Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Tesi di laurea

Verosimiglianza a coppie in un modello per dati binari con dipendenza spaziale.

Relatore Prof.ssa Manuela Cattelan
Dipartimento di Scienze Statistiche

Laureando: Gianmarco Panizza
Matricola N. 1183731

Anno Accademico 2018/2019

A Valentina,
che anche da lontano
mi è sempre stata vicino.

Alla mia famiglia,
che in questi cinque anni
mi ha sempre appoggiato.

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

La stima della struttura di dipendenza presente nei processi spaziali ha suscitato molto interesse nella produzione scientifica degli ultimi vent'anni. I campi casuali rappresentano la struttura matematica per l'analisi statistica di dati spaziali e l'inferenza parametrica basata sul metodo di massima verosimiglianza (ML) è generalmente considerata la miglior opzione per la stima della dipendenza spaziale in questa struttura. Tuttavia, l'adattamento dei modelli tramite ML può risultare problematico. Nel caso di campi gaussiani, ad esempio, il metodo di ML non è praticabile quando l'insieme di dati è grande, poiché la sua valutazione richiede un onere computazionale dell'ordine $O(K^3)$, dove K è il numero di posizioni spaziali. Questo problema si aggrava nella stima del modello di covarianza associato a campi casuali multivariati. Per questo motivo, negli ultimi anni c'è stato un crescente interesse nel proporre nuovi metodi computazionalmente realizzabili per la stima di modelli per dati spaziali gaussiani. In particolare, l'obiettivo è trovare metodi che consentano un ragionevole compromesso tra efficienza statistica e computazionale. L'assunzione di normalità del campo spaziale sottostante semplifica notevolmente la situazione, poiché permette di focalizzarsi solamente sulla struttura del campo di secondo ordine. Sfortunatamente, molte applicazioni all'ecologia e all'epidemiologia mostrano che l'assunzione di normalità è spesso violata (si veda Adler 2008). Un esempio ampiamente discusso in letteratura, e considerato in questa tesi, è il caso dei dati binari con struttura di dipendenza spaziale. A riguardo, Lin e Clayton (2005) considerano un metodo di quasi-verosimiglianza mentre Albert e McShane (1995) utilizzano un approccio basato sulle equazioni di stima generalizzate (GEE) nel contesto dei modelli lineari generalizzati. In ambito geostatistico,

il caso dei dati binari è stato affrontato attraverso diversi approcci alternativi al classico metodo del kriging, come il *disjunctive kriging* o il *trans-Gaussian kriging*. Heagerty e Lele (1998), seguendo invece la letteratura sui modelli gerarchici, hanno derivato un modello basato sulle probabilità congiunte mediante la specificazione di un processo spaziale gaussiano latente. In particolare, con una soglia opportunamente selezionata è possibile passare da un campo gaussiano latente a un campo binario, dal quale possono essere derivate le probabilità marginali attraverso calcoli algebrici. La derivazione della distribuzione congiunta comporta però una somma di integrali gaussiani bidimensionali, quindi anche in questo caso la valutazione della verosimiglianza all'aumentare di K diventa impegnativa dal punto di vista computazionale. Esigenze di questa natura hanno portato numerosi autori a studiare possibili ampliamenti alla definizione di funzione di verosimiglianza, in modo tale da affrontare le difficoltà legate alla sua specificazione. Un'alternativa ai metodi di verosimiglianza ordinaria consiste nell'adottare pseudo-verosimiglianze più semplici, come quelle appartenenti alla classe delle *verosimiglianze composite* (Lindsay, 1988). La verosimiglianza composita si ottiene combinando validi oggetti di verosimiglianza, relativi solitamente a piccoli sottoinsiemi di dati; alcuni esempi sono la *pseudo-verosimiglianza di Besag* (1974) per l'inferenza in modelli spaziali e la *verosimiglianza parziale di Cox* (1975) suggerita per l'inferenza in modelli con rischi proporzionali. L'utilizzo della verosimiglianza composita si riscontra anche nel campo della genetica, delle serie storiche e nell'analisi dei dati di sopravvivenza. In ambito spaziale, Heagerty e Lele (1998) hanno proposto un caso particolare di verosimiglianza composita basato su coppie, noto anche con il nome di *verosimiglianza a coppie*. La verosimiglianza composita è una classe di funzioni di stima che contiene e quindi generalizza i metodi classici di verosimiglianza (si veda Varin *et al.*, 2011 per una revisione completa). Una delle caratteristiche più vantaggiose della verosimiglianza composita è la possibilità di eseguire l'inferenza sui parametri di interesse ad un costo computazionale ridotto. Ciò lo rende un utile metodo di stima quando si trattano insiemi di dati di grandi dimensioni e si considerano modelli statistici assai complessi; inoltre è utilizzabile anche in situazioni dove non è possibile calcolare la verosimiglianza completa. Negli

ultimi anni si è sviluppato un crescente interesse attorno alle *verosimiglianze composite marginali*, ovvero pseudo-verosimiglianze ottenute componendo adeguatamente densità marginali. Quando sono coinvolte distribuzioni marginali di piccole dimensioni, il calcolo degli oggetti di verosimiglianza risulta relativamente semplice; inoltre, sotto determinate condizioni di regolarità, le varie procedure inferenziali hanno proprietà teoriche simili a quelle dei metodi basati sulla verosimiglianza completa, nonostante ci si aspetti una perdita in efficienza.

Nella presente tesi verrà approfondito il metodo della verosimiglianza a coppie e sarà applicato a dati binari organizzati in cluster che presentano una struttura di associazione spaziale. Più precisamente verrà indagata la dipendenza spaziale presente in questo tipo di dati e, con l'ausilio di adeguati modelli parametrici, si cercherà di descriverne la forma. Ciò permetterà di determinare quali coppie di osservazioni includere nella funzione di verosimiglianza e quali escludere perché altrimenti porterebbero a stime distorte e ad una perdita in efficienza. Ampio spazio verrà dedicato anche al calcolo degli errori standard, la cui valutazione risulta spesso problematica a causa di difficoltà legate al calcolo della funzione punteggio di verosimiglianza composita. L'elaborato è organizzato come segue. Il primo capitolo è di rassegna ed introduce gli strumenti e i concetti che verranno impiegati nel seguito della tesi. Sono trattati dettagliatamente i concetti di *verosimiglianza composita* e di *verosimiglianza a coppie* e vengono fornite tutte le procedure inferenziali del metodo; sono inoltre riportati i principali test statistici con i corrispondenti intervalli di confidenza. Il capitolo 2 si apre con una breve revisione dei modelli presenti in letteratura e successivamente vengono definiti il modello proposto per la trattazione di dati binari spaziali organizzati in cluster e la procedura di adattamento derivata dal metodo di verosimiglianza a coppie. Nel capitolo 3 sono riportati gli esiti di due studi di simulazione: il primo considera dei dati generati da una griglia spaziale regolare, mentre nel secondo i dati sono campionati da posizioni casuali nella medesima area. I due studi verranno comparati per valutare se il metodo della verosimiglianza a coppie rimane valido anche in situazioni di disposizione spaziale non regolare.

In particolare verranno confrontate le coperture degli intervalli per i parametri di dipendenza $\boldsymbol{\alpha}$ e regressione $\boldsymbol{\beta}$ sulle due griglie spaziali considerate. Infine, nel capitolo 4, il modello proposto in questo elaborato viene adattato al dataset di un'indagine sulla prevalenza della malaria infantile nel Gambia, descritto in Diggle e Ribeiro (2007) e al dataset *loa-loa* riguardante la diffusione del parassita *Loa loa filariasis* in Africa centrale, esposto in Diggle *et al.* (2007b). In appendice vengono riportate le principali funzioni R utilizzate per la massimizzazione della verosimiglianza a coppie e la derivazione degli errori standard delle stime dei parametri. Come punto di partenza sono state impiegate le funzioni riportate nel *Supplementary materials* di Cattelan e Varin (2018). Tutte le analisi presentate in questa tesi sono state eseguite con il software statistico R (R Core Team, 2018), nella versione 3.5.1.

Capitolo 1

La verosimiglianza composita

1.1 Introduzione

Il presente capitolo è essenzialmente di rassegna, con l'obiettivo di introdurre gli strumenti e le metodologie utilizzate in questa tesi. Il seguente paragrafo è dedicato all'introduzione della verosimiglianza composita e alle sue caratteristiche. Successivamente verranno presentate la *verosimiglianza composita marginale*, costruita a partire da particolari densità marginali, ed un suo caso specifico: la *verosimiglianza a coppie*, che è stata impiegata nello svolgimento delle analisi che verranno mostrate nei seguenti capitoli.
Nel quarto paragrafo vengono definite alcune quantità associate alla verosimiglianza composita, come ad esempio, l'equazione di stima, dove si evidenzia l'importanza della sua proprietà di non distorsione al fine di avere stimatori consistenti, lo stimatore di massima verosimiglianza composita e nel paragrafo successivo le sue proprietà asintotiche.
Gli argomenti presentati nel sesto e settimo paragrafo sono rispettivamente la scelta dei pesi ω_t associati a ciascun sottoinsieme di osservazioni e il calcolo degli errori standard. Infine vengono introdotti i test basati sulla verosimiglianza composita impiegati nella risoluzione di problemi di verifica d'ipotesi ed i relativi intervalli di confidenza.

1.2 Verosimiglianza composita

Il termine *verosimiglianza composita* indica una classe di pseudo-verosimiglianze basate su oggetti di verosimiglianza.

Sia data una variabile casuale $Y = (Y_1, \ldots, Y_n)^T$ con densità congiunta $f(y;\theta)$; inoltre, sia dato un modello statistico parametrico $\mathcal{F} = \{f(y;\theta), y \in \mathcal{Y} \subseteq \mathbb{R}^n, \theta \in \Theta \subseteq \mathbb{R}^q\}$ e un insieme di eventi misurabili $\{\mathcal{A}_1, \ldots, \mathcal{A}_T\}$.
Supponiamo che $f(y;\theta)$ sia difficile da calcolare, ma che per qualche sottoinsieme di dati le verosimiglianze siano facilmente ottenibili. Allora, una *verosimiglianza composita* (CL) è il prodotto delle verosimiglianze corrispondenti a ogni singolo sottoinsieme opportunamente pesato, ovvero è la combinazione di validi oggetti di verosimiglianza:

$$\mathcal{C}L(\theta;y) = \prod_{t=1}^{T} f(y \in \mathcal{A}_t;\theta)^{\omega_t} = \prod_{t=1}^{T} \mathcal{C}L_t(\theta;y)^{\omega_t}, \tag{1.1}$$

dove $\mathcal{C}L_t(\theta;y)$ rappresenta la componente t-esima della verosimiglianza composita e ω_t il relativo peso non negativo. È necessario che ogni componente abbia almeno un parametro in comune con almeno un'altra componente: ciò è essenziale affinché $\mathcal{C}L$ sia una verosimiglianza composita. Se non ci fossero quantità ignote in comune tra le parti, ottimizzare $\mathcal{C}L$ equivarrebbe a massimizzare i singoli $\mathcal{C}L_t$ separatamente, poiché i T insiemi di parametri sarebbero indipendenti tra loro.

Si indica con

$$\mathcal{C}l(\theta;y) = \log \mathcal{C}L(\theta;y),$$

la funzione di log-verosimiglianza composita.

La classe delle verosimiglianze di tipo composito è un insieme di metodi molto vasto ed in continuo sviluppo, che nel corso degli anni ha trovato applicazione in diversi ambiti. Uno di questi è il *metodo di omissione*, nel quale la pseudo-verosimiglianza si ottiene rimuovendo alcuni termini che rendono

complicato il calcolo della verosimiglianza completa. L'auspicio è che tale riduzione di complessità avvenga con una perdita d'informazione su θ nulla o trascurabile, ovvero che la perdita in efficienza di questo metodo risulti tollerabile. Alcuni esempi di verosimiglianza appartenenti a questo metodo sono la *pseudo-verosimiglianza di Besag* (Besag, 1974, 1977) per dati spaziali, la *verosimiglianza di ordine m* per processi stazionari (Azzalini, 1983) e la *verosimiglianza di Stein* (2004) per grandi insiemi di dati spaziali. Anche la famosa *verosimiglianza parziale di Cox* (1975) può essere inclusa in questo gruppo. Un metodo diverso di verosimiglianza composita è costituito dalla *verosimiglianza composita marginale*, costruita combinando verosimiglianze marginali a partire dalla specificazione diretta della distribuzione congiunta di coppie, triplette, o insiemi di osservazioni di ordine superiore. Viene fatto notare che in questo approccio l'interesse è focalizzato sulla composizione di distribuzioni marginali di piccole dimensioni, dato l'importante risparmio computazionale. Perciò questo metodo, consistente nella combinazione di densità marginali, risulta un'alternativa assai attraente, soprattutto nel caso di applicazioni complesse, che vanno dai dati raggruppati in *cluster* e dai dati longitudinali o spaziali fino alle serie temporali e all'analisi di dati di sopravvivenza.
Di seguito verrà approfondito il concetto di verosimiglianza composita marginale attraverso alcune sue specificazioni.

1.3 Verosimiglianza composita marginale

La verosimiglianza composita marginale comporta un notevole risparmio computazionale poiché non richiede la distribuzione congiunta dei dati ma solo una particolare combinazione di densità marginali fino ad un certo ordine. Inoltre, sotto adeguate condizioni di regolarità, le varie procedure inferenziali hanno proprietà teoriche simili a quelle dei metodi basati sulla verosimiglianza completa, sebbene ci si aspetti una perdita di efficienza.

Il tipo di verosimiglianza composita marginale più semplice è la pseudo-verosimiglianza costruita utilizzando le sole marginali univariate, come sotto

l'ipotesi di indipendenza. Si ottiene quindi

$$\mathcal{I}L(\theta; y) = \prod_{i=1}^{n} f(y_i; \theta)^{\omega_i},$$

alla quale in letteratura spesso ci si riferisce con il termine *verosimiglianza di indipendenza* (Chandler e Bate, 2007). La verosimiglianza di indipendenza permette di fare inferenza solamente sui parametri marginali. Qualora siano di interesse anche i parametri responsabili della dipendenza è necessario modellare blocchi di osservazioni, per esempio la verosimiglianza composita costruita utilizzando marginali bivariate è

$$\mathcal{P}L(\theta; y) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} f(y_i, y_j; \theta)^{\omega_{ij}}, \tag{1.2}$$

che viene denominata *verosimiglianza a coppie*. In particolare $f(\cdot)$ è la funzione di densità per la coppia di osservazioni y_i e y_j costruita opportunamente partendo dalle distribuzioni marginali di Y. Di conseguenza si indica con

$$\mathcal{P}l(\theta; y) = \log \mathcal{P}L(\theta; y),$$

la funzione di log-verosimiglianza a coppie.

Un problema interessante consiste nel combinare in modo ottimale le verosimiglianze basate sulle marginali univariate e bivariate (Cox e Reid, 2004). Tuttavia in questa tesi si è limitato l'utilizzo solamente alla verosimiglianza a coppie ($\mathcal{P}L$).

1.4 Quantità associate alla verosimiglianza composita

La funzione punteggio basata sulla verosimiglianza composita si ottiene calcolando la derivata prima della log-verosimiglianza composita

$$U_{\mathcal{C}}(\theta) = \frac{\partial \mathcal{C}l(\theta; y)}{\partial \theta} = \sum_{t=1}^{T} \omega_t \frac{\partial \log \mathcal{C}L(\theta; y)}{\partial \theta}.$$

L'equazione di stima basata sulla verosimiglianza composita, detta equazione di verosimiglianza composita, è data da

$$U_{\mathcal{C}}(\theta) = 0.$$

Lo stimatore di massima verosimiglianza composita, $\hat{\theta}_{\mathcal{C}}$, è la soluzione, se esiste ed è unica, dell'equazione di verosimiglianza composita. Essendo ogni singola componente della verosimiglianza composita una verosimiglianza propria, risulta che la funzione punteggio composita è una combinazione lineare di funzioni punteggio associate ad ogni singolo termine della verosimiglianza composita. Quindi l'equazione di stima ad essa associata risulta non distorta. Tuttavia i singoli contributi $\mathcal{C}L_t(\theta; y)$, $t = 1, \ldots, T$, alla verosimiglianza composita sono moltiplicati tra loro: che essi siano indipendenti o meno, l'inferenza che ne deriva ha le medesime proprietà di quella ottenuta da un modello non correttamente specificato. Di conseguenza, la seconda identità di Bartlett non regge ed è perciò necessario fare una distinzione tra matrice di sensibilità $H(\theta)$ e matrice di variabilità $J(\theta)$:

$$H(\theta) = E_\theta\left(-\frac{\partial U_{\mathcal{C}}(\theta)}{\partial \theta}\right)$$

$$J(\theta) = E_\theta\big(U_{\mathcal{C}}(\theta) U_{\mathcal{C}}(\theta)^T\big) = Var_\theta\big(U_{\mathcal{C}}(\theta)\big)$$

Attraverso la teoria delle equazioni di stima non distorte, si dimostra che la distribuzione asintotica dello stimatore $\hat{\theta}_{\mathcal{C}}$ è Normale di media θ e varianza

$G(\theta)^{-1}$, dove $G(\theta)$ è la *matrice d'informazione di Godambe* associata alla verosimiglianza composita. Questa matrice, detta anche *sandwich information matrix* si può scrivere come

$$G(\theta) = H(\theta)J(\theta)^{-1}H(\theta).$$

Se $\mathcal{C}L(\theta; y)$ fosse una vera verosimiglianza, $H(\theta)$ e $J(\theta)$ coinciderebbero e sarebbero quindi uguali alla matrice di informazione di Fisher. Essendo lo stimatore $\hat{\theta}_{\mathcal{C}}$ non distorto, la disuguaglianza di Cramér-Rao implica che $G(\theta)^{-1} - I(\theta)^{-1}$ sia semi-definita positiva. Questa è invece strettamente positiva se $\hat{\theta}_{\mathcal{C}}$ non è funzione di una statistica sufficiente minimale; in questo caso tale stimatore è sempre meno efficiente del corrispettivo di massima verosimiglianza $\hat{\theta}_{MV}$. Da un punto di vista pratico $G(\theta)$ è solitamente calcolata in $\hat{\theta}_{\mathcal{C}}$; inoltre se il valore atteso rispetto a θ è complesso da calcolare, si possono usare delle approssimazioni empiriche.

1.5 Proprietà asintotiche

In caso di campione indipendente ed identicamente distribuito da una v.c. in $\mathbf{R}^n$, e sotto le usuali condizioni di regolarità delle componenti della verosimiglianza, i seguenti risultati asintotici sono validi (Molenberghs e Verbeke, 2005, pp. 189-202):

- lo stimatore $\hat{\theta}_{\mathcal{C}}$ converge in probabilità al vero ed ignoto valore del parametro:
$$\hat{\theta}_{\mathcal{C}} \xrightarrow{p} \theta;$$

- vale inoltre la teoria asintotica sulle equazioni di stima non distorte, portando al risultato asintotico di normalità (convergenza in distribuzione) dello stimatore di massima verosimiglianza composita:
$$\sqrt{n}(\hat{\theta}_{\mathcal{C}} - \theta) \xrightarrow{d} N\big(0, G(\theta)^{-1}\big). \tag{1.3}$$

Tale risultato rimane valido se si calcola $H(\theta)$ in $\hat{\theta}_{\mathcal{C}}$ o in un altro stimatore consistente di θ.

1.6 La scelta dei pesi ω_t

Quando si considera la verosimiglianza composita si deve affrontare il problema riguardante la scelta dei pesi, in modo tale da suggerire una strategia che migliori la funzione di stima composita. Si ricorda che il ruolo dei pesi nella funzione $\mathcal{C}L$ è di risparmiare tempo di calcolo computazionale e migliorare l'efficienza statistica.

Spesso la stima dei pesi risulta difficile specialmente quando il parametro θ non è unidimensionale (Lindsay, 1988). Inoltre i pesi non sono sempre scelti secondo criteri ottimi. Alcuni autori hanno proposto dei pesi che possono essere considerati delle variabili indicatrici, le quali selezionano i termini da includere nella verosimiglianza composita. In particolare nel caso si consideri la verosimiglianza a coppie in contesto spaziale, come in questa tesi, una funzione di pesi compatta del tipo $w_{ij}(d) > 0$ se $s_i - s_j \leq d$, e 0 altrimenti, presenta evidenti vantaggi computazionali, dove s_i è il vettore di coordinate dell'i-esima osservazione e d è una generica distanza prefissata. Inoltre, anche una semplice funzione di riduzione dei pesi, $w_{ij}(d) = 1$ se $s_i - s_j \leq d$, e 0 altrimenti, può migliorare l'efficienza come è stato mostrato in Joe e Lee (2009), Davis e Yau (2011) e Bevilacqua *et al.* (2012). L'intuizione alla base di questo approccio è che le correlazioni tra coppie di osservazioni distanti sono spesso quasi nulle. Pertanto l'uso di tutte le coppie può perdere efficienza poiché troppe coppie di osservazioni ridondanti possono distorcere le informazioni contenute in coppie di osservazioni vicine e portare a stime non consistenti.

1.7 Calcolo degli errori standard

Per calcolare gli errori standard è necessario disporre di stime consistenti delle matrici $H(\theta)$ e $J(\theta)$, le quali compongono la matrice d'informazione di Godambe. Solitamente le matrici $H(\theta)$ e $J(\theta)$ vengono stimate empiricamente, sfruttando gruppi di dati indipendenti o pseudo-indipendenti, oppure attraverso simulazione. Tipicamente tali matrici sono stimate in $\hat{\theta}_{\mathcal{C}}$. Quando sono disponibili gruppi di osservazioni indipendenti, come ad esempio quando i dati sono divisi in K cluster, è possibile stimare $J(\theta)$ come

$$\hat{J}^E(\theta) = \frac{1}{K}\sum_{k=1}^{K} U_{\mathcal{C}_k}(\theta)U_{\mathcal{C}_k}(\theta)^T,$$

dove $U_{\mathcal{C}_k}(\theta)$ indica gli elementi dello score di verosimiglianza composita che coinvolgono solamente osservazioni del cluster k. In ambiti di serie storiche e dati spaziali, quando è possibile identificare gruppi di dati a bassa dipendenza, questo metodo può essere applicato a tali gruppi.
Uno stimatore empirico per la matrice di sensibilità $H(\theta)$ è

$$\hat{H}^E(\theta) = -\frac{1}{\Omega_K}\sum_{k=1}^{K} \omega_k \frac{\partial\, U_{\mathcal{C}_k}(\hat{\theta}_{\mathcal{C}})}{\partial\, \theta},$$

dove $\Omega_K = \sum_{k=1}^{K} \omega_k$, oppure se l'Hessiano è difficile da trattare

$$\hat{H}^E(\theta) = -\frac{1}{\Omega_K}\sum_{k=1}^{K} \omega_k U_{\mathcal{C}_k}(\hat{\theta}_{\mathcal{C}})U_{\mathcal{C}_k}(\hat{\theta}_{\mathcal{C}})^T,$$

dato che la seconda identità di Bartlett rimane valida per ogni singolo termine di verosimiglianza. La stima empirica di $H(\theta)$ e $J(\theta)$ non richiede ulteriori assunzioni rispetto a quelle fatte per la funzione di verosimiglianza a coppie, che consistono solamente nella specificazione delle probabilità marginali.

Un modo alternativo per stimare la matrice di informazione di Godambe è attraverso simulazione, che richiede assunzioni riguardo la completa distribuzione dei dati. Tuttavia tali assunzioni non sono sempre possibili. Sebbene

il presupposto della distribuzione completa dei dati possa apparire una limitazione importante di questo metodo, nella maggior parte delle applicazioni della verosimiglianza composita si presume un modello completo per i dati, ma le difficoltà nel calcolare la funzione di verosimiglianza portano all'uso di una verosimiglianza composita. In casi come questi la funzione di verosimiglianza è difficile da ottenere, ma può essere più semplice simulare dal modello completo.

Sia $y^m, m = 1, \ldots, M$, l'm-esimo vettore risposta simulato da $f(y;\theta)$, la distribuzione completa dei dati. Allora, le stime Monte Carlo di $H(\theta)$ e $J(\theta)$ sono

$$\hat{J}^S(\theta) = \frac{1}{M}\sum_{m=1}^{M} U_{\mathcal{C}}(\theta; y^m) U_{\mathcal{C}}(\theta; y^m)^T$$

e

$$\hat{H}^S(\theta) = -\frac{1}{M}\sum_{m=1}^{M} \frac{\partial\, U_{\mathcal{C}}(\theta; y^m)}{\partial\, \theta}.$$

Di nuovo, nella stima di $H(\theta)$ è possibile sfruttare la seconda identità di Bartlett, che è valida per ogni componente della verosimiglianza composita. Inoltre queste approssimazioni via Monte Carlo verranno calcolate in $\hat{\theta}_{\mathcal{C}}$; ciò significa che durante il calcolo, ad esempio, di $\hat{J}^S(\hat{\theta}_{\mathcal{C}})$ i dati vengono simulati da $f(y;\hat{\theta}_{\mathcal{C}})$.

1.8 Test d'ipotesi

In questo paragrafo vengono illustrate le modifiche che consentono l'applicazione al campo della verosimiglianza composita dei classici test d'ipotesi, per definire ipotesi nulle puntiformi contro ipotesi alternative bidirezionali del tipo

$$\begin{cases} H_0 \colon \theta = \theta_0 \\ H_1 \colon \theta \neq \theta_0\,. \end{cases} \tag{1.4}$$

Per quanto riguarda il test di Wald si può usare la (1.3) mentre per la statistica punteggio basta considerare che $Var_\theta(U_{\mathcal{C}}(\theta)) = J(\theta)$; si arriva quindi

alle usuali distribuzioni asintotiche χ^2. Diversamente dai precedenti, il test del rapporto di verosimiglianza si presenta nella solita forma, ma con una differente distribuzione asintotica: vale a dire una somma pesata di χ^2_1 indipendenti (si veda Varin et al., 2011). Esistono però numerose correzioni del test che consentono di fare riferimento, come per gli altri due, al χ^2_q (Pace, Salvan e Sartori, 2011); di seguito sarà illustrata in dettaglio una possibile correzione.
Nel caso in cui si sia interessati solo ad un sottovettore ψ di $\theta = (\psi, \tau)$, di dimensione d, abbiamo a che fare con un test profilo, dove il sistema d'ipotesi diventa

$$\begin{cases} H_0 \colon \psi = \psi_0 \\ H_1 \colon \psi \neq \psi_0 \, . \end{cases} \tag{1.5}$$

Definiamo quindi le seguenti quantità che saranno utili nel prosieguo del paragrafo:

- $\hat{\theta}_{\mathcal{C}\psi_0} = (\psi_0, \hat{\tau}_{\mathcal{C}\psi_0})$, la stima vincolata sotto H_0;
- $H(\theta) = \begin{bmatrix} H_{\psi\psi} & H_{\psi\tau} \\ H_{\tau\psi} & H_{\tau\tau} \end{bmatrix}$ e $H(\theta)^{-1} = \begin{bmatrix} H^{\psi\psi} & H^{\psi\tau} \\ H^{\tau\psi} & H^{\tau\tau} \end{bmatrix}$,

dove, ad esempio, $H_{\psi\tau} = E_\theta\Big(-\frac{\partial U_{\mathcal{C}\psi}(\theta)}{\partial \tau^T}\Big)$ con $U_{\mathcal{C}\psi}(\theta) = \frac{\partial \mathcal{C}l(\theta;y)}{\partial \psi}$.
Notazione analoga vale per le matrici $J(\theta)$ e $G(\theta)$.

1.8.1 Il test di Wald

Grazie alla normalità asintotica di $\hat{\theta}_{\mathcal{C}}$, sotto H_0 vale

$$W^e_{\mathcal{C}}(\theta) = (\hat{\theta}_{\mathcal{C}} - \theta_0)^T G(\theta)(\hat{\theta}_{\mathcal{C}} - \theta_0) \dot{\sim} \chi^2_q.$$

La formulazione corrispondente al test d'ipotesi profilo (1.5) è invece

$$W^e_{\mathcal{C}_P}(\psi) = (\hat{\psi}_{\mathcal{C}} - \psi_0)^T (G^{\psi\psi})^{-1} (\hat{\psi}_{\mathcal{C}} - \psi_0) \dot{\sim} \chi^2_d.$$

Come nel caso della classica verosimiglianza, il test di Wald non è invariante rispetto a riparametrizzazioni; inoltre le regioni di confidenza che ne derivano

presentano forma ellittica, vincolo stringente e non sempre adeguato rispetto alla forma della verosimiglianza composita.

1.8.2 Il test di Rao

La sua applicazione si basa sulla normalità asintotica della funzione $U_{\mathcal{C}}(\theta)$ (Rotnitzky e Jewell, 1990). Il test ha la seguente forma

$$W^u_{\mathcal{C}}(\theta) = U_{\mathcal{C}}(\theta)^T J(\theta)^{-1} U_{\mathcal{C}}(\theta),$$

con distribuzione asintotica χ^2_q sotto H_0 (Geys et al., 1999). Il medesimo test nel caso in cui il sistema d'ipotesi coinvolga solo una parte dei parametri, diviene

$$W^u_{\mathcal{C}_P}(\psi) = U_{\mathcal{C}\psi}(\hat{\theta}_{\mathcal{C}\psi_0})^T H^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})\{G^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})\}^{-1} H^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})\, U_{\mathcal{C}\psi}(\hat{\theta}_{\mathcal{C}\psi_0}) \,\dot{\sim}\, \chi^2_d.$$

Tale statistica ha i vantaggi di poter essere ottenuta lavorando solo con il modello nullo e di essere invariante rispetto a riparametrizzazioni monotone. Ciò nonostante potrebbe essere numericamente instabile nell'ambito della verosimiglianza composita e più in generale in quello delle equazioni di stima generalizzate (Rotnitzky e Jewell, 1990).

1.8.3 Il test di Wilks

Entrambi i test sopra descritti presentano evidenti limiti; il test del rapporto di verosimiglianza rappresenta un'alternativa in grado di garantire invarianza e stabilità dei risultati. La formulazione non varia rispetto alla sua applicazione nell'ambito classico della verosimiglianza. Senza e con parametri di disturbo, facenti rispettivamente riferimento ai sistemi d'ipotesi (1.4) e (1.5), il test è uguale a

$$W_{\mathcal{C}}(\theta) = -2\{\mathcal{C}l(\theta_0) - \mathcal{C}l(\hat{\theta}_{\mathcal{C}})\} \quad \text{e} \quad W_{\mathcal{C}_P}(\psi) = -2\{\mathcal{C}l(\hat{\theta}_{\mathcal{C}\psi_0}) - \mathcal{C}l(\hat{\theta}_{\mathcal{C}})\}.$$

La differenza rispetto alla classica verosimiglianza risiede nella sua distribuzione asintotica nulla; per quanto riguarda $W_{\mathcal{C}}(\theta)$, essa è uguale a:

$$\sum_{\alpha=1}^{q} \mu_\alpha(\theta) X_\alpha,$$

dove $X_1, \ldots, X_q$ sono v.c. i.i.d. χ^2_1, con $\mu_1(\theta), \ldots, \mu_q(\theta)$ autovalori di $J(\theta)H(\theta)^{-1} = H(\theta)G(\theta)^{-1}$ (Geys et al., 1999).
Per $W_{\mathcal{C}_P}(\psi)$, q (la dimensione di θ) è sostituito da d (la lunghezza di ψ) e $\mu_1(\theta), \ldots, \mu_q(\theta)$ da $\nu_1(\theta), \ldots, \nu_d(\theta)$, autovalori di $(H^{\psi\psi})^{-1}G^{\psi\psi}$.
È possibile calcolare gli autovalori, nel primo caso in θ_0, e nel secondo in $\hat{\theta}_{\mathcal{C}\psi_0}$; tuttavia, sebbene il massimo vincolato sarebbe più appropriato, i risultati asintotici sono ugualmente validi se viene utilizzato in entrambe $\theta = \hat{\theta}_{\mathcal{C}}$, ossia la stima del modello non ristretto, in quanto stimatore consistente di θ.
Per applicare il test è quindi necessario stimare, attraverso una simulazione, la sua distribuzione asintotica nulla; alternativamente è possibile applicare una correzione affinché questa sia ricondotta al classico χ^2, con q (o d) gradi di libertà.

La correzione del test

In letteratura sono state proposte una serie di correzioni. In particolare esponiamo quella presentata in Pace, Salvan e Sartori (2011)

$$W_{\mathcal{C}}(\theta)_{INV} = \frac{U_{\mathcal{C}}(\theta)^T J(\theta)^{-1} U_{\mathcal{C}}(\theta)}{U_{\mathcal{C}}(\theta)^T H(\theta)^{-1} U_{\mathcal{C}}(\theta)} W_{\mathcal{C}}(\theta) = \frac{W^u_{\mathcal{C}}(\theta)}{U_{\mathcal{C}}(\theta)^T H(\theta)^{-1} U_{\mathcal{C}}(\theta)} W_{\mathcal{C}}(\theta).$$

Essa risulta invariante e riporta il test $W_{\mathcal{C}}(\theta)_{INV}$ ad avere distribuzione asintotica nulla χ^2_q; inoltre gli intervalli di confidenza (IC) relativi presentano buona copertura. L'analogo in caso di parametri di disturbo τ , con solo

interesse nel sottoinsieme ψ di θ diviene

$$W_{\mathcal{C}_P}(\psi)_{INV} = \frac{U_{\mathcal{C}\psi}(\hat{\theta}_{\mathcal{C}\psi_0})^T H^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})\{G^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})\}^{-1} H^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})\, U_{\mathcal{C}\psi}(\hat{\theta}_{\mathcal{C}\psi_0})}{U_{\mathcal{C}\psi}(\hat{\theta}_{\mathcal{C}\psi_0})^T H^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})\, U_{\mathcal{C}\psi}(\hat{\theta}_{\mathcal{C}\psi_0})} W_{\mathcal{C}_P}(\psi)$$
$$= \frac{W^u_{\mathcal{C}_P}(\psi)}{U_{\mathcal{C}\psi}(\hat{\theta}_{\mathcal{C}\psi_0})^T H^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})\, U_{\mathcal{C}\psi}(\hat{\theta}_{\mathcal{C}\psi_0})} W_{\mathcal{C}_P}(\psi),$$

che ha distribuzione approssimata χ^2_d sotto H_0. Nel caso in cui il parametro di interesse ψ sia unidimensionale i test globale e profilo si semplificano rispettivamente nelle forme

$$W_{\mathcal{C}}(\theta)_{INV} = \frac{H(\theta)}{J(\theta)} W_{\mathcal{C}}(\theta) \quad \text{e} \quad W_{\mathcal{C}_P}(\psi)_{INV} = \frac{H^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})}{G^{\psi\psi}(\hat{\theta}_{\mathcal{C}\psi_0})} W_{\mathcal{C}_P}(\psi).$$

1.9 Intervalli di confidenza

Illustriamo la struttura delle regioni di confidenza derivate dai test appena descritti; concentriamoci sui soli IC profilo unidimensionali: definiamo ψ come una costante ignota.
L'IC basato sul test di Wald si fonda sul risultato (1.3); dal quale segue

$$IC^{Wald}_{1-\alpha}(\psi) = \left[\hat{\psi} \pm z_{1-\alpha/2}\sqrt{G(\theta)_{\psi\psi}}\right],$$

dove $z_{1-\alpha/2}$ indica il quantile di livello $1-\alpha/2$ della v.c. $N(0,1)$.
L'IC derivato dal test del punteggio è invece ottenuto numericamente, cercando i valori soglia che definiscono il minor intervallo

$$IC^{Rao}_{1-\alpha}(\psi) = [\psi^L, \psi^U] \quad \text{t.c.} \quad W^u_{\mathcal{C}_P}(\psi) < \chi^2_{1,1-\alpha}\, \forall\, \psi \in [\psi^L, \psi^U].$$

Quello relativo al rapporto di verosimiglianza sostituisce $W_{\mathcal{C}_P}(\psi)$ a $W^u_{\mathcal{C}_P}(\psi)$.

Capitolo 2

Modelli per dati binari in cluster spaziali

2.1 Modelli presenti in letteratura

In ambito epidemiologico spesso si incontrano dati binari organizzati in cluster derivanti dal campionamento gerarchico di soggetti all'interno di unità più grandi. Un caso importante è dato dalle indagini sulla prevalenza di malattie, utilizzate per studiare la variabilità spaziale di una malattia in una regione attraverso il campionamento di soggetti all'interno di diverse unità geografiche, come comunità o villaggi. Ad esempio, Thomson *et al.* (1999) hanno combinato dati epidemiologici e satellitari per valutare l'efficacia dell'impiego della zanzariera da letto per prevenire la malaria nei bambini campionati da vari villaggi del Gambia. Clements *et al.* (2006) hanno studiato i fattori che influenzano le infezioni da schistosomiasi nella Tanzania nord-occidentale, raccogliendo dati sui bambini nelle scuole primarie. Getachew *et al.* (2013) hanno esaminato l'effetto di una diga idroelettrica sull'incidenza della malaria nei bambini campionati dai villaggi dell'Etiopia meridionale.
La tipica assunzione di indipendenza tra cluster risulta dubbia nelle indagini di prevalenza di malattie, infatti non prendere in considerazione la dipendenza spaziale tra cluster può avere un'influenza sostanziale sulla correttezza delle conclusioni inferenziali (Thomson *et al.*, 1999; Clements *et al.*, 2006).

Sono stati proposti vari approcci per l'analisi di dati binari raggruppati spazialmente. Diggle *et al.* (2002b) ad esempio hanno assunto un modello di regressione logistica con due effetti casuali, per gestire sia l'eterogeneità all'interno che la dipendenza spaziale tra cluster. Per quanto riguarda la modellazione marginale, le equazioni di stima generalizzate (GEE), rese popolari da Liang e Zeger (1988), hanno trovato largo impiego in analisi di dati raggruppati in cluster e longitudinali. Più precisamente il metodo delle equazioni di stima generalizzate del primo ordine (GEE1) calcola le stime dei parametri a partire da un modello per le previsioni marginali della risposta e da una matrice contenente la struttura di correlazione per le osservazioni appartenenti allo stesso cluster. Uno dei vantaggi del metodo GEE1 è la sua semplicità computazionale. Inoltre, porta a stime consistenti dei parametri di regressione anche in situazioni di errata specificazione della struttura di associazione, sebbene possa esserci qualche perdita di efficienza. Tuttavia, anche se il modello assunto per la struttura di correlazione è adeguato, GEE1 può essere inefficiente per la stima dei parametri delle associazioni a coppie. Un'alternativa alle GEE1 è il metodo delle equazioni di stima generalizzate del secondo ordine (GEE2), che stima simultaneamente i parametri di regressione marginale e i parametri di associazione a coppie. Uno svantaggio del metodo GEE2 è che le stime GEE2 dei parametri di regressione non sono robuste rispetto ad un'errata specificazione della struttura di associazione. Inoltre, come sottolineato da Carey *et al.* (1993), l'implementazione delle GEE2 comporta l'inversione di matrici di dimensione $O(n_k^2) \times O(n_k^2)$ e quindi il metodo è proibitivo dal punto di vista computazionale anche quando le numerosità dei cluster n_k sono moderate. Un'altra opzione per modellare dati binari correlati è l'*alternating logistic regression* (ALR), proposta da Carey *et al.* (1993). Il modello alterna l'uso di GEE1 per stimare i coefficienti di regressione date le stime degli odds ratio con una regressione logistica di ciascuna risposta sulle altre appartenenti allo stesso cluster per aggiornare le stime degli odds ratio. Pertanto, contrariamente al metodo GEE1 in cui la struttura di correlazione è considerata un parametro di disturbo, qui può essere esplorata e stimata combinando le GEE1 con un'equazione di regressione logistica con offset. Il vantaggio dell'approccio ALR è che com-

porta solo l'inversione di matrici di dimensione $O(n_k) \times O(n_k)$. Inoltre, le stime ALR dei parametri di regressione marginale sono robuste rispetto ad un'errata specificazione della struttura di correlazione.

In generale sono possibili due ampi approcci analitici: (1) modelli marginali o sulla media della popolazione, di solito stimati tramite GEE e (2) modelli lineari generalizzati ad effetti misti (GLMM), ovvero modelli la cui interpretazione è soggetto-specifica. Quindi i modelli che considerano effetti casuali sono appropriati quando l'interesse scientifico risiede nella valutazione degli effetti soggetto-specifici (Diggle *et al.*, 2002a). Al contrario, i modelli marginali risultano più appropriati quando l'obiettivo dell'analisi consiste nel cogliere gli effetti sul soggetto medio della popolazione. Diversamente dai riferimenti precedenti che utilizzano equazioni di stima generalizzate per l'analisi di dati spazio-dipendenti, l'approccio impiegato in questo elaborato modella la dipendenza spaziale in termini di odds ratio a coppie per evitare problemi che insorgono quando la dipendenza tra i dati binari viene modellata in funzione della correlazione. Nei seguenti paragrafi questa differenza verrà approfondita e dimostrata. Il modello considerato viene stimato con il metodo della verosimiglianza a coppie, descritto nel capitolo precedente, e si limita quindi a fare assunzioni sulle distribuzioni univariate e bivariate. Nei seguenti paragrafi viene presentata per esteso la sua formulazione con l'introduzione delle quantità necessarie.

2.2 Notazione e formulazione

I dati organizzati in cluster sono generalmente indicati usando due pedici, uno dei quali rappresenta le osservazioni e l'altro il cluster al quale l'osservazione appartiene. Tale notazione è conveniente quando osservazioni in cluster differenti sono indipendenti. Nelle applicazioni spaziali, dati in cluster distinti potrebbero non essere indipendenti ed è molto conveniente adottare la seguente notazione a singolo indice. Sia n_k il numero di osservazioni nel cluster $k = 1, \ldots, K$, e $n = \sum_{k=1}^{K} n_k$ sia il numero totale di osservazioni; allora $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ è il vettore che si ottiene impilando le osservazioni in modo tale che quelle provenienti dallo stesso cluster siano contigue.

Si suppone che l'interesse scientifico stia nella stima dei coefficienti di un modello marginale logistico

$$\text{logit}\{\pi_i(\boldsymbol{\beta})\} = \mathbf{x}_i^T\boldsymbol{\beta}, \tag{2.1}$$

dove $\pi_i(\boldsymbol{\beta}) = \Pr(Y_i = 1)$, $\mathbf{x}_i$ è un vettore p-dimensionale di covariate e $\boldsymbol{\beta}$ sono i corrispondenti coefficienti di regressione. Le equazioni di stima ottimali per $\boldsymbol{\beta}$ sono

$$\mathrm{D}^T\mathrm{V}^{-1}(\mathbf{y} - \pi) = \mathbf{0}, \tag{2.2}$$

dove $\mathrm{D} = \partial\pi/\partial\boldsymbol{\beta}$, V è la matrice di covarianza di $\mathbf{Y}$, $\mathbf{y} = (y_1, \ldots, y_n)^T$ è il vettore dei valori osservati di $\mathbf{Y}$ e $\pi = (\pi_1, \ldots, \pi_n)^T$ è il corrispondente vettore di probabilità marginali. La dipendenza di V e π dai parametri marginali $\boldsymbol{\beta}$ è stata tolta dalla notazione per convenienza. Nel modello di regressione logistica (2.1), la matrice Jacobiana è $\mathrm{D} = \mathrm{A}\mathbf{X}$, dove $\mathrm{A} = \text{diag}\{\pi_i(1 - \pi_i)\}$ e $\mathbf{X}$ è la matrice del disegno. Nel metodo molto diffuso delle equazioni di stima generalizzate (Zeger *et al.*, 1988) per dati di tipo longitudinale e di cluster, la matrice di covarianza è calcolata come $\mathrm{V}_{GEE} = \mathrm{A}^{1/2}\mathrm{R}\mathrm{A}^{1/2}$, dove $\mathrm{R} = \text{diag}\{R_1, \ldots, R_K\}$ è una matrice di correlazione blocco-diagonale costruita assumendo una parametrizzazione di lavoro ortogonale agli effetti medi. Diversi ricercatori hanno severamente sconsigliato l'uso delle matrici di correlazione nelle equazioni di stima generalizzate per dati binari perché, come verrà spiegato nel seguente paragrafo, la correlazione tra due osservazioni binarie è notevolmente vincolata dalle medie marginali (Diggle *et al.*, 2002a). Al contrario, l'odds ratio a coppie è svincolato dalle medie marginali e fornisce una misura naturale e flessibile dell'associazione a coppie in dati binari. Sia $\pi_{ij} = \Pr(Y_i = 1, Y_j = 1)$; allora l'odds ratio a coppie di Y_i e Y_j è

$$\psi_{ij} = \frac{\pi_{ij}(1 - \pi_i - \pi_j + \pi_{ij})}{(\pi_i - \pi_{ij})(\pi_j - \pi_{ij})}. \tag{2.3}$$

La formula (2.3) può essere invertita per ottenere π_{ij} dati π_i, π_j e ψ_{ij} (Mardia, 1967):

$$\pi_{ij}(\psi_{ij}, \pi_i, \pi_j) = \begin{cases} \pi_i \pi_j, & \text{se } \psi_{ij} = 1, \\ \frac{1+(\pi_i+\pi_j)(\psi_{ij}-1)-G(\pi_i,\pi_j,\psi_{ij})}{2(\psi_{ij}-1)}, & \text{se } \psi_{ij} \neq 1, \end{cases} \tag{2.4}$$

con

$$G(\pi_i, \pi_j, \psi_{ij}) = \sqrt{[1 + (\pi_i + \pi_j)(\psi_{ij} - 1)]^2 + 4\psi_{ij}(1 - \psi_{ij})\pi_i \pi_j}.$$

Date le probabilità marginali π_i e π_j e l'odds ratio a coppie ψ_{ij}, le entrate della matrice di covarianza V, necessarie per risolvere le equazioni di stima ottimali in (2.2), vengono quindi calcolate come $\text{Cov}(Y_i, Y_j) = \pi_{ij}(\psi_{ij}, \pi_i, \pi_j) - \pi_i \pi_j$. Visto che l'odds ratio a coppie assume valori non negativi, risulta naturale descrivere la struttura associativa di dati binari modellando il logaritmo del rapporto di probabilità a coppie $\gamma_{ij} = \log(\psi_{ij})$.

2.3 Il lorelogramma

Al fine di osservare la struttura associativa tra dati binari organizzati in cluster spaziali, e quindi dipendenti, si è fatto ricorso in questo elaborato ad una particolare rappresentazione grafica chiamata lorelogramma. Di seguito viene riportata una breve introduzione di questo strumento e dei vantaggi che comporta il suo utilizzo rispetto ad altri più conosciuti.

Heagerty e Zeger coniarono il termine *lorelogramma* nell'anno 1998 per indicare il logaritmo dell'odds ratio a coppie visto come una funzione di una qualche distanza tra le osservazioni, in alternativa al correlogramma e al variogramma che hanno trovato largo impiego nelle serie storiche e nella geostatistica rispettivamente.
Visto il ruolo fondamentale che ha avuto il lorelogramma nello svolgimento di questa tesi, viene prima fornita una breve introduzione ai concetti

di variogramma e correlogramma, ed in seguito viene mostrato il motivo dell'introduzione di quest'altra funzione.
Nello specifico sono stati considerati modelli isotropici nei quali la dipendenza tra osservazioni si assume dipenda solamente dalla distanza tra queste.

Il variogramma (Cressie 1993; Diggle 1990) è una funzione descrittiva della struttura di dipendenza isotropica in serie storiche, osservazioni spaziali e dati longitudinali. Senza perdita di generalità, è possibile richiamare il variogramma e la sua stima specificatamente per risposte longitudinali continue. Si considerino le misurazioni Z_{ij}, $j = 1, \ldots, n_i$ sui soggetti $i = 1, \ldots, N$ rilevate al tempo t_{ij}. I residui sono definiti come $R_{ij} = Z_{ij} - E[Z_{ij}]$. Per $j \neq k$ il variogramma è definito come

$$\delta(|t_{ij} - t_{ik}|) = \frac{1}{2} E[(R_{ij} - R_{ik})^2].$$

Un'altra funzione spesso utilizzata insieme a questa è il covariogramma

$$C(|t_{ij} - t_{ik}|) = \mathrm{Cov}(R_{ij}, R_{ik}).$$

Infine, il correlogramma è una versione riscalata del covariogramma

$$r(|t_{ij} - t_{ik}|) = \frac{C(|t_{ij} - t_{ik}|)}{V},$$

dove $V = \mathrm{Var}(R_{ij})$. Ognuna di queste funzioni cattura la dipendenza isotropica equivalentemente se la varianza marginale esiste. Le principali differenze stanno nell'interpretazione di queste funzioni e nelle proprietà statistiche dei loro stimatori. Per risposte continue, il variogramma viene stimato in modo non parametrico a partire dal variogramma empirico, il quale è definito come un lisciamento del quadrato delle differenze residue, $\hat{\delta}(u) = \frac{1}{2}(R_{ij} - R_{ik})^2$, rispetto alla differenza temporale $u = |t_{ij} - t_{ik}|$ (Diggle *et al.* 2002a). Di conseguenza il correlogramma empirico si ottiene da $\hat{r}(u) = 1 - \hat{\delta}(u)/\hat{V}$.
Variogramma e correlogramma empirico forniscono una rappresentazione grafica della struttura di dipendenza isotropica, che può essere impiegata per guidare la selezione di un appropriato modello parametrico.

Per dati binari o categoriali, la varianza è una funzione della media, pertanto il variogramma o il covariogramma non sono appropriati a descrivere la dipendenza. Il correlogramma potrebbe essere utilizzato, ma per risposte binarie è noto che la correlazione è vincolata dalle medie a causa della disuguaglianza di Frechet: $\Pr(Z_1 = 1, Z_2 = 1) \leq \min(\Pr(Z_1 = 1), \Pr(Z_2 = 1))$ (Lipsitz, Laird, e Harrington 1991). Per esempio, sia $\mu_1 = E[Z_{i1}]$, $\mu_2 = E[Z_{i2}]$ e si assuma $\mu_1 \leq \mu_2$. Allora $E[Z_{i1}Z_{i2}] \leq \mu_1$ e attraverso dei semplici calcoli si arriva a dimostrare che $\mathrm{Corr}(Z_{i1}, Z_{i2}) \leq \left(\frac{\mu_1(1-\mu_2)}{(1-\mu_1)\mu_2}\right)^{1/2}$.
Per valori delle medie equivalenti non ci sono restrizioni, tuttavia, se le due medie differiscono, l'intervallo ammissibile per la correlazione viene drasticamente ridotto. Contrariamente alla correlazione, l'odds ratio a coppie è completamente svincolato dalle medie marginali; per i dati binomiali è definito come

$$\psi(Z_{i1}, Z_{i2}) = \frac{\Pr(Z_{i1} = 1, Z_{i2} = 1)\Pr(Z_{i1} = 0, Z_{i2} = 0)}{\Pr(Z_{i1} = 0, Z_{i2} = 1)\Pr(Z_{i1} = 1, Z_{i2} = 0)},$$

ed operativamente viene calcolato con la formula (2.3).

2.4 Modelli proposti per il lorelogramma spaziale

In seguito all'introduzione del lorelogramma ad opera di Heagerty e Zeger, il suo impiego è stato esteso alla modellazione dell'associazione spaziale tra dati binari e combinato con i più tradizionali concetti della geostatistica.
Sia u_i il vettore di coordinate dell'i-esima osservazione. In questa tesi, ci si concentra su situazioni dove tutte le osservazioni in un cluster hanno le stesse coordinate, come nell'applicazione al caso della malaria infantile presentata nel capitolo 4, dove sono note solamente le coordinate dei villaggi a cui i bambini appartengono. Perciò, la distanza tra le osservazioni è diversa da 0 solo se queste appartengono a cluster differenti.
Il lorelogramma spaziale tra l'osservazione i e l'osservazione j viene definito

come la funzione $\gamma_{ij} = \gamma(u_i, u_j)$ e viene detta *shift invariant* se $\gamma_{ij} = \gamma(u_i - u_j)$. Una situazione di isotropia corrisponde al caso speciale in cui $\gamma_{ij} = \gamma(d_{ij})$, dove $d_{ij} = \|u_i - u_j\|$ è una certa distanza tra le due coordinate. Il comportamento atteso del lorelogramma isotropico assomiglia a quello di una funzione di covarianza spaziale isotropica: assume il valore più grande all'inizio per poi tendere a 0 all'aumentare della distanza tra cluster.

Come proposto da Cattelan e Varin (2018), in questo elaborato il logaritmo dell'odds ratio a coppie è stato modellato come una funzione della distanza tra le osservazioni; questo metodo esplorativo è stato impiegato per permettere la valutazione della struttura di dipendenza presente in risposte binarie. Con lo scopo di ottenere una più adatta descrizione dell'associazione tra questo tipo di dati, sono stati considerati diversi modelli parametrici presenti in letteratura per modellare la covarianza spaziale.
La forma generale dei modelli per il lorelogramma spaziale isotropico considerati è la seguente:

$$\gamma(d_{ij}; \boldsymbol{\alpha}) = \alpha_1 \mathbb{1}(d_{ij} = 0) + \alpha_2 \rho(d_{ij}/\alpha 3), \tag{2.5}$$

dove $\mathbb{1}(\mathbf{E})$ è la funzione indicatrice dell'evento $\mathbf{E}$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$ è un vettore di parametri non negativi relativi a *nugget*, *sill* e *range*, e $\rho(\cdot)$ è la funzione di dipendenza spaziale tra cluster. I parametri $\boldsymbol{\alpha}$, presi in prestito dalla letteratura geostatistica, colgono rispettivamente la variabilità casuale non correlata alla distanza (ad esempio errori di misura), la variabilità correlata alla distanza e la distanza oltre la quale non è più presente correlazione spaziale tra coppie di osservazioni. Ponendo $\alpha_2 = 0$, viene considerato un modello che include solamente il parametro di *nugget* e corrisponde ad assumere un valore dell'odds ratio interscambiabile tra le coppie di osservazioni all'interno dello stesso cluster ed uno indipendente per le coppie appartenenti a cluster differenti. Al contrario, valori di $\alpha_2 > 0$ indicano la presenza di dipendenza spaziale tra osservazioni provenienti da cluster diversi. Per la scelta di $\rho(\cdot)$, alcuni esempi di modello spaziale da applicare al lorelogramma empirico sono elencati in Tabella 2.1 e il loro adattamento è illustrato di seguito in Figura 2.1.

Modello	*Funzione di dipendenza* $\rho(x)$	*Range*	*Monotona*
Esponenziale	exp(-x)	Infinito	Sì
Gaussiano	exp(-x^2)	Infinito	Sì
Sferico	$(1 - 3/2x + 0.5x^3)\mathbb{1}(x < 1)$	Finito	Sì
Wave	$\sin(x)/x$	Infinito	No

Tabella 2.1: Proprietà dei modelli considerati per il lorelogramma spaziale.

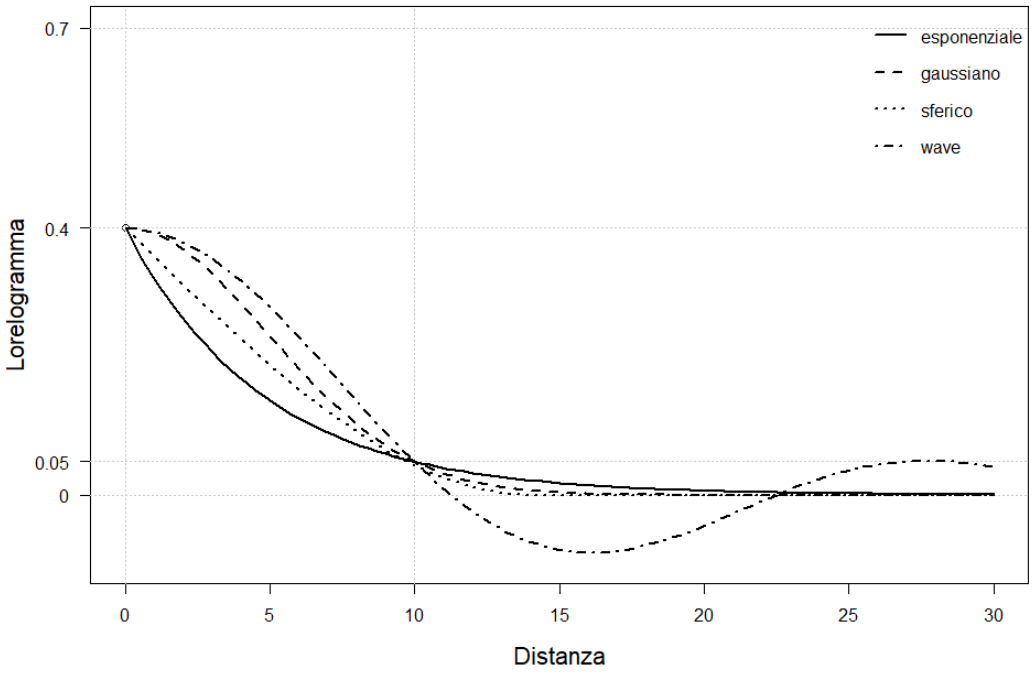

Figura 2.1: Esempi di modelli per il lorelogramma spaziale.

I modelli Esponenziale, Gaussiano e Wave presentano un *range* infinito, siccome il corrispondente lorelogramma converge a 0 solo asintoticamente, mentre il modello Sferico ha un *range* finito, pari alla distanza α_3. Tra i quattro considerati il modello Wave è l'unico modello non monotono, ed è anche l'unico che può assumere valori negativi. Anche il concetto di *practical range* viene preso in prestito dalla geostatistica, ma viene qui ridefinito come

la distanza oltre la quale il lorelogramma assume un valore inferiore alla soglia arbitraria 0.05, usata per indicare un'evidente mancanza di dipendenza spaziale. I lorelogrammi presentati in Figura 2.1 sono stati 'standardizzati', nel senso che tutti i modelli hanno lo stesso livello di dipendenza entro i cluster, lo stesso effetto *nugget* e lo stesso *practical range* di 10 unità.

2.5 Il lorelogramma spaziale empirico

Analogamente al variogramma empirico della letteratura geostatistica, il lorelogramma spaziale empirico è stato introdotto come uno stimatore iniziale del lorelogramma spaziale. Supponendo che la dipendenza spaziale diventi trascurabile oltre una certa distanza massima d_{max}, l'intervallo $[0, d_{max}]$ è stato suddiviso da Cattelan e Varin (2018) in $P + 1$ intervalli parzialmente sovrapposti. Nello specifico il primo intervallo include le coppie appartenenti allo stesso cluster, ovvero quelle a distanza 0, mentre i restanti P intervalli sono costruiti attorno a una sequenza di punti medi equispaziati $0 < m_1 < m_2 < \cdots < m_{P-1} < m_P < d_{max}$. Il p-esimo intervallo comprende quindi quelle coppie di osservazioni la cui distanza cade in un intervallo centrato in m_p e di raggio h.

Il lorelogramma spaziale empirico consiste perciò in una serie di stime $\hat{\gamma}_p$ del logaritmo dell'odds ratio a coppie $\gamma_p = \gamma(m_p)$, calcolate utilizzando tutte le coppie di osservazioni nel p-esimo intervallo. Il diagramma di dispersione del lorelogramma empirico, che stima quindi $\hat{\gamma}_p$ rispetto al punto medio m_p di ciascun intervallo, è utile per un'identificazione preliminare delle principali caratteristiche del lorelogramma spaziale, come la forma e il *range* della dipendenza spaziale, o la presenza di un effetto *nugget*. A livello pratico, la distanza d_{max} alla quale la dipendenza spaziale diventa trascurabile non è nota a priori, ma un'identificazione approssimata può essere basata sull'ispezione grafica del lorelogramma empirico. Cattelan e Varin (2018) raccomandano inoltre una certa attenzione nella scelta di d_{max} per evitare sovrainterpretazioni a lunghe distanze nell'andamento del lorelogramma empirico: gli intervalli oltre una certa distanza potrebbero contenere un numero

non sufficientemente elevato di coppie da stimarlo con una precisione accettabile. Poiché l'odds ratio a coppie è identificato dalla funzione di probabilità bivariata, allora uno stimatore $\hat{\gamma}_p$ è ottenibile attraverso la massimizzazione della verosimiglianza a coppie composta da tutte le coppie di osservazioni appartenenti al p-esimo intervallo.

$$l(\boldsymbol{\beta}, \gamma_p) = \sum_{(i,j)\in I_p} \log\{\Pr(Y_i = y_i, Y_j = y_j)\}, \tag{2.6}$$

dove il set I_p indica le coppie appartenenti all'intervallo corrispondente:

$$I_p = \{(i,j)\colon i < j \text{ e } \max(0, m_p - h) \leq d_{ij} \leq m_p + h\},$$

e la funzione di probabilità bivariata è

$$\begin{aligned}&\Pr(Y_i = y_i, Y_j = y_j) = \\ &\quad \pi_{ij}^{y_i y_j}(\pi_i - \pi_{ij})^{y_i(1-y_j)}(\pi_j - \pi_{ij})^{(1-y_i)y_j}(1 - \pi_i - \pi_j + \pi_{ij})^{(1-y_i)(1-y_j)},\end{aligned} \tag{2.7}$$

dove π_{ij} viene calcolata dall'equazione (2.4) con $\psi_{ij} = \exp(\gamma_p)$ per tutti gli i e j nel p-esimo intervallo. Poiché lo scopo è quello di utilizzare il lorelogramma per misurare la dipendenza spaziale residua in un modello di regressione logistica marginale, allora $\boldsymbol{\beta}$ viene sostituito nella verosimiglianza a coppie (2.6) con lo stimatore di massima verosimiglianza $\hat{\boldsymbol{\beta}}_{\text{ind}}$, ottenuto sotto l'assunzione di osservazioni indipendenti.

La Figura 2.2 mostra cinque lorelogrammi spaziali empirici ottenuti da altrettante simulazioni di un lorelogramma spaziale di tipo esponenziale senza effetto *nugget*. I dati simulati si è presunto potessero variare a livello spaziale su un quadrato unitario, inoltre i parametri del lorelogramma sono stati fissati a $\alpha_1 = 0$, $\alpha_2 = 1$ e $\alpha_3 = 0.07$. Il lorelogramma empirico ottenuto conta $P = 13$ intervalli e un raggio tale da produrre una sovrapposizione al 50% tra questi. In altri termini, ogni intervallo si estende dal punto medio del precedente (m_{p-1}) al punto medio del successivo (m_{p+1}).

Il lorelogramma empirico venne considerato anche da Bevilacqua *et al.* (2015) per descrivere l'associazione spaziale tra dati binari in un modello

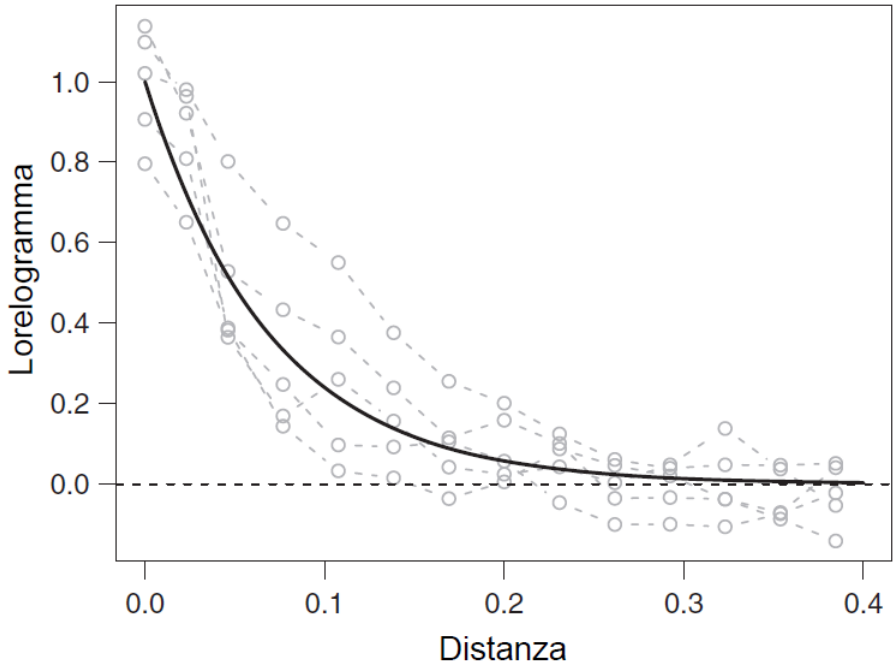

Figura 2.2: Lorelogramma spaziale empirico di cinque realizzazioni simulate da un lorelogramma spaziale con decadimento esponenziale della dipendenza.

probit con una funzione di correlazione esponenziale; si veda anche l'implementazione in **R** nel pacchetto `CompRandFld` (Padoan e Bevilacqua, 2015).

2.5.1 Calcolo del lorelogramma spaziale empirico

Le stime $\hat{\gamma}_p$ del lorelogramma spaziale empirico sono state ottenute con l'algoritmo di *Fisher scoring*. La (r+1)-esima iterazione dell'algoritmo è

$$\hat{\gamma}_p^{(r+1)} = \hat{\gamma}_p^{(r)} + E\left\{ -\frac{\partial^2 l(\hat{\boldsymbol{\beta}}, \hat{\gamma}_p^{(r)})}{\partial \gamma_p^2} \right\}^{-1} \frac{\partial l(\hat{\boldsymbol{\beta}}, \hat{\gamma}_p^{(r)})}{\partial \gamma_p}, \tag{2.8}$$

dove

$$\frac{\partial l(\boldsymbol{\beta}, \gamma)}{\partial \gamma_p} = \sum_{\{(i,j) \in I_p\}} \frac{\phi_{ij}}{\lambda_{ij}}$$

e

$$E\left\{ -\frac{\partial^2 l(\boldsymbol{\beta}, \gamma)}{\partial \gamma_p^2} \right\} = \sum_{\{(i,j) \in I_p\}} \frac{1}{\lambda_{ij}},$$

con

$$\phi_{ij} = \frac{y_i y_j}{\pi_{ij}} - \frac{y_i(1-y_j)}{\pi_i - \pi_{ij}} - \frac{(1-y_i)y_j}{\pi_j - \pi_{ij}} + \frac{(1-y_i)(1-y_j)}{1-\pi_i-\pi_j+\pi_{ij}}, \tag{2.9}$$

e

$$\lambda_{ij} = \frac{1}{\pi_{ij}} + \frac{1}{\pi_i - \pi_{ij}} + \frac{1}{\pi_j - \pi_{ij}} + \frac{1}{1-\pi_i-\pi_j+\pi_{ij}}. \tag{2.10}$$

In questo caso l'algoritmo è particolarmente efficiente, perché nella verosimiglianza a coppie (2.6) i parametri $\boldsymbol{\beta}$ e γ_p sono ortogonali secondo Cox e Reid (1987) poiché

$$E\left\{-\frac{\partial^2 l(\boldsymbol{\beta}, \gamma_p)}{\partial \boldsymbol{\beta}\, \partial \gamma_p}\right\} = \mathbf{0}. \tag{2.11}$$

Questa condizione di ortogonalità può essere verificata usando gli stessi argomenti di Palmgren (1989).

2.5.2 Minimi quadrati pesati

Allo stesso modo in cui il variogramma empirico viene usato nell'ambito della geostatistica per l'adattamento di modelli al variogramma, il lorelogramma empirico può essere utilizzato per l'adattamento di modelli parametrici al lorelogramma spaziale. Viene di seguito definito lo stimatore ai minimi quadrati pesati $\hat{\boldsymbol{\alpha}}_{\text{wls}}$ come l'insieme di valori che minimizza la somma pesata dei quadrati delle differenze tra il lorelogramma spaziale empirico $\hat{\gamma}_p$ ed i corrispondenti valori $\gamma(m_p; \boldsymbol{\alpha})$ attesi dal modello spaziale scelto,

$$\hat{\boldsymbol{\alpha}}_{\text{wls}} = \arg\min_{\boldsymbol{\alpha}} \sum_{p=0}^{P} n_p \{\hat{\gamma}_p - \gamma(m_p; \boldsymbol{\alpha})\}^2,$$

dove n_p è il numero di coppie appartenenti al p-esimo intervallo.

2.6 Stima di α e β con la verosimiglianza a coppie

Per la stima dei parametri del modello adattato al lorelogramma spaziale empirico, è necessario considerare non solo le coppie appartenenti allo stesso cluster, come nell'originale formulazione della verosimiglianza a coppie, ma anche le coppie appartenenti a cluster differenti. Studi precedenti hanno dimostrato che un'efficiente implementazione dell'inferenza basata sulla verosimiglianza a coppie nei processi spaziali e temporali coinvolge solo le coppie di osservazioni sufficientemente vicine tra loro da essere informative rispetto ai parametri di dipendenza $\boldsymbol{\alpha}$; si veda Varin *et al.* (2011). Pertanto, la log-verosimiglianza a coppie è stata costruita usando tutte le coppie di osservazioni divise da una distanza massima di d unità:

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{\{(i,j) \in S_d\}} \log\{\Pr(Y_i = y_i, Y_j = y_j)\}, \tag{2.12}$$

dove S_d indica tutte le coppie di osservazioni che distano d o meno unità,

$$S_d = \{(i,j) \colon i < j \text{ e } d_{ij} \leq d\}.$$

Il valore soglia d è stato selezionato con il *practical range* del modello adattato precedentemente al lorelogramma con il metodo dei minimi quadrati pesati, discusso nel paragrafo precedente. La probabilità bivariata $\Pr(Y_i = y_i, Y_j = y_j)$ viene calcolata attraverso l'espressione (2.7), questa volta con l'odds ratio a coppie espresso in termini del modello assunto per il lorelogramma spaziale: $\psi_{ij} = \exp\{\gamma(d_{ij}; \boldsymbol{\alpha})\}$.

In seguito all'identificazione del modello spaziale che meglio si adatta al lorelogramma empirico, è possibile passare alla fase di stima vera e propria, sia dei parametri $\boldsymbol{\alpha}$, responsabili della dipendenza spaziale, che dei parametri $\boldsymbol{\beta}$, associati alle esplicative considerate informative per la risposta binaria. La stima dei parametri può essere ottenuta attraverso diversi metodi, nei seguenti paragrafi vengono presentati il metodo a coppie ibrido (per una descrizione più dettagliata si veda Cattelan e Varin, 2018) e il metodo a

coppie.

2.6.1 Metodo a coppie ibrido

La verosimiglianza a coppie ibrida è un metodo iterativo che passa da equazioni di stima ottimali (2.2) per trovare delle stime di $\boldsymbol{\beta}$ dati $\boldsymbol{\alpha}$ a stime di massima verosimiglianza a coppie per i parametri $\boldsymbol{\alpha}$ dati $\boldsymbol{\beta}$. Il metodo della verosimiglianza a coppie ibrida originariamente era stato disegnato per dati provenienti da cluster indipendenti, ma in tali circostanze è più adeguato considerare la versione che contempla situazioni di dipendenza spaziale tra cluster.

Kuk (2007) derivò la distribuzione asintotica del primo ordine dello stimatore di verosimiglianza a coppie ibrida per dati raggruppati in cluster con il classico approccio di espandere la funzione punteggio attorno al vero valore dei parametri, sostituendo l'informazione osservata con quella attesa:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \end{pmatrix} \doteq \begin{pmatrix} \mathrm{D}^T\mathrm{V}^{-1}\mathrm{D} & \mathbf{0} \\ E\big(-\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\alpha}^T}\big) & E\big(-\frac{\partial^2 l}{\partial \boldsymbol{\alpha}\, \partial \boldsymbol{\alpha}^T}\big) \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{D}^T\mathrm{V}^{-1}(\mathbf{y} - \boldsymbol{\pi}) \\ \frac{\partial l(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} \end{pmatrix}. \quad (2.13)$$

L'applicazione di questo metodo ibrido al modello di dipendenza spaziale è molto attraente a causa del risultato di ortogonalità (2.11) descritto nel paragrafo 2.5.1, il quale implica

$$E\left\{ -\frac{\partial^2 l(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\alpha}^T} \right\} = \mathbf{0}.$$

Pertanto, la matrice di informazione nell'equazione (2.13) è blocco-diagonale, il che significa che $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ sono ortogonali e le distribuzioni dei loro stimatori sono asintoticamente indipendenti.

Diversamente dagli ambiti classici, in contesto spaziale i cluster non sono tra loro indipendenti e sono perciò necessarie ulteriori assunzioni per garantire la consistenza e la normalità asintotica degli stimatori di massima verosimiglianza a coppie ibrida. In Lin e Clayton (2005) e Lin (2008) so-

no contenuti alcuni risultati interessanti. Questi autori hanno dimostrato che il teorema del limite centrale rimane valido per gli stimatori di quasi-verosimiglianza di un modello di regressione logistica spaziale sotto adeguate condizioni di regolarità, la più importante delle quali implica che la correlazione decresca esponenzialmente con la distanza. Poiché i modelli considerati per il lorelogramma decadono esponenzialmente con la distanza e presentano una corrispondenza biunivoca con la funzione di covarianza, si può concludere che l'assunzione di decadimento esponenziale della correlazione è valida in questo contesto. Anche la consistenza degli stimatori di massima verosimiglianza a coppie dei parametri di dipendenza $\boldsymbol{\alpha}$ dipende da un decadimento esponenziale della funzione di correlazione (Heagerty e Lele, 1998) e dall'uso di un'adeguata soglia per i pesi (Bevilacqua e Gaetan, 2015) nella specificazione della verosimiglianza a coppie (2.12), dove sono state considerate solamente le coppie di osservazioni fino a una certa distanza d.

Assunto un modello per il lorelogramma spaziale, l'espansione della funzione punteggio (2.13) mostra che $\hat{\boldsymbol{\beta}}$ ha la stessa varianza asintotica che si sarebbe ottenuta risolvendo le equazioni di stima ottimali (2.2) con degli $\boldsymbol{\alpha}$ noti:

$$\text{asymvar}(\hat{\boldsymbol{\beta}}) = (\mathrm{D}^T \mathrm{V}^{-1} \mathrm{D})^{-1}. \tag{2.14}$$

Il metodo di verosimiglianza a coppie ibrida di Kuk (2007) itera tra stime di $\boldsymbol{\beta}$ dati $\boldsymbol{\alpha}$ e stime di $\boldsymbol{\alpha}$ dati $\boldsymbol{\beta}$ fino a convergenza. Tuttavia, se l'algoritmo viene inizializzato con uno stimatore consistente di $\boldsymbol{\beta}$, allora lo stimatore ottenuto dopo un singolo ciclo della procedura di stima ibrida ha la stessa distribuzione asintotica dello stimatore che si otterrebbe dalla procedura iterativa completa (Lehmann (1983), pag. 422). Nel modello di regressione logistica marginale considerato in (2.1), lo stimatore consistente più naturale di $\boldsymbol{\beta}$ è quello che si ottiene dalla massimizzazione della verosimiglianza di indipendenza, ovvero la verosimiglianza calcolata sotto l'assunzione di osservazioni indipendenti.

2.6.2 Metodo a coppie

Analogamente alla verosimiglianza a coppie ibrida, anche il presente è un metodo iterativo, ma i due approcci presentano sostanziali differenze. Con questo metodo infatti la stima di $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ avviene simultaneamente, massimizzando numericamente la verosimiglianza a coppie. In particolare la massimizzazione numerica della verosimiglianza a coppie avviene attraverso la funzione `nlminb` (R Core Team, 2018), che si basa su `PORT Mathematical Subroutine Library` ed è più stabile, robusta ed affidabile di altre funzioni come `optim`. Essendo un algoritmo iterativo il costo computazionale è decisamente superiore rispetto al metodo ibrido con un solo ciclo riportato in Cattelan e Varin (2018), ed aumenta esponenzialmente all'aumentare del numero di coefficienti $\boldsymbol{\beta}$ relativi alle variabili esplicative. Altre quantità che richiedono uno sforzo computazionale non indifferente sono gli errori standard delle stime. Ricordando che la seconda identità di Bartlett non regge in situazioni come quella qui considerata, la stima della matrice di sensibilità $H(\boldsymbol{\theta})$ avviene empiricamente come descritto nel paragrafo 1.7, mentre la stima della matrice di variabilità $J(\boldsymbol{\theta})$ risulta più problematica. Dalla prima identità di Bartlett infatti deriva la non distorsione dello stimatore di massima verosimiglianza composita, ovvero $U_{\mathcal{C}}(\hat{\boldsymbol{\theta}}_{\mathcal{C}}) = \mathbf{0}$, pertanto non è possibile stimare $J(\boldsymbol{\theta})$ usando la derivata della verosimiglianza composita calcolata in $\hat{\boldsymbol{\theta}}_{\mathcal{C}}$.

Al fine di ottenere una matrice di variabilità invertibile ed utilizzabile per il calcolo della matrice di informazione di Godambe, Heagerty e Lele (1998) suggerirono di ottenere delle stime dello score composito su B sotto-regioni di ampiezza S_b dello spazio considerato, dalle quali si ottiene

$$\hat{J}(\boldsymbol{\theta}_{\mathcal{C}}) = \frac{1}{B} \sum_{b=1}^{B} S_b U_{\mathcal{C}_b}(\hat{\boldsymbol{\theta}}_{\mathcal{C}}) U_{\mathcal{C}_b}(\hat{\boldsymbol{\theta}}_{\mathcal{C}})^T.$$

Le B sotto-regioni, o blocchi, sono intese come zone tra loro indipendenti. Per quanto riguarda il calcolo di $U_{\mathcal{C}_b}(\hat{\boldsymbol{\theta}}_{\mathcal{C}})$, ovvero del vettore di score di verosimiglianza a coppie, limitato alle coppie di osservazioni appartenenti al blocco b, la sua determinazione è avvenuta attraverso un algoritmo di derivazione

numerica contenuto nella funzione `grad`, all'interno della libreria `numDeriv` (Gilbert e Varadhan, 2019). Per determinare la matrice $\hat{J}(\boldsymbol{\theta}_{\mathcal{C}})$ è necessario quindi calcolare numericamente la funzione punteggio separatamente per ogni blocco b, perciò ci si aspetta che il costo computazionale aumenti all'aumentare del numero di blocchi in cui si divide la regione spaziale considerata. La scelta di derivazione degli $U_{\mathcal{C}_b}(\hat{\boldsymbol{\theta}}_{\mathcal{C}})$ attraverso un algoritmo numerico è stata presa rispetto all'alternativa di calcolare le derivate in maniera analitica. Il calcolo numerico presenta purtroppo lo svantaggio di non essere sempre affidabile, la derivazione numerica infatti non è altro che un'approssimazione della derivazione analitica, e ciò si traduce nella possibilità di ottenere comunque delle matrici $\hat{J}(\boldsymbol{\theta}_{\mathcal{C}})$ non invertibili o delle matrici $\hat{H}(\hat{\boldsymbol{\theta}}_{\mathcal{C}})$ non a rango pieno. Questi aspetti di instabilità numerica nella fase di stima della verosimiglianza a coppie verranno ripresi nel prossimo capitolo, che sarà basato su studi di simulazione di questo approccio di verosimiglianza.

Riassumendo, dopo aver assunto un modello per il lorelogramma spaziale, il metodo di stima della verosimiglianza a coppie si articola nelle seguenti fasi:

(a) calcolo di $\hat{\boldsymbol{\beta}}_{ind}$ ignorando la struttura di dipendenza spaziale;

(b) calcolo di $\hat{\boldsymbol{\theta}}_{\mathcal{C}}$ massimizzando numericamente la verosimiglianza a coppie con `nlminb` e fornendo come punti iniziali $(\hat{\boldsymbol{\alpha}}_{\text{wls}}, \hat{\boldsymbol{\beta}}_{ind})$;

(c) suddivisione della regione spaziale in B blocchi di ampiezza S_b e calcolo di $U_{\mathcal{C}_b}(\hat{\boldsymbol{\theta}}_{\mathcal{C}})$ per $b = 1, \ldots, B$;

(d) calcolo di $\hat{J}(\boldsymbol{\theta}_{\mathcal{C}})$ come media pesata delle varie $\hat{J}_b(\boldsymbol{\theta}_{\mathcal{C}}) = S_b U_{\mathcal{C}_b}(\hat{\boldsymbol{\theta}}_{\mathcal{C}}) U_{\mathcal{C}_b}(\hat{\boldsymbol{\theta}}_{\mathcal{C}})^T$ e di $\hat{H}(\hat{\boldsymbol{\theta}}_{\mathcal{C}})$;

(e) calcolo degli errori standard di $\hat{\boldsymbol{\theta}}_{\mathcal{C}}$ come $se(\hat{\boldsymbol{\theta}}_{\mathcal{C}}) = \sqrt{G(\hat{\boldsymbol{\theta}}_{\mathcal{C}})^{-1}_{[i,i]}}$.

Oltre alle differenze nella fase di stima dei parametri, i due metodi di verosimiglianza riportati differiscono anche nel modo in cui considerano i parametri responsabili della dipendenza $\boldsymbol{\alpha}$: nel metodo a coppie sono considerati come parametri di interesse, dai quali si cerca di dedurre la struttura di dipendenza spaziale presente nei dati per trarne delle conclusioni attraverso procedure inferenziali; nel metodo ibrido invece sono trattati come dei veri e propri parametri di disturbo, a cui il ricercatore non è interessato poiché la sua attenzione è focalizzata completamente sui parametri di regressione $\boldsymbol{\beta}$. Da ciò si evince che i due approcci presentati non sono direttamente confrontabili e nelle applicazioni a dati binari che presentano una struttura di dipendenza spaziale un metodo viene scelto rispetto all'altro solamente sulla base dell'obiettivo dello studio.

Capitolo 3

Studi di simulazione

In questo capitolo si cerca di valutare, attraverso due studi di simulazione, il comportamento e le proprietà delle stime dei parametri relativi alla dipendenza ($\boldsymbol{\alpha}$) e alle variabili esplicative considerate informative ($\boldsymbol{\beta}$). Il modello viene stimato con il metodo della massima verosimiglianza a coppie. I dati binari sono simulati dal modello scelto per il lorelogramma spaziale, utilizzando l'algoritmo descritto in Emrich e Piedmonte (1991) che permette la simulazione di variabili binarie multivariate ad alta dimensionalità con determinate probabilità marginali univariate π_i e bivariate π_{ij}. Più specificamente, l'algoritmo di simulazione è costituito dai seguenti passi:

(a) disegnare le posizioni dei k cluster e fissare il numero di osservazioni per cluster;

(b) simulare le variabili esplicative, calcolare le probabilità marginali π_i e gli odds ratio a coppie ψ_{ij};

(c) derivare le probabilità bivariate π_{ij} date le marginali e gli odds ratio a coppie con la formula (2.4);

(d) calcolare la correlazione a coppie come $\text{Corr}(Y_i, Y_j) = \frac{\pi_{ij} - \pi_i \pi_j}{\sqrt{\pi_i \pi_j (1-\pi_i)(1-\pi_j)}}$;

(e) simulare i dati binari utilizzando l'algoritmo di Emrich e Piedmonte (1991).

L'algoritmo di Emrich e Piedmonte permette quindi di generare osservazioni binarie con una correlazione data, ma i tempi di esecuzione sono piuttosto lunghi a causa della stima di una matrice di correlazione tetracorica Σ. Questa matrice contiene le stime delle correlazioni di tutte le coppie di osservazioni binarie se queste fossero misurate su una scala continua. Condizione necessaria per passare alla fase successiva dell'algoritmo è che la matrice Σ sia sempre semidefinita positiva, condizione purtroppo non sempre garantita. Con lo scopo di calcolare la matrice definita positiva più vicina rispetto a quella cercata, la funzione R `nearPD` nel pacchetto `Matrix` (Bates e Maechler, 2018) viene impiegata per permettere all'algoritmo di passare alla fase seguente e poter quindi generare i dati desiderati.

Le simulazioni sono effettuate partendo dal modello di regressione logistica

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{3.1}$$

dove x_1 è un trend lineare spaziale tra i cluster lungo la direzione est e x_2 è una variabile soggetto-specifica simulata dalla distribuzione di una normale standard. Il trend lineare spaziale è calcolato come l'ascissa riscalata dei cluster, ovvero l'ascissa di ciascun cluster divisa per il valore massimo di ascissa osservato nei dati generati. I valori dei coefficienti di regressione sono stati fissati pari a $\beta_0 = -0.5$, $\beta_1 = 1.5$ e $\beta_2 = 1$.

In questi studi sono state disegnate due griglie di 144 cluster ciascuna su un quadrato di superficie $[0, 1] \times [0, 1]$. Come presentato in Figura 3.1, nel primo studio i cluster sono stati disposti su una griglia regolare 12×12 dove i punti sono equi-spaziati lungo le coordinate X e Y, mentre nel secondo è stata utilizzata una griglia con un incremento di 0.01 lungo entrambi gli assi sul quadrato unitario; i punti sono stati perturbati tramite l'aggiunta di un valore proveniente dalla variabile casuale uniforme $U[-0.005, 0.005]$ a ogni coordinata ed è stato campionato casualmente un sottoinsieme di 144 punti. I punti della griglia rappresentano le coordinate dei cluster e tutte le osservazioni appartenenti allo stesso cluster hanno le stesse coordinate.

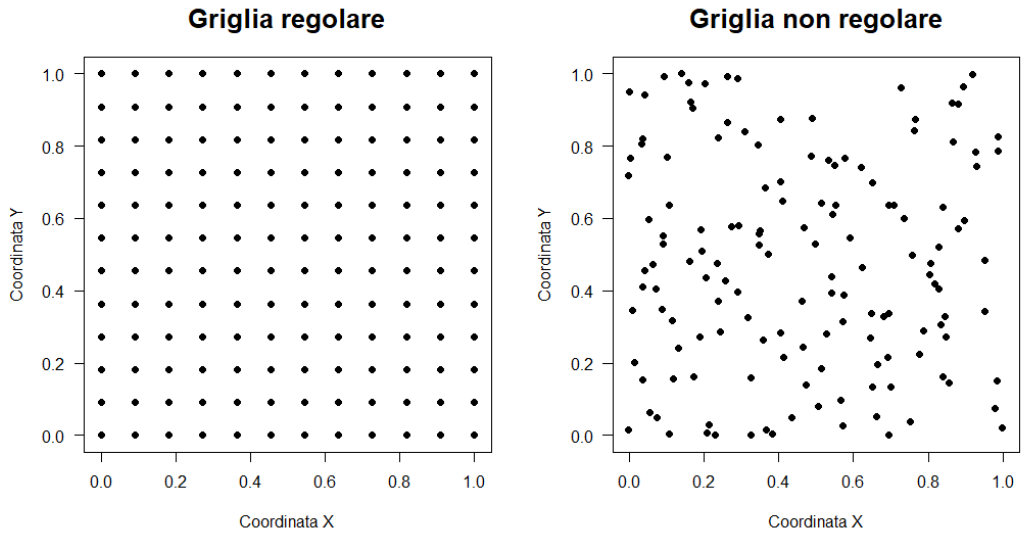

Figura 3.1: Griglie di cluster considerate negli studi di simulazione.

Sono stati esaminati cluster di dimensione $n_k = 5, 10, 15$, e si è considerato un modello per il lorelogramma spaziale con decadimento della dipendenza di tipo esponenziale senza effetto *nugget* ($\alpha_1 = 0$). Il parametro di *sill* è stato fissato a $\alpha_2 = 1$, inoltre per analizzare differenti livelli di dipendenza spaziale sono stati assunti valori diversi per il parametro di *range* α_3, ovvero 0.05 e 0.07, che corrispondono rispettivamente a *practical range* di 0.15 e 0.21 circa. Al fine di valutare il comportamento delle stime e dei relativi errori standard dei parametri $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$, il numero di cluster è stato scelto in modo tale da permettere suddivisioni della griglia per varie scelte di B; nel caso qui considerato la regione spaziale è quindi divisibile in $B = 4, 9, 16, 36$ blocchi.

Riassumendo, le simulazioni sono state calcolate per quarantotto diverse configurazioni ottenute variando:

(a) la struttura spaziale dei dati considerata (regolare, non regolare);

(b) il numero di osservazioni per cluster n_k (5, 10, 15);

(c) i valori di α_3 (0.05, 0.07);

(d) il numero di blocchi in cui dividere la griglia (4, 9, 16, 36).

Queste configurazioni sono state scelte in modo da riflettere situazioni in cui la quantità di informazione e la misura di dipendenza spaziale presente nei dati siano differenti, inoltre a parità di n_k, α_3 e B, è di particolare interesse valutare il comportamento delle stime e degli errori standard passando da una struttura di cluster a griglia regolare ad una struttura a griglia libera.

Per ogni configurazione sono state effettuate 1000 simulazioni e i risultati relativi ai parametri di dipendenza $\boldsymbol{\alpha} = (\alpha_2, \alpha_3)$ e di regressione $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ sono riassunti di seguito nelle Tabelle 3.1 - 3.6 in termini di distorsione, deviazione standard delle stime, errori standard derivati dalla matrice di informazione di Godambe e copertura nominale al 95% degli intervalli di confidenza alla Wald.

GRIGLIA		REGOLARE					NON REGOLARE				
α_3	B	Bias	SD	SE	SE MED	CI	Bias	SD	SE	SE MED	CI
					$\hat{\alpha}_2$						
0.05	4	0.029	0.217	0.165	0.159	80.7	0.007	0.275	0.202	0.186	83.8
	9	0.029	0.217	0.210	0.208	92.6	0.007	0.275	0.242	0.231	92.3
	16	0.029	0.217	0.231	0.228	94.9	0.007	0.275	0.289	0.275	96.7
	36	0.029	0.217	0.278	0.279	97.8	0.007	0.275	0.363	0.347	98.5
0.07	4	0.033	0.213	0.182	0.174	83.9	0.007	0.272	0.213	0.200	84.2
	9	0.033	0.213	0.235	0.227	94.4	0.007	0.272	0.259	0.255	92.5
	16	0.033	0.213	0.275	0.272	96.9	0.007	0.272	0.332	0.322	97.7
	36	0.033	0.213	0.417	0.415	99.7	0.007	0.272	0.454	0.449	99.0
					$\hat{\alpha}_3$						
0.05	4	0.006	0.057	0.017	0.014	85.7	0.003	0.094	0.015	0.012	70.3
	9	0.006	0.057	0.019	0.017	95.1	0.003	0.094	0.017	0.014	78.2
	16	0.006	0.057	0.021	0.018	96.6	0.003	0.094	0.018	0.014	81.3
	36	0.006	0.057	0.022	0.018	97.9	0.003	0.094	0.019	0.015	83.7
0.07	4	0.013	0.039	0.017	0.014	69.4	0.006	0.089	0.016	0.013	58.9
	9	0.013	0.039	0.020	0.017	80.1	0.006	0.089	0.019	0.015	65.6
	16	0.013	0.039	0.022	0.019	84.4	0.006	0.089	0.020	0.017	69.7
	36	0.013	0.039	0.026	0.021	95.6	0.006	0.089	0.022	0.018	75.9

Tabella 3.1: Distorsione, deviazione standard, errori standard medi e mediani e copertura nominale al 95% dei parametri di dipendenza ottenute con il metodo della verosimiglianza a coppie, 5 osservazioni per cluster.

GRIGLIA		REGOLARE					NON REGOLARE				
α_3	B	Bias	SD	SE	SE MED	CI	Bias	SD	SE	SE MED	CI
					$\hat{\beta}_0$						
0.05	4	0.007	0.152	0.088	0.083	68.1	0.003	0.172	0.100	0.095	69.7
	9	0.007	0.152	0.110	0.108	81.9	0.003	0.172	0.116	0.112	78.9
	16	0.007	0.152	0.107	0.107	82.0	0.003	0.172	0.104	0.103	74.9
	36	0.007	0.152	0.084	0.084	72.6	0.003	0.172	0.083	0.083	65.0
0.07	4	0.010	0.185	0.100	0.093	66.2	0.004	0.209	0.107	0.100	64.9
	9	0.010	0.185	0.109	0.107	73.0	0.004	0.209	0.117	0.115	71.7
	16	0.010	0.185	0.093	0.093	67.9	0.004	0.209	0.097	0.096	63.0
	36	0.010	0.185	0.059	0.058	46.3	0.004	0.209	0.066	0.066	47.4
					$\hat{\beta}_1$						
0.05	4	0.022	0.482	0.317	0.301	74.5	0.023	0.576	0.372	0.355	75.5
	9	0.022	0.482	0.375	0.371	84.2	0.023	0.576	0.431	0.422	85.2
	16	0.022	0.482	0.369	0.365	85.9	0.023	0.576	0.391	0.384	81.1
	36	0.022	0.482	0.293	0.291	77.2	0.023	0.576	0.318	0.318	73.0
0.07	4	0.026	0.581	0.377	0.364	74.2	0.031	0.698	0.426	0.405	73.1
	9	0.026	0.581	0.390	0.381	77.7	0.031	0.698	0.456	0.445	79.0
	16	0.026	0.581	0.336	0.331	73.1	0.031	0.698	0.378	0.368	70.5
	36	0.026	0.581	0.213	0.210	51.3	0.031	0.698	0.264	0.261	55.4
					$\hat{\beta}_2$						
0.05	4	0.020	0.108	0.082	0.078	80.8	0.021	0.112	0.080	0.079	77.6
	9	0.020	0.108	0.087	0.086	86.5	0.021	0.112	0.091	0.089	85.8
	16	0.020	0.108	0.087	0.086	86.9	0.021	0.112	0.082	0.081	83.6
	36	0.020	0.108	0.074	0.074	82.0	0.021	0.112	0.069	0.069	75.9
0.07	4	0.024	0.112	0.082	0.077	79.3	0.026	0.116	0.078	0.077	77.3
	9	0.024	0.112	0.079	0.078	80.4	0.026	0.116	0.084	0.082	81.5
	16	0.024	0.112	0.072	0.071	77.3	0.026	0.116	0.072	0.071	75.5
	36	0.024	0.112	0.050	0.049	60.6	0.026	0.116	0.053	0.053	61.3

Tabella 3.2: Distorsione, deviazione standard, errori standard medi e mediani e copertura nominale al 95% dei parametri di regressione ottenute con il metodo della verosimiglianza a coppie, 5 osservazioni per cluster.

GRIGLIA		REGOLARE					NON REGOLARE				
α_3	B	Bias	SD	SE	SE MED	CI	Bias	SD	SE	SE MED	CI
					$\hat{\alpha}_2$						
0.05	4	0.026	0.149	0.119	0.114	81.9	0.023	0.164	0.144	0.137	82.9
	9	0.026	0.149	0.147	0.145	92.4	0.023	0.164	0.174	0.167	92.8
	16	0.026	0.149	0.162	0.162	94.7	0.023	0.164	0.202	0.199	96.0
	36	0.026	0.149	0.196	0.196	98.5	0.023	0.164	0.255	0.248	98.1
0.07	4	0.030	0.148	0.130	0.126	83.4	0.019	0.181	0.162	0.154	83.9
	9	0.030	0.148	0.168	0.165	93.7	0.019	0.181	0.200	0.192	92.7
	16	0.030	0.148	0.198	0.198	97.0	0.019	0.181	0.245	0.236	96.9
	36	0.030	0.148	0.297	0.295	99.8	0.019	0.181	0.340	0.327	99.4
					$\hat{\alpha}_3$						
0.05	4	0.006	0.035	0.013	0.012	85.2	0.006	0.016	0.012	0.010	69.1
	9	0.006	0.035	0.015	0.013	93.6	0.006	0.016	0.013	0.011	78.2
	16	0.006	0.035	0.016	0.014	96.1	0.006	0.016	0.013	0.011	79.8
	36	0.006	0.035	0.016	0.014	96.8	0.006	0.016	0.014	0.012	81.8
0.07	4	0.012	0.019	0.013	0.011	63.1	0.013	0.020	0.014	0.011	56.8
	9	0.012	0.019	0.015	0.013	72.0	0.013	0.020	0.015	0.013	63.1
	16	0.012	0.019	0.016	0.015	77.0	0.013	0.020	0.015	0.014	65.5
	36	0.012	0.019	0.017	0.016	88.5	0.013	0.020	0.016	0.014	68.9

Tabella 3.3: Distorsione, deviazione standard, errori standard medi e mediani e copertura nominale al 95% dei parametri di dipendenza ottenute con il metodo della verosimiglianza a coppie, 10 osservazioni per cluster.

GRIGLIA		REGOLARE					NON REGOLARE				
α_3	B	Bias	SD	SE	SE MED	CI	Bias	SD	SE	SE MED	CI
					$\hat{\beta}_0$						
0.05	4	0.005	0.139	0.085	0.079	71.2	0.006	0.157	0.092	0.085	68.6
	9	0.005	0.139	0.099	0.097	80.3	0.006	0.157	0.108	0.107	78.8
	16	0.005	0.139	0.095	0.094	81.4	0.006	0.157	0.096	0.095	72.0
	36	0.005	0.139	0.075	0.075	71.1	0.006	0.157	0.077	0.077	62.2
0.07	4	0.006	0.175	0.096	0.089	66.1	0.006	0.195	0.102	0.095	63.6
	9	0.006	0.175	0.100	0.099	72.0	0.006	0.195	0.111	0.110	70.0
	16	0.006	0.175	0.085	0.085	64.5	0.006	0.195	0.091	0.090	62.8
	36	0.006	0.175	0.053	0.053	44.7	0.006	0.195	0.061	0.061	46.1
					$\hat{\beta}_1$						
0.05	4	0.005	0.421	0.291	0.275	74.8	0.022	0.520	0.334	0.316	73.1
	9	0.005	0.421	0.331	0.328	84.0	0.022	0.520	0.398	0.389	82.4
	16	0.005	0.421	0.324	0.321	84.7	0.022	0.520	0.351	0.348	80.8
	36	0.005	0.421	0.262	0.260	77.3	0.022	0.520	0.292	0.289	75.1
0.07	4	0.013	0.515	0.344	0.327	73.0	0.026	0.637	0.397	0.371	73.1
	9	0.013	0.515	0.355	0.352	77.2	0.026	0.637	0.431	0.419	79.5
	16	0.013	0.515	0.305	0.305	73.6	0.026	0.637	0.349	0.343	72.4
	36	0.013	0.515	0.193	0.192	52.6	0.026	0.637	0.244	0.240	54.0
					$\hat{\beta}_2$						
0.05	4	0.014	0.081	0.059	0.057	77.3	0.018	0.080	0.059	0.057	79.6
	9	0.014	0.081	0.063	0.063	84.7	0.018	0.080	0.062	0.061	83.5
	16	0.014	0.081	0.061	0.061	85.5	0.018	0.080	0.058	0.057	81.6
	36	0.014	0.081	0.052	0.051	78.7	0.018	0.080	0.048	0.047	74.4
0.07	4	0.019	0.083	0.058	0.056	74.9	0.021	0.083	0.058	0.055	76.9
	9	0.019	0.083	0.057	0.057	77.8	0.021	0.083	0.058	0.057	80.2
	16	0.019	0.083	0.050	0.050	73.7	0.021	0.083	0.051	0.050	76.3
	36	0.019	0.083	0.035	0.035	58.7	0.021	0.083	0.037	0.037	63.5

Tabella 3.4: Distorsione, deviazione standard, errori standard medi e mediani e copertura nominale al 95% dei parametri di regressione ottenute con il metodo della verosimiglianza a coppie, 10 osservazioni per cluster.

GRIGLIA		REGOLARE					NON REGOLARE				
α_3	B	Bias	SD	SE	SE MED	CI	Bias	SD	SE	SE MED	CI
					$\hat{\alpha}_2$						
0.05	4	0.017	0.145	0.104	0.097	79.5	0.019	0.150	0.124	0.120	82.9
	9	0.017	0.145	0.127	0.125	90.9	0.019	0.150	0.150	0.147	90.9
	16	0.017	0.145	0.142	0.140	94.5	0.019	0.150	0.176	0.171	96.6
	36	0.017	0.145	0.170	0.169	97.3	0.019	0.150	0.219	0.215	98.4
0.07	4	0.026	0.134	0.114	0.106	81.3	0.020	0.165	0.141	0.133	82.2
	9	0.026	0.134	0.146	0.141	92.5	0.020	0.165	0.177	0.169	92.0
	16	0.026	0.134	0.173	0.170	96.6	0.020	0.165	0.217	0.210	97.4
	36	0.026	0.134	0.259	0.254	99.4	0.020	0.165	0.297	0.290	99.0
					$\hat{\alpha}_3$						
0.05	4	0.005	0.046	0.012	0.010	83.3	0.006	0.014	0.011	0.009	69.3
	9	0.005	0.046	0.013	0.012	90.4	0.006	0.014	0.012	0.010	76.7
	16	0.005	0.046	0.014	0.012	93.9	0.006	0.014	0.012	0.011	79.3
	36	0.005	0.046	0.014	0.012	94.6	0.006	0.014	0.012	0.012	83.7
0.07	4	0.012	0.016	0.012	0.010	62.2	0.013	0.018	0.012	0.010	55.8
	9	0.012	0.016	0.013	0.012	70.5	0.013	0.018	0.014	0.013	63.7
	16	0.012	0.016	0.014	0.014	76.1	0.013	0.018	0.014	0.013	66.8
	36	0.012	0.016	0.015	0.014	83.4	0.013	0.018	0.015	0.014	74.5

Tabella 3.5: Distorsione, deviazione standard, errori standard medi e mediani e copertura nominale al 95% dei parametri di dipendenza ottenute con il metodo della verosimiglianza a coppie, 15 osservazioni per cluster.

GRIGLIA		REGOLARE					NON REGOLARE				
α_3	B	Bias	SD	SE	SE MED	CI	Bias	SD	SE	SE MED	CI
					$\hat{\beta}_0$						
0.05	4	0.003	0.139	0.082	0.077	67.8	4e-04	0.156	0.089	0.085	67.6
	9	0.003	0.139	0.095	0.094	80.2	4e-04	0.156	0.104	0.103	77.0
	16	0.003	0.139	0.092	0.091	79.6	4e-04	0.156	0.095	0.096	73.9
	36	0.003	0.139	0.072	0.072	67.5	4e-04	0.156	0.074	0.075	60.5
0.07	4	0.005	0.179	0.095	0.088	65.1	0.001	0.196	0.100	0.095	63.5
	9	0.005	0.179	0.097	0.096	69.1	0.001	0.196	0.109	0.108	70.0
	16	0.005	0.179	0.083	0.082	61.4	0.001	0.196	0.090	0.091	62.2
	36	0.005	0.179	0.051	0.051	42.4	0.001	0.196	0.060	0.059	41.2
					$\hat{\beta}_1$						
0.05	4	0.013	0.426	0.283	0.267	73.7	0.006	0.514	0.327	0.312	71.8
	9	0.013	0.426	0.320	0.308	83.1	0.006	0.514	0.389	0.380	83.7
	16	0.013	0.426	0.315	0.313	84.6	0.006	0.514	0.349	0.340	81.2
	36	0.013	0.426	0.255	0.254	74.4	0.006	0.514	0.284	0.275	72.1
0.07	4	0.013	0.527	0.340	0.325	73.2	0.010	0.634	0.388	0.365	72.3
	9	0.013	0.527	0.343	0.334	77.2	0.010	0.634	0.426	0.412	80.1
	16	0.013	0.527	0.298	0.295	72.3	0.010	0.634	0.348	0.342	72.4
	36	0.013	0.527	0.188	0.186	52.0	0.010	0.634	0.239	0.233	53.9
					$\hat{\beta}_2$						
0.05	4	0.011	0.064	0.049	0.046	79.4	0.012	0.068	0.049	0.047	76.0
	9	0.011	0.064	0.053	0.053	87.3	0.012	0.068	0.053	0.052	86.0
	16	0.011	0.064	0.052	0.052	88.3	0.012	0.068	0.048	0.048	82.1
	36	0.011	0.064	0.043	0.043	82.2	0.012	0.068	0.041	0.040	78.3
0.07	4	0.014	0.067	0.048	0.045	77.8	0.017	0.071	0.049	0.047	76.5
	9	0.014	0.067	0.048	0.048	82.0	0.017	0.071	0.049	0.049	80.9
	16	0.014	0.067	0.043	0.042	78.2	0.017	0.071	0.042	0.042	73.4
	36	0.014	0.067	0.029	0.029	63.1	0.017	0.071	0.032	0.032	56.9

Tabella 3.6: Distorsione, deviazione standard, errori standard medi e mediani e copertura nominale al 95% dei parametri di regressione ottenute con il metodo della verosimiglianza a coppie, 15 osservazioni per cluster.

Come visibile nelle precedenti tabelle, dalle colonne di distorsione e deviazione standard dei parametri $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ si deduce che le stime ottenute per qualsiasi combinazione di α_3 e n_k sono non distorte e la loro variabilità risulta leggermente superiore nello studio di simulazione su griglia non regolare. Inoltre, per entrambe le griglie e a parità di n_k, passando da $\alpha_3 = 0.05$ a $\alpha_3 = 0.07$ la variabilità delle stime diminuisce, ma aumenta lievemente la distorsione. Per i parametri $\boldsymbol{\alpha}$ si può notare che gli errori standard presentano un andamento monotono crescente all'aumentare del numero di blocchi B. Se confrontati con le rispettive deviazioni standard delle stime, per α_2 si passa da una situazione di sottostima ad una dove gli errori standard sovrastimano la vera variabilità delle stime, mentre per α_3 non si arriva mai a sovrastimare la variabilità effettiva, anche per elevate numerosità di blocchi. Ciò ha delle ripercussioni sulla copertura degli intervalli di confidenza alla Wald, che, seguendo lo stesso andamento degli errori standard, passano da una copertura inferiore al livello nominale del 95% ad una superiore. Per i parametri di regressione $\boldsymbol{\beta}$ si è osservato invece un comportamento degli errori standard differente da quello riscontrato per gli $\boldsymbol{\alpha}$: all'aumentare del numero di blocchi gli errori standard inizialmente presentano un andamento crescente, ma oltrepassato un certo valore di B, questo subisce un'inversione e gli errori stimati dalla matrice di informazione di Godambe tendono a ridursi. Comparando questi valori con le deviazioni standard corrispondenti, si può notare che i primi risultano sistematicamente inferiori rispetto alla variabilità effettiva delle stime $\hat{\boldsymbol{\beta}}$, ciò significa che gli errori standard vengono sempre sottostimati e il livello di copertura degli intervalli non è garantito. All'aumentare del numero di osservazioni per cluster n_k, non è stato rilevato alcun miglioramento significativo nella copertura degli intervalli, per $\alpha_3 = 0.05$ è invece evidente che la copertura è nettamente superiore a quella che si ottiene per $\alpha_3 = 0.07$. La stessa differenza è emersa anche per il parametro α_3. Infine, un risultato particolarmente interessante e di diretto interesse per le applicazioni svolte è il confronto delle coperture degli intervalli nei due studi di simulazione compiuti: a parità di n_k, α_3 e B, è possibile osservare che nella maggior parte dei casi i valori per la griglia regolare sono sistematicamente, ma in misura lieve, più alti di quelli osservati per lo studio su griglia non regolare.

Capitolo 4

Applicazione

Per l'applicazione del metodo della verosimiglianza a coppie descritto nei precedenti capitoli sono stati utilizzati due insieme di dati, il primo proviene da un'indagine sulla malaria a livello infantile in Gambia ampiamente descritta in Thomson *et al.* (1999), mentre il secondo tratta il caso *loa-loa* discusso in Diggle *et al.* (2007b).

4.1 Dataset Gambia

4.1.1 Introduzione ai dati

Il dataset considerato in questa applicazione contiene informazioni su 2035 bambini provenienti da 65 villaggi, il numero di bambini campionati per villaggio varia da 8 a 63, con una media di 31.3 e una mediana di 30 bambini. La variabile risposta è l'indicatore binario della presenza di parassiti della malaria nel sangue del bambino. Il dataset comprende anche tre covariate soggetto-specifiche, vale a dire l'età del bambino in giorni (*Age*), l'indicatore del fatto che il bambino dorma regolarmente sotto una zanzariera (*Net-use*) e l'indicatore che la rete sia stata trattata con insetticidi (*Treated net*). Sono disponibili inoltre delle covariate specifiche per villaggio, ovvero l'indicatore della presenza di un centro sanitario (*PHC*) e una misura satellitare della quantità di verde della vegetazione nei pressi del villaggio (*Green*).

L'obiettivo principale di questa indagine è quello di valutare l'associazione tra la risposta e l'uso delle zanzariere da letto. In precedenti analisi di questi dati sono state impiegate equazioni di stima generalizzate (GEE) (Thomson *et al.*, 1999), modelli lineari generalizzati con effetti casuali spazialmente correlati (Diggle *et al.*, 2002b) e regressioni probit spaziali (Bai *et al.*, 2014). La Figura 4.1 mostra una cartina geografica del Gambia, sulla quale sono state riportate anche le posizioni dei 65 villaggi considerati. Il criterio della colorazione di ciascun pixel, descritto in Thomson *et al.* (1999), deriva dalla scansione satellitare della zona ed il colore è un indicatore della quantità di *Green* dell'area circostante. In particolare i colori blu-beige identificano le zone a basso contenuto, i colori giallo-marrone corrispondono alle zone con medio contenuto, e i colori tendenti al verde sono abbinati alle zone ad alto contenuto di *Green*.

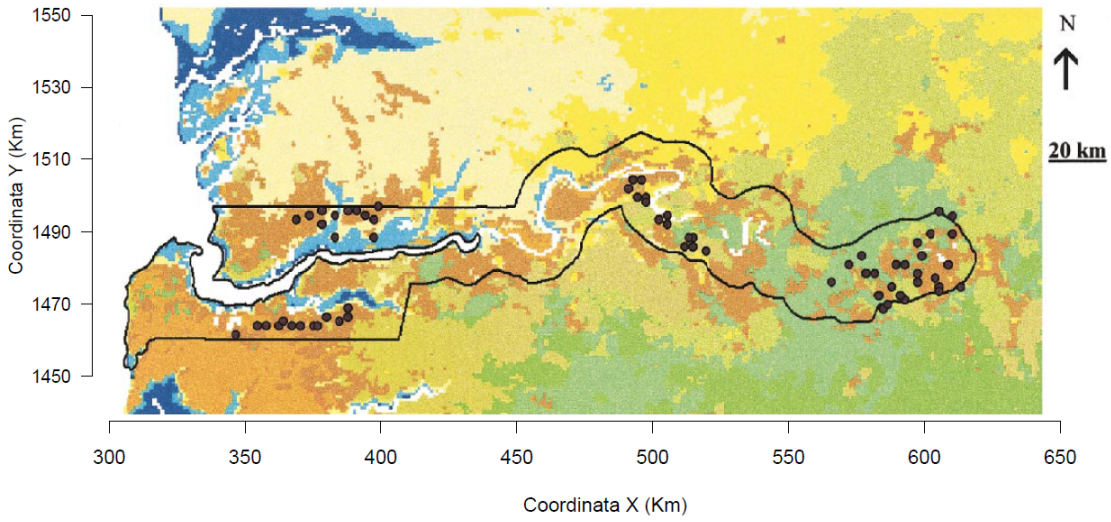

Figura 4.1: Dislocazione spaziale dei villaggi in Gambia.

In Figura 4.2 è riportata la distribuzione delle distanze tra i villaggi considerati, la massima distanza osservata è di 273.3 Km, con una mediana e una media di 109.8 Km e 114.8 Km rispettivamente. Dal grafico è inoltre evidente che le distanze possono essere classificate in tre gruppi; nel primo sono incluse tutte le distanze tra i villaggi che condividono la stessa posizione geografica e non sono distanti più di 60 Km approssimativamente.

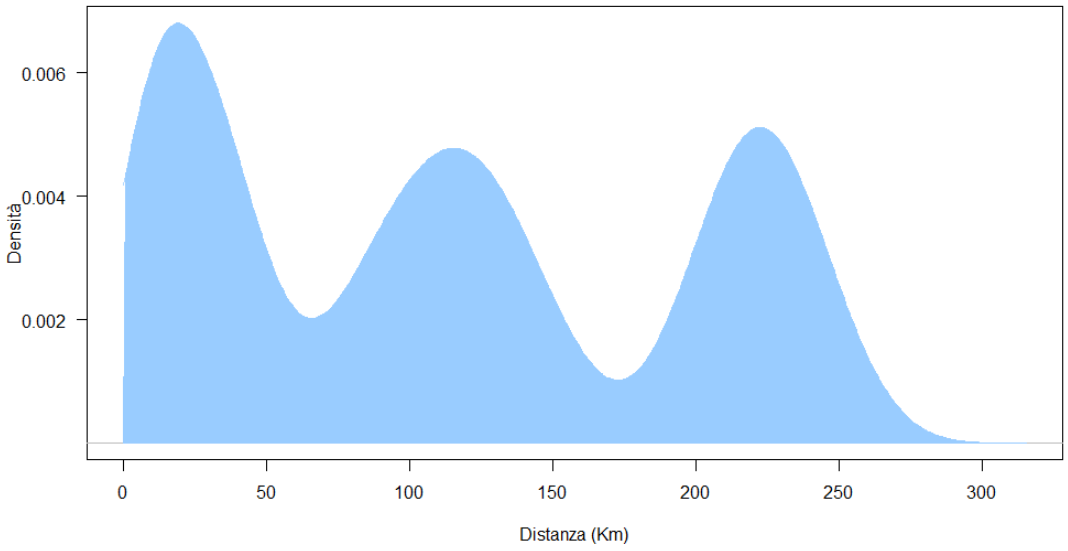

Figura 4.2: Distribuzione delle distanze tra i villaggi in Gambia.

4.1.2 Risultati

Per prima cosa è stato calcolato il lorelogramma spaziale empirico, che permette di valutare la presenza di dipendenza spaziale nei dati e il suo andamento all'aumentare della distanza. Facendo riferimento a Diggle *et al.* (2002b), la distanza massima d_{max} per il calcolo del lorelogramma empirico è stata fissata a 30 Km, questa distanza è stata divisa in $B = 13$ intervalli di raggio $h = 3.5$ Km ciascuno.

I quattro modelli spaziali descritti nel paragrafo 2.4 sono stati adattati al lorelogramma spaziale empirico con ($\alpha_1 \neq 0$) e senza ($\alpha_1 = 0$) effetto *nugget*, utilizzando il metodo dei minimi quadrati pesati (paragrafo 2.5.2). Per avere un confronto con un modello che non considera la correlazione spaziale tra i villaggi è stato impiegato anche il metodo GEE. L'adattamento dei modelli al lorelogramma empirico è stato valutato in termini di AICc (*corrected Akaike information criterion*, Hurvich e Tsai (1989)) e i valori ottenuti sono riportati in Tabella 4.1. Per cogliere le differenze tra i modelli stimati, la terza colonna

della tabella riporta anche il peso di ciascun modello, calcolato come

$$\zeta_m = \exp\{-0.5\Delta\text{AICc}(m)\} / \sum_m \exp\{-0.5\Delta\text{AICc}(m)\},$$

dove

$$\Delta\text{AICc}(m) = \text{AICc}(m) - \min_m \text{AICc}(m),$$

per il modello $m = 1, \ldots, 9$. Come visibile in Tabella 4.1, il modello esponenziale senza effetto *nugget* ha fornito il miglior adattamento in termini di AICc. Da notare che il modello stimato con il metodo GEE ha fornito l'adattamento peggiore, ciò significa che l'inclusione di una qualsiasi funzione di dipendenza spaziale tra quelle riportate in Figura 2.1 comporta un miglioramento nell'adattamento del modello ai dati. In Figura 4.3 viene riportato il lorelogramma empirico relativo ai dati del Gambia, dove " ◯ " sono le stime del logaritmo degli odds ratio e la loro grandezza è proporzionale al numero di coppie appartenenti a ciascun intervallo; la curva più marcata si riferisce all'adattamento del modello esponenziale senza effetto *nugget*.

Modello	*AICc*	*Peso di Akaike*
Esponenziale	-1978.73	0.64
Esponenziale + *nugget*	-1976.09	0.17
Sferico	-1974.62	0.08
Sferico + *nugget*	-1973.79	0.05
Gaussiano + *nugget*	-1972.03	0.02
Gaussiano	-1971.85	0.02
Wave	-1966.65	<0.01
Wave + *nugget*	-1966.58	<0.01
Non spaziale (GEE)	-1945.92	<0.001

Tabella 4.1: Selezione del modello per il lorelogramma spaziale empirico stimato ai minimi quadrati pesati, applicazione al dataset del Gambia.

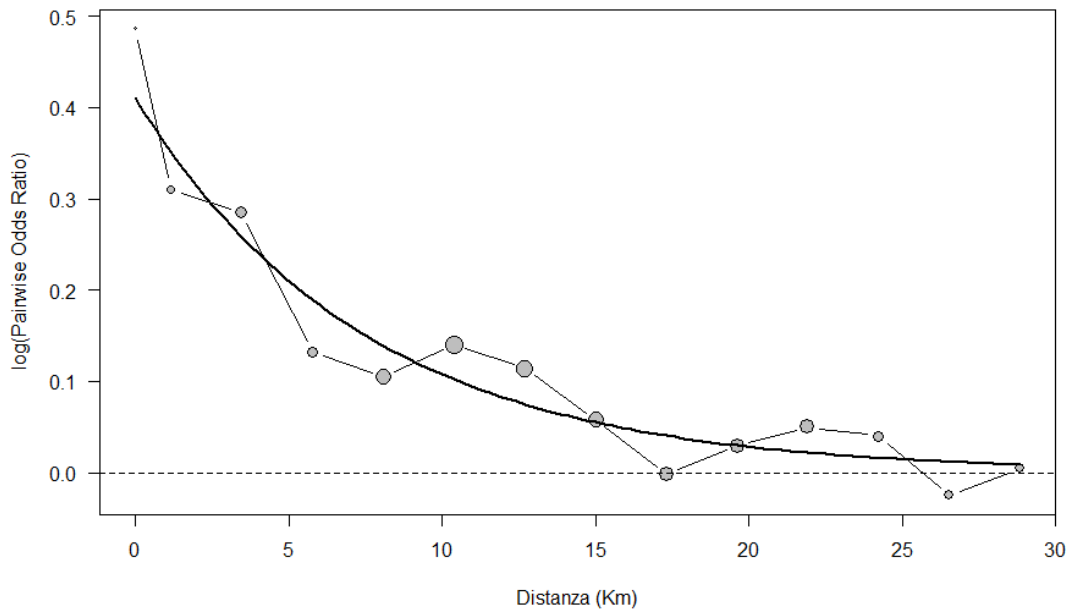

Figura 4.3: Adattamento al lorelogramma empirico del modello esponenziale senza effetto *nugget*, applicazione al dataset del Gambia.

Le stime ai minimi quadrati pesati dei parametri relativi alla struttura di dipendenza esponenziale sono $\hat{\alpha}_{\text{wls},2} = 0.410$ e $\hat{\alpha}_{\text{wls},3} = 7.47$, che corrispondono a un *practical range* di 15.73 Km.

Il modello esponenziale senza effetto *nugget* è stato ristimato con il metodo della verosimiglianza a coppie, nel quale la distanza massima d è stata fissata pari al *practical range* appena stimato e i parametri $\hat{\boldsymbol{\alpha}}_{\text{wls}}$ sono stati forniti all'algoritmo di massimizzazione numerica come valori iniziali. Vista la particolare dislocazione spaziale dei villaggi, si è deciso di dividere la regione in quattro blocchi, evidenziati dai colori diversi dei 65 villaggi in Figura 4.4, in modo tale da evitare assunzioni di indipendenza troppo stringenti. Le stime di massima verosimiglianza a coppie dei parametri di dipendenza sono simili a quelle derivate dal metodo dei minimi quadrati pesati: $\hat{\alpha}_2 = 0.423$ e $\hat{\alpha}_3 = 6.29$. Sono stati stimati quindi un odds ratio entro i cluster di exp(0.423)=1.527 e un *practical range* di 13.43 Km. Come suggerito anche

dal lorelogramma empirico, la stima del *practical range* indica la presenza di una dipendenza spaziale a corto raggio, inoltre solamente il 10.03% delle coppie di osservazioni appartenenti a differenti cluster sono separate da una distanza inferiore o uguale al *practical range*.

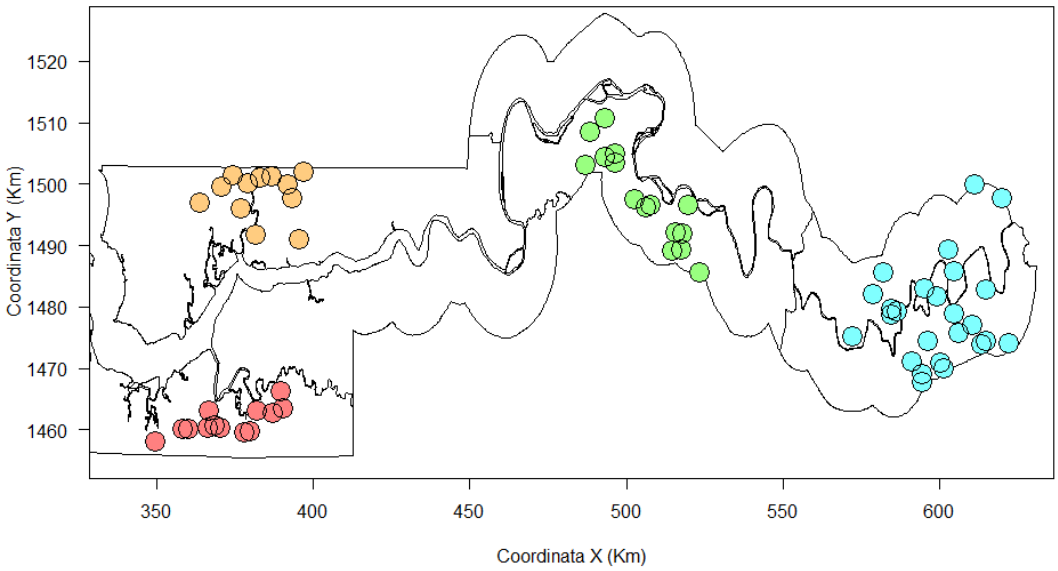

Figura 4.4: Divisione in blocchi dei villaggi in Gambia.

Le stime dei coefficienti di regressione $\boldsymbol{\beta}$ con relativi errori standard sono riportati in Tabella 4.2 ed indicano un'associazione positiva della malaria con l'età, associazione negativa con l'utilizzo della zanzariera, un effetto quadratico di *Green* ed effetti non significativi per la variabile *PHC* e il trattamento della zanzariera con un insetticida. Più precisamente il coefficiente relativo all'uso della zanzariera da letto indica che l'odds di avere la malaria, piuttosto che non averla, tra coloro che ne fanno uso è exp(-0.615) = 0.541 volte l'odds di chi non ne fa utilizzo, cioè è circa del 46% inferiore per chi fa uso della rete rispetto a chi non ne fa uso.

Coefficienti	*Risultati del modello adattato al lorelogramma spaziale*			
	Stima	*SE*	*z*	*p-value*
Intercetta	5.906	0.6970	8.472	<2e-05
Age $\times 10^3$	0.611	1.7e-04	3.697	2e-04
Net-use	-0.615	0.2600	-2.366	0.018
Treated net	-0.204	0.2985	-0.687	0.492
Green	-0.332	0.0126	-26.39	<2e-05
Green^2 $\times 10^2$	0.402	5.5e-04	7.274	<2e-05
PHC	-0.227	0.2609	-0.869	0.385

Tabella 4.2: Risultati relativi ai coefficienti di regressione del modello adattato sui dati del Gambia.

4.2 Dataset *loa-loa*

4.2.1 Introduzione ai dati

In questa seconda applicazione viene considerato il dataset *loa-loa*, che contiene informazioni provenienti da 197 villaggi sparsi in Africa centrale, prevalentemente in Camerun e Nigeria. Il numero di persone campionate per ciascun villaggio è molto variabile: va da 24 a 432, con media e mediana di 135.26 e 132 rispettivamente ed una popolazione totale di 26646 persone. La variabile risposta è l'indicatore binario della presenza di parassiti *Loa loa* nell'organismo dei soggetti. Questi parassiti sono la causa della malattia *Loa loa filariasis*, che colpisce la pelle e gli occhi dei soggetti infettati. Gli esseri umani possono contrarre questa malattia attraverso il morso di Chrysops callidus (anche nota con il nome di *deer fly*) o di Cordylobia Anthropophaga (detta anche *tumbu fly* o *mango fly*), i portatori del verme parassita *Loa loa*. Nel dataset sono presenti anche tre variabili esplicative: l'altitudine del villaggio, *elev*, un indice relativo alla vegetazione, *NDVI*, misurato attraverso

scansioni satellitari dell'area, e la sua deviazione standard *sd_NDVI*. Queste variabili sono quindi villaggio-specifiche e sono state inserite in maniera additiva nel modello.
La Figura 4.5 mostra una cartina geografica di Camerun e Nigeria, sulla quale sono state riportate le posizioni dei 197 villaggi considerati.

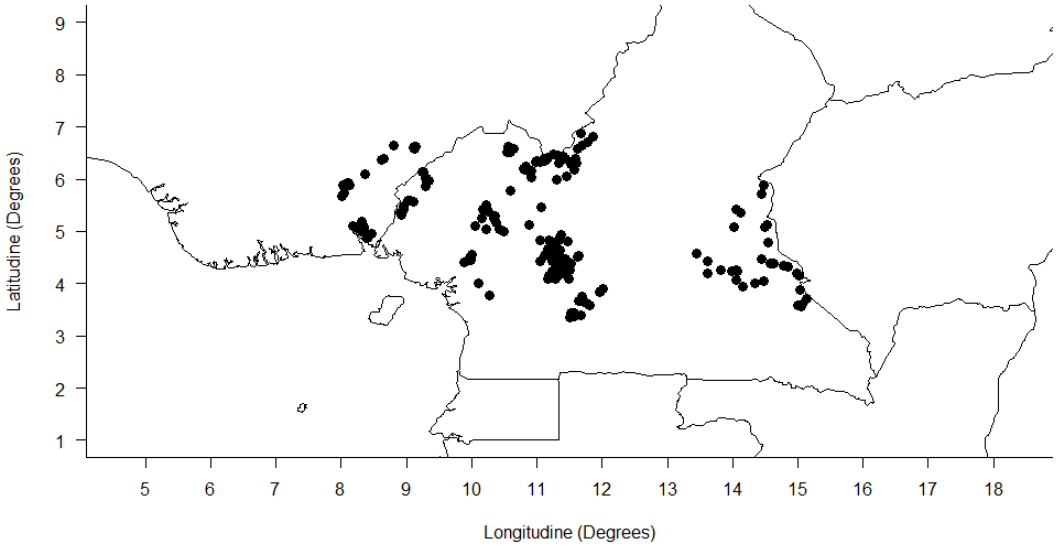

Figura 4.5: Posizione spaziale dei villaggi in Camerun e Nigeria.

In Figura 4.6 è riportata la distribuzione delle distanze in chilometri tra i villaggi considerati. Queste distanze sono state ottenute a partire dai valori di latitudine e longitudine attraverso la funzione R `spDistsN1` nel pacchetto `sp` (Pebesma, E.J. e R.S. Bivand, 2005). La massima distanza osservata è di 824.6 Km, con una mediana e una media di 261.5 Km e 270.7 Km rispettivamente. La massima distanza osservata è di 824.6 Km, con media e mediana rispettivamente di 270.7 Km e 261.5 Km. Per quel che concerne l'applicazione della malaria infantile in Gambia, si può notare che i villaggi sono dislocati su un'area molto più vasta rispetto a quella considerata in precedenza.

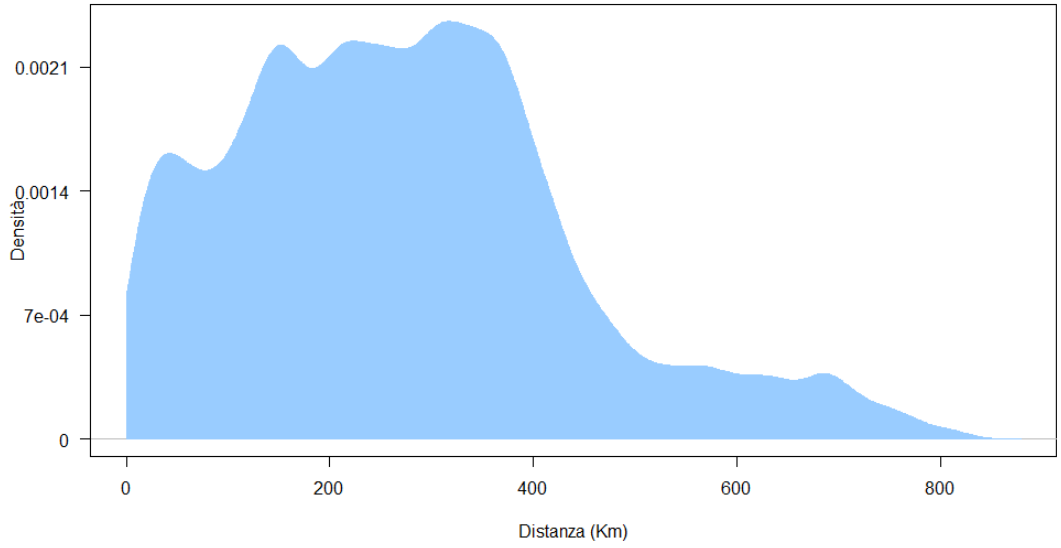

Figura 4.6: Distribuzione delle distanze tra i 197 villaggi.

4.2.2 Risultati

Ricordando che il metodo impiegato in questo elaborato considera coppie di osservazioni, per motivi di natura computazionale dalle 26646 osservazioni iniziali ne sono state campionate 5000, mantenendo la medesima prevalenza per ciascuno dei 197 villaggi rilevati. Per valutare la presenza di dipendenza spaziale in questi dati è stato calcolato il lorelogramma spaziale empirico, dove la distanza massima d_{max} è stata fissata pari a 320 Km. L'intervallo $[0, d_{max}]$ è stato diviso in $B = 9$ intervalli di raggio 30 Km e con una sovrapposizione leggermente inferiore al 50%.

Anche in questo caso, con il metodo dei minimi quadrati pesati riportato nel capitolo 2, sono stati adattati al lorelogramma empirico i modelli parametrici presentati nel paragrafo 2.4, con ($\alpha_1 \neq 0$) e senza ($\alpha_1 = 0$) effetto *nugget*. Con il metodo GEE è stato stimato anche un modello che non considera la correlazione spaziale tra i vari villaggi, così da avere un riferimento. In Tabella 4.3 sono riportati i valori di AICc dei modelli adattati al

Modello	*AICc*	*Peso di Akaike*
Esponenziale	-1123.56	0.34
Sferico	-1123.28	0.29
Gaussiano	-1121.40	0.19
Wave	-1120.88	0.09
Non spaziale (GEE)	-1119.60	0.05
Esponenziale + *nugget*	-1117.62	0.02
Sferico + *nugget*	-1117.36	0.01
Gaussiano + *nugget*	-1116.67	0.01
Wave + *nugget*	-1115.24	$<$0.01

Tabella 4.3: Selezione del modello per il lorelogramma spaziale empirico stimato ai minimi quadrati pesati, applicazione al dataset *loa-loa*.

lorelogramma empirico, di nuovo il miglior adattamento è stato fornito dal modello esponenziale senza effetto *nugget*, visibile in Figura 4.7. Dal grafico è possibile notare che la dipendenza spaziale decresce molto velocemente con l'aumentare della distanza, e dopo i 90 Km sono visibili delle oscillazioni molto più ampie rispetto al caso precedente del Gambia.

Le stime ai minimi quadrati pesati dei parametri relativi alla struttura di dipendenza esponenziale sono $\hat{\alpha}_{\text{wls},2} = 0.337$ e $\hat{\alpha}_{\text{wls},3} = 48.93$, che corrispondono a un *practical range* stimato di 93.38 Km. Questo valore è stato utilizzato come distanza massima d_{max} nella fase di stima del modello con il metodo della verosimiglianza a coppie, al quale i parametri $\hat{\boldsymbol{\alpha}}_{\text{wls}}$ e $\hat{\boldsymbol{\beta}}_{\text{ind}}$ sono stati forniti come punti iniziali dell'algoritmo. Per quanto riguarda il calcolo degli errori standard delle stime, si è scelto di dividere l'area in cinque blocchi, visibili in Figura 4.8 dalla colorazione assegnata ai 197 villaggi.
Le stime di massima verosimiglianza a coppie dei parametri di dipendenza ottenute sono quindi $\hat{\alpha}_2 = 0.425$ e $\hat{\alpha}_3 = 25.70$, di conseguenza sono stati stimati un odds ratio entro i cluster di $\exp(0.425) = 1.53$ e un *practical range* di 54.99 Km.

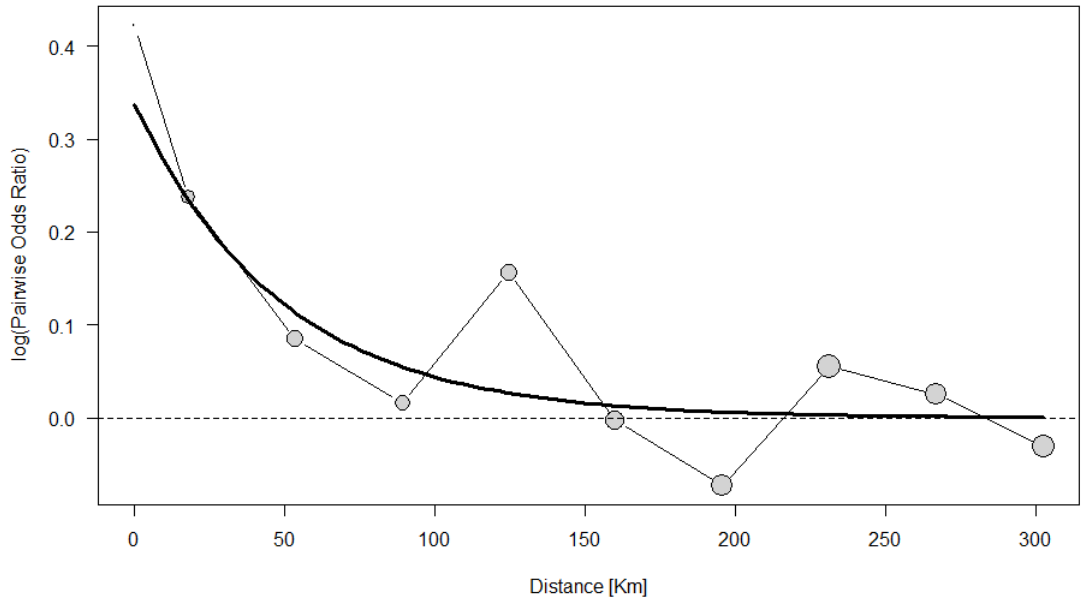

Figura 4.7: Adattamento al lorelogramma empirico del modello esponenziale senza effetto *nugget*, applicazione al dataset *loa-loa*.

Considerando la vastità dell'area su cui sono sparsi i vari villaggi, la stima del *practical range* indica la presenza di una dipendenza spaziale a corto raggio, e come suggerito dal lorelogramma empirico decresce molto rapidamente con la distanza; inoltre solamente il 9.23% delle coppie di osservazioni appartenenti a differenti cluster sono separate da una distanza inferiore o uguale al *practical range*.

Le stime dei coefficienti di regressione $\boldsymbol{\beta}$ con relativi errori standard sono riportati in Tabella 4.4 ed indicano un'associazione positiva della presenza del parassita *loa loa* con *NDVI* ed effetti non significativi per le variabili *elev* e *sd_NDVI*. In particolare il coefficiente relativo all'indice della vegetazione indica che il logaritmo dell'odds di essere infettati dal parassita *Loa loa* aumenta di 11.071 unità all'aumentare di un'unità di *NDVI*.

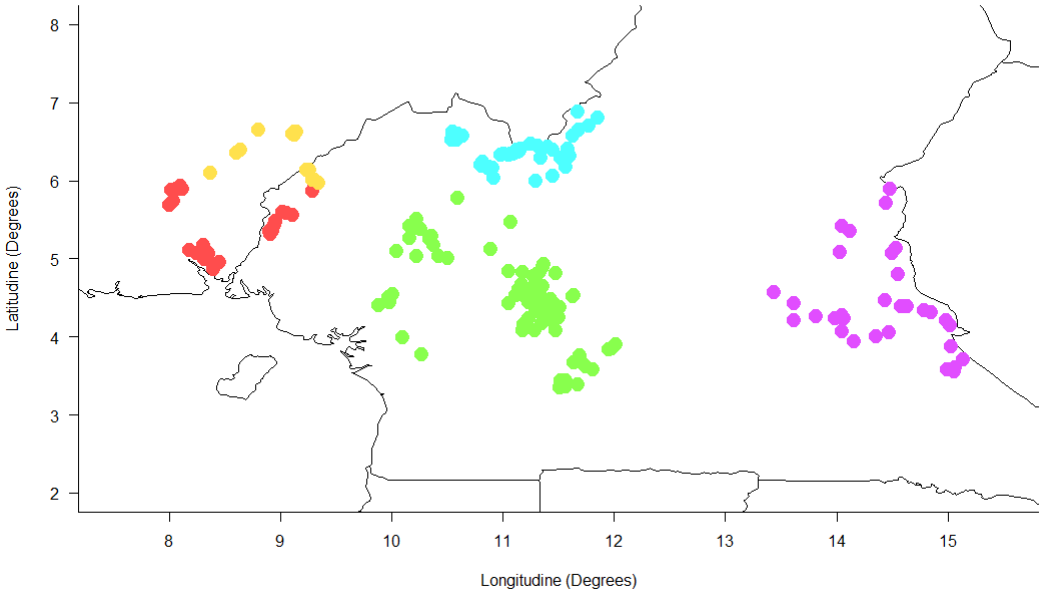

Figura 4.8: Divisione in blocchi dei 197 villaggi in Camerun e Nigeria.

Coefficienti	*Risultati del modello adattato al lorelogramma spaziale*			
	Stima	*SE*	*z*	*p-value*
Intercetta	-11.323	1.475	-7.675	<1e-06
elev	-3.6e-04	4.7e-04	-0.767	0.443
NDVI	11.071	1.961	5.647	<1e-06
sd_NDVI	7.472	9.040	0.827	0.408

Tabella 4.4: Risultati relativi ai coefficienti di regressione del modello adattato al dataset *loa-loa*.

Conclusioni

In questa tesi è stato presentato il modello di regressione logistica applicato a dati binari con dipendenza spaziale ed organizzati in cluster. La forma della dipendenza spaziale è stata catturata dal lorelogramma empirico, una funzione descrittiva della struttura di dipendenza basata su odds ratio. Sono stati proposti vari modelli parametrici per descrivere e modellare la dipendenza spaziale presente nei dati e, in seguito alla selezione del modello più adeguato attraverso il metodo dei minimi quadrati pesati, la stima dei parametri di dipendenza $\boldsymbol{\alpha}$ e regressione $\boldsymbol{\beta}$ è stata eseguita con il metodo della verosimiglianza a coppie. Particolare attenzione è stata prestata al calcolo degli errori standard: il metodo della verosimiglianza a coppie appartiene infatti a una classe di verosimiglianze composite per le quali la seconda identità di Bartlett non è verificata e la derivazione delle matrici H e J è risultata piuttosto complessa. Nelle analisi svolte la matrice di sensibilità H è stata stimata con l'hessiano della log-verosimiglianza a coppie, ottenuto numericamente con la funzione `optimHess` (R Core Team, 2018). Per il calcolo della matrice J è stata adottata una strategia che prevede la divisione in blocchi della regione spaziale considerata, con l'assunzione non sempre realistica che questi siano indipendenti. Lo score di verosimiglianza composita è stato calcolato numericamente attraverso la funzione `grad` nella libreria `numDeriv` ed ha portato ad una stima di J per ciascun blocco; una media pesata di queste matrici ha permesso di giungere ad una stima empirica della matrice di variabilità. Al fine di valutare le proprietà dello stimatore di verosimiglianza a coppie in campioni finiti sono stati condotti due studi di simulazione, la numerosità dei cluster K è stata fissata a 144 e come regione spaziale si è considerato il quadrato unitario. Nel primo studio i cluster sono stati disposti su griglia

regolare 12×12, mentre nel secondo sono stati distribuiti casualmente nell'area considerata. Come è stato mostrato nel capitolo 3, le stime ottenute da entrambi gli studi di simulazione risultano non distorte, sia per i parametri di dipendenza che per quelli di regressione. Per quanto riguarda gli errori standard, derivati dalla stima della matrice di Godambe G al variare del numero di blocchi B, è possibile fare alcune osservazioni. Se comparati con le deviazioni standard delle rispettive stime (fisse al variare della numerosità dei blocchi), per i parametri di dipendenza $\boldsymbol{\alpha}$ si può notare che gli errori standard aumentano all'aumentare di B, passando da una situazione di sottostima ad una dove gli errori standard sovrastimano la vera variabilità delle stime $\hat{\boldsymbol{\alpha}}$. Ciò ha delle ripercussioni sulla copertura degli intervalli di confidenza alla Wald riportati: per basse numerosità di blocchi la copertura è inferiore al livello nominale del 95%, mentre la situazione risulta invertita quando la regione spaziale viene divisa in un numero elevato di blocchi. Per i parametri di regressione $\boldsymbol{\beta}$ si è osservato invece un comportamento negli errori standard leggermente diverso da quello riscontrato per gli $\boldsymbol{\alpha}$: gli errori presentano un andamento crescente, ma dopo un certo numero di blocchi questo subisce un'inversione e gli errori stimati dalla matrice di informazione di Godambe cominciano a ridursi. Si è inoltre notato che questi errori risultano sistematicamente inferiori rispetto alla variabilità effettiva delle stime $\hat{\boldsymbol{\beta}}$, ciò significa che gli errori standard vengono sempre sottostimati e il livello di copertura degli intervalli di confidenza non è garantito. Per quanto riguarda l'andamento generale degli errori standard rispetto alle deviazioni delle stime $\hat{\boldsymbol{\theta}}_{\mathcal{C}}$, nei due studi di simulazione non si sono osservate sostanziali differenze, tuttavia a livello di copertura degli intervalli alla Wald è possibile notare che i valori per la griglia regolare sono sistematicamente, seppur in lieve misura, più alti di quelli osservati per lo studio su griglia non regolare. Infine nel capitolo 4 il metodo di verosimiglianza a coppie è stato applicato ai dataset sulla prevalenza della malaria infantile misurata in Gambia e sulla prevalenza del parassita *Loa loa filariasis* in Camerun e Nigeria. Nel primo caso, vista la naturale separazione geografica dei villaggi in quattro gruppi, si è deciso di mantenere tale suddivisione per il calcolo empirico della matrice J, così da evitare assunzioni di indipendenza irrealistiche. Le stime ottenute

indicano che le variabili esplicative *PHC* e *Treated net* non presentano un effetto significativo sulla prevalenza della malaria, mentre la variabile *net-use*, di diretto interesse nell'indagine svolta, indica che l'odds di avere la malaria tra coloro che fanno uso della zanzariera da letto è exp(-0.615) = 0.541 volte l'odds di chi non ne fa utilizzo. Nella seconda applicazione la dislocazione geografica dei villaggi non presenta una separazione netta come nel caso del Gambia, perciò l'area è stata divisa in cinque blocchi come riportato in Figura 4.8. Le stime ottenute mostrano che l'altitudine del villaggio non influisce sulla presenza del parassita, al contrario l'indice di vegetazione *NDVI* ha un'importante influenza sul fenomeno considerato.

Appendice A

Appendice: Codice R

A.1 Funzioni per la stima del modello su griglia regolare

```
lor.r <- function(y, X, lor.model = c("nonspatial", "exponential", "gaussian
    ", "spherical", "wave"), nugget = FALSE, alpha.start, dmat, dmax, stderr
    =F, n.clust, n.block) {

    require(Matrix)
    lor.model <- match.arg(lor.model)
    nalpha <- 1 + nugget + (lor.model != "nonspatial")
    if (length(alpha.start) != nalpha)
        stop("Wrong size of 'alpha.start'")
    ## starting values beta
    mod0 <- glm.fit(X, y, family = binomial())
    beta.start <- coef(mod0)

    if (lor.model == "nonspatial") dmax <- 0.0
    mypairs <- which(dmat <= dmax & upper.tri(dmat), arr.ind = TRUE)
    dmat.sub <- dmat[mypairs]
    ## compute the maximum pairwise likelihood estimates
    alpha_beta <- .compute.alpha.beta(alpha.start, beta.start, y, X, mypairs
        , lor.model, nugget, dmat.sub)
    alpha <- alpha_beta$par[1:nalpha]
    beta <- alpha_beta$par[-c(1:nalpha)]
    conv <- alpha_beta$convergence
    ifelse(class(alpha_beta)=='optim.obj', iter <- alpha_beta$counts[1],
        iter <- alpha_beta$iterations)
    if(stderr){
        H <- alpha_beta$hessian

```

```
24          # compute J
25          score <- matrix(0, nrow = n.block, ncol = length(alpha_beta$par))
26          Ji <- list()
27          for(i in 1:n.block){
28              score[i,] <- grad(pairllik2, alpha_beta$par, y=y, X=X, mypairs=
                    mypairs, lor.model=lor.model, nugget=nugget, dmat=dmat.sub,
                    n.clust=n.clust, n.block=n.block, b=i)
29              Ji[[i]] <- (n.clust/n.block) * score[i,] %*% t(score[i,])
30          }
31          J <- Reduce('+', Ji)/n.block
32
33          # compute the inverse of Godambe matrix information
34          if(rankMatrix(H)==5 & rankMatrix(J)==5){
35              Ginv <- try(solve(H) %*% J %*% solve(H), silent = T)
36              if(inherits(Ginv, 'try-error')){
37                  ans <- list(alpha = alpha, beta = beta, se.par = rep(NA,
                        NROW(J)), H = H, J = J,      iter = iter, convergence =
                        conv)
38                  class(ans) <- "lor.fit"
39                  return(ans)
40              }
41
42              # standard errors
43              se.par <- sqrt(diag(Ginv))
44              ans <- list(alpha = alpha, beta = beta, se.par = se.par, H = H,
                    J = J, iter = iter, convergence = conv)
45              class(ans) <- "lor.fit"
46              return(ans)
47          }
48
49          ans <- list(alpha = alpha, beta = beta, se.par = rep(NA, NROW(J)), H
                 = H, J = J, iter = iter, convergence = conv)
50          class(ans) <- "lor.fit"
51          return(ans)
52      }
53      else{
54      ans <- list(alpha = alpha, beta = beta, iter = iter, convergence = conv)
55      class(ans) <- "lor.fit"
56      return(ans)
57      }
58  }
59
60
61  .compute.alpha.beta <- function(alpha.start, beta.start, y, X, mypairs, lor.
        model = c("nonspatial", "exponential", "gaussian", "spherical", "wave"),
         nugget, dmat){
62      require(numDeriv)
63
64      lor.model <- match.arg(lor.model)
65      nalpha <- 1 + nugget + (lor.model != "nonspatial")
```

```
66      par <- c(log(alpha.start), beta.start)
67      mple <- nlminb(par, npairllik.ok, y=y, X=X, mypairs=mypairs, lor.model=
            lor.model, nugget=nugget, dmat=dmat)
68      if(mple$convergence!=0){
69          mple <- optim(par, npairllik.ok, method = 'BFGS', hessian = F, y=y,
                X=X, mypairs=mypairs, lor.model=lor.model, nugget=nugget, dmat=
                dmat)
70          class(mple) <- 'optim.obj'
71      }
72      mple$hessian <- optimHess(mple$par, npairllik.ok, y=y, X=X, mypairs=
            mypairs, lor.model=lor.model, nugget=nugget, dmat=dmat)
73      mple$par[1:nalpha] <- al(mple$par[1:nalpha])
74      p <- c(mple$par[1:nalpha],rep(1,length(par)-nalpha)) %*% t(c(mple$par[1:
            nalpha],rep(1,length(par)-nalpha)))
75      mple$hessian <- mple$hessian/p
76      mple
77  }
78
79
80  ta <- function(alpha) log(alpha)
81  al <- function(tau) exp(tau)
82
83  npairllik.ok <- function(param, y, X, mypairs, lor.model = c("nonspatial", "
        exponential", "gaussian", "spherical", "wave"), nugget, dmat){
84
85      nalpha <- 1 + nugget + (lor.model != "nonspatial")
86      npairllik(param=c(al(param[1:nalpha]),param[-c(1:nalpha)]), y=y, X=X,
            mypairs=mypairs, lor.model=lor.model, nugget=nugget, dmat=dmat)
87
88  }
89
90
91  npairllik <- function(param, y, X, mypairs, lor.model = c("nonspatial", "
        exponential", "gaussian", "spherical", "wave"), nugget, dmat){
92
93      lor.model <- match.arg(lor.model)
94      nalpha <- 1 + nugget + (lor.model != "nonspatial")
95      alpha <- param[1:nalpha]
96      mu <- plogis(X %*% param[-c(1:nalpha)])
97      pair.stats <- .compute.stats(y, mu)
98      ## pairwise statistics up to distance dmax
99      pair.stats.ok <- lapply(pair.stats, function(x) x[mypairs])
100     gamma <- lor(dmat, lor.model, nugget, alpha)
101     p11 <- pmin(pmax(.compute.p11(gamma, pair.stats.ok$sum.mu, pair.stats.ok
            $prod.mu), 1e-6), 1-1e-6)
102     p10 <- pair.stats.ok$p1. - p11
103     p01 <- pair.stats.ok$p.1 - p11
104     p00 <- 1 + 1e-3 - p01 - p10 - p11
105     probs <- with(pair.stats.ok, y11 * log(p11) + y01 * log(p01) + y10 * log
            (p10) + y00 * log(p00))
```

```
106     -sum(probs)
107 }
108 
109 
110 pairllik2 <- function(param, y, X, mypairs, lor.model = c("nonspatial", "
        exponential", "gaussian", "spherical", "wave"), nugget, dmat, n.clust, n
        .block, b){
111 
112     lor.model <- match.arg(lor.model)
113     nalpha <- 1 + nugget + (lor.model != "nonspatial")
114     alpha <- param[1:nalpha]
115     mu <- plogis(X %*% param[-c(1:nalpha)])
116     pair.stats <- .compute.stats(y, mu)
117     ## pairwise statistics up to distance dmax
118     pair.stats.ok <- lapply(pair.stats, function(x) x[mypairs])
119     gamma <- lor(dmat, lor.model, nugget, alpha)
120     p11 <- pmin(pmax(.compute.p11(gamma, pair.stats.ok$sum.mu, pair.stats.ok
            $prod.mu), 1e-6), 1-1e-6)
121     p10 <- pair.stats.ok$p1. - p11
122     p01 <- pair.stats.ok$p.1 - p11
123     p00 <- 1 + 1e-3 - p01 - p10 - p11
124     probs <- with(pair.stats.ok, y11 * log(p11) + y01 * log(p01) + y10 * log
            (p10) + y00 * log(p00))
125     llik.val <- matrix(0, nrow = length(y), ncol = length(y))
126     llik.val[mypairs] <- probs
127 
128     nobs <- length(y)/n.clust
129     cluster <- matrix(0, n.block, n.clust/n.block)
130     pos <- array(0, c(n.block,n.clust/n.block,nobs))
131     a <- matrix(0, nrow = sqrt(n.clust), ncol = sqrt(n.clust))
132     for(i in 1:sqrt(n.clust)){
133         a[,i] <- c((i*sqrt(n.clust)):(((i-1)*sqrt(n.clust))+1))
134     }
135     uni <- matrix(1, nrow = sqrt(n.clust/n.block), ncol = sqrt(n.clust/n.
            block))
136     zeri <- matrix(0, nrow = sqrt(n.clust), ncol = sqrt(n.clust))
137 
138     ans <- seq(1, sqrt(n.clust), by=sqrt(n.clust/n.block))
139     ans2 <- NULL
140     for(i in 1:length(ans)){
141         ans2 <- c(ans2,ans+(ans[i]-1) * sqrt(n.clust))
142     }
143 
144     pos.block <- which(a==ans2[b], arr.ind = T)
145     zeri[pos.block[1]:(pos.block[1]-sqrt(n.clust/n.block)+1),pos.block[2]:(
            pos.block[2]+sqrt(n.clust/n.block)-1)] <- uni
146 
147     #there is a row for each block and in every row there are cluster
            belonging to the relative block
148     cluster[b,] <- a[zeri==1]
```

```
149
150     for(j in 1:(n.clust/n.block)){
151         pos[b,j,] <- ((cluster[b,j]-1)*nobs+1):(cluster[b,j]*nobs)
152     }
153     val <- 0
154     for(i in 1:(n.clust/n.block)){
155         val <- val + sum(llik.val[pos[b,i,],pos[b,,]])
156     }
157     return(val)
158 }
```

A.2 Funzioni per la stima del modello su griglia non regolare

```
1  lor.nr <- function(y, X, coords, lor.model = c("nonspatial", "exponential",
       "gaussian", "spherical", "wave"), nugget = FALSE, alpha.start, dmat,
       dmax, stderr=F, n.clust, n.block) {
2
3      lor.model <- match.arg(lor.model)
4      nalpha <- 1 + nugget + (lor.model != "nonspatial")
5      if (length(alpha.start) != nalpha)
6          stop("Wrong size of 'alpha.start'")
7      ## starting values beta
8      mod0 <- glm.fit(X, y, family = binomial())
9      beta.start <- coef(mod0)
10
11     if (lor.model == "nonspatial") dmax <- 0.0
12     mypairs <- which(dmat <= dmax & upper.tri(dmat), arr.ind = TRUE)
13     dmat.sub <- dmat[mypairs]
14     ## compute the maximum pairwise likelihood estimates
15     alpha_beta <- .compute.alpha.beta(alpha.start, beta.start, y, X, mypairs
           , lor.model, nugget, dmat.sub)
16     alpha <- alpha_beta$par[1:nalpha]
17     beta <- alpha_beta$par[-c(1:nalpha)]
18     conv <- alpha_beta$convergence
19     ifelse(class(alpha_beta)=='optim.obj', iter <- alpha_beta$counts[1],
           iter <- alpha_beta$iterations)
20     if(stderr){
21         H <- alpha_beta$hessian
22
23         # compute J
24         score <- matrix(0, nrow = n.block, ncol = length(alpha_beta$par))
25         Ji <- list()
26         block2remove <- ctrl.cluster(coords, n.block)
27         ifelse(length(block2remove)==0, tmp <- 1:n.block, tmp <- (1:n.block)
               [-block2remove])
28         for(i in tmp){
```

```
29              wi <- wt(y=y, coords = coords, mypairs = mypairs, n.clust = n.
                    clust, n.block = n.block, b=i)
30              score[i,] <- grad(pairllik3, alpha_beta$par, y=y, X=X, coords=
                    coords, mypairs=mypairs, lor.model=lor.model, nugget=nugget,
                     dmat=dmat.sub, n.clust=n.clust, n.block=n.block, b=i)
31              Ji[[i]] <- wi * score[i,] %*% t(score[i,])
32          }
33          if(length(block2remove)!=0) Ji <- Ji[-block2remove]
34          J <- Reduce('+', Ji)/(n.block-length(block2remove))
35
36          # compute the inverse of Godambe matrix information
37          if(rankMatrix(H)==5 & rankMatrix(J)==5){
38              Ginv <- try(solve(H) %*% J %*% solve(H), silent = T)
39              if(inherits(Ginv, 'try-error')){
40                  ans <- list(alpha = alpha, beta = beta, se.par = rep(NA,
                        NROW(J)), H = H, J = J,     iter = iter, convergence =
                        conv)
41                  class(ans) <- "lor.fit"
42                  return(ans)
43              }
44
45              # standard errors
46              se.par <- sqrt(diag(Ginv))
47              ans <- list(alpha = alpha, beta = beta, se.par = se.par, H = H,
                    J = J, iter = iter, convergence = conv)
48              class(ans) <- "lor.fit"
49              return(ans)
50          }
51
52          ans <- list(alpha = alpha, beta = beta, se.par = rep(NA, NROW(J)), H
                 = H, J = J, iter = iter, convergence = conv)
53          class(ans) <- "lor.fit"
54          return(ans)
55      }
56      else{
57          ans <- list(alpha = alpha, beta = beta, iter = iter, convergence =
                conv)
58          class(ans) <- "lor.nr"
59          return(ans)
60      }
61  }
62
63
64  ctrl.cluster <- function(coords, n.block){
65
66      q1 <- seq(range(coords[,1])[1], range(coords[,1])[2], length.out = sqrt(
            n.block)+1)
67      q2 <- seq(range(coords[,2])[1], range(coords[,2])[2], length.out = sqrt(
            n.block)+1)
68      M <- matrix(0, nrow = sqrt(n.block), ncol = sqrt(n.block))
```

```
69      for(i in 1:sqrt(n.block)){
70          M[,i] <- c((i*sqrt(n.block)):(((i-1)*sqrt(n.block))+1))
71      }
72      clust.coord <- unique(coords)
73      rownames(clust.coord) <- 1:NROW(clust.coord)
74      bl2rm <- NULL
75      for(j in 1:n.block){
76          pos <- which(M==j, arr.ind = T)
77          cx <- as.numeric(cut(clust.coord[,1], breaks = q1, include.lowest =
                T))
78          cy <- as.numeric(cut(clust.coord[,2], breaks = q2, include.lowest =
                T))
79          cluster <- as.numeric(rownames(clust.coord)[cx==pos[2] & cy==sqrt(n.
                block)+1-pos[1]])
80          if(length(cluster)==0) bl2rm <- c(bl2rm,j)
81      }
82      return(bl2rm)
83  }
84
85
86  pairllik3 <- function(param, y, X, coords, mypairs, lor.model = c("
        nonspatial", "exponential", "gaussian", "spherical", "wave"), nugget,
        dmat, n.clust, n.block, b){
87
88      nobs <- length(y)/n.clust
89      q1 <- seq(range(coords[,1])[1], range(coords[,1])[2], length.out = sqrt(
            n.block)+1)
90      q2 <- seq(range(coords[,2])[1], range(coords[,2])[2], length.out = sqrt(
            n.block)+1)
91      M <- matrix(0, nrow = sqrt(n.block), ncol = sqrt(n.block))
92      for(i in 1:sqrt(n.block)){
93          M[,i] <- c((i*sqrt(n.block)):(((i-1)*sqrt(n.block))+1))
94      }
95      pos <- which(M==b, arr.ind = T)
96      clust.coord <- unique(coords)
97      rownames(clust.coord) <- 1:NROW(clust.coord)
98      cx <- as.numeric(cut(clust.coord[,1], breaks = q1, include.lowest = T))
99      cy <- as.numeric(cut(clust.coord[,2], breaks = q2, include.lowest = T))
100     cluster <- as.numeric(rownames(clust.coord)[cx==pos[2] & cy==sqrt(n.
            block)+1-pos[1]])
101     pos.obs <- matrix(0, nrow = length(cluster), ncol = nobs)
102     for(j in 1:length(cluster)){
103         pos.obs[j,] <- ((cluster[j]-1)*nobs+1):(cluster[j]*nobs)
104     }
105
106     lor.model <- match.arg(lor.model)
107     nalpha <- 1 + nugget + (lor.model != "nonspatial")
108     alpha <- param[1:nalpha]
109     mu <- plogis(X%*%param[-c(1:nalpha)])
110     pair.stats <- .compute.stats(y, mu)
```

```
111     ## pairwise statistics up to distance dmax
112     pair.stats.ok <- lapply(pair.stats, function(x) x[mypairs])
113     gamma <- lor(dmat, lor.model, nugget, alpha)
114     p11 <- pmin(pmax(.compute.p11(gamma, pair.stats.ok$sum.mu, pair.stats.ok
            $prod.mu), 1e-6), 1-1e-6)
115     p10 <- pair.stats.ok$p1. - p11
116     p01 <- pair.stats.ok$p.1 - p11
117     p00 <- 1 + 1e-3 - p01 - p10 - p11
118     probs <- with(pair.stats.ok, y11 * log(p11) + y01 * log(p01) + y10 * log
            (p10) + y00 * log(p00))
119     llik.val <- matrix(0, nrow = length(y), ncol = length(y))
120     llik.val[mypairs] <- probs
121 
122     val <- 0
123     for(i in 1:length(cluster)){
124         val <- val + sum(llik.val[pos.obs[i,],pos.obs])
125     }
126     return(val)
127 }
128 
129 
130 wt <- function(y, coords, mypairs, n.clust, n.block, b){
131     nobs <- length(y)/n.clust
132     q1 <- seq(range(coords[,1])[1], range(coords[,1])[2], length.out = sqrt(
            n.block)+1)
133     q2 <- seq(range(coords[,2])[1], range(coords[,2])[2], length.out = sqrt(
            n.block)+1)
134     M <- matrix(0, nrow = sqrt(n.block), ncol = sqrt(n.block))
135     for(i in 1:sqrt(n.block)){
136         M[,i] <- c((i*sqrt(n.block)):(((i-1)*sqrt(n.block))+1))
137     }
138     pos <- which(M==b, arr.ind = T)
139     clust.coord <- unique(coords)
140     rownames(clust.coord) <- 1:NROW(clust.coord)
141     cx <- as.numeric(cut(clust.coord[,1], breaks = q1, include.lowest = T))
142     cy <- as.numeric(cut(clust.coord[,2], breaks = q2, include.lowest = T))
143     cluster <- as.numeric(rownames(clust.coord)[cx==pos[2] & cy==sqrt(n.
            block)+1-pos[1]])
144     pos.obs <- matrix(0, nrow = length(cluster), ncol = nobs)
145     for(j in 1:length(cluster)){
146         pos.obs[j,] <- ((cluster[j]-1)*nobs+1):(cluster[j]*nobs)
147     }
148     NROW(mypairs)/NROW(mypairs[mypairs[,1] %in% pos.obs & mypairs[,2] %in%
            pos.obs,])
149 }
```

A.3 Funzioni per la stima del modello al dataset del Gambia

```
lor.nrG <- function(y, X, coords, lor.model = c("nonspatial", "exponential",
     "gaussian", "spherical", "wave"), nugget = FALSE, alpha.start, dmat,
    dmax, stderr=F, n.clust, n.block, nobs) {

    lor.model <- match.arg(lor.model)
    nalpha <- 1 + nugget + (lor.model != "nonspatial")
    if (length(alpha.start) != nalpha)
        stop("Wrong size of 'alpha.start'")
    ## starting values beta
    mod0 <- glm.fit(X, y, family = binomial())
    beta.start <- coef(mod0)

    if (lor.model == "nonspatial") dmax <- 0.0
    mypairs <- which(dmat <= dmax & upper.tri(dmat), arr.ind = TRUE)
    dmat.sub <- dmat[mypairs]
    ## compute the maximum pairwise likelihood estimates
    cat('compute the mple...\n')
    alpha_beta <- .compute.alpha.beta(alpha.start, beta.start, y, X, mypairs
        , lor.model, nugget, dmat.sub)  #algoritmo pag 10
    cat('------> computed!\n')
    alpha <- alpha_beta$par[1:nalpha]
    beta <- alpha_beta$par[-c(1:nalpha)]
    conv <- alpha_beta$convergence
    ifelse(class(alpha_beta)=='optim.obj', iter <- alpha_beta$counts[1],
        iter <- alpha_beta$iterations)
    if(stderr){
        cat('compute the stderr of mple...\n')
        H <- alpha_beta$hessian

        # compute J
        score <- matrix(0, nrow = n.block, ncol = length(alpha_beta$par))
        Ji <- list()
        block2remove <- ctrl.clusterG(coords, n.block)
        ifelse(length(block2remove)==0, tmp <- 1:n.block, tmp <- (1:n.block)
            [-block2remove])
        for(i in tmp){
            cat(i,')\t')
            wi <- wt(y=y, coords = coords, mypairs = mypairs, n.clust = n.
                clust, n.block = n.block, b=i)
            score[i,] <- grad(pairllik3G, alpha_beta$par, y=y, X=X, coords=
                coords, mypairs=mypairs, lor.model=lor.model, nugget=nugget,
                 dmat=dmat.sub, n.clust=n.clust, n.block=n.block, b=i, nobs
                = nobs)
            Ji[[i]] <- wi * score[i,] %*% t(score[i,])
        }
```

```
37          if(length(block2remove)!=0) Ji <- Ji[-block2remove]
38          J <- Reduce('+', Ji)/(n.block-length(block2remove))
39
40          # compute the inverse of Godambe matrix information
41          if(rankMatrix(H)==5 & rankMatrix(J)==5){
42              Ginv <- try(solve(H) %*% J %*% solve(H), silent = T)
43              if(inherits(Ginv, 'try-error')){
44                  ans <- list(alpha = alpha, beta = beta, se.par = rep(NA,
                        NROW(J)), H = H, J = J,      iter = iter, convergence =
                        conv)
45                  class(ans) <- "lor.fit"
46                  return(ans)
47              }
48
49              # standard errors
50              se.par <- sqrt(diag(Ginv))
51              ans <- list(alpha = alpha, beta = beta, se.par = se.par, H = H,
                    J = J, iter = iter, convergence = conv)
52              class(ans) <- "lor.fit"
53              return(ans)
54          }
55
56          ans <- list(alpha = alpha, beta = beta, se.par = rep(NA, NROW(J)), H
                 = H, J = J, iter = iter, convergence = conv)
57          class(ans) <- "lor.fit"
58          return(ans)
59      }
60      else{
61          ans <- list(alpha = alpha, beta = beta, iter = iter, convergence =
                conv)
62          class(ans) <- "lor.nr"
63          return(ans)
64      }
65  }
66
67
68  ctrl.clusterG <- function(coords, n.block){
69
70      q1 <- c(349, 420, 550, 623)
71      q2 <- c(1456, 1467.5, 1511)
72      M <- matrix(0, nrow = (length(q2)-1), ncol = (length(q1)-1))
73      for(i in 1:(length(q1)-1)){
74          M[,i] <- c((i*(length(q2)-1)):(((i-1)*(length(q2)-1))+1))
75      }
76      clust.coord <- unique(coords)
77      rownames(clust.coord) <- 1:NROW(clust.coord)
78      bl2rm <- NULL
79      for(j in 1:n.block){
80          pos <- which(M==j, arr.ind = T)
81          cx <- as.numeric(cut(clust.coord[,1], breaks = q1, include.lowest =
```

```
            T))
82          cy <- as.numeric(cut(clust.coord[,2], breaks = q2, include.lowest =
                T))
83          cluster <- as.numeric(rownames(clust.coord)[cx==pos[2] & cy==length(
                q2)-pos[1]])
84          if(length(cluster)==0) bl2rm <- c(bl2rm,j)
85      }
86      return(bl2rm)
87  }
88
89
90  pairllik3G <- function(param, y, X, coords, mypairs, lor.model = c("
        nonspatial", "exponential", "gaussian", "spherical", "wave"), nugget,
        dmat, n.clust, n.block, b, nobs){
91
92      q1 <- c(349, 420, 550, 623)
93      q2 <- c(1456, 1467.5, 1511)
94      M <- matrix(0, nrow = (length(q2)-1), ncol = (length(q1)-1))
95      for(i in 1:(length(q1)-1)){
96          M[,i] <- c((i*(length(q2)-1)):(((i-1)*(length(q2)-1))+1))
97      }
98      pos <- which(M==b, arr.ind = T)
99      clust.coord <- unique(coords)
100     rownames(clust.coord) <- 1:NROW(clust.coord)
101     cx <- as.numeric(cut(clust.coord[,1], breaks = q1, include.lowest = T))
102     cy <- as.numeric(cut(clust.coord[,2], breaks = q2, include.lowest = T))
103     cluster <- as.numeric(rownames(clust.coord)[cx==pos[2] & cy==length(q2)-
            pos[1]])
104     pos.obs <- matrix(0, nrow = length(cluster), ncol = max(nobs[cluster]))
105     for(j in 1:length(cluster)){
106         pos.obs[j,(1:nobs[cluster[j]])] <- (sum(nobs[1:(cluster[j]-1)])+1):(
                sum(nobs[1:(cluster[j]-1)])+nobs[cluster[j]])
107         if(b==1) pos.obs[1,] <- c(1:33, rep(0,(max(nobs[cluster])-33)))
108     }
109
110     lor.model <- match.arg(lor.model)
111     nalpha <- 1 + nugget + (lor.model != "nonspatial")
112     alpha <- param[1:nalpha]
113     mu <- plogis(X %*% param[-c(1:nalpha)])
114     pair.stats <- .compute.stats(y, mu)
115     ## pairwise statistics up to distance dmax
116     pair.stats.ok <- lapply(pair.stats, function(x) x[mypairs])
117     gamma <- lor(dmat, lor.model, nugget, alpha)
118     p11 <- pmin(pmax(.compute.p11(gamma, pair.stats.ok$sum.mu, pair.stats.ok
            $prod.mu), 1e-6), 1-1e-6)
119     p10 <- pair.stats.ok$p1. - p11
120     p01 <- pair.stats.ok$p.1 - p11
121     p00 <- 1 + 1e-3 - p01 - p10 - p11
122     probs <- with(pair.stats.ok, y11 * log(p11) + y01 * log(p01) + y10 * log
            (p10) + y00 * log(p00))
```

```
123     llik.val <- matrix(0, nrow = length(y), ncol = length(y))
124     llik.val[mypairs] <- probs
125 
126     val <- 0
127     for(i in 1:length(cluster)){
128         val <- val + sum(llik.val[pos.obs[i,],pos.obs])
129     }
130     return(val)
131 }
```

A.4 Funzioni per la stima del modello al dataset *loa-loa*

```
1  lor.nrN <- function(y, X, coords, lor.model = c("nonspatial", "exponential",
       "gaussian", "spherical", "wave"), nugget = FALSE, alpha.start, dmat,
     dmax, stderr=F, n.clust, n.block, nobs) {
2      require(Matrix)
3  
4      lor.model <- match.arg(lor.model)
5      nalpha <- 1 + nugget + (lor.model != "nonspatial")
6      if (length(alpha.start) != nalpha)
7          stop("Wrong size of 'alpha.start'")
8      ## starting values beta
9      mod0 <- glm.fit(X, y, family = binomial())
10     beta.start <- coef(mod0)
11 
12     if (lor.model == "nonspatial") dmax <- 0.0
13     mypairs <- which(dmat <= dmax & upper.tri(dmat), arr.ind = TRUE)
14     dmat.sub <- dmat[mypairs]
15     ## compute the maximum pairwise likelihood estimates
16     cat('compute the mple...\n')
17     alpha_beta <- .compute.alpha.beta(alpha.start, beta.start, y, X, mypairs
           , lor.model, nugget, dmat.sub)
18     cat('------> computed!\n')
19     alpha <- alpha_beta$par[1:nalpha]
20     beta <- alpha_beta$par[-c(1:nalpha)]
21     conv <- alpha_beta$convergence
22     ifelse(class(alpha_beta)=='optim.obj', iter <- alpha_beta$counts[1],
           iter <- alpha_beta$iterations)
23     if(stderr){
24         cat('compute the stderr of mple...\n')
25         H <- alpha_beta$hessian
26 
```

```
27          # calcolo di J a blocchi
28          score <- matrix(0, nrow = n.block, ncol = length(alpha_beta$par))
29          Ji <- list()
30          block2remove <- ctrl.clusterN(coords, n.block)
31          ifelse(length(block2remove)==0, tmp <- 1:n.block, tmp <- (1:n.block)
                [-block2remove])
32          for(i in tmp){
33              cat(i,')\t')
34              wi <- wtN(y=y, coords = coords, mypairs = mypairs, n.clust = n.
                    clust, n.block = n.block, nobs = nobs, b=i)
35              score[i,] <- grad(pairllik3N, alpha_beta$par, y=y, X=X, coords=
                    coords, mypairs=mypairs, lor.model=lor.model, nugget=nugget,
                     dmat=dmat.sub, n.clust=n.clust, n.block=n.block, nobs =
                    nobs, b=i)
36              Ji[[i]] <- wi * score[i,] %*% t(score[i,])
37          }
38          if(length(block2remove)!=0) Ji <- Ji[-block2remove]
39          J <- Reduce('+', Ji)/(n.block-length(block2remove))
40
41          # compute the inverse of Godambe matrix information
42          if(rankMatrix(H)==6 & rankMatrix(J)==6){
43              Ginv <- try(solve(H) %*% J %*% solve(H), silent = T)
44              if(inherits(Ginv, 'try-error')){
45                  ans <- list(alpha = alpha, beta = beta, se.par = rep(NA,
                        NROW(J)), H = H, J = J, iter = iter, convergence = conv)
46                  class(ans) <- "lor.fit"
47                  return(ans)
48              }
49
50              # standard errors
51              se.par <- sqrt(diag(Ginv))
52              ans <- list(alpha = alpha, beta = beta, se.par = se.par, H = H,
                    J = J, iter = iter, convergence = conv)
53              class(ans) <- "lor.fit"
54              return(ans)
55          }
56
57          ans <- list(alpha = alpha, beta = beta, se.par = rep(NA, NROW(J)), H
                 = H, J = J, iter = iter, convergence = conv)
58          class(ans) <- "lor.fit"
59          return(ans)
60      }
61      else{
62          ans <- list(alpha = alpha, beta = beta, iter = iter, convergence =
                conv)
63          class(ans) <- "lor.nr"
64          return(ans)
65      }
66  }
67
```

```
ctrl.clusterN <- function(coords, n.block){

    q1 <- c(8, 9.5, 13, 15.2)
    q2 <- c(3.3, 5.95, 6.9)
    M <- matrix(0, nrow = (length(q2)-1), ncol = (length(q1)-1))
    for(i in 1:(length(q1)-1)){
        M[,i] <- c((i*(length(q2)-1)):(((i-1)*(length(q2)-1))+1))
    }
    clust.coord <- unique(coords)
    rownames(clust.coord) <- 1:NROW(clust.coord)
    bl2rm <- NULL
    for(j in 1:n.block){
        pos <- which(M==j, arr.ind = T)
        cx <- as.numeric(cut(clust.coord[,1], breaks = q1, include.lowest =
            T))
        cy <- as.numeric(cut(clust.coord[,2], breaks = q2, include.lowest =
            T))
        cluster <- as.numeric(rownames(clust.coord)[cx==pos[2] & cy==length(
            q2)-pos[1]])
        if(length(cluster)==0) bl2rm <- c(bl2rm,j)
    }
    return(bl2rm)
}


pairllik3N <- function(param, y, X, coords, mypairs, lor.model = c("
    nonspatial", "exponential", "gaussian", "spherical", "wave"), nugget,
    dmat, n.clust, n.block, nobs, b){

    q1 <- c(8, 9.5, 13, 15.2)
    q2 <- c(3.3, 5.95, 6.9)
    M <- matrix(0, nrow = (length(q2)-1), ncol = (length(q1)-1))
    for(i in 1:(length(q1)-1)){
        M[,i] <- c((i*(length(q2)-1)):(((i-1)*(length(q2)-1))+1))
    }
    pos <- which(M==b, arr.ind = T)
    clust.coord <- unique(coords)
    rownames(clust.coord) <- 1:NROW(clust.coord)
    cx <- as.numeric(cut(clust.coord[,1], breaks = q1, include.lowest = T))
    cy <- as.numeric(cut(clust.coord[,2], breaks = q2, include.lowest = T))
    cluster <- as.numeric(rownames(clust.coord)[cx==pos[2] & cy==length(q2)-
        pos[1]])
    pos.obs <- matrix(0, nrow = length(cluster), ncol = max(nobs[cluster]))
    for(j in 1:length(cluster)){
        pos.obs[j,(1:nobs[cluster[j]])] <- (sum(nobs[1:(cluster[j]-1)])+1):(
            sum(nobs[1:(cluster[j]-1)])+nobs[cluster[j]])
        if(cluster[1]==1) pos.obs[1,] <- c(1:nobs[cluster[1]], rep(0,(max(
            nobs[cluster])-nobs[cluster[1]])))
    }

```

```
110     lor.model <- match.arg(lor.model)
111     nalpha <- 1 + nugget + (lor.model != "nonspatial")
112     alpha <- param[1:nalpha]
113     mu <- plogis(X%*%param[-c(1:nalpha)])
114     pair.stats <- .compute.stats(y, mu)
115     ## pairwise statistics up to distance dmax
116     pair.stats.ok <- lapply(pair.stats, function(x) x[mypairs])
117     gamma <- lor(dmat, lor.model, nugget, alpha)
118     p11 <- pmin(pmax(.compute.p11(gamma, pair.stats.ok$sum.mu, pair.stats.ok
            $prod.mu), 1e-6), 1-1e-6)
119     p10 <- pair.stats.ok$p1. - p11
120     p01 <- pair.stats.ok$p.1 - p11
121     p00 <- 1 + 1e-3 - p01 - p10 - p11
122     probs <- with(pair.stats.ok, y11 * log(p11) + y01 * log(p01) + y10 * log
            (p10) + y00 * log(p00))
123     llik.val <- matrix(0, nrow = length(y), ncol = length(y))
124     llik.val[mypairs] <- probs
125
126     val <- 0
127     for(i in 1:length(cluster)){
128         val <- val + sum(llik.val[pos.obs[i,],pos.obs])
129     }
130     return(val)
131 }
132
133 wtN <- function(y, coords, mypairs, n.clust, n.block, nobs, b){
134
135     q1 <- c(8, 9.5, 13, 15.2)
136     q2 <- c(3.3, 5.95, 6.9)
137     M <- matrix(0, nrow = (length(q2)-1), ncol = (length(q1)-1))
138     for(i in 1:(length(q1)-1)){
139         M[,i] <- c((i*(length(q2)-1)):(((i-1)*(length(q2)-1))+1))
140     }
141     pos <- which(M==b, arr.ind = T)
142     clust.coord <- unique(coords)
143     rownames(clust.coord) <- 1:NROW(clust.coord)
144     cx <- as.numeric(cut(clust.coord[,1], breaks = q1, include.lowest = T))
145     cy <- as.numeric(cut(clust.coord[,2], breaks = q2, include.lowest = T))
146     cluster <- as.numeric(rownames(clust.coord)[cx==pos[2] & cy==length(q2)-
            pos[1]])
147     pos.obs <- matrix(0, nrow = length(cluster), ncol = max(nobs[cluster]))
148     for(j in 1:length(cluster)){
149         pos.obs[j,(1:nobs[cluster[j]])] <- (sum(nobs[1:(cluster[j]-1)])+1):(
                sum(nobs[1:(cluster[j]-1)])+nobs[cluster[j]])
150         if(cluster[1]==1) pos.obs[1,] <- c(1:nobs[cluster[1]], rep(0,(max(
                nobs[cluster])-nobs[cluster[1]])))
151     }
152     NROW(mypairs)/NROW(mypairs[mypairs[,1] %in% pos.obs & mypairs[,2] %in%
            pos.obs,])
153 }
```

Bibliografia

Adler, Robert J. (2008). "Some new random field tools for spatial analysis". In: *Stochastic Environmental Research and Risk Assessment*. Vol. **22**. Stochastic Environmental Research e Risk Assessment, pp. 809–822.

Albert, Paul S. e Lisa M. McShane (1995). "A Generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data". In: *Biometrics* **51**, pp. 627–638.

Azzalini, Adelchi (1983). "Maximum likelihood estimation of order m for stationary stochastic processes". In: *Biometrika* **70**, pp. 381–387.

Bai, Yun, Jian Kang e Peter X.K. Song (2014). "Efficient pairwise composite likelihood estimation for spatial-clustered data". In: *Biometrics* **70**, pp. 661–670.

Bates, Douglas e Martin Maechler (2018). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-14.

Besag, Julian (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* **36**, pp. 192–225.

— (1977). "Efficiency of Pseudolikelihood Estimation for Simple Gaussian Fields". In: *Biometrika* **64**, pp. 616–618.

Bevilacqua, Moreno, Federico Crudu e Emilio Porcu (2015). "Combining Euclidean and composite likelihood for binary spatial data estimation". In: *Stochastic Environmental Research and Risk Assessment* **29**, pp. 335–346.

Bevilacqua, Moreno e Carlo Gaetan (2015). "Comparing composite likelihood methods based on pairs for spatial Gaussian random fields". In: *Statistics and Computing* **25**, pp. 877–892.

Bevilacqua, Moreno, Carlo Gaetan, Jorge Mateu e Emilio Porcu (2012). "Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach". In: *Journal of the American Statistical Association* **107**, pp. 268–280.

Carey, Vincent, Scott L. Zeger e Peter Diggle (1993). "Modelling multivariate binary data with alternating logistic regressions". In: *Biometrika* **80**, pp. 517–526.

Cattelan, Manuela e Cristiano Varin (2018). "Marginal logistic regression for spatially clustered binary data". In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* **67**, pp. 939–959.

Chandler, Richard E. e Steven Bate (2007). "Inference for clustered data using the independence loglikelihood". In: *Biometrika* **94**, pp. 167–183.

Clements, Archie C.A., Nicholas J.S. Lwambo, Lynsey Blair, Ursuline Nyandindi, Godfrey Kaatano, Safari Kinung'hi, Joanne P. Webster, Alan Fenwick e Simon Brooker (2006). "Bayesian spatial analysis and disease mapping: Tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania". In: *Tropical Medicine and International Health* **11**, pp. 490–503.

Cox, D. R. e N. Reid (1987). "Parameter Orthogonality and Approximate Conditional Inference". In: *Journal of the Royal Statistical Society: Series B (Methodological)* **49**, pp. 1–18.

— (2004). "A note on pseudolikelihood constructed from marginal densities". In: *Biometrika* **91**, pp. 729–737.

Cressie, Noel (1991). "Statistics for spatial data". In: *Terra Nova* **4**, pp. 613–617.

Davis, Richard A. e Chun Yip Yau (2011). "Comments on pairwise likelihood in time series models". In: *Statistica Sinica*. Vol. **21**. Institute of Statistical Science, Academia Sinica, pp. 255–277.

Diggle, P, P J Diggle, P Heagerty e KY Liang (2002a). *Analysis of longitudinal data, 2nd Edition*. Clarendon Press.

Diggle, P J et al. (2007). "Annals of Tropical Medicine & Parasitology Spatial modelling and the prediction of Loa loa risk: decision making under un-

certainty". In: *Annals of Tropical Medicine & Parasitology* **101**, pp. 499–509.

Diggle, Peter, Rana Moyeed, Barry Rowlingson e Madeleine Thomson (2002b). "Childhood malaria in the Gambia: A case-study in model-based geostatistics". In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* **51**, pp. 493–506.

Diggle, Peter. e Paulo J. Ribeiro (2007). *Model-based geostatistics*. Springer.

Diggle, PJ (1990). "Time series; a biostatistical introduction". In: *Biometrics* **49**, p. 1286.

Emrich, Lawrence J. e Marion R. Piedmonte (1991). "A method for generating high-dimensional multivariaten binary variates". In: *American Statistician* **45**, pp. 302–304.

Geys, Helena, Geert Molenberghs e Louise M. Ryan (1999). "Pseudolikelihood Modeling of Multivariate Outcomes in Developmental Toxicology". In: *Journal of the American Statistical Association* **94**, pp. 734–745.

Gilbert, Paul e Ravi Varadhan (2019). *numDeriv: Accurate Numerical Derivatives*. R package version 2016.8-1.1.

Heagerty, Patrick J. e Subhash R. Lele (1998). "A composite likelihood approach to binary spatial data". In: *Journal of the American Statistical Association* **93**, pp. 1099–1111.

Heagerty, Patrick J. e Thomas Lumley (2000). "Window Subsampling of Estimating Functions with Application to Regression Models". In: *Journal of the American Statistical Association* **95**, pp. 197–211.

Heagerty, Patrick J. e Scott L. Zeger (1998). "Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses". In: *Journal of the American Statistical Association* **93**, pp. 150–162.

Hurvich, Clifford M. e Chih Ling Tsai (1989). "Regression and time series model selection in small samples". In: *Biometrika* **76**, pp. 297–307.

Joe, Harry e Youngjo Lee (2009). "On weighting of bivariate margins in pairwise likelihood". In: *Journal of Multivariate Analysis* **100**, pp. 670–685.

Karim, Ali Mehryar, Kesetebirhane Admassu, Joanna Schellenberg, Hibret Alemu, Nebiyu Getachew, Agazi Ameha, Luche Tadesse e Wuleta Be-

temariam (2013). "Effect of Ethiopia's Health Extension Program on Maternal and Newborn Health Care Practices in 101 Rural Districts: A Dose-Response Study". In: *PLoS ONE* **8**.

Kaufman, Cari G., Mark J. Schervish e Douglas W. Nychka (2008). "Covariance tapering for likelihood-based estimation in large spatial data sets". In: *Journal of the American Statistical Association* **103**, pp. 1545–1555.

Kuk, Anthony Y C (2007). "A hybrid pairwise likelihood method". In: *Biometrika* **94**, pp. 939–952.

Lehmann, E L e George Casella Springer (1983). *Theory of Point Estimation, Second Edition*. Rapp. tecn.

Lin, Pei Sheng (2008). "Estimating equations for spatially correlated data in multi-dimensional space". In: *Biometrika* **95**, pp. 847–858.

Lin, Pei Sheng e Murray K. Clayton (2005). "Analysis of binary spatial data by quasi-likelihood estimating equations". In: *Annals of Statistics* **33**, pp. 542–555.

Lindsay e B. G. (1988). "Composite Likelihood Methods". In: *Contemporary Mathematics* **80**, pp. 221–239.

Lipsitz, Stuart R., Nan M. Laird e David P. Harrington (1991). "Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association". In: *Biometrika* **78**, pp. 153–160.

Mardia, K. V. (1967). "Some contributions to contingency-type bivariate distributions." In: *Biometrika* **54**, pp. 235–249.

Molenberghs, Geert. e Geert. Verbeke (2005). *Models for discrete longitudinal data*. Springer Science+Business Media, Inc.

Pace, Luigi, Alessandra Salvan e Nicola Sartori (2011). "Adjusting composite likelihood ratio statistics". In: *Statistica Sinica*. Vol. **21**. Institute of Statistical Science, Academia Sinica, pp. 129–148.

Padoan, SA e M Bevilacqua (2015). "Analysis of random fields using comprandfld". In: *Journal of Statistical Software* **63**, pp. 1–27.

Palmgren, Juni (2011). *Regression Models for Bivariate Binary Responses*. Rapp. tecn., pp. 11–12. URL: `http://biostats.bepress.com/uwbiostat/paper101`.

Pebesma, Edzer J. e Roger S. Bivand (2005). "Classes and methods for spatial data in R". In: *R News* **5**, pp. 9–13.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rotnitzky, Andrea e Nicholas P. Jewell (1990). "Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data". In: *Biometrika* **77**, pp. 485–497.

Thomson, Madeleine C., Stephen J. Connor, Umberto D'Alessandro, Barry Rowlingson, Peter Diggle, Mark Cresswell e Brian Greenwood (1999). "Predicting malaria infection in Gambian children from satellite data and bed net use surveys: The importance of spatial correlation in the interpretation of results". In: *American Journal of Tropical Medicine and Hygiene* **61**, pp. 2–8.

Varin, Cristiano (2007). "On composite marginal likelihoods". In: *Advances in Statistical Analysis. A Journal of the German Statistical Society*. Vol. **92**. Advances in Statistical Analysis. A Journal of the German Statistical Society, pp. 1–28.

Varin, Cristiano, Nancy Reid e David Firth (2011). "An overview of composite likelihood methods". In: *Statistica Sinica*. Vol. **21**. Institute of Statistical Science, Academia Sinica, pp. 5–42.

Zeger, Scott L., Kung-Yee Liang e Paul S. Albert (1988). "Models for Longitudinal Data: A Generalized Estimating Equation Approach". In: *Biometrics* **44**, pp. 1049–1060.