Università degli Studi di Firenze
Dipartimento di Statistica "G. Parenti"
Scuola di Dottorato di Ricerca in Statistica Applicata XXI ciclo

Modelli Bayesiani gerarchici per il controllo della distorsione ecologica

Laura Grisotto

Tutor	Prof. Annibale Biggeri
Co-tutor	Dott.ssa Emanuela Dreassi Prof. Marco Breschi
Coordinatore	Prof. Guido Ferrari

Alla mia famiglia

Ringraziamenti

Giunta al termine di questo lavoro desidero ringraziare tutte le persone che, in modi diversi, mi sono state vicine e mi hanno sempre incoraggiato in questi tre anni di dottorato.

I maggiori ringraziamenti vanno a chi mi ha seguito durante la realizzazione del lavoro di tesi, e in particolare al relatore, Professor Annibale Biggeri, per avermi pazientemente seguito durante il lavoro, con consigli e insegnamenti che mi hanno aiutato ad intraprendere, ogni volta, le scelte più appropriate. Un ringraziamento anche a Emanuela Dreassi, che mi ha seguito e aiutato soprattutto nelle analisi, e al Professor Marco Breschi, per i suoi preziosi consigli e per avermi spronato ad iniziare il dottorato. Ringrazio anche Gabriele Accetta per il suo aiuto nelle analisi del primo capitolo.

Un ringraziamento particolare a Dolores Catelan, che più di tutti mi ha seguito e sostenuto durante il dottorato.

Ringrazio inoltre gli amici e colleghi del XXI ciclo con i quali ho condiviso questi tre anni di dottorato; i membri del “trio M” con i quali ho condiviso non solo le ore di studio ma anche e soprattutto i piacevoli svaghi; tutti i colleghi dell’Unità operativa di Biostatistica dell’Istituto per lo Studio e la Prevenzione Oncologica che mi hanno supportato quotidianamente, in particolare a Giancarlo che mi deve sopportare anche dopo l’orario di lavoro.

Per ultimi, ma di certo non per importanza, ringrazio la mia famiglia, a cui dedico questo lavoro di tesi: i miei genitori, che con i loro sacrifici hanno reso possibile questo mio traguardo, Sandro, Roberto, Samuela e Roberta che mi sono sempre stati vicini appoggiandomi ed incoraggiandomi nel perseguire i miei obiettivi.

Ringrazio Emanuele, che in questi anni oltre ad avermi sempre “supportato”, mi ha più di tutto “sopportato”. Se ho raggiunto questo traguardo lo devo anche alla sua continua presenza; per avermi fatto capire che potevo farcela, incoraggiandomi sempre a fare del mio meglio.

Sommario

Indice delle tabelle

Indice delle figure

Introduzione

Nella letteratura epidemiologica è stata documentata da molto tempo l'associazione tra salute e deprivazione materiale, misurate sia a livello individuale (Townsend e Davidson, 1982[**]) che aggregato (St Leger, 1995).

Quando lo scopo è fare inferenza sull'effetto di covariate misurate a livello individuale, l'utilizzo di informazioni a livello aggregato può portare a ciò che è noto in letteratura come distorsione ecologica. L'effetto della covariata stimato sui dati aggregati non necessariamente riproduce i risultati dell'analisi di dati individuali, a meno che i soggetti rispetto alla covariata di interesse non siano omogenei entro livello di aggregazione.

Tuttavia, anche gli studi che utilizzano dati individuali possono portare a stime potenzialmente distorte, in quanto non considerano la presenza di un possibile effetto contestuale (Diez-Roux, 1988). L'effetto della povertà sociale ed economica sulle condizioni di salute si esplica infatti anche a livello ecologico, non solo a livello individuale. Ad esempio, la mortalità che osserviamo nelle aree socialmente ed economicamente più degradate è maggiore di quella predetta dalla somma di tutte le condizioni di svantaggio registrate a livello individuale (Riva *et al.*, 2007).

È necessario considerare adeguatamente la struttura gerarchica dei dati nella modellazione statistica.

Dal punto di vista pratico le statistiche correnti sia di mortalità che quelle relative alle variabili socio-economiche, come i censimenti della popolazione, non sono disponibili con dettaglio individuale per motivi di confidenzialità,. È quindi necessario ricorrere a dei metodi che permettano di controllare la distorsione ecologica in tali condizioni.

Le soluzioni possono variare dal definire aggregati di piccola dimensione, in modo tale che la popolazione sia omogenea entro area; al cercare di stimare la distribuzione entro area della variabile di interesse ed eventualmente di altre covariate potenzialmente confondenti, utilizzando informazioni campionarie. I

modelli di analisi che utilizzano appropriatamente informazioni a livello aggregato sulla frequenza di malattia con informazioni raccolte a livello individuale sui fattori di rischio (Prentice e Sheppard, 1995, Wakefield e Salway, 2001, 2008) si fermano alla stima dell'effetto a livello individuale perdendo di vista la struttura gerarchica dei dati e di conseguenza le possibili distorsioni dovute alla presenza di effetti contestuali.

La mancata disponibilità di dati individuali è fonte di un'ulteriore distorsione nelle regressioni ecologiche. Il problema, noto come *mutual standardization bias,* è stato sollevato da Rosenbaum e Rubin (1984) ma è stato trascurato nella letteratura epidemiologica. È usuale nei modelli di regressione ecologica utilizzare come variabile di risposta indicatori di mortalità o frequenza di malattia standardizzati per età e come variabili esplicative indicatori socio-economici non aggiustati per lo stesso confondente. Collegandosi a quanto discusso prima ci riportiamo ancora al problema dell'integrazione delle inferenze a livello aggregato con informazioni a livello individuale che permettano di stimare la distribuzione congiunta ad esempio della deprivazione materiale e dell'età.

Gli studi epidemiologici osservazionali hanno una fondamentale rilevanza nella ricerca epidemiologica, e risolvere i problemi appena descritti ha una grande rilevanza pratica. Si pensi ad esempio al monitoraggio delle diseguaglianze sociali nella salute attraverso l'uso di indicatori del livello socio-economico (Subramanian *et al.,* 2006); oppure alla definizione di formule per il riparto del Fondo Sanitario Nazionale che considerino non solo la struttura demografica della popolazione ma anche il livello di deprivazione materiale (Carr-Hill *et al.,* 1997).

Il presente lavoro è articolato in tre parti distinte, ciascuna delle quali propone soluzioni ad alcuni dei problemi emersi nella letteratura epidemiologica sull'associazione tra salute e deprivazione materiale, e sopra descritti.

Nel primo capitolo affrontiamo il problema del *mutual standardization bias* con l'intento di valutare l'adeguatezza dell'utilizzo di un indice di deprivazione materiale come predittore del bisogno sanitario quando non siamo in grado di standardizzare tale indice per gli stessi confondenti per i quali è stato standardizzato l'esito di interesse, tenendo conto nel contempo dell'effetto contestuale dello stesso.

A questo scopo abbiamo utilizzato i dati dello Studio Longitudinale Toscano (SLTo) in particolare della coorte censuaria del 1991.

Nel secondo capitolo vogliamo quantificare l'associazione tra deprivazione materiale e incidenza per tumore del polmone in Toscana usando informazioni socio-economiche disponibili a livello di sezione di censimento. A questo scopo ci siamo avvalsi dei dati individuali raccolti dal Registro Tumori Toscano, per l'anno 2004 e per i fattori socio-economici, i dati aggregati a livello di sezione del Censimento ISTAT 2001. L'uso della sezione di censimento è un esempio di disegno volto a controllare la distorsione ecologica mediante la definizione di aggregati di piccole dimensioni ed omogenei rispetto alla covariata di interesse.

Infine, nel terzo lavoro, proponiamo un modello Bayesiano gerarchico che integra l'informazione a livello aggregato sulla variabile di risposta con l'informazione campionaria individuale sui fattori di rischio (deprivazione materiale) e che diversamente da quanto già presente in letteratura, permette di stimare sia l'effetto individuale che contestuale. Il modello di regressione proposto mira a studiare la relazione tra deprivazione materiale e mortalità generale a livello italiano. A questo scopo sono stati utilizzati i dati relativi ai certificati di causa di morte ISTAT per gli anni 2000-2004, per tutta Italia a livello provinciale, e per quanto riguarda i fattori socio-economici, i dati individuali dell'Indagine Multiscopo ISTAT del 2000.

---

[*] Townsend P, Davidson N (1982). Inequalities in health: the black report. *Penguin Books*. Baltimora.

Capitolo 1

Stima del valore predittivo sul bisogno sanitario di indicatori socio-economici ai fini del loro uso nelle formule allocative

1.1 Introduzione

In Epidemiologia descrittiva quando si eseguono regressioni ecologiche è comune usare come variabile di risposta un indicatore di salute standardizzato per età (considerata come confondente rilevante) perché questo permette una forte riduzione della dimensionalità.

Se la variabile risposta è standardizzata per età, o, più in generale, per un qualsiasi confondente, allora anche le covariate inserite nel modello dovrebbero essere standardizzate pena la stima distorta dei coefficienti di regressione. Tale distorsione nota in letteratura come *mutual standardization bias* è stata discussa da Rosenbaum e Rubin (1984). Quando non si aggiustano per età e le variabili indipendenti, si sta implicitamente assumendo che esse siano costanti nelle diverse classi d'età. Se questo non è vero si genera una distorsione assimilabile formalmente alla distorsione ecologica (Morgenstern, 1998). Un punto importante da tener presente è che per poter stimare correttamente l'effetto dell'esposizione è sufficiente disporre della distribuzione congiunta età-regressore {CX} (C = confondente, X = repressore) non è necessaria la distribuzione congiunta completa {YCX}, risposta-età-regressore (Wakefield, 2007).

Queste considerazioni assumono rilevanza se pensiamo all'esempio concreto di introduzione di indicatori socio-economici nelle formule di allocazione delle risorse

sanitarie (Carr Hill *et al.*, 1994, Ministero della Salute, 1998). In Italia il problema si pone sia a livello della ripartizione del Fondo Sanitario Nazionale che dei Fondi Sanitari Regionali.

In termini molto generali, i criteri di allocazione delle risorse si basano in parte sulla spesa storica e in parte su criteri oggettivi basati sulla quota capitaria determinata con vari meccanismi di pesatura (Stone e Galbraith, 2006). I pesi dovrebbero essere costruiti in modo da riflettere sia la struttura demografica che socio-economica della popolazione, per quella parte che determina un bisogno sanitario assistenziale.

I primi passi sulla introduzione di criteri formali oggettivi si sono compiuti in Inghilterra dopo l'istituzione nel 1948 del *National Health Service* (NHS). I criteri di ripartizione sono mutati nel corso del tempo; si è partiti da una mera considerazione delle caratteristiche demografiche della popolazione (sostanzialmente la percentuale di anziani) per arrivare all'inserimento nella formula di indicatori socio-economici complessi, come l'indice di deprivazione di Jarman (correttivi volti a dare di più in ragione della carenza di risorse proprie) (Jarman, 1983; Carr Hill *et al.*, 1994).

In Italia esistono numerosi studi che evidenziano la relazione tra indicatori socio-economici e indicatori sanitari (Costa *et al.*, 2004) ma solo di recente si è iniziato a discutere del loro inserimento nella formula per il riparto del Fondo Sanitario Nazionale, in sede di Conferenza Stato-Regioni (Convegno: "Criteri per la determinazione dei fabbisogni regionali ai fini del riparto delle risorse economiche per la sanità", Roma, 20 aprile 2006).

Nel presente lavoro ci proponiamo di valutare l'adeguatezza dell'utilizzo di un indicatore socio-economico (ad esempio l'indice di deprivazione materiale) come predittore del bisogno sanitario quando non siamo in grado di standardizzare tale indice per gli stessi confondenti per i quali è stato standardizzato l'esito di interesse. Inoltre, si deve considerare il fatto che l'effetto della deprivazione materiale è sia diretto, cioè prodotto sul singolo individuo dal suo proprio livello di deprivazione, sia mediato dal livello medio di deprivazione dell'area in cui risiede, (effetto contestuale) (Firebaugh, 1978).

Ad esemplificazione si trae vantaggio dello Studio Longitudinale Toscano, in particolare della coorte censuaria 1991 della città di Firenze (Biggeri *et al.*, 2001). La disponibilità dei dati individuali, ci permette di prendere in considerazione tutti i livelli gerarchici, e valutare sia l'effetto individuale sia l'effetto contestuale e confrontare i risultati di differenti modelli in cui l'indice di deprivazione e l'esito sanitario vengano o no standardizzati per lo stesso confondente.

Nel paragrafo 1.2 sono descritti i dati oggetto dell'analisi. Nel paragrafo 1.3 vengono presentati i metodi statistici utilizzati: modelli individuali e modelli su dati aggregati (ecologici). I risultati, la discussione e le conclusioni sono riportate rispettivamente nei paragrafi 1.4, 1.5 e 1.6.

1.2 Materiali

1.2.1 Dati di mortalità

I dati analizzati si riferiscono alla coorte censuaria della città di Firenze, che è parte dello Studio Longitudinale Toscano SLTo (Biggeri *et al*., 2001). Consideriamo i residenti censiti al 31 ottobre 1991 seguiti fino al 31 dicembre 1995 per quanto riguarda lo stato di vita.

Al fine del presente studio si è scelto di analizzare la mortalità generale per tutte le cause nei maschi dai 18 ai 75 anni di età per un totale di 144.001 soggetti (568.713 anni persona) e 4.312 decessi (un tasso grezzo di 7,6 per mille) su 2.752 sezioni di censimento.

1.2.2 Deprivazione

In base alle informazioni individuali contenute nel censimento ISTAT del 1991 per ogni individuo è stato definito un indice di deprivazione come somma di quattro eventi sfavorevoli: bassa istruzione (licenza elementare o inferiore), disoccupazione, abitazione piccola (<25 mq), assenza di bagno nell'appartamento. L'indice vale 0 se all'individuo non corrisponde nessuno degli eventi sfavorevoli,

vale 1 se l'individuo al censimento 1991 presentava solo un evento sfavorevole e 2 se l'individuo presentava due o più eventi.

La tabella 1.1 riporta la distribuzione per indice di deprivazione del numero di individui arruolati, dei decessi osservati e attesi e dei rapporti standardizzati di mortalità (SMR). Il numero di casi attesi è stato calcolato seguendo la standardizzazione interna indiretta e classificando la popolazione in 16 classi d'età (0-4, 5-9,…, 75 e più) (Breslow e Day, 1975). È evidente l'andamento crescente della mortalità rispetto alla condizione di deprivazione dell'individuo; se si considera l'indice come variabile ordinale il rischio relativo (RR) è pari a 1,31 (logRR=0,27) per ogni incremento di livello (Armitage, 1955; Cochran, 1954).

1.3 Metodi

1.3.1 Modelli su dati individuali

Sia Y_{jsa} una variabile bernoulliana che vale 1 se il *j*-esimo individuo che risiede nella *s*-esima sezione di censimento e che appartiene alla *a*-esima classe d'età è deceduto e vale 0 altrimenti.

Considerando che la variabile esplicativa deprivazione materiale è di fatto discreta (0, 1, 2+ condizioni di svantaggio) possiamo considerare i dati elementari come costituiti dalle frequenze dei deceduti (Y_{isa}) per ogni *i*-esima categoria di deprivazione, *s*-esima sezione e *a*-esima classe d'età. Assumiamo che Y_{isa} segua una distribuzione di Poisson con valore atteso , $(E_{isa}\theta_{isa})$, dove θ_{isa} rappresenta il rischio relativo ed E_{isa} è il numero di persone anno di osservazione o il numero di eventi attesi se applichiamo la standardizzazione indiretta interna.

Sono stati specificati tre modelli lineari generalizzati sul logaritmo del rischio relativo:

Modello individuale grezzo

$$\log(\theta_{isa}) = \alpha + \beta^T x_{isa} + \sum \xi_a age_a \qquad (1.1)$$

dove α è l'intercetta, x_{isa} è l'indice di deprivazione materiale individuale, age_a sono variabili fittizie che valgono 1 se l'individuo *i*-esimo appartiene alla classe d'età *a*-esima e 0 altrimenti e β^T rappresenta l'effetto individuale grezzo e ξ_a rappresenta l'effetto della classe d'età *a*-esima ($\xi_{a=1} = 0$ per identificabilità).

Modello contestuale (Firebaugh, 1978)

$$\log(\theta_{isa}) = \alpha + \beta^I x_{isa} + \beta^C \bar{x}_s + \sum \xi_a age_a \qquad (1.2)$$

dove $\bar{x}_s$ è la deprivazione media della *s*-esima sezione di censimento, i coefficienti β^I e β^C sono rispettivamente, l'effetto individuale netto o effetto *within*, e l'effetto contestuale. Tali modelli sono stati ampiamente utilizzati nei settori della ricerca volti a studiare gli effetti di gruppo sul comportamento degli individui (Kreft e De Leeuw, 1998). Questi tipi di analisi prevedono l'introduzione nei modelli di variabili a livello di gruppo, o contestuali, definite come medie della variabile a livello individuale.

Modello di Cronbach (Cronbach e Webb, 1975)

$$\log(\theta_{isa}) = \alpha + \beta^I \left(x_{isa} - \bar{x}_s\right) + \beta^A \left(\bar{x}_s - \bar{x}\right) + \sum \xi_a age_a \qquad (1.3)$$

dove $\bar{x}$ è la deprivazione media della popolazione oggetto dello studio, il coefficiente β^I è l'effetto individuale netto, mentre il coefficiente β^A è l'effetto aggregato (chiamato anche globale o *between*). Nel caso di un modello lineare l'effetto aggregato è esattamente la somma algebrica dell'effetto individuale e dell'effetto contestuale ($\beta^A = \beta^I + \beta^C$). L'indice di deprivazione individuale viene centrato rispetto alla media di sezione e la media di sezione rispetto alla media

generale. L'uso degli scarti dalla media permette di limitare i problemi dovuti alla collinearità tra le variabili esplicative che in questo caso sono misurate sulla stessa unità statistica anche se ad un livello di aggregazione diverso.

1.3.2 Modelli su dati aggregati

In questo tipo di analisi i dati sono rappresentati dalle frequenze per unità gerarchicamente superiori all'individuo. Assumiamo che il numero di casi osservati nella *s*-esima sezione di censimento Y_s segua una distribuzione di Poisson con valore atteso, $(E_s\theta_s)$, dove θ_s rappresenta il rischio relativo della sezione di censimento ed E_s è il relativo numero di persone-anno di osservazione o il numero di casi attesi. Possiamo specificare diversi modelli lineari generalizzati per il logaritmo del rischio relativo.

Modello su dati non standardizzati ed età come confondente :

$$\begin{aligned} E_s &= PY_s \\ \log(\theta_s^g) &= \alpha + \beta^A \overline{x}_s + \xi \overline{age_s} \end{aligned} \qquad (1.4)$$

dove, θ_s^g è il rischi relativo grezzo della *s*-esima sezione di censimento, α è l'intercetta, β^A rappresenta l'effetto aggregato. In questo modello, sia la mortalità che l'indice di deprivazione materiale non sono aggiustate per età. L'età media della sezione di censimento, $\overline{age_s}$, viene inserita nel modello come variabile continua, ξ è l'effetto lineare dell'età e PY_s gli anni-persona di osservazione.

Modello su dati standardizzati per età:

$$\begin{aligned} E_s &= \sum \xi_a^R PY_{sa} \\ \log(\theta^{age}{}_s) &= \alpha + \beta^A \overline{x}_s{}^{age} \end{aligned} \qquad (1.5)$$

dove θ_s^{age} è il rischio relativo aggiustato per età, β^A rappresenta l'effetto aggregato; $\bar{x}_s^{age}$ è l'indice di deprivazione standardizzato per età analogamente usando il metodo indiretto. ξ_a^R sono i tassi specifici per età di riferimento e PY_{sa} gli anni-persona di osservazione.

Modello su risposte standardizzate per età e covariata non standardizzata:

$$\log(\theta_s^{age}) = \alpha + \beta^A \bar{x}_s \qquad (1.6)$$

dove l'indice di deprivazione $\bar{x}_s$ è l'indice di deprivazione medio non standardizzato.

Modello su risposte standardizzate per età, covariata grezza e età come confondente:

$$\log(\theta_s^{age}) = \alpha + \beta^A \bar{x}_s + \xi \overline{age_s} \qquad (1.7)$$

dove $\overline{age_s}$ l'età media della sezione di censimento.

Tutti i modelli sono stati adattati anche utilizzando l'indice di deprivazione diviso per la sua deviazione standard. Questo accorgimento permette di ottenere l'effetto per unità di deviazione standard e quindi di poter direttamente confrontare tra loro i coefficienti individuali e aggregati.

1.4 Risultati

I logaritmi dei rischi relativi, stimati dai modelli individuali sono riportati in tabella 1.2.

La stima dell'effetto individuale della deprivazione materiale si riduce da 0,238 a 0,166 passando dal modello individuale grezzo al modello che considera anche gli

effetti contestuali. Il coefficiente individuale grezzo è dunque una stima distorta dell'effetto della deprivazione materiale a livello individuale in quanto non si tiene conto dell'effetto contestuale veicolato dal valor medio per sezione di censimento.

Nel caso del modello lineare si può dimostrare che:

$$\beta^T = \beta^I + \beta^C \frac{Var(\bar{x})}{Var(x)}.$$

Poiché nei nostri dati:

$Var(x)$ =0,152

$Var(\bar{x})$ =0,015

$\frac{Var(\bar{x})}{Var(x)}$ =0,10

otteniamo un effetto totale: $0{,}166 + 0{,}672 \times 0{,}10 = 0{,}233$, prossimo a quello stimato dal modello loglineare pari a 0,238. Abbiamo un differenza tra le due stime 0,238-0,166=0,072 contro un errore standard misurato a $0{,}04 \times \sqrt{2}$ e nel caso di effetti espressi per unità di deviazione standard la differenza tra le due stime è 0,093-0,065=0,028 con un errore standard $0{,}015 \times \sqrt{2}$.

L'effetto contestuale stimato è più forte di quello individuale ($\hat{\beta}^C$ =0,672 contro $\hat{\beta}^I$ =0,166), anche se si considera l'effetto per unità di deviazione standard ($\hat{\beta}^C$ =0,082 contro $\hat{\beta}^I$ =0,065). L'effetto complessivo della deprivazione (modello di Cronbach) è pari a $\hat{\beta}^A$ =0,838, ed è esattamente la somma dei coefficienti stimati nel modello contestuale, (1.2). L'effetto complessivo per unità di deviazione standard è pari a $\hat{\beta}^A$ =0,102.

In tabella 1.3 sono riportati i risultati dei modelli aggregati. L'analisi sui dati aggregati non riproduce i risultati dell'analisi sui dati individuali, se non per il modello su dati aggregati standardizzati per età (1.5), dove abbiamo usato la *mutual standardization* ($\hat{\beta}^A_{(5)}$ =0,119).

Il modello sui dati aggregati grezzi, (1.4), fornisce anch'esso stime vicine al valore ottenuto sui dati individuali ($\hat{\beta}^{A}_{(4)}$ =0,097). Al contrario, l'effetto stimato dai modelli (1.6) e (1.7) è sovrastimato (rispettivamente $\hat{\beta}^{A}_{(6)}$ =0,143 e $\hat{\beta}^{A}_{(7)}$ =0,140).

1.5 Discussione

Introdurre l'indice di deprivazione in una formula allocativa, che già contenga l'età come predittore, può essere fuorviante perché, se le due variabili sono correlate, finiremmo per usare due volte la stessa informazione. A tal proposito bisogna tener presente che l'indice di deprivazione viene calcolato come somma di diverse condizioni di svantaggio, tra cui la bassa istruzione. Come dimostrano i dati del censimento ISTAT 2001 la frequenza di individui che non hanno conseguito il titolo di studio previsto attualmente per la scuola dell'obbligo è maggiore nelle classi d'età più elevate (14° Censimento generale della popolazione ISTAT 2001).

Ciò significa che la deprivazione materiale calcolata a livello aggregato potrebbe essere inserita nelle formule allocative come predittore del bisogno di cure aggiuntivo oltre l'età, cioè $\hat{Y}^{age}{}_{s} = f(\beta\ \bar{x}_{s}{}^{age})$ purché il parametro β sia stimato secondo il modello (1.5).

Si noti che introdurre l'indice di deprivazione materiale nelle formule allocative implica assumere che sia plausibile un effetto contestuale della variabile, dato che entra sempre come dato aggregato per area geografica.

In questo lavoro emerge chiaramente che vi è un aspetto della povertà sociale ed economica che si manifesta ad un livello ecologico e non più individuale. La mortalità che osserviamo nelle zone o quartieri più degradati è maggiore di quello che ci aspettiamo se sommiamo tutte le condizioni di svantaggio registrate a livello individuale. Tuttavia, ci sono alcune difficoltà nello studio degli effetti contestuali: il livello più appropriato sembra essere molto piccolo (sezione di censimento), poiché livelli di aggregazione maggiori sembrano diluire l'effetto e non bisogna eccedere nell'interpretare in senso causale questa associazione. Per gli scopi di una

analisi a fini allocativi basta infatti identificare dei predittori del bisogno più che i suoi determinanti.

La possibilità di condurre tali valutazioni e quindi di stimare β^A associata a $\bar{x}_s^{age}$ non è sempre realizzabile. A questo proposito ricordiamo che i dati ISTAT disponibili non prevedono, per ragioni di confidenzialità, la distribuzione congiunta età-deprivazione materiale (comunque la si voglia definire) per area sufficientemente piccola (sezione di censimento).

Tuttavia, anche in questo caso, restano da chiarire alcuni problemi.

L'indice potrebbe avere significati diversi in differenti contesti geografici e in differenti istanti temporali (Dreassi *et al.*, 2005, Catelan *et al.*, 2006). È necessario valutare che l'associazione tra l'indice di deprivazione e la mortalità sia costante nelle differenti Regioni. In alcune realtà territoriali l'associazione potrebbe avere segno inverso, non rispecchiando più una condizione di povertà o di condizione sociale svantaggiata in termini di bisogno sanitario. Questo potrebbe condurre ad una distribuzione delle risorse non in linea con i principi di oggettività, equità ed efficienza.

1.6 Conclusioni

La deprivazione sembra predire una quota significativa, al netto dell'età, del bisogno sanitario. Tale quota è relativa anche ad effetti contestuali, fatto che ne accresce il significato in quanto permetterebbe di distribuire le risorse su microaree geografiche svantaggiate.

Per usare la deprivazione materiale come ponderazione della quota capitaria per il riparto del fondo sanitario occorre che si utilizzi una stima dell'associazione tra deprivazione materiale e bisogno sanitario al netto del contributo dell'età pena una sovrastima della sua importanza e inoltre che si usi sempre un indice standardizzato per età.

Infine, i coefficienti dovrebbero essere specifici per Regione e rivalutati nel tempo.

1.7 Appendice (Modelli di Poisson)

Il modello di Poisson assume che $Y_{isa} \sim \text{Poisson}(PY_{isa}\theta_{isa})$, dove PY_{isa} sono gli anni persona (*Person-Year*), θ_{isa} è il rischio relativo, *i* indica l'*i*-esimo individuo, *s* indica la *s*-esima sezione di censimento ed *a* è l'*a*-esima classe d'età. In questo modo ogni individuo conta per il numero di anni persona in cui è a rischio di decesso.

Alternativamente, possiamo scalare gli anni persona rispetto alla mortalità complessiva usando la standardizzazione indiretta interna sul totale della coorte in studio: $PY_{isa} \times \theta^*_{sa} = E_{isa}$, dove θ^*_{sa} è il tasso di riferimento e E_{isa} sono i casi attesi. In questo caso ogni soggetto conta per il numero di casi attesi, che varia tra 0 e 1. In quest'ultimo caso assumiamo valido il modello moltiplicativo per cui $\theta_{isa} = \theta_{is} \times \theta_{sa}$ (Breslow e Day 1978, 1986). Il modello di Poisson diventa: $Y_{isa} \sim \text{Poisson}(E_{isa}\theta_{isa})$, dove θ_{isa} rappresenta il rischio relativo ed E_{isa} è il numero di casi attesi.

1.8 Bibliografia

Armitage P (1955). Test for linear trend in proportions and frequencies. *Biometrics*; **11:** 375–86.

Biggeri A, Gorini G, Dreassi E, Kalala N, Lisi C (2001). Condizione socio-economica e mortalità in Toscana. *Studi e Ricerche*, n. 7, Edizioni Regione Toscana, Centro Stampa Giunta Regionale, Firenze.

Breslow NE, Day NE (1975). Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *Journal of Chronic Diseases;* **28**(5,6),289–303.

Carr-Hill RA, Sheldon TA, Smith PC, Martin S, Peacock S, Hardman G (1994). Allocating resources to health authorities: development of method for small area analysis of use of inpatient services. *British Medical Journal*; **309**: 1046-1049.

Catelan D, Biggeri A, Dreassi E, Lagazio C (2006). Space-cohort Bayesian models in ecological studies. *Statistical Modelling*; **6**: 1–15

Cochran WG (1954). Some methods for strengthening the common chi-squared tests. *Biometrics*; **10:** 417–54.

Costa G, Spadea T, Cardano M (2004). (a cura di) Diseguaglianze di salute in Italia. *Epidemiologia & Prevenzione*; **28** (3).

Cronbach LJ, Webb J (1975). Between-class and within-class effects in a reported aptitude X treatment interaction. *Journal of Educational Psychology*; **67**, 6:717.

Dreassi E, Biggeri A, Catelan D (2005). Space-time models with time dependent covariates for the analysis of the temporal lag between socio-economic factors and lung cancer mortality. *Statistics in Medicine*; **24:** 1–19

Firebaugh G (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*. **43,** 557-572.

Jarman B (1983). Identification of underprivileged areas. *British Medical Journal*; 1705-1709.

Kreft I, De Leeuw J (1998). Introducing multilevel modeling. *SAGE*, London.

Ministero della Salute (1998). Piano Sanitario Nazionale 1998-2000. Un patto di solidarietà per la salute. Roma.

Morgenstern H (1998). Ecologic Studies. In Rothman KJ, Greenland S Modern Epidemiology – second edition. *Lippincott Williams & Wilkins*, Philadelphia.

Rosenbaum P R, Rubin D B (1984). Difficulties with regression analyses of age-adjusted rates. *Biometrics;* **40**, 2, 437-443.

Stone M, Galbraith J (2006). How not fund hospital and community health services in England. *Journal of the Royal Statistical Society A*; **169** Part 1, 143-164.

Wakefield J (2007). Disease mapping and spatial regression with count data. *Biostatistics*; **8,** 2, 158-183.

Tabelle e Figure

Tabella 1.1 Studio Longitudinale Toscano. Mortalità per tutte le cause. Maschi. Coorte censuaria di Firenze 1991-1995, maschi. Numero di individui, frequenza, numero di decessi osservati e attesi, SMR (limiti di confidenza 90%) per livello di deprivazione materiale (numero di condizioni di svantaggio).

	n	*%*	*Decessi Osservati*	*Decessi attesi*	*SMR (LC 90%)*
0=Nessuna	123.111	85	3.710	3.828,2	0,97 (0,94-1,00)
1=Una	19.429	13	549	451,3	1,22 (1,13-1,31)
2=Due o più	1.461	1	53	32,5	1,63 (1,28-2,05)

Tabella 1.2 Studio Longitudinale Toscano. Mortalità per tutte le cause, Firenze 1991-1995, maschi. Studio Longitudinale Toscano.Coefficiente di regressione (errore standard) per livello medio di deprivazione materiale per sezione di censimento (logaritmo del Rischio Relativo). Risultati dei modelli individuali (1.1) (1.2) (1.3) (vedi testo).

Modello	*Covariate*	β	β *per unità di d.s.*
(1.1) Individuale	x_{isa}	0,238 (0,039)	0,093 (0,015)
(1.2) Contestuale	x_{isa}	0,166 (0,041)	0,065 (0,016)
	$\bar{x}_{sa}$	0,672 (0,120)	0,082 (0,015)
(1.3) Cronbach	$\left(x_{isa} - \bar{x}_{sa}\right)$	0,166 (0,041)	0,062 (0,015)
	$\left(\bar{x}_{sa} - \bar{x}\right)$	0,838 (0,113)	0,102 (0,014)

Tabella 1.3 Studio Longitudinale Toscano. Mortalità per tutte le cause. Maschi. Coorte censuaria di Firenze 1991-1995, maschi.Coefficiente di regressione (errore standard) per livello medio di deprivazione materiale per sezione di censimento (logaritmo del Rischio Relativo). Risultati dei modelli Aggregati (1.4) (1.5) (1.6) (1.7) (vedi testo).

Modello	*Covariate*	β^A	β^A *per unità di d.s.*
(1.4)	$\overline{x}_s$	0,792 (0,112)	0,097 (0,014)
(1.5)	$\overline{x}_s^{\,age}$	0,830 (0,116)	0,119 (0,017)
(1.6)	$\overline{x}_s$	0,799 (0,113)	0,143 (0,020)
(1.7)	$\overline{x}_s$	0,787 (0,114)	0,140 (0,020)

Capitolo 2

Deprivazione materiale e incidenza per tumore del polmone. Un'analisi per sezioni di censimento.

2.1 Introduzione

In letteratura Epidemiologica si trovano esempi dell'uso di indicatori ecologici di deprivazione materiale per sezione di censimento come surrogato del valore individuale (Krieger, 1992; e per il nostro paese Michelozzi *et al.*, 1999). La giustificazione di questo modo di procedere consiste nel rilievo che non si ha distorsione ecologica quando i soggetti entro aggregato sono omogenei per esposizione in studio (Morgenstern, 2008). In questa situazione, la sezione di censimento è considerata un aggregato sufficientemente piccolo per ritenere plausibile che l'indice di deprivazione medio sia una buona approssimazione del livello di svantaggio sociale degli individui che vi risiedono (Costa *et al.*, 2004).

Nel nostro paese problemi legati alla riservatezza dei dati individuali hanno fatto si che ci siano pochi studi che confrontano il livello individuale e quello per sezione di censimento (Biggeri *et al.*, 2001). In Italia questo tipo di analisi sono state condotte nell'ambito di studi longitudinali che coinvolgono aree ristrette, in genere una città (ad esempio Torino, Studio Longitudinale Torinese (Marinacci *et al.*, 2004)) o poche città (ad esempio Firenze, Livorno e Prato, Studio Longitudinale Toscano (Biggeri *et al.*, 2004)). I risultati di questi studi evidenziano

un’associazione tra livello di deprivazione materiale e esiti sanitari, che non è solo frutto delle condizioni di svantaggio in cui versano i singoli individui, ma anche delle condizioni medie della sezione di censimento in cui vivono. La presenza di questo effetto di contesto, quando plausibile biologicamente, complica la situazione. Infatti, le stime di effetto relative alla variabile definita sul livello gerarchicamente più alto saranno maggiori di quelle individuali mentre si avranno stime distorte dell’effetto individuale nei modelli dove non si considera adeguatamente il livello gerarchico dei dati (Diez-Roux, 1988).

L’uso di indicatori socio-economici calcolati su aree geografiche, sufficientemente piccole, proposto per il monitoraggio delle diseguaglianze nella salute in mancanza di dati individuali (Subramanian *et al.*, 2006), dipende tuttavia dalla bontà di attribuzione della residenza, in particolare dalla correttezza dei dati sugli indirizzi. Le analisi, infatti, possono essere soggette ad una distorsione da selezione in quanto è stato osservato che gli individui più poveri hanno più probabilità di avere un indirizzo non referenziabile geograficamente (Krieger *et al.*, 2002).

L’obiettivo di questo studio è quantificare l’associazione tra deprivazione materiale e incidenza per tumore del polmone in Toscana usando informazioni socio-economiche disponibili a livello di sezione di censimento. A questo scopo ci siamo avvalsi dei dati individuali raccolti dal Registro Tumori Toscano per l’anno 2004 e, nell’impossibilità di un record-linkage individuale, dei dati aggregati per sezione di censimento del 2001 sui fattori socio-economici. Inoltre, per valutare la completezza dei dati del Registro Tumori Toscano abbiamo a disposizione i dati relativi ai certificati di morte ISTAT per il periodo 2003-2005.

Nel prossimo paragrafo verranno descritti i dati utilizzati per le analisi. Nel paragrafo 2.2 vengono presentati i metodi statistici: i modelli per l’analisi della completezza di registrazione e i modelli di analisi ecologica. I risultati, la discussione e le conclusioni sono riportati rispettivamente nei paragrafi 2.3, 2.4 e 2.5.

2.2 Dati

2.2.1 Dati di Incidenza e Mortalità

I dati di incidenza provengono dal Registro Tumori Toscano (RTT). Si riferiscono al 2004, il primo anno in cui è stata validata la registrazione a livello regionale. Il RTT infatti dal 1984 copre le province di Firenze e Prato e solo recentemente ha esteso la copertura all'intera Toscana.

I record individuali relativi ai nuovi casi di neoplasie diagnosticate nei cittadini residenti contengono informazioni riguardanti la malattia e le caratteristiche demografiche (genere, anno e luogo di nascita, residenza). Il totale di nuovi casi per tumore del polmone registrati dal RTT nel 2004 per la popolazione maschile è 2097, con un tasso grezzo pari a 124,75 × 100.000. Utilizzando un apposita procedura[1] presso la Regione Toscana, Dipartimento del Territorio si è risaliti dall'indirizzo presente nell'archivio RTT alla sezione di censimento 2001. 1545 record sono stati correttamente imputati. Nessuno dei casi linkati correttamente, è stato attribuito alla provincia di Arezzo. I motivi sono legati alla procedura regionale che ancora non copre l'intera Regione. Per questo motivo abbiamo deciso di escludere dall'analisi la provincia di Arezzo. La distribuzione per età dei casi esclusi dall'analisi è simile a quella dei casi in studio come mostra la figura 2.1. I Comuni oggetto dello studio sono 248 su 287 per un totale di 23.182 sezioni di censimento, la tabella 2.1 mostra la frequenza dei casi per provincia.

La figura 2.2 mostra la distribuzione spaziale del rapporto di incidenza standardizzato per i comuni dell'area oggetto dello studio. Il numero di casi attesi è stato calcolato seguendo la standardizzazione indiretta interna e classificando la popolazione in 16 classi d'età (0-4, 5-9,…, 75 e più). Le popolazioni per sezione di

[1] La sezione di censimento viene attribuita grazie all'utilizzo di SISTER che è un sistema informatico, composto da software e dati, in grado di riconoscere, normalizzare e codificare automaticamente indirizzi fino al livello di sezione di censimento per tutto il territorio nazionale, e di gestire le relative base dati di riferimento, http://ius.regione.toscana.it.

censimento hanno una struttura per età a volte molto particolare e nella standardizzazione si ha un'imperfetta eguaglianza del totale dei casi osservati con il totale dei casi attesi. L'intercetta dei modelli di regressione sarà perciò non interpretabile.

Per verificare la completezza di registrazione dei dati del Registro Tumori Toscano abbiamo condotto una analisi preliminare sul rapporto tra mortalità e incidenza (M/I). Abbiamo utilizzato i dati di mortalità ISTAT per il periodo 2003-2005, relativi ai decessi per tumore del polmone, uomini, per i 248 Comuni Toscani per i quali si dispone dei dati di incidenza. Il rapporto M/I dovrebbe essere intorno all'unità perché in condizioni stabili il numero di decessi per tumore polmonare equivale al numero di nuovi casi. La malattia è infatti letale e la sopravvivenza mediana inferiore ad un anno. Un valore maggiore di 1 è indicativo di una perdita nella registrazione di nuovi casi, un rapporto inferiore a 1 suggerisce che la casistica sia comprensiva di casi prevalenti, diagnosticati in precedenza e lungo sopravviventi per i quali il Registro non è riuscito a determinare correttamente la data di diagnosi (Parkin *et al.*, 1994)

2.2.2 Dati sulla deprivazione materiale

L'indice di deprivazione materiale è stato costruito sulla base di alcune variabili disponibili a livello di sezione al Censimento del 2001, resi disponibili dall'Istituto Nazionale di Statistica (ISTAT) (Ministero della Salute, Regione Autonoma Valle d'Aosta). In particolare le variabili considerate per definire l'ambiente sociale sono state:

- l'istruzione: percentuale di analfabeti, alfabeti o con licenza elementare sulla popolazione con più di 6 anni;
- la condizione professionale: percentuale di disoccupati sulla popolazione attiva;
- l'indice di affollamento: numero medio di occupanti per stanza
- le abitazioni in affitto: proporzione di abitazioni in affitto sul totale delle abitazioni.

È quindi stato calcolato per ognuna delle variabili il rispettivo punteggio *Z*:

$$Z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k}$$

dove x_{ik} è il valore osservato nell'*i*-esima sezione di censimento della *k*-esima variabile; μ_k è la media della *k*-esima variabile; e σ_k è lo scarto quadratico medio della *k*-esima variabile.

L'indice di deprivazione materiale per tutte le sezioni di censimento dell'area Toscana oggetto dello studio è stato calcolato come somma dei punteggi Z (Carstairs, 1995, Grisotto *et al.*, 2008):

$$ID_i = \sum_{k=1}^{4} Z_{ik} \ .$$

La figura 2.4 mostra la distribuzione dell'indice di deprivazione medio per comune dell'area oggetto dello studio.

2.3 Metodi

2.3.1 Modelli per la stima della completezza dei Registri Tumori, stima del rapporto tra mortalità e incidenza

Il rapporto tra mortalità e incidenza (M/I) è un indicatore di completezza della registrazione (Parkin *et al*. 1994).

In questo studio proponiamo una stima del rapporto M/I basata su un modello Bayesiano congiunto. Consideriamo la variabile di risposta Y_{lk}, dove *l*=1,2 indicizza le frequenze osservate di decessi o casi incidenti, rispettivamente, e *k* indica il comune o l'area geografica, distribuita $Y_{lk} \sim Poisson(\theta_{lk} Pop_k)$, dove θ_{lk} è il rischio relativo e Pop_k gli anni persona a rischio.

Un modello di regressione bivariato semplice è il seguente:

$$\log \theta_{lk} = \alpha + \delta_k \; I_l \qquad (2.1)$$

dove α è l'intercetta, I_l è una variabile indicatrice che vale 1 per la mortalità e 0 per l'incidenza, δ_k è un parametro incognito normalmente distribuito, che rappresenta il logaritmo del rapporto M/I.

Alternativamente, abbiamo inserito una componente di eterogeneità u_k, normalmente distribuito con media zero e varianza $\sigma_u^2 \sim Gamma(a_{\sigma_u^2}; b_{\sigma_u^2})$ ($a_{\sigma_u^2}$ e $b_{\sigma_u^2}$ tali che la a priori sia debolmente informativa):

$$\log \theta_{lk} = \alpha + u_k + \delta_k \; I_l \qquad (2.2)$$

o considerato nel modello una componente spazialmente strutturata c_k, nella definizione di Besag *et. al.* (1991) (*Improper Conditional Autoregressive* ICAR):

$$\begin{aligned} &\log \theta_{lk} = \alpha + c_k + \delta_k \; I_l \\ &c_k \left| c_{v \sim k} \right. \sim Normale\left(\bar{c}_k, \lambda_c n_k\right) \end{aligned} \qquad (2.3)$$

dove $v \sim k$ indica le aree adiacenti al *k*-esimo comune, $\bar{c}_k = \sum_{v \sim k} c_v \Big/ n_k$ è la media dei valori delle aree adiacenti al *k*-esimo comune e n_k è la loro numerosità, $\lambda_c \sim IGamma(a_c, b_c)$ è la precisione strutturata spazialmente.

Infine, abbiamo definito un modello dove è presente la componente di eterogeneità u_k e il parametro δ_k è strutturato spazialmente (ICAR):

$$\begin{aligned} &\log \theta_{lk} = \alpha + u_k + \delta_k \; I_l \\ &u_k \sim Normale\left(0, \sigma_u^2\right), \; \delta_k \left| \delta_{v \sim k} \right. \sim Normale\left(\bar{\delta}_k, \lambda_\delta n_k\right). \end{aligned} \qquad (2.4)$$

Le a priori sugli iperparametri dei modelli sono tutte vagamente informative.

2.3.2 Modelli gerarchici per l'analisi dell'incidenza

Definiamo x_{jk} l'indice di deprivazione materiale per la *j*-esima sezione di censimento (*j* = 1,…, 23.182) del *k*-esimo comune (*k* = 1,…,248); $\bar{x}_k$ denota l'indice di deprivazione medio del *k*-esimo comune. Assumiamo che il numero di nuovi casi osservati di tumore del polmone, nella *j*-esima sezione di censimento del *k*-esimo comune (Y_{jk}) segua una distribuzione di Poisson con valore atteso $E_{jk}\theta_{jk}$, $Y_{jk} \sim \text{Poisson}(E_{jk}\theta_{jk})$, dove E_{jk} è il numero di casi attesi, e θ_{jk} è il rischio relativo.

Sono stati specificati i seguenti modelli lineari generalizzati sul logaritmo del rischio relativo:

Modello ecologico definito solo a livello di sezione di censimento:

$$\log \theta_{jk} = \beta^0 + \beta^1 x_{jk} \qquad (2.5)$$

dove β^0 è l'intercetta, β^1 è l'effetto ecologico del livello di deprivazione medio della sezione di censimento.

Modello ecologico definito solo a livello di comune:

$$\log \theta_{jk} = \beta^0 + \beta^2 \bar{x}_k \qquad (2.6)$$

dove β^2 è l'effetto ecologico del livello di deprivazione media del comune.

Modello contestuale (Firebaugh, 1978):

$$\log \theta_{jk} = \beta^0 + \beta^{1*} x_{jk} + \beta^{2*} \bar{x}_k \qquad (2.7)$$

dove β^{1*} è l'effetto del livello di deprivazione della sezione di censimento entro comune, effetto *within*, β^{2*} è l'effetto contestuale della deprivazione media dei comuni.

Modello di Cronbach (Cronbach e Webb, 1975):

$$\log\theta_{jk} = \beta^0 + \beta^{1*}\left(x_{jk} - \bar{x}_k\right) + \beta^2\left(\bar{x}_k - \bar{x}\right) \tag{2.8}$$

dove β^{1*} è l'effetto del livello di deprivazione della sezione di censimento entro comune, effetto *within*, β^2 è l'effetto complessivo della deprivazione, effetto *between*, (effetto contestuale più l'effetto *within*).

2.3.3 Modelli Bayesiani gerarchici multilivello.

I precedenti modelli non considerano la eventuale variabilità del rischio di base, assumendo un'intercetta costante. Questa assunzione è inadatta quando considero un insieme di comuni in una regione.

Un semplice approccio parametrico consiste nello specificare una intercetta per ogni area (comune). L'intercetta può essere parametrizzata come una realizzazione di una variabile casuale Normale con media e varianza incognite.

$$\begin{aligned} &\log\theta_{jk} = \beta_k^0 + \beta^{1*}\left(x_{jk} - \bar{x}_k\right) + \beta^2\left(\bar{x}_k - \bar{x}\right) \\ &\beta_k^0 \sim Normale\left(\mu_{\beta^0}, \sigma^2_{\beta^0}\right). \end{aligned} \tag{2.9}$$

Un'altra specificazione nell'approccio parametrico prevede sia l'intercetta che la pendenza casuali, definite come la realizzazione di una variabile casuale Normale multivariata con media e matrice di covarianza incognite.

$$\begin{aligned} &\log\theta_{jk} = \beta_k^0 + \beta_k^{1*}\left(x_{jk} - \bar{x}_k\right) + \beta^2\left(\bar{x}_k - \bar{x}\right) \\ &\left(\beta_k^0, \beta_k^{1*}\right) \sim MNormale\left(\boldsymbol{\gamma}, \mathbf{T}^{-1}\right). \end{aligned} \tag{2.10}$$

dove i coefficienti di regressione $\left(\beta_k^0, \beta_k^{1*}\right)$ seguono una distribuzione multi normale, iperparametri $\boldsymbol{\gamma} \sim MNormale(\mathbf{a}_\gamma, \mathbf{b}_\gamma)$ e $\mathbf{T} \sim Wishart(\mathbf{a}_T, \mathbf{b}_T)$ con a priori vagamente informative.

Gli effetti casuali possono avere una struttura spaziale. Per questo abbiamo deciso di utilizzare i modelli condizionatamente auto regressivi nella formulazione di Besag (1974).

Abbiamo stimato due modelli con effetti casuali spazialmente strutturati, il primo modello presenta solo l'intercetta casuale:

$$\begin{aligned} &\log\theta_{jk} = \beta_k^0 + \beta^{1*}\left(\bar{x}_{jk} - \bar{x}_k\right) + \beta^2\left(\bar{x}_k - \bar{x}\right) \\ &\beta_k^0 \left| \beta_{v\sim k}^0 \right. \sim Normale\left(\bar{\beta}_k^0, \lambda_{\beta^0} n_k\right) \end{aligned} \tag{2.11}$$

dove $v \sim k$ indica le aree adiacenti al *k*-esimo comune, $\bar{\beta}_k^0 = \sum_{v\sim k} \beta_v \big/ n_k$ è la media dei valori delle aree adiacenti al *k*-esimo comune e n_k è la loro numerosità e $\lambda_\beta \sim IGamma(a_\beta, b_\beta)$ è la precisione (modello ICAR).

Nel secondo modello è stata inserita anche la pendenza casuale, non strutturata spazialmente:

$$\begin{aligned} &\log\theta_{jk} = \beta_k^0 + \beta_k^{1*}\left(\bar{x}_{jk} - \bar{x}_k\right) + \beta^2\left(\bar{x}_k - \bar{x}\right) \\ &\beta_k^0 \left| \beta_{v\sim k}^0 \right. \sim Normale\left(\bar{\beta}_k^0, \lambda_{\beta^0} n_k\right),\ \beta_k^{1*} \sim Normale\left(\mu_{\beta^{1*}}, \sigma^2_{\beta^{1*}}\right) \end{aligned} \tag{2.12}$$

In tutti i modelli le distribuzioni a priori sugli iperparametri sono vagamente informative.

2.3.4 Scelta del modello

Nella letteratura Bayesiana sono stati formalizzati differenti approcci per risolvere il problema della scelta del modello. L'indicatore più utilizzato è il DIC (*Deviance Information Criterion*) proposto da Spiegelhalter *et al.* (2002), definito come somma di un indicatore di bontà di adattamento del modello ai dati ($\overline{D}$) e di un indice di complessità (p_D):

$$\text{DIC} = \overline{D} + p_D \tag{2.13}$$

dove $\overline{D}$ è l'*expected posterior deviance* che riassume la bontà di adattamento del modello:

$$\overline{D} = E_{\theta|y}[D]$$

e p_D è l'effettivo numero di parametri del modello che ne studia la complessità:

$$p_D = E_{\theta|y}[D] - D\left(E_{\theta|y}[\theta]\right) = \overline{D} - D(\bar{\theta}) .$$

Il DIC viene definito a partire dalla devianza:

$$D(\theta) = 2\sum_j \sum_k Y_{jk} \log \frac{Y_{jk}}{\theta_{jk} E_{jk}} - Y_{jk} + \theta_{jk} E_{jk}$$

dove θ rappresenta il vettore dei parametri incogniti del modello.

Un recente studio di Plummer (2008) mette in discussione l'uso del DIC se si ha un numero di osservazioni indipendenti minore del numero di parametri da stimare, come accade in genere nelle analisi geografiche (*disease mapping*).

Per questo Plummer (2008) propone l'utilizzo di una misura del "grado di ottimismo" da utilizzare in una funzione di perdita penalizzata basata su una *cross-validation*. Nel nostro caso, la numerosità delle osservazioni rende complicata la

stima della funzione di perdita penalizzata di Plummer, tuttavia l'autore propone un'alternativa più semplice che abbiamo scelto di seguire. Abbiamo quindi calcolato il DIC corretto come:

$$\begin{aligned} &\mathrm{DIC}c = \overline{\mathrm{D}} + \mathrm{r}_{\mathrm{opt}} \\ &\mathrm{r}_{\mathrm{opt}} \approx \sum_{j=1}^{J} \mathrm{p}_{\mathrm{D}\,j} / (1 - \mathrm{p}_{\mathrm{D}\,j}) \end{aligned} \qquad (2.14)$$

dove *j* indica la generica sezione di censimento, $\mathrm{r}_{\mathrm{opt}}$ (*resisual optimism*) misura la complessità del modello.

Inoltre, abbiamo calcolato l'EPD (*Expected Predictive Deviance*) che misura la discrepanza tra i dati osservati e i dati predetti:

$$EPD = 2\sum (Y_{jk} + 0.05)\log((Y_{jk} + 0.05)/(Y_{jk}^{*} + 0.05)) - Y_{jk} + Y_{jk}^{*} \qquad (2.15)$$

dove i dati predetti da modello, Y_{jk}^{*}, sono stati campionati dalla distribuzione predittive (Gelfand e Ghosh, 1998).

2.4 Risultati

Completezza dei dati di incidenza:

Il rapporto tra mortalità e incidenza stimato è in media pari a 1. La distribuzione è leggermente asimmetrica con alcuni comuni con valori massimi compresi tra 1,7 e il 3,5%, a seconda del modello scelto (figura 2.5)

La mortalità evidenzia eccessi nella Lunigiana e Garfagnana (figura 2.3) non corrispondenti nell'incidenza che invece marca l'area della Val d'Orcia dove la mortalità è inferiore (figura 2.2). Una lieve incompletezza di registrazione è ascrivibile all'area Nord Occidentale della regione.

Deprivazione e incidenza di tumore polmonare negli uomini.

Vi è evidenza di un effetto del livello di deprivazione materiale a livello di sezione di censimento sull'incidenza del tumore del polmone negli uomini. All'aumentare delle condizioni di svantaggio aumenta l'incidenza di malattia. Il coefficiente di regressione della deprivazione materiale per sezione di censimento è pari a 0,024 (RR=1,024 per incremento unitario, IC al 90%: 1 – 1,048). Se consideriamo l'effetto per unità di deviazione standard otteniamo valori di RR= 1,035 (IC al 90%: 0,982-1,092 (tabella 2.2). L'analisi a livello comunale mostra un effetto della deprivazione materiale pari a 0.166 (RR=1,181, IC al 90%: 1,050 – 1,329). L'effetto è paragonabile se riportato per unità di deviazione standard (0,064).

Il coefficiente di regressione per la deprivazione a livello di sezione censuaria stimata dai modelli contestuale e di Cronbach è di 0,018 (IC al 90%: -0,007 – 0,045), e standardizzato di 0.038 (IC al 90%: -0,013 – 0,063) (tabella 2.2).

L'effetto contestuale della deprivazione materiale media per comune sembra essere più forte dell'effetto per sezione di censimento (0.039 vs 0.058, coefficienti per unità di deviazione standard, modello (2.7)).

I modelli Bayesiani gerarchici permettono di includere la dipendenza spaziale dei comuni vicini tramite intercette casuali e spazialmente strutturata (figure 2.6, 2.7, 2.8, 2.10, 2.11). L'effetto della deprivazione media comunale si attenua passando da RR=1,18 a RR=1,16 (tabella 2.3).

Il modello con pendenza casuale mostra alcune aree in cui l'effetto della deprivazione per sezione di censimento è negativo (figura 2.9, 2.13). I comuni sono nel caso del modello (2.10): Lucca, Carrara, Prato, Cascina, Empoli, Siena, Ponsacco. Il modello (2.12), con intercetta spazialmente strutturata, mostra un effetto della deprivazione invertito per i comuni di Prato, Carrara, Lucca e Cascina (in tabella 2.4 sono riportati i RR).

In tabella 2.5 sono confrontati i modelli tra loro. Nonostante un peggior adattamento il modello (2.11) risulta migliore in quanto meno complesso.

2.5 Discussione

In questo lavoro la georeferenziazione della residenza è stata eseguita per mezzo di una procedura automatica. Data la mancata copertura abbiamo eliminato la Provincia di Arezzo. Sui rimanenti la percentuale di mancata attribuzione delle coordinate geografiche è del 10%.

Il profilo geografico di mortalità è leggermente diverso da quello di incidenza. Le spiegazioni possono essere due: una minor completezza e accuratezza di registrazione trattandosi dei primi anni di copertura regionale del Registro Tumori Toscano, oppure una variazione del rischio di malattia per coorte di nascita.

La modellazione statistica da noi adottata è un'estensione dei modelli di regressione ecologica in presenza di un campo markoviano latente (Besag e Higdon, 1993, Clayton *et al.*, 1993).

Nell'ipotesi di una modificazione di effetto legata al contesto urbano/rurale sono stati introdotti modelli a pendenza casuale formulando una struttura gerarchica semplice entro e tra comune. Entro comune le sezioni di censimento sono state modellate con termini casuali scambiabili. In letteratura, ma per zone geografiche molto più piccole di quelle considerate da noi (l'intera Regione Toscana), sono state usate strutture parametriche di correlazione spaziale (Best *et al.*; 2001).

Non sono stati considerati tutti i possibili modelli perché l'interesse è rivolto ad alcuni parametri e l'inclusione di molti termini casuali porta a problemi di identificabilità. Abbiamo perciò ristretto l'attenzione solo ad alcuni modelli che rispondevano a precise ipotesi di ricerca.

Nondimeno, un confronto tra i modelli che abbiamo adattato è stato necessario. Abbiamo valutato il valor atteso della devianza a posteriori come misura di adattamento e l'ottimismo residuo come definito da Efron (1983) e Plummer (2008) come stima della complessità del modello. Altre soluzioni sono proposte in letteratura ma potrebbero essere molto pesanti computazionalmente. Nel caso avessimo un modello di riferimento, un misura di tipo Kullback-Leibler sarebbe di aiuto.

Gli effetti osservati sono di ordine di grandezza simile a quelli attesi sulla base della letteratura. Compare qui un rischio relativo dell'ordine di 1,16 per il valore

medio comunale, anche considerando una struttura latente spazialmente strutturata. Questo risultato è interessante e va letto alla luce delle differenze tra Comuni Toscani rispetto a stili di vita e esposizioni lavorative.

La modificazione d'effetto legata al gradiente urbano/rurale non è molto supportata dai dati. Il modello è molto complesso (tabella 2.5). Tuttavia, emergono alcune aree con gradienti inversi che meritano considerazioni specifiche e richiedono ulteriori indagini.

2.6 Conclusioni

La deprivazione materiale è associata ad un incremento di incidenza di tumore polmonare di circa 3,8% per unità di deviazione standard. Complessivamente, considerando anche il contesto a livello comunale si ha un eccesso di rischio intorno al 16%.

Questi risultati sono molto solidi basandosi su una popolazione regionale di 1.523.903 (popolazione maschile dell'area Toscana oggetto dello studio al censimento 2001) e 1.545 casi.

Eventuali confondenti potranno essere inclusi nel modello che nella formulazione proposta considera soltanto effetti latenti spazialmente strutturati.

2.7 Bibliografia

Besag J (1974). Spatial Interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B*, **36**, 192-236.

Clayton DG, Bernardinelli L, Montmoli C (1993). Spatial Correlation in Ecological Analysis. *International Journal of Epidemiology*, **22**, 1193–1202.

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

Best N, Cocking S, Bennett J, Wakefield J, Elliott P (2001). Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society A*, **164**, 155-174.

Biggeri A, Gorini G, Dreassi E, Kalala N, Lisi C (2001). Condizione socio-economica e mortalità in Toscana, *Studi e Ricerche*, n. 7, Edizioni Regione Toscana, Centro Stampa Giunta Regionale, Firenze.

Clayton DG, Bernardinelli L, Montmoli C (1993). Spatial Correlation in Ecological Analysis. *International Journal of Epidemiology*; 22, 1193–1202.

Costa G, Cardano M, Demaria M (1998). Torino. Storie di salute di una grande città. Osservatorio socio-economico torinese. Ufficio di Statistico città di Torino.

Diez-Roux AV (1988). Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *American Journal of Public Health*, **88**, 216-222.

Efron B (1983). Estimating the error rate of a prediction rule: improvements on cross-validation. *Journal of the American Statistical Association*, **78**, 316-331.

Firebaugh G (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*; **43,** 557-572.

Gelfand AE, Ghosh SK (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*; **85**, 1–11.

Grisotto L., Catelan D., Lagazio C., Biggeri A. (2007). L'uso dell'indice di deprivazione materiale in epidemiologia descrittiva, *Rapporto ISTISAN* **07/05**, 123-134.

Marinacci C, Spadea T, Biggeri A, Demaria M, Caiazzo A, Costa G (2004). The role of individual and contextual socioeconomic circumstances on mortality: analysis of time variations in a city of north west Italy. *Journal of Epidemiology and Community Health*; **58**, 199-207.

Michelozzi P, Perucci CA, Forastiere F, Fusco D, Ancona C, Dell'Orco V (1999). Inequality in health: socioeconomic differentials in mortality in Rome, 1990-95. *Journal of Epidemiology and Community Health*; **53**, 687-693.

Ministero della Salute, Regione Autonoma Valle d'Aosta (2004-2008). Ricerca ex art. 12 "Valorizzazione dei dati del censimento 2001 per il monitoraggio e l'analisi delle diseguaglianze sociali nella salute in Italia". Responsabile Scientifico A. Biggeri.

Krieger N (1992). Overcoming the absence of socio-economic data in medical records: validation and application of a census-based methodology. *American Journal of Public Health*, **82**, 703-710.

Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R (2002). Geocoding and monitoring of US socio-economic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? *American Journal of Epidemiology*; **156**, 471-482.

Morgenstern H (1998). Ecologic Studies. In Rothman K and Greenland S (eds) Modern Epidemiology, *Lippincott-Raven,* Boston.

Parkin DM, Chen VW, Ferlay J, Galceran J, Storm H, Whelan S (1994). Comparability and quality control in cancer registration. *IARC Technical Report No. 19*. Lyon.

Plummer M (2008). Penalized loss functions for Bayesian model comparison. *Biostatistic*s; **9**, 523-539.

Spiegelhalter DJ, Best N, Calrlin BP, van der Linde A (2002). Bayesian measures of model complexity an fit (with discussion). *Journal of the Royal Statistical Society B*; **64**, 583-639.

St Leger S (1995). Use of deprivation indices in small area studies of environment and health. *Journal of Epidemiology and Community Health;* **49** (Suppl 2).

Subramanian SV, Chen JT, Rehkopf DH, Waterman PD, Krieger N (2006). Comparing individual- and area-based socioeconomic measures for the surveilliance of health disparities: a multilevel analysis of Massachussetts birth, 1989-1991. *American Journal of Epidemiology*; **164**, 823-834.

Tabelle e figure

Tabella 2.1 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Distribuzione per provincia dei casi in studio.

Provincia	*Casi*	*Frequenze*
Massa Carrara	96	0.06
Lucca	204	0.13
Pistoia	150	0.10
Firenze	381	0.25
Livorno	177	0.11
Pisa	195	0.13
Siena	145	0.09
Grosseto	100	0.06
Prato	97	0.06
Totale	1545	1

* Esclusa la provincia di Arezzo.

Tabella 2.2 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Rischio relativo, logaritmo del rischio relativo, errore standard del logaritmo del rischio relativo e logaritmo del rischio relativo per unità di deviazione standard per livello di deprivazione nei modelli di Poisson.

Modelli		*RR*	*Log(RR)*	*es del Log(RR)*	*Log(RR) - Indice/s.e.*
(2.5) Ecologico	Sezione di censimento	1,024	0,024	0,014	0,035
(2.6)	Comune	1,181	0,166	0,072	0,064
(2.7) Contestuale	Sezione di censimento	1,019	0,018	0,015	0,039
	Comune	1,163	0,151	0,073	0,058
(2.8) Cronbach	Sezione di censimento	1,019	0,018	0,015	0,038
	Comune	1,185	0,169	0,072	0,065

* Esclusa la provincia di Arezzo.

Tabella 2.3 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Stime dei coefficienti fissi dei modelli Bayesiani gerachici. Rischio Relativo, Logaritmo del rischio relativo, errore standard del Log(RR) e logaritmo del rischio relativo per unità di deviazione standard.

Modelli		*RR*	*Log(RR)*	*es del Log(RR)*	*Log(RR) - Indice/s.e.*
(2.9) Intercetta casuale non strutturata spazialmente	Sezione di censimento	1,016	0,016	0,014	0,033
	Comune	1,182	0,167	0,091	0,063
(2.10) Intercetta e coefficiente casuale non strutturati spazialmente	Sezione di censimento	-	-		-
	Comune	1,183	0,168	0,091	0,067
(2.11) Intercetta casuale ICAR	Sezione di censimento	1,016	0,016	0,014	0,033
	Comune	1,161	0,149	0,094	0,057
(2.12) Intercetta casuale ICAR e coefficiente casuale non strutturato spazialmente	Sezione di censimento	-	-		-
	Comune	1,160	0,149	0,094	0,057

* Esclusa la provincia di Arezzo.

Tabella 2.4 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004.. Stime dei coefficienti casuali, RR per indice di deprivazione materiale nei comuni che presentano un effetto per indice di deprivazione negativo.

Modelli:	*Comune*	*RR*
(2.10) Intercetta e coefficiente casuale non strutturati spazialmente	Carrara	0.974
	Cascina	0.983
	Empoli	0.988
	Lucca	0.973
	Ponsacco	0.999
	Prato	0.977
	Siena	0.998
(2.12) Intercetta casuale ICAR e coefficiente casuale non strutturato spazialmente	Carrara	0.986
	Cascina	0.996
	Lucca	0.993
	Prato	0.984

* Esclusa la provincia di Arezzo.

Tabella 2.5 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Modelli Bayesiani gerachici. Indici di bontà di adattamento: $\overline{D}$, DIC, pd, DIC corretto (DICc), r-opt, EPD (vedi testo).

Modelli	$\overline{D}$	*DIC*	p_D	*DICc*	r_{opt}	*EDP*
(2.9) Intercetta casuale non strutturata spazialmente	4578	4670	92	4666	88	8469
(2.10) Intercetta e coefficiente casuale non strutturati spazialmente	4580	4680	100	4684	104	8466
(2.11) Intercetta casuale ICAR	4583	4646	63	4648	65	8471
(2.12) Intercetta ICAR e coefficiente casuale non strutturati spazialmente	4570	4663	93	4667	97	8471

* Esclusa la provincia di Arezzo.

Figura 2.1 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Distribuzione di frequenza per classe d'età dei casi inclusi ed esclusi dall'analisi.

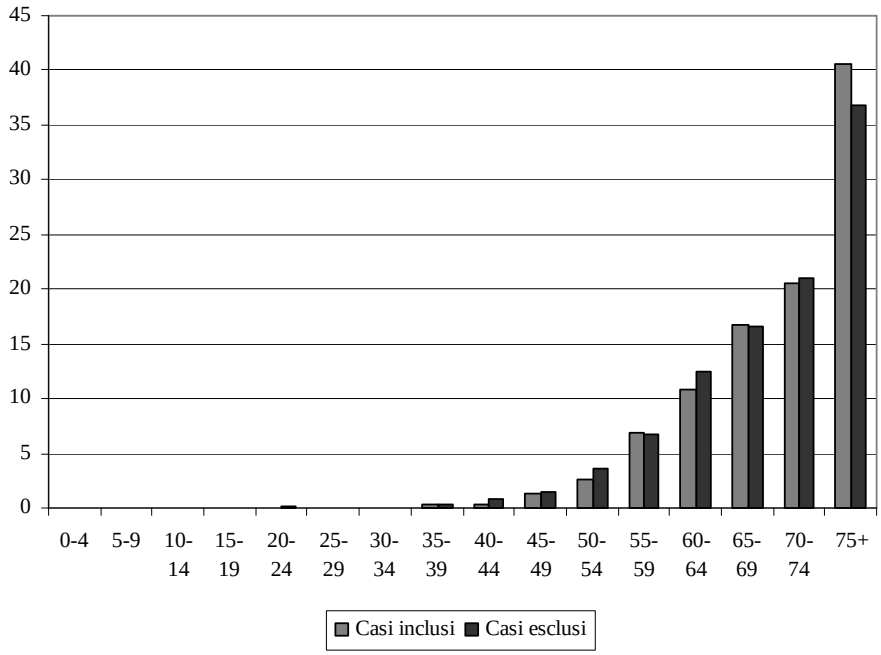
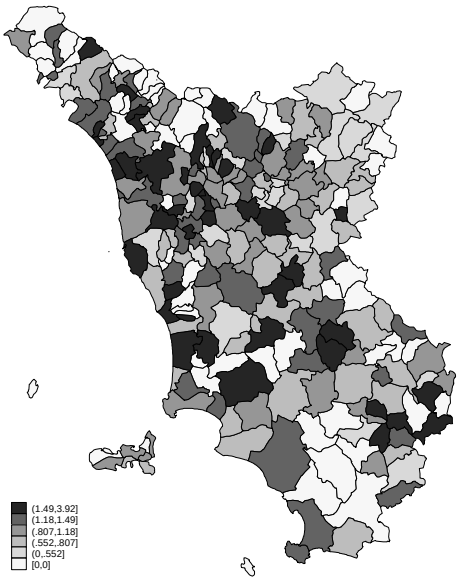
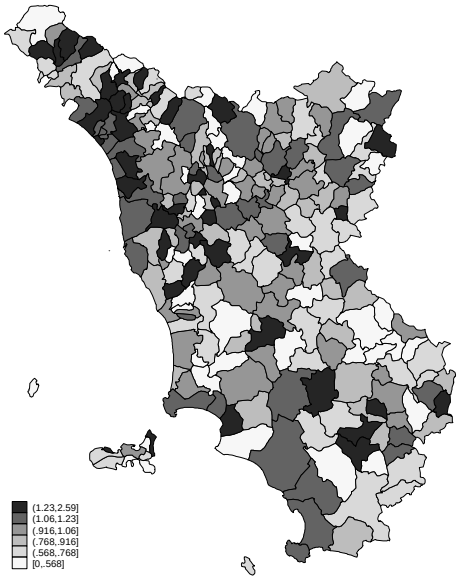


Figura 2.2. Registro Tumori Toscano. Distribuzione spaziale dell'incidenza del Tumore del polmone, maschi, Toscana* 2004.



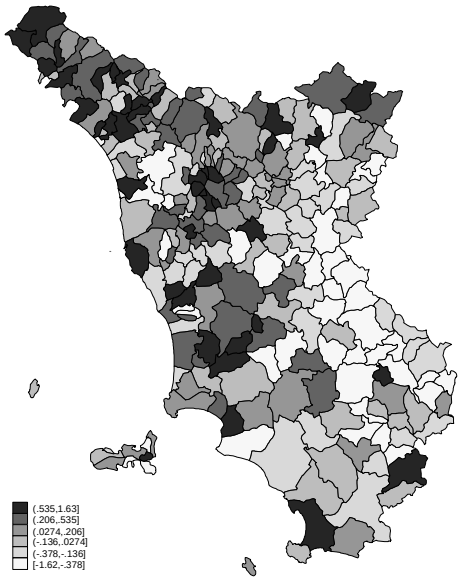
* Esclusa la provincia di Arezzo.

Figura 2.3. ISTAT. Distribuzione spaziale della mortalità per Tumore del polmone, maschi, Toscana* 2003-2005.



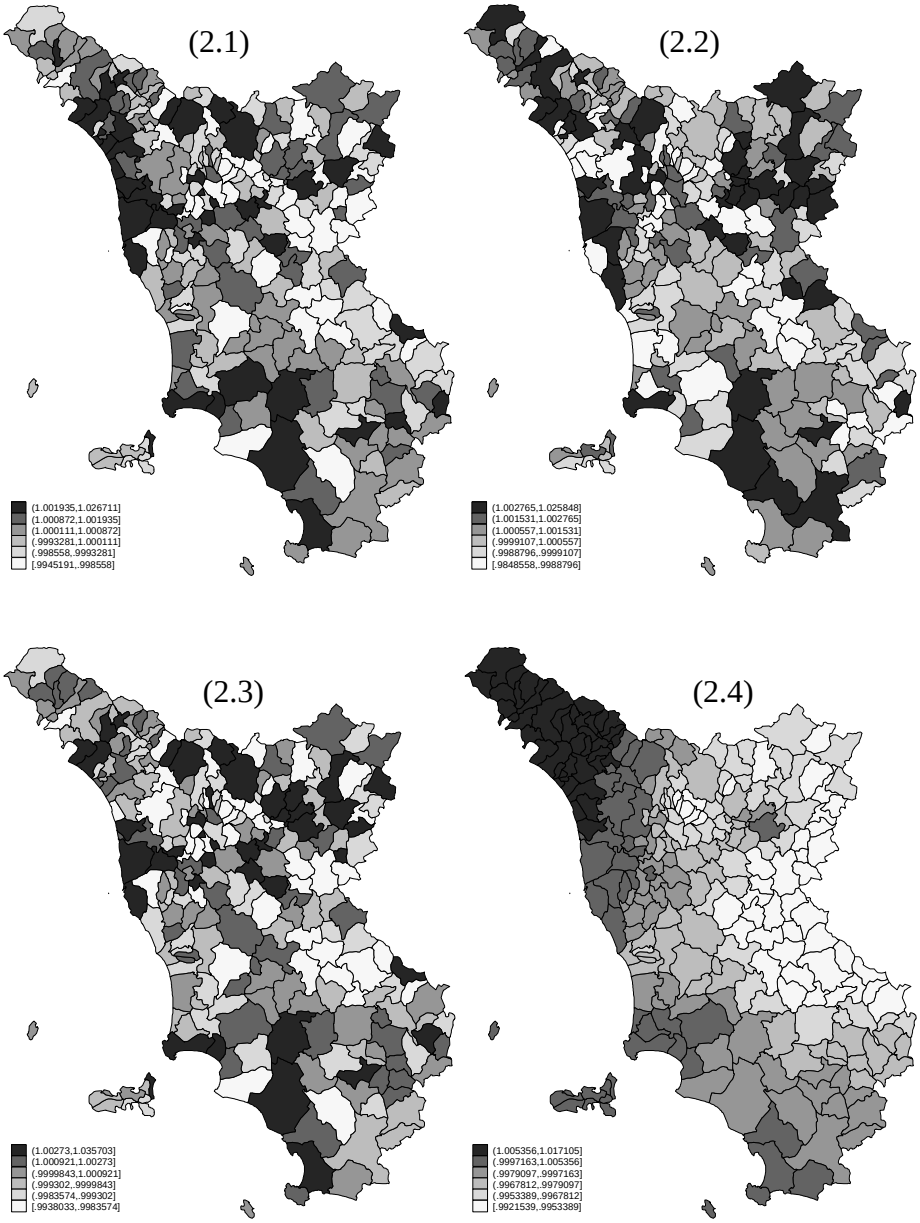
* Esclusa la provincia di Arezzo.

Figura 2.4 ISTAT. Censimento 2001, Toscana*. Distribuzione territoriale dell'indice di deprivazione materiale medio per comune.



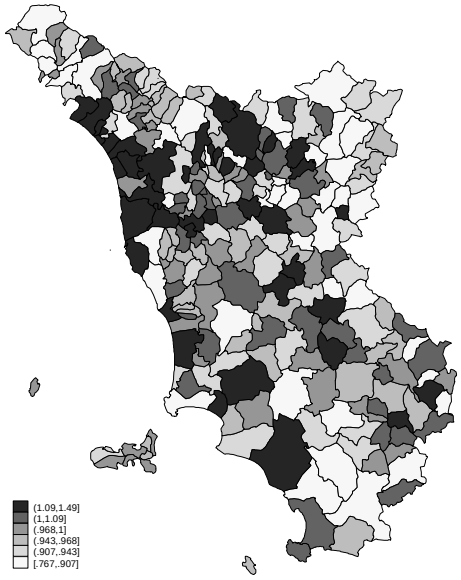
* Esclusa la provincia di Arezzo.

Figura 2.5 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Distribuzione spaziale del rapporto M/I stimato da modello (2.1), (2.2), (2.3), (2.4) (vedi testo).



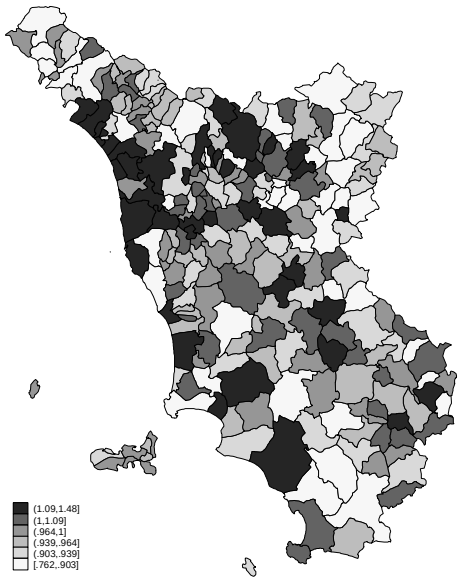
* Esclusa la provincia di Arezzo.

Figura 2.6 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Distribuzione spaziale dei rischi relativi *baseline* per comune. Modello con intercetta casuale non spazialmente strutturata (modello (2.9)) (vedi testo).



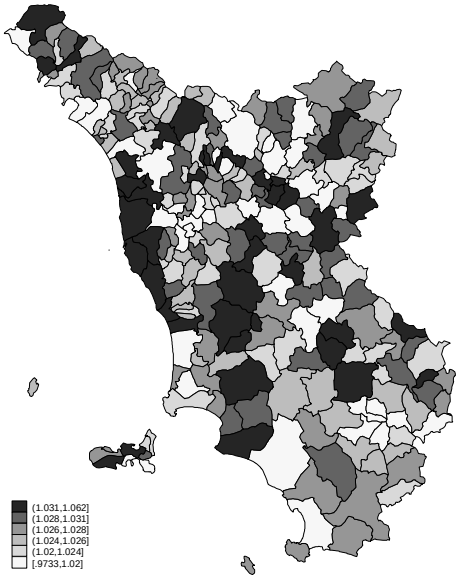
* Esclusa la provincia di Arezzo.

Figura 2.7 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Distribuzione spaziale dei rischi relativi *baseline* per comune. Modello con intercetta casuale e coefficiente non strutturati spazialmente (modello (2.10)) (vedi testo).



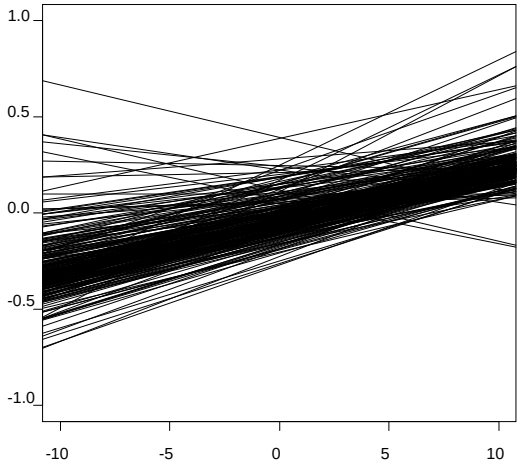
* Esclusa la provincia di Arezzo.

Figura 2.8 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Distribuzione spaziale dei rischi relativi per livello di deprivazione per comune. Modello con intercetta casuale e coefficiente non strutturati spazialmente (modello (2.10)) (vedi testo).



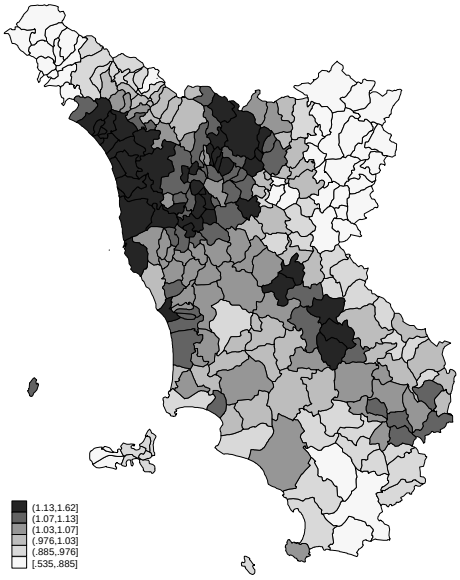
* Esclusa la provincia di Arezzo.

Figura 2.9 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Effetto *within* per ogni comune stimati dal modello (2.10) (vedi testo).



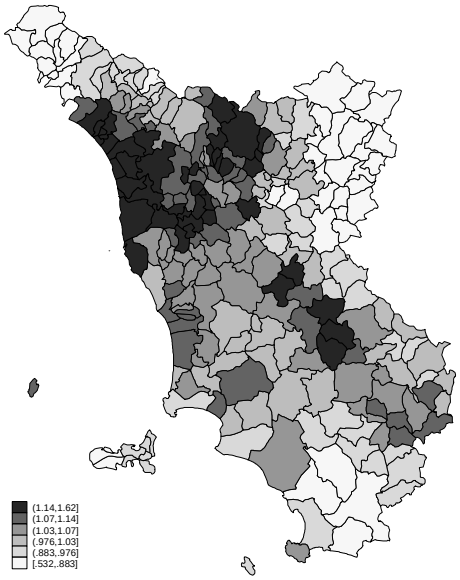
* Esclusa la provincia di Arezzo.

Figura 2.10 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Distribuzione spaziale dei rischi relativi *baseline* per comune. Modello con intercetta casuale spazialmente strutturata (modello (2.11)) (vedi testo).



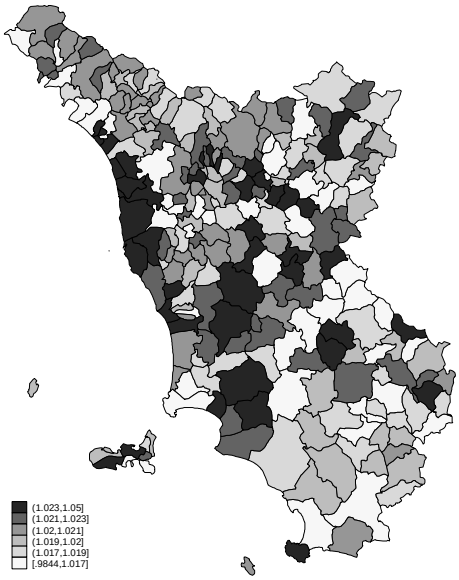
* Esclusa la provincia di Arezzo.

Figura 2.11 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Distribuzione spaziale dei rischi relativi *baseline* per comune. Modello con intercetta casuale ICAR e coefficiente non strutturato spazialmente (modello (2.12)) (vedi testo).



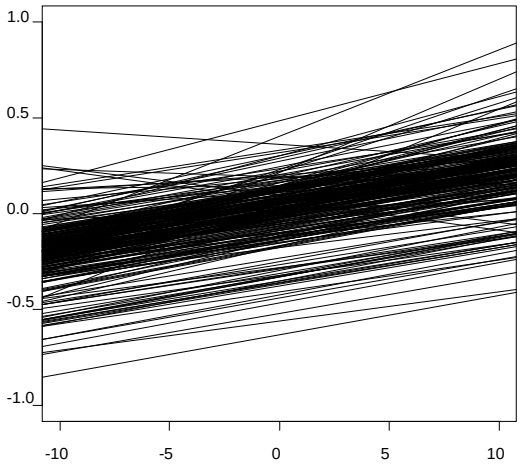
* Esclusa la provincia di Arezzo.

Figura 2.12 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Distribuzione spaziale dei rischi relativi per livello di deprivazione per comune. Modello con intercetta casuale ICAR e coefficiente non strutturato spazialmente (modello (2.12)) (vedi testo).



* Esclusa la provincia di Arezzo.

Figura 2.13 Registro Tumori Toscano. Incidenza per Tumore del Polmone, Toscana*, maschi, 2004. Effetto *within* per ogni comune stimati dal modello (2.12) (vedi testo).



* Esclusa la provincia di Arezzo.

Capitolo 3

Modelli Bayesiani gerarchici per integrare dati individuali nelle regressioni ecologiche

3.1 Introduzione

La variabilità nella frequenza di malattia tra gruppi di popolazioni è generalmente maggiore della variabilità entro popolazioni (Prentice e Sheppard, 1995). Tuttavia gli studi epidemiologici si limitano a valutare le differenze individuali di rischio di malattia entro popolazione e perdono di potenza nell'identificare le associazioni con i potenziali fattori di rischio.

Prentice e Sheppard (1995) hanno proposto un disegno dello studio e un modello di analisi che utilizzano appropriatamente informazioni a livello aggregato sulla frequenza di malattia e le integrano con informazioni sui fattori di rischio raccolte a livello individuale. Questo perché l'uso di informazioni a livello aggregato sia per le variabili di risposta e che per le covariate non è raccomandabile dato che le stime risulterebbero potenzialmente distorte (fallacia ecologica) (Morgenstern, 2008).

Il modello di Prentice e Sheppard è fondato su una procedura basata sulle equazioni generalizzate di stima (GEE) (Liang e Zeger, 1986) e non è robusto. È stato dimostrato che per alcune combinazioni di covariate le equazioni di stima, applicate a questo modello, non arrivano a convergenza (Lancaster *et al.*, 2006). Per risolvere alcuni di questi problemi sono state sviluppate versioni Bayesiane

gerarchiche del modello, in particolare nel caso in cui i campioni di dati individuali sui fattori di rischio sono di piccola numerosità (Wakefield e Salway, 2001; 2008).

Nelle analisi epidemiologiche che studiano l'associazione tra deprivazione materiale e salute è stata documentata la presenza di un effetto di contesto (ad esempio per il nostro paese Biggeri *et al.*, 1999, Biggeri *et al.*, 2004, Costa *et al.*, 2004) che non viene preso in considerazione nella modellistica citata sopra. L'effetto della povertà sociale ed economica sulle condizioni di salute si esplica infatti anche a livello ecologico, non solo a livello individuale. Ad esempio, la mortalità che osserviamo nelle aree socialmente ed economicamente più degradate è maggiore di quella predetta dalla somma di tutte le condizioni di svantaggio registrate a livello individuale (Riva *et al.*, 2007).

Nel presente lavoro proponiamo un modello Bayesiano gerarchico che integra l'informazione a livello aggregato sulla variabile di risposta (frequenza di decessi) con l'informazione campionaria individuale sui fattori di rischio (deprivazione materiale) e che diversamente dai modelli presenti in letteratura, permette di stimare sia l'effetto individuale che contestuale.

Il modello di regressione proposto mira a studiare la relazione tra deprivazione materiale e mortalità generale a livello italiano. A questo scopo si sono considerati i dati relativi ai certificati di causa di morte ISTAT per gli anni 2000-2004, per tutta Italia a livello provinciale, e per quanto riguarda i fattori socio-economici, i dati individuali dell'Indagine Multiscopo ISTAT del 2000.

Nel paragrafo 3.2 vengono descritti i dati utilizzati. Nel paragrafo 3.3 vengono presentati i metodi statistici utilizzati e sono introdotti i modelli Bayesiani per la stima dell'effetto individuale e dell'effetto contestuale. I risultati, la discussione e le conclusioni sono riportati rispettivamente nei paragrafi 3.4, 3.5 e 3.6.

3.2 Deprivazione materiale e mortalità in Italia 2000-2004

La presenza di un differenziale socio-economico nella mortalità è uno dei problemi maggiori nell'agenda politica dei paesi sviluppati (Mackenbach, 2006). I dati

utilizzati per la stima dei differenziali nel nostro paese provengono sostanzialmente sugli studi longitudinali su base censuaria che sono tuttavia limitati a poche realtà urbane (Kunst e Mackenbach, 1994, Kunst *et al.*, 2001). Nell'ambito del progetto di ricerca sanitaria finalizzata ex articolo 12 "Valorizzazione dei dati del censimento 2001 per il monitoraggio e l'analisi delle diseguaglianze sociali nella salute in Italia" (Ministero della Salute, Regione Autonoma Valle d'Aosta, 2004-2008) ci siamo posti l'obiettivo di quantificare, per il periodo più recente possibile, la relazione tra deprivazione materiale, quale indicatore sintetico di diseguaglianza sociale ed economica, e la mortalità generale per tutta l'Italia.

3.2.1 Dati di Mortalità

I dati di mortalità derivano dall'archivio ISTAT delle certificazioni di causa di morte 2000-2004. Nel presente lavoro è stata analizzata la mortalità per tutte le cause (ICD IX 001-999), maschi e femmine, per un totale di 2.750.013 decessi. Il livello di aggregazione scelto è la provincia. Sono stati calcolati i casi attesi utilizzando la standardizzazione indiretta interna per età e genere (Breslow e Day, 1975). La figura 3.1 mostra la distribuzione dei rapporti standardizzati di mortalità delle 103 province italiane. Il differenziale di mortalità oscilla oggi nel nostro paese tra -12% e +24% considerando la Provincia come livello di aggregazione omogeneo. Si mantengono quindi importanti differenze di mortalità (nel 1975-1977 i rapporti standardizzati di mortalità provinciali oscillavano tra 0,72 e 1,25) ma la distribuzione geografica è cambiata (Cislaghi *et al.*, 1986). Il Nord-est non appare più come l'area a maggiore mortalità. Il Sud e in particolare la Campania e la Sicilia condividono con il Nord-ovest i livelli di rischio più alti (sopra il 6% rispetto alla media nazionale) mentre le regioni del centro Italia e il versante adriatico mostrano una mortalità inferiore alla media (figura 3.2).

3.2.2 Deprivazione

I dati sui fattori socio-economici provengono dall'Indagine Multiscopo ISTAT del 2000 (Istituto Nazionale di Statistica, 2001) che può essere tabulata per Provincia,

mantenendone la rappresentatività. I record si riferiscono a 140.011 individui, per i quali è stato definito un indice di deprivazione come somma degli indicatori di quattro eventi sfavorevoli:

- bassa istruzione (licenza elementare o inferiore),
- disoccupazione (disoccupato o in cerca di prima occupazione),
- abitazione in affitto,
- indice di affollamento (numero di individui per stanza) >1,5.

L'indice può assumere valori tra 0 e 4 a seconda delle condizioni di svantaggio in cui versa l'individuo. Quindi, vale 0 se all'individuo non corrisponde nessuno degli eventi sfavorevoli, vale 1 se l'individuo presenta solo un evento sfavorevole e 2 se l'individuo presenta due eventi sfavorevoli, e così via (Biggeri *et al.*, 2004).

Mediamente si anno 1,15 eventi sfavorevoli per persona (figura 3.3). La deviazione standard assumendo il modello binomiale sarebbe 0,82 mentre quella empirica è 0,86, con una sovradispersione quindi dell'11,2%. Vi è una certa evidenza di disomogeneità nella distribuzione della deprivazione materiale. Considerando il livello provinciale (figura 3.4) troviamo una oscillazione dei valori medi dell'indice con *range* compreso tra 0,74 e 1,62. La dimensione provinciale spiega sostanzialmente la sovradispersione osservata. Infatti, entro provincia la variabilità empirica è sempre inferiore a quella da modello (deviazione standard media di 0,81, l'indice di dispersione oscilla tra 0,57 e 1,00 con un valore medio di 0,82). La correlazione tra media e varianza è 0,34.

In figura 3.6 è riportata la distribuzione geografica dell'indice. Il Sud, alcune province del Centro e il Nord Ovest risultano su livelli superiori alla media nazionale.

La deviazione standard della distribuzione dei valori medi provinciali è 0,17 più di quattro volte inferiore alla deviazione standard media entro Provincia.

La modellistica che proponiamo farà tesoro della variabilità entro provincie per stimare l'effetto della deprivazione materiale sulla mortalità.

3.3 Metodi

I modelli di analisi si riferiscono ai dati aggregati (nell'esempio le Provincie). Coerentemente con Sheppard definiamo questi modelli come ecologici (Sheppard, 2003). Si assuma che il numero di casi osservati nella *j*-esima provincia Y_j segua una distribuzione di Poisson con valore atteso ($E_j\theta_j$), dove E_j sono le frequenze attese calcolate utilizzando la standardizzazione indiretta interna per età e genere, e θ_j è il rischio relativo.

Modello Ecologico

Se si denota con X_{ij} la covariata di interesse, cioè l'indice di deprivazione materiale dell'*i*-esimo individuo residente nella *j*-esima provincia e si indica con μ_{X_j} la media dell'indice X per la *j*-esima provincia, allora il modello di regressione ecologico è definito come:

$$\log\theta_j = \alpha + \beta_A\mu_{X_j} \qquad (3.1)$$

dove α è l'intercetta e β_A è l'effetto ecologico della deprivazione materiale.

Modello per la stima dell'effetto individuale

Se vogliamo considerare il livello individuale, possiamo specificare il modello di regressione ecologico come:

$$\log\theta_j = \alpha + K(\beta)$$

dove $K(\beta)$ è la funzione generatrice dei momenti cumulata della distribuzione entro area della covariata (Richardson *et al.*, 1987), e β indica il generico coefficiente di regressione per la covariata misurata a livello individuale. Se

assumiamo gaussiana la distribuzione della covariata $X_{ij} \sim Normale\left(\mu_{X_j}, \sigma^2_{X_j}\right)$, allora $K(\beta) = \beta\,\mu_{X_j} + (\beta^2\sigma^2_{X_j}/2)$ [1] e il modello di regressione diviene:

$$\log\theta_j = \alpha + \beta\,\mu_{X_j} + \frac{\beta^2}{2}\sigma^2_{X_j}. \qquad (3.2)$$

Non è quindi necessario avere la variabile di risposta misurata a livello individuale per stimare l'effetto individuale, purché si disponga di informazioni sulla distribuzione della covariata.

Modelli per la stima dell'effetto individuale e contestuale

Nella formulazione di Firebaugh (1978) il modello contestuale prevede l'inserimento della media di gruppo come ulteriore covariata nella regressione ecologica:

$$\log\theta_j = \alpha + \left[\beta_I\,\mu_{X_j} + \frac{\beta_I^2}{2}\sigma^2_{X_j}\right] + \beta_C\,\mu_{X_j}$$

dove β_C è l'effetto contestuale, e β_I è l'effetto individuale (*within*).

Chiaramente, β_C e β_I non sono identificabili in quanto la stessa covariata non può informare contemporaneamente sui due parametri. La loro somma è però identificabile. Possiamo quindi riparametrizzare il modello come:

$$\log\theta_j = \alpha + \left(\frac{\beta_I^2}{2}\sigma^2_{X_j}\right) + \left(\beta_C + \beta_I\right)\mu_{X_j}$$

---

[1] Consideriamo i primi due momenti. Per situazioni più complesse dove la distribuzione è asimmetrica si veda Wakefield e Salway (2001).

ovvero:

$$\log \theta_j = \alpha + \beta_A \, \mu_{X_j} + \frac{\beta_I^2}{2} \sigma_{X_j}^2 \qquad (3.3)$$

dove $\beta_A = \beta_I + \beta_C$ è noto come effetto ecologico. Il modello è noto in letteratura come modello di Cronbach (Cronbach e Webb, 1975):

3.3.1 Modelli Bayesiani gerarchici

Modello ecologico

Riconsiderando il modello (3.1) la specificazione del modello Bayesiano avviene elicitando le distribuzioni a priori sui parametri ignoti α e β_A. Il modello gerarchico è definito come:

$$\log \theta_j = \alpha + \beta_A \, x_j \qquad (3.4)$$

$$\alpha \sim Normale(\mu_\alpha, \sigma_\alpha^2)$$
$$\beta_A \sim Normale(\mu_{\beta_A}, \sigma_{\beta_A}^2)$$

dove x_j è la media campionaria dell'indice di deprivazione materiale per la *j*-esima provincia ed è considerata un dato fisso e privo di errore (un nodo deterministico nella formulazione grafica del modello). L'interpretazione dei parametri del modello resta invariata e le coppie di iperparametri $(\mu_\alpha, \sigma_\alpha^2)$ e $(\mu_{\beta_A}, \sigma_{\beta_A}^2)$ sono scelte in modo tale che le a priori su α e β_A siano debolmente informative.

Modelli per la stima dell'effetto individuale

La formulazione Bayesiana del modello (3.2) può avvenire sia seguendo un approccio Bayesiano "empirico" che completo (Carlin e Louis, 2008).

In prima approssimazione siano x_j e s_j^2 rispettivamente media e varianza campionarie della covariata (indice di deprivazione) per la *j*-esima provincia. Un modello Bayesiano "plug-in" è definito analogamente al modello ecologico, come:

$$\log\theta_j = \alpha + \beta\, x_j + \frac{\beta^2}{2} s_j^2 \qquad (3.5)$$

dove le a priori sui parametri α e β del modello sono normali debolmente informative (vedi modello (3.4)) e β è l'effetto della deprivazione a livello individuale. La struttura gerarchica del modello Bayesiano risulta più chiara se si rappresenta il grafo orientato aciclico ad esso associato (figura 3.6).

Questo modello non tiene conto della variabilità in x_j e s_j^2.

Un modello Bayesiano empirico (EB) può essere formulato ipotizzando un livello latente in cui media e varianza campionarie siano parametri su cui vengono definite delle appropriate distribuzioni a priori. Si tratta di applicare una filosofia Bayesiana empirica quando l'informazione campionaria su x_j e s_j^2 entra nella definizione delle distribuzioni a priori:

$$\log\theta_j = \alpha + \beta\, \mu_{X_j} + \frac{\beta^2}{2} \sigma_{X_j}^2 \qquad (3.6)$$

e

$$\mu_{X_j} \sim N\left(x_j, s_j^2 / n_j\right)$$
$$\sigma_{X_j}^2 \sim \chi_{\upsilon_j}^2 \left(s_j^2 / n_j\right).$$

Come nelle precedenti formulazioni le a priori su α e β sono assunte normali debolmente informative. Chiameremo questo modello "plug-in EB". La figura 3.7 mostra il DAG associato.

Nella formulazione completamente Bayesiana si specificano due modelli: un modello per la covariata individuale (*exposure model*) e un modello per la variabile dipendente (*response model*) (Gilks *et al.*, 1996) (modello Two-stage) (figura 3.8).

I parametri del modello di esposizione entrano come covariate nel modello per la risposta secondo il seguente modello gerarchico:

Response model:

$$\log \theta_j = \alpha + \beta\, \mu_{X_j} + \frac{\beta^{\,2}}{2} \sigma^2_{X_j} \tag{3.7}$$

Exposure model

$$X_{ij} \sim Normale\left(\mu_{X_j}, \sigma^2_{X_j}\right)$$

e

$$\mu_{X_j} \sim Normale\left(0{,}100000\right)$$

$$\sigma^2_{X_j} \sim Gamma\left(0.001{,}0.001\right)$$

Le distribuzioni a priori sui parametri α e β sono normali debolmente informative.

Modelli per la stima dell'effetto individuale e contestuale

Nel caso del modello (3.3) media e varianza della distribuzione della covariata entro aggregato entrano nel modello con coefficienti non vincolati. Possiamo formulare il modello come:

$$\log \theta_j = \alpha + \beta_A\, x_j + \frac{\beta_I^2}{2} s_j^2 \tag{3.8}$$

Le distribuzioni a priori sui parametri α, β_I, β_A sono normali debolmente informative. La rappresentazione grafica del modello di Cronbach è riportata nella figura 3.9.

Alternativamente possiamo definire una distribuzione a priori informativa su β_A, $\beta_A \sim Normale(\hat{\mu}_{\beta_A}, \hat{\sigma}^2_{\beta_A})$ dove $\hat{\mu}_{\beta_A}, \hat{\sigma}^2_{\beta_A}$ sono i parametri della distribuzione a posteriori di β_A dal modello ecologico (3.4).

3.3.2 Scelta del modello e aspetti computazionali

Per determinare le performance dei modelli abbiamo calcolato l'EPD (*Expected Predictive Deviance*) nella formulazione di Gelfand e Ghosh (1998) come somma tra una misura della bontà di adattamento del modello ai dati (*Expected Posterior Deviance*, $\overline{D}$) e la complessità del modello, in un senso molto particolare cioè facendo riferimento alla distribuzione predittiva di *Y*.

Per approssimare le distribuzioni a posteriori abbiamo utilizzato il software WinBUGS (Spiegelhalter *et al.*, 2000). Per ogni modello abbiamo definito due catene indipendenti; al fine di saggiare la convergenza dell'algoritmo è stato usato il test di Gelman-Rubin (1992). Abbiamo considerato 100.000 iterazioni come burn-in e utilizzato le successive 50.000 per la stima. I modelli sono complessi e computazionalmente intensivi (nell'ordine di un'iterazione ogni 3 secondi).

3.4 Risultati

I logaritmi dei rischi relativi stimati dal modello ecologico (3.4) e dai modelli (3.5), (3.6) e (3.7) di regressione ecologica per la stima dell'effetto individuale sono riportati nella tabella 3.1. Il confronto tra le stime dei rischi relativi del modello ecologico e dei modelli di regressione ecologica per la stima dell'effetto individuale evidenzia che la stima di β del modello (3.5) coincide praticamente con il rischio relativo "ecologico" stimato dal modello (3.4) ($\hat{\beta} = 0{,}18$ vs $\hat{\beta}_A = 0{,}18$).

L'introduzione della media e varianza campionaria, come quantità fisse non riesce a veicolare l'informazione necessaria sulla variabilità individuale dell'esposizione in studio.

I modelli (3.6) e (3.7) permettono di stimare l'effetto individuale. Il modello (3.7) permette inoltre di considerare appropriatamente più fonti di incertezza, l'errore standard stimato per l'effetto individuale è più elevato rispetto alla stima ottenuta dal modello (3.6) (rispettivamente pari a 0,0106 e 0,0076).

Il modello di Cronbach, i cui risultati sono riportati nella tabella 3.2, permette di stimare sia l'effetto individuale che ecologico. Il logaritmo del rischio relativo individuale è risultato $\hat{\beta}_I = 0{,}46$ mentre l'effetto ecologico $\hat{\beta}_A = 0{,}16$ se si definiscono distribuzioni a priori debolmente informative sui parametri del modello. I risultati non sono molto differenti se imponiamo una distribuzione a priori informativa sull'effetto ecologico ($\hat{\beta}_I = 0{,}56$, $\hat{\beta}_A = 0{,}18$).

Dai risultati appena descritti si evince che il livello provinciale porta a diluizione dell'effetto. Se prendiamo in considerazione le rispettive deviazioni standard si nota che la variabilità tra Province è 4,7 volte minore rispetto alla variabilità entro provincia ($ds(\bar{x}) = 0{,}171$, $\overline{ds}(x) = 0{,}806$). Se calcoliamo l'effetto per unità di deviazione standard[2] otteniamo: ${}^s\hat{\beta}_I \approx 0{,}45$ vs ${}^s\hat{\beta}_A \approx 0{,}03$.

Per quanto riguarda il confronto tra i modelli (tabella 3.3) l'EPD favorisce il modello di Cronbach (rispettivamente 1005 e 1006 per il modello (3.8) e (3.9)).

3.5 Discussione

L'indice di deprivazione è stato assunto come variabile continua e nella modellazione probabilistica abbiamo specificato una distribuzione Gaussiana.

[2] La stima si ottiene moltiplicando il coefficiente $\hat{\beta}$ con la deviazione standard della corrispondente variabile esplicativa.

Questa assunzione può essere criticata. La correlazione empirica tra media e varianza è abbastanza modesta (0,34). Gli eventi sfavorevoli che compongono l'indice sono considerati tra loro indipendenti, e vi sono evidenze empiriche che questa assunzione sia valida (Biggeri e Grisotto, 2009). La probabilità per ciascun evento sfavorevole è però plausibile che sia differente e questo potrebbe spiegare la sottodispersione osservata entro provincia.

Uno sviluppo dei modelli per la stima dell'effetto individuale e contestuale consiste nello specificare un modello di errore di misura per l'indice di deprivazione $\{X_{ij}\}$, assumendo una distribuzione binomiale($n = 4$, p_{ij}) per $\{X_{ij}\}$ e formulando un modello Gaussiano per una variabile latente $\text{logit}(p_{ij})$.

La grande dimensione campionaria nella nostra applicazione rende abbastanza piccolo l'impatto di queste scelte modellistiche a fianco di un grande carico computazionale (Wakefield e Salway (2001)).

Il profilo di verosimiglianza bivariato per i parametri β_I e β_A del modello di Cronbach (figura 3.10) mostra chiaramente che i dati in esame non supportano valori diversi da 0,18 per modelli che impongano un vincolo tra i due coefficienti. Confinare o addirittura fissare i valori di μ_{X_j} e $\sigma^2_{X_j}$ per mezzo di distribuzioni a priori troppo informative finisce per restringere lo spazio parametrico in una regione di non identificabilità (modello (3.5)) la elicitazione di *prior* meno vincolanti (modello (3.6)) o per mezzo del modello sull'esposizione (modello (3.7)) è una soluzione migliore sotto il profilo della identificabilità (Gustafson, 2005)

L'EPD è giustificato dalla considerazione che l'adeguatezza di un modello dovrebbe essere valutata tramite la fedeltà di repliche simulate a partire dalle distribuzioni a posteriori dei parametri del modello, ai dati originariamente osservati (Gelman *et al.* 2002). È stato notato che l'EPD tende a penalizzare poco i modelli più complessi (Spiegelhalter *et al.*, 2001), ma nel nostro esempio l'introduzione di indici più severi non cambierebbe l'ordinamento.

L'effetto contestuale nei modelli fin qui considerati è confuso con l'eventuale variazione del livello medio (*baseline*) di mortalità provinciale. Quest'ultimo può riflettersi in una sovradispersione poissoniana. Non sorprende quindi che ogni modello che includa la media provinciale dell'indice di deprivazione con il

coefficiente di regressione parametrizzato come funzione dell'effetto contestuale abbia una miglior bontà di adattamento. In alternativa si potrebbe tentare di introdurre una serie di intercette casuali $\{\alpha_j\}$, eventualmente spazialmente strutturate (Best *et al.*, 2001).

Estensioni del modello di Cronbach che considerano il modello di esposizione pongono problemi di identificabilità non facili da risolvere. È interessante notare che hanno un forte incremento come bontà di adattamento, $\overline{D}$ dell'ordine di 1:5 rispetto al modello (3.6), (3.7).

Si deve considerare che trattandosi comunque di dati aggregati l'informazione effettiva sugli effetti individuali non è mai grande (Sheppard, 2003) e dipende dal particolare insieme di dati analizzato. Inoltre, l'efficienza è maggiore sotto i modelli (3.6) e (3.7) che sono derivati dal modello che tiene conto delle distorsioni da specificazione (Wakefielde e Salway, 2001, Richardson *et al.*, 1987) quando è valida l'assunzione di assenza di altre tipologie di distorsione ecologica.

Nel nostro esempio non è così e i modelli (3.6) e (3.7) hanno una peggiore bontà di adattamento rispetto ai modelli di Cronbach.

Non possiamo escludere ed è anzi lo scopo di sviluppi ulteriori di ricerca, che sia possibile innestare questi modelli in una classe più ampia. Mentre l'estensione dei modelli di Cronbach (3.8) e (3.9) con il livello relativo all'*exposure model* o anche al modello di errore di misura porta si ad un incremento di bontà di adattamento ma anche a una non identificabilità (Gelfand e Sahu, 1999), pensiamo che l'introduzione di una ulteriore classe di termini casuali $\{\alpha_j\}$ in questa situazione potrebbe portare a una soluzione. Gustafson (2005) discute appunto un problema relativo a un modello di errore di misura in cui la mancata identificabilità è risolta specificando un modello più complesso. Riteniamo di non essere lontani da questa situazione quando covariate specifiche per Provincia sono confuse con intercette casuali. Non introdurre quest'ultime nel modello può essere una soluzione ma con modelli più complessi perdiamo la possibilità di interpretare correttamente le stime.

3.6 Conclusioni applicative

Nel presente lavoro abbiamo mostrato che anche in Italia vi è un effetto importante della deprivazione materiale (RR=1.58÷1.75) sulla mortalità. Tale stima risente probabilmente del livello di aggregazione utilizzato (la Provincia), talmente ampio che l'effetto individuale è plausibile che comprenda anche l'effetto contestuale di livelli inferiori. Non può neanche essere esclusa una sovrastima legata all'effetto di confondenti non misurati e spazialmente strutturati. I risultati sono comunque coerenti con i dati in letteratura (ad esempio Carstairs 1995 mostra variazioni di SMR da -20% a +40%).

Linee di lavoro future riguardano la identificabilità degli effetti (Xie e Carlin, 2006) e l'informazione sugli effetti individuali contenuta nei dati aggregati, estendendo il lavoro di Sheppard (2003) ai modelli Bayesiani.

3.7 Bibliografia

Best N, Cocking S, Bennett J, Wakefield J, Elliott P (2001). Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society A*, **164**, 155-174.

Biggeri A, Dreassi E, Marchi M (2004). A multilevel Bayesian model for contextual effect of material deprivation. *Statistical Methods & Applications*, **13**, 87–101

Biggeri A, Grisotto L (2009). Fonti di distorsione nella misura delle disuguaglianze di salute: la validazione, il confronto temporale e spaziale, l'aggiustamento per altre covariate, il bias ecologico. *Salute e Società*, (in fase di stampa)**.**

Biggeri A, Marchi M, Dreassi E, Baldi P, Benvenuti A, Merler E (1999). Studio Longitudinale Toscano: un'analisi per quartiere per le città di Firenze e Livorno. *Epidemiologia & Prevenzione*, **23**, 161-174.

Carlin PB, Louis TA (2008).Bayesian methods for data analysis – 3° edition. Chapman & Hall/CRC. Boco Raton.

Carstairs V (1995). Deprivation indices: their interpretation and use in relation to health. *Journal of Epidemiology and Community Health*, **49** (Suppl 2) S3-S8.

Cislaghi C, Decarli A, La Vecchia C, Laverda N, Mezzanotte G, Smans M (1986). Dati, indicatori e mappe di mortalità tumorale. Italia 1975-1977. Pitagora Editrice. Bologna.

Costa G, Spadea T, Cardano M (2004). (a cura di) Diseguaglianze di salute in Italia. *Epidemiologia & Prevenzione*, **28** (3).

Cronbach LJ, Webb J (1975). Between-class and within-class effects in a reported aptitude X treatment interaction. *Journal of Educational Psychology*, **67**, 6:717.

Firebaugh G (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*. **43**, 557-572.

Gelfand AE, Ghosh SK (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, **85**, 1–11.

Gelfand AE, Sahu SK (1999). Identifiability, improper prios, an Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, **94**, 445, 247-253.

Lancaster GA, Green M, Lane S (2006) Reducing bias in ecological studies: an evaluation of different methodologies. *Journal of the Royal Statistical Society A*, **169**, 4, 681-700

Liang KY, Zeger J (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Mackenbach JP (2006) Health inequalities: Europe in profile. *Expert report by UK Presidency of the EU*.

Marinacci C, Spadea T, Biggeri A, Demaria M, Caiazzo A, Costa G (2004) The role of individual and contextual socioeconomic circumstances on mortality: analysis of time variations in a city of north west Italy. *Journal of Epidemiology and Community Health*, **58**, 199-207.

Ministero della Salute, Regione Autonoma Valle d'Aosta (2004-2008) Ricerca ex art. 12 "Valorizzazione dei dati del censimento 2001 per il monitoraggio e l'analisi delle diseguaglianze sociali nella salute in Italia". Responsabile Scientifico A. Biggeri.

Morgenstern H (2008) Ecologic Studies. In Rothman KJ, Greenland S Modern Epidemiology – 3° edition. Lippincott Williams & Wilkins, Philadelphia.

Prentice R L, Sheppard L (1995) Aggregate data studies of disease risk factors. *Biometrika*, **82**, 1, 113-25.

Richardson S, Stucker I, Hémon D (1987) Comparison of relative risks obtained in ecological and individual studies: some methodological consideration. *International Journal of Epidemiology*, **16**, 111-120.

Riva M, Gauvin L, Barnett TA (2007) Toward the next generation of research into small area effects on health: a synthesis of multilevel investigations published since July 1998. *Journal of Epidemiology and Community Health*; 61:853–861.

Sheppard L (2003), Insight on bias and information in group-level studies. *Biostatistics*, **4**, 2, 265-278.

Wakefield J, Salway R (2001) A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society A*, 164 Part 1, 119-137.

Wakefield J, Salway R (2008) A hybrid model for reducing ecological bias. *Biostatistics*, **9**, 1, 1-17.

Xie Y, Carlin BP (2006). Measures of Bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, **136**, 3458-3477.

Tabelle e figure

Tabella 3.1 Mortalità per tutte le cause, maschi e femmine, 2000-2004. Italia. Modelli Bayesiani gerarchici. Coefficienti di regressione per l'indice di deprivazione materiale β_A = effetto ecologico, *between*, β = effetto individuale, totale(vedi testo).

Modello:	β_A	β
(3.4) Plug-in	0.18 (0.0033)	
(3.5) Plug-in		0.18 (0.0033)
(3.6) Plug-in EB		0.42 (0.0076)
(3.7) Two Stage		0.42 (0.0106)

Tabella 3.2 Mortalità per tutte le cause, maschi e femmine, 2000-2004. Italia. Modelli Bayesiani gerarchici per la stima dell'effetto contestuale. Coefficienti di regressione per l'indice di deprivazione materiale β_A = effetto ecologico, *between*, β_I = effetto individuale, *within* (vedi testo).

Modello:	*(3.8)*	*(3.9)*
	Prior	
	Non informativa	Informativa
β_I	0.46 (0.0101)	0.56 (0.0105)
β_A	0.16 (0.0035)	0.18 (0.0044)

Tabella 3.3 Mortalità per tutte le cause, maschi e femmine, 2000-2004. Italia. Modelli Bayesiani gerarchici. Indici di bontà di adattamento. $\overline{D}$ =Expected Posterior Deviance, EPD= Expected Predictive Deviance (vedi testo).

Modello:	$\overline{D}$	*EPD*
(3.4)	1005	1015
(3.5)	1002	1012
(3.6)	1614	1712
(3.7)	1370	1470
(3.8)	995	1005
(3.9)	995	1006

Figura 3.1 Mortalità per tutte le cause, maschi e femmine, 2000-2004. Italia. Istogramma dei rapporti standardizzati di mortalità.

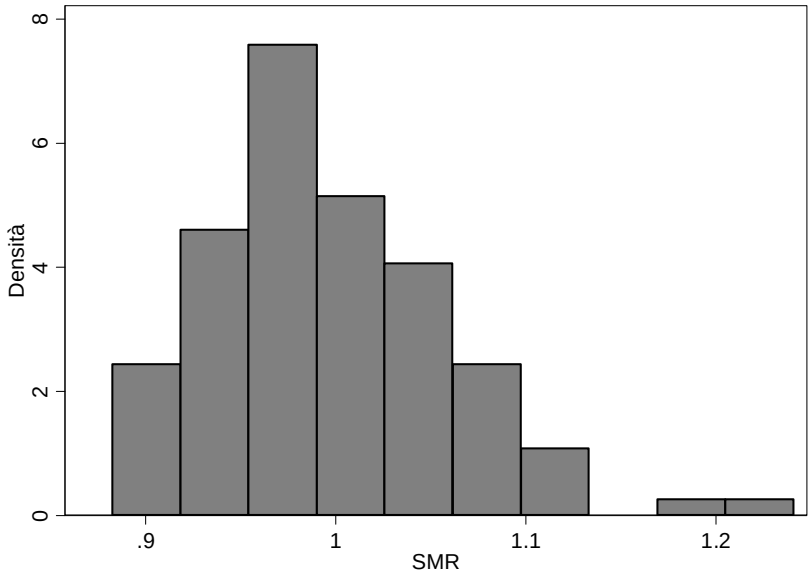


Figura 3.2 Mortalità per tutte le cause, maschi e femmine, 2000-2004. Italia. Distribuzione spaziale dei rapporti standardizzati di mortalità a livello provinciale.

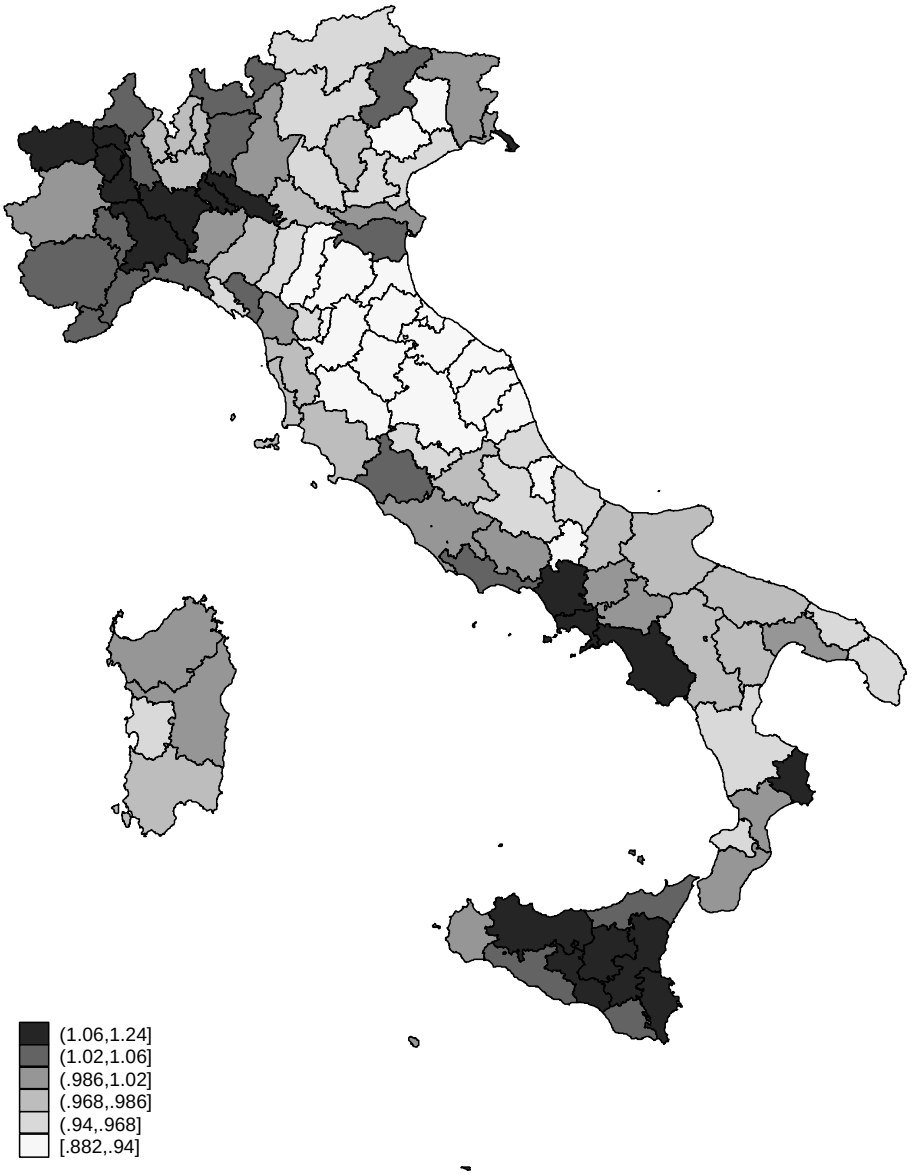


Figura 3.3 Distribuzione dell'indice di deprivazione individuale. Indagine Multiscopo ISTAT 2000.

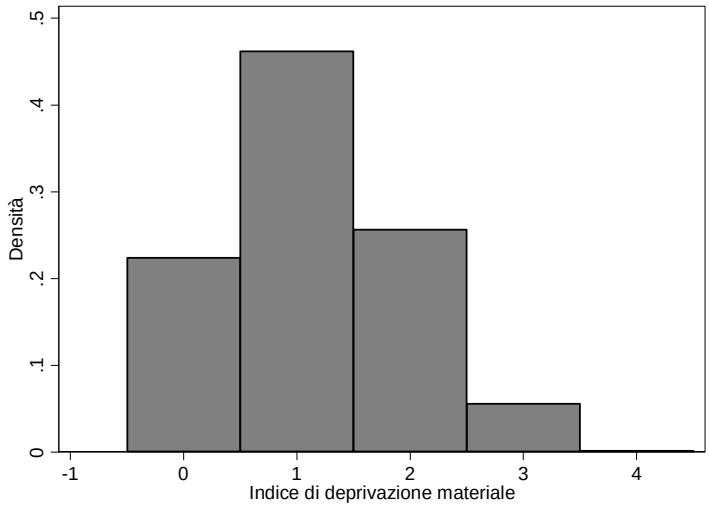


Figura 3.4 Distribuzione dell'indice di deprivazione materiale medio provinciale. Indagine Multiscopo ISTAT 2000.

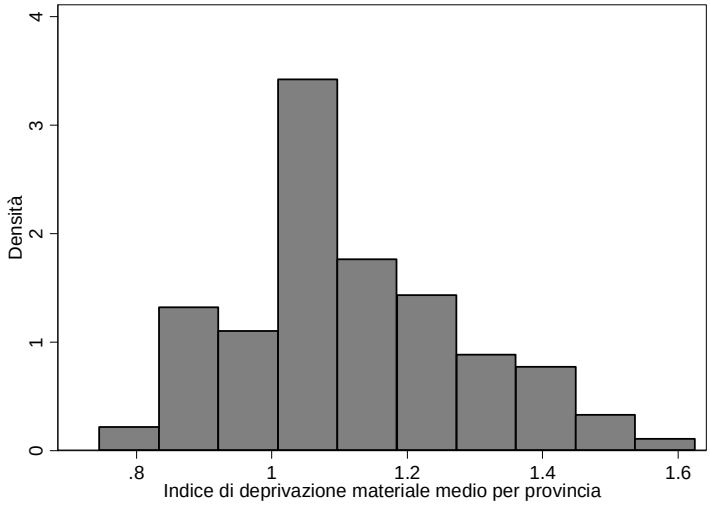


Figura 3.5 Distribuzione spaziale della deprivazione materiale per provincia. Indagine Multiscopo ISTAT 2000.

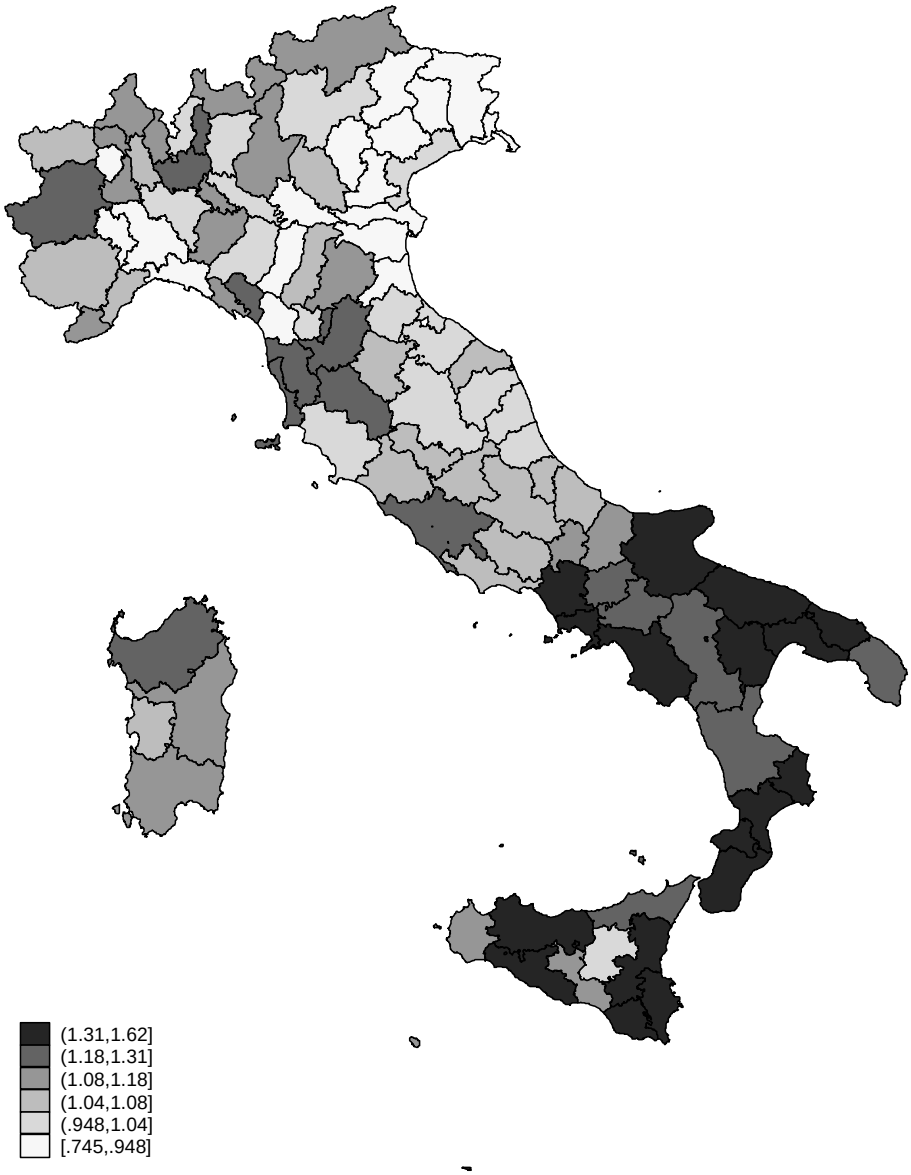


Figura 3.6 Grafo orientato aciclico associato al modello (3.5).

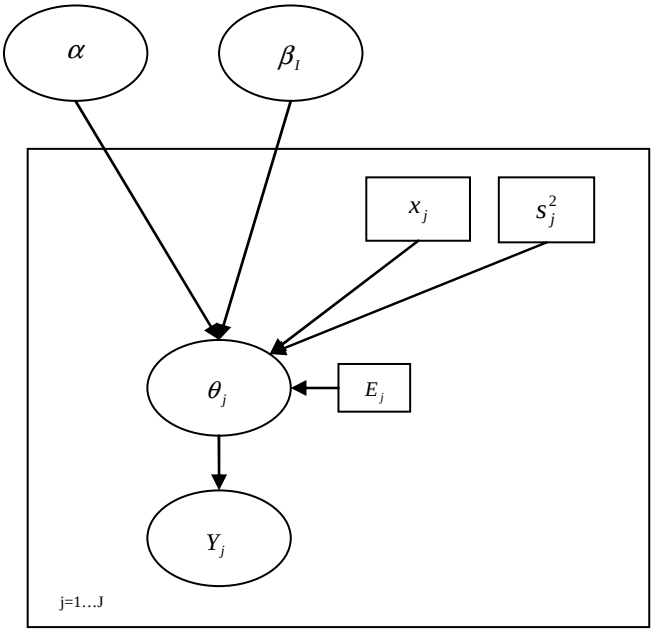


Figura 3.7 Grafo orientato aciclico associato al modello (3.6).

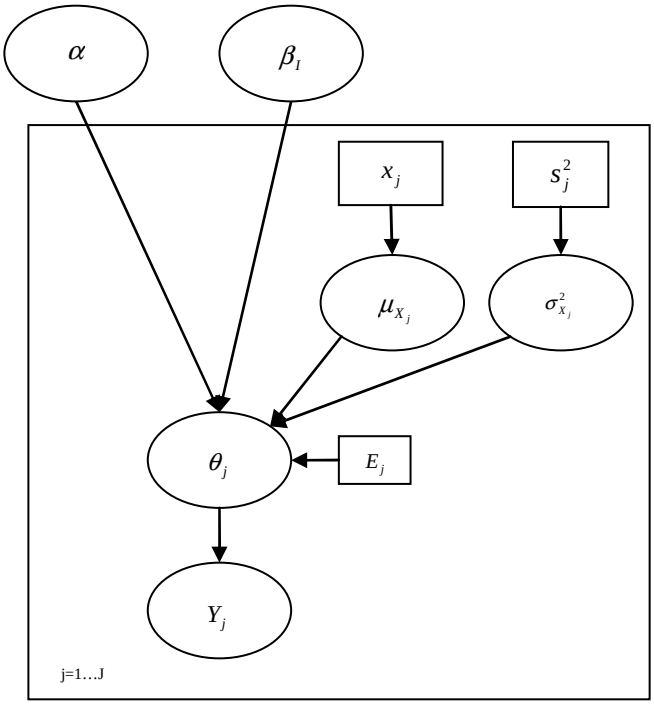


Figura 3.8 Grafo orientato aciclico associato al modello (3.7).

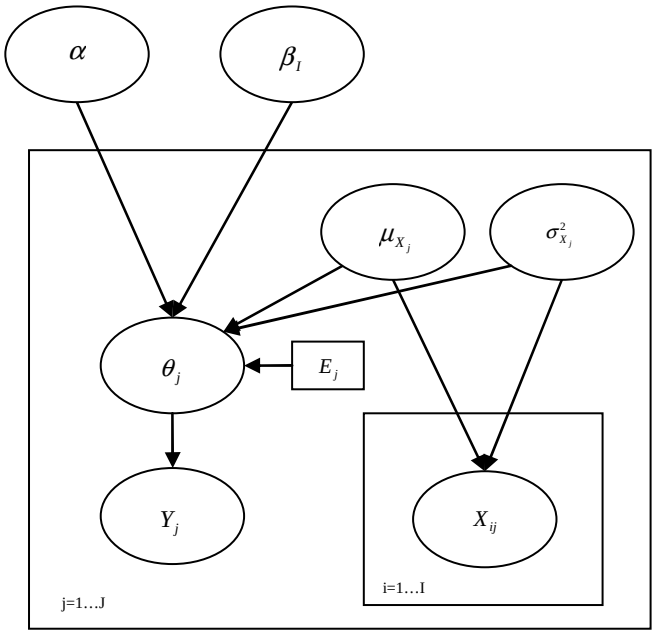


Figura 3.9 Grafo orientato aciclico associato ai modelli (3.8) (3.9).

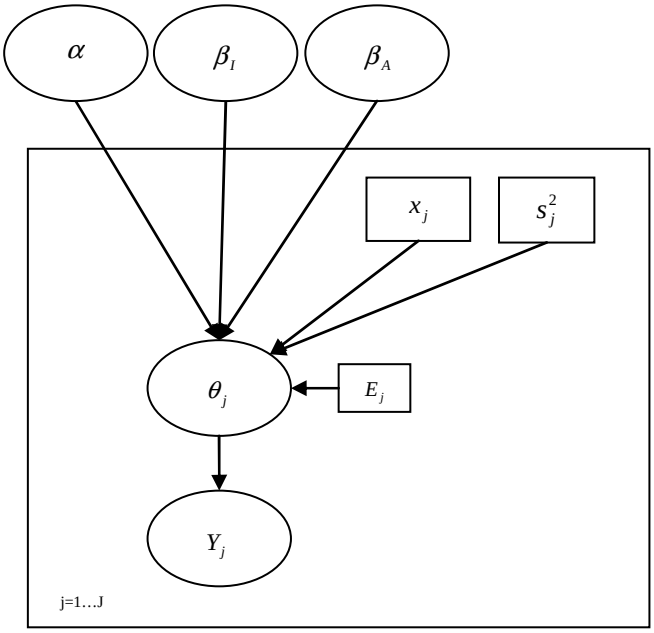
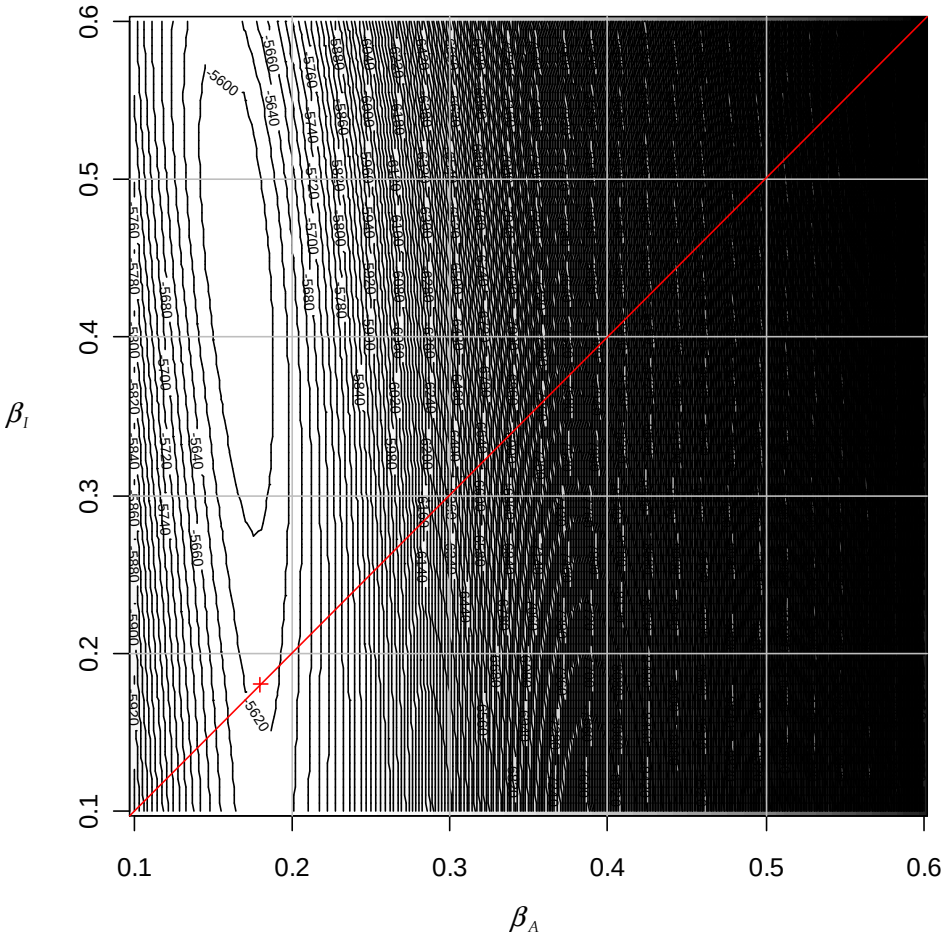


Figura 3.10 Profilo di verosimiglianza bivariato del modello (3.8). β_A **= effetto ecologico,** ***between*****,** β_I **= effetto individuale,** ***within*** **(vedi testo).**



Conclusioni

Il lavoro qui presentato è una collazione di tre distinti lavori metodologici con forte valenza applicativa.

Lo studio della relazione tra fattori socio-economici e salute è complicato dalla presenza di effetti indiretti e mediati. Modelli di analisi che tengano conto della struttura gerarchica dei dati sono pertanto necessari per valutare il ruolo di variabili contestuali. Quando usiamo dati aggregati alcuni effetti sono stimabili ma l'inferenza è debole essendo potenzialmente soggetta alla fallacia ecologica.

Abbiamo considerato l'indice di deprivazione materiale come sintesi dei fattori socio-economici in tutte le analisi presentate. Non sono state invece considerate altre variabili perché abbiamo privilegiato lo sviluppo di metodi di analisi. Eventuali altre covariate possono essere facilmente incluse nei modelli di regressione sviluppati.

Il capitolo 1 mostra gli errori che si compierebbero qualora si volesse includere l'indice di deprivazione nelle formula allocative di riparto del Fondo Sanitario senza un'adeguata standardizzazione per età. Dalla comparazione di una serie di modelli sia per dati individuali che per dati aggregati è stata quantificata la distorsione e sono suggerite alcune soluzioni.

Il capitolo 2 presenta un'analisi ecologica a livello di sezione di censimento. Per la prima volta viene presentata un'analisi su una grande popolazione che permette di valutare la bontà dell'uso surrogato del dato medio di piccola area al posto del dato individuale. Sono stati sviluppati modelli Bayesiani gerarchici per stimare i parametri di interesse tenendo conto di variabili di confondimento latenti, che sono molto plausibili data l'ampiezza dell'area complessivamente analizzata.

Il capitolo 3 affronta il problema dell'integrazione di dati a differente livello di aggregazione. Informazioni aggregate sulla variabile di risposta sono integrate con

informazioni provenienti da studi campionari individuali sulle covariate e i determinanti in studio. Il disegno è altamente innovativo e la serie di modelli descritti in letteratura è stata completata con proposte originali. Nell'applicazione presentata per la prima volta a copertura nazionale sono state prodotte stime dell'effetto della deprivazione materiale sulla mortalità generale. del livello di aggregazione utilizzato (la Provincia è un aggregato piuttosto ampio per cui l'effetto individuale potrebbe comprendere anche l'effetto contestuale di livelli inferiori), e di una possibile sovrastima legata all'effetto di confondenti non misurati e spazialmente strutturati.