Università degli studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

Modellazione bayesiana spazio-temporale della diffusione del Covid-19 in Italia

Relatore: Prof. Nicola Sartori
Dipartimento di Scienze Statistiche

Laureando: Marco Franchin
Matricola n°: 1202694

Anno Accademico 2019/2020

A me e a chi in me ha creduto.
Non conta battere gli altri, ma vincere se stessi.

Indice

Elenco delle tabelle

Introduzione

Dal 23 febbraio 2020 in Italia si sono iniziati a rilevare un gran numero di soggetti infettati dal virus SARS-CoV-2. La malattia a lui legata, il Covid-19, ha colpito tutta la nazione e si è diffusa in diversi altri stati in tutto il mondo. Questo evento ha modificato radicalmente la vita della popolazione dato che si è dovuto attuare un confinamento forzato per evitare che il contagio crescesse sempre di più in forma esponenziale. Da un punto di vista statistico può essere d'interesse riuscire ad analizzare questi dati tramite una modellazione opportuna. Si cerca di comprendere al meglio la dinamica della diffusione sfruttando questo fenomeno fortunatamente raro.

In questa tesi si andrà ad applicare una serie di modelli spazio-temporali sui dati del contagio del Covid-19 in Italia, volendo valutare l'adeguatezza di questi strumenti e analizzando la distribuzione del rischio di contagio da un punto di vista spazio-temporale. Successivamente, si osserverà l'impatto di alcuni fattori di rischio di tipo socio-economico e si valuterà la capacità previsiva di questi modelli.

I modelli spazio-temporali sono utilizzati nel caso in cui si vogliano analizzare dei dati che presentano una localizzazione geografica e una evoluzione temporale temporale. Obiettivo principale di questi strumenti è quello di considerare le diverse forme di dipendenza intrinseche nelle osservazioni, ovvero quella spaziale e quella temporale. Il loro utilizzo principale risiede nel *disease mapping*, grazie al quale si vuole valutare se dei fattori di rischio si possono considerare influenti sulla diffusione di una malattia. Oltre a questo, essi valutano come il rischio del contagio si distribuisca nelle diverse zone dell'area geografica d'interesse.

Dovendo utilizzare delle particolari forme di dati spaziali, l'uso della statistica bayesiana risulta conveniente sotto diversi aspetti, in particolare nel caso in cui le osservazioni indicano l'incidenza di un fenomeno in un'area delimitata. Nell'approccio classico si dovrebbero definire un gran numero di parametri per descrivere le relazioni fra le varie aree, mentre il paradigma bayesiano alleggerisce questa struttura utilizzando delle variabili casuali. La stima delle distribuzioni di queste variabili, sviluppata tramite simulazioni o approssimazioni, consente di non dover affrontare problemi numerici in cui si può incorrere nella stima frequentista dei parametri nei casi di alta dimensionalità. Questa metodologia evita anche l'uso di strumenti più complessi in cui la dipendenza fra le osservazioni deve essere esplicitata direttamente nella verosimiglianza. Infine la stima di una intera distribuzione di probabilità fornisce delle potenzialità maggiori dal punto di vista inferenziale come si vedrà nel seguito della tesi. Verranno quindi utilizzate delle metodologie legate al paradigma bayesiano per la stima di questi modelli. Per attuare l'inferenza in questo contesto non si utilizzeranno le classiche metodologie basate sulla simulazione, ma si introdurrà un metodo basato su approssimazioni analitiche chiamato INLA.

Il lavoro è suddiviso in quattro capitoli. Nel primo si andrà a illustrare il paradigma dell'inferenza bayesiana, spiegandone le caratteristiche principali dal punto di vista teorico, partendo dal Teorema di Bayes fino alla specificazione di un modello gerarchico. Infine si discuterà degli aspetti computazionali, legati all'utilizzo dell'inferenza bayesiana.

Nel Capitolo 2 si introdurrà INLA, una procedura basata su approssimazioni analitiche per attuare inferenza bayesiana nel caso di modelli gerarchici con distribuzione normale degli effetti latenti. Questo strumento è alternativo ai metodi basati sulla simulazione comunemente utilizzati. Nel Capitolo 3 si introdurranno i dati spaziali con le loro principali caratteristiche. Successivamente si presenteranno i diversi modelli spaziali e spazio-temporali per alcune tipologie di questi dati. Infine nel Capitolo 4 si mostreranno le applicazioni di quanto presentato utilizzando i dati della diffusione del Covid-19 in Italia, adattando i modelli e effettuando uno studio sulla capacità predittiva dei modelli.

Capitolo 1

Statistica bayesiana

In questo capitolo viene presentato il paradigma dell'inferenza bayesiana, il quale considera ogni quantità coinvolta nei modelli, osservata o meno, come una variabile casuale. La scelta dell'uso dell'approccio bayesiano è legata alla maggior praticità e immediatezza nello svolgere l'operazione di inferenza sui parametri, andando a sfruttare diverse tecniche di simulazione o di approssimazione che possono fornire con buona accuratezza una stima delle densità. Su queste densità si può calcolare qualsiasi quantità d'interesse, come la probabilità in eccesso rispetto a un valore o qualsiasi altra misura di sintesi. Il vantaggio di tale approccio è legato anche all'utilizzo del modello per previsioni, dato che si otterrà una distribuzione predittiva sulla quale si potranno effettuare calcoli probabilistici.

Nell'ambito dei modelli gerarchici, trattare i parametri come variabili aleatorie può portare a dei vantaggi nella loro stima. In particolare, anche in presenza di poca informazione derivante da osservazioni di numerosità esigua il metodo bayesiano fornisce stime sufficientemente stabili andando a sfruttare altre osservazioni utilizzando il principio del *borrowing strengh* (Ancelet *et al.*, 2012). Infine la costruzione gerarchica che può esserci fra i vari parametri è rappresentata dalla dipendenza condizionata che può essere sfruttata in questo ambito con tecniche di stima che verranno presentate in seguito. La stesura di questo capitolo è basata principalmente sui testi Liseo (2010) e Lawson (2018). Si rimanda a questi testi per approfondimenti nel dettaglio.

1.1 Teorema di Bayes

Il teorema di Bayes è la base fondamentale dell'intera statistica bayesiana. Esso mostra la filosofia di questo approccio, fornendo una visione alternativa di come si può trarre informazione dagli eventi. In particolare sintetizza come razionalmente si possano aggiornare le proprie conoscenze in modo empirico tramite l'osservazione di nuovi eventi.

Teorema di Bayes. *Sia E un evento contenuto in $F_1 \cup ... \cup F_k$ dove gli eventi F_i con $i = 1, ...k$, sono a due a due indipendenti e necessari. Allora si ha*

$$Pr(F_i|E) = \frac{Pr(E|F_i)Pr(F_i)}{\sum_{i=1}^{k} Pr(E|F_i)Pr(F_i)}. \tag{1.1}$$

Come si evince dal teorema, dato che l'evento F_i dipende dall'evento E la sua probabilità dovrà essere aggiornata rispetto alla informazione portata da esso. E non solo, essa dipenderà anche da tutta l'informazione pregressa in nostro possesso.

Il membro di sinistra della (1.1) è chiamato **probabilità a posteriori** dell'evento F_i perché esso deriva dopo l'osservazione dell'evento E, mentre nel membro di destra sono presenti due parti: il denominatore è detto costante di normalizzazione, mentre il numeratore rappresenta la probabilità congiunta dell'informazione passata sull'evento F_i e l'evento E. Questa è costituita da $Pr(F_i)$, chiamata **probabilità a priori**, e da $Pr(E|F_i)$ che, dà una indicazione di quanto verosimile sia l'evento E condizionatamente all'evento F_i d'interesse. Quest'ultima è chiamata **verosimiglianza** come verrà spiegato della sezione successiva.

1.2 Verosimiglianza e distribuzione a priori

Per utilizzare il teorema di Bayes in ambito inferenziale, per prima cosa è necessario definire il **modello statistico**. Indicando con y i dati osservati, si assume che questi siano una realizzazione di una variabile casuale Y con densità $p(y;\theta)$, dove θ è un parametro ignoto. Il modello statistico è quindi $\mathcal{F} = \{p(y;\theta), \theta \in \Theta \subseteq \mathcal{R}^p\}$, dove Θ è detto spazio parametrico. La funzione

di probabilità è $p(y;\theta)$ viene considerata nota a meno del parametro θ. La caratteristica principale nel contesto bayesiano è che il parametro viene assunto a sua volta come una variabile aleatoria, quindi Θ sarà uno spazio di probabilità. Al parametro viene associata una densità, detta **distribuzione a priori** e indicata con $\pi(\theta)$, che riassume l'informazione preliminare su θ. L'intero modello statistico in questo caso si baserà su $\mathcal{F}$ e su $\pi(\theta)$. La distribuzione a priori $\pi(\theta)$ generalmente possiede anch'essa dei parametri, definiti iperparametri, i quali possono essere noti o meno, il caso in cui essi siano considerati a loro volta come variabili casuali verrà analizzato nel Paragrafo 1.4.

In generale, con θ si definirà sempre una variabile casuale continua. Le stesse considerazione, con i dovuti aggiustamenti, si possono generalizzare anche nel caso discreto.

L'informazione a priori sul parametro è data dalla sua distribuzione a priori, mentre la conoscenza che ci viene data dalle osservazioni deriva dalla, $L(\theta;y)$, **verosimiglianza**, la quale è data da

$$L(\theta;y) = p(y;\theta), \tag{1.2}$$

considerando y come fissato. Nel caso in cui y sia costituito da n componenti indipendenti la verosimiglianza ha la forma

$$L(\theta;y) = \prod_{i=1}^{n} p(y_i;\theta).$$

Si nota quindi come in un contesto di inferenza bayesiana vi siano due protagonisti principali, la verosimiglianza che esprime l'informazione derivante dai dati y e la distribuzione a priori che esprime una conoscenza precedente all'esperimento statistico. La scelta dell'informazione a priori risulta quindi cruciale per l'inferenza in quanto può farne variare i risultati a parità di y. La soggettività che si inserisce nel modello viene a volte considerata problematica dal punto di vista della comunicazione scientifica: lo stesso risultato sperimentale potrebbe infatti condurre a conclusioni inferenziali sostanzialmente diverse, qualora le informazioni a priori introdotte nel modello fossero anche parzialmente differenti. Di contro, l'influenza dell'informazione a priori sull'oggettività dei dati non è generalmente così determinante, in quanto a fronte di una dimensione campionaria elevata, le distribuzioni a posteriori

relative a due diverse distribuzioni iniziali risulteranno più simili rispetto alle distribuzioni a priori.

L'applicazione statistica del teorema di Bayes avviene nel contesto di un modello statistico parametrico in cui il parametro non viene più considerato come una costante non osservabile, ma come una variabile aleatoria. Questa visione cambia radicalmente il metodo di inferenza su tale parametro rispetto all'inferenza di tipo frequentista. Si deve infatti andare a stimare la sua distribuzione dopo che si sono osservati i dati, descritta come $\pi(\theta|y)$. Si arriva quindi alla seguente formulazione

$$\pi(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{\int_\Theta p(y|\theta)\pi(\theta)d\theta} = \frac{L(\theta;y)\pi(\theta)}{\int_\Theta L(\theta;y)\pi(\theta)d\theta}. \tag{1.3}$$

Si possono riconoscere degli elementi descritti nel paragrafo precedente. Infatti $\pi(\theta|y)$ prende il nome di **distribuzione a posteriori** del parametro d'interesse θ, $\pi(\theta)$ è la sua distribuzione a priori. Visto che il membro di destra della (1.3) presenta una costante di normalizzazione che non dipende dal parametro θ, data la presenza dell'integrale in Θ, la distribuzione a posteriori si esprime anche come $\pi(\theta|y) \propto L(\theta;y)\pi(\theta)$.

Di seguito verranno accennate diverse tipologie di distribuzioni a priori tra le quali le distribuzioni a priori non informative, che creano un ideale punto d'incontro tra la metodologia frequentista e bayesiana.

1.2.1 Distribuzioni a priori coniugate

Il problema che per molto tempo ha limitato l'uso dell'inferenza bayesiana è legato al fatto che essa comporta il calcolo di integrali che possono risultare molto complessi o impossibili da ottenere in modo analitico. Dunque in molti casi, senza ricorrere ad approssimazioni, risulta complicato riuscire a trovare la distribuzione a posteriori dei parametri d'interesse soprattutto se essa non presentava una forma funzionale nota. Per risolvere tale problematica si può ricorrere all'uso di distribuzioni a priori tali per cui la distribuzione a posteriori possiede la medesima forma. Queste distribuzioni a priori si definiscono **a priori coniugate** rispetto alla verosimiglianza $L(\theta;y)$. Ciò va a significare che, dato un modello statistico, si possono identificare delle distribuzioni

a priori che fanno sì che la distribuzione a posteriori abbia la stessa forma funzionale con un opportuno aggiornamento dei parametri.

La scelta degli iperparametri della distribuzione a priori può essere soggettiva, in modo tale che si possa inserire l'informazione in possesso dello sperimentatore all'interno del processo inferenziale.

1.2.2 Distribuzioni a priori non informative

È possibile che lo sperimentatore non possieda una particolare informazione a priori riguardante l'esperimento, ma comunque l'inferenza bayesiana può essere utilizzata attraverso delle distribuzioni a priori non informative. Ci sono diversi modi per definire delle distribuzioni a priori non informative. L'idea generale è quella di minimizzare la sua influenza sulla distribuzione a posteriori. Questo viene fatto spesso nel caso in cui non si voglia influenzare eccessivamente l'informazione fornita dai dati stessi, ma potendo utilizzare lo stesso i vantaggi di considerare il parametro come una variabile aleatoria. L'uso di queste distribuzioni a priori è idealmente un punto d'incontro tra l'approccio frequentista e quello bayesiano, dato che l'informazione è principalmente apportata dalla verosimiglianza, ma viene mantenuto l'impianto probabilistico della statistica bayesiana.

Una particolare distribuzione a priori non informativa è quella di Jeffreys

$$\pi^J(\theta) \propto |i(\theta)|^{\frac{1}{2}}.$$

dove $i(\theta)$ definisce la matrice d'informazione attesa di Fisher (si veda ad esempio Pace e Salvan, 2001 Cap. 3. In questa formulazione si va direttamente ad utilizzare l'informazione presente nel modello per descrivere la distribuzione a priori, inoltre essa possiede la proprietà di invarianza rispetto a riparametrizzazioni.

Un'altra caratteristica delle distribuzioni non informative è che esse possono essere improprie, ovvero l'integrale definito nel loro dominio non risulta finito

$$\int_\Theta \pi(\theta) d\theta = \infty.$$

È da segnalare che la distribuzione a posteriori può comunque essere propria,anche se la la distribuzione a priori è impropria. Tuttavia, questo va verificato di volta in volta e non è vero in generale.

1.3 Procedure inferenziali

Nell'ambito bayesiano si avverte meno l'esigenza di differenziare le diverse procedure inferenziali rispetto all'ambito classico. Questo deriva dal fatto che tutte le operazioni a scopo inferenziale derivano da opportune sintesi della distribuzione a posteriori.

Per iniziare, la stima puntuale del parametro deriva dall'uso di un indice di posizione applicato alla distribuzione a posteriori. Questo viene utilizzato per identificare il valore più plausibile del parametro che identifica la legge del modello statistico parametrico che abbia generato il risultato sperimentale espresso dai dati y. L'indice più usato è il valore atteso a posteriori, $E(\theta|y)$, calcolato come $\int_\Theta \theta\pi(\theta|y)d\theta$. Dato che questo si presenta come un integrale, a volte il suo calcolo risulta complesso. Altre volte esso, addirittura, può non esistere. In questi casi è opportuno ricorrere a stime puntuali alternative come, ad esempio, la mediana o la moda della distribuzione a posteriori.

Come in ogni analisi statistica è bene poter esprimere l'incertezza associata alle conclusioni inferenziali, ad esempio tenendo conto della variabilità della stima puntuale. Il modo più logico per attuare tale pratica è descrivere un insieme di valori $\hat{\Theta}_{1-\alpha}$ tali che $Pr(\theta \in \hat{\Theta}_{1-\alpha}|y) = 1-\alpha$. Questa è detta regione di credibilità di livello $1-\alpha$. Tale regione nel caso scalare può essere individuata tramite i rispettivi quantili della distribuzione a posteriori $(\hat{\theta}_{\alpha/2}, \hat{\theta}_{1-\frac{\alpha}{2}})$, dove $Pr(\theta < \hat{\theta}_\alpha|y) = \alpha$. Se la distribuzione a posteriori fosse asimmetrica, in questo modo si rischierebbe di non considerare dei punti che possiedono una valore della densità maggiore di quelli esclusi. Un'alternativa è quindi utilizzare un intevallo HPD (*Highest Posterior Density*): si tratta in pratica di inserire nell'insieme $\hat{\Theta}_{1-\alpha}$ tutti i valori di θ la cui densità a posteriori risulta più elevata, fino a raggiungere una probabilità a posteriori di copertura complessiva non inferiore al livello prescelto $1-\alpha$. Tale procedura non può essere generalmente costruita per via analitica. Nei casi in cui la

densità a posteriori risulti simmetrica, i due tipi di intervalli di credibilità forniscono gli stessi risultati.

Coerentemente con l'approccio bayesiano qualsiasi sintesi d'interesse può essere reperita dalla distribuzione a posteriori, come ad esempio la probabilità a posteriori che il parametro θ sia maggiore o minore di una particolare soglia k, come $Pr(\theta > k|y)$. Queste misure sono molto utilizzate in ambito epidemiologico se si assume che θ definisca ad esempio il rischio relativo della popolazione. In ambito frequentista questa operazione non è altrettanto immediata ed è generalmente sostituita dal calcolo di un *p-value* in una verifica d'ipotesi sul rischio relativo, la quale però non fornisce o stesso tipo di informazione quantitativa del fenomeno d'interesse.

1.3.1 Distribuzione predittiva

La distribuzione a posteriori sintetizza la conoscenza riguardante il parametro, dopo aver osservato i dati derivanti da un esperimento. Spesso può essere utile voler valutare delle previsioni sulla base delle informazioni a noi note. Definita una nuova osservazione y^*, possiamo determinarne la distribuzione predittiva nel seguente modo

$$p(y^*|y) = \int_\Theta p(y^*|y,\theta)\pi(\theta|y)d\theta.$$

Quindi la previsione si basa sulla marginalizzazione del parametro contenuto nella verosimiglianza calcolata sulla nuova osservazione pesata con la distribuzione a posteriori di θ. Così facendo si considera in modo intrinseco l'incertezza del parametro, perché utilizzando la distribuzione a posteriori si considera anche la variabilità. In ambito frequentista è più comune usare una stima puntuale, dove il candidato naturale per $p(y^*|y)$ è $p(y^*|y,\hat{\theta})$, in cui $\hat{\theta}$ rappresenta ad esempio la stima di massima verosimiglianza. Nel caso in cui si voglia fare una previsione su un valore mancante presente tra le proprie osservazioni si può utilizzare una tecnica diversa di previsione detta **data augmentation** (Tanner e Wong, 1987), tale metodo è presente in svariati altri contesti, come per esempio la stima di massima verosimiglianza calcolata con l'algoritmo EM, ma essa può essere utilizzata anche in un contesto bayesiano sfruttando la distribuzione a posteriori del parametro.

L'idea di base è la seguente: ai dati osservati y viene aggiunta la quantità latente z, assumendo che entrambe siano note così da poter formulare la distribuzione a posteriori aumentata $p(\theta|y,z)$. Si procede successivamente generando un set di valori, $z^{(1)},...,z^{(m)}$, dalla distribuzione predittiva $p(z|y)$, così facendo la densità a posteriori completa $p(\theta|y)$ può essere definita tramite la media di $p(\theta|y,z)$ per il set di valori generati precedentemente. Come si può notare vi è una dipendenza fra $p(z|y)$ e $p(\theta|y)$, infatti questa dipendenza genera una operazione iterativa che, sotto semplici condizioni di regolarità, porta a convergenza. Così si ottiene una distribuzione predittiva per il valore mancante e la distribuzione a posteriori del parametro è aggiornata per tale valore.

L'algoritmo può essere descritto tramite due step successivi che si possono iterare:

1. *Imputation step*
   (a) Si ottiene un'approssimazione di $p_i(\theta|y)$ con i dati osservati;
   (b) Si genera un campione $z^{(1)},...,z^{(m)}$, dall'approssimazione della densità predittiva $p(z|y) = \int_\Theta p(z|\theta,y)p_i(\theta|y)d\theta$;
2. *Posterior step*
   (a) Si aggiorna la distribuzione a posteriori del parametro, $p_{i+1}(\theta|y) = \frac{1}{m}\sum_{j=1}^{m} p_i(\theta|z^j,y)$

Tale tecnica può essere utilizzata per verificare la capacità di un modello di effettuare delle previsioni "nascondendo" il valore della variabile d'interesse e adattandolo rispetto ai restanti dati. Questa tecnica verrà utilizzata per i test previsivi nel Capitolo 4.

1.4 Modellazione gerarchica

La creazione di modelli bayesiani gerarchici va in qualche modo a "diluire" il peso della scelta della distribuzione a priori. Questo può essere fatto nel caso in cui si ipotizzi che si sia osservato il campione $y_1,...,y_n$ estratto dalla

distribuzione $p(y;\theta)$, dove θ sia dotato di una legge di probabilità $\pi(\theta|\omega)$. In questo caso il parametro d'interesse θ dipende a sua volta da un altro parametro ω, al quale a sua volta viene associata una legge di probabilità di secondo stadio $\zeta(\omega)$. In questo modo, lo stesso modello statistico può essere rappresentato in vari modi, tutti equivalenti, ma in grado di mettere in risalto aspetti diversi del modello stesso. In tale contesto la distribuzione a posteriori si presenterà nel seguente modo

$$\pi(\theta|y,\omega) \propto p(y|\theta)\pi(\theta|\omega)\zeta(\omega).$$

Ovviamente la gerarchia potrebbe continuare descrivendo la legge di ω tramite un altro parametro. Si nota che più si va a avanti nella gerarchia più l'effetto delle gerarchie più avanzate è debole. Questo schema gerarchico può essere descritto utilizzando un semplice esempio

$$\begin{gathered} y_1,...,y_n|\theta \overset{iid}{\sim} Po(\theta), \\ \theta|\alpha,\beta \sim Ga(\alpha,\beta), \\ \alpha \sim Esp(\nu_0), \\ \beta \sim Esp(\rho_0), \end{gathered}$$

in cui si definisce un campione $y_1,...,y_n$ derivante da un modello di Poisson, dove il parametro θ ha una distribuzione a priori gamma, i cui parametri hanno una distribuzione a priori Esponenziale con un parametri fissati, ν_0 e ρ_0. Una comoda visualizzazione di tale gerarchia deriva dall'uso dei grafi direzionali aciclici, dove si mostrano le relazioni a priori fra le varie variabili e la loro natura, come si mostra in Figura 1.1. In questo caso, le variabili casuali sono indicate con un cerchio, mentre i quadrati identificano i valori noti.

Questa possibile descrizione del modello ben si lega alla definizione di un modello gerarchico in cui l'aspetto essenziale è che la generica osservazione y_{ij} effettuata sull'i-esima unità, nel j-esimo gruppo viene utilizzata per stimare la distribuzione dei vari θ_j. I questo caso l'uso della specificazione gerarchica della distribuzione a priori per il vettore $\theta = (\theta_1,...,\theta_n)$ permette di porre che i vari θ_j siano specifici per ogni individuo, ma provengano da una distribuzione comune descritta da un parametro ω a cui verrà associata una distribuzione a priori dipendente a sua volta da un parametro ν_0 che viene considerato fissato.

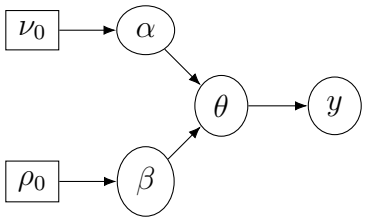


Figura 1.1: Rappresentazione della gerarchia del modello bayesiano usato in esempio

Alla base della costruzione gerarchica del modello c'è la condizione di scambiabilità dei parametri, ovvero che gli elementi che compongono il vettore θ sono scambiabili. Ciò significa che la distribuzione di θ non cambia permutandone gli elementi. In sostanza, si definisce che l'informazione portata dai vari θ_j è la medesima. Si può far sì che vi sia una dipendenza fra vari θ_j tramite la distribuzione a priori marginale per θ nel seguente modo

$$\pi(\theta|\nu_0) = \int_\Omega \pi(\theta|\omega)\zeta(\omega|\nu_0)d\omega.$$

In questo caso si presenta la caratteristica di *borrowing strength* sulle stime dei vari θ_j data dalla dipendenza che si è indotta tra questi elementi.

1.5 Inferenza bayesiana approssimata

Come detto in precedenza la distribuzione a posteriori non può essere sempre trovata in modo analitico. Questo è il motivo principale per cui la statistica bayesiana non ha avuto un grande impatto nella comunità scientifica ai suoi albori, mentre con l'avvento del *personal computer* il suo utilizzo è aumentato notevolmente. Nella maggior parte nei casi, soprattutto utilizzando distribuzioni a priori non coniugate, non si riesce ad ottenere una distribuzione a posteriori in forma chiusa. Per risolvere questa problematica si possono percorrere due diverse strade: ottenere la distribuzione a posteriori tramite simulazione o approssimazione analitica. Il primo metodo, più diffuso, verrà presentato brevemente nel paragrafo seguente, mentre l'uso di approssimazioni analitiche come l'approssimazione di Laplace e le sue modificazioni verranno descritte nel dettaglio nel capitolo seguente e saranno le tecniche che verranno utilizzate in questa tesi.

1.5.1 *Markov Chain Monte Carlo*

I metodi più comunemente utilizzati per ottenere un'approssimazione della distribuzione a posteriori si basano su simulazioni di valori casuali della stessa. Tra i metodi di simulazione, detti metodi Monte Carlo, i più utilizzati sono i metodi *Markov Chain Monte Carlo* (MCMC) i quali basandosi su opportune catene di Markov generano dei valori (dipendenti) appartenenti alla distribuzione a posteriori del parametro d'interesse.

La logica alla base di tale metodologia è quella di costruire una serie di variabili aleatorie $\theta^{(0)}, \theta^{(1)}, ..., \theta^{(j)}, ..., \theta^{(N)}$, detta catena markoviana di primo ordine definite in uno spazio Θ, tale per cui, la variabile $\theta^{(t+1)}$ condizionatamente alla variabile immediatamente precedente $\theta^{(t)}$ risulta indipendente dalle restanti variabili, sotto opportune condizioni, tale catena possiederà una distribuzione limite, la quale nel nostro contesto sarà $\pi(\theta|y)$. Quindi si andrà a generare una serie abbastanza lunga di valori tale per cui essi convergeranno alla distribuzione a posteriori. Si avrà perciò un'approssimazione Monte Carlo di tale distribuzione. Si deve considerare l'importanza che la distribuzione limite sia **invariante**, ovvero che al passo j della catena la distribuzione raggiunta rimanga la medesima per il passo $j+1$ e tutti i successivi passi.

Questo algoritmo presenta due ordini di problemi:

- i valori generati non sono indipendenti ma legati, appunto, da una struttura markoviana;
- non è chiaro fino da quando possiamo iniziare a considerare i valori come realizzazioni da $\pi(\theta|y)$, ovvero quando si è raggiunta la convergenza alla distribuzione limite.

Il primo, come detto, non è veramente importante nel momento in cui la catena è ergodica. Inoltre, volendo, è possibile filtrare i valori della catena considerando solo un sottoinsieme di valori equispaziati così da ridurre la correlazione tra questi. Il secondo problema ha una soluzione meno chiara: dal punto di vista delle applicazioni occorre stabilire dei criteri diagnostici, di monitoraggio della realizzazione della catena che ci consentano di stabilire, con una certa approssimazione, se e quando la catena ha raggiunto la convergenza. È comunque buona norma scartare un insieme di valori iniziali della

catena, dato che molto probabilmente la convergenza non è ancora avvenuta. Il periodo iniziale in cui la catena sta raggiungendo la convergenza è detto *burn-in* o *warm-up*.

Nei prossimi due paragrafi verranno illustrati due metodi che permettono di creare delle catene ergodiche al fine di riuscire ad ottenere un'approssimazione della distribuzione limite.

1.5.2 Metropolis - Hastings

Il primo criterio per generare una catena di Markov con una determinata distribuzione limite $\pi(\theta|y)$ si basa essenzialmente sulla generazione di una serie di valori da una distribuzione nota chiamata *proposal*, $q(.|\theta)$ i quali vengono accettati secondo una condizione che garantisce che i valori generati siano una catena markoviana e che la distribuzione limite sia invariante pari a $\pi(\theta|y)$. L'algoritmo di Metropolis - Hastings può essere descritto nel seguente modo:

1. Al tempo 0, la catena si trova a $\theta^{(0)} = \theta_0$, un valore di partenza noto;
2. Al tempo t, si genera un valore dalla *proposal*, $\theta^* \sim q(z|\theta^{(t-1)})$;
3. Si calcola la probabilità di transizione, ovvero la probabilità che il valore venga accettato:
$$\alpha = \alpha(\theta^*, \theta^{(t-1)}) = \min\left(1; \frac{q(\theta^{(t-1)}|\theta^*)\pi(\theta^*|y)}{q(\theta^*|\theta^{(t-1)})\pi(\theta^{(t-1)}|y)}\right)$$
4. Definire
$$\theta^{(t)} = \begin{cases} \theta^*, & \text{con probabilità } \alpha \\ \theta^{(t-1)}, & \text{con probabilità } 1-\alpha \end{cases}$$

Si nota come il seguente algoritmo genera ad ogni iterazione un valore, esso potrà essere un valore già osservato o meno, a seconda dalla probabilità di accettazione. Nel caso particolare il cui la *proposal* abbia densità simmetrica, il calcolo di α si semplifica ed si ottiene

$$\alpha = \min\left(1; \frac{\pi(\theta^*)}{\pi(\theta^{(t-1)})}\right)$$

Vi sono diverse possibilità per la distribuzione $q(\theta^*|\theta)$. Essa deve essere scelta a seconda della facilità di generazione e dalla efficacia che porta nell'algoritmo. A seconda che il parametro sia o meno multidimensionale vi possono essere diverse strategie per simulare la distribuzione a posteriori. Una scelta comune è il *random walk* in cui si ipotizza che $\theta^* \sim U(\theta^{(t)} - \epsilon, \theta^{(t)} + \epsilon)$ oppure $\theta^* \sim N(\theta^{(t)}, \epsilon^2)$, in cui il parametro ϵ deve essere scelto al fine di ottimizzare l'efficienza dell'algoritmo. Un buon modo per scegliere il valore di ϵ è valutare il tasso medio di accettazione dei valori prodotti il quale dovrebbe essere tra il 0.25 e il 0.50 nel caso mono-parametrico e intorno a 0.25 nel caso multidimensionale. Questo deve essere fatto per far sì che la catena esplori tutto il dominio della distribuzione a posteriori mantenendo una correlazione non troppo elevata.

1.5.3 *Gibbs sampler*

L'algoritmo di Gibbs può essere considerato un caso particolare del Metropolis-Hastings in cui i valori proposti sono sempre accettati. Quindi la probabilità di transizione risulta pari a 1 ad ogni iterazione. Esso però può essere utilizzato solo in specifici casi. In particolare, si usa nel caso multivariato quando è disponibile generare valori dalle densità condizionate delle variabili considerate, condizionatamente a tutte le altre. In particolare, le *proposal* avranno la formulazione

$$q_j(\theta_j|\boldsymbol{\theta}) = \pi(\theta_j|\boldsymbol{\theta}_{(j)}, y).$$

dove j identifica la j-esima componente del parametro d'interesse. Dunque la *proposal* diviene la densità a posteriori di θ_j condizionata a tutte le altre componenti del vettore $\boldsymbol{\theta}$, indicate con $\boldsymbol{\theta_{(j)}}$

Generalmente questo algoritmo produce dei campioni con buone proprietà, anche se la controindicazione è che può essere usato solo se le funzioni di densità condizionate sono note. In ogni caso, se questo non fosse possibile, si può sempre utilizzare un passo di Metropolis-Hastings, o un altro algoritmo della famiglia di metodi MCMC, per generare il valore dalla funzione di densità non nota, a meno di una costante di proporzionalità.

I metodi di simulazione sono ampiamente utilizzati nelle applicazioni. Tuttavia, spesso richiedono sforzi computazionali intensivi, specialmente in

modelli complessi. Per questo motivo altri tipi di approssimazioni, di natura analitica, possono risultare più convenienti. Nel capitolo seguente verrà illustrata in dettaglio la metodologia INLA, che verrà utilizzata poi nelle applicazioni di questa tesi.

Capitolo 2

Integrated Nested Laplace Approximation (INLA)

L'inferenza bayesiana, da un punto di vista concettuale, è abbastanza semplice e lineare: si aggiorna la conoscenza soggettiva descritta dalla distribuzione a priori dell'ignoto parametro con l'informazione disponibile dai dati, ottenendo così la distribuzione a posteriori. Nel momento in cui si è ottenuta tale densità, si possono calcolare tutte le statistiche d'interesse tramite delle semplici operazioni sulla distribuzione.

Da un punto di vista pratico questo risulta più complesso, perché a meno di analisi semplici o scegliendo particolari distribuzioni a priori come descritto nel Paragrafo 1.2.1, si deve spesso applicare un algoritmo numerico della famiglia degli MCMC le cui simulazioni possono essere computazionalmente molto intensive e possono richiedere un lungo periodo prima che la catena di Markov raggiunga la convergenza in modo soddisfacente. Queste problematiche emergono soprattutto in modelli più complessi, come possono essere quelli gerarchici, in cui i parametri sono presenti in diverse gerarchie e possono essere in numerosità elevata. Sono stati creati appositi software come WinBUGS (Unit, 2020) per rendere l'approccio via MCMC più efficiente, ma i costi computazionali risultano ancora non ottimali.

inla hanno proposto una diversa tecnica di per fare inferenza bayesiana più velocemente: essa non si basa sulla convergenza asintotica di catene di Markov ergodiche, ma su una serie di approssimazioni analitiche della

distribuzione a posteriori. Tale tecnica si chiama *Integrated Nested Laplace Approximations*(INLA). Come si evince dal nome, essa sfrutterà delle approssimazioni di Laplace per stimare le distribuzioni a posteriori di modelli bayesiani gerarchici, calcolate tramite integrali numerici. Alla fine del processo di approssimazione si otterranno le varie marginali a posteriori dei parametri d'interesse e non la distribuzione congiunta, il che porta a delle semplificazioni notevoli dal punto di vista computazionale. Il metodo INLA può essere applicato anche a casi d'inferenza bayesiana più semplici, ma con miglioramenti più limitati rispetto alle procedure MCMC.

In particolare INLA è stato creato per essere applicato per la famiglia dei *Latent Gaussian Model*, i quali si possono presentare in diversi casi dato che modelli come: GLMM, GAMM, serie temporali, modelli spaziali, appartengono a tale famiglia e vengono utilizzati negli ambiti più disparati, da quello epidemiologico a quello economico. Nelle prossime sezioni si andranno a presentare in sequenza gli elementi fondamentali di INLA, ovvero

- i *Latent Gaussian Model*;
- i *Gaussian Markov Random Fields* (GMRF), una classe di modelli gaussiani latenti che velocizzano il processo di stima di INLA;
- l'approssimazione di Laplace.

Infine verrà presentato l'algoritmo di INLA e si concluderà con una discussione sulle problematiche legate a tale procedura. A supporto della stesura del seguente capitolo sono stati utilizzati i riferimenti Rue *et al.* (2009), Rue, Riebler *et al.* (2016) e Wang *et al.* (2018).

2.1 *Latent Gaussian Model*

Il concetto di modello gaussiano latente rappresenta una famiglia che riunisce un'ampia classe di modelli statistici. Essi sono descritti da una formulazione gerarchica a tre livelli nella quale le osservazioni y possono essere considerate indipendenti condizionatamente a un effetto latente Gaussiano multidimensionale $\mathbf{x}$ e iperparametro vettoriale $\boldsymbol{\theta}$, il quale esprime

$$f(y|\mathbf{x},\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(y_i|x_i,\boldsymbol{\theta}),$$

dove x assume ha distribuzione normale. In genere l'effetto latente Gaussiano si presenta nella media della distribuzione di y tramite una funzione di legame appropriata, quindi il suo ruolo è di predittore della media. La distribuzione di x sarà

$$\mathbf{x}|\boldsymbol{\theta} \sim N(\mu(\boldsymbol{\theta}),\, Q^{-1}(\boldsymbol{\theta}))$$

in cui con Q si definisce la matrice di precisione, ovvero l'inversa della matrice di varianza-covarianza. In questo caso con $\boldsymbol{\theta}$ si definisce l'insieme degli iperparametri appartenenti sia alla conoscenza a priori su $\mathbf{x}$ che ad eventuali altri iperparametri legati alla verosimiglianza. A posteriori la distribuzione congiunta di $\mathbf{x}$ e $\boldsymbol{\theta}$ sarà:

$$\pi(\mathbf{x}, \boldsymbol{\theta}|y) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i|x_i, \boldsymbol{\theta})$$

Di seguito si elencano delle assunzioni sui modelli che si vogliono adattare affinché il metodo INLA sia accurato ed efficiente:

- la dimensione di $\boldsymbol{\theta}$ deve essere contenuta (minore di 6) e tipicamente non superiore a 20;
- i dati y sono indipendenti condizionatamente a $\mathbf{x}$ e $\boldsymbol{\theta}$, il che implica che y_i dipende da una componente dell'effetto latente x_i;
- la distribuzione dell'effetto latente $\mathbf{x}|\boldsymbol{\theta}$ necessita, oltre ad essere Gaussiana anche di possedere la proprietà di dipendenza markoviana, quindi di essere un *Gaussian Markov random field* se la dimensione di y è maggiore di 10^3.

Come detto precedentemente i modelli gaussiani latenti sono una classe che considera una grande varietà di modelli, infatti l'effetto latente $\mathbf{x}$ nelle sue componenti x_i può essere interpretato come un predittore η_i,

$$\eta_i = \mu + \sum_j \beta_j z_{ij} + \sum_k f_{k,j_k(i)} \tag{2.1}$$

in cui μ è l'intercetta, $\mathbf{z}$ sono le covariate con i relativi parametri $\boldsymbol{\beta}$ e $\mathbf{f}$ rappresentano specifici processi Gaussiani. La notazione legata a $\mathbf{f}$ descrive che il k-esimo elemento di $\mathbf{f}$ contribuisce all'i-esimo predittore lineare con l'elemento j.

Si assume che le tre macro componenti del modelli siano indipendenti a priori e che gli effetti fissi descritti da μ e $\boldsymbol{\beta}$ possiedano distribuzione a priori normale. Questo fa si che vi sia un legame tra il predittore descritto in (2.1) e i *Latent Gaussian Model*. La distribuzione congiunta di

$$\mathbf{x} = (\boldsymbol{\eta}, \mu, \boldsymbol{\beta}, \mathbf{f}) \tag{2.2}$$

si presenta come un effetto latente Gaussiano gerarchico con dimensione pari a n. Anche qui si sottolinea come $\boldsymbol{\theta}$ comprenda sia i parametri dell'ultima gerarchia del modello che quelli legati alla verosimiglianza di y, come ad esempio possibili parametri di dispersione. La sua dimensione ridotta risulta sempre cruciale per l'efficienza computazionale del metodo.

2.2 Gaussian Markov Random Fields

Un GMRF non è altro che un vettore di variabili descritte da una densità multidimensionale normale che possiede la proprietà di indipendenza condizionata, il che significa che gli elementi x_i e x_j sono indipendenti condizionatamente ai rimanenti elementi $\mathbf{x}_{-ij}$ del vettore latente $\mathbf{x}$. Un semplice esempio è descritto da un modello autoregressivo del primo ordine $x_t = \phi x_{t-1} + \epsilon_t$, con $t = 1, ..., n$; in questo caso gli elementi x_t e x_s sono correlati con correlazione pari a $\phi^{|s-t|}$, ma questi due elementi sono condizionatamente indipendenti dato il vettore $\mathbf{x}_{-ij}$ per $|s - t| > 1$.

È noto che nel caso di un vettore normale si ha il seguente risultato

$$x_i \perp x_j \iff Q^{-1}_{i,j} = 0$$

tale forte condizione spiega l'indipendenza marginale all'interno di $\mathbf{x}$, ma non sempre ragionevole. Viceversa, l'indipendenza condizionata è più comune e va a descrivere la seguente condizione:

$$x_i \perp x_j | \mathbf{x}_{-ij} \iff Q_{i,j} = 0$$

che va a definire una matrice di precisione sparsa per $\mathbf{x}$. Tale caratteristica porta con sé un grande beneficio computazionale, infatti nell'esempio autoregressivo precedente la matrice di precisione può essere fattorizzata con un costo dell'ordine $\mathcal{O}(n)$ se è sparsa, altrimenti avrebbe un costo computazione dell'ordine $\mathcal{O}(n^3)$. Da un punto di vista puramente informatico il costo di memorizzazione di una matrice sparsa si abbassa da un ordine di $\mathcal{O}(n^2)$ a $\mathcal{O}(n)$, gli ordini descritti non sono generali, ma dipendono fortemente dalla sparsità della matrice Q. Rimane comunque costante il grande beneficio computazionale. Per dettagli aggiuntivi sul guadagno nella fattorizzazione di Cholesky, si rimanda a Rue e Martino (2007).

Nella costruzione di modelli additivi la presenza di un GMRF porta una conseguenza che verrà successivamente sfruttata in INLA, ovvero che la distribuzione di $\mathbf{x}$ in (2.2) rimane un GMRF la cui matrice di precisione risulta essere la somma delle matrici di precisione di tutte le varie componenti del modello. Questo risultato fa sì che si possa sfruttare l'efficienza computazionale portata da una matrice Q sparsa.

2.3 Approssimazione di Laplace

L'approssimazione di Laplace viene utilizzata in ambito bayesiano per il calcolo della distribuzione a posteriori. In particolare, data una densità, applicando il Teorema 1.1:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta}.$$

il calcolo del denominatore può risultare problematico, perciò si può ricorrere a una sua approssimazione, la quale consiste nell'utilizzare uno sviluppo di Taylor attorno ad un certo valore x_0 del tipo

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + ...$$

dove $f', f''ef'''$ sono le prime tre derivate di f, fino ad arrivare all'ordine desiderato affinché l'approssimazione sia soddisfacente.

Ponendo la costante di normalizzazione della distribuzione a posteriori pari a $\int_a^b g(t)dt$, assumendo che esista un punto t_0 tale per cui la funzione g

sia massima e definendo $h = \log g$, allora l'approssimazione di Laplace di h attorno alla moda t_0 sarà

$$\begin{aligned}
\int_a^b g(t)dt &= \int_a^b \exp(h(t))dt \\
&\approx \int_a^b \exp\left[h(t_0) + h^{'}(t_0)(t-t_0) + \frac{1}{2}h^{''}(t_0)(t-t_0)^2\right]dt, \\
&\approx \int_a^b \exp\left[h(t_0)\frac{1}{2}h^{''}(t_0)(t-t_0)^2\right]dt \\
&\approx \exp[h(t_0)]\int_a^b \exp\left[\frac{1}{2}h^{''}(t_0)(t-t_0)^2\right]dt \\
&\approx \exp[h(t_0)]\int_a^b \exp\left[-\frac{1}{2}\frac{(t-t_0)^2}{-h^{''}(t_0)^{-1}}\right]dt
\end{aligned}$$

Dove si riconosce all'interno dell'integrale una quantità proporzionale alla densità di una distribuzione normale con media pari a t_0 e varianza pari a $-h^{''}(t_0)^{-1}$. Per ottenere esattamente una forma chiusa dell'integrale si dovranno aggiungere degli elementi della funzione di ripartizione della Normale con media t_0 e varianza $-h^{''}(t_0)^{-1}$

$$\int_a^b g(t)dt = exp[h(t_0)]\sqrt{\frac{2\pi}{-h^{''}(t_0)^{-1}}}[\Phi(b|t_0, -h^{''}(t_0)^{-1}) - \Phi(a|t_0, -h^{''}(t_0)^{-1})], \tag{2.3}$$

dove, se $a \approx \infty$ e $b \approx \infty$ come spesso accade, si ha che la parte contenuta nelle parentesi quadre è pari a 1. Perciò infine si ottiene:

$$\int_{-\infty}^{\infty} g(t)dt = exp[h(t_0)]\sqrt{\frac{2\pi}{-h^{''}(t_0)}}.$$

Quindi l'operazione di integrazione della costante di normalizzazione è stata sostituita da un problema di massimizzazione.

L'idea di base è molto semplice, si va ad approssimare la distribuzione target con una Normale, identificando la moda e la curvatura sulla moda le quali se necessario vengono trovate iterativamente tramite un algoritmo come Newton-Raphson; ovviamente tale approssimazione è più raffinata tanto più la distribuzione target si avvicina alla normale. In generale l'errore di approssimazione è del tipo $\mathcal{O}(n^{-1})$, assumendo che $h(t)$ sia di ordine $\mathcal{O}(n)$.

Di seguito viene riportato un esempio dell'applicazione dell'approssimazione di Laplace alla distribuzione Chi-quadrato, $\mathcal{X}^2$ e di come questa migliori a mano a mano che la distribuzione target si avvicina alla distribuzione normale. Gli elementi necessari sono

$$g(x;k) = \frac{x^{(\frac{k}{2})}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}$$

$$h(x;k) = \left(\frac{k}{2} - 1\right) log(x) - \frac{x}{2}$$

$$h^{'}(x;k) = \frac{\left(\frac{k}{2} - 1\right)}{x} - \frac{1}{2}$$

$$h^{''}(x;k) = -\frac{\left(\frac{k}{2} - 1\right)}{x^2}$$

da cui si ottiene che la moda t_0 è pari a $k-2$ e la varianza $-h^{''}(t_0)^{-1}$ è pari a $2(k-2)$. Dal punto di vista della densità, l'approssimazione porta alla seguente formulazione della distribuzione target $\mathcal{X}^2_k \sim N(k-2; 2(k-2))$. Dalla Figura 2.1 si può apprezzare graficamente l'approssimazione al avriare dei gradi di libertà k, mentre dalla Tabella 2.1 si può notare come l'approssimazione dell'integrale $\int_{-\infty}^{\infty} g(x;k)dx$ tramite il metodo descritto nell'equazione (2.3), migliori all'aumentare dei gradi di libertà.

Tabella 2.1: Approssimazione dell'integrale con il metodo di Laplace al variare dei gradi di libertà k

	Valore reale	k=6	k=16	k=30	k=50
Integrale	1	0.959502	0.988174	0.994066	0.996534

L'utilizzo di questa tecnica risulta ancora più vantaggioso nel caso in il numero di parametri sia abbastanza elevato. Assumiamo di dover calcolare la distribuzione marginale $\pi(\gamma_1)$ dalla distribuzione congiunta multidimensionale $\pi(\gamma)$, utilizzando il teorema di Bayes e l'approssimazione di Laplace, si ha

$$\pi(\gamma_1) = \frac{\pi(\gamma)}{\pi(\gamma_{-1}|\gamma_1)} \approx \frac{\pi(\gamma)}{\pi_G(\gamma_{-1}; \mu(\gamma_1), Q^{-1}(\gamma_1))},$$

in cui il denominatore è stato approssimato con una distribuzione normale i cui parametri sono la moda e la curvatura alla moda. In Tierney e Kadane (1986) viene dimostrato come se $\pi(\gamma) \propto \exp(nf_n(x))$, dove $f_n(x)$ è la media della log-verosimiglianza e se siamo in possesso di n osservazioni dallo stesso modello, l'errore di approssimazione è dell'ordine di $\mathcal{O}(n^{-\frac{3}{2}})$. Questo risultato è notevole, ma ha delle assunzioni stringenti, in quanto in genere

- invece di possedere diverse osservazioni dallo stesso modello, potremmo possedere una sola osservazione per modello oppure varie osservazioni da vari modelli;
- la condizione che la dimensione di γ sia fissata per $n \to \infty$ non sempre è rispettata, come per esempio nei modelli gerarchici.

È comunque possibile ottenere dei buoni risultati applicando l'approssimazione di Laplace in un caso multimensionale sfruttando le capacità del modello. In particolare si può sostituire il caso di "osservazioni dallo stesso modello" con "osservazioni da modelli simili". Questo aiuta l'approssimazione utilizzando il *borrow in strength* tra le variabili che assumiamo come simili a priori. Inoltre, si può ridurre la relazione tra la dimensionalità del modello e n, dato che la dimensione effettiva crescerà più lentamente di n dato che si assumerà che le osservazioni sono simili tra loro.

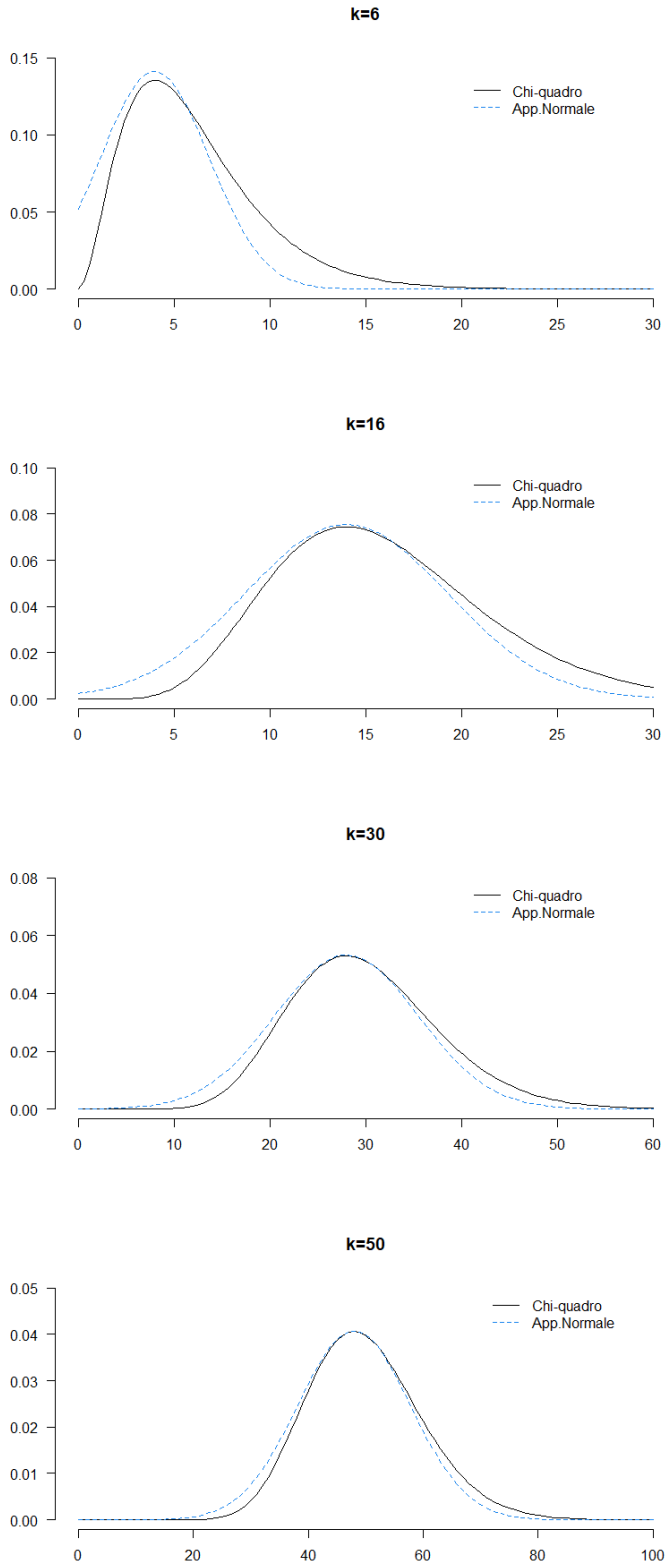


Figura 2.1: Approssimazioni di Laplace della distribuzione Chi-quadro al variare dei gradi di libertà k

2.4 *Integrated Nested Laplace Approximations*

Dopo che si sono analizzati tutti gli elementi caratteristici inseriti all'interno di INLA si prosegue andando a studiare il reale meccanismo con cui si può calcolare un'approssimazione delle distribuzioni a posteriori dei parametri d'interesse. In questo contesto si vogliono ottenere le distribuzioni a posteriori marginali degli iperparametri $\pi(\theta_j|y)$ e degli effetti latenti $\pi(x_i|y)$, ovvero

$$\pi(\theta_j|y) = \int\int \pi(\mathbf{x}, \boldsymbol{\theta}|y) d\mathbf{x} d\boldsymbol{\theta}_{-j} = \int \pi(\boldsymbol{\theta}|y) d\theta_{-j}, \tag{2.4}$$

$$\pi(x_i|y) = \int\int \pi(\mathbf{x}, \boldsymbol{\theta}|y) d\mathbf{x}_{-i} d\boldsymbol{\theta} = \int \pi(x_i|\boldsymbol{\theta}, y)\pi(\boldsymbol{\theta}|y) d\boldsymbol{\theta}. \tag{2.5}$$

Come spiegato precedentemente gli integrali rispetto a $\mathbf{x}$ sono in una dimensione maggiore rispetto a quelli su $\boldsymbol{\theta}$ per costruzione del modello. Il cuore di INLA risiede quindi nell'utilizzare una efficiente integrazione numerica di $\pi(\boldsymbol{\theta}|y)$ e di $\pi(x_i|\boldsymbol{\theta}, y)$ evitando integrazioni poco maneggevoli.

Per prima cosa si deve sfruttare la scomposizione

$$\pi(\boldsymbol{\theta}|y) = \frac{\pi(\mathbf{x}, \boldsymbol{\theta}|y)}{\pi(\mathbf{x}|\boldsymbol{\theta}, y)} \propto \frac{\pi(y|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x}|\boldsymbol{\theta}, y)},$$

in cui il numeratore risulta semplice da calcolare, in quanto presenta la verosimiglianza del modello, la distribuzioni a priori dell'effetto latente e la distribuzione a priori dell'iperparametro, ma il denominatore risulta invece complesso. Tale situazione è la medesima riportata nel Paragrafo 2.3, infatti si deve usare un'approssimazione di Laplace ottenendo

$$\widetilde{\pi}(\boldsymbol{\theta}|y) \propto \left.\frac{\pi(y|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\widetilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, y)}\right|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} \tag{2.6}$$

dove al denominatore ora vi è l'approssimazione a una distribuzione Gaussiana della costante di normalizzazione e $\mathbf{x}^*(\boldsymbol{\theta})$ è la moda di $\mathbf{x}$ per una data configurazione di $\boldsymbol{\theta}$. Questa approssimazione risulta generalmente poco accurata dato che $\widetilde{\pi}(\boldsymbol{\theta}|y)$ si allontana dalla distribuzione normale.

Per migliorare l'accuratezza del risultato si dovrà esplorare e manipolare la distribuzione ottenuta in un modo non parametrico tramite la seguente procedura:

1. Per prima cosa si deve identificare la moda di $\widetilde{\pi}(\boldsymbol{\theta}|y)$ ottimizzando il suo logaritmo rispetto a $\boldsymbol{\theta}$, questo processo può essere fatto utilizzando un algoritmo come Newton-Raphson. La moda verrà definita come $\boldsymbol{\theta}^*$.

2. Si dovrà quindi calcolare la curvatura della funzione alla moda tramite l'Hessiana H, così facendo si può definire con $\Sigma = H^{-1}$. Di tale matrice si esegue la decomposizione spettrale $\Sigma = \mathbf{V}\Lambda\mathbf{V}^T$ per poter esplorare la funzione nel seguente modo:

$$\boldsymbol{\theta}(\mathbf{z}) = \boldsymbol{\theta}^* + \mathbf{V}\Lambda^{1/2}\mathbf{v}$$

   dove $\mathbf{v} \sim N(0, \mathbf{I})$. L'uso di tale riparametrizzazione corregge la distribuzione per la scala e per la rotazione.

3. La Figura 2.2 presa da **inla** rappresenta graficamente la procedura nel caso bimodale. L'esplorazione viene fatta partendo dalla moda, in cui $\mathbf{v}$=0 e si segue la direzione positiva v_1 con passi unitari, finché è rispettata la condizione $log(\widetilde{\pi}(\boldsymbol{\theta}(0)|y)) - log(\widetilde{\pi}(\boldsymbol{\theta}(\mathbf{v})|y)) < \delta_\pi$, in cui solitamente δ_π=2.5. I punti neri che si osservano sono i valori in cui la condizione è rispettata, successivamente si eseguono delle combinazioni di tali punti, ottenendo i punti grigi controllando sempre la condizione precedente. Questo prosegue finché non si trovano d punti.

4. Infine le varie distribuzioni $\pi(\theta_j|y)$ sono calcolate tramite integrazione numerica da $\widetilde{\pi}(\boldsymbol{\theta}|y)$, così facendo si risolve l'equazione (2.4).

Successivamente si dovrà ottenere un'approssimazione per $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ ossia l'altro fattore utile per ottenere $\pi(x_i|y)$. Per proseguire vi sono tre strade possibili che si differenziano per l'accuratezza e il costo computazionale. La prima via è di utilizzare l'approssimazione Gaussiana descritta in (2.6) marginalizzando per le n componenti di $\mathbf{x}$. Tale via anche se è la più breve porta generalmente a errori di simmetria, posizione e pesantezza nelle code. Un'altra alternativa consiste nel seguire l'approccio che si è fatto per $\pi(\boldsymbol{\theta}|y)$, quindi effettuare l'approssimazione di Laplace:

$$\widetilde{\pi}(x_i|\boldsymbol{\theta}, y) = \frac{\pi(\mathbf{x}|\boldsymbol{\theta}, y)}{\pi(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, y)} \propto \left.\frac{\pi(y|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\widetilde{\pi}_G(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, y)}\right|_{\mathbf{x}_{-i}=\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})}.$$

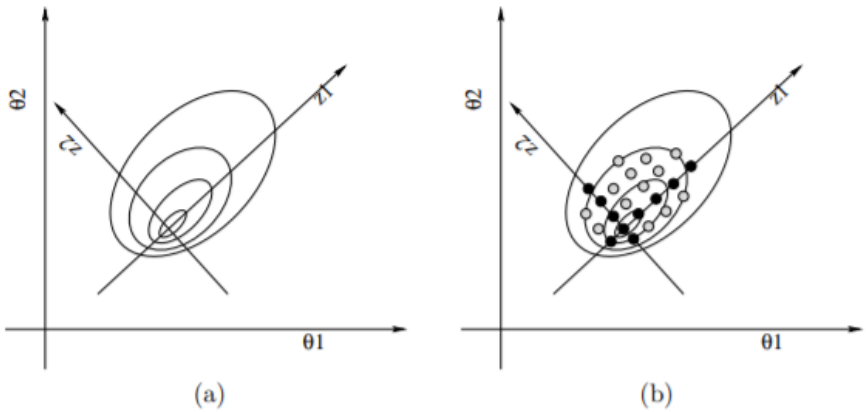


Figura 2.2: Illustrazione dell'esplorazione della distribuzione a posteriori per $\boldsymbol{\theta}$. In (a) si è fatto il passo 1 e 2, mentre in (b) si esegue il passo 3 esplorando la densità.

Dove ancora una volta il denominatore viene calcolato nella moda di $\mathbf{x}_{-i}$ per una data configurazione di $\boldsymbol{\theta}$. Questo approccio fornisce degli ottimi risultati dato che $\mathbf{x}$ si assume con distribuzione normale, ma è computazionalmente molto intensivo.

La terza strada che risulta essere un compromesso tra l'accuratezza dell'approssimazione e l'efficienza è chiamata **approssimazione di Laplace semplificata** la quale migliora l'approssimazione normale dagli errori di posizione e asimmetria tramite l'uso di una espansione di Taylor attorno alla moda definita dall'approssimazione di Laplace.

Infine si può passare all'ultimo passaggio per riuscire a calcolare la quantità espressa nell'equazione (2.5), si esegue l'integrazione numerica rispetto a $\boldsymbol{\theta}$ secondo la formula

$$\widetilde{\pi}(x_i|y) = \sum_k \widetilde{\pi}(x_i|\theta_k, y)\widetilde{\pi}(\theta_k|y)\Delta_k$$

in cui si sono utilizzate le approssimazioni trovate precedentemente e Δ_k sono dei pesi appropriati, i quali sono pari a 1 nel momento i cui i punti descritti nella griglia in Figura 2.2 sono equispaziati.

Alla fine è chiaro il significato dell'acronimo INLA: in questo approccio si usa l'integrazione numerica (*integrated*) di distribuzioni annidate dato che il

modello bayesiano è gerarchico (*nested*), le quali sono approssimate tramite varie approssimazioni di Laplace (*Laplace approximations*).

2.5 Limiti di INLA

INLA è un metodo per praticare inferenza bayesiana in modo deterministico a differenza dell'approccio di simulazione che si esegue con MCMC. Il suo vantaggio è che è computazionalmente meno intensivo, sia per il tempo di valutazione della distribuzione a posteriori sia per la memoria necessaria per eseguire l'inferenza. I risultati sono generalmente comparabili con il metodo basato sulle catene di Markov. Esso presenta però alcune problematiche sia a livello statistico che informatico. Tali limitazioni sono:

- **Iperparametri**: Il primo limite riguarda la dimensione degli iperparametri, la quale deve essere moderata dato che l'integrazione numerica risulta impraticabile con alte dimensioni. Questa condizione non può essere sempre soddisfatta, per esempio se si impone una struttura di correlazione fra le varie componenti del modello, oppure se si inserisce una componente che varia nel tempo.

- **Riproducibilità**: Dato che INLA risulta comunque una procedura intesiva essa ha bisogno molta capacità di calcolo possibile. Infatti essa va a sfruttare OpenMP, un'applicazione per la parallelizzazione del calcolo. Sfruttando tale programma si incorre però a delle differenze di risultato legato al numero di processori e alle loro caratteristiche. In questo caso non vi è la problematica dell'impostazione di un seme, come per i metodi basati sulle simulazioni, perché il problema non è legato alla generazione casuale, ma è legato a quante parallelizzazioni sono disponibili. Le variazioni fra diversi dispositivi sono in genere molto leggere, nell'ordine della settima cifra decimale. Ma nel caso questo risultasse importante si può imporre alla procedura di utilizzare un solo processore.

- **Concavità**: L'uso di INLA è limitato ai casi in cui la log-verosimiglianza è concava rispetto al predittore lineare, ovvero che l'informazione attesa

di Fisher è definita positiva, altrimenti la ricerca della moda nell'approssimazione di Laplace non arriva a convergenza.

- **Black box**: Anche se vi sono varie pubblicazioni sui dettagli del funzionamento di INLA, per l'utente non è possibile conoscere tutte le informazioni della sua implementazione in R. Pur essendo semplice da utilizzare, in molti casi è complicato comprendere le problematiche dei vari modelli. Oltre a questo, a differenza di un metodo MCMC si possiede meno controllo sulle operazioni che si stanno eseguendo e quindi non è possibile effettuare delle prove di adattamento. Le operazioni di *debugging* legate ai messaggi di errore, possono risultare complesse e richiedono spesso una conoscenza troppo approfondita della procedura INLA. Queste problematiche accadono soprattutto se l'insieme di dati presenta dimensioni considerevoli e se la struttura del modello è complessa, come può accadere nei modelli spazio-temporali utilizzati in questa tesi. Ciò può portare a dei risultati non previsti in cui è complicato utilizzare le diagnostiche fornite per evidenziare eventuali problemi.

Capitolo 3

Modelli spaziali

Il termine analisi spaziale identifica un insieme di tecniche atte a studiare le relazioni fra diversi fenomeni utilizzando le loro proprietà topologiche, geometriche o geografiche. Questo tipo di analisi si applica in diversi ambiti dall'astronomia, in cui si studia la posizione nello spazio delle diverse galassie, alla epidemiologia in cui si studiano come alcune malattie si comportano in località vicine.

Alla base della statistica spaziale vi è la definizione di dati spaziali, i quali si definiscono come una realizzazione di un processo stocastico indicizzato dallo spazio

$$Y(s) \equiv \{y(s), s \in \mathcal{D}\},$$

dove $\mathcal{D}$ è un sottoinsieme fissato in $\mathbb{R}^d$.

Il concetto fondamentale che riassume il significato della statistica spaziale è la prima legge della geografia (Tobler, 1979), la quale recita "elementi vicini sono più relazionati di elementi distanti". Essa è tanto semplice quanto efficace nello spiegare la dipendenza intrinseca fra gli elementi quando si conosce la posizione nello spazio di questi.

Di seguito il capitolo proseguirà con la presentazione della statistica spaziale in ambito epidemiologico, la descrizione di diverse tipologie di dati spaziali e un elenco di modelli che si possono utilizzare per analizzarli, partendo dal solo ambito spaziale fino a quello spazio-temporale. I testi di riferimento

utilizzati nella stesura del seguente capitolo sono Lawson (2018), Diggle e Giorgi (2019) e Blangiardo e Cameletti (2015).

3.1 *Disease mapping*

Il *disease mapping*, ovvero l'attività di mappare geograficamente la diffusione della malattia può essere definita tramite diversi nomi come: epidemiologia spaziale, epidemiologia ambientale o in modo più appropriato biostatistica spaziale, che è un termine che enfatizza l'ampiezza di tale argomento. Comunque il fulcro di queste differenti accezioni sono essenzialmente due, la distribuzione spaziale e la presenza di una diffusione. La prima gioca un ruolo importante nel fornire informazioni sulla località in cui sono raccolti i dati ed è una parte fondamentale dell'analisi. La seconda pone l'attenzione su cosa si studia osservando la concentrazione e l'incidenza di tale diffusione.

L'uso della posizione geografica nell'analisi della diffusione delle malattie ha una lunga storia. Uno dei primi casi riportati risale al 1854 a Londra dove il medico John Snow studiò la distribuzione spaziale dei casi di colera accertati, come si nota dalla Figura 3.1 (Parkes, 2013)i casi si concentrano attorno ad un punto. Grazie a questa mappatura realizzata dal medico inglese si scoprì che l'epidemia era stata causata e diffusa da un'unica pompa d'acqua situata a Broad Street (Lawson, 2018).

3.1.1 Dati spaziali

Vi sono diverse tipologie di dati spaziali, essi si differenziano principalmente rispetto alla facilità di poter definire con chiarezza l'area d'interesse e sul fatto che le osservazioni derivino da una operazione di aggregazione o meno. Le tre tipologie di dati spaziali sono:

- **Dati area o lattice**: Quando $y(s)$ è frutto del raggruppamento di osservazioni all'interno di una unità s la quale ha dei confini ben definiti nello spazio $\mathcal{D}$. In questo caso $\mathcal{D}$ è definita come una collezione d-dimensionale di unità spaziali. La differenza tra i dati area o lattice deriva dalla conformazione delle aree interessate, la prima va a definire delle zone irregolari e basate su confini amministrativi come comuni,

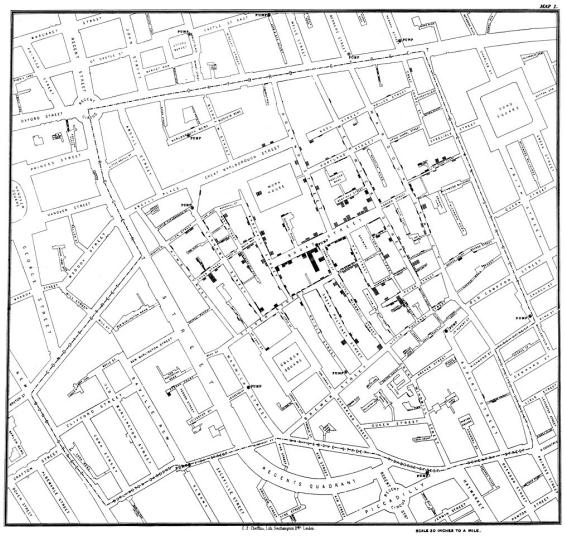


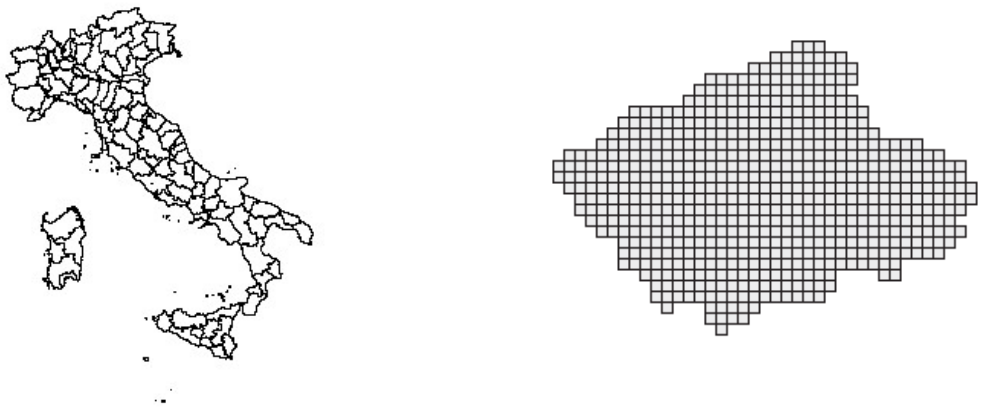
Figura 3.1: Distribuzione dei casi di colera accertati nel 1854 nel quartiere di Soho a Londra

province o regioni, mentre la seconda definisce delle zone regolari. In genere i *lattice data* sono frutto di una suddivisione arbitraria del territorio in un insieme di aree uguali fra loro. In Figura 3.2 si possono osservare degli esempi di tali strutture di dati;

- **Dati geostatistici**: Quando $y(s)$ è una osservazione casuale con una specifica posizione, solitamente identificata con latitudine e longitudine, che varia in modo continuo nello spazio chiuso $\mathcal{D}$. In genere la posizione di s è definita in due dimensioni, ma occasionalmente può anche considerare l'altitudine. Essi possono essere rappresentati da una collezione di osservazioni $y = (y(s_1), ..., y(s_n))$ dove i punti $s_1, ..., s_n$ indicano la località in cui si è raccolta la misura. In Figura 3.3 si può osservare come si presentano tali dati;

- **Processo puntuale** (*spatial point process*): Quando $y(s)$ rappresenta l'avvenimento o meno di un evento e le località stesse sono casuali. Il dominio spaziale $\mathcal{D}$ è un set di punti in $\mathbb{R}^d$ dove le osservazioni vengono raccolte. Per esempio, noi potremmo essere interessati alle posizioni di alcune specie di alberi in una foresta, in questo caso le località s sono casuali e le misure $y(s)$ possono prendere il valore 0 o 1 a seconda

della presenza o meno della caratteristica d'interesse. Come esempio di processi puntuali si può osservare la Figura 3.4.



(a) Area data **(b)** Lattice data

Figura 3.2: Esempio di area e lattice data, rispettivamente le province italiani e la Svizzera (Blangiardo e Cameletti, 2015)

Nel momento in cui si lavora con i dati spaziali è bene tenere conto della componente geografica, dato che questa ulteriore informazione può far evitare distorsioni nelle stime delle quantità d'interesse. In questa tesi si considererà la prima tipologia di dati, in particolare i **dati area**.

Nel momento in cui si utilizzano delle osservazioni di tipo spaziale è bene sottolineare come queste non possano essere indipendenti fra di loro. Come detto precedentemente se gli eventi si sviluppano in alcune zone ci sarà una componente di dipendenza che le lega e questa è chiamata **autocorrelazione spaziale** di cui si discuterà i modo più approfondito nel successivo paragrafo. In questa circostanza è però possibile considerare una indipendenza condizionata. Tale assunzione è meno restrittiva rispetto all'indipendenza totale tra le osservazioni, ma va a descrivere come ci possa essere un effetto non osservato che introduce dipendenza e che quindi deve essere considerato.

Questo accade molto spesso nel *disease mapping* dove alcune variabili che hanno una componente spaziale possono essere escluse erroneamente e introdurre nei dati una componente di dipendenza spaziale. Un buon approccio è

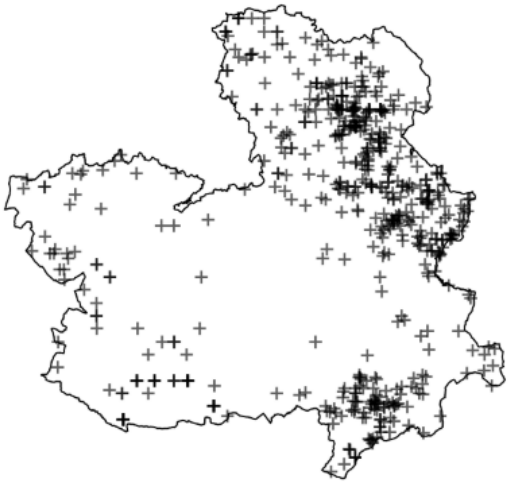


Figura 3.3: Esempio di dati geostatistici, in cui sono state rilevati gli incedi nella Castiglia (Blangiardo e Cameletti, 2015)

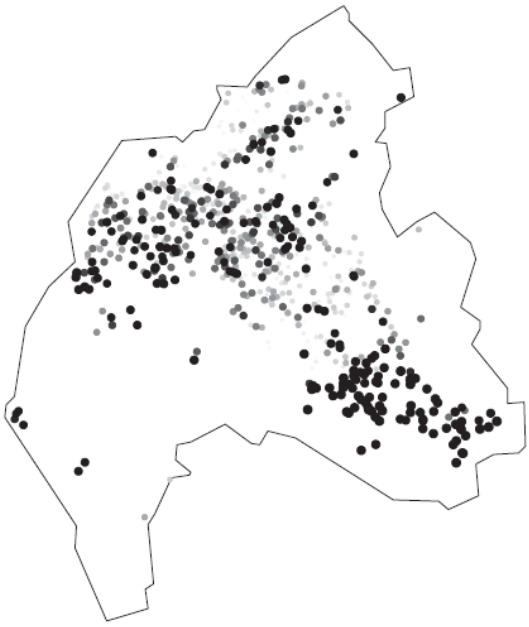


Figura 3.4: Esempio di un processo puntuale, in cui sono state rilevati i casi di problemi dentali nella regione della *North Cumbria* (Blangiardo e Cameletti, 2015)

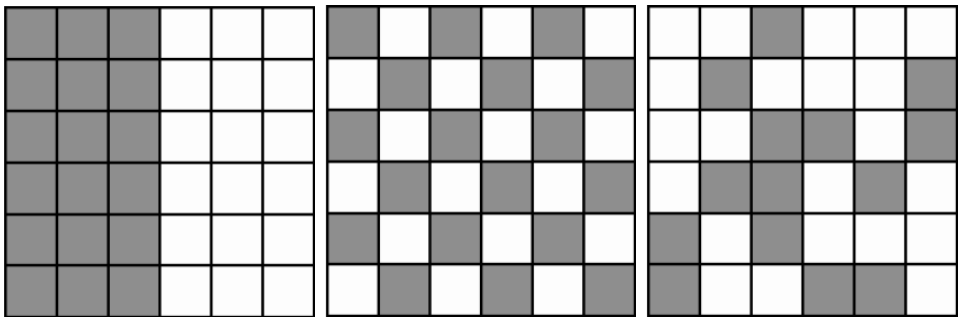
dunque considerare questo legame nella variabile d'interesse e la modellazione bayesiana gerarchica descritta nel Paragrafo 1.4 si adatta perfettamente a questo problema. L'operazione che si dovrà fare sarà costruire la verosimiglianza nel modo classico come nell'equazione (1.2), perché le variabili casuali con densità $p(y_i;\theta)$ saranno condizionatamente indipendenti se la struttura che lega le diverse osservazioni è considerata all'interno della gerarchia che si è costruita sul parametro θ.

3.1.2 Autocorrelazione spaziale

L'autocorrelazione spaziale si riferisce alla presenza di una variazione sistematica di una variabile definita in uno spazio. Essa può essere di carattere positivo o negativo, tale differenza è legata a come si influenzano le varie aree: si è in un caso di autocorrelazione positiva se aree vicine tendono ad assumere i medesimi valori per una variabile d'interesse e viceversa nel caso di autocorrelazione negativa. Essa si definisce come autocorrelazione perché identifica come i valori della stessa variabile sono dipendenti fra loro, anche se tali valori sono osservati in regioni diverse. In Figura 3.5 si può notare come si potrebbero disporre dei valori nello spazio a seconda della presenza e del tipo di autocorrelazione presente nelle osservazioni. Come si nota l'assenza di autocorrelazione è portata dal fatto che la variabile non si comporti in modo sistematico, ma del tutto casuale. È bene però tenere presente che vi possono essere caratteristiche peculiari dell'area d'interesse che fanno apparire il fenomeno come casuale, quindi si deve tenere in considerazione la similitudine generale fra le varie zone adiacenti prima di asserire la mancanza di autocorrelazione.

È importante identificare questo tipo di autocorrelazione perché fornisce un'indicazione della presenza di una meccanica interessante nella distribuzione della variabile nello spazio che può portare ad eseguire ulteriori investigazioni sul fenomeno in questione. Soprattutto perché la presenza di tale dipendenza introduce una ridondanza nell'informazione sulla variabile. Tale effetto distorsivo deve essere considerato al fine di ottenere delle buone analisi.

Per comprendere se vi sia presente o meno l'autocorrelazione spaziale



(a) Autocorrelazione positiva **(b)** Autocorrelazione negativa **(c)** Assenza di autocorrelazione

Figura 3.5: Disposizione spaziali di una variabile a seconda della presenza e tipo di autocorrelazione

per una data variabile si deve essere in grado di specificare la struttura di vicinanza che vi è fra le varie aree. Per semplificare la notazione precedente le aree $(s_1, ..., s_n)$ si definiscono $(1, ..., n)$, dove i vicini dell'area i si indicano con $\mathcal{N}_{(i)}$. In questi casi gli elementi di $\mathcal{N}_{(i)}$ sono le aree che condividono un confine con l'area i essi sono anche chiamati vicini di primo ordine. Si può decidere inoltre di inserire in questo gruppo anche i confini di secondo ordine, ovvero le aree che confinano con i vicini di primo ordine. Tramite questo principio si può costruire una matrice di vicinanza, W, che per l'intero spazio d'interesse $\mathcal{D}$ definisce quali sono le aree che si possono influenzare direttamente. La matrice W è quadrata e simmetrica e le sue celle si riempiono nel seguente modo:

$$w_{i,j} = \begin{cases} 1, & \text{se i} \sim \text{j}, \\ 0, & \text{altrimenti}, \end{cases}$$

dove con $i \sim j$ si indica che l'area i confina con l'area j. Gli elementi diagonali della matrice non vengono considerati, dato che una zona non può essere vicina di se stessa.

Nelle varie applicazioni la matrice W può essere standardizzata per riga o per colonna se vi è una particolare differenza di numero di vicini nelle diverse aree.

Si può notare come la matrice W possa essere molto sparsa soprattutto in presenza di un buon numero di aree. Questo elemento sarà fondamentale per

esprimere successivamente l'effetto di autocorrelazione spaziale utilizzando un GMRF descritto nel Paragrafo 2.2.

La costruzione della matrice di vicinanza ha un impatto importante sulle conclusioni inferenziali successive e sulla presenza dell'autocorrelazione spaziale. La criticità maggiore della definizione da parte dello sperimentatore della matrice W è legata al fatto che non considera fattori ambientali esterni, soprattutto nel caso in cui la diffusione di un certo fenomeno possa essere influenzato da tali fattori. Sono quindi state presentate diverse metodologie tra cui Lee e Mitchell (2011) e Ejigu e Wencheko (2020), i quali propongono di stimare la matrice W direttamente nel modello, utilizzando una funzione $f(z_j)$ dove i z_j sono indici di dissimilarità basati su alcune covariate. La problematica principale legata a questi metodi è dovuta alla scelta delle covariate, le quali devono saper esprimere coerentemente la similarità delle diverse aree a seconda delle caratteristiche del fenomeno che si sta diffondendo. Nel caso in cui il fenomeno non sia noto totalmente o non sia semplice recuperare delle variabili esplicative che influenzano la diffusione questi metodi sono di difficile utilizzo.

3.1.3 Statistica di Moran

L'autocorrelazione spaziale fra le osservazioni può essere valutata tramite alcuni indici. Uno di questi è la **statistica di Moran** o indice di Moran (Moran, 1950). Esso è un valore che fornisce una indicazione sul livello di autocorrelazione spaziale globale presente nei dati. Vi è un parallelo tra questo indice e quello di correlazione di Pearson: si deve però prestare attenzione che la statistica di Moran non definisce la forza della dipendenza spaziale, ma la sua presenza e il segno infatti sul valore che esso assume si deve verificare tramite un test di significatività. La sua formulazione è

$$I = \frac{n}{S} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2}, \tag{3.1}$$

dove con n si identifica il numero di aree considerate, $S = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}$ rappresenta il numero totale di collegamenti presenti fra le diverse aree, e infine con z_i si intende la deviazione dalla media del fenomeno d'interesse calcolato su ogni area. Questo indice, grazie alla quantità $\frac{n}{S}$, è normalizzato,

infatti esso assume i valori $-1 \leq I \leq 1$, dove per $I = 0$ si identifica l'assenza di autocorrelazione spaziale e per gli altri valori si definisce coerentemente presenza di autocorrelazione spaziale negativa o positiva.

Come specificato precedentemente l'indice di Moran deve essere valutato tramite una verifica d'ipotesi, dell'ipotesi $H_0 : I = 0$ contro l'alternativa $H_1 : I \neq 1$, in cui quindi si indica sotto l'ipotesi nulla che i dati sono distribuiti casualmente nello spazio e nell'ipotesi alternativa la presenza di una struttura di autocorrelazione positiva o negativa. La valutazione del test viene fatta in un contesto di statistica classica. Per l'indice I sono disponibili le formulazioni di valore atteso e varianza sotto l'ipotesi nulla per poter costruire la statistica test e tali valori sono

$$E(I) = -\frac{1}{n-1},$$

$$Var(I) = \left(\frac{n^2 C_1 - nC_2 + 3S^2}{S^2(n^2-1)} - \frac{1}{(n-1)^2}\right),$$

$$C_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2}{2},$$

$$C_2 = \sum_{i=1}^{n}(\sum_{j=1}^{n} w_{ij} + \sum_{i=1}^{n} w_{ij})$$

La statistica test ha la seguente forma

$$Z = \frac{I - E(I)}{\sqrt{Var(I)}} \underset{H_0}{\sim} N(0,1).$$

L'approssimazione normale della statistica test Z vale generalmente se la numerosità è adeguata ($n \geq 50$) e questo non sempre è disponibile. Un modo alternativo per valutare la significatività del test è l'approccio Monte Carlo: vengono generati dei valori di I sotto l'ipotesi nulla permutando casualmente gli elementi della matrice W e si confronta il valore osservato i valori generati. Questa strada è preferibile nel momento in cui non si è sicuri che le assunzioni distributive su Z siano verificate, quindi si può scegliere un approccio non parametrico.

La statistica di Moran può essere anche visualizzata diversamente dalla semplice formula mostrata in (3.1), infatti si può presentare una interpretazione alternativa che fornisce più informazione sull'autocorrelazione spaziale.

Esso è pari al coefficiente angolare di una regressione lineare semplice dove come variabile esplicativa si utilizzano i valori osservati y_i e come variabile d'interesse si utilizzano i valori $\sum_{j=1}^{n} y_{ij} w_{ij}$, ovvero la somma dei valori osservati dei vicini per ogni area i. Così facendo si può visualizzare tramite un grafico di dispersione se l'autocorrelazione è influenzata da particolari punti leva oppure quali aree sono circondate da aree che presentano un comportamento simile.

3.2 Modelli spaziali per dati di conteggio

Nella modellazione spaziale è frequente incontrare dei dati di conteggio su una popolazione in diverse aree. Nel caso in cui le osservazioni che presentano la caratteristica d'interesse siano relativamente poche se rapportate alla popolazione totale si è soliti formulare un modello di Poisson

$$Y_i \sim Po(\mu_i),$$

dove il parametro μ_i viene scomposto in due parti: una che esprime il rischio relativo dell'area in questione, θ_i, e l'altra rappresenta il numero di casi attesi in tale area, e_i. La relazione è quindi $E(Y_i) = \mu_i = e_i \theta_i$.

Il valore e_i viene considerato come un *offset*, dunque viene inserito all'interno della regressione per considerare le diversità nel numero di persone a rischio nelle diverse regioni. Dato che è un *offset* deve avere un valore noto. Questo dipende generalmente dalla disponibilità di informazioni in nostro possesso o dalle conoscenze sulle dinamiche di diffusione del fenomeno. Nel caso in cui non si posseggano abbastanza dati il suo valore potrebbe essere la popolazione totale della regione d'interesse. La definizione di e_i è comunque importante al fine del *disease mapping* perché andrà a influenzare la stima del rischio relativo nelle località che è di fondamentale importanza nella mappatura della malattia.

Un metodo più efficace per scegliere il valore di e_i è la standardizzazione indiretta che va a definire una popolazione standard calcolando un rischio globale nella macro zona interessata rapportandolo poi al numero di elementi

presenti nelle regioni. Quello che si ottiene è quindi

$$e_i = p_i R = p_i \frac{\sum y_i}{\sum p_i}$$

dove con p_i si identifica la popolazione totale della regione i-esima e la somma è calcolata su tutte le aree considerate nello studio. Essenzialmente questo calcolo distribuisce proporzionalmente il rischio in base alla popolazione presente. Anche il questo caso la selezione di p_i è cruciale per le conclusioni dello studio per quanto riguarda il rischio relativo. Esiste la possibilità di eseguire una standardizzazione più sofisticata aggiustando la stima di e_i utilizzando dei diversi substrati nella popolazione, magari considerando fattori di rischio rilevanti come età o sesso. Questa pratica può essere molto efficace nel caso in cui vi siano ampie differenze nella distribuzione spaziale nei diversi substrati e soprattutto se tali fattori di rischio sono incisivi sulla presenza o meno del fenomeno nelle diverse regioni.

Per quanto riguarda l'altra componente ovvero θ_i, esso è il parametro d'interesse e sarà definito tramite un predittore e una funzione di legame d'interesse, la quale generalmente è logaritmica

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

In η_i si potranno inserire diverse covariate d'interesse ed effetti casuali specifici per area. È da ricordare che essendo in un contesto bayesiano l'inserimento di effetti casuali piuttosto che covariate classiche è semplice dato che tutti i parametri presenti sono variabili casuali. Una importante caratteristica di η_i è che esso sarà un GMRF, quindi tutte le variabili che saranno inserite nel predittore avranno distribuzione normale. Mentre la dipendenza markoviana verrà spiegata nel Paragrafo 3.2.3.

3.2.1 Specificazioni distributive

Assumere che la variabile risposta abbia distribuzione Poisson deve essere fatto dopo un'attenta analisi preliminare. Gli errori possibili sono legati all'assunzione di identità tra la media e la varianza della variabile e al legame tra media e moda. Questi due casi possono essere trattati sviluppando dei

modelli diversi da quello di Poisson, rispettivamente un modello Binomiale Negativo e uno a inflazione di zeri (Salvan *et al.*, 2020).

Per quanto riguarda il primo caso, può accadere che i dati presentino varianza maggiore rispetto alla media per esempio nel caso in cui vengano trascurate delle variabili esplicative rilevanti. Una via per evitare questa problematica è assumere che la variabile risposta si distribuisca come una binomiale negativa, la quale mantiene la caratteristica di descrivere i dati di conteggio, ma presenta un parametro di dispersione che fa sì che la sua varianza sia maggiore rispetto alla media.

Nel secondo caso si può osservare che la frequenza dei conteggi pari a zero è superiore rispetto a quella attesa dal modello considerato. In questo caso si osserva che la distribuzione empirica della variabile risposta presenta un'inflazione di zeri. Questo fenomeno si manifesta quando i conteggi riguardano caratteristiche sui soggetti che possono risultare del tutto inattive, oppure per quanto riguarda l'ambito spaziale, nel momento in alcune zone non è stato controllato il fenomeno d'interesse. Per risolvere tale questione si può usare un modello mistura tra tra una distribuzione che descriva i conteggi e una distribuzione degenere in zero. In questo caso, per esempio, il modello di Poisson con inflazione di zeri (ZIP) assume che:

$$Y_i \sim \begin{cases} 0, & \text{con probabilità } 1-\phi_i, \\ \text{Po}(\mu_i), & \text{con probabilità } \phi_i. \end{cases}$$

Una modellazione di questo tipo tiene anche conto di due diversi tipi di "zero", uno descritto dalla mancanza di malattia e uno che descrive la mancanza della ricerca sulla presenza del fenomeno.

Inoltre anche questo modello presenta sovradispersione se paragonato al modello di Poisson, dato che presenta un valore medio inferiore alla varianza.

3.2.2 Modelli con specificazione della matrice di varianza

Un metodo per modellare delle osservazioni che presentano autocorrelazione spaziale preso dalla geostatistica è di specificare tale dipendenza direttamente

nella matrice di varianza-covarianza di un effetto casuale (Best *et al.*, 2005). Il parametro θ_i sarà allora definito come:

$$\log(\theta_i) = \beta_0 + S_i,$$

dove per semplicità non vi saranno altre covariate, ma solo l'intercetta e $S = (S_1, ..., S_n)$ sarà un effetto spaziale multidimensionale definito come $S \sim MVN(\mathbf{0}, \Sigma)$.

Si nota come ora le osservazioni $y_i|\theta_i$ possano essere definite condizionatamente indipendenti dato che nella definizione di θ_i viene inserita una struttura che spiega l'autocorrelazione nei dati. Infatti $\Sigma = \sigma^2\Omega$ e l'elemento ω_{ij}di Ω descrive la correlazione che vi è fra S_i e S_j e quindi anche delle osservaioni y_i e y_j. Dato che il numero di parametri da stimare sarebbe $\frac{n(n+1)}{2}$, di solito per rendere più parsimonioso il modello gli elementi della matrice che esprimono i vari legami vengono descritti tramite una funzione della distanza d_{ij} fra i centroidi delle varie aree, cioè $\omega_{ij} = f(d_{ij}, \phi)$. Vi possono essere diverse specificazioni di f, dalla funzione esponenziale alla Mater Diggle e Giorgi, 2019, Cap. 3. In tutte le scelte il parametro ϕ controlla la velocità di indebolimento della correlazione al variare della distanza.

Vi sono due problematiche nell'utilizzo di questi modelli nell'ambito dei dati ad area, uno computazionale e uno prettamente statistico. Il primo è legato alla stima di tale modello, dato che le regioni possono essere qualche centinaio l'implementazione via *Markov chain Monte Carlo* è computazionalmente molto intensiva data l'inversione della matrice di varianza-covarianza di dimensione $n \times n$. Inoltre la specificazione di tale modello non soddisfa le assunzioni richieste per poter applicare il metodo INLA in modo efficiente.

Dal punto di vista statistico dato che questo modello deriva dell'ambito geostatistico, presenta delle caratteristiche che mal si adattano ai dati area, in particolare l'assunzione che tutti i conteggi siano concentrati nel centroide della regione e la conseguente modellazione della correlazione tramite la distanza fra questi punti. Questa formulazione originariamente è nata per spiegare come un fenomeno di possa distribuire in uno spazio conoscendo le coordinate geografiche dei punti in cui sono state fatte le rilevazioni. L'utilizzo della distanza nella spiegazione della correlazione è utile sia per distanze brevi che grandi, mentre nei dati ad area questa potenzialità viene persa per-

ché che la maggior parte delle aree presenteranno distanze simili. Inoltre con tale metodologia si otterranno dei rischi diversi nella medesima zona, ma non corrisposti da una effettiva osservazione nei diversi luoghi.

3.2.3 Modelli a effetti casuali

Una strada diversa per poter implementare dei modelli che tengano conto della dipendenza fra le varie osservazioni è tramite l'aggiunta di effetti casuali. Gli effetti che si aggiungono possono anche non presentare una struttura di autocorrelazione, ma essere indipendenti tra loro. Allora una prima modificazione del modello può essere quella di inserire nel predittore η_i un effetto casuale non correlato:

Effetto casuale indipendente

$$\begin{aligned} Y_i &\sim Po(e_i\theta_i) \\ \log(\theta_i) &= \beta_0 + \nu_i \\ \nu_i &\sim N(0, \sigma_\nu^2) \end{aligned}$$

Si costruisce così un modello gerarchico bayesiano, inserendo nel secondo livello della gerarchia delle variabili in numero pari alle osservazioni, che spiegheranno il fenomeno. L'uso di un effetto casuale indipendente è un'altra alternativa per considerare la possibile sovradispersione nei dati dato che con tale effetto si agisce sulla variabilità del fenomeno.

Per completare la specificazione del modello, la distribuzione a priori di β_0 sarà normale e non informativa, quindi con una varianza ampia. La medesima cosa verrà fatta per quanto riguarda l'iperparametro σ_ν^2, il quale non viene considerato fissato, ma possiede una distribuzione a priori; essa sarà la distribuzione coniugata gamma-inversa non informativa.

Modello CAR proprio

Il modello può essere modificato inserendo una componente casuale che riesca ad esprimere l'autocorrelazione spaziale. Però deve possedere delle caratteristiche differenti rispetto al modello illustrato nel Paragrafo 3.2.2. In particolare, dovrà presentare una struttura più sparsa così da poter alleggerire

il peso computazionale mantenendo l'efficacia esplicativa del metodo. Una prima diversa specificazione per l'effetto S_i visto in precedenza è fornita da Besag *et al.* (1991) con la prima formulazione dei modelli CAR (*Conditional Autoregressive*).

In questo contesto le relazioni fra le varie osservazioni sono modellate tramite degli effetti casuali che possiedono la proprietà di dipendenza Markoviana descritta nel Paragrafo 2.2, i vari S_i si presenteranno quindi come

$$S_i|S_{-i} \sim N\left(\sum_{i=1}^{n} a_{ij}S_j, \frac{\sigma_S^2}{\sum_{j=1}^{n} w_{ij}}\right),$$
$$a_{ij} = \frac{\alpha w_{ij}}{\sum_{j=1}^{n} w_{ij}}.$$

Dove σ_S^2 è un parametro che identifica un elemento di variabilità comune dei vari effetti casuali e α è un parametro che indica il peso dell'influenza dei vicini su un elemento. Come si è spiegato in precedenza, un singolo effetto casuale S_i avrà distribuzione normale e condizionatamente a tutti gli altri effetti casuali dipenderà solamente dai suoi vicini descritti dalla matrice W. Infatti solo gli effetti casuali che corrispondono a delle aree confinanti con l'area in questione possederanno un valore di w_{ij} pari a 1 e andranno ad influenzare la distribuzione dell'effetto dell'area i-esima. Viene anche dimostrato che la distribuzione congiunta di S sarà $S \sim MVN(\mathbf{0}, Q^{-1})$. Affinché questa distribuzione congiunta esista la matrice Q dovrà essere simmetrica e definita positiva. Essa in questo modello viene scomposta in $Q = D(I - \alpha W)$, dove D è una matrice diagonale che i cui elementi indicano la precisione per i vari effetti, I è la matrice identità e il termine α controlla l'ammontare della dipendenza spaziale. L'uso di questo elemento garantisce la proprietà a Q di essere definita positiva nel caso in cui esso sia contenuto nell'intervallo $\lambda_{min}^{-1}, \lambda_{max}^{-1}$, dove λ_{min} e λ_{max} sono il più piccolo e il più grande autovalore della matrice W. In questa formulazione si utilizza una distribuzione a priori gamma-inversa per σ_S^2 e Uniforme tra λ_{min}^{-1} e λ_{max}^{-1} per α.

ICAR

La specificazione precedente presenta un problema proprio per il parametro α in quanto risulta generalmente ostico da stimare. Vi è quindi una modi-

ficazione che di fatto elimina il parametro ponendolo pari a 1 e ottenendo il modello ICAR (*Intrinsic Conditional Autoregressive*). A questo punto la matrice Q risulta pari a $D - W$, portandone però il determinante a 0. Per questo motivo tale formulazione risulta impropria, la matrice Q non è più semi-definita positiva quindi non garantisce che l'integrale della densità di S sia finito e unitario. Essa a priori non è una densità. La densità congiunta sarà:

$$\pi(S|\sigma_S^2) \propto exp\left(-\frac{1}{2\sigma_S^2}S^T[D-W]S\right) \propto exp\left(\sum_{i\sim j}(S_i - S_j)^2\right)$$

Dato che tali differenze a coppie risulterebbero identiche anche aggiungendo qualsiasi costante a S il modello risulta non identificabile. Si deve quindi imporre il vincolo $\sum_{i=1}^{n} S_i = 0$.

Dal punto di vista computazionale l'eliminazione del parametro α va a limitare il costo nel calcolo del determinante, quindi il numero di operazioni per il calcolo della densità passa da un ordine $\mathcal{O}(n^3)$ a $\mathcal{O}(n^2)$, rendendo possibile l'utilizzo del metodo MCMC senza particolari problematiche. Per quanto riguarda la procedura INLA questa variazione fornisce delle differenze minime, vi è una leggera velocizzazione data dalla mancanza di un parametro per cui effettuare l'approssimazione, ma il guadagno risulta impercettibile.

BYM

Il modello BYM (Besag *et al.*, 1991) nasce dall'idea che gli elementi non osservati in uno studio possono influenzarne la variabilità, quindi per tenere conto di questo il predittore lineare legato a θ_i può essere descritto come:

$$\log(\theta_i) = \beta_0 + \nu_i + S_i,$$

includendo sia l'effetto casuale indipendente che l'effetto casuale correlato. In questo caso la formulazione che si utilizza per il termine S_i è quella ICAR definita precedentemente. Così facendo si riesce a cogliere sia la componente di autocorrelazione spaziale presente nelle osservazioni, sia l'eterogeneità in eccesso non spiegata da S_i e da eventuali covariate.

In generale i due effetti non sono ben identificabili, ovvero si può constatare l'effetto della somma di queste due componenti, ma questi non possono

essere osservati singolarmente. Un metodo per poter avere una misura di quanto le due componenti spiegano della varianza globale è il calcolo del contributo relativo della varianza o anche detta correlazione intraclasse. I valori di σ_S^2 e σ_ν^2 non possono essere usate direttamente per questa operazione in quanto una è legata alla varianza della specificazione autoregressiva condizionata, mentre l'altra deriva da una parte marginale. Quello che si può fare è calcolare la componente empirica marginale di σ_S^2

$$s_S^2 = \frac{\sum_{i=1}^{n}(S_i - \bar{S})^2}{n-1},$$

dove $\bar{S}$ è la media di S. Utilizzando questo valore comparandolo con la varianza dell'effetto casuale indipendente si ha

$$\sigma_E^2 = \frac{s_S^2}{s_S^2 + \sigma_\nu^2}.$$

Così facendo si riconosce la quantità di varianza che viene spiegata dalla componente autocorrelata.

Leroux

Il modello di convoluzione descritto in precedenza ha il difetto di possedere un problema di identificabilità. Inoltre dato che le due componenti di variabilità degli effetti casuali sono descritte a diversi livelli risulta più complessa la scelta delle distribuzioni a priori. Per risolvere queste problematiche è stata quindi proposta un'alternativa: il modello di Leroux (Leroux *et al.*, 2000). Questa nuova modificazione vuole rendere più esplicita la relazione fra i due effetti casuali. Si definirà quindi un unico effetto

$$u \sim N(\mathbf{0}, \tau^2[Q(\rho)]^{-1})$$

Diversamente dal modello BYM, in cui la varianza della componente $S_i + \nu_i$ era pari a $\sigma_\nu^2 I + \sigma_S^2 Q^{-1}$, ora la varianza presenta una diversa formulazione, in cui è presente un peso ρ e una varianza unica τ^2:

$$Var(u|\tau^2, \rho) = \tau^2(\rho Q + (1-\rho)I)^{-1}$$

Il modello risulta più flessile del precedente e tramite l'imposizione del valore del parametro ρ si può ottenere un modello ICAR ($\rho = 1$) o un modello ad effetti indipendenti ($\rho = 0$). In genere se si lascia il parametro libero,

esso viene stimato all'interno del modello e fornisce una indicazione del peso che le due componenti possiedono. Esso non può essere tuttavia visto come la frazione di varianza spiegata dal modello come vedremo nella sezione seguente. Generalmente come distribuzione a priori di questo parametro è assunta una uniforme nell'intervallo $(0, 1)$.

Dal punto di vista dell'effetto casuale spaziale, condizionatamente la sua distribuzione sarà diversa da quella descritta nel Paragrafo 3.2.3 dato che è stato aggiunto un parametro che va a mediare fra le componenti. La sua formulazione sarà quindi

$$S_i|S_{-i} \sim N\left(\frac{\rho\sum_{j=1}^{n} w_{ij}S_i}{\rho\sum_{j=1}^{n} w_{ij} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{j=1}^{n} w_{ij} + 1 - \rho}\right).$$

BYM2

Il modello BYM2 (Riebler *et al.*, 2015) deriva da una diversa parametrizzazione del modello BYM, la quale migliora il controllo dei parametri. Questo viene fatto utilizzando la tecnica *penalized complexity* (PC) che va a favorire i modelli dove i parametri hanno una chiara interpretazione, garantendo così la possibilità di definire delle distribuzioni a priori ben definite.

Il primo step che viene fatto nella costruzione del seguente modello è lo *scaling* della matrice Q dell'effetto casuale autocorrelato. Questo viene fatto per evitare la troppa influenza che la struttura delle osservazioni può avere sulla varianza globale. In particolare si costringe la media geometrica della varianza marginale del modello a essere pari a τ^2:

$$\sigma^2_{GV}(S) = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\tau^2 Q^{-1}_{ii}\right)\right) = \tau^2\exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(Q^{-1}_{ii}\right)\right).$$

Come si vede dall'equazione questo può avvenire solo se gli elementi diagonali della matrice Q^{-1} sono pari a 1, quindi se i vari S_i hanno varianza unitaria.

La seconda modificazione che viene fatta è di definire u nel seguente modo:

$$u = \tau^2(\sqrt{1-\rho}\nu + \sqrt{\rho}S_*)$$

in questo modo la sua matrice di varianza sarà pari a $\tau^2((1-\rho)I + \rho Q^{-1}_*)$, dove l'uso dell'asterisco a pedice sottolinea che in tali elementi è

stato effettuato lo *scaling*. Grazie a questa formulazione ora il parametro ρ può essere interpretato come la frazione di varianza spiegata dalla struttura di dipendenza spaziale.

Il *framework* PC è basato su quattro punti principali:

- il rasoio di Occam: modello semplici dovrebbero essere preferiti a modelli più complessi a meno che questi non portino un' adeguata capacità esplicativa;
- una misura di complessità: la distanza di Kullbac-Leibler;
- una penalizzazione: una deviazione dal modello più semplice è penalizzata con un decadimento costante;
- dimensione nota: l'utente ha un'idea della dimensione dei parametri d'interesse.

Grazie all'utilizzo di queste linee guida, si riescono a definire le distribuzioni a priori per i parametri τ^2 e ρ. Quello che si ottiene è una distribuzione di Gumbel di secondo tipo per τ e una distribuzione ad hoc definita nell'intervallo (0,1) per ρ. Ulteriori dettagli qui omessi possono essere osservati in Riebler *et al.*, 2015 Cap. 4. Questo modello è nato all'interno del progetto di INLA e risulta quindi applicabile solo in questo contesto. Esso è ancora in fase di sperimentazione. Tuttavia in seguito verrà comunque utilizzato per modellare i dati sul Covid-19 nel Capitolo 4.

3.3 Modelli spazio-temporali per dati di conteggio

Investigare solo la componente spaziale della diffusione di un fenomeno risulta riduttivo nel caso in cui esso si ripresenti a certe distanze temporali. In questo caso si necessita di uno strumento che tenga conto anche della variazione nel tempo del rischio di diffusione. In particolare tenere conto di questa componente apra alla possibilità di eseguire delle previsioni a diverse distanze temporali sul fenomeno d'interesse, come può essere per esempio il numero di persone che presentano una certa malattia o la quantità di pioggia

caduta in diverse aree. Tenere conto di quest'altra componente permette di verificare se vi sia effettivamente un cambiamento nei diversi periodi o se il rischio abbia una variazione soltanto spaziale.

La generalizzazione dei modelli precedentemente presentati è abbastanza agevole. Il processo deve solo essere aggiornato con l'aggiunta della componente temporale e prede la forma

$$Y(s,t) \equiv \{y(s,t), (s,t) \in \mathcal{D}\},$$

dove $\mathcal{D}$ è un sottoinsieme fissato in $\mathbb{R}^2 \times \mathbb{R}$, le osservazioni sono quindi rilevate in n aree e in T istanti temporali diversi. Questa variazione porta a diverse implicazioni a seconda del tipo di dato spaziale che si possiede. Per quanto riguarda i dati ad area l'uso dei GMRF può essere esteso includendo una matrice di precisione definita anche nella nuova dimensione, mantenendo però la struttura di relazione spaziale costante nel tempo.

In questa nuova formulazione i modelli presentati nel Paragrafo 3.2, variano dipendentemente dal fatto che la matrice dei dati passa da una dimensione $n \times 1$ a $n \times T$, quindi l'estensione si presenta come:

$$\begin{aligned} Y_{it} &\sim Po(\mu_{it}), \\ \mu_{it} &= e_{it}\theta_{it}. \end{aligned}$$

Ancora una volta vi è la scomposizione della media della distribuzione di Poisson in due componenti, dove il logaritmo del rischio relativo $log(\theta_{it})$ rimane il focus della modellazione. Per quanto riguarda la stima dell'*offset* e_{it} valgono le considerazioni fatte in precedenza e rimane valida la pratica della standardizzazione indiretta, la quale si effettua

$$e_{it} = p_{it}R_t = p_{it}\frac{\sum_i \sum_t y_{it}}{\sum_i \sum_t p_{it}},$$

in cui p_{it} identifica la popolazione totale della regione i-esima all'istante t-esimo. A seconda della grandezza degli istanti temporali questa può variare in modo significativo o meno. In generale ora il predittore lineare legato al logaritmo del rischio relativo si presenterà nella seguente forma:

$$\log(\theta_{it}) = \eta_{it} = \beta_0 + S_i + \nu_i + \delta_t + \phi_{it},$$

dove i nuovi termini δ_t e ϕ_{it} rappresentano rispettivamente il trend temporale e una interazione spazio-temporale. Vi possono essere diverse specificazioni per questi due elementi di natura parametrica o meno.

Trend parametrico

L'inserimento di un trend parametrico nel predittore lineare va a definire che δ_t assumerà una struttura fissata del tipo βt, il cui t è un indicatore dell'istante temporale considerato, per esempio se i dati sono annuali esso rappresenterà l'anno in cui è stata fatta l'osservazione, mentre β sarà il suo parametro associato. La formulazione del trend t può essere di diverso tipo, come per esempio parabolica. Un modello presentato da Bernardinelli *et al.* (1995) per la mappatura della diffusione di un fenomeno nel tempo presenta il seguente predittore lineare:

$$\eta_{it} = \beta_0 + S_i + \nu_i + (\beta + \gamma_i) \times t,$$

il quale presenta un trend temporale parametrico unico, al quale viene aggiunta una interazione spazio-temporale che va a modificare nelle varie aree il trend globale. La distribuzione di γ_i è la medesima di un effetto casuale indipendente, esso va quindi a rappresentare lo scostamento del trend di un'area rispetto all'andamento globale.

Trend non parametrico

Nella precedente formulazione l'imposizione di una struttura fissa al trend temporale può risultare parsimoniosa, ma troppo restrittiva. È possibile rilassare questa assunzione utilizzando un trend non parametrico, ovvero un trend descritto da un effetto casuale. Quindi il predittore sarà

$$\eta_{it} = \beta_0 + S_i + \nu_i + \gamma_t,$$

in cui la modellazione dinamica del trend verrà eseguita dal parametro γ_t, la sua distribuzione può essere di vario tipo:

- $\gamma_t \sim N(0, \sigma^2_\gamma)$, ovvero un effetto casuale indipendente;

- $\gamma_t \sim N(\gamma_{t-p}, \sigma^2_\gamma)$, un effetto casuale autoregressivo di ordine p, in genere questo viene definito di primo ordine. L'inserimento di questa componente è atta a cogliere una eventuale autocorrelazione temporale nei dati.

- $\gamma_t|\gamma_{t-1} \sim N(0, \sigma^2_\gamma)$, un effetto casuale *random walk* di prim'ordine. Anche in questo caso si possono utilizzare diversi ordini. A differenza del precedente questa struttura tiene conto di una possibile non stazionarietà del dati.

Tutte le distribuzioni sono sempre considerate Gaussiane, per mantenere la proprietà del predittore lineare di essere trattato come un GMRF.

Interazione spazio-temporale

Un'ultima possibile variazione temporale che si può inserire è legata al parametro ϕ_{it}, ovvero l'inserimento di una interazione spazio-temporale non parametrica che vada a modificare il trend temporale a seconda della diversa area d'interesse. Il vettore $\boldsymbol{\phi}$ possiede una distribuzione normale in cui la matrice di precisione è data da $\sigma_\phi^{-2}\mathbf{R}_\phi$ in cui $\mathbf{R}_\phi$ identifica la struttura di dipendenza spaziale o temporale che esso può esprimere. Tale matrice di precisione è fattorizzabile come il prodotto di Kronecker delle matrici di precisione dei corrispondi effetti casuali che interagiscono.

Le diverse specificazioni di tale interazione vengono chiamate:

- Type I: se rappresenta l'interazione due effetti casuali indipendenti, uno spaziale e uno temporale;

- Type II: se rappresenta l'interazione tra un effetto spaziale indipendente e uno temporale *random walk*;

- Type III: se rappresenta l'interazione tra un effetto spaziale della famiglia dei modelli CAR e uno temporale indipendente;

- Type IV: se rappresenta l'interazione tra un effetto spaziale della famiglia dei modelli CAR e uno temporale *random walk*.

In questa tesi viene considerato solo il primo tipo di interazione dato che i restanti impongono una struttura che risulta computazionalmente molto pesante e in più rendono molto complicato il modello. Per quanto riguarda l'interazione Type I la matrice di precisione per $\boldsymbol{\phi}$ è:

$$\mathbf{R}_\phi = \mathbf{R}_\nu \otimes \mathbf{R}_\gamma = I \otimes I = I$$

conseguentemente non si assume nessun tipo di dipendenza nell'interazione e i singoli ϕ_{it} avranno distribuzione $N(0, \sigma^2_\phi)$.

3.4 Validazione del modello

Come nella statistica classica, anche l'approccio bayesiano necessita di appositi indici per poter confrontare la bontà di adattamento di più modelli che si possono differenziare dalle distribuzioni a priori dei vari parametri o per le variabili esplicative inserite nel modello. Dato che i vari parametri sono descritti come variabili aleatorie risulta complesso usare i criteri d'informazione come l'AIC, perciò sono stati formulati diversi indicatori appositi basati sulla verosimiglianza.

Il primo di questi è il DIC (*Deviance Information Criterion*), il quale rappresenta esattamente la generalizzazione dell'AIC in ambito bayesiano. Esso si basa sulla devianza del modello, $D(\theta) = -2\log(p(y|\theta))$, la quale essendo una variabile aleatoria può essere utilizzata attraverso diverse misure di sintesi. La sua formulazione è

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\bar{\theta}),$$

dove con $\bar{D} = E_{\theta|y}(D(\theta))$ si intende il valore atteso della devianza, e $D(\bar{\theta})$ è la devianza calcolata alla media a posteriori del parametro, $\bar{\theta} = E(\theta|y)$ e p_D è una misura del numero effettivo di parametri. La sua derivazione viene mostrata in Spiegelhalter *et al.* (1998) e deriva dal valore atteso dell'espansione di Taylor della devianza attorno alla media a posteriori del parametro.

Un'altro criterio che può essere utilizzato è il WAIC (*Watanabe-Akaike Information Criterion*) che è la log densità predittiva a posteriori valutata per vari valori del parametro, penalizzata per il numero effettivo di parametri, in modo da evitare il sovradattamento. Questo indice risulta più stabile

rispetto al precedente e inoltre ha la proprietà di utilizzare la distribuzione a posteriori piuttosto di condizionarsi ad un punto stimato. Il valore si ottiene nel seguente modo:

$$WAIC = lppd - pWAIC$$

dove i vari elementi sono definiti dalle seguenti equazioni:

$$lppd = \sum_{i=1}^{n} \log \left(\frac{1}{G} \sum_{g=1}^{G} p(y_i|\theta^g) \right),$$
$$pWAIC = \sum_{i=1}^{n} V_{g=1}^{G}(\log(p(y_i|\theta^g)))$$

dove $V_{g=1}^{G}(a) = (G-1)^{-1} \sum_{g=1}^{G}(a_g - \bar{a})^2$ e con θ^g si identificano G punti della distribuzione a posteriori di θ. Per entrambi questi indici vale la regola classica dell'AIC, ovvero un valore minore di DIC o WAIC fornisce l'indicazione di un miglior adattamento del modello.

Una via diversa per poter valutare il modello è tramite la sua capacità predittiva, in particolare con un metodo legato alla convalida incrociata. In questa prospettiva si può utilizzare un indice detto CPO (*Conditional Predictive Ordinate*) (Lewis *et al.*, 2013), esso viene calcolato per ogni osservazione e valuta la capacità del modello di prevedere quella particolare unità. Il metodo di valutazione deriva dalla *cross validation leave-one-out* quindi si stima il modello senza una osservazione e la si prevede. Lo sviluppo semplificato del calcolo dei CPO è il seguente

$$CPO_i = p(y_i|y_{(-i)}) = \int p(y_i|\theta)p(\theta|y_i)d\theta = \left(\int \frac{1}{p(y_i|\theta)} p(\theta|y)d\theta \right)^{-1},$$

per lo sviluppo completo si veda Lewis *et al.* (2013). I valori dei CPO possono essere rappresentati in modo empirico sostituendo l'integrale con una media per diversi valori a posteriori del parametro. La formulazione indica che i CPO possono essere calcolati senza dover eseguire n volte la stima del modello. Quello che si ottiene è dunque una probabilità, più essa sarà vicina a 1 più il modello sarà in grado di stimare correttamente le osservazioni. Dunque questa misura può essere utilizzata come indicatore della capacità previsiva del modello.

Dato che in questo caso si è in possesso di n valori del CPO, generalmente si utilizza una misura di sintesi che tramite un valore riassuma tutta l'informazione dei CPO. Questa è la log verosimiglianza pseudomarginale, il cui nome deriva dalla scomposizione che si ottiene calcolando il logaritmo del CPO, essa si presenta come:

$$LPML = \sum_{i=1}^{n} log(\widehat{CPO}_i).$$

dove ai CPO teorici è sostituita loro stima empirica. I valori della LPML forniti da diversi modelli possono essere confrontati per identifica quale modello abbia una maggiore capacità predittiva. Anche in questo caso si segue la regola che il modello con minor valore della LPML sarà quello migliore.

Capitolo 4

Modelli spazio-temporali per la diffusione del Covid-19 in Italia

All'inizio del 2020 il mondo intero è stato colpito da una grave pandemia, il Covid-19, causata dal virus SARS-CoV-2. Tale virus si è diffuso a partire dalla città di Wuhan situata nella provincia di Hubei in Cina. Non è ancora chiaro quando il virus abbia iniziato a diffondersi tra le varie nazioni, alcuni studi da parte dell'Istituto Superiore di Sanità affermano di aver trovato tracce del suo materiale genetico già a dicembre nelle città di Milano e Roma, (La Rosa *et al.*, 2020). Per quanto riguarda i dati ufficiali, il primo contagio rilevato in Italia risale al 23 febbraio in Lombardia, di lì poco molti altri Stati inizieranno ad identificare un gran numero di contagiati tanto che l'Organizzazione Mondiale della Sanità (OMS) dichiarerà lo stato di pandemia, (OMS, 2020).

Dato l'effetto così importante che questo virus ha prodotto sulla vita della popolazione mondiale, tutta la comunità scientifica si è prodigata a studiare, tramite diversi metodi, la diffusione del contagio del virus, analizzando come questo sia avvenuto, quali fattori possono aver contribuito e se le azioni attuate dai vari governi per limitare la diffusione del virus siano state efficaci. In questi termini sono stati prodotti vari elaborati che, utilizzando una grande varietà di modelli statistici ed epidemiologici, hanno cercato di dare il loro contributo nella comprensione del fenomeno. Alcuni di questi sono stati prodotti da ricercatori dell'Università di Padova, tra cui Lavezzo *et al.* (2020)

e Gatto *et al.* (2020), i quali hanno analizzato rispettivamente il processo di diffusione tracciando i contatti fra i cittadini della comunità di Vo' Euganeo e l'efficacia delle misure di contenimento attuate in Italia. Oltre a questi, vari elaborati, tra cui J. Wang *et al.* (2020), Briz-Redon e Serrano-Aroca (2020), Dominici *et al.* (2020) e Setti *et al.* (2020), hanno studiato la relazione tra la diffusione del virus e diversi fattori ambientali come temperatura, umidità e qualità dell'aria. Da queste analisi svolte in diverse nazioni che presentano caratteristiche differenti, sembra evincere che vi sia una sostanziale relazione positiva tra questi fattori e la diffusione del Covid-19, al netto di altre variabili socioeconomiche e demografiche.

Infine in altre pubblicazioni come Kang *et al.* (2020), Fronterrè *et al.* (2020) e Giuliani *et al.* (2020) si è posta l'attenzione sulla dinamica spazio-temporale della diffusione del Covid-19, andando ad analizzare accuratamente se vi fosse una dipendenza fra i contagi rilevati in diverse regioni di un singolo stato oppure la presenza di particolari *hotspot*.

In quest'ottica questo capitolo presenterà l'analisi dei dati ufficiali italiani Lanera *et al.* (2020) riguardanti i contagiati da SARS-CoV-2 al fine di identificare se alcuni modelli spazio-temporali possano essere degli strumenti idonei per fare delle previsioni sul numero di contagi giornaliero, sulla base della dipendenza spazio-temporale e sulla base di altre informazioni ausiliarie.

In questo capitolo vengono utilizzati gli strumenti presentati precedentemente per analizzare i dati relativi alla diffusione del Covid-19 tramite l'adattamento dei modelli spazio-temporali e la loro validazione tramite uno studio previsivo. Si inizia presentando i dati e descrivendone le caratteristiche principali, successivamente si adattano modelli prima spaziali e poi spazio-temporali senza l'uso di covariate. Si presentano i risultati delle capacità previsive degli stessi ed infine si inseriscono delle covariate legate a caratteristiche economiche e demografiche. Purtroppo non si è potuto utilizzare delle variabili esplicative che rappresentassero i fattori ambientali com'è stato fatto in altre analisi perché non erano disponibili i dati per tutte le province italiane.

Tutte le analisi di questo capitolo sono effettuate sul server Hactar messo a disposizione dal Dipartimento di Scienze Statistiche dell'Università di Padova.

4.1 I dati

I dati analizzati in questa tesi sono forniti dalla Protezioni Civile Italiana, in particolare è stata utilizzata la libreria `covid-19-ita` (Lanera *et al.*, 2020) contente le osservazioni dell'epidemia di coronavirus per varie suddivisioni del territorio. Per quanto riguarda i dati provinciali sono disponibili solo i contagiati, mentre per le osservazioni nazionali e regionali sono disponibili informazioni relative ai morti e al numero di ricoveri in terapia intensiva. I dati utilizzati nelle seguenti analisi si riferiscono al numero di contagiati da Covid-19 giornalieri in tutte le province italiane dal 24 Febbraio al 28 Giugno 2020.

Le rilevazioni derivano dai tamponi effettuati sia per casi clinici, ovvero chiunque si sia rivolto alle strutture ospedaliere, sia per i controlli effettuati dalle varie regioni seguendo diverse politiche di *screening*. Una problematica legata a queste osservazioni deriva dal fatto che per alcuni soggetti positivi sono stati effettuati più tamponi, tendenzialmente tre, e ogni volta che il soggetto risultava positivo veniva segnalato. Questo implica che i vari positivi osservati non rappresentato del tutto nuovi soggetti che sono risultati positivi, ma piuttosto il numero di positivi attivi presenti in un dato giorno. Inoltre per le province non è disponibile il numero di tamponi effettuati in un singolo giorno, questo poteva essere utilizzato nel caso si fosse deciso di assumere una distribuzione binomiale per la variabile d'interesse. È da notare come le politiche di controllo del contagio siano state diverse nelle varie regioni, ciò può aver portato a sottostimare l'effettiva portata del contagio in diverse zone.

Le osservazioni vengono aggiornate ogni giorno e fornite tramite conteggio cumulato, quindi per ottenere il formato desiderato si è dovuta applicare una differenza prima nel numero di contagiati per provincia.

Data la situazione critica nella quale sono stati raccolti i dati essi presentano delle incoerenze: in 229 casi il numero cumulato di contagiati per un giorno era inferiore al numero di contagiati del giorno precedente. Come si può vedere dalla Tabella 4.1 questi valori sono abbastanza equidistribuiti nello spazio e data la natura di conteggio dai dati si è deciso di imputare il loro valore a 0. In totale le osservazioni sono 13482, divise nelle 107 province

italiane per 126 giorni di osservazione.

Tabella 4.1: Numero di incongruenze osservate nelle diverse regioni Italiane

Regione	Abruzzo	Basilicata	Calabria	Campania
Valore	11	4	12	11
Regione	Emilia-Romagna	Friuli Venezia Giulia	Lazio	Liguria
Valore	28	11	12	4
Regione	Lombardia	Marche	Molise	Trentino Alto Adige
Valore	23	7	1	4
Regione	Piemonte	Puglia	Sardegna	Sicilia
Valore	22	18	7	13
Regione	Toscana	Umbria	Veneto	
Valore	22	4	15	

Dal punto di vista spaziale i conteggi cumulati alla fine della raccolta dei dati presentano la distribuzione mostrata in Figura 4.1. Come si nota vi è una forte diversità tra le province del Nord e quelle del Sud, sia per quanto riguarda il numero di contagi sia per il comportamento fra province adiacenti. Nel Nord si nota la presenza di zone che condividono alte numerosità, mentre al Centro-Sud vi è più dissimilarità fra le diverse aree.

Come descritto nei Paragrafi 3.1.2 e 3.1.3 si deve verificare la presenza di autocorrelazione spaziale fra le osservazioni.

La matrice di vicinanza W è stata costruita tenendo conto della condivisione di un confine fra le varie province, considerando dunque solo i vicini di prim'ordine. Per osservare quanto la matrice W risulti sparsa si calcola la sua densità:

$$D = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}{n(n-1)} = \frac{476}{107 \times 106} = 0.042.$$

Si può confermare la sparsità della matrice W necessaria per l'efficienza delle operazioni di calcolo di INLA, infatti solo poco più del 4% delle celle di W descrivono un collegamento fra le province.

Questa operazione può essere fatta utilizzando la statistica di Moran, ma dato che questo indice non tiene conto della componente temporale, si deve fotografare la diffusione del contagio e successivamente calcolare l'indice. In questo caso verranno mostrati i risultati della regressione lineare semplice

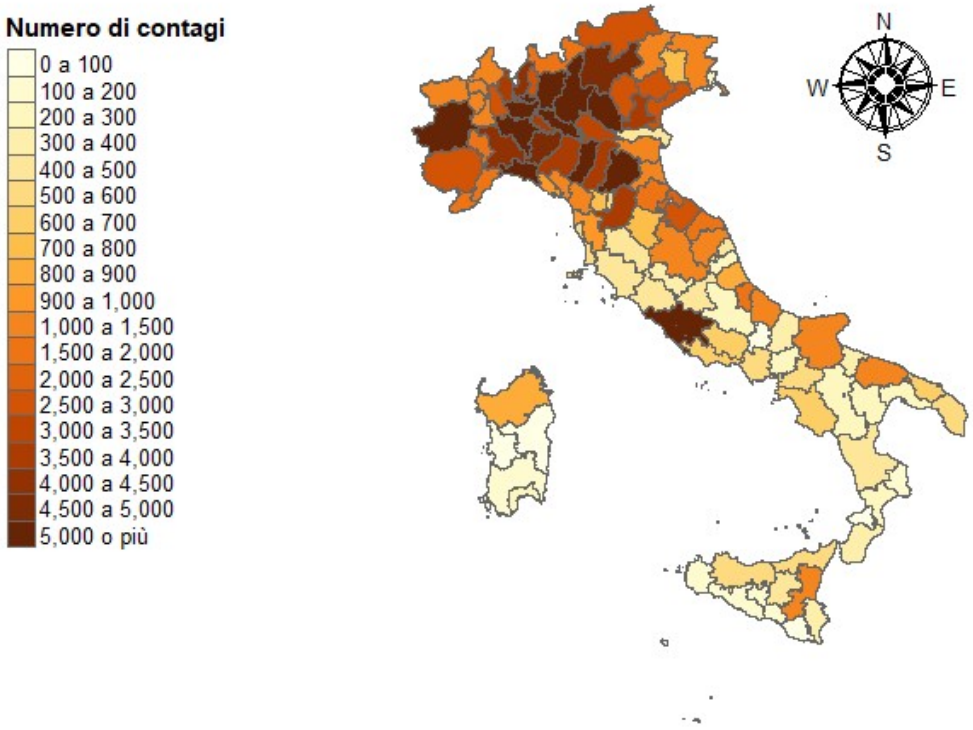


Figura 4.1: Numero cumulato di contagiati al 28 Giugno nelle varie province italiane

che descrive la statistica di Moran per i dati cumulati alla fine del periodo di osservazione. In Tabella 4.2 si mostrano i valori e *p-value* della verifica d'ipotesi per validare la presenza di autocorrelazione spaziale, calcolati a diversi intervalli temporali, sempre cumulando le osservazioni alla data definita. I valori dei *p-value* vengono moltiplicati per 10^2 per renderli più leggibili. Essi si assumono tutti lo stesso valore perché sono stati calcolati tramite il metodo Monte Carlo con 10000 replicazioni, perciò il valore che si ottiene deriva da un limite numerico.

La rappresentazione in Figura 4.2 oltre a fornirci una idea dell'autocorrelazione spaziale presente nei dati, è in grado di mostrare quali siano le province che più influenzano tale valore. Infatti sono presenti dei punti leva che corrispondono alle province di Milano, Brescia, Torino e Bergamo, le aree che sono state più colpite dall'epidemia. Esse tendono a far diminuire il valore del coefficiente angolare della retta di regressione, distanziandosi dalla nuvola di punti principale. Questo è dovuto al fatto che sono state le prime provincia a subire con forza la diffusione della malattia, mentre le altre zone sono più simili tra loro probabilmente per la chiusura dei confini. La retta mantiene comunque una buona pendenza pari a 0.45, il che fa sembra che all'interno delle osservazioni vi sia un buon livello di autocorrelazione.

Tabella 4.2: Valori della statistica di Moran e relativo *p-value* per diversi intervalli temporali

Data	03-03-2020	15-03-2020	30-03-2020	12-04-2020
Stat. Moran	0.52	0.62	0.54	0.49
p-value	0.0099	0.0099	0.0099	0.0099
Data	28-04-2020	20-05-2020	02-06-2020	28-06-2020
Stat.Moran	0.45	0.44	0.44	0.45
p-value	0.0099	0.0099	0.0099	0.0099

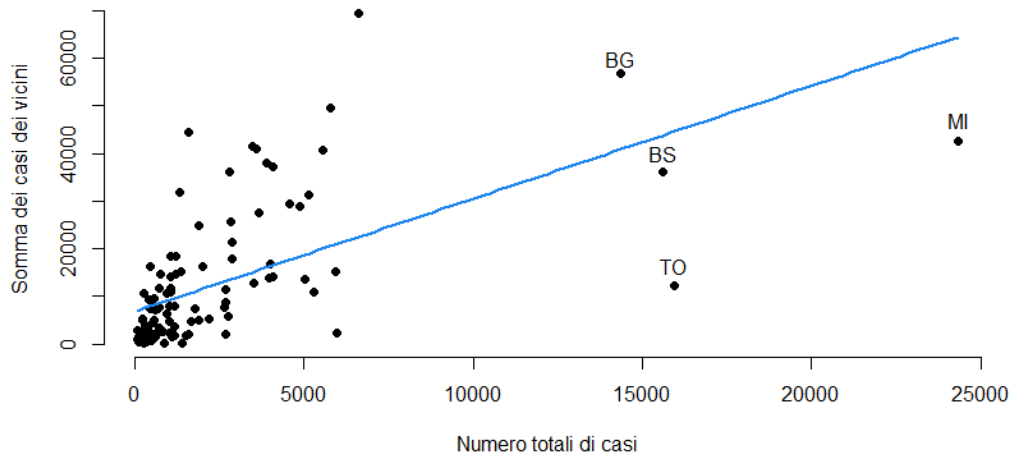


Figura 4.2: Regressione lineare semplice corrispettiva al calcolo della statistica di Moran

4.2 Analisi tramite effetti spazio-temporali

4.2.1 Analisi spaziale

Per prima cosa si è effettuata un'analisi puramente spaziale sui valori cumulati alla fine del periodo di rilevazione. I diversi modelli in questo caso non presentano alcuna covariata e differiscono tra loro rispetto alla componente casuale di dipendenza spaziale. Quest'analisi preliminare è fatta per comprendere il comportamento di questi effetti casuali e per determinare quale sia la distribuzione a priori migliore per gli elementi S_i, la quale viene poi utilizzata per l'adattamento spazio-temporale. I modelli adattati sono: GLM senza componente spaziale, un effetto casuale non correlato, ICAR, BYM, Leroux, BYM2. Al primo passo è utilizzata una distribuzione di Poisson per la variabile risposta, successivamente viene usata una distribuzione Binomiale Negativa per osservare se vi fossero dei miglioramenti nell'adattamento. Infine per poter ottenere un confronto fra il metodo MCMC e INLA dal punto di vista della somiglianza dei risultati ottenuti gli adattamenti sono utilizzati entrambi i metodi, in particolare il metodo MCMC utilizzato è il

Metropolis-Hastings tramite la libreria `CARBayes` (Lee, 2013). Per quanto riguarda le metriche per la stima delle distribuzioni a posteriori si sono generati 94000 valori, partendo da 200000 unità alle quali è stato tolto un *burn-in* di 30000 e successivamente si è applicato un filtro tenendo un valore ogni 5 per diminuire l'autocorrelazione.

Nella Tabella 4.3 vengono mostrati i risultati dei vari modelli tramite diversi indici di bontà di adattamento dove il *p-value* del test di Moran è calcolato sui residui di Pearson.

Tabella 4.3: Indicatori della bontà di adattamento dei vari modelli con distribuzione di Poisson

Modello	DIC	pD	$WAIC$	$LMPL$	Moran *p-value*
GLM	186230.41	1.00	86087.28	-40467.44	0.000
IND.	1167.99	114.81	1144.86	-742.78	0.000
ICAR	1153.00	106.80	1124.55	-659.30	0.998
BYM	1157.96	109.39	1129.03	-660.75	0.997
LEROUX	1166.51	113.60	1143.33	-706.02	0.999
BYM2	1152.45	106.70	1120.82	-661.05	0.997

Si nota come l'inserimento di un effetto casuale spaziale riesca a cogliere l'autocorrelazione presente nelle osservazioni. I primi cinque modelli sono stati adattati tramite il classico algoritmo di Metropolis-Hastings, ma i risultati degli stessi modelli adattati con INLA sono del tutto analoghi. Per quanto riguarda il BYM2 essendo stato formulato all'interno del progetto INLA esso è stimato solamente con questo metodo.

Ora si può andare a confrontare i modelli sulla base dell'assunzione scelta nel modello statistico probabilistico, ovvero se la distribuzione migliore per descrivere il fenomeno di diffusione sia Poisson o Binomiale Negativa. Da qui in poi tutti i modelli verranno stimati tramite INLA, dato che questa specificazione non è utilizzabile tramite la libreria `CARBayes`. I risultati sono riportati in Tabella 4.4

Si sono omessi i *p-value* dato che le specificazioni spaziali erano le medesime. Dai valori degli indicatori sembra che la distribuzione di Poisson sia più adatta a spiegare i dati rispetto alla Binomiale Negativa. In particolare, si

Tabella 4.4: Indicatori della bontà di adattamento dei vari modelli con distribuzione Binomiale Negativa

Modello	DIC	pD	$WAIC$	$LMPL$
GLM	1803.22	2.01	1802.89	-901.45
IND.	1803.13	2.02	1802.86	-901.43
ICAR	1718.63	20.41	1716.54	-1377.45
BYM	1718.30	20.52	1716.31	-1037.79
BYM2	1718.67	25.25	1720.95	-1389.03

possono osservare diverse particolarità: i primi due modelli forniscono risultati pressoché identici, come avviene per gli ultimi modelli, il che dimostra come l'inserimento di una componente di sovradispersione aggiuntiva alla specificazione del modello risulta inutile. Questo avviene perché è già presente nel modello una componente che governa la sovradispersione. Inoltre i valori di pD sono estremamente più bassi rispetto alla tabella precedente indicando come la diversa specificazione porti a una parsimonia maggiore da parte del modello, la quale però non corrisponde a una capacità esplicativa altrettanto adeguata. Non trattandosi di parametri nel senso classico del termine si considera come migliore la specificazione Poisson del modello. Successivamente si attuerà comunque l'analisi spazio-temporale anche con diverse assunzioni distributive. Perciò osservando la Tabella 4.3 sembra che la migliore specificazione della distribuzione a priori della componente spaziale sia l'ICAR. Dato che il modello BYM2 possiede delle buone statistiche di sintesi possiamo osservare come il parametro che denota la quantità di variabilità colta dalla componente autocorrelata al suo interno sia pari a 0.229 con intervallo di credibilità (0.118;0.375), quindi più del 20% della variabilità è colta dal questa componente.

Globalmente si può osservare in Figura 4.3 come la distribuzione a posteriori dei diversi η_i sia molto simile tra i due diversi metodi di stima. Quindi le stime che si possono ottenere dagli adattamenti in un caso o nell'altro non portano grandi differenze.

Se si osserva più nello specifico si possono trovare però delle differenze. Per esempio osservando la distribuzione a posteriori dell'effetto casuale spaziale

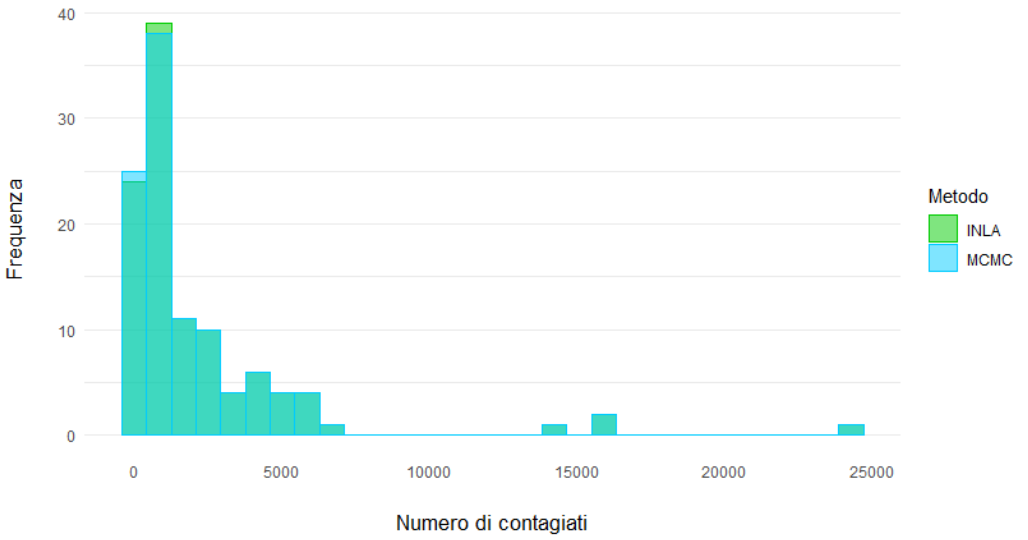


Figura 4.3: Distribuzioni a posteriori stimate di η_i nella metodologia MCMC e INLA

per la provincia di Padova in Figura 4.4, si può osservare come queste siano diverse in vari termini: la varianza non risulta la medesima, in quanto INLA porta a una distribuzione molto più ampia con una varianza pari a 0.219 contro 0.00026 descritta dal metodo MCMC, la curtosi non è la stessa infatti vi è più pesantezza nella distribuzione di destra. Le media a posteriori è invece molto simile, pari rispettivamente a 0.0616 per MCMC e 0.0614 per INLA. Questa differenza di risultati può essere dovuta alla difficoltà che INLA può aver riscontrato nell'approssimazione normale dell'effetto casuale. Infatti sembra che la correzione per la pesantezza nelle code non abbia portato a dei buoni risultati. Questo è accaduto per la distribuzione di varie province, segno che tale metodologia in questo contesto porta a dei risultati corretti per la posizione della distribuzione, ma la sua forma sembra abbastanza instabile.

Una successiva analisi in Figura 4.5 può essere osservare la probabilità di rischio in eccesso che è presente nelle diverse aree che può essere trovata con $Pr(\eta_i > 0|y)$ oppure equivalentemente osservando $Pr(\lambda_i > 1|y)$, utilizzando il rischio relativo. Così facendo si può inferire quali aree presentano una probabilità maggiore rispetto ad una situazione stazionaria. Questa misura fornisce più informazione, dato che va a utilizzare tutta la distribuzione a

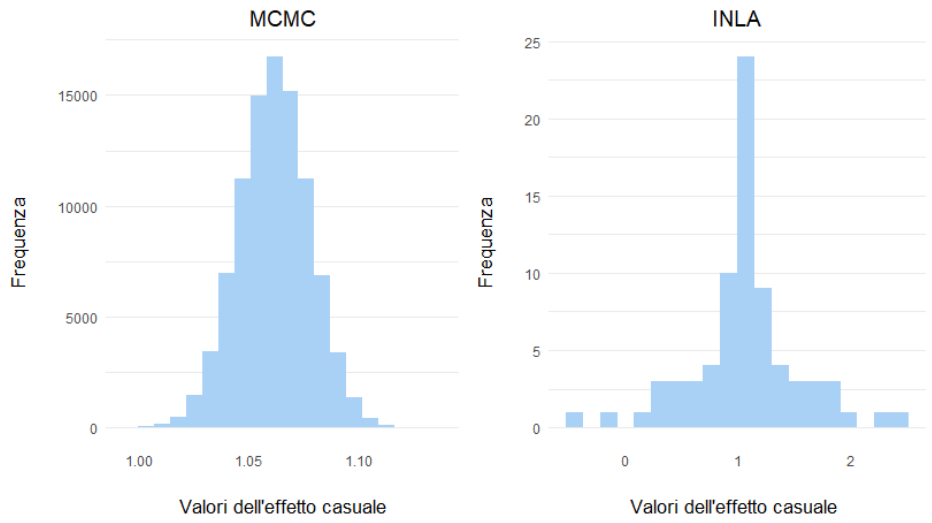


Figura 4.4: Distribuzioni a posteriori stimate delll'effetto casuale spaziale per la provincia di Padova nella metodologia MCMC e INLA

posteriori e non sono un valore puntuale. Dalla mappa si nota come vi sia una clusterizzazione che va a dividere il Nord dal Sud, dove si passa da una probabilità molto elevata di rischio in eccesso a una quasi nulla. Un caso particolare è in Sicilia in cui la provincia di Enna, dove il numero di contagi era più elevato delle altre zone se rapportato alla popolazione a rischio.

Infine si può osservare in Figura 4.6 direttamente il rischio relativo stimato dal modello ICAR così da poter inferire sulla diffusione della malattia in tutta Italia. Le ultime due analisi sono sempre effettuate utilizzando i contagi cumulati al 28 Giugno. I risultati sembrano coerenti con quanto ci si potesse aspettare, in Lombardia vi è un gruppo di province che mostrano un rischio elevato e le altre zone limitrofe tendono ad avere un rischio minore tanto più sono distanti dall'area che si può definire il fulcro dell'epidemia. Non tutto il Nord risulta colpito nello stesso modo, come si nota dal comportamento di Udine e Rovigo. Potrebbero esserci delle caratteristiche di queste zone che non hanno aiutato il contagio. Il resto dell'Italia presenta un rischio relativo abbastanza controllato, a parte in alcune sporadiche zone questo potrebbe essere un segno di come la chiusura al movimento degli individui possa aver portato dei benefici.

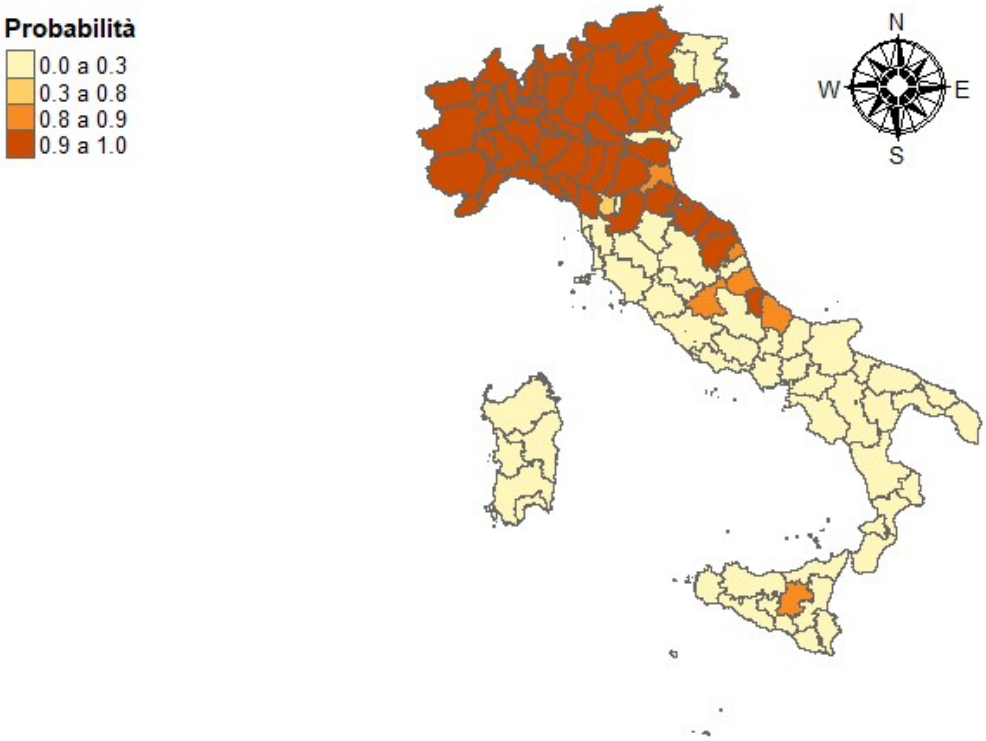


Figura 4.5: Distribuzione spaziale della probabilità di rischio in eccesso

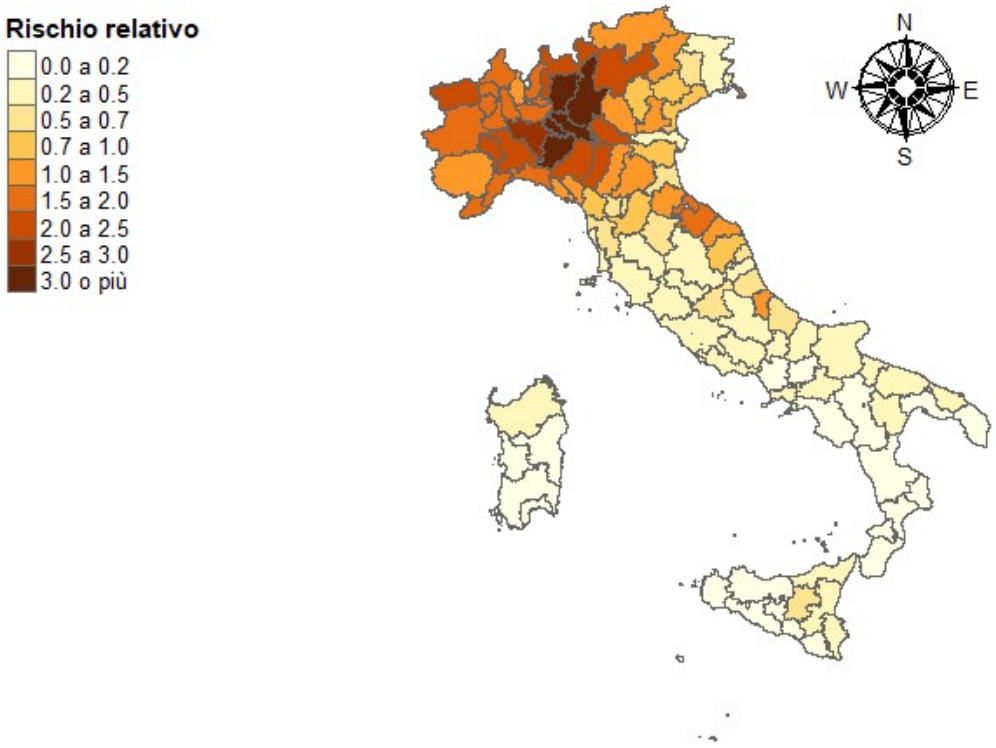


Figura 4.6: Distribuzione spaziale del rischio relativo stimato dal modello ICAR

4.2.2 Analisi spazio-temporali

L'adattamento spazio-temporale è utilizzato solo il metodo INLA dato che è possibile inserire diverse specificazioni per la variabile d'interesse. Inoltre l'uso di INLA in questi casi risulta più agevole in quanto il predittore lineare può essere adattato a proprio piacimento dato che la struttura può essere decisa dall'utente, inserendo o togliendo effetti fissi o casuali.

Per tre diverse definizioni del modello statistico, ovvero Poisson, Binomiale Negativo e ZIP, si sono formulate diverse definizioni del predittore lineare utilizzando diverse specificazioni riguardo alla parte temporale δ_t e ϕ_{it}. Esse sono:

1. trend parabolico;
2. Bernardinelli (trend parabolico singolo per ogni provincia);
3. trend stocastico;
4. trend autoregressivo;
5. trend *random walk*;
6. interazione spazio-temporale type I;
7. trend parabolico e interazione spazio-temporale type I;
8. trend stocastico e interazione spazio-temporale type I;
9. trend autoregressivo e interazione spazio-temporale type I;
10. trend *random walk* e interazione spazio-temporale type I;

Il trend è descritto come parabolico per ottenere una migliore descrizione del fenomeno dato che nel periodo considerato esso possiede una forma campanulare. Per quanto riguarda l'effetto casuale autocorrelato si è utilizzato il modello ICAR dato che dall'analisi precedente risultava la migliore specificazione. Data l'estensione del modello alla componente temporale si inserisce una diversa struttura di autocorrelazione, ovvero quella temporale, come si nota dalla Figura 4.7 per la provincia di Padova. Sarà quindi da

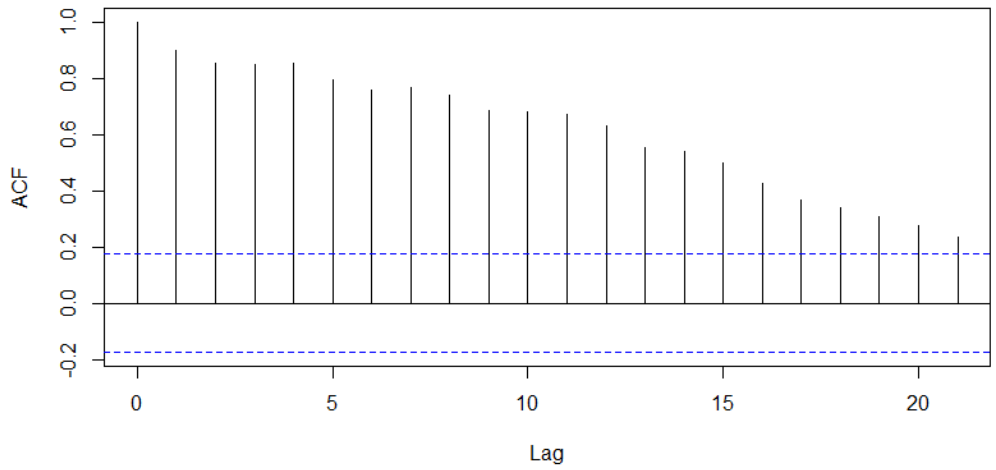


Figura 4.7: Autocorrelazione per le osservazioni della provincia di Padova

considerare se le diverse specificazioni temporali riescano a cogliere tale tipo di dipendenza.

Ora si presenta il modello migliore per ognuna delle specificazioni distributive della variabile risposta, questo modello è sempre stato scelto in base alla statistiche di sintesi.

Tabella 4.5: Modelli migliori per le diverse assunzioni distributive della variabile risposta

Modello	DIC	pD	$WAIC$	$LMPL$
Poisson 9	57619.71	8245.03	56233.43	-49252.03
Binomiale Negativa 9	70384.00	3440.60	70226.91	-88991.44
ZIP 7	63096.43	6925.28	62310.51	-42554.44

Le prime due assunzioni (Poisson, Binomiale Negativa) sono concordi nel definire la specificazione 9. Come la migliore, mentre l'ultima (ZIP) definisce la 7. Dai risultati in Tabella 4.5 si nota come sia ancora presente la tendenza alla parsimonia nel caso in cui si assuma una Binomiale Negativa, dato che presenta un numero di parametri effettivi molto minore rispetto agli altri casi. Però ancora una volta, i restanti indici non risultano molto soddisfacenti.

L'aggiunta dell'interazione comporta un forte aumento del numero di parametri effettivi che sembra corrispondere a una adeguata capacità esplicativa del modello come mostrato da DIC e WAIC. Infatti questi modelli si adattano molto bene ai dati grazie agli effetti casuali non parametrici. Di certo, anche se non si intendono i parametri come nel caso frequentista i valori riportati da pD sono estremamente elevati. Nei casi di studio riportati da Lawson (2018), anche rapportandosi al numero di osservazioni non si raggiungevano certi livelli. Probabilmente questo è anche dovuto alla frequenza giornaliera e alla volatilità dei valori che richiedono una sforzo maggiore al modello per poterne comprendere l'andamento. Non vi sono molti casi in letteratura dove i dati presentano una struttura temporale così densa. Perciò questo può essere un esempio dell'adattamento di questi modelli a dati di questo tipo e si potrà scegliere con cura tra la parsimonia e la bontà. In particolare, i valori di $LMPL$ sono molto bassi, il che indica che i CPO possiedono valori prossimi allo zero probabilmente per lo stesso motivo. La difficoltà di ottenere una buona distribuzione predittiva togliendo un valore può essere difficoltoso data la complesità del dato. Successivamente si osserveranno le simulazioni effettuate proprio per valutare questa struttura in ambito previsivo con degli indici diversi per confermare o meno questo risultato.

Per quanto riguarda la dipendenza temporale, essa risulta complessa da cogliere e i modelli presentano comportamenti diversi tra loro osservando l'autocorrelazione sui residui dei vari modelli. Esse si dividono in tre gruppi, il primo legato ai modelli che presentano l'uso di un trend parametrico, ovvero le specificazioni 1, 2 e 7 i quali non colgono eccessivamente l'autocorrelazione. Questa indicazione porta a ipotizzare che la struttura di autocorrelazione delle osservazioni sia in realtà più complessa. Il secondo gruppo formato dai modelli che presentano una componente autoregressiva o di *random walk* mostrano un buon risultato nel calcolo della funzione di autocorrelazione. Essi tramite gli effetti casuali autocorrelati riescono a cogliere gran parte della dipendenza. In questo caso la moda a posteriori dell'effetto autoregressivo di prim'ordine è pari a 0.955, con intervallo di credibilità (0.905;0.99) si spiega dunque perché fornisca un risultato simile all'effetto *random walk*. Infine l'ultimo gruppo sono i modelli 3 e 6, i quali presentano un trend non parametrico o una interazione temporale che colgono una parte dell'autocorrelazione,

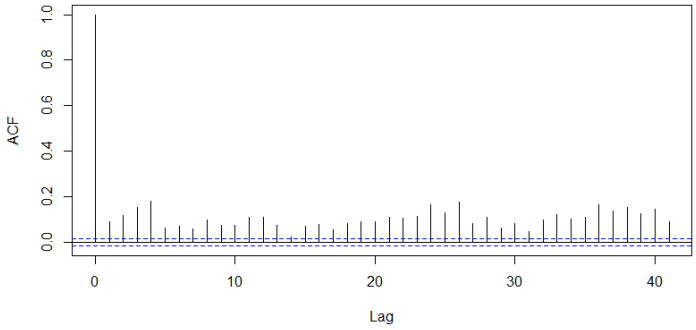
ma non in maniera soddisfacente per la parte iniziale. In Figura 4.8 si può osservare l'adattamento di questi tre gruppi.

I risultati non sono comunque pienamente soddisfacenti, segno che queste specificazioni non riescono a cogliere adeguatamente questa dipendenza temporale sia perché essa risulta complessa, sia per la limitazione dei modelli stessi.

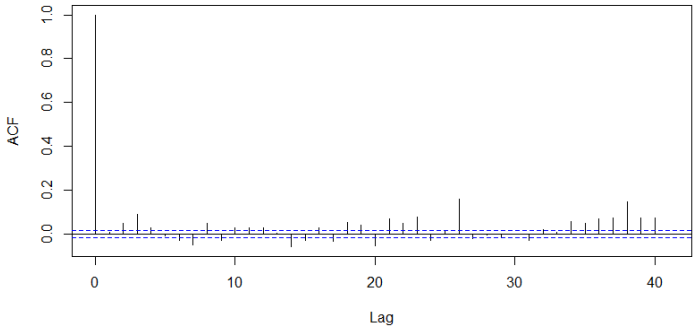
Date queste premesse si è comunque deciso di utilizzare la specificazione 9 con distribuzione di Poisson per descrivere il rischio relativo nell'Italia in ambito spaziale e temporale. In Figura 4.9, in cui la scala è la medesima usata in Figura 4.6, si osserva come sia avvenuta la diffusione dell'epidemia nei diversi periodi descritti. In particolare, il *cluster* lombardo-piemontese è l'ultimo a riuscire a raggiungere un rischio relativo basso. Vi sono stati vari momenti in cui anche nel Sud la situazione era critica e poteva replicare il Nord, ma questo non è avvenuto dato che la malattia non si era radicata così profondamente nel territorio. Si possono osservare comunque delle zone non sospette in cui il rischio risultava abbastanza elevato come in Sardegna, Sicilia e Puglia. Nell'ultima mappa vi sono tre province dove sembra che la malattia stia accelerando il contagio dato che presenta un rischio relativo tra 1.5 e 2. Queste aree sono la provincia autonoma di Trento, Chieti e Teramo che hanno segnalato il 24 Giugno rispettivamente 387, 186 e 154 nuovi positivi, ma questo probabilmente è dovuto ad un accumulo di tamponi.

4.2.3 Inserimento covariate

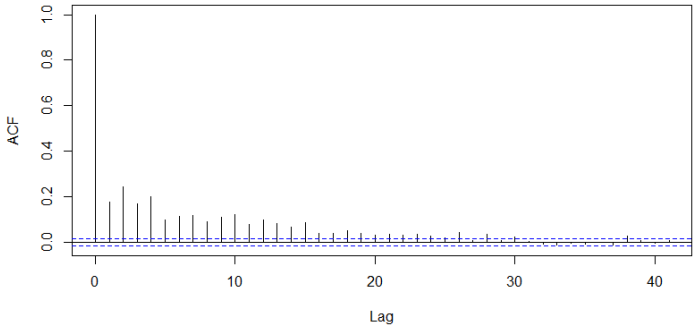
Le osservazioni delle variabili esplicative a livello provinciale sono state raccolte dal sito dell'ISTAT e sono le seguenti: altitudine media, PIL, percentuale dell'occupazione, utilizzo del trasporto pubblico. Queste variabili vengono utilizzate come delle *proxy* per definire le relazioni e gli spostamenti degli abitanti della provincia in questione. L'altitudine dovrebbe definire la distizione tra le zone italiane in cui la diffusione del trasporto di merci e persone è maggiore, il PIL e la percentuale di occupazione identificano le aree più ricche in cui l'economia può portare più spostamenti e l'utilizzo del trasporto pubblico va a definire quanto gli abitanti della zone tendono ad fare viaggi in gruppo. Queste covariate non soddisfano tutte le casistiche per cui il con-



(a) Primo gruppo



(b) Secondo gruppo



(c) Terzo gruppo

Figura 4.8: Autocorrelogrammi dei tre diversi gruppi di modelli

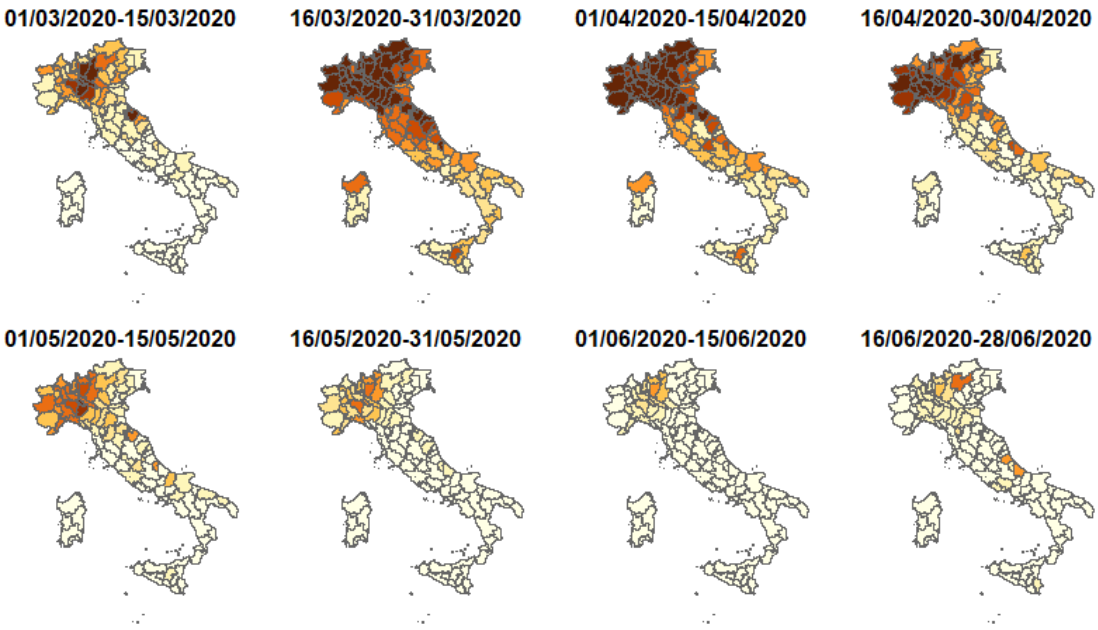


Figura 4.9: Distribuzione spazio-temporale del rischio relativo stimato

tagio può essere influenzato, ma in questo caso le si utilizzano per verificare il comportamento dei modelli spazio-temporali.

Si possono osservare ora delle statistiche di sintesi riguardanti le covariate dal modello che possiede gli indici di bontà di adattamento migliori. In questo caso si ha una conferma di quanto visto in Tabella 4.5: il modello 9 Poisson con l'aggiunta di covariate possiede un DIC pari a 57491.43, un $WAIC$ di 56101.88 e un valore di $LMPL$ di -47244.95. L'uso delle covariate ha migliorato gli indici del modello, segno che questi elementi sembrano fornire una spiegazione aggiuntiva rilevante anche tenendo conto che l'aggiunta di parametri è veramente esigua rispetto a quelli già presenti. I risultati ottenuti sono riassunti nella Tabella 4.6. Si può osservare come l'elemento che incide maggiormente sia il livello di occupazione, il quale se riportato alla sua scala naturale definisce che all'aumentare di una unità percentuale di tale livello, il rischio di contagio aumenta del 6% con un intervallo di credibilità che va da 3.6% al 9%. Le altre covariate non presentano un effetto così decisivo sulla diffusione del contagio: PIL, altitudine e uso dei trasporti pubblici forse sono delle variabili troppo vaghe che non riescono ad incidere nella spiegazione del

Tabella 4.6: Statistiche di sintesi della distribuzione a posteriori delle covariate utilizzando il modello 9 Poisson

	Media	SD	2.5%	50%	97.5%	Moda
Intercetta	-9.18	1.45	-12.55	-9.04	-6.64	-8.87
PIL	9.36×10^{-5}	2.54×10^{-5}	4.48×10^{-5}	9.30×10^{-5}	1.44×10^{-4}	9.21×10^{-5}
Occupazione	6.17×10^{-2}	1.30×10^{-2}	3.58×10^{-2}	6.19×10^{-2}	8.70×10^{-2}	6.20×10^{-2}
Altezza	2.89×10^{-4}	2.41×10^{-4}	-1.80×10^{-4}	2.87×10^{-4}	7.68×10^{-4}	2.84×10^{-4}
Trasporti	-2.15×10^{-3}	8.31×10^{-4}	-3.83×10^{-3}	-2.14×10^{-3}	-5.60×10^{-3}	-2.11×10^{-3}

fenomeno come *proxy*. Questo sottolinea ancora una volta come il fenomeno in questione sia complesso da analizzare e per poterlo comprendere più a fondo è necessario conoscere gli elementi più importanti che possono influenzare la diffusione. In questo caso è anche possibile che la grande quantità di parametri effettivi portata dalla componente temporale abbia, in qualche modo, contenuto la capacità esplicativa delle variabile esplicative.

Si anticipa che nel paragrafo successivo le analisi effettuate non considereranno l'uso delle covariate dato che si è visto che in questo caso non forniscono capacità esplicativa aggiuntiva.

4.3 Studio delle previsioni

Come detto in precedenza, solitamente questi modelli bayesiani adattati ai dati di tipo area vengono utilizzati per stimare il rischio relativo delle varie zone in un lasso di tempo oppure per visualizzare l'effetto di alcune covariate d'interesse considerando l'autocorrelazione spaziale. Raramente questi modelli vengono presi in considerazione nel momento in cui si voglia eseguire una previsione per giorni successivi, in questi casi si utilizzano strumenti più specifici come i modelli per diffusione.

Dopo aver adattato i vari modelli ai dati e aver visualizzato il loro comportamento può essere comunque d'interesse studiarne le caratteristiche anche per uno scopo diverso. Per questo si sono fatte diverse previsioni per le specificazioni descritte del paragrafo precedente. Queste simulazioni sono state fatte in diversi casi. Qui verranno presentati solo i più rilevanti. Dal punto di vista tecnico le previsioni sono state eseguite con INLA utilizzando la *data*

augmentation presentata nel Paragrafo 1.3.1. Questa tecnica risulta computazionalmente onerosa dato che per eseguire diverse previsioni si deve stimare l'intero modello.

Nello specifico le previsioni sono valutate tramite le metriche RMSE (*Root Mean Squared Error*) e MAE (*Mean Absolute Error*) e si confronteranno questi valori per le 10 specificazioni del modello. Verrà effettuato anche un confronto tra assunzioni distributive tra Poisson, Binomiale Negativa e ZIP.

Per prima cosa si sono fatte delle previsioni con poca informazione, ovvero si è stimato il modello con i dati fino al 9 Marzo, data del decreto che ha istituito il *lockdown*, e si è fatta una previsione fino al 16 Marzo. Un primo esperimento è stato eseguito cercando di prevedere il comportamento dell'intera Italia come mostrato in Tabella 4.7.

Tabella 4.7: Confronto dei risultati delle previsioni fra le diverse specificazioni del modello con distribuzione Poisson

Modello	1	2	3	4	5
RMSE	44.02	102.91	38.66	68.79	15652.03
MAE	15.86	25.84	15.60	23.36	27895.18
Modello	6	7	8	9	10
RMSE	46.71	672.87	62.91	552.62	815.56
MAE	17.79	174.86	23.13	147.51	1452.25

In questo caso sembra che i modelli migliori siano l'1 o il 3. Questo mostra come delle specificazioni più semplici sembrano essere più vantaggiose. In particolare il *random walk* sembra soffrire più di tutti la mancanza di informazione nei dati e risulta complicato fare delle previsioni in modo consecutivo, in quanto il modello tende costantemente a sovrastimare l'effettiva magnitudine della diffusione.

Ora si possono confrontare i risultati dello stesso esperimento per le altre due assunzioni distributive presenti in Tabella 4.8 e 4.9. La specificazione Binomiale Negativa riduce gli errori del *random walk*, ma aumenta quelli delle altre specificazioni, ottenendo dei valori delle metriche peggiori di prima. Viceversa utilizzare la distribuzione ZIP porta a dei grandi vantaggi in tutte le diverse specificazioni, probabilmente dato dal fatto che essendo costruito

ad hoc per dati con molti zeri riesce a cogliere l'andamento dell'epidemia in questo contesto. Il modello che presenta i risultati migliori è il primo, nel quale è inserito un trend parametrico.

Tabella 4.8: Confronto dei risultati delle previsioni fra le diverse specificazioni del modello con distribuzione Binomiale Negativa

Modello	1	2	3	4	5
RMSE	419.71	319.08	50.61	420.87	9023.01
MAE	103.54	94.12	19.69	117.08	1609.45
Modello	6	7	8	9	10
RMSE	53.58	523.24	53.58	512.53	588.18
MAE	20.17	139.23	20.55	140.02	104.19

Tabella 4.9: Confronto dei risultati delle previsioni fra le diverse specificazioni del modello con distribuzione ZIP

Modello	1	2	3	4	5
RMSE	42.42	75.95	53.48	52.52	58.09
MAE	17.89	19.78	20.30	19.98	22.20
Modello	6	7	8	9	10
RMSE	52.92	60.02	48.44	50.79	81.79
MAE	19.60	24.42	18.29	19.11	33.60

Proseguendo nelle simulazioni si sono fatte due diverse previsioni per osservare quanto la dipendenza tra le aree portasse vantaggi anche in ambito previsivo. Si sono previsti i valori della regione Campania nel periodo considerato prima stimando il modello solo con le osservazioni della stessa e poi con i dati delle regioni confinanti, in Tabella 4.10 si possono osservare i risultati per i due migliori modelli nei due diversi casi. Come si nota, l'aumento dell'informazione fornito dai vicini della regione porta a dei benefici anche dal punto di vista previsivo, sfruttando la potenzialità della struttura di dipendenza presente nel modello. I modelli migliori differiscono da quelli precedenti, segno che non si riconosca un modello migliore in generale,

ma piuttosto sembra che ad ogni caso corrisponda uno strumento adeguato. In questo caso gli errori sono molto più piccoli dei precedenti, ma questo è dovuto anche alla diversità dei valori reali.

Una ulteriore conferma è data dalla stessa simulazione effettuata con la regione Veneto presente in Tabella 4.11.

Tabella 4.10: Confronto dei risultati delle previsioni per la Campania considerando le regioni limitrofe o meno

Modello	4 no vicini	8 no vicini	7 vicini	10 vicini
RMSE	9.90	22.30	10.93	8.53
MAE	6.13	16.94	7.01	5.95

Tabella 4.11: Confronto dei risultati delle previsioni per il Veneto considerando le regioni limitrofe o meno

Modello	1 no vicini	2 no vicini	1 vicini	3 vicini
RMSE	22.79	26.89	20.79	21.44
MAE	15.13	19.05	13.65	16.04

Come si è osservato in questo caso non sembra che vi sia un modello che riesca a prevedere il numero di nuovi contagi in modo migliore rispetto agli altri. In questo contesto la specificazione ZIP sembra la più adeguata al fine di ridurre il valore delle metriche. Può essere quindi interessante valutare le capacità previsive del modello nelle sue specificazioni in un altro contesto, ovvero nel caso in cui si possiede più informazione.

In questo caso si confronteranno le *performance* dei modelli adattati con i dati fino al 12 Aprile e si esegue una previsione fino al 19 Aprile. In Tabella 4.12 si presentano i modelli migliore fra tutte le tre diverse assunzioni distributive; l'uso della Binomiale Negativa sembra confermare quanto visto in precedenza: essa infatti risulta inadeguata per questo scopo. Invece le distribuzioni Poisson e ZIP hanno un comportamento simile tra loro ed è da notare come le metriche della distribuzione Poisson siano migliorate di molto ora che l'informazione contenuta nei dati è maggiore.

Tabella 4.12: Confronto dei risultati delle previsioni per l'Italia per i migliori modelli nelle diverse assunzioni

Modello	1 Poisson	2 Poisson	9 Binomiale negativa	4 ZIP
RMSE	41.86	40.81	84.31	39.82
MAE	19.42	19.45	38.32	21.09

È da notare come i modelli riescano a mantenere lo stesso livello del valore delle metriche sia in una fase di esplosione della malattia, sia in una fase discendente come quella appena analizzata. Volendo osservare la bontà delle previsioni in questo contesto, ma con un caso più specifico, in Figura 4.10 si può osservare il confronto fra l'andamento dell'epidemia reale e quello previsto dal modello 1 e 2 con distribuzione Poisson per la sola provincia di Padova utilizzando i dati del Veneto e delle regioni limitrofe. In questo caso i modelli con trend parametrico forniscono dei risultati migliori. Diversamente accade invece per le simulazioni riguardanti Bari, in questo caso come si osserva in Figura 4.11 i modelli migliori risultano quelli non parametrici. In entrambi i grafici sono rappresentati gli intervalli di credibilità tramite i quantili delle distribuzioni predittive nei diversi punti, come si nota nel caso di Bari gli intervalli sono molto più variabili e molto più larghi rispetto al caso parametrico.

Da questi due grafici si nota molto bene le differenze che vi sono fra i due approcci. In particolare i modelli non parametrici dato il loro utilizzo di effetti casuali senza nessuna struttura prestabilita quasi sovradattano i dati stessi utilizzando anche un gran numero di parametri effettivi. Questo è molto frequente nel caso in cui vi sia una interazione spazio-temporale. Questo li porta a poter essere utilizzati nel caso in cui si voglia adattare un insieme di dati, ma non funziona altrettanto bene se si vuole eseguire una previsione. Infatti sembra che i modelli non colgano molto efficacemente l'andamento dei nuovi contagi giornalieri. Viceversa il modello parametrico presenta dei buoni risultati nel caso in cui si vogliano prevedere dei valori, ma hanno comunque bisogno di una definizione strutturale: se si fosse cambiata la specificazione del trend deterministico i risultati sarebbero diversi.

Inoltre si riscontra una problematica riguardante i modelli non parametri-

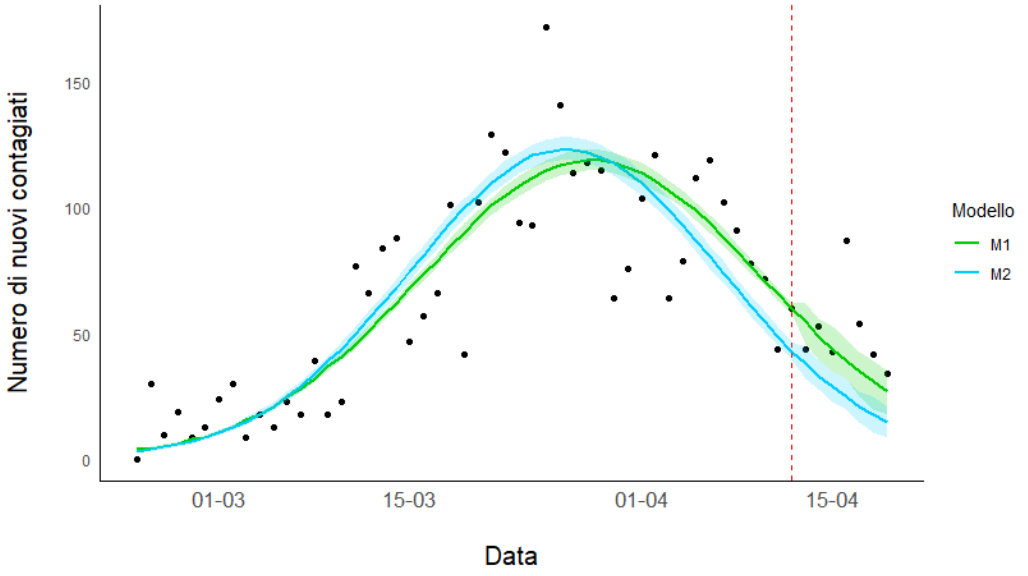


Figura 4.10: Andamento delle previsioni del modello 1 e 2 Poisson per Padova confrontato con il reale numero di nuovi contagi. La linea rossa identifica l'inizio della previsione in data 13 Aprile.

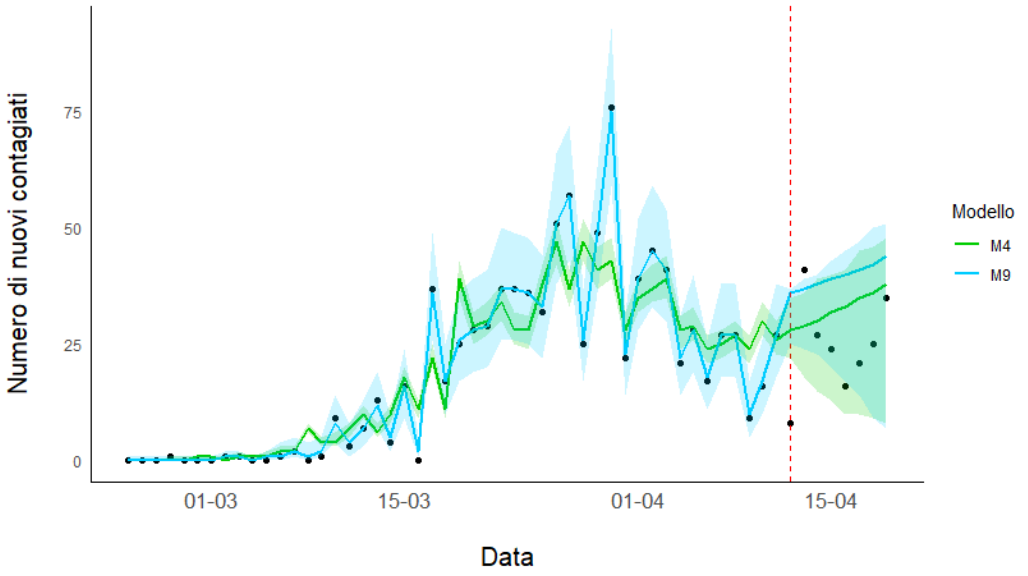


Figura 4.11: Andamento delle previsioni del modello 4 e 9 Poisson per Bari confrontato con il reale numero di nuovi contagi. La linea rossa identifica l'inizio della previsione in data 13 Aprile.

ci che presentano un effetto casuale indipendente o solamente una interazione Type I per quanto riguarda le previsioni: questi modelli non possedendo una forma di dipendenza come può essere quella autoregressiva. Essi non fanno variare le previsioni per istanti temporali diversi. Le previsioni a sette giorni successivi risultavano tutte uguali per una data provincia segno che questa semplice struttura dovrebbe essere accompagnata da altri elementi per poter ottenere dei risultati soddisfacenti in questo ambito.

Osservando il comportamento delle diverse previsioni non risulta conveniente eseguirle per un periodo di lunghezza maggiore. Il trend parametrico non riuscirà a prevedere la volatilità nella code destra della distribuzione della diffusione in quanto essa tenderà a valori molto bassi e nemmeno il trend stocastico nelle sue diverse forme ha l'elasticità adatta per comprendere questa dinamica in periodi più lunghi.

Conclusioni

Nel corso di questa tesi sono stati presentati vari strumenti per l'analisi dei dati spaziali a partire dalla loro definizione fino ai modelli che possono essere implementati per poterli analizzare. In particolare, si è posta l'attenzione su uno specifico tipo di dato spaziale: i dati ad area che rappresentano solo una parte di questa tipologia. Partendo da questi dati si sono presentati una serie di modelli spaziali e spazio-temporali atti a considerare l'autocorrelazione spaziale e temporale che è intrinseca in queste osservazioni. Infine si è presentata una metodologia per la stima di questi modelli dal punto di vista bayesiano che si diversifica dalla famiglia dei metodi MCMC per il fatto che utilizza delle approssimazioni delle distribuzioni dei vari parametri per eseguire l'inferenza. Infine si è mostrata l'applicazione di quanto illustrato ai dati della diffusione del Covid-19 che ha colpito l'Italia dalla prima parte del 2020. Parte dei commenti sui metodi e sui risultati sono già stati fatti nelle diverse sezioni, ora vengono affrontati con spirito critico alcuni limiti di queste applicazioni e si proporranno delle scelte alternative a questo lavoro.

La prima criticità è legata alla complessità temporale dei dati: come mostrato in altri lavori, come Khan *et al.* (2018), la grandezza della dimensione spaziale non incide eccessivamente sul costo computazionale e sulle difficoltà di convergenza dei risultati. Sembra invece che la parte temporale influisca sull'efficienza ed efficacia dei vari modelli. In questo contesto l'alto numero di osservazioni porta ad un aumento consistente del numero di parametri effettivi e porta a problemi di approssimazioni in INLA e ne aumenta il costo computazionale diminuendo il guadagno rispetto al metodo MCMC. Legata a questa problematica c'è la difficoltà nell'uso di INLA, che è un *framework* di lavoro che dovrebbe essere approfondito per poterne comprendere tutte

le sfaccettature e poter essere utilizzato nel pieno delle sue potenzialità, le risorse da poter consultare non sono molte e questo può portare a utilizzare altre librerie più diffuse.

La seconda criticità è espressa dai dati stessi, in quanto essi si riferiscono ad un fenomeno abbastanza complicato e legato a un molti fattori. Inoltre la raccolta dei dati è stata di secondaria importanza nella gestione dell'epidemia e questo ne ha in parte intaccato la qualità e questo potrebbe aver influenzato i risultati ottenuti. In futuri sviluppi si potrebbe utilizzare un diverso dataset per valutare complessivamente questo tipo di modellazione e l'utilizzo di modelli alternativi in INLA, dopo uno studio che vada a valutare anche la dinamica del fenomeno in questione.

Altri modelli, anche già implementati in diverse librerie con metodo MCMC, sarebbero interessanti, come utilizzare una distribuzioni a priori della famiglia dei CAR anche per la componente temporale, cercando di abbattere il costo computazionale utilizzando diverse catene. Strumenti che possono essere implementati per fenomeni del genere potrebbero essere legati ai modelli grafici ciclici, in cui ogni nodo è una provincia e il collegamento è la vicinanza. Sempre in quest'ottica una formulazione alternativa della matrice dei pesi potrebbe essere considerata, magari utilizzando covariate che si definiscono importanti per la diffusione.

Nel complesso l'adattamento dei modelli è risultato abbastanza buono e utile per dati di questo tipo, i quali possono essere frequenti per esempio in ambito medico dove l'aggregazione può essere utilizzata come strumento di anonimizzazione. L'uso delle covariate può essere uno strumento utile per questi modelli dato che esse potrebbero cogliere molto dell'eterogeneità dei dati così da poter valutare efficacemente l'effetto delle altre variabili.

La capacità previsiva di questi modelli risulta instabile e dipende da diverse caratteristiche come la zona interessata, la lunghezza temporale e il modello scelto. Essa può essere comunque utilizzata per fornire una indicazione di come potrebbe proseguire la diffusione del fenomeno considerato se si fosse in possesso di dati in tale formato, dato che si è visto come sia necessario considerare l'autocorrelazione spaziale delle osservazioni.

Bibliografia

Ancelet, S., J. Abellan, V. Vilas, C. Birch e S. Richardson (2012). «Bayesian shared spatial-component models to combine and borrow strength across sparse disease surveillance sources». In: *Biometrical journal. Biometrische Zeitschrift* , **54**, 385–404.

Bernardinelli, L., D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi e M. Songini (1995). «Bayesian analysis of space-time variation in disease risk». In: *Statistics in Medicine* , **14**, 2433–2443.

Besag, J., J. York e A. Mollié (1991). «Bayesian image restoration, with two applications in spatial statistics». In: *Annals of the Institute of Statistical Mathematics* , **43**, 1–20.

Best, N., S. Richardson e A. Thomson (2005). «A Comparison of Bayesian Spatial Models for Disease Mapping». In: *Statistical Methods in Medical Research* , **14**, 35–59.

Blangiardo, M. e M. Cameletti (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley.

Briz-Redon, A. e A. Serrano-Aroca (2020). «A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain». In: *Science of The Total Enviroment* , **728**, 138811–138818.

Diggle, P. e E. Giorgi (2019). *Model-based Geostatistic for Global Public Health*. Chapman & Hall/CRC.

Dominici, F., Xiao, R. Nethery, M. Benjamin e D. Braun (2020). «Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study». URL: `https://www.medrxiv.org/content/10.1101/2020.04.05.20054502v2`.

Ejigu, B. e E. Wencheko (2020). «Introducing covariate dependent weighting matrices in fitting autoregressive models and measuring spatio-environmental autocorrelation». In: *Spatial Statistics* , **38**.

Fronterrè, C., J. Read, B. Rowlingson, J. Bridgen, S. Alderton, P. Diggle e C. Jewell (2020). «COVID-19 in England: spatial patterns and regional outbreaks». URL: https://www.medrxiv.org/content/10.1101/2020.05.15.20102715v1.

Gatto, M., E. Bertuzzo, L. Mari, S. Miccoli, L. Carraro, R. Casagrandi e A. Rinaldo (2020). «Spread and dynamics of the COVID-19 epidemic in Italy: Effects of emergency containment measures». In: *Proceedings of the National Academy of Sciences* , **117**, 10484–10491.

Giuliani, D., M. Dickson, G. Espa e F. Santi (2020). «Modelling and predicting the spread of Coronavirus (COVID-19) infection in NUTS-3 Italian regions». URL: https://arxiv.org/abs/2003.06664.

Kang, D., H. Choi, J. Kim e J. Choi (2020). «Spatial epidemic dynamics of the COVID-19 outbreak in China». In: *International Journal of Infectious Diseases* , **94**, 96–102.

Khan, D., L. Rossen, H. Hedegaard e M. Warner (2018). «A Bayesian Spatial and Temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with R-INLA». In: *Journal of Data Science : JDS* , **16**, 147–182.

La Rosa, G., M. Iaconelli, G. Ferraro, P. Mancini e C. Veneri (2020). *Studio ISS su acque di scarico*. URL: https://www.iss.it/primo-piano/-/asset_publisher/o4oGR9qmvUz9/content/cs-n%25C2%25B039-2020-studio-iss-su-acque-di-scarico-a-milano-e-torino-sars-cov-2-presente-gi%25C3%25A0-a-dicembre.

Lanera, C., F. Pirotti e I. Prosepe (2020). URL: https://github.com/UBESP-DCTV/covid19ita/.

Lavezzo, E., E. Franchin e C. Ciavarella (2020). «Suppression of COVID-19 outbreak in the municipality of Vo, Italy».

Lawson, A. (2018). *Bayesian Disease Mapping*. Chapman & Hall/CRC.

Lee, D. (2013). *CARBayes version 5.2: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors*. URL: https://cran.r-project.org/web/packages/CARBayes/vignettes/CARBayes.pdf.

Lee, D. e R. Mitchell (2011). «Boundary detection in disease mapping studies». In: *Biostatistics* , **13**, 415–426.

Leroux, B., X. Lei e N. Breslow (2000). «Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence». In: *Institute for Mathematics and Its Applications* , **116**, 179–191.

Lewis, P., W. Xie, M. Chen, Y. Fan e L. Kuo (2013). «Posterior Predictive Bayesian Phylogenetic Model Selection». In: *Systematic biology* , **63**, 309–315.

Liseo, B. (2010). «Introduzione alla Statistica Bayesiana». In: *Dispensa didattica*.

Moran, P. (1950). «Notes on continuos stochastic phenomena». In: *Biometrika* , **37**, 17–23.

OMS, Ufficio stampa (2020). *WHO Director-General's opening remarks at the media briefing on COVID-19*. URL: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

Pace, L. e A. Salvan (2001). *Introduzione alla Statistica: Inferenza, Verosimiglianza, Modelli*. Cedam.

Parkes, E. (2013). «Mode of Communication of Cholera. By John Snow». In: *International journal of epidemiology* , **42**, 1543–1552.

Riebler, A., S. Sørbye, D. Simpson e H. Rue (2015). «An intuitive Bayesian spatial model for disease mapping that accounts for scaling». In: *Statistical Methods in Medical Research* , **25**, 1079–1084.

Rue, H. e S. Martino (2007). «Approximate Bayesian inference for hierarchical Gaussian Markov random fields». In: *Journal of Statistical Planning and Inference* , **137**, 3177–3192.

Rue, H., S. Martino e N. Chopin (2009). «Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations». In: *Journal of the Royal Statistical Society Series B* , **71**, 319–392.

Rue, H., A. Riebler, S. Sørbye, J. Illian, D. Simpson e F. Lindgren (2016). «Bayesian computing with INLA: A review». In: *Annual Review of Statistics and Its Application* , **4**, 395–421.

Salvan, A., N. Sartori e L. Pace (2020). *Modelli Lineari Generalizzati*. Springer.

Setti, L., F. Passarini, G. De Gennaro, P. Barbieri, M. Perrone, A. Piazzalunga, M. Borelli, J. Palmisani, A. Gilio, P. Piscitelli e A. Miani (2020). «The Potential role of Particulate Matter in the Spreading of COVID-19 in Northern Italy: First Evidence-based Research Hypotheses». URL: https://www.medrxiv.org/content/10.1101/2020.04.11.20061713v1.full.pdf.

Spiegelhalter, D., N. Best e B. Carlin (1998). «Bayesian Deviance, the Effective Number of Parameters, and the Comparison of Arbitrarily Complex Models». In: *Journal of Royal Statistical Society* , **64**, 583–639.

Tanner, M. e W.H. Wong (1987). «The Calculation of Posterior Distributions by Data Augmentation». In: *Journal of the American Statistical Association* , **82**, 528–540.

Tierney, L. e J. Kadane (1986). «Accurate Approximations for Posterior Moments and Marginal Densities». In: *Journal of the American Statistical Association* , **81**, 82–86.

Tobler, W. (1979). «Cellular Geography». In: *Philosophy in Geography* , **20**, 379–386.

Unit, MRC Biostatistics (2020). *The BUGS project: WinBUGS*. URL: https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/.

Wang, J., K. Tang, K. Feng e W. Lv (2020). «High Temperature and High Humidity Reduce the Transmission of COVID-19». In: URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3551767.

Wang, Y. Yue e J. Faraway (2018). *Bayesian regression modeling with INLA*. Chapman & Hall/CRC.

Appendice A

Codice R utilizzato

Codice A.1: Caricamento e pulizia dei dati

```
library(remotes)
remotes::install_github("UBESP-DCTV/covid19ita")
library(covid19ita)
library(tidyr)
library(dplyr)

dati <- dpc_covid19_ita_province
#Eliminazione delle note

dati <- dati[,-c(2,11,12)]

#Effetto Napoli

dati[which(dati$denominazione_provincia == 'Napoli'), 'sigla_provincia'] =
     'NAP'
dati[dati$denominazione_regione == 'Campania', 'sigla_provincia']

#Sistemazione data

dati <- dati%>%separate(data, into=c('giorno','ora'), sep=' ', convert=T)
```

```
#Eliminazione NA (da controllare se le aggiornano o meno ste righe)

dati <- na.omit(dati)

#Creazione della colonna "Contagi giornalieri" e assegnazione

temp <- dati%>%arrange(giorno)%>%arrange(sigla_provincia)

head(temp)

nuovi_casi <- rep(0,NROW(temp))

for(i in 2:NROW(temp)) {
        if(temp$sigla_provincia[i] == temp$sigla_provincia[i-1]) {
                nuovi_casi[i] <- temp$totale_casi[i] - temp$totale_casi[i
                    -1]
        } else {
        nuovi_casi[i] <- 0
        }
}

nuovi_casi[which(nuovi_casi < 0)] <- 0

#Unione col precedente dataset

temp['nuovi_casi'] <- nuovi_casi
temp <- temp[,c(1,7,11)]

dati <- left_join(dati,temp, by = c('giorno','sigla_provincia'), keep = F)

rm('temp')

dati.spatial <- dati%>%group_by(sigla_provincia)%>%
                            summarise(lat=min(lat), long=min(long), nuovi_
    casi=sum(nuovi_casi),denominazione_prov=min(denominazione_provincia),
    denominazione_reg=min(denominazione_regione))
```

```
dati.spatial <- as.data.frame(dati.spatial)
```

Codice A.2: Accorpamento del dataset al file shp che descrive i limiti geografici

```
library(sf)
library(spdep)

#Carico la mappa multipoligono delle province Italiane

Italy.prov <- st_read("File_Italia/Limiti01012018/ProvCM01012018")

#Sistemazione province

colnames(Italy.prov)[9] <- 'sigla_provincia'
Italy.prov$sigla_provincia <- as.factor(Italy.prov$sigla_provincia)
levels(Italy.prov$DEN_PCM)[levels(Italy.prov$DEN_PCM)=='Forli\' - Cesena']
    <- 'Forli-Cesena'
levels(Italy.prov$sigla_provincia)[levels(Italy.prov$sigla_provincia)=='NA
    '] <- 'NAP'

#Bisogna avere il medesimo ordine tra il file spaziale e il dataset

Italy.prov <- Italy.prov[order(Italy.prov$sigla_provincia),]

#Standardizzazione:

R <- sum(dati.spatial$nuovi_casi)/sum(dati.spatial$Popolazione)
dati.spatial$e <- round(R*dati.spatial$Popolazione)

#Codifica corretta delle variabili:

dati.spatial$sigla_provincia <- as.factor(dati.spatial$sigla_provincia)
dati.spatial$denominazione_prov <- as.factor(dati.spatial$denominazione_
    prov)
```

```
dati.spatial$denominazione_reg <- as.factor(dati.spatial$denominazione_reg
    )

Italy.prov.st <- Italy.prov
Italy.prov.st$reg <- dati.spatial$denominazione_reg
Italy.prov.st <- Italy.prov.st[order(Italy.prov.st$reg),]



dati.st <- dati[,-c(2,3)]
dati.st$codice_provincia <- as.factor(dati.st$codice_provincia)
levels(dati.st$codice_provincia)[104:107] <- c('104','105','106','107')
dati.st$codice_provincia <- as.numeric(dati.st$codice_provincia)
colnames(dati.st)[3] <- 'id.provincia'
dati.st$id.temp <- as.numeric(as.factor(dati.st$giorno))

#Codifica corretta delle variabili

dati.st$giorno <- as.Date(dati.st$giorno)
dati.st$denominazione_regione <- as.factor(dati.st$denominazione_regione)
dati.st$denominazione_provincia <- as.factor(dati.st$denominazione_
    provincia)
dati.st$sigla_provincia <- as.factor(dati.st$sigla_provincia)

#Aggiungo le variabili utili per il join

tmp <- dati.spatial[,c(1,7)]

dati.st <- dati.st%>%
inner_join(tmp, by='sigla_provincia')%>%
arrange(giorno,denominazione_provincia)

rm(tmp)

#Aggiungo l'offset corretto

R <- sum(dati.st$nuovi_casi)/sum(dati.st$Popolazione)
```

```
dati.st$e <- R*dati.st$Popolazione
dati.st <- dati.st[,-11]
```

Codice A.3: Creazione della matrice di vicinanza e calcolo dell'indice di Moran

```
Wnb <- poly2nb(Italy.prov)
Wnb.st <- Wnb

Wb.mat <- nb2mat(Wnb, style='B')
colnames(Wb.mat) <- row.names(Wb.mat)

nb2INLA("map.adj", Wnb)
g <- inla.read.graph(filename = "map.adj")

Wb.moran <- nb2listw(Wnb, style='B')

moran.sim <- moran.mc(dati.spatial$nuovi_casi, Wrs.moran, nsim=10^4,
    return_boot = F)
moran.plot(dati.spatial$nuovi_casi, Wrs.moran, labels = dati.spatial$sigla
    _provincia)
```

Codice A.4: Adattamento dei modelli spaziali con MCMC e INLA. Si presenta solo l'esempio per una specificazione distributiva e senza covariate.

```
library(CARBayes)
library(INLA)
library(Matrix)

burnin <- 30000
n.sample <- 500000
thin <- 5

lin_pred <- as.formula(nuovi_casi~offset(log(e)))

mod.glm <- S.glm(lin_pred, data=dati.spatial, family='poisson',
burnin = burnin, n.sample = n.sample ,thin = thin)
```

```
mod.ind <- S.CARleroux(lin_pred, data=dati.spatial, family='poisson', W=Wb
    .mat,
burnin = burnin, n.sample = n.sample,thin = thin,rho = 0)

mod.icar <- S.CARleroux(lin_pred, data=dati.spatial, family='poisson', W=
    Wb.mat,
burnin = burnin, n.sample = n.sample,thin = thin,rho = 1)

mod.bym <- S.CARbym(lin_pred, data=dati.spatial, family='poisson', W=Wb.
    mat,
burnin = burnin, n.sample = n.sample,thin = thin)

mod.ler <- S.CARleroux(lin_pred, data=dati.spatial, family='poisson', W=Wb
    .mat,
burnin = burnin, n.sample = n.sample,thin = thin)

formula <- nuovi_casi ~ 1 + f(id_inla, model = 'bym2', scale.model = T,
    graph=g,
hyper = list(
phi = list(
prior = "pc",
param = c(0.5 , 2/3),
initial = -3) ,
prec = list(
prior = "pc.prec ",
param = c(0.2 /0.31 , 0.01) ,
initial = 5))) + PIL + Occupazione + altezza + Trasporti

mod.bym2 <- inla(formula, family = "poisson", data = dati.spatial,
E = e, control.predictor = list(compute = TRUE),
control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE, config=T))


#Modelli con INLA
```

```
formula <- nuovi_casi ~ 1 + f(id_inla, model = 'iid')

mod.ind.inla <- inla(formula, family = "poisson", data = dati.spatial,
E = e, control.predictor = list(compute = TRUE),
control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE, config=T))

formula <- nuovi_casi ~ 1 + f(id_inla, model='besag', graph = g)

mod.icar.inla <- inla(formula, family = "poisson", data = dati.spatial,
E = e, control.predictor = list(compute = TRUE),
control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE, config=T))

formula <- nuovi_casi ~ 1 + f(id_inla, model='bym', graph = g)

mod.bym.inla <- inla(formula, family = "poisson", data = dati.spatial,
E = e, control.predictor = list(compute = TRUE),
control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE, config=T))
```

Codice A.5: Adattamento dei modelli spazio-temporali con INLA. Si presenta solo l'esempio per una specificazione distributiva e senza covariate.

```
dati.st <- dati.st%>%arrange(giorno,sigla_provincia)

# Adatto il trend parabolico

id.temp2 <- (dati.st$id.temp)^2

fo <- nuovi_casi ~ 1 + id.temp + id.temp2 + f(id.provincia, model='besag',
     graph = g)

m.1 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor =
     list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE,
     cpo = TRUE, config=T))

# Adatto Bernardinelli
```

```
id.area1 <- dati.st$id.provincia

fo <- nuovi_casi ~ 1 + id.temp + id.temp2 + f(id.provincia, model='bym',
    graph=g)+ f(id.area1, id.temp, model='iid')

m.2 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor =
     list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE,
     cpo = TRUE, config=T))

# Adatto trend stocastico

id.temp1 <- dati.st$id.temp

fo <- nuovi_casi ~ 1 + f(id.provincia, model='besag', graph = g) + f(id.
    temp1, model='iid')

m.3 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor =
     list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE,
     cpo = TRUE, config=T))

# Adatto trend autoregressivo

fo <- nuovi_casi ~ 1 + f(id.provincia, model='besag', graph = g) + f(id.
    temp1, model='ar1')

m.4 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor =
     list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE,
     cpo = TRUE, config=T))

# Adatto trend random walk 1

fo <- nuovi_casi ~ 1 + f(id.provincia, model='besag', graph = g) + f(id.
    temp1, model='rw1')

m.5 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor =
     list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE,
```

```
    cpo = TRUE, config=T))

# Adatto interazione spazio–temporale type I

id <- 1:nrow(dati.st)

fo <- nuovi_casi ~ 1 + f(id.provincia, model='besag', graph = g) + f(id,
    model='iid')

m.6 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor =
     list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE,
     cpo = TRUE, config=T))

# Adatto trend parabolico e interazione

fo <- nuovi_casi ~ 1 + f(id.provincia, model='besag', graph = g) + f(id,
    model='iid') +
id.temp + id.temp2

m.7 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor =
     list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE,
     cpo = TRUE, config=T))

# Adatto trend stocastico e interazione

fo <- nuovi_casi ~ 1 + f(id.provincia, model='besag', graph = g) + f(id,
    model='iid') +
f(id.temp, model='iid')

m.8 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor =
     list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE,
     cpo = TRUE, config=T))

# Adatto trend autoregressivo e interazione

fo <- nuovi_casi ~ 1 + f(id.provincia, model='besag', graph = g) + f(id,
```

```
    model='iid') +
f(id.temp, model='ar1')

m.9 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor =
     list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE,
     cpo = TRUE, config=T))

# Adatto trend random walk e interazione

fo <- nuovi_casi ~ 1 + f(id.provincia, model='besag', graph = g) + f(id,
    model='iid') +
f(id.temp, model='rw1')

m.10 <- inla(fo, data=dati.st, family = 'poisson', E=e, control.predictor
    = list(compute = TRUE), control.compute = list(dic = TRUE, waic = TRUE
    , cpo = TRUE, config=T))
```

Codice A.6: Esempio di simulazione. Per effettuare altre simulazioni si devono identificare altra date e altre zone.

```
#Usando i dati del Nord, vado a prevedere Padova. Sempre stesse date
    definite precedentemente

Nord <- c(which(dati.st_sort$denominazione_regione=='Emilia-Romagna'),
    which(dati.st_sort$denominazione_regione=='Friuli Venezia Giulia'),
which(dati.st_sort$denominazione_regione == 'Liguria'), which(dati.st_sort
    $denominazione_regione == 'Lombardia'),
which(dati.st_sort$denominazione_regione=='P.A. Bolzano'), which(dati.st_
    sort$denominazione_regione=='P.A. Trento'),
which(dati.st_sort$denominazione_regione == 'Piemonte'), which(dati.st_
    sort$denominazione_regione=='Valle d\'Aosta'),
which(dati.st_sort$denominazione_regione == 'Veneto'))

nord <- dati.st_sort[Nord, ]
nord <- nord[order(nord$giorno),]
nord <- nord[nord$giorno < '2020-03-17',]
```

```
y.oss <- nord[nord$giorno >= '2020-03-09' & nord$denominazione_provincia
    == 'Padova',9]
nord <- nord[-which(nord$giorno >= '2020-03-09' & nord$denominazione_
    provincia != 'Padova'),]
nord[nord$giorno >= '2020-03-09' & nord$denominazione_provincia == 'Padova
    ',9] <- NA
nord$id.provincia <- c(rep(1:47,14),rep(42,8))

righe <- which(is.na(nord$nuovi_casi))

nb2INLA("map.adj", Wb.n)
g <- inla.read.graph(filename = "map.adj")

#Proviamo i modelli:

pd_sim <- previsioni(nord)
err.pd_sim <- lapply(pd_sim, function(x) metrics(y.oss$nuovi_casi,x[righe
    ]))
```

La funzione `previsioni` è formata da un insieme di istruzioni come viene descritto nel Codice A.5.

Codice A.7: Esempio di codice per la creazione di una mappa

```
library(tmap)

csi <- mod.bym.inla$marginals.random$id_inla[1:107]
pr <- lapply(csi, function(x) 1-inla.pmarginal(0,x))

a <- as.vector(unlist(pr))

dati.spatial$pr <- a

dati2 <- dati.spatial
prov <- which(colnames(dati2) == 'sigla_provincia')
dati2 <- dati2[,c(prov,10)]
temp <- inner_join(Italy.prov, dati2, by = 'sigla_provincia')
```

```
list_format <- list('text.separator'='a', 'text.or.more'='o piu', 'text.
    less.than'='Meno di')
legend_breaks <- unique(c(0,0.3,0.8,0.9,1))

tm_shape(temp) +
tm_fill('pr', title = 'Probabilita',
textNA = "None counted", style='fixed',breaks=legend_breaks)+
tm_compass(type = "rose",show.labels = 2, position = c("right", "top"),
    size=4)+
tm_layout(legend.outside=T, legend.outside.position=c('left'),
legend.format = list_format, frame=F, attr.position = c('right','top'),
legend.title.fontface = 2, legend.text.size = 0.8)+
tm_borders()
```

Ringraziamenti

Al termine di un percorso così lungo non posso esimermi dal fare dei ringraziamenti, perché in un modo o nell'altro, il successo arriva da una diversità di fattori ed è giusto riconoscerli quando si è raggiunto un traguardo che sembrava solo un miraggio.

Per prima cosa ringrazio il Dipartimento di Scienze Statistiche che mi ha accolto in quel lontano Ottobre 2015 e non mi ha mai dato motivo di andarmene. Tra i vari alti e bassi sono molto orgoglioso e felice delle scelta di aver studiato a Santa Caterina perché mi ha dato veramente tanto, insegnandomi una disciplina che mi ha colpito per la sua complessità e bellezza conquistandomi fino dalla prima regressione lineare. Ringrazio i professori che più di tutti mi hanno insegnato e fatto appassionare alla Statistica, mostrandosi eccellenti nel loro ruolo accademico e nel loro ruolo umano di guida. Tra questi voglio nominare il mio relatore prof. Sartori, anche per la pazienza e la disponibilità garantitami sia nella laurea triennale che in quella magistrale. Voglio ringraziare tutti i miei compagni di corso con i quali ho condiviso le fatiche di questo percorso, ma anche le gioie dei successi e sopratutto tantissime partite a briscola in aula studio.

Successivamente voglio ringraziare il professor Faccioli grazie al quale ho capito cosa fare del mio futuro, senza il quale non avrei mai amato la matematica e certamente non sarei dove e quello che sono ora.

La lista delle persone con le quali ho condiviso almeno un momento in questi cinque anni sarebbe veramente troppo lunga, ringrazio comunque tutti quelli che sono riusciti a sostenermi, che mi hanno strappato un sorriso e con i quali ho trascorso dei bei momenti. In particolare devo nominarne alcune che sono state fondamentali in questo periodo e spero anche nei prossimi a venire: ringrazio Elisa, una chiave di volta per me, senza la quale sarei rimasto a terra molto tempo fa e grazie alla quale mi sono sempre rialzato superando tutti i miei fallimenti. Grazie per avermi fatto migliorare sotto diversi punti di vista e per tutte le avventure

che abbiamo passato insieme. Ringrazio Crive con il quale condivido 20 anni di amicizia, l'unica persona che riesce a passare da discussioni profonde su diversi temi a battute discutibili con una leggerezza unica. Grazie per essere stato sempre presente e pronto anche quando tutto poteva andare per il verso sbagliato. Infine ringrazio Debba con il quale ho costruito un'amicizia strettissima in pochissimo tempo, una persona che mi fa divertire e passare dei bei momenti con una semplicità e spensieratezza unica.

Per concludere ringrazio il mio MSI per avermi dato la possibilità di lavorare adeguatamente anche con operazioni veramente costose, la birra per i momenti di pausa e per le serate in compagnia e il rap per avermi dato la carica di affrontare qualsiasi avversità.

Ad maiora.