Di particolare interesse sono tre zone: quella tosco-romagnola (R), la Valle del Casentino (C) e la Valle del fiume Tevere (T).

Tabella 4.4: Confronto tra i diversi metodi per le aree lungo il bordo nord-est della Toscana

Comune	SMR	EB	IN	CO	PE	MS	DS	CH
Firenzuola (R)	2.739	2.303	2.306	2.047	2.147	2.152	2.129	2.231
Palazzuolo (R)	1.714	1.449	1.999	1.725	1.991	1.976	1.989	2.004
Marradi (R)	2.414	2.003	2.158	1.978	2.090	2.111	2.121	2.160
S. Godenzo (R)	2.000	1.545	2.094	1.847	2.063	2.097	2.069	2.130
Stia (C)	2.500	1.978	2.278	2.100	2.213	2.241	2.248	2.285
Pratovecchio (C)	2.000	1.671	2.155	2.020	2.147	2.140	2.162	2.164
Poppi (C)	3.082	2.514	2.664	2.634	2.563	2.552	2.617	2.639
Bibbiena (C)	3.086	2.694	2.833	2.796	2.721	2.719	2.787	2.809
Chiusi Verna (C)	1.613	1.456	2.008	1.928	2.014	2.003	2.017	2.004
Pieve S.S. (T)	1.720	1.547	1.807	1.745	1.779	1.815	1.825	1.821
Badia Ted. (T)	1.702	1.471	1.846	1.656	1.755	1.839	1.900	1.847
Sestino (T)	2.157	1.672	2.023	1.631	1.792	1.885	2.036	1.928

Le proprietà del metodo di stima sono evidenti nello *smoothing* spaziale che è presente nelle aree ai confini della Valle del Casentino: il comune di San Godenzo ha per la stima del rischio relativo dei valori vicini a quelli della Valle del Casentino, mentre al confine Sud per il comune di Chiusi della Verna il rischio relativo è stato sottostimato data la prossimità del comune alle aree a basso rischio della Valle del fiume Tevere.

Le stime ottenute usando il modello di Besag, York e Mollié sui dati completi evidenziano come aree a più elevato rischio, quelle della Valle del Casentino. Usando soltanto i dati relativi alla Toscana, invece, si ottengo-

no valori elevati anche per le aree della Valle del fiume Tevere e della zona tosco-romagnola, quindi, contrariamente a quanto studi precedenti hanno evidenziato. I risultati del modello di Besag, York e Mollié applicato ai soli dati della regione Toscana, ma corretto per l'effetto confine, invece, confermano il reale andamento del rischio sulla regione, ossia un rischio elevato soprattutto nella Valle del Casentino. Nella tabella 4.5 le differenze, in valore assoluto, tra le diverse stime del rischio relativo per i comuni lungo il bordo nord-est della Toscana, adottando il metodo bayesiano nella versione originale e nelle due versioni corrette per l'effetto confine rispetto alla stima ottenuta applicando il modello con la completa informazione per le aree adiacenti. Tali differenze sono quelle riportate nella mappa della figura 4.7.b e in quelle della figura 4.11.

Tabella 4.5: Differenze tra le diverse stime per le aree lungo il bordo nord-est della Toscana rispetto alle stime ottenute con informazione completa CO

Comune	IN	PE	MS	DS	CH
Firenzuola (R)	0.259	0.100	0.105	0.082	0.185
Palazzuolo (R)	0.275	0.267	0.252	0.264	0.279
Marradi (R)	0.180	0.112	0.132	0.142	0.181
S. Godenzo (R)	0.246	0.216	0.250	0.222	0.282
Stia (C)	0.178	0.113	0.141	0.148	0.185
Pratovecchio (C)	0.135	0.127	0.120	0.142	0.144
Poppi (C)	0.030	0.071	0.082	0.017	0.005
Bibbiena (C)	0.037	0.076	0.078	0.010	0.013
Chiusi Verna (C)	0.080	0.086	0.075	0.089	0.076
P. S. Stefano (T)	0.062	0.034	0.070	0.080	0.075
Badia Tedalda (T)	0.190	0.099	0.183	0.244	0.191
Sestino (T)	0.392	0.161	0.254	0.405	0.297

Per l'esempio reale del tumore allo stomaco in Toscana, i metodi di correzione per l'effetto confine sono quindi validi: le procedure di correzione per l'effetto confine riflettono meglio la struttura sottostante del rischio relativo messa in luce dall'uso di informazione completa (la corona esterna), reperita dai comuni confinanti con la Toscana.

Considerazioni conclusive

I metodi proposti nel capitolo 3, quando applicati al caso reale della mappatura del rischio relativo di morte per tumore allo stomaco, permettono la

correzione per l'effetto confine; infatti i risultati mettono in luce il reale andamento del rischio di tumore allo stomaco sulla regione Toscana, ricostruito tramite il reperimento dell'informazione altrimenti mancante.

Il metodo che prevede l'introduzione di un sistema di pesi nel modello fornisce risultati migliori rispetto a quello che considera l'imputazione dell'informazione mancante. Questo è dovuto al fatto che tale metodo è più robusto del secondo, in cui invece si ha una forte influenza dello schema spaziale seguito dai rischi relativi e della struttura delle adiacenze e delle popolazione. Inoltre, entrambi i metodi usano quale informazione i casi attesi nelle aree adiacenti ed esterne alla Toscana: in questo caso comuni prevalentemente montani, quindi scarsamente popolati. Poichè i due metodi sfruttano tale informazione in modo diverso, in questo caso particolare, il metodo di pesatura dei dati fornisce risultati migliori.

Conclusioni

Nel lavoro sono stati analizzati i problemi che l'effetto confine può provocare nelle analisi di dati spaziali; in particolare, come si presenta il problema nella mappatura del rischio di mortalità in una data regione geografica, quando per la stima dei rischi relativi viene adottato il modello bayesiano gerarchico.

Sia nell'esperimento di simulazione, che nell'esempio del tumore allo stomaco maschile, l'applicazione del modello bayesiano gerarchico, a causa dell'effetto confine, conduce a risultati fuorvianti per quello che concerne l'esatta individuazione delle aree a maggiore rischio; l'effetto confine introduce, infatti, una distorsione nelle stime dei rischi relativi. I risultati, quindi, confermano la necessità di ricorrere a metodi che permettano la correzione per tale effetto. Si noti che l'impossibilità di dare una misura della distorsione introdotta ha portato alla ricerca di metodi che correggano direttamente il modello, senza la possibilità di apportare correzioni sulle stime ottenute.

Sarebbe fondamentale valutare in quale parte del reticolo ed in quale misura l'effetto confine sia presente; questo potrebbe essere fatto procedendo a specifici esperimenti di simulazione, oppure operando una disaggregazione dei dati per suddividere diversamente la regione in aree, in modo da riportarsi all'analisi di reticoli regolari, dove tentativi in questo senso sembrano destinati ad avere maggiore successo.

Per la correzione dell'effetto confine per il modello bayesiano gerarchico di Besag, York e Mollié, adottato per la stima dei rischi relativi di morte, sono stati proposti due approcci. Il primo si basa sull'applicazione di un sistema di pesi, l'altro rientra nelle tecniche per la trattazione dei dati mancanti; in particolare, sono stati proposti tre algoritmi originali, estendendo gli usuali algoritmi di imputazione di dati.

Il primo algoritmo, chiamato EM doppiamente stocastico, prevede un passo E eseguito tramite campionamento dalla distribuzione predittiva e un passo M eseguito mediante campionamento dalle distribuzioni condizionali complete (*Gibbs sampler*). Si noti che per tale algoritmo è stato necessario

considerare anche un metodo originale per verificarne la convergenza; questo prende spunto dal test di Gelman e Rubin (1992a, 1992b) usato per la convergenza di algoritmi di campionamento, quali ad esempio di *Gibbs sampler*.

Il secondo algoritmo, chiamato EM stocastico nel passo M, rappresenta una versione semplificata del precedente. Il passo E non è stocastico, cioè il valore imputato non deriva da una procedura di campionamento, bensì da una misura di locazione della distribuzione predittiva.

Il terzo algoritmo, chiamato *chained data augmentation* a tre vettori, prevede l'imputazione dell'informazione all'interno dell'algoritmo *Gibbs sampler*; in questo caso i dati mancanti sono trattati come parametri del modello bayesiano.

Il due metodi suggeriti, ed in particolare i diversi algoritmi di imputazione proposti, sono stati valutati mediante un esperimento di simulazione e la loro applicazione a dati reali. In base ai risultati ottenuti in questo lavoro, i metodi di correzione proposti danno dei buoni risultati; essi forniscono delle mappe che rivelano andamenti del rischio relativo conformi a quelli che si evincono dall'uso di un'informazione completa.

Nel caso reale analizzato nel capitolo 4, relativo alla mappatura del rischio di mortalità per tumore allo stomaco in Toscana, i due metodi forniscono risultati simili, anche se quelli ottenuti introducendo un sistema di pesi nel modello sembrano migliori. L'esperimento di simulazione porta a preferire invece i metodi di imputazione dell'informazione mancante, in special modo l'algoritmo *chained data augmentation* a tre vettori.

Dai risultati delle simulazioni del capitolo 3, si evince che la bontà dei metodi dipende dallo schema spaziale seguito dal rischio relativo, dalla struttura di adiacenze fra le aree e dalla struttura della loro popolazione. Per tale motivo, i risultati dell'esperimento di simulazione operato su un reticolo regolare e con una popolazione fissa trovano riscontro solo in parte nelle analisi fatte sulla popolazione della regione Toscana.

Da quanto ottenuto nel presente lavoro si deduce che non esiste un metodo unico valido per la correzione dell'effetto confine; vanno, quindi, valutate le caratteristiche del problema prima di adottare uno specifico schema di correzione. In tal senso si può pensare di considerare ulteriori esperimenti di simulazione, o su altri tipi di schemi spaziali, o su strutture di adiacenza più complicate o per diverse strutture di popolazione.

Tra gli altri possibili sviluppi di ricerca, si segnala la costruzione di un algoritmo di imputazione dell'informazione mancante quando per le aree esterne

ai confini non siano noti i casi attesi. In questo caso si potrebbe pensare di imputare i rischi relativi delle aree di una corona esterna e non i casi osservati; infatti, in questo caso, è impossibile specificare la distribuzione predittiva per operare l'imputazione dei decessi nelle aree esterne alla regione esaminata.

Appendice A

Correzione per l'effetto confine nei metodi che usano distanze tra eventi

A.1 Processi spaziali di punto e metodi che usano distanze tra e da eventi

Un processo spaziale di punto è un modello stocastico che determina la collocazione di un insieme di eventi (*spatial point pattern*) $\{\boldsymbol{s}_i\}$, con $i = 1, \ldots, n$, in una certa regione $A \subset I\!R^2$; in questo senso i dati individuali possono essere considerati come una realizzazione di un processo di questo tipo (si veda Cox e Isham 1980, Diggle 1983, Cressie 1991). Si denoti con eventi la realizzazione del processo $\boldsymbol{s}$ e con punti, i punti di $I\!R^2$. Generalmente si considera una misura di conteggio ϕ su A; $\phi(R)$ indica il numero di eventi all'interno della regione R, con $R \subset A$. Si denoti con $\mid R \mid$ la superficie di R e con $d\boldsymbol{s}$ una regione infinitesima contenente il punto $\boldsymbol{s}$. Le proprietà di primo ordine del processo possono essere descritte tramite la funzione di intensità di primo ordine (si veda per esempio Diggle, 1983)

$$\lambda(\boldsymbol{s}) = \lim_{|ds| \to 0} \frac{\mathrm{E}[\phi(d\boldsymbol{s})]}{\mid d\boldsymbol{s} \mid}.$$

Un processo di punto è stazionario ed isotropico se la sua struttura probabilistica in una qualunque regione del piano è invariante per traslazioni e rotazioni della regione R; quindi non esistono effetti direzionali. Si noti che per un processo stazionario la funzione di intensità di primo ordine $\lambda(\boldsymbol{s})$ assume un valore costante λ; questo rappresenta il numero medio di eventi in una area di superficie unitaria. La funzione di intensità di secondo ordine è

definita come

$$\lambda_2(\boldsymbol{s}_1, \boldsymbol{s}_2) = \lim_{|ds_1|,|ds_2|\to 0} \frac{\mathrm{E}\left[\phi(d\boldsymbol{s}_1)\phi(d\boldsymbol{s}_2)\right]}{\mid d\boldsymbol{s}_1 \mid\mid d\boldsymbol{s}_2 \mid}.$$

Per un processo stazionario, $\lambda_2(\boldsymbol{s}_1, \boldsymbol{s}_2) \equiv \lambda_2(\boldsymbol{s}_1 - \boldsymbol{s}_2)$; per un processo stazionario e isotropico la funzione di intensità di secondo ordine è data da $\lambda_2(r)$, dove r è la distanza tra $\boldsymbol{s}_1$ ed $\boldsymbol{s}_2$. Una caratterizzazione alternativa delle proprietà di secondo ordine di un processo stazionario e isotropico è data dalla funzione $K(r)$ definita come

$$K(r) = \lambda^{-1}\,\mathrm{E}\left[\text{numero eventi entro una distanza } r \text{ da un arbitrario evento}\right].$$

Per stabilire un legame tra la funzione $K(r)$ e $\lambda_2(r)$, si assume che il processo sia ordinato, in altre parole si assume che non si possano verificare eventi coincidenti, cioè $P\left[\phi(d\boldsymbol{s}) > 1\right]$ è di un ordine di infinitesimo superiore a $\mid d\boldsymbol{s} \mid$. Questo significa che

$$\mathrm{E}[\phi(d\boldsymbol{s})] \simeq P\left[\phi(d\boldsymbol{s}) = 1\right] \qquad \text{cioè} \qquad \lim_{|ds|\to 0} \frac{\mathrm{E}[\phi(d\boldsymbol{s})]}{P\left[\phi(d\boldsymbol{s}) = 1\right]} \to 1,$$

si veda Diggle, 1983. In modo analogo si assume anche che

$$\mathrm{E}\left[\phi(d\boldsymbol{s}_1)\phi(d\boldsymbol{s}_2)\right] \simeq P\left[\phi(d\boldsymbol{s}_1) = \phi(d\boldsymbol{s}_2) = 1\right].$$

Sotto tali assunzioni, l'intensità condizionale che ci sia un evento in $\boldsymbol{s}$ data la presenza di un evento in 0, è $\lambda_2(0, \boldsymbol{s})/\lambda$. Segue che il numero atteso di ulteriori eventi entro una distanza r da un arbitrario evento è

$$\lambda K(r) = \int_0^{2\pi} \int_0^r \frac{\lambda_2(t)t}{\lambda} dt d\theta = \frac{2\pi}{\lambda} \int_0^r \lambda_2(t)\, t\, dt,$$

oppure

$$\lambda_2(r) = \frac{\lambda^2}{2\pi r} K'(r).$$

Da un punto di vista teorico è più conveniente lavorare con $\lambda_2(r)$ piuttosto che con $K(r)$, ma nella pratica si usa $K(r)$ dato che, come vedremo, tale funzione può essere stimata dai dati. Si noti che la funzione $K(r)$ può essere considerata, a meno di un fattore λ, come una distanza tra eventi: essa rappresenta il numero atteso di eventi entro una distanza r da un evento arbitrario.

Le proprietà di secondo ordine, malgrado rappresentino un buon punto di partenza per la descrizione di un processo spaziale di punto, non ne danno

una completa descrizione: processi diversi possono avere, infatti, proprietà di secondo ordine identiche. Per questo si definiscono due funzioni: la funzione $G(r)$ con r la distanza *nearest-neighbour* e la funzione $F(r)$ con r la distanza *point-nearest-event*.

La distanza r_i da ognuno degli eventi $\boldsymbol{s}_i$ di una regione A all'evento della stessa regione a lui più vicino, è detta distanza *nearest-neighbour*. Si definisce $G(r)$ come la probabilità che la distanza tra un evento scelto casualmente ed il suo vicino più prossimo sia minore o uguale ad r, quindi $G(r)$ rappresenta la funzione di ripartizione delle distanze *nearest-neighbour*.

Si definisce con $F(r)$ la funzione di ripartizione delle distanze *point-to-nearest-event* r_j, $j = 1, \ldots, z$, calcolate da ognuno degli z punti campionati su A all'evento ad esso più vicino tra gli n in A; in altre parole la probabilità che la distanza fra un punto scelto casualmente e l'evento ad esso più vicino, sia minore o uguale ad r.

Molte tecniche adottate per l'analisi di dati individuali, quali test di completa casualità spaziale (CSR), l'individuazione di *cluster* geografici o la stima dei parametri del processo, fanno uso delle funzioni K, G e F. Queste possono essere utilmente stimate attraverso le distanze tra e da eventi osservati sul piano.

Lo stimatore per la funzione K si ottiene sostituendo l'operatore di valore atteso con la media empirica. Si definisca con $I_r(r_{ij})$ l'indicatore di distanze r_{ij} minori di r tra due eventi $\boldsymbol{s}_i$ e $\boldsymbol{s}_j$; una stima di $K(r)$, per $r > 0$ e per λ noto (o stimato), è data da

$$\hat{K}(r) = \lambda^{-1} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, \neq i}^{n} I_r(r_{ij}). \tag{A.1}$$

Si consideri ancora l'indice $I_r(r_i)$, questa volta come funzione indicatrice di distanze *nearest-neighbour* minori o uguali ad r. La funzione $G(r)$ può essere stimata attraverso la corrispondente funzione di ripartizione empirica determinata tramite le osservazioni sul piano; per $r > 0$ si ottiene

$$\hat{G}(r) = \frac{1}{n} \sum_{i=1}^{n} I_r(r_i). \tag{A.2}$$

La funzione F può essere stimata attraverso la sua corrispondente funzione di ripartizione empirica, data da

$$\hat{F}(r) = \frac{1}{z} \sum_{j=1}^{z} I_r(r_j), \tag{A.3}$$

con $I_r(r_j)$ la funzione indicatrice di distanze *point-to-nearest-event* minori di r.

A.2 L'effetto confine e la sua correzione nei metodi che usano distanze tra e da eventi

Quando si determina una distanza tra o da eventi, questa non è valutata correttamente per quelli prossimi al confine, per i quali opera un meccanismo di censura; di conseguenza si ha una distorsione nell'inferenza che si costruisce sopra tali distanze. Ad esempio, la stima della funzione $G(r)$ tramite l'equazione (A.2) sarà distorta, questo perché si ottengono distanze maggiori ai bordi, escludendo la possibilità di eventi più vicini che non possono essere osservati.

Quando queste distanze sono usate, ad esempio, per la stima di un modello o per la costruzione di un test, si possono ottenere risultati distorti. E' quindi necessario considerare e correggere tale distorsione.

Un metodo è quello di considerare spazi infiniti abolendo i confini, ad esempio pensando la regione a forma toroidale, così gli eventi prossimi a bordi opposti, sono considerati vicini. Tale metodo è tuttavia criticabile, poiché irrealistico e valido solo per regioni rettangolari.

Un'altra possibile soluzione al problema può essere quella di considerare le distanze solo per una parte di punti ed eventi della regione, quelli sufficientemente lontani dal confine di A, così l'effetto confine non incide su tali distanze. Seguendo tale metodo si costruisce una corona interna alla regione A e si misurano le distanze solo per punti (o eventi) interni alla regione e non per quelli nella corona, che sono però considerabili come vicini; in tal modo si compensa direttamente il metodo per l'effetto confine, però vengono perse spesso un gran numero di osservazioni.

Un altro metodo è quello di estendere la regione di studio reperendo ulteriori dati, costruendo una corona esterna; gli eventi che si trovano in tale corona sono usati solo per la determinazione delle distanze.

Si può anche adottare un sistema di pesi per le distanze, questo perché esse possono essere distorte quando determinate per eventi vicini al confine della regione.

Un altro metodo è ottenere un campione finito di correzioni per la distribuzione teorica di una specifica statistica, cioè calcolare l'effetto confine (in genere tramite simulazioni) e riproporzionare la statistica con tale quantità. Questo è stato fatto, ad esempio, per la statistica di Clark-Evans (1954)

usata nei test di completa casualità spaziale, da Donnely (1978) relativamente alle distanze *nearest-neighbour* e Doguwa e Upton (1988) per le distanze *point-to-nearest-event*. Tale metodo però è empirico e fortemente legato alla forma della regione analizzata.

A.2.1 La mappa di forma toroidale

Ripley (1979, 1981, 1984, 1988) considera la mappa a forma toroidale: viene operata una sorta di espansione della finestra entro cui viene osservato il processo secondo l'interpretazione della figura A.1, cioè si considera la mappa come al centro di una griglia 3×3 di replicazioni della mappa stessa e si procede misurando le distanze dai punti del rettangolo centrale a punti degli altri rettangoli circostanti; questo per la correzione delle funzioni sulle distanze analizzate nel paragrafo A.1.

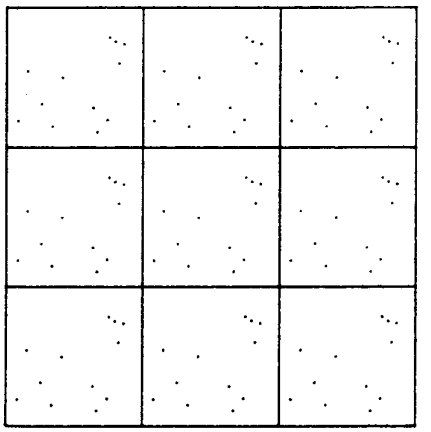


Figura A.1: Un'interpretazione della mappa di forma toroidale

Questo è ragionevole dato che la regione rettangolare osservata rappresenta un campione casuale di tutti i rettangoli che possono essere osservati (per l'assunzione di stazionarietà ed isotropicità); ciò che si può pensare degli altri rettangoli è che su essi il processo abbia identica intensità di primo e secondo ordine.

A.2.2 Eliminazione di una corona interna

Un altro metodo (si veda Ripley 1981, 1988) è quello di considerare solo gli eventi distanti dal confine per una distanza maggiore di r. In questo caso lo stimatore corretto della funzione K (A.1) sarà dato da

$$\hat{K}'(r) = \lambda^{-1} \sum_{i=1}^{n} \sum_{j=1,\neq i}^{n} \{I_r(r_{ij}) [1 - I_r(d_i)]\} / \sum_{i=1}^{n} [1 - I_r(d_i)] ,$$

per $r > 0$ e d_i che indica la distanza di un evento dal bordo.

Nel caso della funzione di ripartizione delle distanze *nearest-neighbour* (lo stimatore distorto per l'effetto confine è dato dall'espressione A.2), questo può essere fatto considerando invece che tutti gli n eventi, soltanto quelli ad una distanza maggiore di r dal confine (si veda Ripley, 1976); cioè, denotando con r_i le distanze di ogni evento $i = 1, \ldots, n$ dall'evento suo vicino più prossimo in A, e con d_i la distanza di ogni evento dal punto più vicino del confine di A, la stima di $G(r)$ è data dalla proporzione delle distanze $r_i \leq r$ determinata considerando solo gli eventi che distano almeno una distanza r dal confine di A, cioè con $d_i > r$. Definendo una funzione $I_r(d_i)$ in modo analogo a quanto fatto precedentemente, si ha

$$\hat{G}'(r) = \sum_{i=1}^{n} \{I_r(r_i) [1 - I_r(d_i)]\} / \sum_{i=i}^{n} [1 - I_r(d_i)] .$$

Tale stimatore non considera però le distanze *nearest-neighbour* fra gli eventi nella corona interna (che possono essere minori di r). Gli eventi di tale corona sono solo considerati come eventi vicini più prossimi di quelli interni alla regione, ma non sono inserite nel calcolo le loro distanze *nearest-neighbour*; inoltre la funzione $\hat{G}(r)$ non è necessariamente crescente per r. Hanish (1984) propone uno stimatore alternativo, che considera le distanze *nearest neighbour* di un insieme più grande di eventi, quelli che distano dal bordo di una distanza maggiore della propria distanza *nearest-neighbour*. Cioè

$$\hat{G}''(r) = \sum_{i=1}^{n} \{I_r(r_i) [1 - I_{r_i}(d_i)]\} / \sum_{i=i}^{n} [1 - I_{r_i}(d_i)] ,$$

si noti che questa funzione è non decrescente per r.

Per stimare correttamente $F(r)$ (lo stimatore distorto per l'effetto confine è dato dall'espressione A.3), si denoti con r_j la distanza da ognuno degli z punti campionari $j = 1, \ldots, z$ dall'evento più vicino in A e con d_j la distanza

di ogni punto dal punto più vicino del confine di A. Lo stimatore corretto per l'effetto confine della funzione $F(r)$ è dato dalla proporzione delle distanze $r_j \leq r$, fra gli z punti che distano almeno una distanza r dal confine di A, cioè con $d_j > r$, e l'evento a loro più vicino, cioè

$$\hat{F}'(r) = \sum_{j=1}^{z} \{I_r(r_j)\,[1 - I_r(d_j)]\} \,/\, \sum_{j=1}^{z} [1 - I_r(d_j)]\,.$$

A.2.3 Estensione dell'area di studio

Una correzione per l'effetto confine della stima di K consiste nell'estendere l'area con una corona esterna e quindi considerare, per ognuno degli n eventi in A, le distanze da un numero $n+m$ di eventi che si trovano in parte (n) in A ed il resto (m) esterni ad A, si veda Ripley (1981). Per lo stimatore della funzione K si ottiene

$$\hat{K}''(r) = \lambda^{-1} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1,\neq i}^{n+m} I_r(r_{ij}),$$

con un ampiezza della corona esterna maggiore di r. Tale metodo può essere applicato anche per la correzione della distribuzione empirica delle distanze *nearest-neighbour*, cioè per la stima della funzione G; essa sarà data da una formula analoga alla precedente (A.2), ma in questo caso le distanze saranno calcolate non più relativamente agli eventi più vicini tra gli n della regione A, ma tra quelli in un area espansa (rilevando altre osservazioni su un area intorno ad A).

A.2.4 Schema di pesi

Un'altra alternativa, per la correzione della stima della funzione K (si veda Ripley 1981, 1984, 1988) è quella di considerare uno schema di pesi per la determinazione delle distanze fra eventi. Si costruisce una circonferenza centrata nell'evento $\boldsymbol{s}_i$ di raggio pari alla distanza tra esso e l'evento $\boldsymbol{s}_j$; la proporzione della circonferenza che giace entro la regione (si veda la figura A.2) è usata per determinare il peso g_{ij} da attribuire alle distanze r_{ij} minori di r tra $\boldsymbol{s}_i$ ed $\boldsymbol{s}_j$. Per ogni processo stazionario e isotropico, g_{ij} è la probabilità condizionale che un evento sia osservato, dato soltanto che la sua distanza dall'evento $\boldsymbol{s}_i$ è $r_{ij} < r$, da notare che, in generale, $g_{ij} \neq g_{ji}$. Poiché il numero atteso di eventi in A è $\lambda \mid A \mid$, segue che il numero atteso di coppie ordinate

di eventi ad una distanza minore di r, con il primo evento di ogni coppia in A, è $\lambda^2 \mid A \mid K(r)$. Uno stimatore non distorto di $\lambda^2 \mid A \mid K(r)$ è dato quindi da

$$\sum_{i=1}^{n} \sum_{j=1,\neq i}^{n} g_{ij}^{-1} I_r(r_{ij}).$$

Sostituendo l'intensità ignota λ con una stima, data dall'intesità osservata $n/ \mid A \mid$, si ottiene lo stimatore di Ripley per $K(r)$ con $r > 0$,

$$\hat{K}'''(r) = \frac{\mid A \mid}{n^2} \sum_{i=1}^{n} \sum_{j=1,\neq i}^{n} g_{ij}^{-1} I_r(r_{ij}).$$

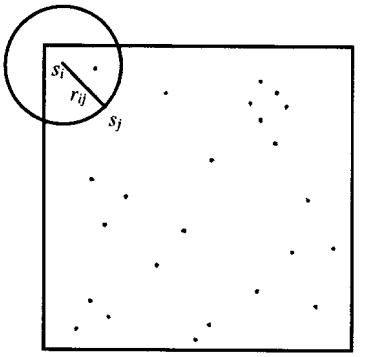


Figura A.2: Determinazione dei pesi da assegnare alle distanze

Questo è uno stimatore approssimativamente corretto per l'effetto confine per valori di r sufficientemente piccoli (l'approssimazione deriva dal fatto che $n/ \mid A \mid$ è uno stimatore leggermente distorto di λ, per λ noto lo stimatore è corretto). La restrizione ad r piccoli è necessaria dato che i pesi g_{ij}^{-1} possono divenire illimitati quando r aumenta. Per forme semplici della regione A (ad esempio rettangoli o cerchi) è possibile scrivere formule esplicite per g_{ij}, che fanno uso delle distanze tra un evento ed il suo vicino più prossimo e le distanze dal confine della regione. Anche questo metodo di correzione si basa sull'assunzione di isotropia: dato un evento $\boldsymbol{s}_j$ ad una distanza minore di r da $\boldsymbol{s}_i$, si ipotizza l'esistenza di altri eventi esterni al confine alla stessa distanza, il numero di eventi esterni è pari alla proporzione di circonferenza che giace entro il confine.

Anche nel caso della distribuzione empirica delle distanze *nearest-neigbour* si può considerare un sistema di ponderazione, in questo caso si riassegnano

i valori delle distanze per gli eventi vicini al bordo. Si considera, per ogni distanza r da un certo evento $\boldsymbol{s}_i$, l'area del cerchio centrato sull'evento e di raggio r; se il confine della regione analizzata cade dentro tale cerchio, un modo possibile di correggere per l'effetto confine è riassegnare il valore alla distanza nel seguente modo

$$r = \sqrt{\frac{\mid C^* \mid}{\pi}},$$

con $\mid C^* \mid$ l'area della parte del cerchio che si trova interna al confine. Tale metodo ha però il limite di essere di difficile applicazione per regioni di forma complessa.

Appendice B

Correzione per l'effetto confine nei modelli autoregressivi spaziali

B.1 Stima di massima verosimiglianza per modelli autoregressivi spaziali gaussiani di primo ordine

Il modello autoregressivo simultaneo (SAR) di Whittle (1954) è definito nel seguente modo

$$X_i = \sum_{j \neq i} g_{ij} X_j + \epsilon_i,$$

per $i = 1, \ldots, n$, dove gli ϵ_i rappresentano dei termini di errore incorrelati con $\mathrm{E}(\epsilon_i) = 0$ e $\mathrm{var}(\epsilon_i) = \sigma_i^2$. I termini g_{ij} definiscono la struttura autoregressiva del modello. Per semplicità si assume $\mathrm{E}(X_i) = 0$ per ogni i. In notazione matriciale il modello può essere scritto come

$$\boldsymbol{X} = \boldsymbol{G}\boldsymbol{X} + \boldsymbol{\epsilon},$$

con

$$\boldsymbol{X} = (X_1, \ldots, X_n)^T, \qquad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \qquad \text{e} \qquad \boldsymbol{G}_{n \times n} = \{g_{ij}\},$$

assumendo

$$\mathrm{E}(\boldsymbol{\epsilon}) = 0 \qquad \text{e} \qquad \mathrm{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix}.$$

Il modello può anche essere scritto come

$$\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{G})^{-1}\boldsymbol{\epsilon}, \qquad \text{con} \qquad \mathrm{E}(\boldsymbol{X}) = 0$$

e quindi

$$\begin{aligned} \boldsymbol{V} = \mathrm{var}(\boldsymbol{X}) &= \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T) = (\boldsymbol{I}-\boldsymbol{G})^{-1}\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)(\boldsymbol{I}-\boldsymbol{G}^T)^{-1} = \\ &= (\boldsymbol{I}-\boldsymbol{G})^{-1}\boldsymbol{\Sigma}(\boldsymbol{I}-\boldsymbol{G}^T)^{-1}. \end{aligned}$$

Se $\boldsymbol{\epsilon}$ è distribuito come un vettore normale multivariato, anche $\boldsymbol{X}$ ha una distribuzione normale multivariata, cioè $\boldsymbol{X} \sim \mathrm{NMV}(0, \boldsymbol{V})$; naturalmente, si può generalizzare operando una traslazione della distribuzione con un vettore di medie $\boldsymbol{\mu}$.

Bartlett (1971) e Besag (1974) specificano il modello autoregressivo condizionale (CAR) come segue

$$\mathrm{E}(X_i \mid \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) = \sum_{j \neq i} g_{ij} x_j \quad \text{e} \quad \mathrm{var}(X_i \mid \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) = \sigma_i^2,$$

per $i = 1, \ldots, n$, con $\boldsymbol{X}_{-i}$ e $\boldsymbol{x}_{-1}$ che indicano rispettivamente il vettore $\boldsymbol{X}$ ed il vettore dei valori osservati, senza considerare la componente i-esima. Quando le distribuzioni condizionali sono tutte normali, la distribuzione congiunta è normale multivariata $\boldsymbol{X} \sim \mathrm{NMV}(0, \boldsymbol{V})$ dove

$$\boldsymbol{V} = (\boldsymbol{I}-\boldsymbol{G})^{-1}\boldsymbol{\Sigma} \qquad \text{con} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix}.$$

La matrice $\boldsymbol{G}$ è scritta nella forma $\rho\boldsymbol{W}$, con $\boldsymbol{W}$ una matrice delle adiacenze con elemento generico w_{ij} unitario quando le aree i e j sono adiacenti, nullo in caso contrario.

Se si considera, per semplicità, $\sigma_i^2 = \sigma^2$ per ogni i, nel modello autoregressivo simultaneo, la matrice di varianza-covarianza sarà data da

$$\boldsymbol{V} = \sigma^2(\boldsymbol{I}-\boldsymbol{G})^{-1}(\boldsymbol{I}-\boldsymbol{G}^T)^{-1}, \tag{B.1}$$

dove $\boldsymbol{G}$ non è necessariamente simmetrica e $(\boldsymbol{I}-\boldsymbol{G})$ è non singolare e $\boldsymbol{V}$ è definita positiva; nel modello autoregressivo condizionale la matrice di varianza-covarianza ha invece la forma

$$\boldsymbol{V} = \sigma^2(\boldsymbol{I}-\boldsymbol{G})^{-1}, \tag{B.2}$$

dove $\boldsymbol{G}$ deve essere simmetrica perché lo sia $(\boldsymbol{I}-\boldsymbol{G})$ e quindi $\boldsymbol{V}$, ed inoltre $(\boldsymbol{I}-\boldsymbol{G})$ deve essere positiva definita sempre perché lo sia $\boldsymbol{V}$. È interessante notare che nel caso autoregressivo gaussiano di primo ordine, ogni

modello simultaneo con matrice $\boldsymbol{G}_s$, può essere espresso tramite un modello autoregressivo condizionale con matrice

$$\boldsymbol{G}_c = \boldsymbol{G}_s + \boldsymbol{G}_s^T - \boldsymbol{G}_s\boldsymbol{G}_s^T.$$

La stima di massima verosimiglianza dei parametri del modello autoregressivo simultaneo (SAR) e del modello autoregressivo condizionale (CAR), nel caso in cui $\boldsymbol{\epsilon} \sim \text{NMV}(0, \sigma^2\boldsymbol{I})$, quindi $\boldsymbol{X} \sim \text{NMV}(0, \boldsymbol{V})$, è la seguente. Si supponga di avere le osservazioni $\boldsymbol{x} = (x_1, \ldots, x_n)^T$. La log-verosimiglianza è data da

$$\ell(\boldsymbol{V} \mid \boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log \mid \boldsymbol{V} \mid -\frac{1}{2}\boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{x}.$$

Nel modello autoregressivo condizionale, considerando l'espressione (B.2) per la matrice di varianza-covarianza (con $\boldsymbol{G} = \rho\boldsymbol{W}$), dato che

$$\mid \sigma^2(\boldsymbol{I} - \rho\boldsymbol{W})^{-1} \mid = \sigma^{2n} \mid \boldsymbol{I} - \rho\boldsymbol{W} \mid^{-1}$$

si ottiene

$$\begin{aligned} \ell(\sigma^2, \rho \mid \boldsymbol{W}, \boldsymbol{x}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) + \\ &+ \frac{1}{2}\log \mid \boldsymbol{I} - \rho\boldsymbol{W} \mid -\frac{\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}}{2\sigma^2}. \end{aligned} \qquad \text{(B.3)}$$

L'equazione di stima per σ^2 è

$$\frac{\partial\ell(\sigma^2, \rho \mid \boldsymbol{W}, \boldsymbol{x})}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}}{2\sigma^4} = 0$$

da cui si ottiene lo stimatore di massima verosimiglianza di σ^2 per ρ fissato

$$\hat{\sigma}^2 = \frac{\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}}{n}.$$

Sostituendo nella funzione di log-verosimiglianza (B.3) σ^2 con la sua stima $\hat{\sigma}^2$ si ottiene il profilo di verosimiglianza

$$\ell(\rho \mid \boldsymbol{W}, \boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\frac{\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}}{n} + \frac{1}{2}\log \mid \boldsymbol{I} - \rho\boldsymbol{W} \mid,$$

che massimizzato rispetto a ρ, fornisce la stima di massima verosimiglianza del coefficiente di autocorrelazione spaziale.

Nel modello autoregressivo simultaneo, la matrice di varianza-covarianza è data dall'espressione (B.1), la log-verosimiglianza assume quindi la forma

$$\begin{aligned}\ell(\sigma^2, \rho \mid \boldsymbol{W}, \boldsymbol{x}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) + \log \mid \boldsymbol{I} - \rho\boldsymbol{W} \mid + \\ &- \frac{1}{2\sigma^2}\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W}^T)(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}. \qquad \text{(B.4)}\end{aligned}$$

L'equazione di stima per σ^2 è

$$\frac{\partial\ell(\sigma^2, \rho \mid \boldsymbol{W}, \boldsymbol{x})}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W}^T)(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}}{2\sigma^4} = 0$$

da cui si ottiene lo stimatore di massima verosimiglianza di σ^2 per ρ fissato

$$\hat{\sigma}^2 = \frac{\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W}^T)(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}}{n}.$$

Sostituita nella log-verosimiglianza (B.4) riduce il problema alla massimizzazione del profilo di log-verosimiglianza

$$\begin{aligned}\ell(\rho \mid \boldsymbol{W}, \boldsymbol{x}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2} + \\ &- \frac{n}{2}\log\frac{\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W}^T)(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x}}{n} + \log \mid \boldsymbol{I} - \rho\boldsymbol{W} \mid,\end{aligned}$$

rispetto a ρ. Si noti che

$$\begin{aligned}\boldsymbol{x}^T(\boldsymbol{I} - \rho\boldsymbol{W}^T)(\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{x} &= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\rho\sum_{j=1}^{n} w_{ij}x_j + \\ &+ \rho^2\sum_{i=1}^{n}(\sum_{j=1}^{n} w_{ij}x_j)^2. \qquad \text{(B.5)}\end{aligned}$$

Per la stima dei parametri dei modelli autoregressivi, sia per quello nella formulazione simultanea che per quello nella formulazione condizionale, la cosa più difficile, computazionalmente, è determinare $\mid \boldsymbol{I} - \rho\boldsymbol{W} \mid$ per diversi valori di ρ, in modo da operare la massimizzazione. A tal proposito Ord (1975) suggerisce il seguente metodo: siano $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ gli autovalori di $\boldsymbol{W}$, allora

$$\mid \boldsymbol{I} - \rho\boldsymbol{W} \mid = \prod_{i=1}^{n}(1 - \rho\lambda_i).$$

In tal modo la massimizzazione del profilo di log-verosimiglianza rispetto a ρ diviene relativamente più semplice, vengono determinati gli autovalori di $\boldsymbol{W}$ una sola volta e vengono sostituiti nell'espressione del profilo di verosimiglianza che diviene più semplice da massimizzare.

B.2 Metodi proposti per la correzione dell'effetto confine nei modelli autoregressivi spaziali

B.2.1 La mappa di forma toroidale

Griffith e Amrhein (1983) e Griffith (1983) correggono per l'effetto confine la stima del coefficiente di autocorrelazione in un modello autoregressivo simultaneo aggirando il problema di finitezza della regione, come proposto nel caso della stima della funzione K (si veda paragrafo A.2.1); pensando la superficie come infinita, sotto l'assunzione di isotropicità del processo generatore del termine di errore $\boldsymbol{\epsilon}$, considerano la mappa a forma geometrica toroidale. Se la condizione di isotropicità è soddisfatta, infatti, i dati possono essere considerati come parte rappresentativa dell'intero processo e quindi, per ricostruire il fenomeno nelle altre aree circostanti, si può pensare di replicare l'andamento del fenomeno osservato. La costruzione della mappa è fatta prima unendo le aree al bordo site su due lati opposti della regione, così la regione assume una forma a cilindro, successivamente unendo le aree poste a lati estremi del cilindro; si veda la figura 1.2 nel capitolo 1. Correggendo l'effetto confine in questo modo, quando si stima il coefficiente di autocorrelazione ρ, la matrice stocastica delle adiacenze $\boldsymbol{W}$ rimane $n \times n$, ma adesso ha alcuni valori non più nulli per le aree al bordo; nella funzione da massimizzare per ottenere la stima di ρ, i termini osservati $\boldsymbol{x}$ continuano ad essere gli n termini originari ma gli autovalori usati per la stima sono quelli della nuova matrice $\boldsymbol{W}$. Per un modello autoregressivo condizionale gaussiano del primo ordine, lo stesso procedimento è stato proposto da Moran (1973).

B.2.2 Eliminazione di una corona interna

Ancora seguendo l'idea di quanto proposto per la correzione dell'effetto confine nel caso di dati individuali (paragrafo A.2.2), Griffith (1983) e Griffith e Amrhein (1983) propongono di considerare una corona interna (*empirical buffer zone* o area di guardia), per un reticolo regolare $p \times q$ essa comprende $2q + 2p - 4$ aree; nella stima del coefficiente di autocorrelazione si continua ad usare tutte le n aree per quanto riguarda la struttura delle adiacenze, ma i valori osservati nelle aree della corona interna non sono considerati nella somma dei quadrati nel primo termine della (B.5), né è considerata, nel secondo e terzo termine della (B.5), la loro equazione autoregressiva. Nel caso di un modello markoviano di ordine superiore ad uno, la corona interna dovrà contenere un numero maggiore di aree, così, adottando tale metodo di corre-

zione, si ha una notevole perdita di informazione, che aumenta considerando modelli markoviani di ordine superiore.

Comunque, espandendo il reticolo all'infinito, la perdita di informazione tende a zero, in quanto ad esso tende la proporzione di aree nella corona interna sul totale; nella realtà, poiché si hanno un numero limitato di osservazioni, la perdita di informazioni diventa notevole. Per un modello autoregressivo condizionale, tale metodo è stato proposto da Gleeson e McGilchrist (1980).

Si noti che tale metodo risulta essere valido nel caso in cui sia possibile reperire informazioni su una corona esterna della regione analizzata, in tal modo la perdita di informazione è relativa al fenomeno all'esterno della regione di interesse. Quando non sia possibile ottenere tali dati, una possibilità è ricostruirli.

B.2.3 Costruzione di un'area esterna artificiale

Per ovviare al problema di perdita di informazioni Griffith e Amrhein (1983) propongono di eliminare non una parte della regione analizzata, ma un'area esterna artificiale (*artificial buffer zone*), costruita come corona esterna composta da un'unica area a cui è assegnato come valore la media di tutti i valori osservati all'interno della regione (si veda la figura 1.3.a nel capitolo 1); ciò è plausibile sotto l'assunzione di isotropicità. Quando si stima ρ usando questa tecnica di correzione, la matrice delle adiacenze $\boldsymbol{W}$ diventa $(n+1) \times (n+1)$, con $w_{n+1,i} = 0$ per ogni $i = 1, \ldots, n$ e $w_{i,n+1} > 0$ solo se l'area i-esima è un'area posta al confine della regione.

Nella funzione di verosimiglianza compaiono gli $n+1$ autovalori della nuova matrice $\boldsymbol{W}$; il valore x_{n+1} non compare nel primo termine dell'espressione (B.5), cioè nella somma dei quadrati degli x_i e, poiché $w_{n+1,i} = 0$, nel secondo e terzo termine dell'espressione (B.5), non viene considerata la sua equazione autoregressiva, benché il valore x_{n+1} venga considerato nelle altre. L'inconveniente di questo metodo è che assumendo un valore costante per i vicinati oltre confine, si altera la variazione spaziale osservata nella regione.

B.2.4 Costruzione di più aree esterne artificiali

Una variante della precedente proposta è quella (si veda Griffith 1983 e Griffith e Amrhein 1983) di costruire più aree esterne artificiali. In questo caso la corona esterna è partizionata in m diverse aree, $2p + 2q$, a cui è assegnato quale valore la media di quelli osservati all'interno della regione (si veda la

figura 1.3.b nel capitolo 1); si noti che questa volta vengono considerate le interazioni tra queste aree e così la matrice $\boldsymbol{W}$ diviene $(n+m) \times (n+m)$. In questo caso, nell'espressione (B.5), la somma dei termini al quadrato riguarda tutte le $n+m$ aree e gli autovalori sono $n+m$, quelli della nuova matrice $\boldsymbol{W}$. Per il modello autoregressivo condizionale tale correzione è stata sviluppata da Haining(1978).

B.2.5 Costruzione di aree esterne usando un modello trend surface

Griffith (1983), propone di assegnare alle aree esterne artificiali i valori ottenuti dalla estrapolazione di una superficie stimata mediante un modello *trend surface*. Ad ogni centroide di ogni area della regione è assegnata una coppia di coordinate cartesiane e tramite un'interpolazione con una superficie si ottiene una media del valore del processo. In altre parole si stima una superficie sui punti campionari (le coordinate del centroide di ogni area della regione) riportandosi allo studio usuale di dati geologici, cioè ad un campione da un processo spaziale dove l'insieme degli indici varia nel continuo (si veda figura B.1).

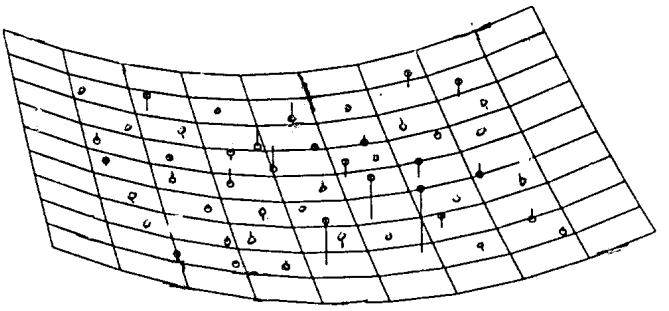


Figura B.1: Assegnazione dei valori per le aree esterne artificiali mediante estrapolazione di una superficie stimata usando un modello *trend surface*

La superficie

$$f(\boldsymbol{s}_i) = f(u_i, v_i) = \sum_{r+s \leq p} a_{rs} u_i^r v_i^s,$$

per $i = 1, \ldots, n$, viene stimata mediante il metodo dei minimi quadrati, con p che rappresenta l'ordine della superficie. Le funzioni adottate in due dimensioni sono solitamente quella costante a (per $p = 0$), la lineare $a+bu+cv$

(per $p = 1$) o la quadratica $a+bu+cv+du^2+euv+fv^2$ (per $p = 2$). Occorre, in generale, stimare $(p+1)(p+2)/2$ coefficienti; questo viene fatto minimizzando

$$\sum_{i=1}^{n} [X(\boldsymbol{s}_i) - f(\boldsymbol{s}_i)]^2 ,$$

per ulteriori spiegazioni si veda, per esempio, Ripley 1981, pag. 29–35). Tale metodo di correzione è tuttavia criticabile, dato che i valori estrapolati risentono anch'essi dell'effetto confine. Infatti ai bordi della regione analizzata la superficie assume valori instabili, poiché stimata su poche osservazioni. Tale effetto è presente anche nella regressione polinomiale nel caso di una dimensione; in più dimensioni diviene più grave, dato che ci sono più confini che ne sono affetti. Quindi il metodo di correzione non sembra valido dato che risente esso stesso del problema dell'effetto confine; tenderà a dare valori o troppo alti o troppo bassi alle aree artificiali.

B.2.6 Metodo dei dati mancanti

Tale metodo di correzione considera il problema come un problema di dati mancanti. Per tale proposta si veda Haining, Griffith e Bennet (1984, 1989), Martin (1984, 1987) e Griffith (1983, 1985, 1988 e 1989). La correzione per l'effetto confine avviene stimando simultaneamente i valori mancanti e i parametri del modello, questo è possibile all'interno di un modello autoregressivo condizionale del primo ordine, dove l'informazione delle aree confinanti può essere usata per stimare i valori mancanti.

Si consideri la costruzione di una corona esterna costituita da aree i cui valori sono trattati come mancanti; per semplicità si analizza prima il problema di un singolo valore mancante x_{n+1}, con $X_i \sim \mathrm{N}(\mu, \sigma^2)$, per ogni $i = 1, \ldots, n+1$. La stima di massima verosimiglianza di x_{n+1} è derivabile dalla funzione di verosimiglianza

$$\begin{aligned} L(x_{n+1}, \mu, \sigma \mid x_{-(n+1)}) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n+1} \exp\left[-\frac{(x_{n+1}-\mu)^2}{2\sigma^2}\right] \times \\ &\times \prod_{i=1}^{n} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right], \end{aligned}$$

risolvendo il sistema di equazioni per σ, μ e x_{n+1}, si ottiene che la stima di massima verosimiglianza per x_{n+1} è data da dalla media aritmetica degli n valori osservati; alla corona esterna può essere assegnato quale valore il

valore medio dei valori delle aree della regione, come è fatto nel metodo della costruzione di un area artificiale visto in precedenza. Ma poiché x_{n+1}, in un modello autoregressivo, è una funzione di $\rho \sum_{j=1}^{n} w_{n+1,j} x_j$, questa informazione addizionale può essere usata per ottenere degli stimatori corretti per l'effetto confine di ρ e x_{n+1}.

Consideriamo il caso generale di n valori osservati ed m valori mancanti, si assuma $\boldsymbol{X}$ normale multivariata con media $\boldsymbol{\mu}$ e matrice di varianza-covarianza $\boldsymbol{V} = \sigma^2(\boldsymbol{I} - \rho\boldsymbol{W})^{-1} = \sigma^2\boldsymbol{\Sigma}$. La matrice $\boldsymbol{\Sigma} = (\boldsymbol{I} - \rho\boldsymbol{W})^{-1}$, può essere partizionata nel seguente modo

$$\boldsymbol{\Sigma} = \begin{bmatrix} (\boldsymbol{I} - \rho\boldsymbol{W})_{oo} & -\rho\boldsymbol{W}_{om} \\ -\rho\boldsymbol{W}_{mo} & (\boldsymbol{I} - \rho\boldsymbol{W})_{mm} \end{bmatrix}^{-1}$$

dove o denota la porzione della matrice associata alle aree della regione con dati osservati ed m quella relativa ai dati mancanti. La stima dei valori mancanti, si veda Orchard e Woodbury (1972), è data da

$$\hat{\boldsymbol{X}}_m = \bar{\boldsymbol{X}}_o + \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}(\boldsymbol{X}_o - \bar{\boldsymbol{X}}_o)$$

con $\bar{\boldsymbol{X}}_o$ il vettore che ha come valori la media dei valori osservati. Questa può essere scritta, considerando la partizione prima definita

$$\hat{\boldsymbol{X}}_m = \bar{\boldsymbol{X}}_o + \rho\boldsymbol{W}_{mo}(\boldsymbol{I} - \rho\boldsymbol{W})_{mm}^{-1}(\boldsymbol{X}_o - \bar{\boldsymbol{X}}_o).$$

Nel caso si consideri un solo valore mancante, allora $(\boldsymbol{I} - \rho\boldsymbol{W})_{mm} = 1$ e quindi l'equazione si riduce a quella di Besag del modello condizionale

$$\mathrm{E}(X_{n+1} \mid \boldsymbol{X}_{-(n+1)}) = \bar{x}_o + \rho \sum_{j=1}^{n} w_{n+1,j}(x_j - \bar{x}_o);$$

se si considera più di un valore mancante, allora posto $\boldsymbol{A} = \boldsymbol{W}_{mo}(\boldsymbol{I} - \rho\boldsymbol{W})_{mm}^{-1}$, il generico valore mancante è dato da

$$\hat{X}_m = \bar{x}_o + \rho \sum_{j=1}^{n} a_{mj}(x_j - \bar{x}_o).$$

I valori mancanti possono essere stimati usando un algoritmo iterativo, inizializzando i valori mancanti delle aree $\bar{\boldsymbol{X}}_m^{(0)}$ con un vettore di elementi $\bar{\boldsymbol{X}}_o$, la r-esima iterazione è data dai seguenti passi:

1. si calcola $\hat{\rho}$ per un modello autoregressivo condizionale usando l'equazione di massima verosimiglianza;

2. si stima $\boldsymbol{X}_m$ con l'equazione

$$\hat{\boldsymbol{X}}_m^{(r)} = \bar{\boldsymbol{X}}_m^{(r-1)} + \rho \boldsymbol{W}_{mo}(\boldsymbol{I} - \rho \boldsymbol{W})_{mm}^{-1}(\boldsymbol{X}_o - \bar{\boldsymbol{X}}_o);$$

3. si determina $\bar{\boldsymbol{X}}_m^{(r)}$ e si sostituisce al valore corrente di $\bar{\boldsymbol{X}}_m^{(r-1)}$, poi si ritorna al passo (1).

Gli ultimi tre passi sono ripetuti finché non è raggiunta la convergenza della stima di ρ. Si noti che i valori mancanti ricostruiti, una volta usati per la stima di ρ, sono esclusi da qualsiasi altra analisi di tipo statistico e che la matrice $(\boldsymbol{I} - \rho \boldsymbol{W})$ deve essere necessariamente simmetrica. Si noti che il metodo tradizionale che prevede la costruzione di un'area artificiale con valore pari alla media delle aree interne, si ferma al passo (2) della prima iterazione.

B.2.7 Modello di regressione con una variabile dummy

Griffith (1983, 1985) propone di discriminare le aree appartenenti alla corona interna da quelle interne alla regione mediante l'inserimento di una variabile *dummy* in un modello di regressione. Il fenomeno viene trattato come una realizzazione di un processo non omogeneo che opera sul piano; l'inomogeneità è tale che la media delle aree sul bordo è diversa da quella delle aree interne alla regione ed ogni differenza, quando diversa da zero, è attribuita all'effetto confine. Le due medie rappresentano i valori stimati attribuiti alle due diverse tipologie di aree, ottenute dalla stima dei minimi quadrati ordinari del modello di regressione che contiene una costante e una variabile *dummy*, che assume valore non nullo per le aree sul confine; da tale suddivisione è stata proposta una procedura iterativa per la stima del coefficiente di autocorrelazione ρ per un modello autoregressivo simultaneo. Quando si stimano i coefficienti del modello di regressione, con costante e la variabile *dummy*, si ottiene, come stima della costante, la media delle aree interne alla regione, e come coefficiente della variabile *dummy* la differenza tra la media calcolata nelle aree nella corona e quelle nelle aree interne. Di conseguenza, i valori stimati per le aree sono distinti, nel senso che le aree nella corona sono stimate dalla loro media $\bar{x}_o$ e quelle interne dalla loro media $\bar{x}_\emptyset$.

Si può correggere la stima del parametro di autocorrelazione nel seguente modo:

1. si determinano gli scarti tra i valori delle aree e i valori stimati di esse, quindi $e_i = x_i - \bar{x}_o$ per i che indica le aree al confine e $e_i = x_i - \bar{x}_\emptyset$ per i che indica le aree interne alla regione;

2. si stima il parametro di autocorrelazione spaziale $\hat{\rho}$ per il modello autoregressivo simultaneo applicato agli scarti ottenuti nel passo (1) usando il metodo della massima verosimiglianza;

3. si costruisce una nuova variabile $x'_i = x_i - \hat{\rho} \sum_{j=1}^{n} w_{ij} x_j$ e poi si determinano le media delle aree al confine e di quelle interne $\bar{x}'_o$ e $\bar{x}'_\emptyset$;

4. si sostituisce x_i, $\bar{x}_o$ e $\bar{x}_\emptyset$ con x'_i, $\bar{x}'_o$ e $\bar{x}'_\emptyset$ e si torna al passo (1).

Si ripetono i quattro passi, sino a convergenza della stima del parametro ρ.

B.2.8 Metodo dei minimi quadrati generalizzati

Griffith (1980, 1983, 1985) propone di determinare una matrice di trasformazione che corregga per l'effetto confine in maniera analoga a quanto è stato proposto in una dimensione nelle serie temporali (si veda Wonnacott e Wonnacott, 1970 a pag. 328); questo è anche il metodo che viene usato per ricondurre una stima fatta con il metodo dei minimi quadrati generalizzati in un modello con errori autocorrelati ad una stima eseguita mediante il metodo dei minimi quadrati.

Si consideri il modello autoregressivo di primo ordine per una serie temporale

$$X_t = \rho \boldsymbol{C} X_t + \epsilon_t \qquad \text{o} \qquad (\boldsymbol{I} - \rho \boldsymbol{C}) X_t = \epsilon_t$$

con $t = 1, \ldots$ e la matrice $\boldsymbol{C}$ con elementi $c_{ij} = 1$ se $i - 1 = j$ e $c_{ij} = 0$ altrimenti e si consideri anche un limite inferiore sulla serie temporale X_t, cioè un termine X_0 sarà presente come effetto ma non fra i dati. Per considerare tale influenza nella stima del parametro di autoregressione si deve ricorrere alla trasformazione $\epsilon_1 = x_1 \sqrt{1 - \rho^2}$; tale trasformazione si ottiene premoltiplicando la matrice $(\boldsymbol{I} - \rho \boldsymbol{C})$ per la matrice Γ, così da ottenere una nuova espressione

$$\Gamma(\boldsymbol{I} - \rho \boldsymbol{C}) X_t = \epsilon_t$$

che soddisfa la condizione

$$\Omega^{-1} \quad = \quad (\boldsymbol{I} - \rho \boldsymbol{C})^T \Gamma^T \Gamma (\boldsymbol{I} - \rho \boldsymbol{C}) =$$

$$= \begin{bmatrix} 1 & -\rho & & & 0 \\ -\rho & 1+\rho^2 & -\rho & & \\ & -\rho & 1+\rho^2 & -\rho & \\ & & \ddots & \ddots & \ddots \\ & & -\rho & 1+\rho^2 & -\rho \\ 0 & & & -\rho & 1 \end{bmatrix} \tag{B.6}$$

con Ω che rappresenta la matrice di varianza covarianza (a meno di un fattore di scala) nella teoria dei minimi quadrati generalizzati relativa ad un modello con disturbi autocorrelati. Risolvendo l'equazione per $\Gamma^T\Gamma$ ignoto si ottiene

$$\Gamma^T\Gamma = (\boldsymbol{I} - \rho\boldsymbol{C})^{-T}\Omega^{-1}(\boldsymbol{I} - \rho\boldsymbol{C})^{-1};$$

il problema viene usualmente risolto trasformando il problema nella soluzione di un equazione che coinvolge la matrice diagonale degli autovalori di $\Gamma^T\Gamma$ e la matrice degli autovettori ad essi associati. Nel caso delle serie temporali si ottiene

$$\Gamma = \begin{bmatrix} \sqrt{1-\rho^2} & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}$$

da cui

$$\Gamma(\boldsymbol{I} - \rho\boldsymbol{C}) = \begin{bmatrix} \sqrt{1-\rho^2} & & & 0 \\ -\rho & 1 & & \\ & \ddots & \ddots & \\ 0 & & -\rho & 1 \end{bmatrix}$$

che soddisfa la relazione (B.6).

Lo stesso metodo può essere generalizzato a più dimensioni per processi spaziali; sfortunatamente però in questo caso si è in presenza di dipendenze multidirezionali e quindi la soluzione è più complessa. La matrice di trasformazione Γ in tal caso è più complicata a causa di dipendenze multiple: la trasformazione operata solo per il primo valore della serie, adesso va estesa a più termini, sarà necessario che sia soddisfatta

$$(\boldsymbol{I} - \rho\boldsymbol{W})^T\Gamma^T\Gamma(\boldsymbol{I} - \rho\boldsymbol{W}) = \Omega^{-1}.$$

Tale metodo di correzione per l'effetto confine è inapplicabile nelle analisi empiriche, dato che è necessario conoscere ρ, così il suo uso è principalmente dedicato ad esperimenti di simulazione, in cui ρ è un parametro di ingresso nell'analisi.

Appendice C

Algoritmi per il trattamento dei dati mancanti

C.1 La condizione di ignorabilità

Per analizzare le distribuzioni a posteriori quando si opera in condizioni di dati mancanti in ambito bayesiano vengono generalmente adottati degli algoritmi (si veda, ad esempio, Tanner 1993, McLachlan e Krishnan 1997, Schafer 1997) nati con lo scopo di ottenere la a posteriori per i dati osservati tramite quella dei dati completi: un problema intrattabile di dati incompleti può essere, infatti, risolto lavorando ripetutamente su problemi di dati completi trattabili. Si indichino con $\boldsymbol{O}_{obs}$ i dati osservati, con $\boldsymbol{O}_{mis}$ quelli mancanti e con $\boldsymbol{O}_{com}$ i completi; i parametri del modello sono dati dal vettore $\boldsymbol{\theta}$.

Quando si analizzano problemi di dati mancanti, è necessario considerare il meccanismo che determina l'assenza di tali informazioni. Infatti, le metodologie qui trattate, se applicate quando tale meccanismo sia non "ignorabile", possono dare origine ad un'inferenza sbagliata. Prima di applicare uno dei metodi qui esposti, è quindi necessario assicurarsi che tale meccanismo soddisfi la condizione di ignorabilità (si veda Little e Rubin 1987, Rubin 1987): essa è soddisfatta quando lo sono le condizioni di MAR e di distinguibilità dei parametri, analizzate di seguito.

Un dato si dice mancante a caso (MAR, *Missing At Random*), nel senso di Rubin (1976), quando la probabilità che l'osservazione sia mancante può dipendere da $\boldsymbol{O}_{obs}$ ma non da $\boldsymbol{O}_{mis}$, ossia quando la probabilità che un dato sia mancante non dipende dal dato, ma solo indirettamente attraverso quantità che sono osservate. Formalmente Rubin (1976) definisce la condizione MAR in termini di un modello probabilistico per i dati mancanti. Sia $\boldsymbol{R}$ una matrice $n \times p$ di variabili indicatrici (con n il numero complessivo dei dati e

p il numero delle variabili) i cui elementi sono zero o uno in relazione al fatto che i dati corrispondenti siano osservati o mancanti. Si definisce, per $\boldsymbol{R}$, il seguente modello probabilistico $p(\boldsymbol{R} \mid \boldsymbol{O}_{com}, \boldsymbol{\xi})$, che dipende da $\boldsymbol{O}_{com}$ e da un vettore di parametri ignoto $\boldsymbol{\xi}$. L'assunzione MAR è che la distribuzione non dipenda da $\boldsymbol{O}_{mis}$, cioè

$$p(\boldsymbol{R} \mid \boldsymbol{O}_{obs}, \boldsymbol{O}_{mis}, \boldsymbol{\xi}) = p(\boldsymbol{R} \mid \boldsymbol{O}_{obs}, \boldsymbol{\xi}).$$

Si noti che nel caso in cui la distribuzione non dipenda neanche da $\boldsymbol{O}_{obs}$, è assunta la condizione particolare di dati mancanti completamente a caso (MCAR, *Missing Completely At Random*).

Si deve anche assumere che i parametri del modello $\boldsymbol{\theta}$ e quelli $\boldsymbol{\xi}$ del meccanismo di mancanza dei dati, siano distinti: questo significa che lo spazio parametrico congiunto di $(\boldsymbol{\theta}, \boldsymbol{\xi})$ è il prodotto cartesiano incrociato degli spazi parametrici singoli ossia, in una prospettiva bayesiana, questo significa che ogni distribuzione a priori congiunta applicata a $(\boldsymbol{\theta}, \boldsymbol{\xi})$ può essere fattorizzata tramite due a priori marginali indipendenti per $\boldsymbol{\theta}$ e per $\boldsymbol{\xi}$. In molte situazioni questo è intuitivamente ragionevole, dato che la conoscenza di $\boldsymbol{\theta}$ dà poca informazione circa $\boldsymbol{\xi}$ e viceversa.

Se sia la condizione di MAR che quella di distinguibilità sono soddisfatte, il meccanismo di generazione dei dati mancanti è detto essere ignorabile.

In ambito bayesiano l'inferenza è basata su una distribuzione a posteriori per i parametri ignoti condizionata ai dati osservati. I parametri ignoti sono $(\boldsymbol{\theta}, \boldsymbol{\xi})$ e le quantità osservate sono $\boldsymbol{O}_{obs}$ ed $\boldsymbol{R}$. Dal teorema di Bayes, la distribuzione a posteriori può essere scritta come

$$p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \boldsymbol{O}_{obs}, \boldsymbol{R}) = \frac{p(\boldsymbol{R}, \boldsymbol{O}_{obs} \mid \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{\theta}, \boldsymbol{\xi})}{\int\int p(\boldsymbol{R}, \boldsymbol{O}_{obs} \mid \boldsymbol{\theta}, \boldsymbol{\xi}) p(\boldsymbol{\theta}, \boldsymbol{\xi}) d\boldsymbol{\theta} d\boldsymbol{\xi}}.$$

Sotto l'assunzione di MAR, poiché

$$\begin{aligned} p(\boldsymbol{R}, \boldsymbol{O}_{obs} \mid \boldsymbol{\theta}, \boldsymbol{\xi}) &= p(\boldsymbol{R} \mid \boldsymbol{O}_{obs}, \boldsymbol{\xi}) \int p(\boldsymbol{O}_{com} \mid \boldsymbol{\theta}) d\boldsymbol{O}_{mis} \\ &= p(\boldsymbol{R} \mid \boldsymbol{O}_{obs}, \boldsymbol{\xi}) p(\boldsymbol{O}_{obs} \mid \boldsymbol{\theta}), \end{aligned}$$

si ha

$$p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \boldsymbol{O}_{obs}, \boldsymbol{R}) \propto p(\boldsymbol{R} \mid \boldsymbol{O}_{obs}, \boldsymbol{\xi}) p(\boldsymbol{O}_{obs} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}, \boldsymbol{\xi}).$$

L'inferenza bayesiana su $\boldsymbol{\theta}$ è basata sulla distribuzione marginale a posteriori ottenuta integrando questa funzione sul parametro di disturbo $\boldsymbol{\xi}$. Quando i

parametri $\boldsymbol{\theta}$ e $\boldsymbol{\xi}$ sono distinguibili, la distribuzione a priori è fattorizzata nel seguente modo:

$$p(\boldsymbol{\theta}, \boldsymbol{\xi}) = p(\boldsymbol{\theta})p(\boldsymbol{\xi}),$$

così la distribuzione marginale a posteriori per $\boldsymbol{\theta}$, sotto la condizione di ignorabilità, diviene

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}, \boldsymbol{R}) &= \int p(\boldsymbol{\theta}, \boldsymbol{\xi} \mid \boldsymbol{O}_{obs}, \boldsymbol{R}) d\boldsymbol{\xi} \\ &\propto p(\boldsymbol{O}_{obs} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \int p(\boldsymbol{R} \mid \boldsymbol{O}_{obs}, \boldsymbol{\xi}) p(\boldsymbol{\xi}) d\boldsymbol{\xi} \\ &\propto L(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) p(\boldsymbol{\theta}), \end{aligned}$$

dove la proporzionalità riguarda un fattore moltiplicativo che non dipende da $\boldsymbol{\theta}$. Dato che $\boldsymbol{R}$ non appare più si ha così

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}, \boldsymbol{R}) = p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}).$$

Sotto la condizione di ignorabilità tutta l'informazione relativa a $\boldsymbol{\theta}$ è contenuta nella a posteriori che ignora il meccanismo di dati mancanti,

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) \propto L(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) p(\boldsymbol{\theta})$$

detta a posteriori dei dati osservati.

C.2 L'algoritmo EM

L'algoritmo EM è tipicamente presentato come una tecnica per determinare le stime di massima verosimiglianza, ma come puntualizzato da Dempster, Laird e Rubin (1977), tale algoritmo può essere usato anche per calcolare le mode delle distribuzioni a posteriori, cioè dei valori per i quali è massimizzata la distribuzione a posteriori dei dati osservati, anziché la verosimiglianza dei dati osservati. L'algoritmo EM si basa sull'interdipendenza tra i dati mancanti $\boldsymbol{O}_{mis}$ ed i parametri $\boldsymbol{\theta}$. Il fatto che $\boldsymbol{O}_{mis}$ contenga rilevanti informazioni per stimare $\boldsymbol{\theta}$, e $\boldsymbol{\theta}$ a sua volta fornisca validi valori per $\boldsymbol{O}_{mis}$, suggerisce il seguente schema per la stima di $\boldsymbol{\theta}$ in presenza soltanto di $\boldsymbol{O}_{obs}$: assegnando dei valori a $\boldsymbol{O}_{mis}$ basati su una prima stima di $\boldsymbol{\theta}$, si ristima $\boldsymbol{\theta}$ basandosi su $\boldsymbol{O}_{obs}$ ed i valori imputati $\boldsymbol{O}_{mis}$ e si itera sino al raggiungimento della convergenza delle stime. Viene considerato, in un primo momento, l'algoritmo nella prospettiva di massimizzazione della log-verosimiglianza, poi esso viene riformulato per la massimizzazione della a posteriori.

C.2.1 Massimizzazione della funzione di verosimiglianza

In ogni problema di dati incompleti, la distribuzione di quelli completi $\boldsymbol{O}_{com}$ può essere fattorizzata come

$$p(\boldsymbol{O}_{com} \mid \boldsymbol{\theta}) = p(\boldsymbol{O}_{obs} \mid \boldsymbol{\theta}) p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}). \tag{C.1}$$

Vedendo i termini in funzione di $\boldsymbol{\theta}$ e passando ai logaritmi si possono considerare le log-verosimiglianze:

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{O}_{com}) = \ell(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) + \log p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}) + c,$$

dove $\ell(\boldsymbol{\theta} \mid \boldsymbol{O}_{com})$ rappresenta la log-verosimiglianza dei dati completi, $\ell(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$ la log-verosimiglianza dei dati osservati e c una costante arbitraria. Il termine $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta})$, chiamato distribuzione predittiva condizionata a $\boldsymbol{\theta}$ dei dati mancanti, gioca un ruolo centrale nell'algoritmo EM, dato che cattura l'interdipendenza tra $\boldsymbol{O}_{mis}$ e $\boldsymbol{\theta}$. Quando visto come distribuzione di probabilità esso riassume l'informazione circa $\boldsymbol{O}_{mis}$ per ogni valore assunto per $\boldsymbol{\theta}$, quando visto come funzione di $\boldsymbol{\theta}$ esso riassume l'evidenza circa $\boldsymbol{\theta}$ contenuta in $\boldsymbol{O}_{mis}$ che non sia già presente in $\boldsymbol{O}_{obs}$. Poiché $\boldsymbol{O}_{mis}$ non è noto, non è possibile determinare tale termine, così si considera la media dell'equazione fatta rispetto alla distribuzione predittiva $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{(t)})$, dove $\boldsymbol{\theta}^{(t)}$ è la stima preliminare del parametro ignoto. Tale media porta a

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \ell(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) + H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) + c$$

dove

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \int \ell(\boldsymbol{\theta} \mid \boldsymbol{O}_{com}) p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}_{(t)}) d\boldsymbol{O}_{mis}$$

e

$$H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \int \log p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}) p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{O}_{mis}.$$

Un importante risultato di Dempster, Laird e Rubin (1977) è che se $\boldsymbol{\theta}^{(t+1)}$ è il valore di $\boldsymbol{\theta}$ che massimizza $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ allora $\boldsymbol{\theta}^{(t+1)}$ è una stima migliore di $\boldsymbol{\theta}^{(t)}$, nel senso che in tale valore la log-verosimiglianza dei dati osservati è almeno come quella in $\boldsymbol{\theta}^{(t)}$, cioè

$$\ell(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{O}_{obs}) \geq \ell(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{O}_{obs}).$$

Questo può essere dimostrato scrivendo

$$\begin{aligned} \ell(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{O}_{obs}) - \ell(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{O}_{obs}) &= Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) + \\ &+ H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}). \end{aligned}$$

La quantità

$$Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})$$

è non negativa dato che, per ogni $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(t+1)}$ soddisfa

$$Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}).$$

La quantità $H(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})$, che può essere scritto come

$$\int \log \frac{p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{(t)})}{p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{(t+1)})} p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{O}_{mis},$$

è non negativa per la disuguaglianza di Jensen e la proprietà di convessità della funzione $x \log x$.

Un'iterazione dell'algoritmo EM consiste nei due seguenti passi distinti:

- al passo E (*Expectation*) si determina la funzione $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ tramite la media della log-verosimiglianza dei dati completi $\ell(\boldsymbol{\theta} \mid \boldsymbol{O}_{com})$ sulla distribuzione predittiva $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{(t)})$;

- al passo M (*Maximization*) si determina il valore $\boldsymbol{\theta}^{(t+1)}$ che massimizza $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$.

C.2.2 Massimizzazione della a posteriori

L'algoritmo EM può anche essere utilizzato per determinare le mode della distribuzione a posteriori dei dati osservati. Poiché la distribuzione a posteriori, dati $\boldsymbol{O}_{com}$, quando si assume una distribuzione a priori sui parametri $p(\boldsymbol{\theta})$ è data da

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}_{com}) \propto p(\boldsymbol{O}_{com} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

passando ai logaritmi e considerando la (C.1) si ottiene

$$\log p(\boldsymbol{\theta} \mid \boldsymbol{O}_{com}) = \ell(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) + \log p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + c.$$

Facendo una media di quest'equazione sulla distribuzione predittiva $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{(t)})$, si ottiene

$$Q^*(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \log p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) + H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) + \log p(\boldsymbol{\theta}) + c,$$

dove

$$Q^*(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) + \log p(\boldsymbol{\theta}^{(t)})$$

e le funzioni $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ e $H(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ sono definite come prima; se all'iterazione successiva si considera il valore $\boldsymbol{\theta}^{(t+1)}$ che massimizza $Q^*(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ in modo che

$$Q^*(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \geq Q^*(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$$

per ogni $\boldsymbol{\theta}$, allora ad ogni iterazione $\log p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$ aumenta e la sequenza delle stime dei parametri convergerà alla moda di $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$. E' evidente che quando la a priori $p(\boldsymbol{\theta})$ è una funzione costante sullo spazio parametrico Θ, quest'algoritmo si riduce alla versione dell'EM per la massimizzazione della verosimiglianza. Se la a priori non è costante il passo M cambierà, richiedendo la massimizzazione di $Q^*(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ invece che di $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$. Il passo E è invece lo stesso dato che non dipende dalla a priori, dipende soltanto da un valore fissato per $\boldsymbol{\theta}$.

C.3 L'algoritmo Monte Carlo EM, EM stocastico ed EM-tipo

Poiché la funzione da massimizzare

$$Q^*(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \int \log p(\boldsymbol{\theta} \mid \boldsymbol{O}_{com}) p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{O}_{mis}$$

non è sempre determinabile, cioè il passo E dell'algoritmo EM non ammette una soluzione in forma chiusa, si procedere tramite metodi Monte Carlo (Wei e Tanner, 1990); un algoritmo dove il passo E è eseguito mediante metodi Monte Carlo è detto algoritmo MCEM (*Monte Carlo Expectation Maximization*). La t-esima iterazione dell'algoritmo MCEM è data dai seguenti passi

- al passo MCE (*Monte Carlo Expectation*) vengono estratti M valori $\boldsymbol{o}_{mis}^{1(t)}, \ldots, \boldsymbol{o}_{mis}^{M(t)}$ dalla distribuzione predittiva $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{(t)})$. Poi si approssima la funzione Q, per $m = 1, \ldots, M$, con

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \frac{1}{M} \sum_{m=1}^{M} \log p(\boldsymbol{\theta} \mid \boldsymbol{o}_{mis}^{m(t)}, \boldsymbol{O}_{obs}).$$

- al passo M (*Maximization*) si massimizza la funzione $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ rispetto a $\boldsymbol{\theta}$ e si ottiene $\boldsymbol{\theta}^{(t+1)}$.

La scelta di M può essere fatta in vari modi, Wei e Tanner suggeriscono di non mantenere M fisso, ma usare piccoli valori alle prime iterazioni ed aumentare il valore quando l'algoritmo si avvicina alla convergenza.

Da sottolineare che l'algoritmo MCEM, nel caso in cui si consideri $M = 1$, coincide con l'algoritmo EM stocastico (SEM, *Stochastic Expectation Maximization*) che Broniatowski *et al.* (1983) e Celeux e Diebolt (1985, 1986a, 1986b) definiscono quale versione modificata dell'algoritmo EM per determinare le stime di massima verosimiglianza per modelli mistura.

Si noti anche che quando $M = 1$ e il valore di $\boldsymbol{O}_{mis}^{1t}$ non è campionato ma bensì è sostituito da una suo indice sintetico, quale la media e la moda, si ottiene il cosiddetto algoritmo EM-tipo. Se si usa la media quale indice e la densità a posteriori dei dati completi per $\boldsymbol{\theta}$ è lineare su $\boldsymbol{O}_{mis}$, allora l'algoritmo MCEM equivale al quello EM.

C.4 L'algoritmo poor man's data augmentation, data augmentation e chained data augmentation

Si supponga che la sequenza $\boldsymbol{\theta}^{(t)}$ nell'algoritmo MCEM converga ad un valore $\boldsymbol{\theta}^*$; allora mediante un algoritmo non iterativo, che prende il nome di *poor man's data augmentation*, si può ottenere un'approssimazione della distribuzione a posteriori e non soltanto la sua moda, tramite i seguenti due passi:

- si estraggono M valori, $\boldsymbol{o}_{mis}^m$ per $m = 1, \ldots, M$, da $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^*)$;

- si determina l'approssimazione della a posteriori come

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) = \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{\theta} \mid \boldsymbol{o}_{mis}^m, \boldsymbol{O}_{obs}).$$

Da questo deriva un algoritmo più generale chiamato *data augmentation* (Tanner e Wong 1987, Wei e Tanner 1990 e Tanner 1993); questo, analogamente all'algoritmo EM, si basa sulla semplificazione analitica di trattare su dati completi, anziché parziali, per l'analisi della distribuzione a posteriori dei parametri, ma al contrario dell'algoritmo EM il suo obiettivo è quello di ottenere l'intera distribuzione a posteriori non solo il suo massimo. L'idea di fondo dell'algoritmo *data augmentation* è quella di aumentare i dati osservati $\boldsymbol{O}_{obs}$ con quelli non osservati $\boldsymbol{O}_{mis}$ in modo da poter campionare dalla distribuzione a posteriori aumentata $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{com})$. Per ottenere poi la distribuzione a posteriori dei parametri $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$, si generano valori multipli, con una procedura di imputazione multipla (MI, *Multiple Imputation*) di $\boldsymbol{O}_{mis}$ dalla distribuzione predittiva $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs})$ e poi si determina la

media di $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{com})$ sui valori imputati. Poiché $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs})$ dipende da $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$, si ottiene un algoritmo iterativo per calcolare $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$.

L'algoritmo *data augmentation* si basa su due identità:
- l'identità a posteriori

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}, \boldsymbol{O}_{mis}) p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}) d\boldsymbol{O}_{mis}$$

dove $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$ denota la densità a posteriori dei parametri $\boldsymbol{\theta}$ date le osservazioni $\boldsymbol{O}_{obs}$, $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs})$ la densità predittiva dei dati mancanti $\boldsymbol{O}_{mis}$ dati $\boldsymbol{O}_{obs}$, e $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{com})$ indica la densità condizionale di $\boldsymbol{\theta}$ dati $\boldsymbol{O}_{com}$ (la a posteriori aumentata) e
- l'identità predittiva

$$p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}) = \int p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \phi) p(\phi \mid \boldsymbol{O}_{obs}) d\phi$$

dove $p(\boldsymbol{O}_{mis} \mid \phi, \boldsymbol{O}_{obs})$ indica la distribuzione predittiva condizionale.

Sostituendo l'identità predittiva nell'identità a posteriori e scambiando l'ordine di integrazione, $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$ soddisfa la seguente equazione integrale

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) = \int \int p(\boldsymbol{\theta} \mid \boldsymbol{O}_{com}) p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \phi) d\boldsymbol{O}_{mis} p(\phi \mid \boldsymbol{O}_{obs}) d\phi. \quad \text{(C.2)}$$

Per risolvere tale equazione può essere usata una procedura di sostituzioni successive: iniziando con una approssimazione iniziale $p^{(0)}(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$, si determina al passo generico t-esimo

$$p^{(t+1)}(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) = \int \int p(\boldsymbol{\theta} \mid \boldsymbol{O}_{com}) p(\boldsymbol{O}_{mis} \mid \phi, \boldsymbol{O}_{obs}) d\boldsymbol{O}_{mis} p^{(t)}(\phi \mid \boldsymbol{O}_{obs}) d\phi.$$

Spesso è necessario ricorrere a metodi di simulazione Monte Carlo per operare l'integrazione rispetto ad $\boldsymbol{O}_{mis}$ nell'espressione (C.2). In particolare, applicando l'integrazione Monte Carlo all'identità a posteriori, ad ogni iterazione viene generato un campione $\boldsymbol{o}_{mis}^{1(t)}, \ldots, \boldsymbol{o}_{mis}^{M(t)}$ dall'approssimazione corrente della distribuzione predittiva $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs})$, poi viene aggiornata l'approssimazione corrente della a posteriori $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$ come mistura di a posteriori aumentate di $\boldsymbol{\theta}$ dati gli $\boldsymbol{o}_{mis}^{m(t)}$, $m = 1, \ldots, M$, imputati

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) = \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}, \boldsymbol{o}_{mis}^{m(t)}).$$

L'algoritmo *data augmentation* alla t-esima iterazione diviene

- al passo I (*Imputation*), per $m = 1, \ldots, M$ si estrae $\boldsymbol{\theta}^{m(t)}$ dall'approssimazione corrente $p^{(t)}(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs})$ e $\boldsymbol{o}_{mis}^{m(t)}$ dall'approssimazione corrente della predittiva condizionata ai valori dei parametri prima campionati $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta}^{m(t)})$ (si noti che ciò equivalente a estrarre $\boldsymbol{o}_{mis}^{m(t)}$ dalla distribuzione predittiva);

- al passo P (*Posterior*) si aggiorna l'approssimazione corrente della distribuzione a posteriori considerando la mistura delle a posteriori aumentate ottenuta considerando gli M valori di $\boldsymbol{O}_{mis}$ estratti al passo precedente

$$p^{(t+1)}(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) = \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}, \boldsymbol{o}_{mis}^{m(t)}).$$

Tale algoritmo iterativo, può essere visto come la combinazione dell'algoritmo EM e quello di imputazione multipla MI dove il passo *expectation* dell'algoritmo EM è sostituito dal passo di imputazione multipla per i dati mancanti ed il passo *maximization* dall'imputazione multipla dei valori per il vettore dei parametri.

Nel caso in cui $m = 1$ allora l'algoritmo si riduce all'applicazione iterativa dei seguenti passi:

- dato $\boldsymbol{O}_{mis}$, si estrae $\boldsymbol{\theta}$ da $p(\boldsymbol{\theta} \mid \boldsymbol{O}_{mis}, \boldsymbol{O}_{obs})$ considerando che

$$p(\boldsymbol{\theta} \mid \boldsymbol{O}_{obs}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{O}_{mis}, \boldsymbol{O}_{obs}) p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}) d\boldsymbol{O}_{mis};$$

- dato $\boldsymbol{\theta}$, si estrae $\boldsymbol{O}_{mis}$ da $p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \boldsymbol{\theta})$ considerando che

$$p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}) = \int p(\boldsymbol{O}_{mis} \mid \boldsymbol{O}_{obs}, \phi) p(\phi \mid \boldsymbol{O}_{obs}) d\phi.$$

L'algoritmo, in questa formulazione, prende il nome di *chained data augmentation*. Esso possiede il pregio di poter ricondurre il problema di dati mancanti ad un problema risolvibile tramite *Gibbs sampler*, infatti esso non è altro che la formulazione dell'algoritmo nel caso in cui si considerino due vettori di parametri, $\boldsymbol{\theta}$ e $\boldsymbol{O}_{mis}$.

Appendice D

Gibbs sampler, adaptive rejection sampling e test di convergenza

D.1 Gibbs sampler

L'algoritmo *Gibbs sampler* è divenuto noto con l'articolo di Geman e Geman (1984) ma il metodo ha i suoi fondamenti già nell'articolo di Metropolis *et al.* (1953), con ulteriori sviluppi in Hastings (1970); tuttavia solo con l'articolo di Gelfand e Smith (1990) ha trovato largo consenso in statistica applicata.

Il *Gibbs sampler* è un algoritmo iterativo che permette di campionare da una distribuzione marginale in modo indiretto (senza cioè doverla determinare), campionando dalle distribuzioni condizionali di ogni variabile date le altre (distribuzioni condizionali complete o *full conditional*) aggiornando i valori cui condizionarsi con quelli campionati precedentemente. Sfruttando le proprietà delle catene di Markov, le distribuzioni empiriche, cioè le distribuzioni dei valori campionati, convergono alle distribuzioni marginali di interesse.

Si considerino le variabili $X_1, \ldots, X_n$, si supponga, inoltre, di conoscere la distribuzione condizionale completa di ogni variabile X_i, per $i = 1, \ldots, n$, cioè $p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ e di essere interessati alla distribuzione marginale di ogni variabile X_i o a qualche loro valore di sintesi.

Quando le distribuzioni marginali non sono né determinabili né approssimabili direttamente, sfruttando la conoscenza delle distribuzioni condizionali complete, tramite l'algoritmo *Gibbs sampler*, si possono generare dei valori (sequenza *Gibbs*) da cui è possibile ricavare informazioni sulle distribuzioni marginali. Ad ogni generica iterazione $j = 1, \ldots, J + k$ dell'algoritmo si generano i valori

$x_1^{(j)}$ da $p(x_1 \mid X_2 = x_2^{(j-1)}, \ldots, X_n = x_n^{(j-1)})$
$x_2^{(j)}$ da $p(x_2 \mid X_1 = x_1^{(j)}, X_3 = x_3^{(j-1)}, \ldots, X_n = x_n^{(j-1)})$
$\ldots$
$x_n^{(j)}$ da $p(x_n \mid X_1 = x_1^{(j)}, \ldots, X_{n-1} = x_{n-1}^{(j)})$.

Un modo di trarre informazioni da ogni sequenza i-esima ottenuta è quello di fissare un valore k e trattare i termini della sequenza campionata $x_i^{(j)}$ per $j \geq k$ come un campione di dimensione J estratto dalla marginale $p(x_i)$. Praticamente sono scartati dal campione i primi k valori campionati (*burn-in*) prima del raggiungimento della convergenza.

La convergenza può essere controllata applicando strumenti diagnostici alla sequenza ottenuta dall'algoritmo, per una rassegna si veda Cowles e Carlin (1996) e Brooks e Roberts (1999). In questo lavoro è stato utilizzato il test di Geweke (1992) (descritto nel paragrafo D.3); il test di Gelman e Rubin (1992a, 1992b), anch'esso nato per il controllo della convergenza dell'algoritmo, quindi considerato in tali rassegne, è descritto nel paragrafo 3.2.1, ma il suo utilizzo, nel presente lavoro, è relativo ad un altro contesto.

La media di X_i, può essere approssimata con la media campionaria determinata sulla sequenza $x_i^{(1)}, \ldots, x_i^{(J)}$ considerando che

$$\lim_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} x_i^{(j)} = \int x_i p(x_i) dx_i = \mathrm{E}(X_i).$$

D.2 Rejection sampling e adaptive rejection sampling

Quando la distribuzione condizionale completa non è in forma standard, è necessario ricorrere ad algoritmi che permettono di campionare da una distribuzione in forma standard che ne sia una buona approssimazione. Un metodo è quello del *rejection sampling*; esso richiede la determinazione di una funzione $h(x)$ che domina la distribuzione condizionale $g(x)$ da cui campionare, in modo che $h(x) \geq g(x)$ per ogni x. Il campionamento è fatto da una distribuzione proporzionale ad $h(x)$ ed il valore campionato è soggetto ad un test di accettazione-rifiuto; la procedura si arresta quando viene accettato un valore. Il test è fatto nel seguente modo: si accetta il valore campionato con probabilità $g(x)/h(x)$; nella pratica si accetta se questo è non minore di un valore estratto da una distribuzione uniforme tra zero ed uno. Spesso sono necessari un gran numero di test accettazione-rifiuto prima di ottenere un valore valido, per ridurre il numero di volte in cui si rifiuta tale valore,

è necessario scegliere un funzione $h(x)$ vicina alla funzione $g(x)$, da cui sia però facile campionare.

Nel caso in cui si debba campionare da delle distribuzione condizionali complete log-concave, si può sfruttare tale proprietà per semplificare l'algoritmo di campionamento e adottare il cosiddetto *adaptive rejection sampling* (ARS), si veda Gilks e Wild (1992). In particolare si semplifica la scelta della funzione che domina la distribuzione condizionale completa $g(x)$, determinandola come spezzata (*rejection envelope*) costruita sul logaritmo della distribuzione. Fissati degli intervalli per la variabile di interesse, la spezzata è costruita con segmenti definiti dall'intersezione delle rette tangenti nell'estremo superiore di ogni intervallo (o dalle rette secanti negli estremi degli intervalli fissati). In questo modo si ottiene, per ogni intervallo, la semplice approssimazione della curva con distribuzioni esponenziali, da cui si può facilmente campionare. Il valore estratto sarà poi sottoposto ad un test di accettazione-rifiuto, utilizzando la funzione *squeezing* costituita dagli archi che congiungono i punti di tangenza. Il logaritmo della distribuzione da cui campionare sarà compreso tra queste due funzioni che con il procedere dell'algoritmo convergono ad essa. Infatti, ogni volta che un valore campionato viene scartato, esso viene considerato come estremo di un ulteriore intervallo, in tal modo le funzioni si accostano sempre più alla curva (in questo senso di parla di metodo adattivo).

Si consideri una distribuzione condizionale completa $g(x)$ continua e differenziabile ovunque e la trasformata $h(x) = \log g(x)$ concava ovunque. Si valuta la funzione $h(x)$ e $h'(x)$ in k valori di ascissa iniziali x definiti da $T_k = \{x_i; i = 1, \ldots, n\}$. La *rejection envelope* su T_k viene definita come $\exp u_k(x)$, dove $u_k(x)$ rappresenta la spezzata (*upper hull*) costituita da segmenti determinati come intersezione delle tangenti di $h(x)$ nei punti di ascissa definiti da T_k. Le tangenti nei punti x_j e x_{j+1}, per $j = 1, \ldots, k-1$, si intersecano nei punti

$$z_j = \frac{h(x_{j-1}) - h(x_j) - x_{j+1}h'(x_{j+1}) + x_j h'(x_j)}{h'(x_j) - h'(x_{j+1})}.$$

Per i valori di x compresi tra gli z_j si definisce l'*upper hull*

$$u_k(x) = h(x_j) + (x - x_j)h'(x_j),$$

considerando quale z_0 il limite inferiore del dominio, oppure il valore $-\infty$ se il dominio non è limitato inferiormente; analogamente, per z_k il limite

superiore del dominio, oppure ∞ se esso non è limitato. Si definisce anche l'*upper hull* esponenzializzato e normalizzato come

$$s_k(x) = \exp u_k(x) / \int \exp u_k(x')dx'.$$

La funzione *squeezing* su T_k è invece definita da $\exp l_k(x)$, con $l_k(x)$ che indica la spezzata (*lower hull*) costruita con le corde che uniscono i valori $h(x)$ corrispondenti alle ascisse adiacenti in T_k. Per ogni valore di x compreso tra due ascisse adiacenti si definisce

$$l_k = \frac{(x_{j+1} - x)h(x_j) + (x - x_j)h(x_{j+1})}{x_{j+1} - x_j}$$

per $j = 1, \ldots, k-1$. Per $x < x_1$ e $x > x_k$ si definisce $l_k(x) = -\infty$. La *rejection envelope* e la funzione *squeezing* sono esponenziali a tratti, esse "contengono" la distribuzione $h(x)$, infatti, la concavità di $h(x)$ assicura che $l_k(x) \leq h(x) \leq u_k(x)$ per ogni x. Dopo la costruzione delle due funzioni, mostrata in figura D.1, per campionare un valore dalla distribuzione $g(x)$ si procede iterativamente attraverso tre passi.

- Passo di inizializzazione: si inizializza il vettore delle ascisse T_k, se il dominio è illimitato inferiormente occorre scegliere x_1 in modo che $h'(x_1) > 0$, se il dominio è illimitato superiormente allora x_k è scelto in modo che $h'(x_k) < 0$. Definiti i k valori iniziali, si determinano le funzioni u_k, $s_k(x)$ e $l_k(x)$. Da notare che nel caso il dominio sia illimitato sia superiormente sia inferiormente è necessario considerare quale partenza un numero di punti maggiore di uno.

- Passo di campionamento: viene campionato un valore x^* dalla funzione $s_k(x)$, poi viene campionato un valore w, indipendentemente, da una distribuzione uniforme, con cui procedere al test *squeezing*, se $w \leq \exp(l_k(x^*) - u_k(x^*))$ allora si accetta il valore campionato x^*, altrimenti si valuta $h(x^*)$ e $h'(x^*)$ e si procede al seguente test *rejection*, se $w \leq \exp(h(x^*) - u_k(x^*))$ allora si accetta x^*, altrimenti si rifiuta.

- Passo di aggiornamento: viene incluso x^* in T_k per creare T_{k+1}, si riordinano in modo crescente i suoi $k+1$ valori e si costruiscono le funzioni $u_{k+1}(x)$, $s_{k+1}(x)$ e $l_{k+1}(x)$. Si torna al passo di campionamento finché non sia stato accettato un valore.

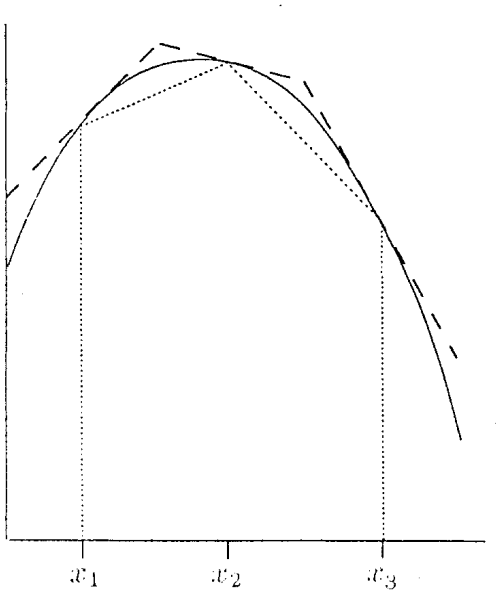


Figura D.1: *Upper* (- - -) e *lower* (. . .) *hull*

D.3 Test di convergenza di Geweke

Geweke (1992) adotta un metodo nato per l'analisi spettrale delle serie storiche, la sequenza *Gibbs* $x_1, \ldots, x_n$ è vista, appunto, come una serie storica. Tale metodo è adatto quando l'interesse sia nella media di una qualche trasformata $g(x)$ della variabile campionata e quando esista la funzione di densità spettrale $S_g(w)$ per tale serie storica senza discontinuità alla frequenza 0. La varianza asintotica per lo stimatore dato dalla media della trasformata della sequenza, è pari a $S_g(0)$. La diagnostica per la convergenza di Geweke su una sequenza di n campioni è determinata come la differenza tra la media della trasformata di una certa porzione di valori campionati inizialmente n_A e la media di una certa porzione finale n_B, diviso per l'errore standard asintotico della differenza, determinato dalla stima della densità spettrale per le due porzioni della sequenza analizzata; Geweke suggerisce di utilizzare $n_A = 0.1n$ e $n_B = 0.5n$. Se l'intera catena è stazionaria la media delle due porzioni di sequenza sarà simile. Se n_A/n e n_B/n sono fissati e $n_A + n_B < n$, allora per il teorema del limite centrale, la distribuzione della diagnostica è asintoticamente quella di una variabile normale standardizzata.

Appendice E

Materiale relativo all'atlante di mortalità toscano

Nella figura E.1 è mostrata la suddivisione amministrativa, a livello comunale, della regione Toscana; ogni comune è identificato tramite il proprio codice ISTAT, la codifica considerata è quell'antecedente la formazione della provincia di Prato.

Il codice ISTAT rappresenta la chiave di lettura della tabella E.1, infatti, in essa, ad ogni codice viene associato il nome del comune, la provincia di appartenenza e sono inoltre elencati (identificati dal loro numero progressivo) i suoi comuni adiacenti. La prima parte della tabella è dedicata ai 287 comuni appartenenti alle province della Toscana: Massa Carrara (MS), Lucca (LU), Pistoia (PT), Firenze (FI), Livorno (LI), Pisa (PI), Arezzo (AR), Siena (SI) e Grosseto (GR); i comuni che si trovano al confine della Toscana, e che quindi confinano con i comuni delle regioni Liguria, Emilia Romagna, Marche, Umbria e Lazio, sono stati evidenziati scrivendo il loro nome con un carattere diverso. La seconda parte della tabella è invece dedicata ai 69 comuni, confinanti con la Toscana, delle province di La Spezia (SP), Parma (PR), Reggio Emilia (RE), Modena (MO), Bologna (BO), Ravenna (RA), Forlì (FO), Pesaro (PS), Perugia (PG), Terni (TR) e Viterbo (VT).

Le adiacenze tra comuni sono definite in base ai loro confini geografici: due comuni sono considerati adiacenti se condividono parte del loro confine. Per i comuni che si trovano nelle isole, le adiacenze sono definite in base all'esistenza di un collegamento, tramite traghetti, agli altri comuni.

Nella tabella E.2 sono riportate le cause di morte analizzate nell'atlante di mortalità, con la relativa codifica ICDVIII e ICDIX.

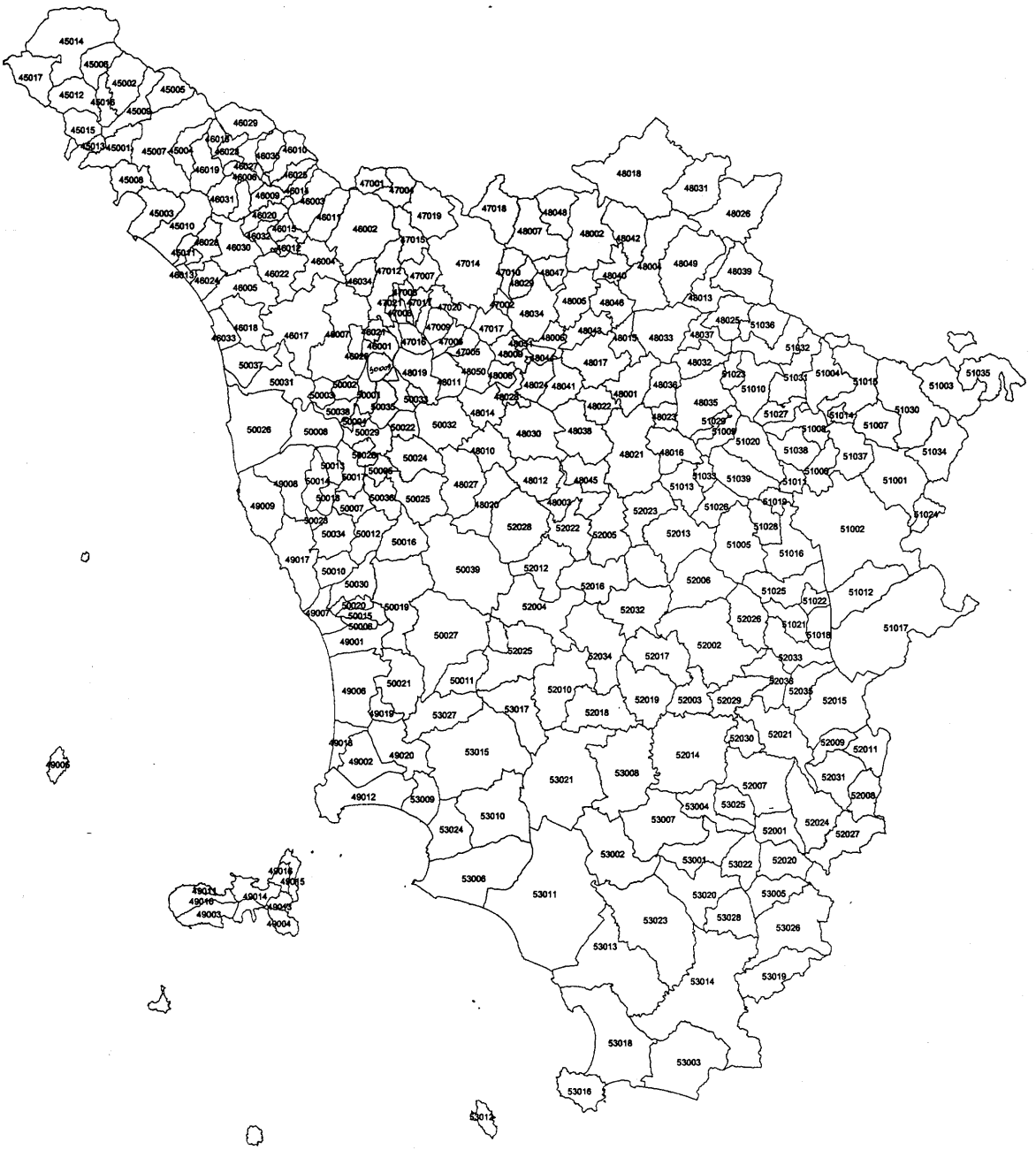


Figura E.1: Mappa a livello comunale della Toscana con codifica ISTAT

Tabella E.1: Comuni della Toscana e confinanti, codici ISTAT e adiacenze

num	cod.	pr.	nome del comune	elenco dei comuni adiacenti
1	45001	MS	**Aulla**	7,8,9,13,288,294,295
2	45002	MS	**Bagnone**	6,9,16,301,302
3	45003	MS	**Carrara**	7,8,10,290,292,295
4	45004	MS	Casola in Lunigiana	7,33,36
5	45005	MS	**Comano**	7,9,302,303,305
6	45006	MS	**Filattiera**	2,12,14,16,301
7	45007	MS	**Fivizzano**	1,3,4,5,8,9,10,33,36,46,303
8	45008	MS	**Fosdinovo**	1,3,7,290,295
9	45009	MS	**Licciana Nardi**	1,2,5,7,13,15,16,302
10	45010	MS	Massa	3,7,11,36,45,47,48
11	45011	MS	Montignoso	10,30,41,45
12	45012	MS	**Mulazzo**	6,14,15,16,17,289,293
13	45013	MS	**Podenzana**	1,9,15,288,289,291
14	45014	MS	**Pontremoli**	6,12,17,298,299,300,301
15	45015	MS	**Tresana**	9,12,13,16,288,289
16	45016	MS	Villafranca in Lunig.	2,6,9,12,15
17	45017	MS	**Zeri**	12,14,293,296,297,298
18	46001	LU	Altopascio	24,38,43,74,93,154
19	46002	LU	**Bagni di Lucca**	21,28,51,53,56,64,67,308,310
20	46003	LU	**Barga**	28,31,32,37,310
21	46004	LU	Borgo a Mozzano	19,24,28,29,32,34,39,51
22	46005	LU	Camaiore	34,35,39,41,47,50
23	46006	LU	Camporgiano	25,26,36,40,42,44,48
24	46007	LU	Capannori	18,21,34,38,43,51,64,146, 147,148,176
25	46008	LU	Careggine	23,26,37,47,48
26	46009	LU	Castelnuovo di Garf.	23,25,31,32,37,42
27	46010	LU	**Castiglione Garf.**	42,52,306,309,310
28	46011	LU	**Coreglia Antelmin.**	19,20,21,32,310
29	46012	LU	Fabbriche di Vallico	21,32,39,49
30	46013	LU	Forte dei Marmi	11,41,45
31	46014	LU	**Fosciandora**	20,26,32,42,310
32	46015	LU	Gallicano	20,21,26,28,29,31,37,49
33	46016	LU	Giuncugnano	4,7,36,40,46
34	46017	LU	Lucca	21,22,24,35,39,176,182
35	46018	LU	Massarosa	22,34,50,182
36	46019	LU	Minucciano	4,7,10,23,33,40,48
37	46020	LU	Molazzana	20,25,26,32,47,49
38	46021	LU	Montecarlo	18,24,43,64,74
39	46022	LU	Pescaglia	21,22,29,34,47,49
40	46023	LU	Piazza al Serchio	23,33,36,44,46
41	46024	LU	Pietrasanta	11,22,30,45,47
42	46025	LU	**Pieve Fosciana**	23,26,26,27,31,44,52,310
43	46026	LU	Porcari	18,24,38
44	46027	LU	S. Romano in Garfagn.	23,40,42,46,52
45	46028	LU	Seravezza	10,11,30,41,47

num	cod.	pr.	nome del comune	elenco dei comuni adiacenti
46	46029	LU	**Sillano**	7,33,40,44,52,303,304,306
47	46030	LU	Stazzema	10,22,25,37,39,41,45,48,49
48	46031	LU	Vagli di Sotto	10,23,25,36,47
49	46032	LU	Vergemoli	29,32,37,39,47
50	46033	LU	Viareggio	22,35,182
51	46034	LU	Villa Basilica	19,21,24,64
52	46035	LU	**Villa Collemandina**	27,42,44,46,306
53	47001	PT	**Abetone**	19,56,308
54	47002	PT	Agliana	62,66,69,108
55	47003	PT	Buggiano	60,64,68,73,74
56	47004	PT	**Cutigliano**	19,53,67,71,307,308
57	47005	PT	Lamporecchio	58,69,72,85,124
58	47006	PT	Larciano	57,61,68,72,85,93
59	47007	PT	Marliana	60,63,64,66,67,72
60	47008	PT	Massa e Cozzile	55,59,63,64,68
61	47009	PT	Monsummano Terme	58,65,68,72
62	47010	PT	Montale	54,66,81,103,108
63	47011	PT	Montecatini Terme	59,60,65,68,72
64	47012	PT	Pescia	19,24,38,51,55,59,60,67,73,74
65	47013	PT	Pieve a Nievole	61,63,68,72
66	47014	PT	**Pistoia**	54,59,62,67,69,70,71,72,81, 315,316,319
67	47015	PT	Piteglio	19,56,59,64,66,71
68	47016	PT	Ponte Buggianese	55,58,60,61,63,65,74,93
69	47017	PT	Quarrata	54,57,66,72,83,108,124
70	47018	PT	**Sambuca Pist.**	66,81,311,313,315
71	47019	PT	**S. Marcello Pist.**	56,66,67,307,316
72	47020	PT	Serravalle Pistoiese	57,58,59,61,63,65,66,69
73	47021	PT	Uzzano	55,64,74
74	47022	PT	Chiesina Uzzanese	18,38,55,64,68,73,93
75	48001	FI	Bagno a Ripoli	89,91,95,96,107,110
76	48002	FI	**Barberino di Mug.**	79,81,92,114,116,121,122,314
77	48003	FI	Barberino Val D'Elsa	86,119,228,245,251
78	48004	FI	Borgo San Lorenzo	89,92,100,105,107,114, 116,120,123
79	48005	FI	Calenzano	76,80,108,114,117,120,121
80	48006	FI	Campi Bisenzio	79,91,108,117,118,125
81	48007	FI	**Cantagallo**	62,66,70,76,103,121,122,311
82	48008	FI	Capraia e Limite	83,88,102,124
83	48009	FI	Carmignano	69,82,98,102,108,118,124,125
84	48010	FI	Castelfiorentino	86,88,94,101,104,177
85	48011	FI	Cerreto Guidi	57,58,88,93,124,177
86	48012	FI	Certaldo	77,84,94,104,119,251
87	48013	FI	Dicomano	99,100,107,111,113,123
88	48014	FI	Empoli	82,84,85,102,104,124,177
89	48015	FI	Fiesole	75,78,91,107,117,120
90	48016	FI	Figline Valdarno	95,97,109,193,197,213,217

num	cod.	pr.	nome del comune	elenco dei comuni adiacenti
91	48017	FI	Firenze	75,80,89,96,115,117,118
92	48018	FI	**Firenzuola**	76,78,105,116,312,314,317, 318,320,322
93	48019	FI	Fucecchio	18,58,68,74,85,154,177,178
94	48020	FI	Gambassi Terme	84,86,101,184,251
95	48021	FI	Greve in Chianti	75,90,96,97,110,112,119, 197,228,246
96	48022	FI	Impruneta	75,91,95,112,115
97	48023	FI	Incisa in Val d'Arno	90,95,109,110
98	48024	FI	Lastra a Signa	83,102,104,115,118
99	48025	FI	Londa	87,111,113,216,220
100	48026	FI	**Marradi**	78,87,105,113,123,321, 324,325,328
101	48027	FI	Montaione	84,94,169,170,177,184
102	48028	FI	Montelupo Fiorentino	82,83,88,98,104
103	48029	FI	Montemurlo	62,81,108,121
104	48030	FI	Montespertoli	84,86,88,98,102,112,115,119
105	48031	FI	**Palazzuolo Senio**	78,92,100,321,322
106	48032	FI	Pelago	107,109,110,111,207
107	48033	FI	Pontassieve	75,78,87,89,106,110,111,123
108	48034	FI	Prato	54,62,69,79,80,83,103,121,125
109	48035	FI	Reggello	90,97,106,110,193,194,207,213
110	48036	FI	Rignano sull'Arno	75,95,97,106,107,109
111	48037	FI	Rufina	87,99,106,107,207,216
112	48038	FI	S. Casciano Val Pesa	95,96,104,115,119
113	48039	FI	**San Godenzo**	87,99,100,220,325,326,327
114	48040	FI	San Piero a Sieve	76,78,79,116,120
115	48041	FI	Scandicci	91,96,98,104,112,118
116	48042	FI	Scarperia	76,78,92,114
117	48043	FI	Sesto Fiorentino	79,80,89,91,120
118	48044	FI	Signa	80,83,91,98,115,125
119	48045	FI	Tavernelle Val Pesa	77,86,95,104,112,228
120	48046	FI	Vaglia	78,79,89,114,117
121	48047	FI	Vaiano	76,79,81,103,108
122	48048	FI	**Vernio**	76,81,311,314
123	48049	FI	Vicchio	78,87,100,107
124	48050	FI	Vinci	57,69,82,83,85,88
125	48051	FI	Poggio a Caiano	80,83,108,118
126	49001	LI	Bibbona	131,132,151,160,164,166
127	49002	LI	Campiglia Marittima	137,143,145
128	49003	LI	Campo nell'Elba	129,135,139
129	49004	LI	Capoliveri	128,138,139
130	49005	LI	Capraia Isola	134,137
131	49006	LI	Castagneto Carducci	126,143,144,145,166
132	49007	LI	Cecina	126,142,151,155,160,165,175
133	49008	LI	Collesalvetti	134,142,153,159,168,171
134	49009	LI	Livorno	130,133,142,171
135	49010	LI	Marciana	128,136,139

num	cod.	pr.	nome del comune	elenco dei comuni adiacenti
136	49011	LI	Marciana Marina	135
137	49012	LI	Piombino	127,130,138,139,140,143,145,268
138	49013	LI	Porto Azzurro	129,137,139,140,141
139	49014	LI	Portoferraio	128,129,135,137,138,141
140	49015	LI	Rio Marina	137,138,141
141	49016	LI	Rio nell'Elba	138,139,140
142	49017	LI	Rosignano Marittimo	132,133,134,155,168,179
143	49018	LI	San Vincenzo	127,131,137,145
144	49019	LI	Sassetta	131,145,166
145	49020	LI	Suvereto	127,131,137,143,144,166, 268,274,286
146	50001	PI	Bientina	24,147,149,154,180,183
147	50002	PI	Buti	24,146,148,183
148	50003	PI	Calci	24,147,176,183
149	50004	PI	Calcinaia	146,153,174,180,183
150	50005	PI	Capannoli	162,169,170,173,174,181
151	50006	PI	Casale Marittimo	126,132,160
152	50007	PI	Casciana Terme	157,162,163,179,181
153	50008	PI	Cascina	133,149,158,159,162,171,174, 176,183
154	50009	PI	Castelfranco di Sotto	18,93,146,167,177,178,180
155	50010	PI	Castellina Marittima	132,142,157,175,179
156	50011	PI	Castelnuovo V. Cecina	172,227,248,286
157	50012	PI	Chianni	152,155,161,175,179,181
158	50013	PI	Crespina	153,159,162,163
159	50014	PI	Fauglia	133,153,158,163,168
160	50015	PI	Guardistallo	126,132,151,164,165
161	50016	PI	Lajatico	157,164,170,175,181,184
162	50017	PI	Lari	150,152,153,158,163,173,174,181
163	50018	PI	Lorenzana	152,158,159,162,168,179
164	50019	PI	Montecatini V. Cecina	126,160,161,165,166,172,175,184
165	50020	PI	Montescudaio	132,160,164,175
166	50021	PI	Monteverdi Marittimo	126,131,144,145,164,172,286
167	50022	PI	Montopoli V. d'Arno	154,169,174,177,180
168	50023	PI	Orciano Pisano	133,142,159,163,179
169	50024	PI	Palaia	101,150,167,170,174,177
170	50025	PI	Peccioli	101,150,161,169,181,184
171	50026	PI	Pisa	133,134,153,176
172	50027	PI	Pomarance	156,164,166,184,248,286
173	50028	PI	Ponsacco	150,162,174
174	50029	PI	Pontedera	149,150,153,162,167,169,173,180
175	50030	PI	Riparbella	132,155,157,161,164,165
176	50031	PI	San Giuliano Terme	24,34,148,153,171,182
177	50032	PI	San Miniato	84,85,88,93,101,154,167,169,178
178	50033	PI	Santa Croce sull'Arno	93,154,177,180
179	50034	PI	Santa Luce	142,152,155,157,163,168
180	50035	PI	Santa Maria a Monte	146,149,154,167,174,178

num	cod.	pr.	nome del comune	elenco dei comuni adiacenti
181	50036	PI	Terricciola	150,152,157,161,162,170
182	50037	PI	Vecchiano	34,35,50,176
183	50038	PI	Vicopisano	146,147,148,149,153
184	50039	PI	Volterra	94,101,161,164,170,172, 227,235,251
185	51001	AR	**Anghiari**	186,191,208,214,218,221,338
186	51002	AR	**Arezzo**	185,190,195,196,200,201, 203,206,208,209,221,340,342
187	51003	AR	**Badia Tedalda**	214,218,219,329,331,333,335
188	51004	AR	**Bibbiena**	192,199,211,215,323
189	51005	AR	Bucine	200,209,210,212,229,236,249
190	51006	AR	Capolona	186,192,195,221,222
191	51007	AR	Caprese Michel.	185,198,199,214,221
192	51008	AR	Castel Focogn.	188,190,199,204,211,215,221,222
193	51009	AR	Castelfr. di Sopra	90,109,194,204,213,217,223
194	51010	AR	Castel S.Niccolò	109,193,204,207,211,215,216
195	51011	AR	Cast. Fibocchi	186,190,203,204,222,223
196	51012	AR	Cast. Fiorentino	186,201,202,206
197	51013	AR	Cavriglia	90,95,210,217,236,246
198	51014	AR	Chitignano	191,199,221
199	51015	AR	**Chiusi della V.**	188,191,192,198,214,215, 221,323,329
200	51016	AR	Civit. Val di Ch.	186,189,203,209,212
201	51017	AR	**Cortona**	186,196,202,238,256,258, 337,340,341,344,345
202	51018	AR	Foiano della Ch.	196,201,205,206,256
203	51019	AR	Laterina	186,195,200,212,223
204	51020	AR	Loro Ciuffenna	192,193,194,195,211,222,223
205	51021	AR	Lucignano	202,206,209,249,256
206	51022	AR	Marciano della Ch.	186,196,202,205,209
207	51023	AR	Montemignaio	106,109,111,194,216
208	51024	AR	**Monterchi**	185,186,338,342
209	51025	AR	Monte S. Savino	186,189,200,205,206,249
210	51026	AR	Montevarchi	189,197,212,217,223,236
211	51027	AR	Ortignano Raggiolo	188,192,194,204,215
212	51028	AR	Pergine Valdarno	189,200,210,203,223
213	51029	AR	Pian di Scò	90,109,193
214	51030	AR	**Pieve S. Stefano**	185,187,191,199,218,329
215	51031	AR	**Poppi**	188,192,194,199,211,216,323
216	51032	AR	**Pratovecchio**	99,111,194,207,215,220,323,327
217	51033	AR	S. Giovanni Valdarno	90,193,197,210,223
218	51034	AR	**Sansepolcro**	185,187,214,331,338,343,343
219	51035	AR	**Sestino**	187,330,331,332,333,334,335,336
220	51036	AR	**Stia**	99,113,216,327
221	51037	AR	Subbiano	185,186,190,191,192,198,199
222	51038	AR	Talla	190,192,195,204
223	51039	AR	Terranuova Bracciol.	193,195,203,204,210,212,217
224	52001	SI	Abbadia San Salvatore	230,243,247,263,281,284
225	52002	SI	Asciano	226,229,240,249,252,255,259

num	cod.	pr.	nome del comune	elenco dei comuni adiacenti
226	52003	SI	Buonconvento	225,237,240,242,252
227	52004	SI	Casole d'Elsa	156,184,233,235,239,248,257
228	52005	SI	Castellina in Chianti	77,95,119,229,239,245,246
229	52006	SI	Castelnuovo Berarden.	189,225,228,236,239,246,249,255
230	52007	SI	Castiglione d'Orcia	224,237,244,247,253,263,284
231	52008	SI	**Cetona**	234,250,254,339,347
232	52009	SI	Chianciano Terme	234,238,244,254
233	52010	SI	Chiusdino	227,241,248,257,276,280
234	52011	SI	**Chiusi**	231,232,238,254,337,339
235	52012	SI	Colle di Val d'Elsa	184,227,239,245,251
236	52013	SI	Gaiole in Chianti	189,197,210,229,246
237	52014	SI	Montalcino	226,230,242,252,253,263,266,267
238	52015	SI	**Montepulciano**	201,232,234,244,258,337
239	52016	SI	Monteriggioni	227,228,229,235,245,255,257
240	52017	SI	Monteroni d'Arbia	225,226,242,255,257
241	52018	SI	Monticiano	233,242,257,267,280
242	52019	SI	Murlo	226,237,240,241,257,267
243	52020	SI	**Piancastagnaio**	224,247,250,264,281,355
244	52021	SI	Pienza	230,232,238,247,252,253, 254,258,259
245	52022	SI	Poggibonsi	77,228,235,239,251
246	52023	SI	Radda in Chianti	95,197,228,229,236
247	52024	SI	Radicofani	224,230,243,244,250,254
248	52025	SI	Radicondoli	156,172,227,233,276,286
249	52026	SI	Rapolano Terme	189,205,209,225,229,256,259
250	52027	SI	**S. Casciano Bagni**	231,243,247,254,339,346,347, 348,355
251	52028	SI	San Gimignano	77,86,94,184,235,245
252	52029	SI	San Giovanni d'Asso	225,226,237,244,253,259
253	52030	SI	San Quirico d'Orcia	230,237,244,252
254	52031	SI	Sarteano	231,232,234,244,247,250
255	52032	SI	Siena	225,229,239,240,257
256	52033	SI	Sinalunga	201,202,205,249,258,259
257	52034	SI	Sovicille	227,233,239,240,241,242,255
258	52035	SI	Torrita di Siena	201,238,244,256,259
259	52036	SI	Trequanda	225,244,249,252,256,258
260	53001	GR	Arcidosso	261,263,266,279,281
261	53002	GR	Campagnatico	260,266,267,270,279,280,282
262	53003	GR	**Capalbio**	273,277,353
263	53004	GR	Castel del Piano	224,230,237,260,266,281,284
264	53005	GR	**Castell'Azzara**	243,281,285,287,355
265	53006	GR	Cast. della Pescaia	269,270,283
266	53007	GR	Cinigiano	237,260,261,263,267
267	53008	GR	Civitella Paganico	237,241,242,261,266,280
268	53009	GR	Follonica	137,145,274,283
269	53010	GR	Gavorrano	265,270,274,280,283
270	53011	GR	Grosseto	261,265,269,272,280,282

num	cod.	pr.	nome del comune	elenco dei comuni adiacenti
271	53012	GR	Isola del Giglio	275
272	53013	GR	Magliano in Toscana	270,273,277,282
273	53014	GR	**Manciano**	262,272,277,278,279,282,285, 287,349,351,353
274	53015	GR	Massa Marittima	145,268,269,276,280,283,286
275	53016	GR	Monte Argentario	271,277
276	53017	GR	Montieri	233,248,274,280,286
277	53018	GR	Orbetello	262,272,273,275
278	53019	GR	**Pitigliano**	273,285,350,351,352,356
279	53020	GR	Roccalbegna	260,261,273,281,282,287
280	53021	GR	Roccastrada	233,241,261,267,269,270,274,276
281	53022	GR	Santa Fiora	224,243,260,263,264,279,287
282	53023	GR	Scansano	261,270,272,273,279
283	53024	GR	Scarlino	265,268,269,274
284	53025	GR	Seggiano	224,230,263
285	53026	GR	**Sorano**	264,273,278,287,348,352,354,355
286	53027	GR	Monterotondo Maritt.	145,156,166,172,248,274,276
287	53028	GR	Semproniano	264,273,279,281,285
288	11004	SP	Bolano	1,13,15,291,294
289	11008	SP	Calice al Cornoviglio	12,13,15,291,293
290	11011	SP	Castelnuovo Magra	3,8,292,295
291	11013	SP	Follo	13,288,289
292	11020	SP	Ortonovo	3,290,295
293	11025	SP	Rocchetta di Vara	12,17,289,297
294	11026	SP	S. Stefano di Magra	1,288,295
295	11027	SP	Sarzana	1,3,8,290,292,294
296	11028	SP	Sesta Godano	17,297,298
297	11032	SP	Zignago	17,293,296
298	34001	PR	Albareto	14,17,296,300
299	34004	PR	Berceto	14,300,301
300	34006	PR	Borgo Val di Taro	14,298,299
301	34012	PR	Corniglio	2,6,14,299,302
302	34022	PR	Monchio delle Corti	2,5,9,301,305
303	35019	RE	Collagna	5,7,46,304,305
304	35025	RE	Ligonchio	46,303,306
305	35031	RE	Ramiseto	5,302,303
306	35045	RE	Villa Minozzo	27,46,52,304,309
307	36011	MO	Fanano	56,71,308,316
308	36014	MO	Fiumalbo	19,53,56,307,310
309	36016	MO	Frassinoro	27,306,310
310	36031	MO	Pievepelago	19,20,27,28,31,42,308,309
311	37010	BO	Camugnano	70,81,122,313,314
312	37014	BO	Castel del Rio	92,318,322
313	37015	BO	Castel di Casio	70,311,315,319
314	37022	BO	Castiglion dei Pepoli	76,92,122,311,320
315	37029	BO	Granaglione	66,70,313,319

num	cod.	pr.	nome del comune	elenco dei comuni adiacenti
316	37033	BO	Lizzano in Belvedere	66,71,307,319
317	37040	BO	Monghidoro	92,318,320
318	37041	BO	Monterenzio	92,312,317
319	37049	BO	Porretta Terme	66,313,315,316
320	37051	BO	S. Ben. Val di Sambro	92,314,317
321	39004	RA	Brisighella	100,105,322,324
322	39005	RA	Casola Valseno	92,105,312,321
323	40001	FO	Bagno di Romagna	188,199,215,216,327,329
324	40022	FO	Modigliana	100,321,328
325	40031	FO	Portico S. Benedetto	100,113,326,328
326	40033	FO	Premilcuore	113,325,327
327	40043	FO	Santa Sofia	113,216,220,323,326
328	40048	FO	Tredozio	100,324,325
329	40049	FO	Verghereto	187,199,214,323,333
330	41005	PS	Belforte all'Isauro	219,332,336
331	41006	PS	Borgo Pace	187,218,219,332,334,343
332	41009	PS	Carpegna	219,330,331,334,335,336
333	41011	PS	Casteldelci	187,219,329,335
334	41025	PS	Mercatello sul Met.	219,331,332,340,343
335	41047	PS	Pennabilli	187,219,332,333
336	41047	PS	Piandimeleto	219,330,332
337	54009	PG	Castiglion del Lago	201,234,238,339,344
338	54011	PG	Citerna	185,208,218,340,342,343
339	54012	PG	Città della Pieve	231,234,250,337,346,347
340	54013	PG	Città di Castello	186,201,334,338,342,343,345
341	54025	PG	Lisciano Niccone	201,344,345
342	54032	PG	Monte S. M. Tiberina	186,208,338,340
343	54044	PG	San Giustino	218,331,334,338,340
344	54055	PG	Tuoro sul Trasimeno	201,337,341
345	54056	PG	Umbertide	201,340,341
346	55002	TR	Allerona	250,339,347,348
347	55011	TR	Fabro	231,250,339,346
348	56001	VT	Acquapendente	250,285,346,354,355
349	56012	VT	Canino	273,351,353
350	56026	VT	Farnese	278,351,356
351	56031	VT	Ischia di Castro	273,278,349,350
352	56032	VT	Latera	278,285,354,356
353	56035	VT	Montalto di Castro	262,273,349
354	56040	VT	Onano	285,348,352
355	56044	VT	Proceno	243,250,264,285,348
356	56053	VT	Valentano	278,350,352

Tabella E.2: Cause di morte analizzate nell'atlante di mortalità toscano

Cause di morte	codifica ICDVIII	codifica ICDIX
tutte le cause	1-999	1-999
tutti i tumori	140-239	140-239
tumore app. aer. diger. sup.	140-149, 150, 161	140-149, 150, 161
tumore all'esofago	150	150
tumore allo stomaco	151	151
tumore al colon-retto	152-154	152-154, 159
tumore al fegato	155, 197.8	155
tumore al pancreas	157	157
tumore alla laringe	161	161
tumore al polm.-trachea-bronchi	162	162
tumore alla pleura	163	163
tumore alla mammella	174	174-175
tumore all'utero	180-182	179-182
tumore all'ovaio	183	183
tumore alla prostata	185	185
tumore alla vescica	188	188
linfomi	200-202	200-202
leucemie	204-207	204-208
diabete	250	250
malattie neurologiche	320-389	320-389
malattie app. circolatorio	390-459	390-459
ipertensione arteriosa	400-404	401-405
malattie ischemiche del cuore	410-414	410-414
infarto	410	410
altre malattie cardiache	420-429	420-429
malattie cerebrali	430-438	430-438
malattie respiratorie	460-519	460-519
polmoniti	480-486	480-486
bronc. cron. enfisema e asma	490-493, 518.9	490-496
malattie apparato digerente	520-579	520-579
cirrosi	571	571
malatt. app. urinario	580-599	580-599

Bibliografia

Bartlett, M.S. (1971), Physical nearest-neighbour models and non-linear time series, *Journal of Applied Probability*, **8**, 222–232.

Bernardinelli, L., Clayton, D., Montomoli, C. (1995a), Bayes estimates of disease maps: how important are priors?, *Statistics in Medicine*, **14**, 2411–2431.

Besag, J. (1974), Spatial interaction and statistical analysis of lattice system, *Journal of the Royal Statistical Society B*, **36**, 192–236.

Besag, J. (1975), Statistical analysis of non lattice data, *The Statistician*, **24**, 179–195.

Besag, J. (1986), On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society B*, **48**, 259–302.

Besag, J., Green, P. (1993), Spatial statistics and bayesian computation, *Journal of the Royal Statistical Society B*, **55**, 25–37.

Besag, J., York, J., Mollié, A. (1991), Bayesian image restoration, with application in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics*, **43**, 1, 1–59.

Biggeri, A., Braga, M., Marchi, M. (1995), Empirical bayes interval estimates: an application to geographical epidemiology, *Journal of the Italian Statistical Society*, **2**, 251–267.

Breslow, N.E. (1984), Extra–Poisson variation in log–linear models, *Journal of the Royal Statistical Society C*, **33**, 38–44.

Breslow, N.E. , Clayton, D.G. (1993), Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9–25.

Breslow, N.E. , Day, N.E. (1975), Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data, *Journal Chronic Diseases*, **28**, 289–303.

Broniatowski, M., Celeux, G., Diebolt, J. (1983), Reconnaissance de densités par un algorithme d'apprentissage probabiliste. In *Data Analysis and Informatics*, vol. 3. Amsterdam: North-Holland, 359–374.

Brooks, S.P., Roberts, G.O. (1999), Assessing convergence of Markov Chain Monte Carlo algorithms, accettato per la pubblicazione su *Statistics and Computing*.

Celeux, G., Diebolt, J. (1985), The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, *Computational Statistics Quartely*, **2**, 73–82.

Celeux, G., Diebolt, J. (1986a), The SEM and EM algorithms for mixtures: numerical and statistical aspects, *Proceeding of the* 7^{th} *Franco-Belgium Meeting of Statistics*, Bruxelles: Publication des Facultés Universitaries St. Louis.

Celeux, G., Diebolt, J. (1986b), L'algorithme SEM: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélanges de densités, *Revue de Statistique Appliquée*, **34**, 35–52.

Cislaghi, C., Braga, M., Biggeri, A. (1995a), Analisi della concentrazione spaziale di eventi per mezzo delle superfici di densità, *Epidemiologia & Prevenzione*, **19**, 142–149.

Cislaghi, C., Braga, M., Luppi, G., Tasco, C. (1995b), Un metodo per l'identificazione automatica di aggregati di casi in mappe di eventi sanitari, *Epidemiologia & Prevenzione*, **19**, 150–160.

Clayton, D.G., Kaldor, J. (1987), Empirical bayes estimates of age—standardized relative risks for use in disease mapping, *Biometrics*, **43**, 671–681.

Clark, P.J., Evans, F.C. (1954), Distance to nearest neighbor as a measure of spatial relationships in populations, *Ecology*, **35**, 445–453.

Cliff, A.D., Ord, J.K. (1981), *Spatial processes: models and applications*, Pion, Londra.

Cowles, M.K., Carlin, B.P. (1996), Markov Chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association*, **91**, 883–904.

Cox, D.R., Isham, V. (1980), *Point Processes*, Chapman & Hall, Londra.

Cressie, N. (1991), *Statistics for spatial data*, Wiley, New York.

Dahlhaus, R., Künsch, H.R. (1987), Edge-effect and efficient parameter estimation for stationary random fields, *Biometrika*, **74**, 877–882.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM Algorithm, *Journal of the Royal Statistical Society B*, **39**, 1–38.

Diggle, P.J. (1983), *Statistical analysis of spatial point patterns*, Academic Press, Londra.

Doguwa, S.I., Upton, G.J.G. (1988), On edge corrections for the point-event analogue of the Clark-Evans statistic, *Biometrical Journal*,**30**, 957–963.

Donnelly, K. (1978), Simulation to determine the variance and edge-effect of total nearest neighbour distance. In *Simulation Methods in Archeology*, I.R. Hodder, ed. Cambridge University Press, Cambridge, 91–95.

Efron, B., Morris, C. (1973), Stein's estimation rule and its competitors — an empirical bayes approach, *Journal of the American Statistical Association*, **68**, 341, 117–130.

Gelfand, A.E., Smith, A.F.M. (1990), Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A. (1996), Inference and monitoring convergence, in *Markov Chain Monte Carlo in practice*, Chapman & Hall.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin D.B. (1995), *Bayesian Data Analysis*, Chapman & Hall.

Gelman, A., Rubin, D. R. (1992a), A single series from the Gibbs sampler provides a false sense of security, in *Bayesian Statistics 4*, editori Bernardo J.M., Berger J.O., Dawid A.P.e Smith A.F.M., Oxford: Oxford University Press, 625–631.

Gelman, A., Rubin, D. R. (1992b), Inference fron iterative simulation using multiple sequences (with discussion), *Statistical Science*, **7**, 457–511.

Geman, S., Geman, D. (1984), Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geweke, J.(1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in *Bayesian Statistics 4*, editori Bernardo J.M., Berger J.O., Dawid A.P.e Smith A.F.M., 625–631.

Gilks, W.R., Wild, P. (1992), Adaptive rejection sampling for Gibbs Sampling, *Applied Statistics*, **41**, 2, 337–348.

Gleeson, A.C., McGilchrist, C.A. (1980), Bilateral processes on a rectangular lattice, *Australian Journal of Statistics*, **22**, 197–206.

Ghosh, M. (1992), Constrained bayes estimation with applications, *Journal of the American Statistical Association*, **87**, 533–540.

Green, P.J. (1984), Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives, *Journal of the Royal Statistical Society B*, **46**, 149–192.

Griffith, D.A. (1980), Towards a theory of spatial statistics, *Geographical Analysis*, **12**, 329–334.

Griffith, D.A. (1983), The boundary value problem in spatial statistical analysis, *Journal of Regional Science*, **23**, 377–387.

Griffith, D.A. (1985), An evaluation of correction techniques for boundary effects in spatial analysis: Contemporary Methods, *Geographical Analysis*, **17**, 81–88.

Griffith, D.A. (1988), A reply to R. Martin's "Some comments on correction techniques for boundary effects and missing value techniques", *Geographical Analysis*, **20**, 70–75.

Griffith, D.A. (1989), Correcting comments on "A Reply to R. Martin's 'Some comments on correction techniques for boundary effects and missing value techniques"', *Geographical Analysis*, **21**, 359.

Griffith, D.A., Amrhein C.G. (1983), An evaluation of correction techniques for boundary effects in spatial analysis: traditional methods, *Geographical Analysis*, **15**, 352–360.

Guyon, X. (1982), Parameter estimation for a stationary process on a d-dimensional lattice, *Biometrika*, **69**, 95–105.

Haining, R. (1977), Model specification in stationary random fields, *Geographical Analysis*, **9**, 107–129.

Haining, R. (1978), Estimating spatial interaction models, *Environment and Planning*, **10**, 305–320.

Haining, R. (1990), *Spatial data analysis in the social and environmental sciences*, Cambridge University Press.

Haining, R., Griffith, D.A., Bennett, R. (1984), A statistical approach to the problem of missing spatial data using a first-order Markov model, *Professional Geographer*, **36**, 338–345.

Haining, R., Griffith, D.A., Bennett, R. (1989), Maximum likelihood estimation with missing spatial data and with application to remotely sensed data, *Communications Statistics Theoric Methods*, **18**, 1875–1894.

Hanisch, K.H. (1984), Some remarks on estimators of the distribution function of nearest neighbour distance in stationary spatial point processes, *Mathematische Operationsforschung und Statistik, Series Statistics*, **15**, 409–412.

Hastings, W.K. (1970), Monte Carlo sampling methods using Markov Chains and their applications, *Biometrika*, **57**, 1, 97–109.

James, W., Stein, C. (1961), Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Symposium*, **1**, 361–379. Berkeley, California: University of California Press.

Künsch, H.R. (1987), Intrinsic autoregressions and related models on the two-dimensional lattice, *Biometrika*, **74**, 517–524.

Laird, N.M., Louis, T. (1987), Empirical bayes confidence intervals based on bootstrap samples, *Journal of the American Statistical Association*, **82**, 739–750.

Little, R.J.A., Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, Wiley and Sons.

Louis, T.A. (1984), Estimating a population of parameter values using empirical bayes methods, *Journal of the American Statistical Association*, **79**, 393–398.

Martin, R.J. (1984), Exact maximum-likelihood for incomplete data from a correlated gaussian process, *Communications Statistics Theoric Methods*, **13**, 1275–1288.

Martin, R.J. (1987), Some comments on correction techniques for boundary effects and missing value techniques, *Geographical Analysis*, **19**, 273–282.

McLachlan, G.J., Krishnan, T. (1997), *The EM algorithm and extensions*, Wiley Series in Probability and Statistics.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087–1091.

Mollié, A. (1996), Bayesian mapping of disease, in *Markov Chain Monte Carlo in practice*, Chapman & Hall.

Moran, P.A.P. (1973), A Gaussian Markovian process on a square lattice, *Journal of Applied Probability*, **10**, 54–62.

Morris, C.N. (1983), Parametric empirical bayes inference: theory and applications, *Journal of the American Statistical Association*, **78**, 381, 47–65.

Orchard, T., Woodbury, M. A. (1972), A missing information principle: theory and applications, Proocedings 6th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Berkeley. University of California Press, 697–715.

Ord, J.K. (1975), Estimation methods for models with spatial interaction, *Journal of the American Statistical Society*, **70**, 120–126.

Ripley, B.D. (1976), The second-order analysis of stationary point processes, *Journal of Applied Probability*, **13**, 255–266.

Ripley, B.D. (1979), Tests of Randomness for spatial point patterns, *Journal of the Royal Statistical Society B*, **41**, 368–374.

Ripley, B.D. (1981), *Spatial Statistics*, Wiley & Sons.

Ripley, B.D. (1984), Spatial statistics: developments 1980-83, *International Statistical Review*, **52**, 141–150.

Ripley, B.D. (1988), *Statistical Inference for spatial processes*, Cambridge University Press.

Rubin, D.B. (1976), Inference and missing data, *Biometrika*, **63**, 581–592.

Rubin, D.B. (1987), *Multiple imputation for non response in survey*, Wiley & Sons.

Schafer, J.L. (1997), *Analysis of incomplete multivariate data*, Chapman & Hall.

Tanner, M.A. (1993), *Tools for statistical inference: methods for the exploration of posterior distribution and likelihood function*, Springer Series in Statistics.

Tanner, M.A., Wong, W.H. (1987), The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**, 528–550.

Upton, G.J.G., Fingleton, B. (1985), *Spatial data analysis by Example*, Wiley and Sons.

Wei, G.C.G., Tanner, M.A. (1990), A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, *Journal of the American Statistical Association*, **85**, 699–704.

Whittle, P. (1954), On stationary processes on the plane, *Biometrika*, **41**, 434–449.

Winer, B.J. (1971), *Statistical principles in experimental design*, Second Edition, McGraw Hill, New York.

Wonnacott, R., Wonnacott, T. (1970), *Econometrics*, New York: Wiley.