Scuola di Dottorato in Scienze Economiche e Statistiche

Dottorato di Ricerca in
Metodologia Statistica per la Ricerca Scientifica
XIX ciclo

Alma Mater Studiorum - Università di Bologna

Analisi spaziale della longevità in Emilia-Romagna

Massimiliano Marino

Tutor: Prof.ssa Paola Monari

Co-tutor: Prof.ssa Rossella Miglio

Coordinatore: Prof.ssa Daniela Cocchi

Settore disciplinare: SECS-S/01

Dipartimento di Scienze Statistiche "P. Fortunati"
Marzo 2008

A me stesso

Indice analitico

Introduzione

Negli ultimi anni la longevità è divenuto un argomento di notevole interesse in diversi settori scientifici. Le ricerche volte ad indagare i meccanismi che regolano i fattori della longevità si sono moltiplicate nell'ultimo periodo interessando, in maniera differente, alcune regioni del territorio italiano. Lo studio presentato nella tesi ha l'obiettivo di identificare eventuali aggregazioni territoriali caratterizzate da una significativa propensione alla longevità nella regione Emilia-Romagna mediante l'impiego di quattro metodologie di *clustering* spaziale, alcune delle quali di recente implementazione.

La tesi è suddivisa in 4 capitoli. Il primo capitolo comprende la descrizione dello studio e dei materiali utilizzati per la ricerca; il secondo e terzo capitolo descrivono le metodologie di *cluster detection* impiegate per l'identificazione delle aggregazioni spaziali ad elevata longevità mentre il quarto capitolo fornisce una descrizione dettagliata dei risultati ottenuti e le conclusioni derivate dalle analisi effettuate.

Capitolo 1

Longevità e mortalità

Introduzione

Nel corso degli ultimi 150 anni, dall'epoca dell'unificazione nazionale ad oggi, la durata media della vita in Italia è passata da 35 ad 80 anni. Il processo di riduzione della mortalità, alla base di questo straordinario processo, è stato studiato in maniera approfondita indagando su alcune delle cause prevalenti legate sia ai progressi della medicina che alle mutate condizioni ambientali e di vita, come i miglioramenti nell'alimentazione ed il progressivo innalzamento del tenore di vita delle popolazioni. Il processo di riduzione della mortalità non si è sviluppato in modo uniforme sul territorio ma ha interessato, in periodi differenti, maggiormente alcune classi di età rispetto ad altre e con differenze di genere. E' ben noto che all'inizio della transizione demografica e sanitaria, nella seconda metà del XIX secolo, le differenze tra i due sessi non erano così accentuate: nelle età infantili (dopo il primo anno di nascita) e nelle età riproduttive, vi era una maggiore prevalenza di mortalità femminile. Nel corso della transizione tale evidenza è radicalmente mutata consentendo alla popolazione femminile in Italia, come nel resto dei paesi sviluppati, di godere di una maggiore durata della vita che (in media) supera quella maschile di almeno cinque anni. Negli ultimi decenni, si è dedicata particolare attenzione alla riduzione della mortalità nelle età anziane e alla crescente proporzione di popolazione, in prevalenza donne, che raggiungono età molto avanzate. La longevità è un fenomeno in crescente evidenza, rappresentando un argomento di interesse in diversi settori di ricerca, che ha favorito, nel corso degli anni, la nascita di studi volti ad indagare i meccanismi che regolano i fattori della longevità. Con il termine invecchiamento il senso comune descrive intuitivamente una serie di manifestazioni tipiche dell'età avanzata, quali la perdita delle funzioni fisiche dell'organismo, una scarsa capacità di adattamento ai fattori ambientali e un'incrementata suscettibilità alle patologie cronico-degenerative. Non è facile dare una definizione rigorosa di invecchiamento, considerata la stretta correlazione tra fattori ambientali, biologici e culturali, ma è possibile darne una definizione classica: "*per tutti gli organismi a riproduzione sessuata in cui si verifica, l'invecchiamento è un insieme di cambiamenti strutturali e funzionali che fa seguito al raggiungimento della maturità riproduttiva, che si traduce in una*

diminuzione della capacità di adattamento, avente come risultato l'aumento della probabilità di morte con l'avanzare dell'età."

Vista da una prospettiva diversa, la senescenza può indicare la propensione di un individuo alla longevità; va da sé che una ridotta mortalità, in particolare nelle classi più anziane, aumenta la possibilità di sopravvivenza fino ad età elevate. Ad ogni modo l'invecchiamento è un fenomeno difficile da inquadrare e la mancanza di una sua definizione univoca ne è una prova evidente. Ancora più difficile è il tentativo di interpretare il fenomeno, ovvero di capire *come* e *perché* si invecchia. Esistono diverse teorie a riguardo spesso in conflitto tra loro. Tra le quelle storiche si collocano la "teoria dell'invecchiamento programmato" e la "teoria del danno accumulato". La prima postula che la senescenza è un processo determinato geneticamente, una naturale prosecuzione dello sviluppo umano, che diventa una forma di caratteristica imposta dalla selezione naturale per apportare un beneficio a tutta la specie a scapito della *fitness* di un individuo anziano. Tale teoria è stata criticata dai biologi e dai medici per il suo contrasto con la teoria della selezione naturale di Darwin ma, nell'aspetto relativo alla genericità del fenomeno, rappresenta ancora oggi un punto di riferimento ed oggetto di studi specifici. La seconda teoria considera, invece, la senescenza non un adattamento ma, bensì, un difetto nell'evoluzione di un individuo derivante da una serie di errori accumulati nel corso della vita, influenzati dall'ambiente, dal metabolismo e dall'effetto dell' entropia. Negli anni '50, il mondo scientifico riteneva l'invecchiamento un "problema irrisolto della natura" e ciò ha favorito la nascita di nuove correnti di pensiero che prendono il nome di teorie tradizionali dell'invecchiamento. Esse cercano di mantenere sia un grado di compatibilità con la teoria di Darwin che spiegare la variabilità della durata della vita tra la specie umana. Queste teorie sono anche dette teorie evoluzionistiche non adattive perché combinano la selezione naturale con il concetto di danno accumulato. La selezione naturale spiega perché gli animali vivono abbastanza a lungo da riprodursi mentre il danno accumulato spiega perchè essi invecchiano dopo il periodo riproduttivo, quando tale fenomeno ha una scarsa influenza sulla *fitness* individuale. Tra le teorie tradizionali vi è quella della "pleiotropia antagonista" (Williams,1957) secondo la quale esiste un'influenza dei geni nella manifestazione fenotipica di un individuo,

considerando l'invecchiamento un inevitabile effetto collaterale della riproduzione.

L'approccio alla senescenza delle teorie tradizionali non spiega in maniera esaustiva le modalità di invecchiamento. L'individuazione e la conoscenza dei processi biologici, che si trovano alla base del fenomeno, faciliterebbero la creazione di una "terapia" che permetterebbe di affrontare il problema in modo mirato. Alla fine degli anni '50, nascono le teorie indicate come teorie molecolari o casuali, secondo le quali la senescenza è il frutto di un processo di deterioramento di un individuo a livello cellulare o molecolare. Tra esse la "teoria dei radicali liberi" secondo la quale il processo di deterioramento del nostro corpo è dovuto alla sua reazione alla presenza di radicali liberi dell'ossigeno al nostro interno. La teoria ipotizza che la reattività ai radicali liberi non è ereditaria ma è il risultato di un danno cumulativo legato al tempo; la presenza di radicali liberi nel nostro corpo provoca alterazioni a livello cellulare scatenando modificazioni strutturali e funzionali potenzialmente irreversibili. Le aggressioni al nostro organismo da parte di agenti esterni contribuiscono alla formazione di radicali liberi tant'è che l'inquinamento dell'aria, il fumo, i farmaci, i raggi ultravioletti sono annoverati tra i fattori che possono contribuire alla loro formazione, accelerando il processo di senescenza delle nostre cellule. Tuttavia, alla fine degli anni '80, anche le teorie casuali dell'invecchiamento iniziano ad essere criticate e smentite con studi specifici. Nascono così le teorie sistemiche dell'invecchiamento che provano a ricercare le cause principali della senescenza ad un livello più alto di organizzazione biologica, occupandosi delle interazioni tra cellule e molecole e studiando il funzionamento di apparati e sistemi endogeni in fase di deterioramento. Nessuna delle ipotesi sull'invecchiamento è in grado di spiegare da sola un fenomeno così complesso. Negli ultimi anni si è cercato di fornire una tesi che riunisse i diversi aspetti delle ipotesi formulate definendo quella che gli studiosi definiscono "*network* dell'invecchiamento". Si ammette, pertanto, che per contrastare i meccanismi lesivi, sia endogeni che esogeni, si mettano in moto una serie di fenomeni cellulari a livelli bassi di organizzazione biologica (membrana cellulare e acidi nucleici) che, nel loro complesso, sono appunto detti "*network* anti-invecchiamento", destinate al mantenimento dell'omeostasi e alla riparazione

dei danni biologici. Secondo questa nuova ipotesi, l'invecchiamento non è geneticamente programmato perché sfavorevole ma è la longevità ad essere sotto controllo genetico in qualche sua parte. Questa recente prospettiva del fenomeno, oltre a spiegare la longevità come specie-specifica e familiare, consente di indirizzare efficacemente le ricerche genetiche, mediche e statistiche, verso lo studio della longevità piuttosto che dell'invecchiamento.

1.1 - Descrizione dello studio

Negli ultimi anni, l'Italia si è collocata tra i primi posti al mondo come paese con la più alta percentuale di popolazione longeva. Nel nostro territorio, la longevità è un fenomeno che ha assunto una dimensione crescente negli ultimi decenni interessando, in misura diversa, alcune zone del nostro Paese. La distribuzione territoriale dell'invecchiamento appare assai diversificata tra le regioni così come all'interno di esse. E' ormai noto, come riportato in numerosi studi scientifici, che la Sardegna è la regione italiana in cui la longevità si manifesta in modo chiaro in alcune aree comunali del suo territorio assumendo un rilevante interesse scientifico. Di recente altre regioni, quali la Lombardia, il Veneto, la Sicilia e la Calabria, sono state oggetto di analisi relative all'identificazione di ultra centenari sul proprio territorio (Robine *et al*.,2006). La regione Emilia-Romagna, che si colloca in una posizione intermedia nella graduatoria italiana per la presenza di centenari con 1.57 casi ogni 10.000 abitanti, non è stata ancora oggetto di analisi specifiche ed approfondite a riguardo. A tale scopo è stato sviluppato uno studio di ricerca multi-disciplinare che, nell'intento di cogliere a pieno l'analisi di un fenomeno complesso come quello della longevità, coinvolge al suo interno diversi ambiti scientifici quali la demografia, la medicina, la genetica e la statistica. L'idea è nata dalla domanda: "*C'è qualcosa di interessante o inatteso sul territorio della nostra Regione in merito alla longevità? E se si, dove e come si può localizzare?"*. L'obiettivo dello studio è consentire un'analisi integrata del fenomeno tesa, in prima istanza, ad ottenere una rappresentazione spaziale della longevità nella regione Emilia-

Romagna, mediante l'identificazione di aree territoriali o aggregazioni di esse caratterizzate da un'evidente espressione dell'invecchiamento demografico, e, in un secondo tempo, ad individuarne i possibili fattori determinanti. L'identificazione di *clusters* spaziali di individui longevi rappresenta un punto di partenza per approfondire, in indagini successive e in modo mirato, gli eventuali fattori socio-demografici, genetici ed economici che hanno caratterizzato tale distribuzione. Per raggiungere l'obiettivo prefissato si è scelto di utilizzare alcune metodologie di *clustering* spaziale, basate sulla teoria della massima verosimiglianza, che si differenziano tra loro per la modalità di ricerca dei potenziali *clusters*: la differenza principale consiste nella capacità di identificare aggregazioni territoriali di forma regolare (*spatial scan statistic*) o dall'andamento geometrico "libero" (*flexible scan statistic,* algoritmo genetico e *greedy growth search*). Le caratteristiche di ciascuna metodologia consentono, in tal modo, di "catturare" le possibili conformazioni geografiche delle aggregazioni presenti sul territorio e la teoria statistica di base, comune ad esse, permette di effettuare agevolmente un confronto tra i risultati ottenuti. La persistenza di un'area della regione caratterizzata da un'elevata propensione alla longevità consente di ritenere il *cluster* identificato di notevole interesse per approfondimenti successivi. La specificità del territorio esaminato, relazionato ad un fenomeno di emergente interesse scientifico come la longevità, rappresenta un aspetto innovativo della ricerca e l'impiego di metodologie di *cluster detection* di recente implementazione aggiungono una caratteristica di ulteriore interesse allo studio.

1.2 - Confronto tra *clusters*

Le metodologie di *cluster detection* impiegate nello studio si differenziano tra loro per le specifiche modalità di definizione dei *clusters* e la soggettività insita nella scelta dei parametri di ricerca può favorire l'identificazione di aree geografiche non perfettamente sovrapponibili e di difficile interpretazione. La definizione di un criterio di confronto tra i diversi risultati rappresenta un ulteriore

obiettivo dello studio ed ha lo scopo di delineare, in maniera più efficace, i confini di un possibile *cluster*. Il criterio di valutazione utilizzato è stato derivato dalla teoria dei grafi, con particolare riferimento ai multigrafi, spesso utilizzati per la classificazione di unità statistiche caratterizzate da numerose variabili, non direttamente confrontabili, in cui ciascun elemento viene rappresentato come vertice di un grafo. L'idea è di confrontare, a parità di parametri di ricerca, i grafi associati ai *clusters* identificati con le diverse metodologie attraverso una valutazione delle occorrenze dei collegamenti esistenti tra le coppie di vertici.

Si definisce multigrafo G_m una coppia (V, E_m) costituita, rispettivamente, da un insieme non vuoto di vertici $V = \{v_1, v_2, ...v_n\}$ e da un insieme multiplo $E_m = \{(e_i, e_j)_t\}$, $1 \le i, j \le n, i \ne j$ e $t = 1, ...m$, di coppie distinte e non ordinate di vertici dette lati. Il valore m identifica la molteplicità totale del multigrafo mentre il numero complessivo n di vertici ne determina l'ordine: una qualsiasi coppia di vertici distinti può essere, quindi, collegata da un numero massimo m di lati. Un multigrafo G_m può essere proiettato in un grafo ${}_pG_m = (V, E_m)$, detto *p*-proiezione di G_m, avente come insieme di lati ${}_pE_m = \{(e_i, e_j) : |\{(e_i, e_j)_1, (e_i, e_j)_2, ...(e_i, e_j)_m \cap E_m, \} \ge p|\}$, $1 \le p \le m$, e come insieme di vertici $V = {}_pV$. Da ciò si deduce che due vertici appartenenti al grafo ${}_pG_m$ risultano collegati da un lato se, e solo se, gli stessi vertici del multigrafo originario risultano connessi da almeno *p* lati.

Nel nostro studio, il multigrafo originario è rappresentato dall'insieme dei centroidi delle aree comunali e dall'insieme dei rispettivi collegamenti mentre la molteplicità m è indicata dal numero di *clusters* a confronto. Una coppia di vertici è considerata persistente se il collegamento ad esse associato si presenta almeno un numero $k(\le m)$ di volte tra i *clusters*. L'occorrenza di un collegamento viene determinata utilizzando le matrici di adiacenza definite dai *clusters*. Considerato un territorio suddiviso in I aree distinte, la matrice di adiacenza A è una matrice simmetrica, di dimensione $(I \times I)$, in cui un generico elemento a_{ij} assume valore 1 in corrispondenza di due aree adiacenti e valore 0 in tutti gli altri casi. In tal modo se, sommando le matrici di adiacenza

dei *clusters* a confronto, gli "1" relativi ad un collegamento risultano superiori o uguali al valore prefissato k, la coppia di vertici associata è considerata persistente. La soglia k è definita a priori e, nella nostra ricerca, assume un valore intero compreso tra 1 e 5[1] che rappresenta la *p*-proiezione del multigrafo; fissando k =4, un collegamento risulta persistente se è presente in almeno 4 aggregazioni su 5. Il criterio utilizzato introduce un ulteriore elemento di soggettività nella fase di definizione del *cluster* legato alla scelta del valore k ma si ritiene che tale approccio possa limitare, in qualche modo, l'effetto legato alla diversità geografica dei *clusters* finali consentendo un confronto più appropriato tra i risultati.

1.3 - Materiali e Metodi

La popolazione in esame è costituita dagli individui residenti in Emilia-Romagna nel quinquennio 2000-2004 (5 anni di calendario) suddivisa in classi di età, sesso e comune. L'analisi è di tipo puramente spaziale, in cui l'unità geografica elementare è identificata dal comune, ed è stata condotta separatamente per i due sessi, in quanto è noto dalla letteratura che numerosi fenomeni epidemiologici e demografici, si manifestano in modi e misure diverse nei due sessi. I valori di popolazione totale, maschile e femminile, sia residente che longeva, sono stati determinati come media dei valori annuali nel periodo di osservazione.

Alcune valutazioni di carattere demografico ed un esame della letteratura esistente sugli studi di longevità hanno indotto alla definizione di due classi (aperte) di età per rappresentare il fenomeno nella nostra ricerca. La prima classe comprende gli individui con età superiore o uguale a 95 anni (indicata con 95+) mentre la seconda, ancora più restrittiva, comprende gli individui con età superiore o uguale a 100 anni (indicata con 100+). Nella fase iniziale dello studio, è stata individuata anche una terza classe comprendente gli individui con

[1] A parità di condizioni sperimentali, viene scelto un *cluster* per ogni metodologia tranne nel caso della *spatial scan statistic* in cui si utilizzano sia *clusters* circolari che ellittici

età superiore o uguale a 105 anni ma, considerata l'esigua numerosità di tale popolazione, è stata immediatamente abbandonata. I risultati preliminari hanno indotto, successivamente, ad escludere anche la classe 100+, seppur di notevole interesse scientifico, per la scarsa numerosità degli individui, in particolare per il sesso maschile. Di conseguenza, gli individui ritenuti longevi ai fini del nostro studio sono rappresentati dalla popolazione media residente, in ciascun comune, con età superiore o uguale a 95 anni.

Un ulteriore argomento di interesse dello studio è la scelta della misura di sintesi impiegata per analizzare il fenomeno sul territorio. Si tratta di un indicatore specifico di longevità, mutuato dalla demografia, indicato con *Centenarian Rate* (CR) (Robine e Caselli,2005). Esso è stato utilizzato in uno studio europeo sul *trend* nell'ultimo secolo di ultra centenari presenti in 12 Stati europei tra cui l'Italia. L'idea alla base dell'indicatore è confrontare la popolazione longeva in un istante temporale con quella presente, nella stessa area, circa 40 anni prima dell'osservazione. In genere, le misure di sintesi comunemente utilizzate per analizzare la longevità sono espresse da un tasso grezzo, ottenuto dal rapporto tra gli individui longevi e la popolazione totale residente in una specifica area, o dal rapporto tra gli individui longevi ed il numero di nascite nella stessa area. Nel primo caso non si può effettuare un confronto tra aree o popolazioni diverse a meno di opportune procedure di standardizzazione mentre, nel secondo caso, non si tiene conto dell'effetto delle migrazioni di una popolazione. Il CR è stato costruito con lo scopo di superare tali limitazioni ipotizzando che l'effetto migratorio di una popolazione possa ritenersi trascurabile oltre i 60 anni di età. In Italia, le migrazioni sono state un fenomeno particolarmente evidente tra coloro che oggi costituiscono la popolazione longeva ed un probabile ritorno nella regione di nascita avveniva, in genere, entro i 60 anni di età quando, a metà del nostro secolo, l'attività professionale di un individuo si concludeva prima di questo limite di età (Toutain,2001; Caselli,2001).

Nel nostro caso, il CR assume il ruolo di un indicatore epidemiologico in quanto il numeratore identifica i casi "osservati" mentre il denominatore, costituito dalla popolazione residente 40 anni prima nello stesso comune, rappresenta la popolazione "a rischio". Seguendo tale approccio si ipotizza che la popolazione

presente in un territorio 40 anni prima risulti esposta al "rischio" di essere longeva o, analogamente, che gli individui longevi, osservati in uno specifico istante, derivino dalla coorte di individui presenti 40 anni prima nello stesso territorio. Valori elevati dell'indicatore indicano una maggiore propensione alla longevità della popolazione mentre valori prossimi allo zero identificano una tendenza contraria.

L'indicatore è stato calcolato distintamente per sesso ed il denominatore, ottenuto dalle tavole ISTAT del 1961 suddivise per sesso, fascia di età e comune, è costituito dagli individui appartenenti alla classe di età 55–59 anni; a differenza del citato studio europeo, al numeratore è stata considerata una classe di età aperta piuttosto che un'età singola. Per gli individui con età 95+, il CR può essere espresso dal rapporto[2]:

$$\frac{\text{Pop. media residente in un comune con età } \geq 95 \text{ anni}}{\text{Pop. residente nello stesso comune al censimento del 1961 (età 55-59 anni)}}$$

[2] Per gli individui 100+, il denominatore è costituito dalla popolazione del 1961 con età 60- 64 anni

Capitolo 2

Metodi di analisi spaziale in epidemiologia

Introduzione

L'epidemiologia spaziale consente di descrivere ed analizzare le informazioni di carattere sanitario indicizzate geograficamente in relazione a fattori di rischio derivanti, ad esempio, dalla demografia, dall'ambiente, dalle abitudini comportamentali, dagli aspetti sociali ed economici e dalla genetica (Elliot e Wartenberg,2004). Essa è parte di una lunga tradizione di analisi di dati geografici quando, già a partire dalla fine dell'800, in alcuni paesi iniziavano ad essere diffuse le prime mappe dei tassi di rischio per la conoscenza e la caratterizzazione delle cause eziologiche di specifiche malattie quali la febbre gialla ed il colera (Walter,2000). L'epidemiologia spaziale ha esteso e continuato la ricca tradizione degli studi ecologici esistenti aventi come obiettivo finale l'individuazione della distribuzione spaziale delle malattie in un ambito territoriale (Doll,1980; Keys,1980). Recenti e rapidi sviluppi delle metodologie statistiche per l'analisi di dati epidemiologici, unitamente alla maggiore disponibilità di informazioni sanitarie e di indicatori epidemiologici geografici, hanno fortemente contribuito alla diffusione di tecniche di investigazione spaziale del rischio di malattie o dei fenomeni ad esse correlate. Si distinguono essenzialmente quattro tipi di studi di epidemiologia spaziale (Elliot *et al.*,2000):

- mappe di rischio *(disease mapping)*
- studi di correlazione geografica *(geographical correlation studies)*
- stima del rischio in relazione a sorgenti puntuali o lineari *(point and point-line source studies)*
- individuazione e definizione di *clusters* spaziali di malattie *(cluster detection)*

Il *disease mapping* consente di ottenere una rapida visione del fenomeno indagato attraverso le mappe di variazione spaziale (o spazio-temporale) del rischio di una malattia e di identificare eventuali andamenti anomali dei dati che altrimenti sfuggirebbero in una semplice rappresentazione tabellare. L'obiettivo del *disease mapping* è fornire indicazioni immediate e di carattere generale per elaborare delle ipotesi investigative da approfondire successivamente in studi

specifici. Nell'ambito della programmazione sanitaria, lo studio delle mappe di rischio consente di allocare in modo mirato le risorse sociali ed economiche destinate alla cura della salute della popolazione di un territorio. Il *disease mapping* fornisce, dunque, una visione spaziale di un fenomeno attraverso alcuni indicatori relazionati al fenomeno stesso. La misura di sintesi comunemente impiegata è il Rapporto Standardizzato di Mortalità (SMR, *Standardized Mortality Ratio*) ovvero il rapporto tra i casi osservati ed i casi attesi di una malattia per una specifica zona e popolazione. Una mappa graduata secondo una scala di colori opportunamente selezionati consente di ottenere un'immediata "fotografia" del fenomeno identificando, in dettaglio, le aree caratterizzate da un livello significativo di rischio. Benché le mappe consentano un'efficace visione di insieme del fenomeno, occorre porre attenzione nella loro interpretazione in particolare in studi su piccole aree. La scelta di elementi di carattere tecnico, come la scala di rappresentazione ed i colori della scala stessa, possono condizionare l'interpretabilità dei risultati ottenuti. L'omogeneità dei gruppi aggregati nell'analisi è un elemento fondamentale per una rappresentazione corretta dei risultati; ad esempio, l'impiego di differenti scale di misura e di strategie di aggregazione dei dati possono condurre alla creazione di mappe altrettanto valide ma riportanti caratteristiche diverse dei dati generando quello che, nella letteratura geografica, si chiama problema della "unità areale modificabile" (Openshaw,1984). Esso è la manifestazione geografica della fallacia ecologica in cui le conclusioni, basate su dati aggregati di un particolare insieme di aree, possono cambiare se si aggregano gli stessi dati in un insieme diverso di aree. Inoltre, le variazioni dei tassi tra le aree di un territorio possono derivare sia da una scarsa qualità delle informazioni, rilevate mediante un'errata classificazione degli indicatori sanitari, che da errori di assegnazione delle unità attraverso i sistemi di informazione geografica (GIS). La presenza di tali errori comporta una classificazione non corretta della popolazione a rischio o dei casi osservati, influenzando in maniera significativa le stime dei relativi tassi di rischio.

Gli studi di correlazione geografica hanno come obiettivo principale l'analisi delle variazioni geografiche derivanti dall'esposizione di un individuo, appartenente ad una popolazione di riferimento, a fattori ambientali (aria, terra e

acqua), a fattori legati agli stili di vita (fumo e alimentazione) o a fattori socio-economici e demografici (reddito e sesso), in relazione ad indicatori di salute misurati su scala geografica o ecologica. Questo approccio consente di utilizzare informazioni già disponibili, raccolte di routine, che possono essere impiegate per investigare esperimenti naturali dove l'esposizione ha una base fisica, come il suolo o l'acqua. La differenza con il metodo precedente consiste nel fatto che il *disease mapping* è principalmente utilizzato per scopi descrittivi mentre gli studi ecologici sono focalizzati su questioni eziologiche.

Gli studi *point and point-line* sono generalmente impiegati in indagini volte ad accertare la presenza di un rischio aumentato nei pressi di sorgenti sospette o dove una sorgente è ritenuta una potenziale fonte di rischio ambientale. L'esposizione può essere di natura puntuale, come in presenza di un inceneritore o di un traliccio radio-trasmittente, o di natura lineare come nel caso di strade o di sorgenti elettriche. In situazioni simili, qualsiasi incremento dell'esposizione dovuto alle sorgenti di rischio può estendersi ad un'area limitrofa ed uno studio focalizzato, con una risoluzione geografica idonea, consente di ottenere una stima del rischio associato.

Le metodologie di *disease cluster o general clustering* consentono di ottenere una prima informazione sul rischio di una malattia quando non si hanno conoscenze a priori sull'eziologia della malattia stessa.

La distinzione tra i quattro tipi di analisi presentati è in qualche modo legata all'aspetto pratico dell'indagine; ad esempio, una buona mappa di incidenza di una malattia gioca spesso un ruolo preliminare in studi di *clustering*. Le mappe di rischio includono solitamente relazioni con specifiche covariate che descrivono i fattori di rischio noti mentre uno studio ambientale può essere il preludio ad una indagine finalizzata all'analisi della relazione esistente tra l'incidenza della malattia ed i fattori di rischio. La classificazione nelle quattro categorie consente di ricoprire molteplici obiettivi di indagini in ambito epidemiologico pur sussistendo, talvolta, una sovrapposizione tra le diverse metodologie. Il *disease mapping* fornisce informazioni sia per la definizione di *clusters* individuali che, in senso più generale, sulla tendenza dei casi ad aggregarsi sul territorio; gli studi epidemiologici sul rischio in eccesso, dovuto alla presenza di fonti lineari o puntuali, sono spesso basati sulle informazioni

aggregate a livello di area a causa della scarsa risoluzione geografica dei dati disponibili; ciò implica che le esposizioni a sorgenti di rischio vengano assunte costanti su piccole aree ipotizzando una distribuzione uniforme degli individui nello spazio e nel tempo che nella realtà spesso non si verifica: un individuo nasce in una specifica zona e in una determinata data e dipende (in termini probabilistici) dalla struttura e dalla densità della popolazione di appartenenza nel periodo di nascita. Ciascun individuo si muove nello spazio e durante la sua vita sarà soggetto a fenomeni di movimento, quali le migrazioni, esponendosi a più fattori di rischio specifici delle aree geografiche di appartenenza. Le caratteristiche individuali, come sesso ed età nonché i fattori genetici e gli stili di vita, possono influenzare la sopravvivenza di un individuo e contribuire allo sviluppo di malattie future.

Un insieme ideale di dati dovrebbe comprendere informazioni dettagliate sulla popolazione in esame, quali le sue caratteristiche individuali, i movimenti demografici, le esposizioni ad eventuali sorgenti di rischio e le registrazioni di carattere sanitario. Naturalmente questo tipo di informazione, sia singola che combinata, non è sempre disponibile e di conseguenza le metodologie statistiche provano ad adeguarsi al livello di dettaglio delle informazioni presenti.

Un elemento discriminante nella scelta di un approccio statistico in epidemiologia geografica è la natura dei dati da analizzare; esistono sostanzialmente quattro tipi di dati frequentemente utilizzati in ambito epidemiologico spaziale (Bailey,2001):

- dati derivanti da una griglia irregolare, come nel caso delle classificazioni censuarie o amministrative, riportanti il numero di casi e la popolazione a rischio nonché alcune misure socio-economiche;
- dati di eventi, come la localizzazione di casi individuali di malattie o di persone che fungono da controllo (popolazione a rischio). Possono essere comprese le covariate per ogni individuo;
- dati geostatistici, quali le rilevazioni puntuali di carattere ambientale;
- dati derivanti da un grigliato regolare, simili al caso irregolare, e spesso derivanti da misure di *remote sensing*.

I dati possono essere ulteriormente distinti in dati puntuali e dati di conteggio; per ogni popolazione, tipo di esposizione e genere di informazione sanitaria, si possono associare caratteristiche spaziali e temporali esatte (dati puntuali) oppure caratteristiche aggregate secondo opportuni criteri (dati di conteggio). I dati puntuali rappresentano sicuramente la fonte di informazione più idonea per essere impiegata in un'indagine di dettaglio ma, come spesso accade, risultano di difficile disponibilità.

2.1 - Metodologie di *clustering* spaziale

Nell'ambito dell'epidemiologia spaziale, i termini *cluster detection*, *clustering* e variazione spaziale del rischio sono usati per indicare argomenti differenti ma il tentativo di distinguerli genera spesso confusione (Diggle,2000). Una possibile spiegazione è la mancanza di una definizione formale di *cluster* o di una definizione matematica precisa e compatibile con i metodi statistici utilizzati. Knox (1989) definisce "*un cluster come un insieme di occorrenze limitato geograficamente di dimensione sufficiente e di concentrazione tale da non essere dovuto al solo caso*". Il riferimento alla casualità è rivolto alla verifica di un'ipotesi nulla; in questo contesto, l'assunzione iniziale postula l'indipendenza distributiva dei casi osservati ipotizzando un rischio uniforme per tutti gli elementi di una popolazione. Le due assunzioni riportate nella definizione risultano spesso invalidate in fase di analisi; è possibile che un singolo caso di una malattia, appartenente ad un'area scarsamente popolata, possa esso stesso rappresentare una forte evidenza statistica del fenomeno, violando il requisito di "sufficiente dimensione". La caratteristica di "concentrazione" può essere invalidata, invece, quando si osserva un eccesso di casi, numericamente modesto, all'interno di una regione molto estesa. Quanto detto non implica necessariamente che la definizione di *cluster* di Knox sia errata ma si intende porre l'attenzione sul fatto che non sempre è facile adattarla ad un contesto statistico formale. Una seconda spiegazione alla mancanza di una distinzione netta tra i termini indicati è probabilmente l'uso della definizione di ipotesi nulla; i

termini possono differenziarsi per lo scopo scientifico dell'indagine e per le ipotesi alternative utilizzate. La differenza tra uno studio di *clustering* e uno studio sulla variazione spaziale del rischio può risultare più chiara in questo modo: nel primo si parla di un allontanamento dall'assunzione di indipendenza dei casi osservati sul territorio mentre nel secondo si parla di allontanamento da una distribuzione uniforme del rischio. Il *clustering* invita ad un'interpretazione dei risultati in un'ottica di trasmissione delle malattie o di suscettibilità genetica degli individui di una popolazione. La variazione spaziale del rischio indirizza, invece, verso un'interpretazione di natura ambientale del fenomeno esaminato, notando altresì che le due distinzioni non sono mutuamente esclusive. Di contro, la *cluster detection* si rivolge allo studio delle caratteristiche della distribuzione spaziale delle malattie ed in questo caso una definizione più appropriata potrebbe essere *anomaly detection* o *surveillance*.

2.2 - *Cluster detection*

L'obiettivo principale della *cluster detection* è l'individuazione di aree territoriali anomale o di particolare interesse, in relazione ad uno o più fenomeni presi in esame. Gli andamenti spaziali anomali sono spesso correlati con fattori diversi tra loro e possono dipendere dal settore di applicazione. Lo scopo della ricerca consiste nell'identificazione delle aree di un territorio caratterizzate da un numero di casi osservati significativamente diverso da quello atteso, in relazione ad opportune informazioni di base. Nell'ambito della salute pubblica, ad esempio, l'interesse può essere rivolto all'individuazione di aggregazioni spaziali in cui si osserva un eccesso di malattia o di quantità derivabili da esse, quali il numero di ricoveri ospedalieri o di cure prestate, che possono risultare indicative di emergenze epidemiche localizzate o diffuse in uno specifico territorio. La *cluster detection* risponde a due domande fondamentali: "C*'è qualcosa di interessante o inatteso sul territorio esaminato? E se si, dove e come si può localizzare?".* La risposta può essere articolata in due parti: la prima consente di rappresentare graficamente "quello che ci aspettiamo di vedere" mentre la

seconda permette di definire le zone del territorio che si discostano significativamente dalle nostre aspettative.

Formalizzando quanto detto, si consideri un insieme di punti z_i nello spazio Ω ad ognuno dei quali è associato un conteggio di dati c_i ed un'informazione di base n_i. La definizione specifica dei parametri indicati dipende dal settore di applicazione; ad esempio, i conteggi c_i possono identificare i casi di una malattia in una specifica area mentre la base n_i può essere rappresentata dalla popolazione a rischio nella stessa area. L'obiettivo è individuare quella zona Z dello spazio, formata da un insieme di aree z_i, tale che i conteggi osservati al suo interno risultino significativamente maggiori di quelli attesi, in relazione alla popolazione di riferimento. Il passo successivo consiste nel valutare se l'aggregazione spaziale identificata è statisticamente significativa o semplicemente dovuta al caso mediante un processo statistico inferenziale. L'ipotesi iniziale postula l'assenza di *clusters* del fenomeno sul territorio contro una serie di ipotesi alternative che postulano l'esistenza di almeno una zona dello spazio Ω in cui tale fenomeno si manifesta in modo evidente.

Da decenni ormai, le metodologie statistiche di *clustering* geografico rappresentano un interesse scientifico in crescente e rapido sviluppo sia per le applicazioni in ambito epidemiologico che per settori diversi di ricerca, quali l'archeologia, le scienze agrarie, la veterinaria. I metodi esistenti differiscono tra loro per le modalità di ricerca delle potenziali aree di interesse; i più utilizzati si basano sull'uso di una finestra mobile di scansione, di forma regolare, imposta sul territorio da analizzare e sulle tecniche di confronto statistico dei *clusters* individuati. Le prime applicazioni spaziali in ambito epidemiologico sono state effettuate da Naus nel 1965 che impiegava celle di forma rettangolare o quadrata di dimensione fissa sovrapposte al territorio. Negli anni successivi, il problema della unidimensionalità dell'informazione è stato esteso in varie direzioni e la necessità di investigare fenomeni diversi tra loro, in aggiunta alla disponibilità di informazioni a livello aggregato, ha indirizzato la ricerca verso lo sviluppo di tecniche simili da applicare in un contesto multidimensionale.

Upton e Fingleton (1985) hanno sviluppato due tra i maggiori approcci nell'ambito dell'analisi spaziale puntuale dei dati, entrambi utilizzati nella ricerca

di *clusters* di malattie. Il primo metodo, detto *metodo delle distanze* (alcuni esempi applicativi sono dovuti a Whittermore *et al.*,1987), si basa sulla misurazione della distanza geografica tra la localizzazione dei casi di malattia e l'ipotesi iniziale di rischio uniforme tra le aree di un territorio è valutata mediante un test basato sulla distanza media tra tutte le coppie di casi. Il metodo è stato criticato perché, in particolari contesti geografici quali le aree rurali o urbane, le distanze possono risultare molto diverse tra loro e quindi poco idonee per essere utilizzate come criterio di confronto. Il secondo metodo, indicato con *metodo dei quadrati* (un esempio è dato da Choynowsky,1959), si basa sullo studio della variabilità del conteggio di casi in specifici sottoinsiemi del territorio chiamati quadrati. Queste tecniche si propongono come metodi di *clustering* totale in quanto non consentono di localizzare geograficamente un *cluster* ma piuttosto permettono di quantificare la presenza del fenomeno sul territorio.

In indagini spaziali in cui l'interesse è rivolto sia alla localizzazione geografica del *cluster* che alla valutazione della sua significatività statistica, è consigliabile utilizzare approcci differenti da quelli di *clustering* totale[3]. I due principali metodi descrittivi in questo ambito sono stati definiti da Openshaw *et al.* (1987,1988) e da Besag e Newell (1991). Entrambi i metodi consentono di identificare *clusters* geografici di un fenomeno mediante la sovrapposizione sul territorio di figure geometriche regolari, come quadrati o cerchi, e di valutare la significatività statistica del numero di casi osservati all'interno di ciascuna zona così definita. Il metodo proposto da Openshaw consiste in una tecnica grafica chiamata "Macchina di Analisi Geografica" (GAM, *Geographical Analysis Machine*) che con l'ausilio di finestre circolari, di dimensione variabile, consente di identificare le aree di un territorio caratterizzate da un livello significativo del fenomeno. La tecnica impone sul territorio un grigliato regolare di I punti in ciascuno dei quali si posiziona una finestra circolare di raggio R di dimensione variabile, solitamente da 5 a 10 volte lo spazio esistente tra le celle, e le linee del grigliato sono tali da permettere la sovrapposizione dell'80% dei cerchi costruiti. Per ogni zona si determinano il numero totale di casi e la popolazione a rischio e le zone circolari di interesse sono identificate dalle aree con un numero di casi osservati c_{iR} eccedenti un valore critico C^*_{iR}, corrispondente al 98-

[3] Una buona revisione dei metodi di *clustering* spaziale è riportata in Kulldorff, 2006b

99esimo percentile della distribuzione della variabile casuale C_{iR}; tale variabile descrive il numero di casi nelle zone circolari sotto l'ipotesi di distribuzione uniforme del fenomeno. Il confronto tra le zone avviene effettuando un *test* di ipotesi per ogni valore di C^*_{iR}. La procedura viene ripetuta facendo variare il raggio R delle finestre per un numero minimo di volte (di solito 4-5). La metodologia è simile a quella dei quadrati ad eccezione della forma della finestra di scansione ed è stata criticata anch'essa per il problema dei *tests* multipli. Il numero di *tests* effettuati è elevato (dipende dal numero di zone considerate) e sono tra loro dipendenti; il problema è stato affrontato utilizzando la correzione di Bonferroni ma il suo effetto può risultare poco idoneo e soprattutto conservativo in presenza di un numero elevato di confronti. L'eterogeneità della densità di popolazione può rappresentare un problema nella corretta interpretazione dei risultati ottenuti sulla base di tassi di incidenza o di massimi locali. L'uso di una distanza geografica fissa nel metodo GAM genera un insieme di aree caratterizzate da un numero variabile di casi osservati e di popolazione a rischio. Un modo migliore per effettuare un confronto tra le finestre di ricerca è fissare la dimensione della popolazione. Turnbull *et al.* (1990) ha sviluppato, sulle indicazioni del metodo precedente, un *test* indicato con il nome di *Cluster Evaluation Permutation Procedure* (CEPP) che consente di identificare direttamente il *cluster* che porta al rifiuto dell'ipotesi iniziale. Le finestre sovrapposte all'area studio sono di forma circolare e posizionate nei centroidi delle I celle in cui sono stati aggregati i dati. La particolarità del metodo è la dimensione fissa della popolazione: ogni cerchio è costruito in modo tale da includere lo stesso ammontare di popolazione P piuttosto che lo stesso raggio R o lo stesso numero di casi. Sotto l'ipotesi iniziale di distribuzione uniforme dei casi nella popolazione, la variabile casuale C_{iP}, $i = 1,2,...,I$, descrive il numero di casi osservati nelle zone circolari. La presenza di *clusters* sul territorio viene valutata confrontando i tassi di incidenza o il numero massimo di casi osservati tra le zone circolari identificate, $Z_P = \max\{C_{iP} : i = 1,2,...I\}$. La significatività statistica del risultato è saggiata utilizzando una procedura di simulazione Monte Carlo che campiona i valori fittizi dei casi dalla distribuzione di Z_P sotto l'ipotesi di rischio uniforme. La tecnica proposta differisce dagli altri metodi perché effettua un solo *test* di

ipotesi basato sul valore di Z_P contro una singola ipotesi alternativa. Anche se derivato da considerazioni diverse, tuttavia, il metodo reintroduce il problema dei *tests* multipli che nasce proprio dalla dimensione fissa della popolazione. Fissata una dimensione P, infatti, l'ipotesi alternativa postula l'esistenza di un *cluster* tra tutte le zone circolari aventi la stessa popolazione ma, poiché non esiste una regola che consente di fissare tale dimensione in indagini diverse, lo stesso fenomeno potrebbe essere ugualmente rappresentato da *clusters* con una differente numerosità. Turnbull suggerisce, quindi, di ripetere la procedura variando la dimensione P ma in tal modo si genera ancora un insieme di *tests* di ipotesi correlati tra loro ed un aggiustamento di Bonferroni può non essere sufficiente a meno di un numero ridotto di tentativi sulla popolazione P.

Il metodo di Besag e Newell nasce come approccio alternativo al controllo dell'eterogeneità delle misure locali di incidenza. Essi propongono di selezionare a priori la dimensione k del *cluster* finale. Per ogni caso presente sul territorio, si posiziona il centro di una finestra circolare il cui raggio varia fino ad includere il k-esimo caso presente nell'intorno più vicino. La sovrapposizione sul territorio di finestre circolari rimane invariata ma, piuttosto che fissarne il raggio e valutare la significatività statistica per ogni zona circolare ottenuta, si mantiene costante il numero di casi entro ogni area circolare e si determina la significatività statistica solo per le zone caratterizzate da un eccesso di rischio in popolazione. La scelta della dimensione k è ovviamente fondamentale e, in genere, si utilizza un *range* di valori.

Un metodo basato sull'impiego di finestre mobili di ricerca è la *scan statistic*. La metodologia è stata proposta la prima volta da Naus (1965) come una soluzione al problema dei *tests* multipli; essa è stata sviluppata per la scansione di un territorio in un'ottica temporale alla ricerca di aree significative di un fenomeno utilizzando, come criterio di confronto, il numero massimo di casi osservati all'interno di una finestra di dimensione fissa. La tecnica è sostanzialmente simile a quelle proposte da Openshaw e da Besag e Newell.

Nel corso degli anni, la letteratura a riguardo è divenuta consistente sia nel caso di approcci temporali che spaziali. Secondo le tecniche sinora presentate, la valutazione della significatività statistica dei *clusters* avviene impiegando un *test* di ipotesi per ciascuna area di interesse. Nel caso di un numero elevato di

confronti, si può generare il problema dei *tests* multipli; sotto l'ipotesi nulla di assenza di aggregazioni spaziali significative, la probabilità di un "falso allarme", ovvero di ipotizzare la presenza sul territorio di un *cluster* significativo che in realtà potrebbe non esistere, coincide al più con il livello di significatività α fissato. Al diminuire del livello di errore scelto, si riduce la probabilità di rifiutare l'ipotesi iniziale ma, di contro, si riduce anche la possibilità di individuare un *cluster* realmente esistente. Quanto detto vale sotto la condizione di indipendenza lineare dei *tests* che, nel caso di *clustering* spaziale, potrebbe non essere verificata. Un fenomeno manifestatosi in una specifica zona può essere influenzato dalle aree circostanti invalidando l'assunzione di indipendenza lineare. Come accennato in precedenza, una possibile soluzione al problema è data dalla correzione di Bonferroni (1935). Il metodo consente di correggere la probabilità di rifiutare l'ipotesi nulla per il numero di *tests* effettuati; allo scopo di lasciare invariato il livello di errore α, le aree effettivamente significative sono rappresentate dalle zone aventi un *p-value* pari ad α / I che, in alcune situazioni, porterebbe a rifiutare l'ipotesi nulla solo in condizioni molto restrittive risultando in tal modo conservativo.

Capitolo 3

Spatial Scan Statistic

3.1 - *Spatial scan statistic* per *clusters* regolari

3.1.1 - Aspetti generali e *circular scan statistic*

Le metodologie sino ad ora descritte hanno rappresentato un punto di riferimento per lo sviluppo di una tecnica di *clustering* geografico, ormai consolidata e largamente diffusa, denominata *Spatial Scan Statistic* (Kulldorff e Nagarwalla,1995; Kulldorff,1997,1999). Nei suoi primi anni di applicazione, la *spatial scan statistic* si poneva l'obiettivo di analizzare la distribuzione spaziale di processi puntuali monodimensionali e di individuare *clusters* geografici da valutare in indagini successive. La proposta di Kulldorff è la metodologia di *cluster detection* più utilizzata in ambito epidemiologico e consente sia di localizzare le aggregazioni spaziali di un fenomeno che di valutarne la significatività statistica attraverso opportune procedure inferenziali. Gli aspetti innovativi della metodologia hanno risolto il problema dei confronti multipli e consentito l'uso di finestre di scansione variabili nella forma e nella dimensione. La possibilità di variare la dimensione dell'area di ricerca consente di esaminare quei fenomeni per i quali non si hanno conoscenze a priori sulla loro distribuzione e, di conseguenza, sull'estensione geografica di un possibile *cluster*. La *spatial scan statistic* può essere impiegata in diversi settori di ricerca. In ambito forestale o agricolo, ad esempio, si possono identificare le aree geografiche caratterizzate da una presenza significativa di piante o di alberi, appartenenti ad una specifica specie o con determinati attributi; in campo astronomico, si può essere interessati a verificare se particolari famiglie di stelle o altri elementi dello spazio, si distribuiscono in modo casuale; in zoologia, si possono identificare eventuali concentrazioni sul territorio di animali appartenenti a determinate specie.

Le tre caratteristiche di base della *spatial scan statistic* sono: (1) la geometria della finestra di ricerca imposta sul territorio, (2) la distribuzione di probabilità degli eventi osservati (sotto l'ipotesi nulla) e (3) la dimensione della finestra di ricerca. La variazione singola o congiunta delle tre caratteristiche consente di ricoprire diverse situazioni di indagine adattando la metodologia alle informazioni disponibili. Un aspetto rilevante della metodologia è la sua flessibilità in merito alla natura dei dati da analizzare che possono essere sia

aggregati che individuali. Dalla sua prima formulazione, la *spatial scan statistic* è stata aggiornata ed arricchita di nuovi aspetti tecnici quali l'introduzione di una finestra di forma ellittica (Kulldorff *et al.*,2006a), l'elaborazione di dati ordinali (Jung *et al.*,2007), l'analisi di dati di sopravvivenza (Huang *et al.*,2007) e la possibilità di essere impiegata in indagini con fonti di informazioni multiple (Kulldorff *et al.,*2007).

Si consideri un territorio suddiviso in I sezioni geografiche minori, chiamate *celle*, per ognuna delle quali sono note le coordinate geografiche del centroide (baricentro geografico o di popolazione), il numero di casi osservati e la popolazione di riferimento. Un esempio reale è rappresentato da uno studio sulla mortalità regionale causa-specifica in cui l'area in esame è identificata dall'intera regione e le celle sono definite dai comuni in cui è suddivisa amministrativamente la regione. I casi osservati sono rappresentati dal numero di decessi per ogni causa e la popolazione di riferimento è data dal numero di residenti all'interno di ciascun comune. Indicata con N la popolazione totale e con C il numero totale di casi osservati sul territorio, l'analisi è condizionata al valore dei casi C assunto come misura nota del fenomeno indagato supponendo che, all'interno di un'area, essi si distribuiscano in maniera uniforme. La *spatial scan statistic* opera su un insieme di finestre circolari, il cui raggio varia da 0[4] ad un limite superiore definito a priori espresso come proporzione massima della popolazione totale da includere nel *cluster* finale[5]. Ogni dimensione, forma e localizzazione geografica di una finestra definisce una zona z che identifica un *cluster* potenziale del fenomeno. L'insieme delle zone circolari è indicato con Z e $z \in Z$: una zona z è definita da un cerchio comprendente tutti gli individui appartenenti alle celle i cui centroidi ricadono all'interno del cerchio stesso. Seguendo tale impostazione potenzialmente si può generare un numero "infinito" di cerchi virtuali, spesso sovrapposti, caratterizzati da un numero diverso di casi osservati e di popolazione. Due zone aventi lo stesso insieme di casi osservati presentano le stesse caratteristiche in termini statistici consentendo di passare, in tal modo, da un numero infinito di zone ad un numero finito di aree circolari. Nel caso di dati disaggregati, una

[4] Il raggio del cerchio più piccolo coincide con la distanza minore osservata tra tutte le coppie di centroidi
[5] E' possibile controllare l'estensione del *cluster* fissando la sua lunghezza massima

zona z è perfettamente circolare ed il centro è posizionato sull'individuo stesso; in presenza di dati aggregati, come nel caso di popolazioni raggruppate per comune o distretto censuario, una zona z può avere confini irregolari che dipendono dalle dimensioni e dalla forma geografica delle aree in essa contenute. In questo caso, la definizione delle zone circolari avviene con un criterio diverso dal precedente: gli individui esterni alla zona, ma appartenenti alle celle i cui centroidi ricadono entro il perimetro del cerchio, sono inclusi nella zona. In maniera analoga, ma opposta, vengono esclusi gli individui interni alla zona circolare appartenenti alle celle i cui centroidi ricadono esternamente ad essa. In situazioni particolari, questo modo di operare potrebbe escludere una o più aree di un territorio nella formazione del *cluster* finale: sebbene buona parte della sua popolazione risulti compresa entro i confini della zona circolare, infatti, i centroidi possono collocarsi in una posizione geografica esterna. Ad ogni variazione del raggio R si definisce una zona z per ciascun centroide; in casi limite, la dimensione variabile della finestra può identificare aree geografiche molto estese tali da ricoprire la maggior parte del territorio esaminato. In questo caso parlare di *cluster* risulta di scarso interesse ai fini dello studio spostando l'obiettivo della ricerca proprio sulle zone escluse dalla definizione del *cluster*. Definito con Z l'insieme di tutte le possibili zone circolari, indichiamo con c_i e n_i, rispettivamente, il numero di casi osservati e la popolazione presente in una specifica area i, con $i = 1,..,I$ e con C_z la variabile casuale che descrive il numero di casi osservati per ogni zona z. Definito con Ω lo spazio parametrico di ricerca e con (z, p, q) un suo punto, $p, q \in [0,1]$, indichiamo con z un vettore tridimensionale contenente le coordinate centrali delle celle e la misura del raggio e con p, q due vettori di probabilità che indicano, rispettivamente, la probabilità di un individuo di essere un caso interno ad una zona z e la probabilità di un individuo di essere un caso al di fuori di tale zona. Mediante la *scan statistic,* l'ipotesi iniziale da valutare è di distribuzione uniforme del fenomeno sul territorio o, equivalentemente, di uguaglianza delle probabilità p e q, $H_0 : p \equiv q$. L'ipotesi alternativa invece postula la presenza di almeno una zona z del territorio tale che il rischio interno sia superiore alle rimanenti zone, $H_1 : z \in Z, p > q$, e la stima dei valori di p e q viene effettuata con un

processo di stima di massima verosimiglianza[6]. Lo scopo della ricerca è individuare quelle zone del territorio caratterizzate da un'espressione elevata o ridotta del fenomeno e verificare se l'aumento dei casi osservati in tali zone è casuale. L'aspetto inferenziale della metodologia è basato sulla teoria della massima verosimiglianza. Posto c_z il numero totale di casi osservati all'interno di una zona z, $c_z = \sum_{i \in z} c_i$, ed $n_z = \sum_{i \in z} n_i$ il numero totale di individui presenti nella stessa zona, il fenomeno può essere descritto utilizzando due modelli distributivi: un modello binomiale di parametri n_z e p ed un modello di Poisson con parametro $n_z p$. La valutazione parametrica è condizionata al numero di casi osservati in una specifica zona attraverso la variabile casuale C_z.

L'aspetto innovativo della *spatial scan statistic* è l'utilizzo del rapporto delle funzioni di verosimiglianza (o log-verosimiglianza, *Log-Likelihood Ratio*, LLR) quale statistica-test per l'identificazione di aggregazioni spaziali significative. Poiché le aree di ricerca sono caratterizzate da numerosità di popolazione differenti, il confronto tra le zone non può avvenire utilizzando il numero massimo di casi osservati o il valore massimo del tasso di incidenza poiché non si terrebbe conto dell'eteroschedasticità delle misure adottate, conducendo all'identificazione di un massimo locale piuttosto che globale, in particolare per le zone caratterizzate da una ridotta popolazione. La definizione del rapporto delle funzioni di verosimiglianza dipende dal modello probabilistico utilizzato per la distribuzione dei casi osservati. Nel caso di un modello binomiale, dove un caso è rappresentato da un individuo avente o meno uno specifico attributo (ad esempio una malattia), per ogni zona z e sotto l'ipotesi nulla $H_0 : p \equiv q$, la variabile casuale C_z si distribuisce secondo la legge $C_z \approx Bin(n_z, p)$. Sotto l'ipotesi alternativa $H_1 : z \in Z, p > q$, invece, i casi osservati si distribuiscono come $C_z \approx Bin(n_z, p)$ per le aree appartenenti alla zona z e come $C_z \approx Bin((N - n_z), q)$ per tutte le rimanenti aree del territorio. La funzione di verosimiglianza $L(z, p, q)$ può essere descritta dall'espressione:

$$L(z, p, q) = [p^{c_z}(1-p)^{n_z - c_z}][q^{C - c_z}(1-q)^{(N-n_z)-(C-c_z)}] \quad (1)$$

[6] La disuguaglianza si inverte quando l'analisi è rivolta all'identificazione di *clusters* caratterizzati da un livello ridotto del fenomeno

Il rapporto delle funzioni di verosimiglianza, che indicheremo sempre con Ts, può essere definito come:

(2) $$Ts = LR = \frac{\sup_{z \in Z,\, p>q} L(z, p, q)}{\sup_{p=q} L(z, p, q)} \qquad \text{con } (p, q \in [0,1])$$

Sotto l'ipotesi nulla $(p \equiv q)$ la funzione di verosimiglianza (1) si riduce a:

(3) $$L(z,p,q) = p^C (1-p)^{N-C} = \left(\frac{C}{N}\right)^C \left(1-\frac{C}{N}\right)^{N-C} = \frac{C^C (N-C)^{N-C}}{N^N} \equiv L_0$$

Nel caso di assenza di *clusters* sul territorio, la funzione di verosimiglianza assume un valore costante che dipende solo dal numero totale di casi osservati e dalla popolazione totale e non dalla loro distribuzione spaziale. La funzione di verosimiglianza posta al numeratore della (2) viene massimizzata su tutti i possibili valori di $0 \le q \le p \le 1$: fissata una zona z, il valore massimo della funzione di verosimiglianza per una distribuzione binomiale si ottiene quando $p = c_z / n_z$ e $q = (C - c_z)/(N - n_z)$ che rappresentano gli stimatori di massima verosimiglianza dei parametri della distribuzione Binomiale. Possiamo dunque scrivere:

(4) $$L(z,p,q) = p^{c_z} (1-p)^{n_z - c_z} q^{C - c_z} (1-q)^{(N-n_z)-(C-c_z)} =$$

$$= \left(\frac{c_z}{n_z}\right)^{c_z} \left(\frac{n_z - c_z}{n_z}\right)^{n_z - c_z} \left(\frac{C - c_z}{N - n_z}\right)^{C - c_z} \left(\frac{N - n_z - (C - c_z)}{N - n_z}\right)^{(N-n_z)-(C-c_z)} \text{, se } \left(\frac{c_z}{n_z}\right) > \left(\frac{C - c_z}{N - n_z}\right)$$

Il processo di ricerca si riduce all'individuazione dei parametri incogniti che massimizzano la funzione di verosimiglianza posta al numeratore della statistica-test. La zona che risolve il problema di massimo riportato nella (4) identifica il *Most*

Likely Cluster (MLC), ovvero il *cluster* più verosimile o probabile, e corrisponde a quella zona $\hat{z} \in Z$ tale che $L(\hat{z}) \geq L(z), \forall z \in Z$. Rapportando la funzione (4) alla (3), la statistica-test può essere espressa sinteticamente da:

(5) $$Ts = \frac{\max_z L(z)}{L_0} = \frac{L(\hat{z})}{L_0}$$

La distribuzione del LR dipende dalla distribuzione non omogenea della popolazione sul territorio e, in genere, non ha una forma analitica di semplice definizione; essa assume valore 1 quando non esistono aggregazioni spaziali sul territorio relative al fenomeno osservato (ipotesi H_0 vera) oppure un valore diverso da 1, determinato dall'espressione (4), nel caso in cui esiste almeno una zona $z \in Z$ caratterizzata da un livello del fenomeno superiore (o inferiore) a quello delle rimanenti zone (ipotesi H_1 vera).

Nel caso di dati aggregati, si ipotizza che gli eventi siano generati da un processo puntuale di Poisson non omogeneo. Un processo spaziale puntuale è un processo stocastico in cui ciascuna variabile casuale rappresenta la localizzazione di un evento nello spazio. Un processo spaziale puntuale di Poisson (omogeneo o non omogeneo) è un processo stocastico in cui il numero di eventi in un'area finita z si distribuisce secondo un modello di Poisson con intensità λ che identifica il numero di casi attesi per ogni unità areale. Nel caso di un processo omogeneo, si può variare l'ammontare di casi osservati per ogni realizzazione del processo tenendo costante un numero non noto di casi attesi: l'intensità λ risulta, in tal modo, costante per ogni zona del territorio. L'assunzione di omogeneità dei tassi è un'ipotesi molto forte, soprattutto in ambito epidemiologico, in quanto la popolazione a rischio può variare in modo non uniforme tra le aree; si preferisce utilizzare, in questi casi, un processo di Poisson non omogeneo in cui il numero di casi presenti in una specifica area z si distribuisce secondo una distribuzione di Poisson con valore atteso $\mu_z = \int_i \lambda_i(x)dx$ e, condizionatamente al numero di casi, le localizzazioni x_i identificano un campione casuale indipendente della distribuzione (definita su

z) avente una funzione di densità di probabilità proporzionale a $\lambda_i(x)$. In questo caso si ipotizza che il numero di eventi in una zona sia maggiore in corrispondenza di un valore elevato di $\lambda_i(x)$. La *scan statistic* di Kulldorff utilizza un processo spaziale puntuale di Poisson non omogeneo ipotizzando che aree diverse abbiano una densità di popolazione diversa. Sotto l'ipotesi nulla $H_0 : p \equiv q$, la variabile che descrive i casi osservati si distribuisce come $C_z \approx Pois(pn_z)$ mentre, sotto $H_1 : z \in Z, p > q$, esiste almeno una zona z tale che $C_z \approx Pois(pn_z + q(N - n_z))$. In questo caso, la funzione di verosimiglianza può essere espressa da una distribuzione di probabilità di Poisson e dalla funzione di densità di probabilità di ciascun caso localizzato sul territorio. La distribuzione di probabilità dei casi C_z è data da un modello di Poisson con parametro $(pn_z + q(N - n_z))$:

(6) $$p(z, p, q) = Pois(pn_z + q(N - n_z)) = \frac{e^{-[pn_z + q(N - n_z)]}[pn_z + q(N - n_z)]^C}{C!}$$

La funzione di densità $f(x)$ di un caso osservato alla locazione x è data da:

(7) $$\frac{pn_x}{pn_z + q(N - n_z)}, \text{ se } x \in z$$

(8) $$\frac{qn_x}{pn_z + q(N - n_z)}, \text{ se } x \notin z$$

Di conseguenza la funzione di verosimiglianza può essere scritta come:

(9) $$L(z, p, q) = \frac{e^{-[pn_z + q(N - n_z)]}[pn_z + q(N - n_z)]^C}{C!}$$

$$\times \prod_{x_i \in z} \frac{pn_{x_i}}{pn_z + q(N - n_z)} \prod_{x_i \notin z} \frac{qn_{x_i}}{pn_z + q(N - n_z)}$$

$$= \frac{e^{-[pn_z + q(N-n_z)]}[pn_z + q(N-n_z)]^C}{C!} p^{c_z} q^{(C-c_z)} \prod_{x_i} n_{x_i}$$

Sotto H_0 essa assume un valore costante anche per il modello di Poisson:

$$L(z, p, q) = L_0 = \sup_p \left[\frac{e^{-pN} p^C}{C!} \prod_{x_i} n_{x_i} \right] = \frac{e^{-C}}{C!} \left(\frac{C}{N} \right)^C \prod_{x_i} n_{x_i}$$

Il valore massimo dell'espressione (9) si ha quando $p = c_z / n_z*$ e $q = (C - c_z)/(N - n_z*)$, dove n_z* identifica il numero di casi attesi nella zona z, ottenuto mediante una standardizzazione indiretta in assenza di covariate, che risulta proporzionale alla popolazione a rischio presente nell'area considerata ($n_z(C/N)$)[7]. Si può dunque scrivere[8]:

(10) $$L(z, p, q) = \frac{e^{-C}}{C!} \left(\frac{c_z}{n_z*} \right)^C \left(\frac{C - c_z}{N - n_z*} \right)^{C-c_z} \prod_{x_i} n_{x_i} \quad \text{,se } \left(\frac{c_z}{n_z*} \right) > \left(\frac{C - c_z}{N - n_z*} \right)$$

La statistica-test può essere definita dal rapporto:

(11) $$Ts = \frac{\max_z L(z)}{L_0} = \frac{\left(\frac{c_z}{n_z*} \right)^{c_z} \left(\frac{C - c_z}{N - n_z*} \right)^{C-c_z}}{\left(\frac{C}{N} \right)^C}$$

[7] Il numero di casi attesi n_z*, aggiustato per la categoria i ($i = 1,...k$) di una covariata, è dato da $\sum_i n_z(C_i / N_i)$

[8] $\hat{p}$ e $\hat{q}$ rappresentano gli stimatori di massima verosimiglianza per una distribuzione di Poisson

Come nel modello binomiale, Ts assume valore 1 in presenza di una distribuzione uniforme del fenomeno ed un valore diverso in tutti gli altri casi. Nell'ambito della *spatial scan statistic*, non si può parlare di *test* uniformemente più potente, in quanto l'ipotesi alternativa composta postula l'esistenza di almeno una zona in cui il fenomeno è statisticamente diverso da quello delle altre aree, ma si parla di *test individualmente* più potente.
Ritornando alla *spatial scan statistic*, il processo di simulazione Monte Carlo impiegato nella procedura di ricerca si può descrivere nei seguenti passi:

1. si calcolano i valori della statistica-test per i dati reali
2. dato il numero totale di casi, si simula un numero elevato di *datasets* fittizi (repliche) sotto l'ipotesi di distribuzione uniforme
3. si determina il valore della statistica-test per ogni replica
4. si ordinano, in senso decrescente, i valori della statistica-test ottenuti per i dati reali e per i dati simulati
5. si determina il rango dei valori della statistica-test per i dati reali e si confronta con quello relativo ai dati simulati. Se il rango reale è compreso nel primo α% dei dati simulati, si rifiuta l'ipotesi iniziale di distribuzione uniforme del fenomeno.

Nel processo di simulazione, ogni replica è una copia esatta delle aree originali aventi la stessa popolazione a rischio n_z mentre il numero di casi c_z è ottenuto campionando da una distribuzione di Poisson con parametro $(C/N)n_z$; in tal modo, i casi sono generati con un tasso di rischio uniforme pari a quello osservato $p = q = (C/N)$. Il numero di repliche è solitamente fissato a 999, 9999, 19999 o simili, ottenendo una stima più attendibile del *p-value* all'aumentare delle simulazioni. La valutazione della significatività avviene confrontando i ranghi dei valori di LLR ordinati in senso decrescente relativi ai dati reali e a quelli simulati; il *p-value* di un *cluster* si ottiene dalla formula $p = r/(1+sim)$, dove r è il rango della statistica-test per i dati reali e sim è il numero di simulazioni effettuate: fissato un livello di significatività α, un *cluster* risulterà significativo se il valore della statistica-test per i dati reali è tra i primi α% valori di LLR simulati ordinati in senso decrescente.

E' importante segnalare che malgrado il *cluster* identificato rappresenti l'aggregazione spaziale più "verosimile" sul territorio, esso potrebbe non coincidere con il *cluster* "reale" in quanto altre zone del territorio, geograficamente simili ad esso, potrebbero essere caratterizzate da una verosimiglianza pressoché uguale tale da rappresentare il fenomeno. Il *cluster* deve essere utilizzato come indicazione della localizzazione del fenomeno sul territorio senza attribuire ad esso una definizione rigorosa e precisa, sottolineando l'incertezza dei confini geografici dei *clusters* realmente esistenti. Solo accurate indagini sulle aggregazioni identificate potranno delineare, con maggiore precisione, le zone realmente interessate da un livello significativo del fenomeno.

3.1.2 - *Elliptic scan statistic*

In particolari situazioni di indagine, la scelta di una finestra di scansione circolare può risultare poco idonea per la corretta identificazione di *clusters c*aratterizzati da un andamento spaziale meno regolare. Si pensi, ad esempio, alle aree di un territorio in cui la popolazione è esposta ad una sorgente lineare di rischio, quale una linea elettrica o un corso d'acqua; in questi casi è molto probabile che un *cluster* assuma una forma geometrica stretta e/o allungata piuttosto che circolare. Per ovviare a tale inconveniente, la *spatial scan statistic* è stata recentemente ampliata con l'introduzione di una finestra di ricerca ellittica (Kulldorff *et al.*,2006a). L'*elliptic spatial scan statistic* è una derivazione speciale della metodologia circolare e la ricerca dei *clusters*, la teoria statistica ed i criteri di valutazione della significatività rimangono invariati rispetto al caso circolare.

Un'ellisse è caratterizzata e identificata da cinque parametri: coordinate cartesiane, x ed y, del suo centroide, andamento (o eccentricità), angolo e dimensione. L'andamento di un'ellisse è dato dal rapporto tra il semiasse maggiore e il semiasse minore; un valore del rapporto uguale ad 1 identifica un cerchio mentre valori superiori ad 1 definiscono un'ellisse di forma sempre più

stretta ed allungata. L'angolo θ di un'ellisse è definito dalla linea orizzontale (direzione est-ovest) e il semiasse maggiore e si ha un angolo di 90° quando quest'ultimo è orientato in direzione nord-sud. Analogamente al caso circolare, si genera un sottoinsieme di ellissi variando opportunamente i parametri di riferimento. Fissati i centroidi, la forma e gli angoli dell'ellisse, la procedura varia la dimensione della finestra da 0 ad un limite superiore, fissato a priori, espresso in percentuale di popolazione da includere nel *cluster* finale. Per analisi con dati aggregati, le aree sono incluse nella zona ellittica se il loro centroide ricade nei limiti geometrici definiti. Nella fase di ricerca, si definisce un *range* di possibili valori per l'andamento dell'ellisse. I valori di s solitamente impiegati nella procedura sono 1 (cerchio), 1.5, 2, 3, 4, 5, ma è possibile variare tale parametro a seconda delle esigenze specifiche dello ricerca e di eventuali conoscenze a priori del fenomeno. L'identificazione di *clusters* ellittici dalla forma particolarmente allungata e stretta non rappresenta sempre un vantaggio per una corretta interpretazione dei risultati; in particolari conformazioni del territorio, un *cluster* molto allungato potrebbe includere aree appartenenti a Stati (o confini) diversi rendendo di difficile interpretazione il *cluster* individuato. Oltre all'andamento dell'ellisse, per la ricerca dei *clusters* si utilizzano diversi valori angolari per ogni ellisse. Fissato un valore di s, si genera un numero di angoli θ_n pari a 3 volte (6 volte) il valore dell'andamento fissato; ad esempio, fissando s=2 si definiscono 6 angoli, θ_n=6. Essi sono determinati ripartendo un angolo di 180° in θ_n parti uguali e l'angolo θ=90° è sempre selezionato dalla procedura. A seconda dell'andamento dell'ellisse, il valore di θ può condizionare la definizione del *cluster* finale; studi su dati reali hanno evidenziato che, per *clusters* eccentrici, il numero di zone che forma l'aggregazione finale può cambiare ad una minima variazione dell'angolo includendo nel *cluster* anche le aree, non necessariamente contigue, caratterizzate da un rischio elevato pur non supportati da un analogo fenomeno nelle zone circostanti. Allo scopo di prevenire la formazione di aggregazioni spaziali di scarso interesse, è stato introdotto un parametro di penalizzazione o di non-compattezza. Il parametro consente di "scoraggiare" l'individuazione di *clusters* dalla forma geometrica eccentrica agendo sull'andamento dell'ellisse. Il parametro di penalizzazione è dato dall'espressione $\left(4s/(s+1)^2\right)^a$, dove s è

l'andamento dell'ellisse ed $a(\geq 0)$ è un valore scalare di regolazione. La funzione di log-verosimiglianza risulta modificata e penalizzata dal parametro come indicato dall'espressione:

(12) $$Ts = LLR_{adj}(z) = LLR * \left(\frac{4s}{(s+1)^2} \right)^a$$

Si noti che per $a \equiv 1$, si ottiene il rapporto inverso tra l'area del rettangolo più piccolo contenente l'ellisse e l'area di un quadrato contenente una circonferenza di perimetro uguale a quello del rettangolo. La penalizzazione ha un effetto direttamente proporzionale ai valori del parametro a; nei casi limiti, quando $a \equiv 0$, la penalità è completamente annullata e i due LLR coincidono mentre, per $a \to \infty$, essa diventa preponderante a tal punto da consentire solo la definizione di *clusters* circolari. L'influenza della penalizzazione aumenta in maniera monotona al crescere di s risultando $LLR_{adj}(z) \to 0$ per $s \to \infty$. In genere, si considera media una penalizzazione pari a 0.5 e forte per $a \to 1$. Qualora si decida di non impiegare una penalizzazione in fase di ricerca, occorre ponderare opportunamente la scelta dei valori di s in quanto andamenti ellittici elevati richiedono un numero maggiore di angoli, influenzando notevolmente il carico computazionale della procedura.

3.1.3 - *Spatial scan statistic* per dati ordinali

Nell'ambito epidemiologico, i fenomeni esaminati sono generalmente espressi da due distribuzioni di probabilità: Poisson, nel caso di dati di conteggio e Bernoulli nel caso di informazioni dicotomiche. In alcune indagini, i due modelli distributivi possono non essere adeguati a descrivere il fenomeno nello spazio probabilistico; si pensi, ad esempio, ad una variabile con livelli ordinati di grandezza, quale lo stadio di un tumore o la classe di età di un gruppo di individui. In questi casi, si sceglie erroneamente di dicotomizzare la variabile osservata e di utilizzare una distribuzione di Bernoulli. Tale approssimazione

genera sia una perdita di informazioni in merito al fenomeno che la mancanza di un riferimento per la scelta del punto di *cut-off* per la dicotomizzazione della variabile; è preferibile utilizzare modelli funzionali che consentono di includere, nella loro definizione, la struttura ordinale dei dati. La *spatial scan statistic* è stata recentemente ampliata con l'aggiunta di modelli probabilistici in grado di analizzare dati in forma ordinale e politomica (Jung *et al.,*2007)

Supponiamo di suddividere un'area studio in I celle per ognuna delle quali si dispone di una variabile risposta classificata in K categorie. Si indichi con c_{ik} il numero di osservazioni nella zona i e nella categoria k, con $i = 1,2....I$ e $k = 1,2,...K$. Le categorie della variabile derivano da un ordinamento naturale del fenomeno; nel caso di una malattia, un valore elevato di k denota uno stadio grave della malattia. Posto $c_i = \sum_k c_{ik}$ il numero di osservazioni nella zona i, $C_k = \sum_i c_{ik}$ il numero totale delle osservazioni nella categoria k e $C = \sum_k \sum_i c_{ik}$ il numero totale delle osservazioni in una specifica area, possiamo definire la funzione di verosimiglianza per il modello ordinale come:

$$L(z, p_1, p_1, ... p_K, q_1, q_2, ... q_K) \propto \prod_k \left(\prod_{i \in z} p_k^{c_{ik}} \prod_{i \notin z} q_k^{c_{ik}} \right) \tag{13}$$

dove p_k è la probabilità che un'osservazione, interna alla finestra z, appartenga alla categoria k e q_k è la probabilità che un'osservazione, esterna a tale zona, appartenga alla stessa categoria. La somma delle singole probabilità è pari ad 1, $\sum_k p_k = 1$ e $\sum_k q_k = 1$. L'ipotesi iniziale postula che la probabilità di essere nella categoria k, appartenendo alla zona z, è uguale alla probabilità di essere nella stessa categoria ma di non appartenere a tale zona. In simboli, possiamo scrivere:

$$H_0 : p_1 = q_1, ..., \; p_K = q_K \tag{14}$$

Di contro, l'ipotesi alternativa composta è definita come:

$$H_1 : \frac{p_1}{q_1} \leq \frac{p_2}{q_2} \leq ..... \leq \frac{p_K}{q_K} \tag{15}$$

con almeno una disuguaglianza in senso stretto. L'ipotesi alternativa postula, invece, la presenza di un *cluster* sul territorio caratterizzato da un tasso di rischio maggiore, rispetto alle aree circostanti[9], tra i livelli più alti della variabile. Questo tipo di restrizione d'ordine è detta ordinamento del rapporto di verosimiglianza (Dykstra *et al.,*1991). Nel caso di $K = 2$ si ritorna alla *spatial scan statistic* con un modello distributivo di Bernoulli. La statistica-test utilizzata nella procedura è ancora il rapporto delle funzioni di verosimiglianza nel quale il numeratore è dato dall'espressione (13) mentre il denominatore (sotto l'ipotesi H_0) è dato dall'espressione:

$$L_0 = \prod_k \prod_i \hat{p}_{ok}^{\,c_{ik}} = \prod_k \left(\frac{C_k}{C} \right)^{\sum_i c_{ik}} = \prod_k \left(\frac{C_k}{C} \right)^{C_k} \tag{16}$$

dove $\hat{p}_{ok} = (C_k / C) = \hat{q}_{ok})$ è lo stimatore di massima verosimiglianza di p_k sotto l'ipotesi nulla. Le stime di massima verosimiglianza, $\hat{p}_k$ e $\hat{q}_k$, sotto l'ipotesi alternativa sono ottenute utilizzando l'algoritmo "*Pool-Adjacent-Violators*" descritto da Barlow *et al.*. Esso combina le categorie adiacenti di una variabile ordinale fino ad ottenere una successione monotona crescente dei tassi di rischio per tutti i livelli della variabile; se la successione dei tassi osservati, internamente ed esternamente ad una zona, $\sum_{i \in z} c_{ik} / \sum_k \sum_{i \in z} c_{ik}$ e $\sum_{i \notin z} c_{ik} / \sum_k \sum_{i \in z} c_{ik}$, risulta monotona non decrescente per $k = 1,2,...K$, le stime di massima verosimiglianza corrispondono ai tassi osservati. Ciò implica che il risultato non identifichi necessariamente un *cluster* caratterizzato da un rischio interno perfettamente ordinato in maniera crescente o decrescente per tutte le categorie della variabile. Ad esempio, nel caso di una malattia con 4 stadi di gravità, si può identificare un *cluster* significativo per lo stadio 4 confrontato con una combinazione dei 3 livelli rimanenti (stadio 4 *vs* stadio1-2-3), anche se lo stadio 3 presenta un tasso di rischio inferiore a quello dei singoli livelli 1 e 2. Similmente, si potrebbe identificare un *cluster* con un rischio elevato per gli stadi 3 e 4 uniti contro lo stadio 2 o lo stadio 1 (stadio 3-4 *vs* stadio1).

[9] Quando si vuole indagare su un livello del rischio ridotto per le categorie inferiori della variabile, occorre invertire il segno della disuguaglianza. Per l'individuazione di *clusters* con alto/basso rischio, invece, la successione dei rapporti può essere sia crescente che decrescente

Dopo aver identificato il *cluster* più probabile, la valutazione della significatività statistica si ottiene mediante un processo di simulazione Monte Carlo, analogamente alle metodologie già illustrate in precedenza. La generazione dei *datasets* fittizi, sotto l'ipotesi nulla, avviene condizionando sul totale dei casi osservati $(C_1,....,C_K)$ in ciascuna categoria k. Si generano inizialmente C_1 individui che si assegnano alla prima categoria; si procede selezionando casualmente C_2 degli individui rimanenti che vengono assegnati alla categoria 2 e così via. Dopo aver generato C_{K-1} individui ed assegnati alla rispettiva categoria, i rimanenti C_K sono attribuiti all'ultima categoria K. Il *p-value* di riferimento è dato da $p = r/(1+sim)$, dove sim è il numero di replicazioni utilizzato ed r è il rango della statistica-test calcolata sui dati reali.

3.1.4 - *Spatial scan statistic* per dati di sopravvivenza

Gli studi di sopravvivenza rappresentano un settore di particolare interesse in ambito epidemiologico e sanitario; l'identificazione di aree geografiche caratterizzate da un'elevata sopravvivenza rappresenta uno degli obiettivi primari nella programmazione sanitaria di un territorio. Le zone con una ridotta sopravvivenza degli individui forniscono indicazioni importanti sull'adeguatezza dei trattamenti sanitari o sul grado di aggressività di alcune malattie. Analogamente, le zone in cui si evidenzia una sopravvivenza maggiore, rispetto alle altre aree del territorio, possono essere utilizzate come indicatori dell'efficacia di specifiche cure sanitarie o evidenziare la presenza di fattori prognostici favorevoli per la salute dell'intera popolazione.

Un primo approccio a questo tipo di analisi è la definizione di un punto di *cut-off* che consente di ripartire i soggetti in due gruppi, lunga e breve sopravvivenza, impiegando successivamente una *scan statistic* con un modello di Bernoulli per la identificazione di eventuali *clusters*. Tale approccio presenta due inconvenienti. Il primo è la scelta di un punto di *cut-off* per i tempi di sopravvivenza e, in presenza di dati censurati, non risulta chiaro come

dicotomizzare i tempi di sopravvivenza antecedenti al punto di *cut-off.* Il secondo problema deriva dalla perdita di informazioni dovuto al passaggio da una variabile continua (tempo di sopravvivenza) ad una variabile dicotomica (lunga o breve sopravvivenza). Huang *et al.*(2007) propone di utilizzare una funzione esponenziale come modello probabilistico per i casi osservati, già ampiamente utilizzata in studi di sopravvivenza tradizionali, per la facilità di trattamento computazionale. Ad ogni individuo presente sul territorio è associata una variabile casuale non negativa (t_i, δ_i) dove $t_i = \min(L_i, T_i)$ ed il parametro δ_i assume valore 1 quando si osserva un effettivo tempo di fallimento T_i e valore 0 in presenza di un tempo censurato L_i. Supponiamo che i tempi osservati per ciascun individuo siano indipendenti ed identicamente distribuiti secondo un modello esponenziale, $f(T_i) = e^{-T_i/\theta}/\theta$, di media θ_{in} e media θ_{out}, rispettivamente, all'interno e all'esterno di una zona z. L'ipotesi nulla postula l'uguaglianza delle due medie $H_0 : \theta_{in} = \theta_{out}$ o, equivalentemente, una sopravvivenza uniforme sull'intero territorio contro un'ipotesi alternativa composta $H_1 : \theta_{in} > \theta_{out}$, per la quale almeno una zona del territorio è caratterizzata da una sopravvivenza maggiore rispetto alle altre. La disuguaglianza nell'ipotesi alternativa si inverte quando si vuole valutare la presenza di aree a sopravvivenza ridotta. Posto con n_{in} e n_{out}, rispettivamente, il numero di individui interni ed esterni alla zona z e con $r_{in} = \sum_{i \in Z} \delta_i$ ed $r_{out} = \sum_{i \notin Z} \delta_i$ il numero di individui non censurati interni ed esterni della stessa zona, la funzione di verosimiglianza per una specifica zona è data dall'espressione:

$$(17) \quad L(z, \theta_{in}, \theta_{out}) = \prod_{i \in Z} \frac{1}{(\theta_{in})^{\delta_i}} e^{-\frac{T_i \delta_i}{\theta_{in}}} e^{-\frac{L_i (1-\delta_i)}{\theta_{in}}} \times \prod_{i \notin Z} \frac{1}{(\theta_{out})^{\delta_i}} e^{-\frac{T_i \delta_i}{\theta_{out}}} e^{-\frac{L_i (1-\delta_i)}{\theta_{out}}}$$

$$= \frac{1}{(\theta_{in})^{r_{in}}} e^{-\sum_{i \in Z} \frac{t_i}{\theta_{in}}} \frac{1}{(\theta_{out})^{r_{out}}} e^{-\sum_{i \notin Z} \frac{t_i}{\theta_{out}}}$$

dove $i \in z$ indica che l'i-esimo individuo è localizzato nella zona z.

Le stime di massima verosimiglianza per θ_{in} e θ_{out} sono espresse da $\hat{\theta}_{in} = r_{in} \Big/ \sum_{i \in z} t_i$ e $\hat{\theta}_{out} = r_{out} \Big/ \sum_{i \notin z} t_i$ per cui, sotto $H_1 : \theta_{in} \neq \theta_{out}$, avremo:

$$L(\hat{z}) = \max_z \frac{1}{\left(\hat{\theta}_{in}\right)^{r_{in}}} e^{-\frac{\sum_{i \in z} t_i}{\hat{\theta}_{in}}} \frac{1}{\left(\hat{\theta}_{out}\right)^{r_{out}}} e^{-\frac{\sum_{i \notin z} t_i}{\hat{\theta}_{out}}} = \max_z \left(\frac{r_{in}}{\sum_{i \in z} t_i} \right)^{r_{in}} e^{-r_{in}} \left(\frac{r_{out}}{\sum_{i \notin z} t_i} \right)^{r_{out}} e^{-r_{out}}$$

mentre per l'ipotesi nulla H_0 si ottiene

$$L_0 = \frac{1}{\left(\hat{\theta}_G\right)^R} e^{-\frac{\sum_{i \in G} t_i}{\hat{\theta}_G}} = \left(\frac{R}{\sum_{i \in G} t_i} \right)^R e^{-R}$$

dove $R = r_{in} + r_{out}$ e G identifica l'intera area studio. La statistica-test è indicata dal rapporto $Ts = L(\hat{z}) / L_0$ dove $\hat{z}$ è la zona che massimizza la funzione di verosimiglianza $L(z, \theta_{in}, \theta_{out})$. Per un'ipotesi alternativa $H_1 : \theta_{in} \neq \theta_{out}$, il rapporto delle funzioni di verosimiglianza si può scrivere:

(18)
$$Ts = \frac{\max_z \left(\frac{r_{in}}{\sum_{i \in z} t_i} \right)^{r_{in}} \left(\frac{r_{out}}{\sum_{i \notin z} t_i} \right)^{n_{out}}}{\left(\frac{R}{\sum_{i \in G} t_i} \right)^R}$$

Nel caso di un'ipotesi alternativa $H_1 : \theta_{in} > \theta_{out}$, la statistica-test è moltiplicata per una funzione indicatrice $I\left(\frac{r_{in}}{\sum_{i\in z} t_i}\right) > \left(\frac{r_{out}}{\sum_{i\notin z} t_i}\right)$ mentre per $H_1 : \theta_{in} < \theta_{out}$ è moltiplicata per $I\left(\frac{r_{in}}{\sum_{i\in z} t_i}\right) < \left(\frac{r_{out}}{\sum_{i\notin z} t_i}\right)$.

Le informazioni censuarie sono incorporate nella funzione di verosimiglianza attraverso i valori di $r_{in} = \sum_{i\in Z} \delta_i$, $r_{out} = \sum_{i\notin Z} \delta_i$. Nel caso di sopravvivenza uniforme, L_0 dipende solo dal numero totale degli individui non censurati e non dalla distribuzione spaziale degli individui stessi. Se nello studio non esistono osservazioni censurate, le funzioni di verosimiglianza possono essere espresse utilizzando il numero di individui presenti nelle zone considerate:

(19)
$$Ts = \frac{\max_z \left(\frac{n_{in}}{\sum_{i\in z} t_i}\right)^{n_{in}} \left(\frac{n_{out}}{\sum_{i\notin z} t_i}\right)^{n_{out}}}{\left(\frac{N}{\sum_{i\in G} t_i}\right)^{N}}$$

La determinazione del *p-value* avviene mediante un processo di simulazione diverso da quello finora descritto. La distribuzione dei tempi di sopravvivenza è sconosciuta sia in termini di sopravvivenza attesa che nel meccanismo di censura. L'algoritmo di simulazione diventa un procedimento iterativo di permutazione, condizionato sull'insieme osservato dei tempi t_i e degli indicatori di censura δ_i, permutando le coppie di valori osservati $\{(t_i, \delta_i), i = 1,2,...N\}$ tra le coordinate geografiche individuali. La randomizzazione non è effettuata campionando le osservazioni da una distribuzione esponenziale ma piuttosto permutando le localizzazioni spazio-tempo ed i relativi attributi (tempo di sopravvivenza/meccanismi di censura) per ciascuna osservazione. Per ottenere la distribuzione esatta della statistica-test, quest'ultima deve essere calcolata per tutte le $n!$ permutazioni delle osservazioni per cui il carico computazionale diventa oneroso anche per piccoli *datasets*. La determinazione del valore del *p-*

value rimane invariata rispetto ai casi finora descritti. I tempi di sopravvivenza possono essere descritti anche da una distribuzione di probabilità Gamma, log-Normale o una distribuzione di Weibull. In questi casi continua a sussistere la validità della statistica-test descritta in quanto la sua distribuzione è ottenuta dalle permutazioni casuali delle localizzazioni geografiche e dei tempi di sopravivenza/meccanismi di censura. L'approccio esponenziale si rivela un ottimo strumento di analisi anche in presenza di dati di sopravvivenza censurati ed è possibile inserire nel modello specifiche covariate, sia continue che discrete, per correggere l'analisi da eventuali fattori di disturbo.

3.1.5 - *Spatial scan statistic* per dati multivariati

La *spatial scan statistic* è diventato uno strumento molto utilizzato in ambito biomedico per l'identificazione di *clusters* di una singola malattia. Spesso l'interesse è però rivolto alla valutazione di fenomeni dipendenti da fonti di informazione multiple; ad esempio, se si vuole valutare la presenza di un eccesso di casi di leucemia infantile in uno specifico territorio non è facile decidere se analizzare solo i casi di leucemia linfatica acuta, quelli di leucemia mieolide acuta o combinare entrambe le informazioni. Un approccio multivariato consente di tenere conto delle diverse fonti di dati individuali e di valutare, contemporaneamente, i diversi sintomi con i quali si manifesta una malattia o di indirizzare la ricerca verso specifici gruppi di individui, come nel caso di malattie in età avanzata o infantile. Un approccio semplice e talvolta poco idoneo a questo tipo di problema consiste nell'applicazione della *spatial scan statistic* ad un unico *dataset* ottenuto sommando i casi osservati presenti nelle diverse fonti di informazione; in tal modo, il risultato potrebbe derivare principalmente da un solo insieme di dati ed essere influenzato dalla variabilità dei dati rimanenti. Un'altra soluzione consiste nell'utilizzare la *scan statistic* separatamente per ciascun *dataset* ma, in questo caso, si può avere una perdita di potenza se il *cluster* risulta ugualmente rappresentato nelle diverse fonti di informazione. Allo scopo di superare tali inconvenienti, la *spatial scan statistic* è stata ampliata per

consentire l'impiego in indagini multivariate (Kulldorff *et al.*,2007). La *spatial scan statistic* può essere impiegata in analisi spazio-temporali in cui la finestra di ricerca assume una forma cilindrica, la cui base è definita allo stesso modo di una *scan statistic* puramente spaziale mentre l'altezza rappresenta il periodo di osservazione temporale. La finestra si muove nello spazio e nel tempo e, per ogni posizione geografica e dimensione dell'area di ricerca, effettua una scansione del territorio per ciascun istante di tempo definito; si genera così un insieme di cilindri, spesso sovrapposti, caratterizzati da andamenti e dimensioni differenti che ricopre l'intera area studio e ciascuno dei quali identifica un possibile *cluster*. Il *cluster* più probabile è identificato dalla zona caratterizzata dal valore più elevato di LLR ottenuto sommando i singoli LLR dei *datasets* esaminati. Per ogni dimensione della finestra di ricerca e per ciascun *dataset*, si determina il LLR evidenziando se si tratta di un eccesso o di un difetto di casi osservati rispetto agli attesi. Nella fase successiva, per ogni finestra si calcolano due valori di LLR: il primo è la somma degli LLR (di ogni *dataset*) che identificano un eccesso di casi mentre il secondo valore è ottenuto sommando i LLR (di ogni *dataset*) relativi ad un difetto di casi osservati. Il valore più elevato tra tutte le somme così ottenute corrisponde al *cluster* più probabile. La procedura descritta vale per un'ipotesi alternativa bidirezionale ma può essere modificata qualora si fosse interessati solo ad un eccesso o difetto di casi osservati, utilizzando solo la prima o seconda somma. Fissata una zona z, la statistica-test può essere riportata come segue:

$$Ts = \max_{z} \max\left(\sum_{i} LLR_i(high, z), \sum_{i} LLR_i(low, z), \right) \tag{20}$$

dove i identifica l'i-esimo *dataset*. Nel caso di ricerca di *clusters* caratterizzati da un rischio elevato, la statistica-test è identificata dalla seguente somma:

$$Ts = \max_{z}\left(\sum_{i} LLR_i(high, z) \right)$$

Il criterio di valutazione della significatività statistica dei risultati è identico a quello illustrato per le procedure precedenti.

In sintesi, la procedura di *spatial scan statistic* può essere riassunta nei seguenti passi:

1. definizione di un insieme di finestre di ricerca (circolari, ellittiche o cilindriche) per ogni area (o individuo) del territorio. Ogni finestra è posizionata nel centroide (o individuo) dell'area considerata.
2. variazione della dimensione della finestra da 0 fino ad un limite superiore prefissato (in percentuale di popolazione totale o come lunghezza lineare massima della finestra)
3. determinazione del valore di LLR per ciascuna finestra
4. ordinamento decrescente dei valori di LLR trovati
5. procedura di simulazione Monte Carlo per la determinazione della distribuzione della statistica-test sotto l'ipotesi di distribuzione uniforme
6. determinazione dei valori di LLR per ciascuna finestra fittizia generata
7. ordinamento decrescente dei valori di LLR ottenuti dalla simulazione
8. valutazione della significatività statistica dei *clusters* trovati mediante confronto tra il rango del LLR dei dati reali e dei dati fittizi

3.1.6 - Considerazioni e limitazioni della *spatial scan statistic*

La *spatial scan statistic* è il metodo di *clustering detection* più utilizzato in diversi settori di ricerca. Numerose applicazioni presenti in letteratura confermano l'efficacia e la potenza della metodologia nell'identificazione di *clusters* di forma geometrica regolare nonché la sua elevata velocità di esecuzione in condizioni medie di analisi. L'utilità della metodologia è, tuttavia, condizionata da alcuni fattori. La procedura effettua una scansione completa del territorio e, nel caso di un numero elevato di aree, la fase computazionale assume un'importanza rilevante in funzione del numero di simulazioni effettuate. Per *datasets* molto grandi, ad esempio con 600.000 *records* e più di 15.000 aree

esaminate, la metodologia necessita di alcuni giorni di elaborazione per l'identificazione del *cluster* finale (Neill,2005). Il limite principale della metodologia rimane, ad ogni modo, il vincolo geometrico imposto alla finestra di ricerca che condiziona fortemente la definizione del *cluster;* essa può risultare, infatti, poco idonea per la ricerca di aggregazioni spaziali dalla forma geometrica allungata ed irregolare.

3.2 - *Spatial scan statistic* per *clusters* irregolari

Introduzione

La stima del massimo valore di verosimiglianza è il problema principale delle metodologie di *scan statistic*; in genere, si utilizzano due differenti strategie per ottenere una soluzione approssimata del massimo della funzione. La prima consiste nella riduzione dello spazio parametrico di ricerca in un suo sottoinsieme che risulti facilmente gestibile da un punto di vista computazionale, come accade nella *spatial scan statistic* di Kulldorff. La seconda soluzione si basa, invece, su metodi di ottimizzazione stocastica quali gli algoritmi genetici (Knjazev,2002; Duczmal *et al.,*2007) o di *simulated annealing* (Aarts e Korst,1989; Winkler,1995; Duczmal e Assunção,2004); si tratta di procedure iterative di notevole complessità computazionale che convergono, sotto certe assunzioni, al massimo ottimo in senso globale attraverso un numero elevato di iterazioni. Sebbene efficace e largamente utilizzata in diversi ambiti di applicazione, la metodologia *circular spatial scan statistic* presenta alcune limitazioni che hanno favorito lo sviluppo di approcci diversi di *cluster detection,* in particolare per l'identificazione di aggregazioni dalla forma geometrica irregolare. Spesso in situazioni reali, i *clusters* presenti sul territorio risultano caratterizzati da un andamento differente da quello circolare o regolare; ad esempio, la presenza di una sorgente di inquinamento che rappresenta una fonte di rischio per l'insorgenza di malattie respiratorie nelle aree geografiche limitrofe ad essa e la presenza di vento, di direzione variabile, può influenzare la

definizione del *cluster* finale invalidando una possibile simmetria dello stesso (Biggeri *et al*.,1996). Un rischio incrementato lungo i corsi di acqua oppure in prossimità di linee ferroviarie, stradali o di tensione elettrica, può contribuire alla definizione di aggregazioni geografiche di forma ben lontana da quella di un *cluster* circolare (Verkasalo,1993). Applicando le usuali tecniche di *clustering* spaziale alle situazioni appena descritte, si può incorrere in un'errata localizzazione geografica del *cluster* reale per due motivi. Il primo è legato ad un'estensione territoriale maggiore del *cluster* individuato rispetto a quello reale; il secondo è l'eventuale esclusione delle aree caratterizzate da un livello significativo del fenomeno ma i cui centroidi risultano esterni alla zona circolare definita. I recenti sviluppi delle metodologie di *scan statistic* si sono concentrati sull'implementazione di procedure di ricerca che consentono di superare tali limitazioni, consentendo l'identificazione di aggregazioni di forma irregolare.

3.2.1 - *Upper Level Set scan statistic*

Un recente metodo di *clustering* spaziale è stato proposto da Patil e Taille (2004) ed è indicato con l'acronimo ULS. La tecnica consente l'identificazione di aggregazioni spaziali irregolari ed è un approccio di ricerca basato sui dati che opera su uno spazio parametrico ridotto. Si consideri uno spazio bidimensionale S ripartito in I celle elementari, indicate con a_i ed $i = 1,...I$, per ciascuna delle quali si dispone di un conteggio di casi non negativo, espresso da una variabile casuale C_a, e di una dimensione n_a fissa e nota a priori. Come nel caso della *spatial scan statistic*, la variabile relativa ai casi osservati può essere descritta da due modelli distributivi: quello di Poisson e quello binomiale. Nel caso di un modello binomiale, il valore di n_a identifica il numero di elementi, appartenenti alla cella a, aventi o meno uno specifico attributo, con una certa probabilità non nota p_a, $p_a \in (0,1)$. La variabile di conteggio C_a indica, invece, il numero di individui osservati all'interno di una cella e si distribuisce secondo la legge di probabilità $C_a \approx Bin(n_a, p_a)$. Per un modello di Poisson, il valore di n_a identifica la popolazione a rischio mentre la

variabile C_a individua un processo di Poisson (omogeneo o non omogeneo) di intensità λ_a, descritta dalla distribuzione $C_a \approx Poi(\lambda_a n_a)$, con $\lambda_a > 0$. Per entrambi i modelli distributivi le assunzioni di base prevedono che le variabili risposta C_a siano indipendenti ed identicamente distribuite e che la variabilità spaziale risulti interamente spiegata dai parametri del modello. Analogamente alla *spatial scan statistic*, la ULS *scan statistic* si basa su tre caratteristiche principali: la geometria dell'area di ricerca, un modello probabilistico di riferimento per i valori osservati e la forma della finestra di scansione. Le prime due caratteristiche sono state già illustrate mentre la terza componente può essere descritta come la capacità di identificare *clusters* spaziali senza l'imposizione di vincoli geometrici alle aree esaminate impiegando, come strumento di ricerca, una funzione dei tassi di risposta osservati in ogni cella del territorio; in tal senso, la ULS *scan statistic* dipende solo dai dati reali osservati. Il punto di partenza della metodologia è una struttura a grafo nella quale risultano connessi i vertici delle aree territoriali adiacenti. La condizione di adiacenza geografica può essere scelta in differenti modi. Un caso limite è rappresentato da una zona adiacente solo a sé stessa; essa individua già un'area di interesse ed il numero massimo di *clusters* coincide esattamente con il numero di zone presenti sul territorio. La situazione opposta si ha quando un vertice risulta connesso con tutti gli altri nodi del grafo determinando un insieme di collegamenti difficilmente gestibile in fase di ricerca. In genere, si preferisce utilizzare una connettività di primo ordine in cui si considerano solo le aree strettamente adiacenti. Tale scelta consente di limitare sia la dimensione dell'insieme delle possibili aggregazioni finali che rendere più agevole la ricerca dei risultati spostandosi tra i vertici del grafo percorrendo i collegamenti esistenti. Si possono definire adiacenti due aree aventi un lato o almeno un punto in comune, oppure quando il confine condiviso ha una lunghezza positiva o, ancora, nel caso di corsi d'acqua, quando il flusso procede in una specifica direzione attraversando due aree adiacenti. L'obiettivo della ricerca è l'individuazione delle zone territoriali che mostrano un elevato tasso di risposta del fenomeno in riferimento alle rimanenti aree. L'idea di base della metodologia ULS è, dunque, simile a quella della *spatial scan statistic* di

Kulldorff ma si distingue da essa per la specifica modalità di trattamento delle aree e per l'identificazione dei *clusters* potenziali.

Prima di definire in dettaglio la metodologia ULS, è opportuno riportare alcune nozioni di teoria dei grafi utilizzate nelle tecniche di *clustering* geografico per aggregazioni non regolari. Si definisce *grafo non-orientato* G o brevemente *grafo,* un insieme V finito non vuoto di elementi detti vertici e un insieme E di coppie non ordinate di punti distinti, dette lati, tali che $G := (V, E)$. In genere, un grafo è rappresentato da un diagramma i cui vertici sono identificati da punti e due vertici sono congiunti mediante un segmento lineare e_i che identifica univocamente i due vertici connessi, $e_i = \{v_{i1}, v_{i2}\}$, con $v_{i1} \neq v_{i2}$, $v_{i1}, v_{i2} \in V$, $i = 1,...m$. La cardinalità dell'insieme V è detta ordine del grafo G mentre la cardinalità di E è detta dimensione del grafo (Mignani e Montanari, 1994). Un grafo orientato (o digrafo) G_o è costituito da un insieme finito V non vuoto di vertici e un insieme (anche vuoto) E di coppie ordinate di vertici distinti, dette archi, tali che $G_o := (V, E)$; un arco è rappresentato da una freccia che unisce due vertici definendo una direzione mentre se nel grafo esistono lati o archi che congiungono un vertice con sé stesso si parla di pseudografi. Un grafo G costituito da n vertici ha un numero massimo di vertici e di lati compreso tra 0 e $n(n-1)/2$ mentre in un digrafo tale valore è compreso tra 0 e $n(n-1)$. Un grafo $G := (V, E)$ si dice connesso se ogni coppia distinta di vertici v_j, v_k è collegata da un percorso, detto cammino, ovvero se esiste una sequenza di vertici $v_{r1},...v_{rp}$ tale che $v_j = v_{r_1}, v_k = v_{r_p}$ e $\{v_{ri}, v_{ri+1}\} \in E, i = 1,...p-1.$ Un elemento $S = (V_1, E_1)$ è un sottografo di G se $V_1 \subseteq V$ e $E_1 \subseteq E$; per ottenere un sottografo è sufficiente eliminare dal grafo iniziale uno o più collegamenti o vertici. Si definisce, invece, componente di un grafo G un suo sottografo connesso che non è contenuto propriamente in nessun altro sottografo connesso di G, ovvero è un sottografo massimale connesso. Una componente in cui ogni coppia di vertici è connessa da un unico cammino è detta albero e rimuovendo un collegamento da un albero si ottiene un sottografo sconnesso. Un esempio di struttura a grafo è riportato in figura (1).

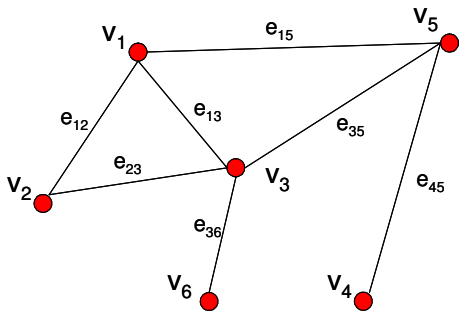


Figura 1. Esempio di grafo con vertici (v_i) e relativi collegamenti (e_{ij})

I concetti di teoria dei grafi illustrati risultano particolarmente utili nei problemi di classificazione. Gli elementi che si desidera classificare in un numero non noto a priori di *clusters* possono essere interpretati come i vertici $v_1,...v_i$ di un grafo generando una corrispondenza tra *clusters* e grafo: l'individuazione delle aggregazioni spaziali equivale alla ricerca di opportuni sottografi del grafo G.

Ritornando alla tecnica ULS, il tasso di risposta empirico, relativo ad una singola cella a, é espresso dal rapporto $G_a = c_a / p_a$. La definizione dell'insieme dei potenziali *clusters*, indicato con *upper level set* (ULS), avviene utilizzando dei valori soglia (o livelli) g del tasso di risposta osservato:

$$U_g = \{a : c_a / p_a \geq g\}$$

Le zone candidate a formare un *cluster* z sono identificate da tutte quelle aree con un tasso di risposta superiore al livello g prefissato. La metodologia ULS consente di definire una struttura ad albero associata allo spazio parametrico Ω_{ULS} in cui i nodi dell'albero rappresentano le zone candidate a formare il *cluster* e le foglie individuano i punti di massimo locale del parametro osservato. La struttura ad albero, indicata con ULS-*tree,* è definita dai collegamenti tra le zone candidate ai diversi livelli della funzione G. La caratteristica principale della procedura è la restrizione dello spazio parametrico di ricerca alle sole zone costituite dalle componenti geografiche connesse, $Z \in \Omega_{ULS}$, identificate dai limiti superiori dei tassi di risposta; tale restrizione consente una ricerca più rapida dei *clusters* in quanto la cardinalità dell'insieme Ω_{ULS} è inferiore al

numero massimo di zone esaminate e le zone appartenenti a tale insieme rappresentano degli *hot-spots*[10] potenziali. Supponiamo di fissare due livelli g e g' dei tassi di risposta e di ottenere, rispettivamente, un *upper level set* individuato da tre componenti connesse, z_1, z_2 e z_3 ed un insieme costituito dalle zone z_4, z_5 e z_6. Il passaggio tra due livelli può consentire la definizione delle zone candidate secondo tre differenti modalità:

- una nuova zona può formarsi dall'unione di due zone già esistenti;
- ad un livello superiore, una zona già esistente può aumentare la sua estensione;
- può emergere una nuova zona non definita al livello precedente.

In fase iniziale, la procedura genera una matrice I di adiacenze delle aree del territorio. Successivamente le righe e le colonne sono ordinate in senso decrescente in base al valore del tasso di risposta definendo, per ogni livello g, una sotto-matrice quadrata triangolare superiore della matrice completa I. Nella fase successiva si delinea lo spazio parametrico di ricerca Ω_{ULS} costituito dall'insieme di aree connesse definito ad ogni livello scelto e si determinano le stime di massima verosimiglianza $L(z, p, q)$[11] per ogni elemento di tale spazio. Sotto l'ipotesi iniziale di distribuzione uniforme, si determina la distribuzione del LLR mediante un processo di simulazione (algoritmo MC) e la significatività statistica dei *clusters* è valutata confrontando il rango della statistica-test per i dati reali con quello ottenuto nella fase di simulazione.
In sintesi, la procedura ULS può essere descritta dal seguente algoritmo:

1. definizione della matrice di adiacenza I e calcolo dei tassi di risposta per ciascuna zona del territorio
2. ordinamento decrescente dei tassi di risposta ottenuti

[10] La definizione di *hot-spot* non è rigorosa e ben definita in letteratura. In genere, con questo termine si indica un'area in cui il valore di verosimiglianza del parametro osservato è troppo grande per essere dovuta al caso
[11] Analogamente alla *circular scan statistic*, i parametri *p* e *q* identificano, rispettivamente, la probabilità di essere un caso internamente ed esternamente ad una zona *z*

3. definizione degli insiemi di nodi connessi ad una zona in riferimento a ciascun livello g[12]
4. determinazione della statistica-test LLR per ogni insieme di zone aggregate
5. ripetizione dei punti (3) e (4) per ogni area del territorio
6. ordinamento decrescente dei valori di LLR ottenuti
7. identificazione dei principali *hot-spots* dalla lista ordinata di LLR
8. verifica della significatività statistica degli *hot-spots* mediante un algoritmo di simulazione MC

3.2.2 - *Simulated annealing* e *spatial scan statistic*

Tra le recenti proposte metodologiche di *clustering* spaziale è presente un algoritmo di ricerca sviluppato da Duczmal e Assunção (2004) che, basandosi su un approccio di *simulated annealing* (SA), consente di superare le limitazioni geometriche imposte dalla *spatial scan statistic* estendendo la ricerca a *clusters* di forma irregolare. L'algoritmo SA prende il nome da un fenomeno termodinamico di ricottura (*annealing*) di materiali solidi (Montanari,1999); se un materiale solido viene riscaldato oltre il proprio punto di fusione e poi successivamente raffreddato (*cooling*), in modo da riportarlo allo stato solido, le sue proprietà strutturali dipendono fortemente dal processo di raffreddamento (*cooling schedule*). In pratica, un algoritmo di *simulated annealing* simula il cambiamento di energia di un sistema, considerato come un insieme di particelle sottoposto a raffreddamento fino a che converge allo stato solido; questo permette di determinare soluzioni ammissibili di problemi di ottimizzazione convergendo verso soluzioni ottime. In un processo termodinamico, la probabilità che un sistema cambi stato, a cui corrisponde un incremento di energia (δE), è pari a

[12] Se il tasso di risposta di una cella è superiore (o inferiore) ad un determinato livello g, la cella viene etichettata come "esposta" (o "non esposta"). Tale classificazione consente, nei passi successivi, di formare un insieme dei nodi "esposti" che identifica la zona di potenziale interesse

$$p(\delta E) = \exp(-\delta E / kt)$$

dove k è la costante di Boltzmann e t è la temperatura del sistema. Gli algoritmi SA utilizzano tecniche di ricerca locale per definire ed esplorare l'intorno di una soluzione corrente; se l'intorno contiene una soluzione migliore, quest'ultima diventa la soluzione corrente e il procedimento viene iterato altrimenti si procede valutando un peggioramento della soluzione stessa. Tale peggioramento viene visto come un aumento di energia del sistema termodinamico associato per cui la probabilità di accettare la mossa peggiorante dipende dalla temperatura di processo e dal peggioramento indotto dalla mossa stessa: la probabilità di accettazione cresce all'aumentare della temperatura e diminuisce all'aumentare del peggioramento.

Nel caso di *clustering* spaziale, la procedura si avvale di una funzione obiettivo per ridurre lo spazio parametrico di ricerca e determinare una "buona" soluzione del problema. Ipotizzando il territorio suddiviso in I zone distinte, la procedura SA propone di selezionare, secondo regole prefissate, solo i sottografi più "promettenti" tra tutti i possibili eliminando le combinazioni di vertici meno interessanti in termini di verosimiglianza: l'obiettivo è identificare quella zona z del territorio, costituita da un insieme di aree connesse, che massimizza il rapporto LLR sotto le ipotesi H_0 e H_1. Analogamente alle procedure sinora descritte, l'ipotesi nulla rimane invariata, $H_0 : p \equiv q$, così come quella alternativa, $H_1 : z \in Z, p > q$. Scelto a caso un nodo iniziale, la procedura seleziona, ad ogni passo, l'area adiacente da aggiungere o sottrarre all'insieme di zone già definito. La scelta avviene in funzione della variazione del LLR ed in considerazione del fatto che la zona deve essere adiacente ad almeno un'area presente in tale insieme; ad un generico passo della procedura, il vertice da aggregare viene individuato secondo opportuni criteri di scelta. Il grado di determinismo della selezione è legato al concetto di temperatura: più è alta la temperatura, maggiore è la casualità nella scelta del prossimo vertice. Il parametro temperatura (indicato con *temp*) assume solo tre valori discreti: alta, media e bassa temperatura. Nel caso di alta temperatura, l'algoritmo seleziona una zona secondo una scelta del tutto casuale, senza una direzione precisa, e con un elevato grado di mobilità sul territorio; nel caso di temperatura media, la

scelta avviene con una probabilità proporzionale al valore di LLR delle aree adiacenti, indirizzando la ricerca verso le zone con un elevato di LLR ma senza, per questo, scartare nessuna altra direzione; infine, nel caso di bassa temperatura, la *routine* opera in maniera puramente deterministica selezionando sempre il nodo adiacente con il più alto valore di LLR. Per poter unificare i tre criteri è stata definita una funzione $F(G,temp)$ che, in base al sottografo G analizzato, individua il vertice adiacente in accordo con il valore del parametro *temp*. L'algoritmo determina il valore della temperatura a seconda dell'andamento della funzione di verosimiglianza nel sottografo corrente. Oltre alla funzione $F(G,temp)$, è stata introdotta una seconda funzione, indicata con $H(G)$, dipendente dallo specifico nodo aggiunto al sottografo analizzato. La funzione $H(G)$ consente di aumentare il numero di vertici del sottografo G aggiungendo il nodo dove si è verificato un recente e significativo incremento del valore di LLR. In pratica, se nell'ultimo passo è stato aggiunto un vertice v_k al sottografo G, la funzione $H(G)$ seleziona, in maniera completamente casuale, un nuovo nodo v_k' adiacente a v_k da aggiungere a G; viceversa, se un nodo v_k è stato scartato al passo precedente, la funzione $H(G)$ fornirà un intorno di G secondo il criterio $F(G,temp=low)$. L'idea è di incrementare l'importanza dei sottografi più promettenti. La definizione dei *clusters* potenziali avviene, quindi, secondo regole decisionali basate sull'uso combinato delle due funzioni obiettivo appena indicate e la scelta della migliore strategia è condizionata da quattro parametri: *hL* (*high-Likelihood function value*), *cs* (*consecutive step*), *vb* (*visited before*), *cv* (*common vertices*). La loro funzione è riportata di seguito:

- *hL*=1 se, al passo corrente, è stato trovato un intorno di G con il più alto valore di verosimiglianza (*hL*=0 altrimenti);
- *cs* è il numero di passi successivi durante i quali non è stato identificato nessun nuovo sottografo con un valore di LLR >1;
- *vb* è il numero di visite sinora effettuate per il sottografo corrente G;
- *cv* è il numero di vertici in comune tra il sottografo corrente G e quello con il più alto valore di verosimiglianza trovato nella ricerca.

Ad ogni passo, l'algoritmo determina il valore dei quattro parametri verificando che risultano entro dei valori soglia prefissati, condizionando dinamicamente il processo di selezione dei sottografi più promettenti. La ricerca di tali sottografi avviene secondo un preciso schema strategico:

- la strategia $F(G, temp = high)$ è scelta se il sottografo corrente G è stato visitato molte volte, ha un valore di verosimiglianza relativamente basso e, per diversi *steps* della *routine*, la verosimiglianza del sottografo non ha subito incrementi;
- la strategia $F(G, temp = medium)$ è adottata in condizioni simili alla precedente ad eccezione di un incremento del valore di verosimiglianza osservato in qualche sottografo visitato di recente;
- la strategia $F(G, temp = low)$, invece, si utilizza quando si è osservato un incremento della verosimiglianza in qualche passaggio recente ed almeno una di queste condizioni risulta verificata: il sottografo corrente è stato visitato più volte oppure ha un valore di verosimiglianza relativamente basso.
- La strategia $H(G)$, infine, viene applicata se il sottografo corrente presenta un elevato valore di verosimiglianza (apparendo dunque "promettente"), non è stato visitato molte volte ed è stato individuato un incremento della verosimiglianza in qualche sottografo visitato di recente.

La procedura termina quando almeno uno dei due parametri, *vb* e/o *cs,* ha raggiunto il limite massimo prefissato. L'uso combinato delle due funzioni consente di eliminare i nodi già visitati che non evidenziano un interesse ai fini dell'identificazione del *cluster* finale limitando, in tal modo, il numero di visite totali e riducendo i tempi di esecuzione della *routine*. L'algoritmo utilizza, come sottografo di partenza, il *cluster* primario individuato dalla *spatial scan statistic*; non è una scelta obbligata ma può risultare conveniente, da un punto di vista computazionale, partire da una zona già identificata come un potenziale *cluster*. L'insieme completo delle aggregazioni di interesse viene definito applicando la procedura per ogni nodo del *cluster* iniziale, selezionato in modo casuale, finché

il 99% dei vertici non è stato visitato almeno una volta. Ciò implica che la *routine* venga eseguita migliaia di volte per differenti sottografi iniziali fino a quando il massimo valore di LLR non subisce incrementi ulteriori per una lunga sequenza di sottografi visitati oppure quando sono stati già visitati tutti i possibili sottografi. Il sottografo associato al massimo valore della statistica-test rappresenta il *cluster* più probabile la cui significatività statistica viene valutata mediante una procedura di simulazione MC.

La metodologia SA si basa su una soluzione quasi-ottimale del massimo valore del LLR a differenza della *spatial scan statistic,* in cui si determina l'esatto valore della statistica-test attraverso una ricerca esaustiva delle zone; in una procedura di *simulated annealing,* tale soluzione potrebbe non essere assicurata o, comunque, non convergere al valore reale. La procedura SA utilizzata in ulteriori studi (Tango e Takahashi,2005) ha mostrato la tendenza ad identificare *clusters* più estesi di quelli realmente esistenti senza peraltro aggiungere informazioni in merito al fenomeno esaminato. La riduzione della dimensione massima del *cluster* è una possibile soluzione a tale inconveniente ma l'operazione si rivela efficace solo per aggregazioni di ridotte dimensioni (circa il 10% delle zone totali) favorendo, per aree più estese, l'individuazione di *clusters* particolarmente irregolari ed una perdita di potenza della *routine*. Generalizzando un'idea già sviluppata per la *elliptic scan statistic* (Kulldorff *et al.,*2006a), la procedura SA è stata modificata introducendo un parametro di penalizzazione per le aggregazioni caratterizzate da andamento fortemente irregolare (Duczmal *et al.*,2006b). Fissata una zona z ed indicata con $A(z)$ la sua area e con $P(z)$ il perimetro del cerchio esterno convesso contenente tale zona[13], il parametro di controllo, detto di "non-compattezza" ed indicato con $K(z)$, è definito dal rapporto tra $A(z)$ e $P(z)$:

$$K(z) = \frac{4\pi A(z)}{P(z)^2} = \frac{A(z)}{\pi\left(\frac{P(z)}{2\pi}\right)^2} \tag{21}$$

[13] Il perimetro della zona convessa viene calcolato mediante l'algoritmo di Quickhull (Schneider e Eberly, 2003)

La dimensione di una zona non influisce sulla sua compattezza che dipende solo dall'andamento geometrico ed il valore di $K(z)$ tende a diminuire all'aumentare dell'irregolarità geometrica dell'aggregazione, intesa come allontanamento dalla forma circolare; ad esempio, un quadrato ha un parametro di non-compattezza pari a $\pi/4$=0.785 mentre, per figure circolari, si ottiene il valore 1. L'obiettivo della penalizzazione è lasciare inalterato il rapporto LLR nel caso di compattezza prossima ad 1 (*clusters* regolari) e di penalizzarlo qualora essa tenda a 0. Il concetto è stato generalizzato introducendo una costante di regolazione $a(\geq 0)$ nella formula (21) facendo assumere alla statistica-test la forma funzionale $LLR(z)^{K(z)^a}$. Quando $a \to 0$, il LLR risulta pressoché invariato mentre la penalizzazione assume un effetto crescente per valori di $a \to 1$ (penalità forte) ammettendo, per $a \to \infty$, solo *clusters* circolari e penalizzando le aggregazioni geografiche la cui estensione è inferiore all'area della zona convessa che le contiene.

In uno studio sul tumore alla mammella nel nord-est degli Stati Uniti (Duczmal *et al*.,2006b), è stata valutata la capacità di ricerca della SA penalizzata in relazione alla SA standard, la *circular* e *elliptic scan statistic*. I risultati hanno evidenziato un ruolo importante della costante di penalizzazione sia in termini di estensione geografica del *cluster* trovato che di LLR; all'aumentare della penalizzazione l'andamento del *cluster* diventa meno irregolare ed il LLR assume valori contenuti in linea con quelli ottenuti dalle *scan statistics* regolari. Inoltre, in uno studio di simulazione è stata valutata la funzione di potenza delle tre metodologie impiegate: non esistono differenze sostanziali in termini di potenza dei *tests* anche se la metodologia di Kulldorff evidenzia prestazioni migliori, rispetto alla SA *scan statistic,* per aggregazioni di forma circolare e/o ellittica. Uno specifico confronto tra le due tecniche di SA ha evidenziato un perdita di potenza della SA non penalizzata all'aumentare della dimensione del *cluster.*

3.2.3 - Algoritmo genetico *e spatial scan statistic*

Un'ottimizzazione della SA è stata proposta dallo stesso Duczmal (2007) e si tratta di un algoritmo genetico che riprende in parte la tecnica di *simulated annealing* penalizzata ed in parte la *spatial scan statistic*. Esso consente di definire aggregazioni spaziali dalla forma meno arbitraria e richiede un carico computazionale minore rispetto alla SA. Gli algoritmi genetici (*Genetic Algorithm,* GA) sono stati implementati per la prima volta da John Holland nel 1960. L'obiettivo principale della loro nascita era lo studio del fenomeno dell'adattamento naturale e biologico degli individui implementato con una logica matematica. Un algoritmo genetico può essere definito come una tecnica adattiva ed euristica il cui obiettivo è determinare il valore reale o una soluzione approssimata di un problema di ottimizzazione e di ricerca. Essi fanno parte di un'ampia famiglia di algoritmi evoluzionistici utilizzati in campo medico e biologico; nel corso degli anni hanno trovato applicazione anche in diversi settori quali l'economia, le analisi di mercato, l'ecologia e nei processi di automazione *cellular automata* o reti neurali. I due principali scopi per cui gli algoritmi genetici sono utilizzati sono la valutazione della bontà di adattamento di modelli quantitativi e l'ottimizzazione della *performance* di sistemi operativi. Nel primo caso, si vogliono identificare i parametri che minimizzano la discrepanza tra i modelli teorici e le loro applicazioni reali; in questo caso, la funzione obiettivo è una funzione di errore che misura la differenza tra i dati osservati e i dati predetti da un modello. Nel secondo caso, invece, si intende valutare la *performance* di sistemi complessi quali, ad esempio i sistemi di erogazione del gas o dell'energia elettrica, i sistemi semaforici, che risultano definiti da un numero elevato di parametri la cui interazione in situazioni realistiche rendono complessa la soluzione analitica. In merito al processo di ottimizzazione, si parla di massimizzazione dell'adattamento di un sistema mentre nel primo caso si parla di minimizzazione tra dati e modello, anche se i due concetti sono utilizzati in maniera interscambiabile. Non esiste una definizione rigorosa e unica per descrivere un algoritmo genetico ma è possibile individuare degli elementi comuni alle diverse definizione esistenti. Gli elementi comuni sono una popolazione iniziale e tre operatori genetici: quest'ultimi sono rappresentati, in

genere, dalla selezione, dal *crossing-over* e dalla mutazione che identificano opportune regole stocastiche che consentono l'evoluzione di una popolazione nella generazione successiva. La popolazione è costituita da un insieme iniziale di elementi candidati alla riproduzione (*cromosomi*) ai quali vengono applicati gli operatori genetici indicati. In termini formali, la popolazione è una sequenza di cromosomi, indicati con una stringa di 0 ed 1 (detti *bit*), ognuno dei quali identifica un punto nello spazio di ricerca delle possibili soluzioni candidate. L'operatore di selezione individua nella popolazione candidata gli elementi più idonei alla riproduzione attraverso la valutazione della *fitness* individuale; quest'ultima è una funzione *score* assegnata ad ogni elemento della popolazione ed è rappresentata da una funzione o un'espressione matematica. L'idea di base è semplice: più è alta la *fitness* di un elemento, maggiore è la sua probabilità di essere selezionato. L'operatore di *crossing-over* combina le informazioni genetiche di due genitori per ottenere due nuovi elementi detti figli. Esso seleziona casualmente il punto (detto *locus*) sulla sequenza di cromosomi di ciascun genitore dal quale iniziare lo scambio (*crossing-over)* del materiale e le due nuove sequenze sono determinate dalla combinazione delle sequenze di cromosomi dal punto di taglio in poi. Esistono diversi tipi di *crossing-over*: singolo (appena descritto), a doppio punto (nel quale sono scelti due *locus* di taglio anziché uno) e uniforme (in cui i cromosomi sono scelti in maniera completamente casuale). Infine, l'operatore di mutazione agisce sul singolo individuo della popolazione effettuando un cambiamento casuale nella sequenza dei cromosomi. La mutazione può avvenire in qualsiasi punto della stringa ed ha una probabilità molto bassa, intorno all'1 per mille, detta tasso di mutazione *(mutation rate*).

Un algoritmo genetico semplice opera secondo lo schema riportato di seguito (Mitchell,1996):

1. genera casualmente una popolazione di n individui (una sequenza di cromosomi di lunghezza l)
2. determina per ogni elemento della popolazione (cromosoma) la sua *fitness*
3. ripete iterativamente i seguenti passi fino alla creazione di n figli:

- seleziona a caso una coppia di genitori dalla popolazione iniziale: maggiore è la *fitness* di un cromosoma, maggiore è la sua probabilità di essere selezionato come genitore. La selezione avviene con reintroduzione per cui uno stesso cromosoma può essere scelto più volte come genitore
- con probabilità p_c (tasso di *crossing-over*), si effettua il *crossing-over* tra i due genitori selezionati in un punto scelto a caso con probabilità uniforme. Se il *crossing-over* non avviene, i figli risultano essere una copia esatta dei loro genitori lasciando inalterato il codice genetico
- effettua la mutazione nei cromosomi dei figli con probabilità p_m (tasso di mutazione) e sposta il cromosoma risultante nella nuova popolazione.

  Se n è dispari, si scarta casualmente un elemento della popolazione.

4. sostituisce la popolazione corrente con la nuova popolazione generata
5. ritorna al passo (2).

Una completa iterazione del ciclo descritto (dal punto 2 al punto 4) è detta generazione; di solito, un algoritmo genetico è costituito da un numero di generazioni compreso tra 50 e 500 e l'insieme completo di generazioni è detto corsa (*run*). Considerata la componente casuale nella procedura di riproduzione, è probabile che due corse giungano a risultati diversi e per tale motivo è opportuno valutare più corse dell'algoritmo confrontando i risultati ottenuti o combinandoli tra loro in qualche modo. La procedura GA si propone, dunque, di modificare un insieme iniziale di individui attraverso le operazioni di selezione, mutazione e *crossing-over* allo scopo di identificare una generazione successiva con caratteristiche migliori. In riferimento al *clustering* spaziale, la selezione individua le zone del territorio più idonee alla formazione di aggregazioni spaziali in base alla funzione di verosimiglianza (*fitness*); il *crossing-over* combina le informazioni derivanti da due zone specifiche (genitori), selezionate casualmente dalla popolazione iniziale, generando due nuove zone (figli); la mutazione agisce sulle singole aree sostituendole, in modo casuale, con nuove aree connesse. In tal modo, i genitori e la relativa prole (*offspring*) formano un insieme di zone connesse che definiscono una struttura a grafo in modo simile

agli approcci *graph-based* sinora illustrati. Le ipotesi utilizzate per la definizione dei *clusters* coincidono con quelle utilizzate nelle precedenti metodologie. Per una i-esima zona, si conoscono il numero di casi osservati e la popolazione a rischio. La variabile casuale che descrive il numero di casi osservati in una zona i è indicata con C_i e, sotto l'ipotesi H_0, risultano indipendenti ed identicamente distribuite secondo un modello di Poisson. L'ipotesi alternativa H_1 postula l'esistenza di almeno una zona del territorio in cui il rischio è superiore a quelle delle rimanenti aree. La statistica-test è data dal rapporto LLR calcolato per ogni aggregazione ritenuta di potenziale interesse. L'insieme di aree che massimizza il valore della statistica-test identifica il c*luster* più probabile e la sua significatività statistica è valutata mediante un algoritmo di simulazione MC.

Il processo di generazione di una nuova popolazione avviene utilizzando una popolazione iniziale costituita da I insiemi (*current generation list*), dove un generico elemento di tale insieme è rappresentato da una serie di aree geografiche connesse ottenuta mediante un processo di aggregazione successiva: per ogni zona della mappa, si definisce un insieme di aree connesse ottenuto aggiungendo, ad ogni passo, il nodo adiacente che massimizza il LLR. In questa fase non si hanno restrizioni a priori sull'andamento del *cluster* finale se non quelle derivate dall'uso di una strategia di controllo utilizzata nella procedura SA penalizzata. L'operazione prosegue fino al raggiungimento di una dimensione massima di popolazione oppure quando il LLR non subisce ulteriori variazioni. Alla fine del processo di ricerca, la popolazione iniziale risulterà costituita da un numero I di zone connesse non necessariamente separate. L'obiettivo principale di un GA è la costruzione delle migliori generazioni discendenti da una popolazione iniziale attraverso un'operazione di selezione degli individui. La selezione si basa sulla *fitness* dei potenziali individui: nel caso di *clustering* spaziale, la *fitness* è rappresentata dal LLR; maggiore LLR, maggiore *fitness*. La lista della generazione corrente (ovvero dei possibili I genitori) è ordinata in senso decrescente per i valori di LLR penalizzati calcolati nell'operazione di aggregazione; da questa lista si estraggono, in modo casuale, le coppie di possibili genitori e, se sussistono le

condizioni per il *crossing-over*[14], si generano le sequenze ordinate di nodi che identificano i figli, creando una nuova lista di elementi (*offspring list*) ordinati in senso decrescente per i valori di LLR calcolati. Dalla lista ordinata dei potenziali genitori, si estrae il primo 10% di elementi che andrà a costituire una parte della nuova generazione (*new generation list*) mentre il rimanente 90% sarà costituito dalla proporzione equivalente di figli, selezionati a partire dal primo della lista ordinata. Tale operazione esclude, in tal modo, il 10% dei figli ed il 90% dei genitori caratterizzati da una *fitness* peggiore e la cardinalità della nuova generazione coinciderà con I. La mutazione interviene in questa fase della *routine*. Essa sostituisce una singola area scelta a caso, appartenente ad una nuova sequenza, con una nuova zona adiacente ad almeno un'area presente nella sequenza. La frazione totale di zone sostituite definisce il tasso di mutazione. Studi di simulazione hanno evidenziato che, per tassi di mutazione superiore al 5%, la variazione del LLR è inferiore allo 0.1%; per tale motivo la *routine* utilizza un tasso di mutazione pari all'1%. La successione delle operazioni finora descritte consente di identificare una nuova generazione G che sostituisce la popolazione iniziale. L'algoritmo riparte da essa generando un numero η di generazioni successive.

Il numero di *crossing-over* effettuati per una generazione condiziona la variazione del LLR: un numero modesto di incroci (ad esempio 50) impatta immediatamente sul LLR già dalle prime generazioni, rimuovendo subito le combinazioni con un valore basso di LLR mentre un aumento graduale del numero di tentativi di *crossing-over* induce uno spostamento della variazione del LLR verso le generazioni più lontane ma con piccoli incrementi. Simulazioni in cui si è utilizzata una strategia intermedia scegliendo un numero basso di incroci per le prime generazioni ed un numero più elevato per quelle successive non hanno fornito buoni risultati per cui è stato suggerito di utilizzare un unico valore di *crossing-over* fissato a 400.

La generazione dei figli avviene selezionando casualmente dalla popolazione iniziale due genitori A e B identificati da due insiemi non disgiunti di zone connesse, $A \cap B \neq 0$. Si definisce un insieme $C = A \cap B$ ed un insieme connesso massimale $D \subseteq C$, scelto in maniera casuale, che identifica

[14] I due insiemi non devono essere disgiunti

la radice r (o *root*) del grafo. Il primo passo è l'assegnazione di un livello[15] ad ogni nodo appartenente ad un genitore; il livello è rappresentato da un numero naturale progressivo positivo (negativo per i nodi dell'altro genitore) assegnato, in maniera iterativa, a partire da un nodo iniziale. Assegnato il livello 0 ai nodi della radice, si seleziona il genitore A e da esso si estrae, in modo casuale, un nodo v_1, $v_1 \in A - A_0$, adiacente all'insieme $A_0 = D$ e si assegna ad esso il livello 1; si prosegue estraendo un nodo v_2 adiacente all'insieme $A_1 = D \cup \{v_1\}$ tale che $v_2 \in A - A_1$ e si assegna ad esso il livello 2; la procedura prosegue assegnando un numero progressivo all'insieme dei nodi $\{v_1, v_2, ..., v_m\}$ la cui cardinalità è data dal numero di elementi di $A - D$. L'insieme così definito, oltre ai nodi della radice e all'insieme di tutti i collegamenti $(v_{i,} v_k)$, (r, v_k), forma una struttura ad albero orientato T_A relativo al genitore A. L'albero gode di una proprietà fondamentale (*lemma* 1): "*per ciascun nodo* $v_i \in A - D$ *esiste un percorso dalla radice* r *al nodo stesso costituito dall'insieme* $\{v_1, ...v_{i-1}\}$*".* In modo analogo si ottiene un albero $T_{\mathbf{B}}$ per il secondo genitore che gode anch'esso dalla proprietà citata.

La generazione dei figli avviene nel seguente modo: posto con $m_A \geq 2$ e $m_B \geq 1$ (con $m_A \geq m_B$), rispettivamente, il numero di elementi degli insiemi $A - D$ e $B - D$, i figli risultano costituiti da $m_B + (m_A - m_B - 1) = m_A - 1$ sequenze ordinate di nodi il cui numero massimo è pari a $(m_A + m_D)$. Le sequenze di nodi sono identificate dai percorsi:

$(m_A - 1, ..., 1, 0, -1)$
$(m_A - 2, ..., 1, 0, -1, -2)$
$(m_A - m_B, ..., 1, 0, -1, ..., -m_B + 1)$ …….
……..……..…
$(m_A - m_B - 1, ..., 1, 0, -1, ..., -m_B + 1)$
……..…..
$(1, 0, -1, ..., -m_B + 1)$

Oltre al *lemma* 1, per un albero vale anche una seconda proprietà (*lemma* 2): "*tutti gli insiemi di nodi generati da due genitori risultano connessi tra loro*". I due

[15] La procedura di assegnazione di un livello non è unica ma può essere implementa in diversi modi

lemmi consentono di velocizzare la fase di ricerca in quanto non sussiste la necessita di verificare la condizione di adiacenza tra i nodi filiali con un guadagno di tempo stimato di circa il 25% circa rispetto alla procedura SA standard. Inoltre, durante la fase di generazione dei figli per ogni zona aggiunta al *cluster* potenziale è sufficiente aggiungere o sottrarre l'ammontare dei casi e della popolazione per ottenerne rapidamente i rispettivi totali con un ulteriore vantaggio in termini computazionali.

La GA *scan statistic* è stata confrontata con la SA *standard*, con e senza l'uso di una penalizzazione, in merito alla capacità di identificare un insieme di *clusters* fittizi[16]. I risultati non hanno evidenziato differenze significative tra le due procedure in termini di potenza: quest'ultima assume un valore elevato, per entrambi gli approcci, nel caso di *clusters* mediamente irregolari mentre, per aggregazioni particolarmente irregolari, la metodologia GA si rivela più potente ed efficace evidenziando l'opportunità di utilizzare una penalizzazione in fase di ricerca. La differenza principale tra le due metodologie consiste nella velocità di esecuzione dell'algoritmo; a parità di parametri di riferimento, il GA ha una velocità di ricerca fino a dieci volte superiore rispetto alla SA. Il vantaggio computazionale dell'algoritmo genetico deriva in particolare dalla fase di generazione dei figli trattandosi di aree già connesse (per definizione dai lemmi 1 e 2). L'operazione di mutazione, anche se necessita della verifica della condizione di adiacenza per le zone sostituite, non aggrava la fase di ricerca che si rivela comunque più lenta della *spatial scan statistic*.
In sintesi, la procedura GA può essere descritta nei seguenti passi:

1. definizione della matrice di adiacenza $W_{(I\times I)}$ per le aree del territorio
2. selezione dei parametri di non-compattezza, del numero di generazioni e del numero di *crossing-over*
3. scelta casuale di un nodo di partenza v_i
4. definizione dell'insieme di aree connesse che massimizzano il LLR mediante un processo di aggregazione successiva

[16] I dati si riferiscono allo studio di Duczmal *et al*.,2006b in cui sono stati ipotizzati 11 *clusters* di forma diversa

5. ripetizione del punto (4), per ogni vertice v_i, fino alla formazione della popolazione iniziale (*current generation list*)
6. ordinamento decrescente dei valori di LLR ottenuti per la popolazione iniziale
7. selezione casuale delle coppie di potenziali genitori mediante una valutazione della funzione *fitness* (LLR)
8. generazione dei figli (*offspring list)* tramite un *crossing-over* tra coppie di genitori scelte a caso dalla popolazione iniziale
9. calcolo del LLR per le sequenze dei figli ed ordinamento della lista in senso decrescente per i valori di LLR
10. estrazione del 10% di elementi (partendo dal primo elemento della lista ordinata) appartenenti alla popolazione iniziale
11. estrazione del 90% di elementi (partendo dal primo elemento della lista ordinata) appartenenti alla lista ordinata dei figli generati al punto (8)
12. sostituzione della popolazione iniziale (*current generation list*) con la nuova popolazione (*new generation list*) ottenuta combinando il 10% dei genitori, scelti al passo (10), ed il 90% dei figli ,selezionati al passo (11)
13. operazione di mutazione per una frazione casuale di elementi appartenenti alla nuova generazione. Si sostituisce una zona con una nuova area adiacente in modo casuale
14. procedura di simulazione Monte Carlo dei casi attesi sotto l'ipotesi nulla
15. verifica della significatività statistica dei *clusters* mediante i risultati ottenuti dal processo di simulazione MC

3.2.4 - *Flexible spatial scan statistic*

La metodologia di *cluster detection* implementata da Tango e Takahashi (2005) si propone come un algoritmo di scansione completa del territorio che consente di identificare aggregazioni di forma irregolare e, contemporaneamente, di limitarne l'estensione entro limiti geografici ridotti. Analogamente alle procedure sinora descritte, l'area in esame è suddivisa in un

numero finito I di zone per ognuna delle quali si conosce il numero di casi osservati e la popolazione; per l'i-esima zona, la variabile casuale che descrive il numero di casi osservati è indicata con C_i e, sotto l'ipotesi nulla H_0, tali variabili sono indipendenti ed identicamente distribuite secondo un modello di Poisson. L'ipotesi alternativa H_1 postula l'esistenza di almeno una zona z dell'intero territorio in cui il rischio interno è superiore a quello delle zone rimanenti. Da un punto di vista formale, si ha :

$$H_0 : p \equiv q \qquad \text{per ogni zona } z$$

$$H_1 : z \in Z, p > q \qquad \text{per almeno una zona } z$$

Per ogni finestra irregolare z, si determina il valore della funzione di verosimiglianza sotto le due ipotesi e si calcola la statistica-test LLR (dove n_z* indica il numero di casi attesi):

(22)
$$Ts = \frac{\max_z L(z)}{L_0} = \frac{\left(\frac{c_z}{n_z*}\right)^{c_z}\left(\frac{C-c_z}{N-n_z*}\right)^{C-c_z}}{\left(\frac{C}{N}\right)^{C}}$$

La finestra z^* che massimizza la (22) rappresenta il *cluster* più probabile. Sotto H_0, la distribuzione del rapporto LLR è determinata da una procedura di simulazione Monte Carlo che consente di valutare la significatività statistica delle aggregazioni definite.

Per ciascuna area del territorio, la *flexible spatial scan statistic* (o *flexibly shaped spatial scan statistic*) utilizza una finestra di ricerca irregolare z, di dimensione variabile k, costituita dalle k zone del territorio connesse geograficamente all'area considerata. Il valore k varia da 1 ad un limite superiore prefissato K che identifica la dimensione massima del *cluster*. Fissato un centroide, si genera un intorno spaziale della cella, inclusa la zona stessa, comprendente le $(K-1)$ celle ad esse più vicine in termini di distanza crescente dal suo centroide; partendo da tale intorno, si costruisce progressivamente una serie di insiemi di

zone connesse ciascuno dei quali identifica un *cluster* potenziale[17]. La ricerca di tali aree avviene in maniera esaustiva sull'intero territorio analizzando tutte le possibili aggregazioni entro il raggio prefissato. In genere, la dimensione massima del *cluster* è limitata a valori inferiori a 25 in quanto K condiziona fortemente la fase computazionale soprattutto quando il numero I di zone è elevato, ad esempio superiore a 200.

Posto con $z_{ik(j)}$ (con $j = 1,... j_{ik}$) la j-esima finestra di ricerca costituita dalle k aree connesse alla zona di partenza i, dove j_{ik} è il valore di j che soddisfa la relazione $z_{ik(j)} \subseteq z_{ik}$, l'insieme completo delle aree potenziali da valutare è dato da:

$$Z_{fin} = \left\{ z_{ik(j)} \middle| 1 \le i \le I, 1 \le k \le K, 1 \le j \le j_{ik} \right\} \tag{23}$$

La metodologia *flexible scan statistic* presenta alcune limitazioni. Essa consente di modellare solo dati di conteggio e di valutare solo la presenza di zone caratterizzate da un rischio elevato del fenomeno, pur definendo *clusters* secondari non sovrapposti a quello principale. Il limite principale della metodologia è però legato alla dimensione K del *cluster* finale. Anche se una ridotta dimensione consente di prevenire la formazione di aggregazioni spaziali che limitano l'inclusione di aree non caratterizzate da un rischio elevato, risulta computazionalmente oneroso estendere la ricerca ad un numero di zone superiore a 30 per la crescita esponenziale dei tempi di elaborazione. Per tale motivo, la *flexible scan statistic* si dimostra più efficace negli studi in cui si ipotizza la presenza di *clusters* di modeste dimensioni. E' opportuno sottolineare, inoltre, che la scelta di K deve essere rapportata allo specifico territorio esaminato e al numero di zone in cui esso è suddiviso. In alcuni casi, la dimensione del *cluster* può condizionare la geometria del *cluster* stesso. Uno studio di simulazione in cui sono state confrontate le metodologie *spatial scan statistic*, *simulated annealing* e la *flexible scan statistic,* ha evidenziato la difficoltà della procedura di Tango nell'identificazione di aggregazioni di forma

[17] La procedura impiega il concetto di adiacenza di primo-ordine e le zone connesse sono rappresentate da una matrice di adiacenza costruita preliminarmente alla procedura di ricerca

allungata, pur evidenziando una buona potenza di ricerca nelle diverse ipotesi di indagine.

In sintesi, l'algoritmo di ricerca può essere descritto dai seguenti passi:

1. definizione della matrice di adiacenza $W_{(I\times I)}$ per le aree esaminate
2. scelta casuale di una area di partenza i
3. costruzione dell'insieme M_i contenente le $K-1$ aree comprese in un intorno dell'area i, ordinato in senso decrescente di distanza dal suo centroide, $M_i = \{i_0, i_1, ...i_{K-1}\}$
4. generazione di tutti i possibili sottoinsiemi $z_{ik} \subset M_i$ comprendenti l'area di partenza i
5. scelta casuale di uno dei possibili sottoinsiemi z_{ik}
6. suddivisione dell'insieme z_{ik} in due sottoinsiemi disgiunti: z_{iko}, contenente la zona iniziale i, e z_{ik1}, costituito dalle zone rimanenti che appartengono a z_{ik}
7. creazione di due nuovi sottoinsiemi, $\overline{z}_{iko}$ e $\overline{z}_{ik1}$, contenenti, rispettivamente, le zone connesse *ad almeno* qualche area di z_{iko} e le zone rimanenti. Essi sostituiscono gli insiemi iniziali z_{iko} e z_{ik1}
8. ripetizione del punto (7) fino a quando almeno uno dei due sottoinsiemi risulta vuoto
9. se diventa vuoto l'insieme $\overline{z}_{iko}$, si parla di insieme sconnesso e z_{ik} viene scartato perchè non rappresenta un *cluster* potenziale; viceversa, se diventa vuoto $\overline{z}_{ik1}$ l'insieme finale risultante dall'unione degli insiemi $\overline{z}_{iko}$, fino ad ora definiti, costituisce una zona connessa $z_{ik(j)}$ che viene aggiunta all'elenco Z_{fin} dei possibili *clusters.*
10. ripetizione dei punti da (5) a (9) per ogni insieme $z_{ik} \subset M_i$
11. ripetizione della procedura dal punto (2) al (10) fino alla definizione completa dell'insieme Z_{fin}
12. calcolo della statistica-test per ogni elemento dell'insieme Z_{fin}
13. ordinamento decrescente dei valori di LLR ottenuti
14. simulazioni Monte Carlo dei casi attesi sotto l'ipotesi nulla
15. verifica della significatività statistica dei *clusters* mediante i risultati ottenuti dal processo di simulazione

3.2.5 - *Greedy Growth Search*

La *Greedy Growth Search* GGS (Yiannakoulias *et al.,*2007) è una recente proposta metodologica di *cluster detection* per aggregazioni spaziali irregolari, nata da un ibrido tra gli algoritmi di *simulated annealing* (Duczmal e Assunção, 2004), *flexible scan statistic* (Tango e Takahashi,2005) e *spatial scan statistic* (Kulldorff,1995,1997,1999), con l'introduzione di due nuovi parametri destinati al controllo dell'andamento geometrico dei *clusters* e dei tempi di elaborazione.

Si consideri un'area geografica suddivisa in I poligoni adiacenti e distinti. Dalla teoria dei grafi, un poligono è individuato dal vertice v_i di un grafo ed è connesso ad uno o più poligoni mediante opportuni collegamenti, $e_1, e_2, ...e_I$, che ne definiscono le adiacenze e riflettono la topologia della regione esaminata. La metodologia GGS utilizza una connettività di primo ordine tra le zone del territorio e le ipotesi utilizzate per la definizione dei *clusters* coincidono con quelle illustrate nelle procedure precedenti. La variabile casuale che descrive i casi osservati all'interno di una zona i si distribuisce secondo un modello di Poisson, l'ipotesi nulla H_0 postula l'uniformità distributiva del fenomeno sul territorio mentre quella alternativa H_1 postula l'esistenza di almeno una zona del territorio in cui il rischio interno è significativamente superiore a quelle delle rimanenti aree. La statistica-test è ancora il rapporto delle funzioni di verosimiglianza calcolate sotto le due ipotesi. Si parte da una zona del territorio scelta a caso e si procede, in maniera sequenziale, aggregando le aree adiacenti che massimizzano il rapporto LLR definendo, ad ogni passo della *routine,* un *cluster* potenziale; tale operazione viene ripetuta per ogni zona del territorio. Tra tutte le aggregazioni così definite, l'insieme di aree connesse che massimizza il valore della statistica-test rappresenta il *cluster* più probabile e la sua significatività statistica è valutata mediante un algoritmo di simulazione MC.

In dettaglio, la procedura GGS opera nel seguente modo. Per ogni area[18] del territorio, ciascuna rappresentata da un vertice v_i, si costruisce una serie di insiemi z_i di zone connesse ciascuno dei quali identifica un possibile c*luster*;

[18] In questo contesto, i termini area, nodo e vertice sono utilizzati come sinonimi per identificare una specifica porzione di territorio

indicato con V l'insieme completo dei vertici, si definiscono progressivamente due insiemi di aree, z_{im} ed il suo complementare $\bar{z}_{im} = V - z_{im}$, $i = 1,...,I$, dove z_{im} contiene le aree adiacenti alla zona di partenza che apportano la variazione maggiore in termini di LLR durante il processo di aggregazione. Al primo *step,* l'insieme z_{i0} contiene solo il vertice v_i di partenza, $z_{i0} = \{v_i\}$, mentre, ad un passo m della *routine,* l'insieme risulta formato dalle aree aggregate fino al passo precedente, aggiunto di un nuovo vertice, $z_{im} = z_{i(m-1)} \cup \{v^*_m\}$. Il nodo $\{v^*_m\}$ candidato ad essere scelto in un generico passo della procedura è individuato da:

(24) $$v^*_m = \arg_{x \in \bar{B}_{i(m-1)}} \max\left[LLR(z_{i(m-1)} \cup x)\right]$$

dove l'insieme $\bar{B}_{i(m-1)} \subseteq \bar{z}_{i(m-1)}$ è un sottoinsieme di aree appartenenti a $\bar{z}_{i(m-1)}$, i cui elementi risultano connessi ad almeno un vertice di $z_{i(m-1)}$. La procedura si arresta quando si raggiunge un limite prefissato espresso come percentuale di popolazione a rischio o quando tutti vertici del grafo risultano inclusi nell'insieme. Supposto che il numero massimo di aree aggregate per una zona v_i sia espresso dal valore M_i, l'insieme dei possibili *clusters* associati è dato da $Z_i = \{z_{i0}, z_{i1},....., z_{iM_i}\}$; al termine della fase di scansione l'insieme $Z_{fin} = \{Z_1, Z_2,.....,Z_I\}$. di cardinalità I, conterrà tutti i *clusters* di potenziale interesse. L'insieme $Z_i \subseteq Z_{fin}$ avente il valore più alto del LLR rappresenta il *Most Likely Cluster* (MLC) la cui significatività statistica viene valutata mediante un algoritmo di simulazione Monte Carlo. Da quanto illustrato si deduce che la procedura di definizione dei potenziali *clusters* è una sorta di ibrido tra le tecniche di Tango (*flexible scan statistic*) e di Duczmal (*simulated annealing*). Il processo di aggregazione successiva delle aree avviene con un criterio deterministico, selezionando l'area che apporta la variazione maggiore in termini di LLR, mentre la definizione degli insiemi z_{im} e del suo complementare è in qualche modo simile al metodo definito da Tango.

La metodologia proposta da Yiannakoulias utilizza in fase di ricerca due nuovi criteri per controllare l'andamento dei *clusters* irregolari e prevenire la formazione di aggregazioni molto estese. I criteri introdotti riguardano un parametro di "non-connettività" (*non-connectivity penalty*) e un parametro di

"profondità" (*depth limit*). La prima penalizzazione si basa sul concetto di non-compattezza già sviluppato nella *elliptic scan statistic* (Kulldorff *et al.*,2006a) e, successivamente, ripreso nella tecnica di *simulated annealing* (Duczmal *et al.*,2006b); il secondo parametro, invece, è stato introdotto con l'obiettivo di velocizzare la *routine* nelle fasi in cui si ha una variazione minima del LLR.
Il parametro di non-connettività, indicato con $K(z)$, tiene conto della struttura geometrica di connessione tra i vertici ed è definito dal rapporto tra il numero $e(z)$ di collegamenti presenti ("osservati") in un *cluster* z ed il numero di collegamenti possibili ("attesi") nello stesso *cluster*[19]:

$$K(z) = \frac{e(z)}{3(v(z)-2)} \tag{25}$$

Analogamente a quanto visto per la *spatial scan statistic* e la metodologia SA, $K(z)$ viene utilizzato per penalizzare la statistica-test

$$Penalized \;\; Likelihood \;\; Ratio \;\; (z,\alpha) = LLR(z)^{K(z)^{\alpha}} \tag{26}$$

mediante un parametro scalare $\alpha(\geq 0)$ definito a priori dall'utente. L'idea di base è simile a quella del parametro di non-compattezza ma, in questo caso, si considera il numero di collegamenti presenti in un *cluster* piuttosto che la sua forma geometrica: le aggregazioni spaziali caratterizzate da un numero elevato di collegamenti tra i vertici assumono una forma geografica più compatta. L'utilizzo di una penalizzazione favorisce l'identificazione di *clusters* più regolari prevenendo la formazione di aggregazioni con un andamento estremo o particolarmente irregolare: per valori di $\alpha \to 0$, la penalità ha una scarsa efficacia sulla statistica-test e, di contro, un effetto maggiore per valori di α crescenti.
Il parametro di profondità, indicato con la lettera u, ha come obiettivo una riduzione dei tempi di elaborazione nelle situazioni in cui la variazione del LLR risulta contenuta; esso può assumere valori da 0 ad I e deve essere fissato a

[19] Questo parametro viene spesso utilizzato come indice di connessione per la stima della quantità di scambi funzionali possibili in un paesaggio

priori dall'utente. Se la finalità di una ricerca è l'individuazione di aree caratterizzate da un livello elevato di rischio, può risultare conveniente arrestare la procedura quando si è nelle cosiddette "valli" del LLR, ovvero quando il valore del LLR non subisce incrementi significativi con l'aggiunta di nuove zone. Un approccio simile può però penalizzare l'identificazione di *clusters* con un livello modesto del fenomeno risultando, in alcuni casi, parzialmente definiti o totalmente esclusi. Il parametro di profondità consente di valutare se, dopo un certo numero di tentativi o visite (il cui valore è espresso da u), si ottiene ancora un aumento consistente del LLR. In caso contrario, si interrompe il processo di ricerca per il nodo v_i e si seleziona un nuovo vertice. In caso di interruzione della procedura, la cardinalità dell'insieme z_{im} potrebbe non coincidere con il valore limite di popolazione imposto a priori ma essere definita dal valore M_i al momento dell'interruzione, definendo aggregazioni spaziali di dimensioni inferiori. Valori crescenti di profondità u incrementano i tempi di elaborazione e consentono di includere nel *cluster* anche le aree che apportano una variazione minima del LLR e la sua dimensione finale sarà presumibilmente limitata dalla percentuale di popolazione massima fissata. Una profondità pari al numero totale di aree I equivale, invece, ad effettuare una ricerca libera dei *clusters* senza nessun controllo sulla statistica-test e ciò potrebbe identificare aggregazioni spaziali molto estese o dalla forma geometrica bizzarra.

La procedura è stata valutata in uno studio di simulazione in cui sono stati ipotizzati 4 *clusters* di forme diverse, posizionati su un territorio suddiviso in 203 aree esagonali in cui la popolazione è stata considerata fissa (1000 o 10000 individui): circolare, a forma di "X", 2 piccoli *clusters* separati e uno a forma di anello. E' stato utilizzato un *range* di valori (da 0 a 4 con incrementi unitari) per i parametri u ed α e il successo della ricerca è stato valutato in termini di sensibilità e di valore predittivo positivo (PPV). Nell'ipotesi con due *clusters*, i risultati hanno evidenziato una maggiore sensibilità e PPV nel caso di popolazione più estesa e, per valori crescenti dei due parametri, anche se, in linea di massima, i risultati sono pressoché simili. Per il *cluster* ad "X", la sensibilità ed il PPV non variano di molto al variare della popolazione e dei due parametri: il PPV per la popolazione maggiore (10000 individui) sembra risentire dell'influenza del parametro di non-compattezza. Al crescere di α il PPV tende a

diminuire in accordo con l'ipotesi di fondo di tale parametro. Nel caso di *cluster* ad anello, i risultati appaiono simili allo scenario precedente ma, per popolazione elevata, il PPV tende ad aumentare al diminuire di α ; un *cluster* ad anello, infatti, ha la stessa forma di un *cluster* circolare anche se al suo interno presenta zone non comprese nel *cluster*. Infine, nello scenario circolare, la procedura si mostra perfettamente idonea alla individuazione del *cluster* imposto determinando i valori di sensibilità e PPV maggiori.

Il limite principale del metodo illustrato è la selezione dei due parametri di controllo e la mancanza di informazioni a priori sul fenomeno rende difficile e soggettiva tale scelta, condizionando il risultato finale. Anche se gli studi di simulazione[20] forniscono una prima indicazione per un uso corretto dei parametri, occorre tener conto della realtà territoriale esaminata per la selezione di tali valori.

In generale l'algoritmo di ricerca *greedy growth search* può essere descritto come segue:

1. definizione della matrice di adiacenza $W_{(I\times I)}$ per le aree del territorio
2. selezione dei valori dei due parametri di controllo u e α
3. scelta casuale di un'area (o vertice) di partenza v_i
4. costruzione degli insiemi di vertici connessi z_{im} e $\overline{z}_{im}$ in funzione dei parametri u e α selezionati
5. determinazione del LLR per ogni *cluster* potenziale
6. ripetizione dei punti (3) ,(4) e (5) per ogni vertice v_i fino alla formazione dell'insieme finale Z_{fin} dei possibili *clusters*
7. ordinamento decrescente dei valori di LLR ottenuti
8. simulazioni Monte Carlo dei casi attesi sotto l'ipotesi nulla
9. verifica della significatività statistica dei *clusters* mediante i risultati ottenuti dal processo di simulazione

[20] L'algoritmo è stato valutato solo su dati simulati e quindi manca di qualsiasi riferimento in applicazioni su dati reali

Capitolo 4

Longevità in Emilia-Romagna: risultati delle tecniche di *clustering* spaziale

4.1 - Descrizione del territorio

L'Emilia-Romagna è una regione dell'Italia settentrionale, con capoluogo Bologna, caratterizzata da un andamento geografico stretto e allungato. Essa è formata dall'unione di due regioni storiche: l'Emilia, che comprende le province di Piacenza, Parma, Reggio, Modena, Ferrara e buona parte della provincia di Bologna, e la Romagna, con le province di Ravenna, Rimini, Forlì-Cesena e la parte orientale della provincia di Bologna (Imola ed il suo circondario). Il territorio regionale è prevalentemente pianeggiante; le sue pianure occupano il 47.8% della regione mentre la zona collinare, con un'estensione di circa il 27.1%, e l'area montana, che occupa il rimanente 25.1%, si trovano nella parte meridionale della regione. La regione è suddivisa in 9 province per un totale di 341 comuni[21] e la provincia con il maggior numero di comuni è quella di Bologna con 60 comuni seguita dalle province di Piacenza (48), Parma (47) e Modena (47) mentre le province di Ravenna (18) e Rimini (20) rappresentano le aree caratterizzate dal minor numero di ripartizioni comunali (tabella *app01* in appendice).

Nel periodo 2000-2004, la popolazione (media) residente in Emilia-Romagna è di circa 4 milioni di abitanti con una maggiore concentrazione nelle province di Bologna (925.764 abitanti) e Modena (638.784) mentre, a fronte di un numero elevato di comuni, la provincia di Piacenza si distingue come l'area meno popolata della regione. La suddivisione per sesso mostra una leggera prevalenza delle donne (2.078.824 abitanti) rispetto agli uomini (1.958.776 abitanti) e tale differenza rimane pressoché invariata in ogni area provinciale. La distribuzione dei comuni per numero di residenti evidenzia una proporzione del 48% di aree con un ammontare di popolazione (media) inferiore ai 5.000 abitanti; in particolare il 43% delle aree comunali ha una popolazione compresa tra 1.000 e 5.000 abitanti mentre solo il 14% di esse è caratterizzata da una popolazione superiore alle 15.000 unità (tabelle *app02* e *app03* in appendice). Il territorio dell'Emilia-Romagna si presenta, dunque, suddiviso in un numero elevato di aree comunali in cui la popolazione assume un'entità modesta nella metà di essi. La popolazione ritenuta longeva ai fini della nostra ricerca è

[21] I riferimenti numerici riportati sono relativi al periodo 2000-2004

identificata dagli individui con età superiore o uguale a 95 anni (95+) e, analogamente alla popolazione residente, sono stati utilizzati i rispettivi valori medi nel quinquennio in esame. La popolazione totale 95+ è di 6.813 individui con una netta prevalenza delle donne (5.509) rispetto agli uomini (1.304) ed un rapporto femmine/maschi (95+) di circa 4.2, in linea con i valori riportati dalle statistiche nazionali. La popolazione con età superiore o uguale a 100 anni invece, si riduce drasticamente rispetto alla classe 95+: si passa dai 6.813 individui 95+ ai 513 individui ultra centenari nell'intera regione. La prevalenza delle donne è ancora più accentuata con 505 femmine contro i 77 individui di sesso maschile ed un rapporto femmine/maschi (100+) pari a 6.6; in particolare, nella provincia di Rimini risultano solo 3 abitanti di sesso maschile appartenenti a questa fascia di età. La popolazione totale di età compresa tra 55 e 59 anni al censimento del 1961 è di 211.218 individui, suddivisa in 102.514 maschi (48%) e 108.704 femmine (52%); i comuni di Migliaro, Goro e Tresigallo della provincia di Ferrara non sono riportati nelle tavole del censimento del 1961 per cui è stato necessario sostituire i dati mancanti scegliendo il valore medio della popolazione di età 55-59 anni residente nei comuni ferraresi aventi una popolazione media inferiore ai 5.000 abitanti, in quanto i tre comuni appartengono a tale categoria.

La distribuzione del *Centenarian Rate* (CR) per la popolazione complessiva evidenzia valori (medi) superiori a quello (medio) regionale (CR=0.032) per le province di Ravenna (CR=0.044), Bologna (CR=0.036) e Forlì-Cesena (CR=0.034) mentre i comuni della provincia di Ferrara risultano caratterizzati da valori molto bassi dell'indicatore (CR=0.021) evidenziando una scarsa propensione alla longevità. Analizzando il fenomeno per sesso, le aree della costa adriatica si distinguono ancora per valori elevati del CR in entrambi i sessi: la provincia di Ravenna è caratterizzata dai valori più elevati dell'indicatore rispetto a tutte le altre province per i due i sessi (CR_{maschi}=0.019; $CR_{femmine}$=0.070). I comuni bolognesi si evidenziano per un elevato valore di longevità femminile ($CR_{femmine}$=0.058) mentre la provincia di Ferrara si conferma l'area regionale con una longevità ridotta anche nella suddivisione per sesso (CR_{maschi}=0.007; $CR_{femmine}$=0.034). In tabella 1 sono riportate le statistiche

relative al CR per provincia di residenza e sesso; il valore medio provinciale è ottenuto come media dei valori di CR calcolato per ciascun comune.

Tabella 1 – Distribuzione del CR 95+ per la provincia e sesso

Provincia		min	max	mean	sd	cv	p25	p50	p75
Piacenza	cr95	.0086957	.0588235	.0305044	.012146	.3981736	.0219475	.027687	.0387816
	cr95m	0	.08	.0146121	.0144124	.9863328	.0077879	.0111166	.0164332
	cr95f	.009901	.0846154	.0467386	.0183961	.3935957	.0362948	.0474782	.0570461
Parma	cr95	.0136986	.0618557	.0300963	.0095432	.3170886	.0232558	.0299065	.0358705
	cr95m	0	.0265487	.0100394	.0067707	.6744124	.0047393	.0106383	.0144231
	cr95f	.0217391	.12	.0503756	.0173047	.343513	.0397351	.0472973	.0625
Reggio Emilia	cr95	.0118343	.0518519	.0302447	.0084348	.2788844	.0250784	.0285714	.0364299
	cr95m	0	.0310078	.0102410	.006831	.6670269	.0057803	.0104712	.014881
	cr95f	.0211268	.0803571	.0496108	.0150538	.3034384	.0387597	.0466667	.0588235
Modena	cr95	.0086957	.0613108	.0317255	.0092925	.292902	.0252366	.0321156	.0380952
	cr95m	0	.0316456	.0129378	.0076358	.5901943	.0077519	.0124224	.0181818
	cr95f	.0074627	.0986547	.0506749	.0176711	.3487159	.038835	.0480000	.0636132
Bologna	cr95	.0077519	.0788644	.0361926	.0130871	.3615964	.0262011	.0349244	.04341
	cr95m	0	.0416667	.0147016	.0088715	.6034408	.0089314	.0127518	.0208356
	cr95f	.016	.1418919	.0581624	.0229648	.394839	.0426283	.0547386	.0724694
Ferrara	cr95	.0039841	.0393701	.0210840	.0076556	.3630992	.0164502	.0210897	.023845
	cr95m	0	.0164835	.0073738	.0054926	.7448797	.0041667	.0069048	.0125
	cr95f	.0088496	.0603015	.0343762	.0125037	.3637307	.0272727	.0342242	.040107
Ravenna	cr95	.0229358	.0641026	.0442519	.0107364	.2426201	.0377604	.0443033	.0521008
	cr95m	0	.0392157	.0190027	.0082512	.4342132	.0166667	.0179052	.0236686
	cr95f	.0353982	.1176471	.0699435	.0201155	.287596	.0573123	.0692084	.0754717
Forlì-Cesena	cr95	.0132743	.0628931	.0341040	.0122718	.3598356	.0253807	.0341723	.0432099
	cr95m	0	.0506329	.0156438	.0101545	.6491107	.0100251	.0133353	.0196078
	cr95f	.0178571	.0970464	.0522788	.0212019	.4055543	.0357143	.0522592	.0697674
Rimini	cr95	.009434	.0652174	.0312772	.0140487	.4491675	.020274	.0294813	.039501
	cr95m	0	.0396825	.0159398	.0119079	.7470548	.0057637	.0159086	.0250156
	cr95f	.0181818	.1168831	.0468307	.0230544	.4922918	.0301587	.0437980	.0579259
Regione ER	cr95	.0039841	.0788644	.0319525	.0117936	.3690968	.0240385	.0305992	.0393701
	cr95m	0	.08	.0130385	.0096639	.7411837	.0073529	.0119048	.0172414
	cr95f	.0074627	.1418919	.0509466	.0200846	.394229	.037037	.0480000	.0636132

4.2 - Risultati delle tecniche di *clustering* spaziale

La prima metodologia utilizzata è la *spatial scan statistic* SSS[22]. E' opportuno ricordare che non si dispone di nessuna informazione a priori in merito al fenomeno esaminato per cui la ricerca ha l'obiettivo primario di fornire indicazioni sulla presenza di eventuali aree con una maggiore propensione alla

[22] La tecnica è stata implementata in un software, *SatSCAN v.7.x*, disponibile on-line in versione gratuita.

longevità e di visualizzarne la localizzazione geografica. Le analisi effettuate sono di tipo puramente spaziale senza l'utilizzo di covariate nel modello distributivo e sono state condotte distintamente per sesso. Il modello di probabilità utilizzato è quello di Poisson in cui i casi osservati sono rappresentati dagli individui di età 95+ residenti in ciascun comune mentre i casi attesi sono stati determinati mediante una standardizzazione indiretta della popolazione a rischio: sotto l'ipotesi nulla di distribuzione uniforme del fenomeno e in assenza di covariate, il numero di casi attesi in ciascuna area è ottenuto moltiplicando la popolazione a rischio per il tasso di longevità dell'intero territorio. Le caratteristiche appena illustrate sono comuni a tutte le metodologie utilizzate mentre i parametri di ricerca sono stati variati allo scopo di cogliere le diverse ipotesi di localizzazione geografica del fenomeno (appendice *app12* e *app1*3). In virtù della conformazione territoriale della regione Emilia-Romagna si è ritenuto procedere impiegando nella procedura SSS, oltre allo standard circolare, anche una finestra di ricerca ellittica per l'identificazione dei *clusters*.

Uno dei parametri principali di ricerca è la dimensione massima del *cluster* finale, in genere, espressa come percentuale della popolazione a rischio presente sull'intero territorio: fissando un limite al 10%, l'eventuale *cluster* risulterà costituito da un insieme di comuni in cui la somma delle singole popolazioni a rischio è inferiore o uguale al 10% della popolazione totale. Nello studio, si è scelto di utilizzare inizialmente i valori 10%, 15%, 20% e 30% ma, in una fase successiva, tali limiti sono stati ridotti al 5% e all'1% in considerazione della rarità del fenomeno esaminato. La *spatial scan statistic* si distingue per la sua velocità di esecuzione[23]: i tempi di ricerca risultano estremamente ridotti per i *clusters* di forma circolare anche con un numero di replicazioni elevato (99999). Nel caso di aggregazioni spaziali ellittiche, la *routine* evidenzia prestazioni peggiori rispetto a quelle circolari, dovute ai tempi richiesti nella fase di costruzione dell'ellisse; nello studio sono stati utilizzati tre valori di penalizzazione, 0 (nessuna penalità), 0.5 (penalità intermedia) e 1 (penalità forte).

[23] In questo studio, è stato utilizzato un notebook Toshiba Satellite M70-244 Centrino 1.7 Ghz, 1Gb RAM, HD 80Gb 7200 rpm

Prima di illustrare alcune mappe relative ai *clusters* identificati, si riportano le tabelle dei risultati ottenuti con i due criteri di ricerca (circolare ed ellittico). I valori si riferiscono al *cluster* primario caratterizzato da un eccesso significativo di individui longevi attesi.

Tabella 2 - Risultati SSS (analisi circolare, solo *cluster* primario)

Popolazione esaminata	Dim.max Popolazione	Repliche	LLR	Rischio relativo	N° comuni cluster	Popolazione cluster	Casi cluster	p-value
Maschi 95+	1%	9999	4.993	2.189	3	761	21	0.7685
"	3%	"	8.971	1.845	18	2612	60	0.0441
"	5%	"	10.885	1.664	26	5125	105	0.0079
"	10%	"	18.418	1.657	23	9623	191	"
"	15%	"	20.622	1.585	27	14443	269	"
"	20%	"	21.922	1.532	59	20121	355	"
"	25%	"	29.873	1.591	85	25182	445	"
"	30%	"	32.499	1.612	94	26488	469	"
Maschi 95+	50%	9999	32.499	1.612	94	26488	469	0.0001
Femmine 95+	1%	9999	18.068	2.175	3	704	77	0.0001
"	3%	"	18.058	2.175	3	704	77	"
"	5%	"	20.989	1.451	26	4996	360	"
"	10%	"	32.990	1.397	33	10867	740	"
"	15%	"	40.194	1.397	47	13682	934	"
"	20%	"	42.327	1.342	62	21542	1372	"
"	25%	"	45.025	1.339	73	24378	1537	"
"	30%	"	45.025	1.339	73	24378	1537	"
Femmine 95+	50%	9999	45.083	1.339	73	24378	1537	0.0001

In riferimento all'analisi circolare, si può notare che, per un limite di popolazione superiore al 20%, i risultati appaiono pressoché simili avvalorando ulteriormente l'ipotesi iniziale di limitare la dimensione del *cluster* finale a valori inferiori ad esso. La figura 2 riporta i *clusters* identificati per la popolazione maschile fissando una dimensione massima di popolazione al 5%. Il *cluster* primario si posiziona geograficamente tra le province di Ravenna e Bologna e si estende per un breve tratto nella provincia di Forlì-Cesena: esso comprende 26 comuni per una popolazione totale a rischio di 5125 individui e 105 casi osservati (RR=1.664;p=0.0079). Le aree incluse nel *cluster* sono rappresentate da 19 comuni bolognesi, appartenenti al Circondario imolese e alla fascia pedemontana, da 1 comune nei pressi di Forlì (Modigliana) e 6 comuni della provincia ravennate. L'aggregazione riportata in rosso nella mappa, invece, si riferisce ad un *cluster* secondario caratterizzato da un rischio ridotto di longevità (RR=0.394;p=0.0096) costituito da 15 comuni della provincia di Ferrara che,

come vedremo anche nel caso ellittico, si distingue sempre come un'area a ridotta longevità[24]. Diminuendo la percentuale di popolazione massima, il *cluster* primario tende a localizzarsi tra i comuni della provincia di Bologna appartenenti alla fascia montana.

Analizzando la popolazione femminile, la situazione nel complesso non appare diversa da quella maschile identificando ancora un *cluster* primario tra le province di Ravenna, Bologna e Forlì-Cesena. Fissando una proporzione massima di popolazione al 5%, il *cluster* primario è costituito da 26 comuni con una popolazione complessiva a rischio di 4996 individui e 360 donne 95+ (RR=1.451;p=0.0001): 19 comuni appartengono alla provincia di Bologna, distribuiti tra la fascia pedemontana e la zona imolese, 1 comune appartiene alla provincia di Forlì-Cesena (Modigliana) e 6 comuni alla provincia ravennate (figura 3). La variazione della dimensione di popolazione identifica *clusters* geograficamente posizionati tra le province indicate mentre, riducendo la frazione di popolazione all'1%, si identifica un aggregazione formata da soli 3 comuni della provincia di Bologna: San Lazzaro di Savena, Sasso Marconi e Pianoro, con un totale di 704 individui a rischio e 77 donne con età 95+ (RR=2.175;p=0.0001).

[24] Nelle analisi di mortalità regionale, la provincia di Ferrara si caratterizza per un tasso di mortalità superiore a quello medio regionale

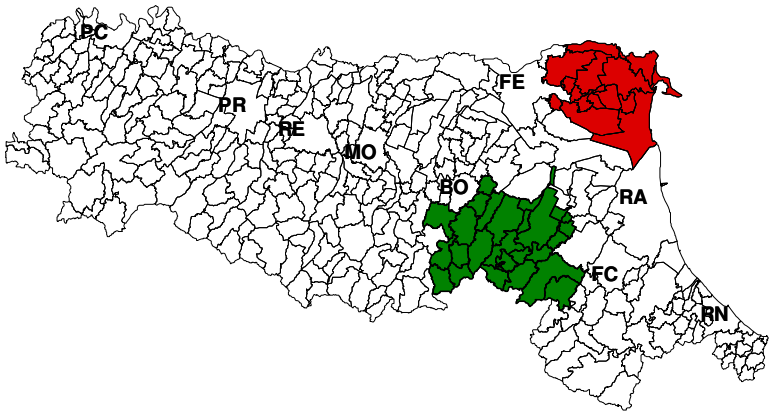


Figura 2 – Mappa CR95+ Maschi SSS (5% dimensione massima di popolazione).
Cluster circolare. 9999 simulazioni.
(verde=elevata longevità; rosso=ridotta longevità)

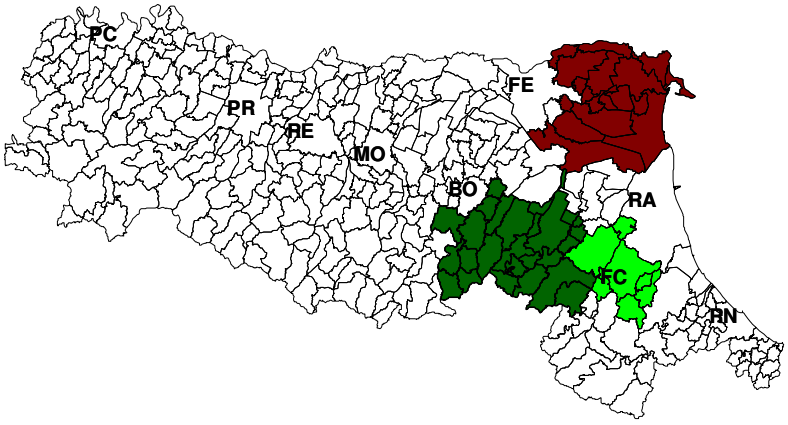


Figura 3 – Mappa CR95+ Femmine SSS (5% dimensione massima di popolazione)
Cluster circolare. 9999 simulazioni.
(verde=elevata longevità; rosso=ridotta longevità (secondario);
verde chiaro=elevata longevità (secondario))

Nel caso di *clusters* di forma ellittica, i risultati finali non si discostano in maniera significativa da quelli ottenuti per le aggregazioni circolari. All'aumentare del parametro di penalizzazione si osserva una diminuzione del valore del LLR e la formazione di *clusters* dalla forma geometrica più compatta, a conferma dell'azione di controllo della penalità sulla geometria del *cluster,* anche se il numero di aree incluse non segue tale andamento. In particolare, per

la popolazione femminile, si nota che il risultato è pressoché simile per tutte le dimensioni di popolazione superiori al 10% in assenza di penalizzazione (a=0) e che la variazione della penalità da 0.5 ad 1 non influisce sull'identificazione del *cluster* finale (tabella 3).

Tabella 3 - Risultati SSS (analisi ellittica, solo *cluster* primario)

Popolazione Esaminata	Dim.max Popolazione	Repliche	LLR[1]	Rischio relativo	Penalità[2]	N° comuni cluster	Popolazione cluster	Casi cluster	p-value
Maschi 95+	1%	9999	7.090	2.427	0.0	5	786	24	0.6524
"	1%	"	6.675	2.536	0.5	4	689	22	0.4940
"	1%	"	6.293	2.536	1.0	4	689	22	0.5108
"	3%	"	9.090	1.896	0.0	17	2370	56	0.2680
"	3%	"	8.971	1.845	1.0	18	2612	60	0.1176
"	3%	"	8.971	1.845	1.0	18	2612	60	0.0824
"	5%	"	12.835	1.764	0.0	14	4657	101	0.0138
"	5%	"	11.792	1.712	0.5	25	4989	105	0.0092
"	5%	"	11.554	1.712	1.0	25	4989	105	0.0073
"	10%	"	20.100	1.672	0.0	37	10181	203	0.0001
"	10%	"	18.951	1.672	0.5	37	10181	203	"
"	10%	"	18.418	1.657	1.0	23	9623	191	"
"	15%	"	28.677	1.693	0.0	39	15153	296	"
"	15%	"	25.260	1.693	0.5	45	13092	259	"
"	15%	"	24.750	1.693	1.0	45	13092	259	"
"	20%	"	31.246	1.659	0.0	66	19640	368	"
"	20%	"	28.406	1.647	0.5	54	19519	364	"
"	20%	"	26.782	1.647	1.0	54	19519	364	"
"	30%	"	36.015	1.657	0.0	91	25699	465	"
"	30%	"	33.391	1.647	0.5	85	26065	469	"
"	30%	"	32.606	1.617	1.0	107	28339	498	"
"	50%	"	36.015	1.657	0.0	91	25699	465	"
"	50%	"	33.391	1.647	0.5	85	26065	469	"
Maschi 95+	50%	9999	32.606	1.617	1.0	107	28339	498	0.0001
Femmine 95+	1%	9999	9.275	1.693	0.0	6	938	80	0.0001
"	1%	"	19.421	2.442	0.5	3	537	66	"
"	1%	"	19.029	2.442	1.0	3	537	66	"
"	3%	"	21.007	2.192	0.0	5	799	88	"
"	3%	"	19.890	1.731	0.5	10	1893	164	"
"	3%	"	19.489	1.731	1.0	10	1893	164	"
"	5%	"	27.996	1.585	0.0	14	4071	320	"
"	5%	"	23.278	1.504	0.5	27	5242	390	"
"	5%	"	21.074	1.508	1.0	17	4503	337	"
"	10%	"	41.758	1.466	0.0	27	10141	722	"
"	10%	"	35.453	1.434	0.5	38	10467	730	"
"	10%	"	33.426	1.434	1.0	38	10467	730	"
"	15%	"	54.587	1.449	0.0	40	15431	1065	"
"	15%	"	44.147	1.404	0.5	46	16232	1089	"
"	15%	"	41.623	1.404	1.0	46	16232	1089	"
"	20%	"	54.587	1.449	0.0	40	15431	1065	"
"	20%	"	46.442	1.393	0.5	48	17666	1172	"
"	20%	"	45.504	1.393	1.0	48	17666	1172	"
"	30%	"	54.587	1.449	0.0	40	15431	1065	"
"	30%	"	47.500	1.349	0.5	85	26814	1688	"
"	30%	"	47.785	1.344	1.0	81	25744	1621	"
"	50%	"	54.587	1.449	0.0	40	15431	1065	"
"	50%	"	47.500	1.349	0.5	85	26814	1688	"
Femmine 95+	50%	9999	45.874	1.344	1.0	81	25744	1621	0.0001

1 La funzione di verosimiglianza è stata penalizzata per il valore del parametro imposto; 2 Penalità nulla=0; media=0.5; forte=1

I *clusters* ellittici, relativi alla popolazione maschile, si collocano geograficamente tra le province di Ravenna e Bologna similmente a quelli circolari e la penalizzazione non modifica in modo evidente la formazione dei *clusters* primari. La figura 4 riporta i risultati ottenuti con una dimensione massima di popolazione a rischio pari al 5% e una penalizzazione uguale ad 1; il *cluster* primario è costituito da 25 comuni suddivisi in 21 comuni appartenenti alla provincia di Bologna, distribuiti tra il Circondario imolese e la fascia appenninica, e 4 comuni della provincia di Ravenna (Brisighella, Casola Valsenio, Riolo Terme e Castel Bolognese) per una popolazione totale a rischio di 4989 individui e con 105 casi osservati (RR=1.712;p=0.073). Una situazione simile vale per la popolazione femminile che risulta caratterizzata da un *cluster* primario posizionato tra le province di Ravenna e Bologna (figura 5); per un limite di popolazione pari al 5% e una penalizzazione uguale ad 1, il *cluster* include 17 comuni di cui 13 della provincia di Bologna, collocati tra il territorio imolese e parte della zona collinare, e 4 dell'area ravennate (Bagnara di Romagna, Castel Bolognese, Russi e Solarolo), per una popolazione a rischio di 4503 individui e 337 casi osservati 95+ (RR=1.508;p=0.0001). Un *cluster* secondario ad elevata longevità, statisticamente significativo, è invece formato da 6 comuni della provincia di Forlì-Cesena e 2 comuni della provincia di Ravenna.

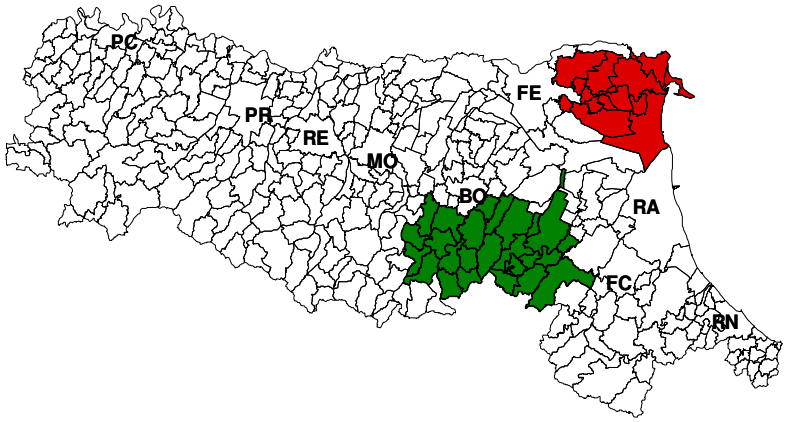


Figura 4 – Mappa CR95+ Maschi SSS (5% dimensione massima di popolazione)
Cluster ellittico. 9999 simulazioni. Penalizzazione a=1
(verde=elevata longevità; rosso=ridotta longevità)

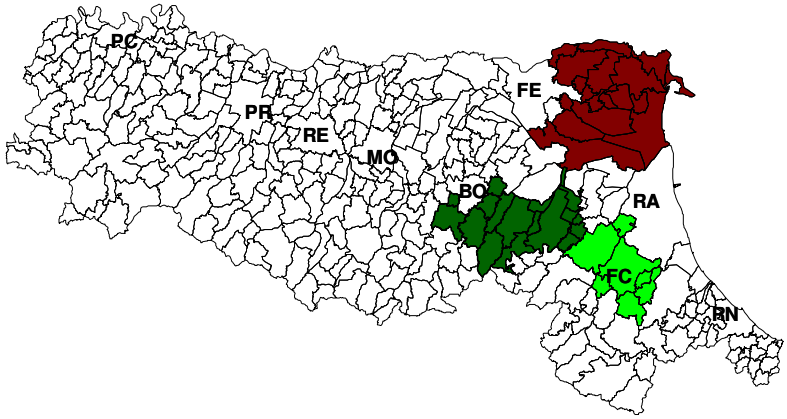


Figura 5 – Mappa CR95+ Femmine SSS (5% dimensione massima di popolazione)
Cluster ellittico. 9999 simulazioni. Penalizzazione a=1
(verde=elevata longevità; rosso=ridotta longevità (secondario);
verde chiaro=elevata longevità (secondario))

Le procedure di *cluster detection* necessitano delle coordinate cartesiane o di proiezione (latitudine e longitudine) dei centroidi di ogni area: la *spatial scan statistic* consente l'uso di entrambi i sistemi di riferimento[25] mentre le altre procedure utilizzano solo le coordinate cartesiane (ad eccezione della *flexible scan statistic)*. In genere, le metodologie di *scan statistic* non dispongono di un

[25] Per i *clusters* ellittici sono necessarie le coordinate cartesiane

output grafico dei risultati ma richiedono ulteriori elaborazioni in ambito GIS per la visualizzazione dei *clusters* identificati[26]; vedremo in seguito, tuttavia, che sia la *flexible scan statistic* che la metodologia GGS forniscono una rappresentazione dei risultati attraverso una struttura a grafo di scarso dettaglio.

La seconda metodologia impiegata nello studio è la *flexible spatial scan statistic* che indicheremo con FSC[27]. A differenza delle altre metodologie in cui si fissa una frazione massima di popolazione a rischio, essa consente di limitare la dimensione finale del *cluster* scegliendo a priori il numero massimo di aree da includere al suo interno; nella nostra ricerca, i valori utilizzati sono stati: 1, 2, 3, 4, 5, 10, 15, 20 e 25 (tabelle 4 e *app14*). La metodologia FSC consente l'identificazione di *clusters* secondari, caratterizzati da un livello di rischio elevato, nonché la definizione di aggregazioni circolari; utilizzando quest'ultima caratteristica sono state effettuate ulteriori ricerche ed i risultati hanno evidenziato una buona sovrapposizione dei *clusters* individuati con i due approcci.

In riferimento alla popolazione maschile, per un numero di comuni massimo pari a 17 (che corrisponde al numero di comuni per una popolazione media a rischio del al 5%), il *cluster* primario si posiziona tra le aree della provincia di Ravenna, con una piccola estensione nelle province di Forli-Cesena e Bologna (figura 6). Esso risulta costituito da 14 comuni così ripartiti: 1 comune della provincia di Bologna (Imola), 1 comune della provincia di Forlì-Cesena (Forlì) e 12 comuni della provincia di Ravenna per un numero di casi osservati pari a 224 (RR=1.531;p=0.001). La procedura identifica anche un *cluster* secondario, non riportato in figura (p=0.115), comprendente 9 comuni dell'area bolognese appartenenti in prevalenza alla fascia montana; la dimensione limitata del *cluster,* infatti, potrebbe aver suddiviso in 2 aggregazioni distinte un'unica area geografica più estesa. La zona riportata in bianco all'interno del *cluster* primario è costituita da 2 piccoli comuni: Bagnara di Romagna (Ra) e Mordano (Bo). La variazione del numero massimo di comuni non modifica sostanzialmente la collocazione geografica dei *clusters* primari: le aree limitrofe ai comuni di Imola,

[26] Nel nostro studio, le mappe sono state elaborate utilizzando un software GIS (ArcView GIS vers.3.3)
[27] La FSC è stata implementata in un software gratuito FlexScan v2.0 disponibile on-line

Ravenna e Forlì si caratterizzano per un significativo livello del fenomeno e il *cluster* singolo identificato dalla procedura corrisponde al comune di Imola.

Tabella 4 - Risultati FSC (solo *cluster* primario)

Popolazione Esaminata	Dimensione max	Repliche	LLR	Rischio relativo	N° comuni cluster	Casi cluster	p-value
Maschi 95+	1	9999	4.610	1.742	1	35	*0.285*
"	2	"	4.610	1.742	1	35	0.3737
"	3	"	5.724	1.787	2	40	0.1833
"	4	"	7.098	1.858	3	44	0.0677
"	5	"	7.785	1.644	5	71	0.0080
"	10	"	13.835	1.516	10	165	0.0007
"	15	9999	17.943	1.604	13	169	0.0003
"	17	999	20.392	1.531	14	224	0.001
"	20	"	22.435	1.546	17	235	0.001
Maschi 95+	25	999	23.536	1.537	19	250	0.001
Femmine 95+	1	9999	16.308	2.800	1	42	0.0001
"	2	"	16.308	2.800	1	42	"
"	3	"	16.869	2.408	2	40	"
"	4	"	19.370	2.519	2	60	"
"	5	"	20.525	1.913	4	118	"
"	10	"	22.422	1.891	5	133	"
"	15	9999	29.181	1.349	13	649	0.0001
"	17	9999	32.455	1.328	15	781	"
"	20	999	33.945	1.340	16	772	0.001
Femmine 95+	25	999	39.783	1.354	19	838	0.001

Analizzando la popolazione femminile, le zone ad elevata longevità si posizionano tra le province di Ravenna, Forlì-Cesena e Bologna. Fissando una dimensione massima pari a 17, il *cluster* primario risulta così costituito: 2 comuni della provincia di Bologna (Imola e Mordano), 1 comune della provincia di Forlì-Cesena (Forlì) e 12 comuni della provincia di Ravenna per un totale di 781 casi (RR=1.328;p=0.001) (figura 7). Il primo dei *clusters* secondari (p=0.001) è rappresentato da un'area formata da 8 comuni appartenenti alla provincia di Bologna collocati nella fascia montana. Si noti l'irregolarità del *cluster* secondario rispetto a quello primario: l'esclusione anche di una sola zona dal *cluster* determina una struttura geografica meno regolare e meno compatta. Diminuendo la dimensione del *cluster* si identifica un numero maggiore di aggregazioni secondarie distribuite tra le province sino ad ora indicate e, nel

caso di *cluster* singolo, si individua il comune di San Lazzaro di Savena[28] appartenente alla provincia di Bologna.

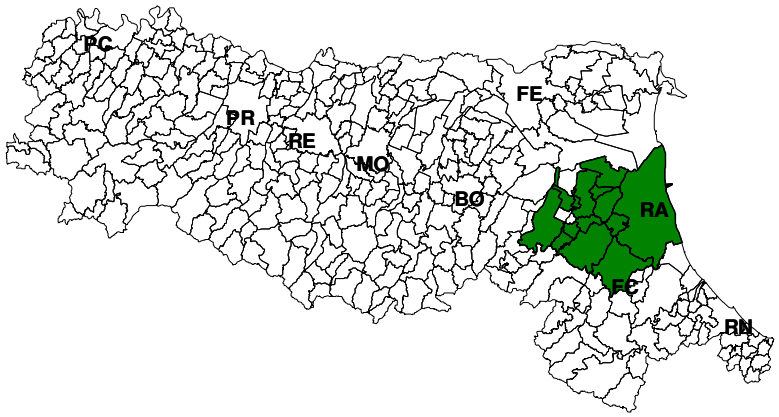


Figura 6 – Mappa CR95+ Maschi FSC (dimensione massima=17 comuni). 999 simulazioni

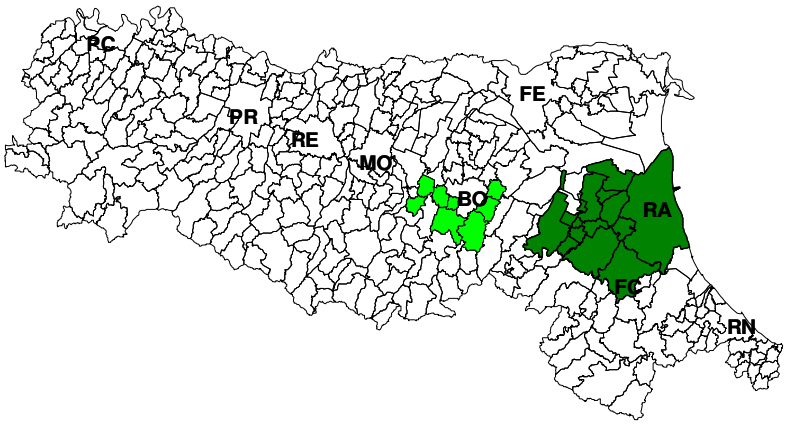


Figura 7 – Mappa CR95+ Femmine FSC (dimensione massima=17 comuni). 999 simulazioni
(verde scuro=elevata longevità (primario); verde chiaro= elevata longevità)

La metodologia FSC consente di visualizzare i risultati ottenuti mediante un grafo del territorio esaminato nel quale è riportato il *cluster* finale: i vertici delle

[28] Il comune di San Lazzaro di Savena (Bo) ha il valore regionale più elevato di CR per la popolazione femminile

aree incluse sono evidenziati in rosso analogamente a tutti i collegamenti esistenti tra le zone (figura 8).

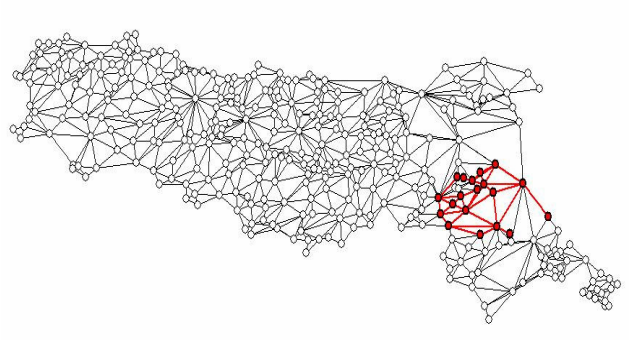


Figura 8 – Esempio di *output* della procedura FSC.
(Maschi 95+; dimensione massima=25 comuni)

La terza metodologia impiegata nella ricerca è l'algoritmo genetico (GA). La *routine* consente di variare i parametri di ricerca quali la dimensione massima del *cluster*, espressa come percentuale di popolazione a rischio, la costante di penalizzazione, il numero G di generazioni ed il numero w di *crossing-over* (appendice *app15*). L'algoritmo GA consente solo l'analisi di dati di conteggio mediante un modello di Poisson e identifica solo il *cluster* primario caratterizzato da un livello elevato del fenomeno. In merito alla scelta del numero di generazioni, di *crossing-over* e del tasso di mutazione, si è ritenuto opportuno procedere seguendo le indicazioni suggerite negli studi di simulazione riportati in letteratura; i parametri sono stati fissati a G=10 generazioni, w=400 *crossing-over* e 1% per il tasso di mutazione. A scopo di ricerca, tuttavia, sono state effettuate ulteriori elaborazioni in cui sono stati modificati tali parametri: la variazione del numero di *crossing-over* e del numero di generazioni ha fornito gli stessi risultati ottenuti con i parametri standard (tabella 5).

La procedura GA è caratterizzata da una buona velocità di esecuzione anche se, aumentando il numero di replicazioni a 9999 e, a parità di parametri di ricerca, l'identificazione del *cluster* finale richiede un tempo fino a 10 volte superiore a quello necessario per 999 simulazioni. Nelle figure 9 e 10 sono riportati due esempi di penalizzazione: una penalità nulla individua un *cluster*

molto esteso ed irregolare a differenza di quello ottenuto con un valore pari ad 1 che appare più compatto e regolare.

Tabella 5 - Risultati GA (solo *cluster* primario)
(G=n°generazioni; w=*crossing-over;* Pop=dimensione massima di popolazione)
(il tasso di mutazione è fissato all'1% salvo indicazioni contrarie)

Pop. esaminata	Pop	Repliche	w	G	LLR[1]	Tasso cluster	Penalità[2]	N° comuni cluster	Pop. cluster	Casi cluster	p-value
Maschi 95+	1%	999	400	10	8.386	0.030	0.0	6	961	29	0.184
"	1%	"	"	"	6.463	0.030	0.5	6	961	29	0.128
"	1%	"	"	"	5.215	0.029	1.0	7	987	28	0.262
"	3%	"	"	"	15.525	0.025	0.0	19	3022	77	0.060
"	3%	"	"	"	10.566	0.026	0.5	15	2519	65	0.025
"	3%	"	"	"	8.688	0.025	1.0	18	2755	68	0.015
"	5%	"	"	"	22.128	0.023	0.0	24	5085	123	0.006
"	5%	"	"	"	14.062	0.023	0.5	23	5122	119	0.004
"	5%	"	"	"	9.243	0.022	1.0	31	5101	112	0.034
"	10%	"	"	"	31.164	0.022	0.0	26	10216	223	0.002
"	10%	"	"	"	18.409	0.021	0.5	38	10239	213	0.003
"	10%	"	"	"	16.574	0.020	1.0	17	9212	186	0.001
"	15%	"	"	"	38.323	0.021	0.0	34	15274	315	"
"	15%	"	"	"	29.319	0.020	0.5	38	15331	314	"
"	15%	"	"	"	21.969	0.020	1.0	19	12473	245	"
"	20%	"	"	"	48.346	0.020	0.0	53	18301	378	"
"	20%	"	"	"	29.798	0.020	0.5	46	16792	338	"
"	20%	"	"	"	24.046	0.020	1.0	44	17225	337	"
"	25%	"	"	"	51.475	0.020	0.0	54	21474	431	"
"	25%	"	"	"	30.960	0.020	0.5	50	18896	373	"
"	25%	"	"	"	24.072	0.020	1.0	46	17719	345	"
"	30%	"	"	"	54.632	0.020	0.0	60	24650	483	"
"	30%	"	"	"	33.605	0.020	0.5	60	23056	447	"
"	30%	999	400	10	24.212	0.020	1.0	48	17867	348	"
"	10%	999	1000	10	16.574	0.020	1.0	17	9212	186	"
"	15%	999	500	10	21.969	0.020	1.0	19	12473	245	"
"	15%	999	750	10	21.969	0.020	1.0	19	12473	245	"
"	15%	999	1000	10	21.969	0.020	1.0	19	12473	245	"
	10%**	999	400	10	16.574	0.020	1.0	17	9212	186	"
"	15%**	999	400	10	29.319	0.020	0.5	38	15331	314	"
"	15%**	999	400	10	21.969	0.020	1.0	19	12473	245	"
"	20%**	999	400	10	23.787	0.020	1.0	46	17410	341	0.001
Maschi 95+	15%	9999	400	10	21.969	0.020	1.0	19	12473	245	0.0001
Femmine 95+	1%	999	400	10	21.212	0.107	0.0	6	888	95	0.001
"	1%	"	"	"	16.309	0.142	0.5	1	296	42	"
"	1%	"	"	"	16.309	0.142	1.0	1	296	42	"
"	3%	"	"	"	30.049	0.084	0.0	19	3133	264	"
"	3%	"	"	"	21.155	0.083	0.5	20	3239	268	"
"	3%	"	"	"	16.722	0.081	1.0	18	3012	243	"
"	5%	"	"	"	37.392	0.079	0.0	25	5435	426	"
"	5%	"	"	"	24.794	0.078	0.5	25	5328	415	"
"	5%	"	"	"	17.124	0.079	1.0	22	3704	222	"
"	10%	"	"	"	58.123	0.074	0.0	31	10860	805	"
"	10%	"	"	"	37.027	0.071	0.5	27	10465	747	"
"	10%	"	"	"	31.469	0.071	1.0	28	10863	768	"
"	15%	"	"	"	74.094	0.071	0.0	46	16303	1165	"
"	15%	"	"	"	50.507	0.070	0.5	36	15764	1106	"
"	15%	"	"	"	44.281	0.068	1.0	40	16297	1114	"
"	20%	"	"	"	78.035	0.071	0.0	52	17240	1229	"
"	20%	"	"	"	50.163	0.070	0.5	36	15764	1106	"
"	20%	"	"	"	45.392	0.070	1.0	43	15764	1106	"
"	25%	"	"	"	78.035	0.071	0.0	52	17240	1229	"
"	25%	"	"	"	50.163	0.070	0.5	36	15764	1106	"
"	25%	"	"	"	45.392	0.069	1.0	43	16681	1147	"
"	30%	"	"	"	78.459	0.071	0.0	53	17429	1241	"
"	30%	"	"	"	50.507	0.070	0.5	37	15818	1110	"
"	30%	999	400	10	45.182	0.069	1.0	43	16681	1147	"
"	10%	999	400	200	31.469	0.071	1.0	28	10863	768	"
"	15%	999	400	50	44.281	0.068	1.0	40	16297	1114	"
"	15%	999	400	100	44.281	0.068	1.0	40	16297	1114	"
"	15%	999	400	200	44.281	0.068	1.0	40	16297	1114	0.001
Femmine 95+	15%	9999	400	10	44.281	0.068	1.0	40	16297	1114	0.0001

1 Il valore di LLR riportato è stato penalizzato per il parametro imposto; 2 Penalità nulla=0;media=0.5;forte=1
** Il tasso di mutazione è stato fissato all'1x1000 (0.001)

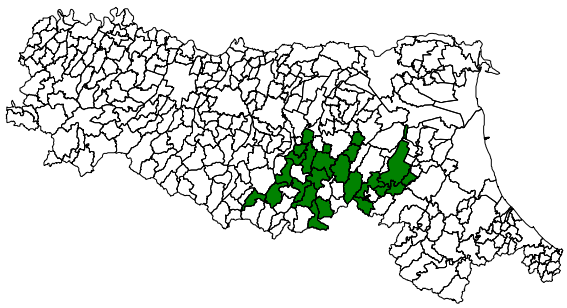


Figura 9 – Mappa CR95+ Maschi GA (5% dimensione massima popolazione)
999 simulazioni. Penalità a=0. *Cluster* primario

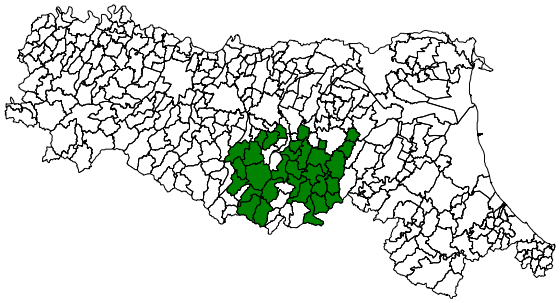


Figura 10 – Mappa CR95+ Maschi GA (5% dimensione massima popolazione)
999 simulazioni. Penalità a=1. *Cluster* primario

Il *cluster* maschile, relativo ad una dimensione massima di popolazione del 5% e una penalità di 0.5, si colloca geograficamente nella provincia di Bologna e comprende una piccola parte nel territorio ravennate (figura 11). Esso è formato da 23 comuni di cui 2 appartenenti alla provincia di Ravenna (Castel Bolognese e Riolo Terme) e 21 alla provincia di Bologna, posizionati tra la fascia montana e la zona imolese, per una popolazione totale a rischio di 5122 individui e 119 casi osservati (p=0.004). La parte mancante nel *cluster* è identificata dai comuni bolognesi di Castel San Pietro Terme, Castel Guelfo di Bologna e Dozza. La diminuzione della dimensione massima di popolazione consente ugualmente di identificare aree appartenenti alle province indicate, con un'estensione geografica variabile a seconda della penalizzazione imposta. La metodologia si dimostra idonea nell'identificazione di aggregazioni di piccole dimensioni: fissando una proporzione massima di popolazione all'1% e con una penalizzazione pari ad 1 o 0.5, si identifica un *cluster* femminile costituito da un

solo comune, San Lazzaro di Savena (Bo), già identificato nella procedura FSC. Il *cluster* riportato in figura 12, invece, si riferisce ad una frazione di popolazione del 5% e una penalizzazione di 0.5; esso risulta costituito da 25 comuni di cui 4 appartenenti alla provincia di Ravenna (Bagnara di Romagna, Castel Bolognese, Cotignola e Solarolo) e 21 alla provincia di Bologna, distribuiti tra la zona del Circondario imolese e la fascia appenninica, per una popolazione totale a rischio di 5328 donne e 415 casi osservati 95+ (p=0.001).

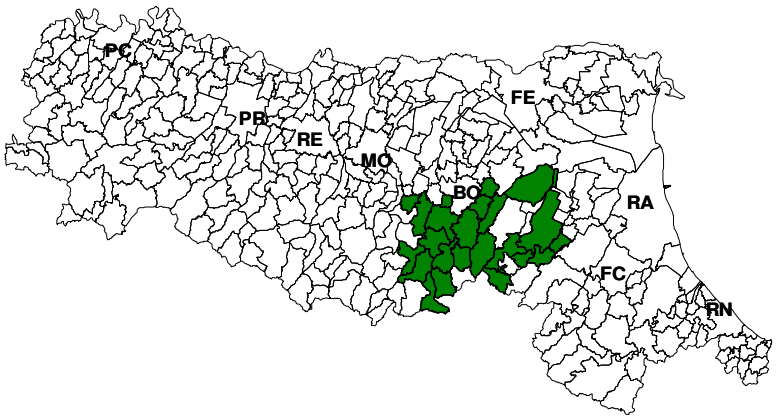


Figura 11 – Mappa CR95+ Maschi GA (5% popolazione). 999 simulazioni. Penalità a=0.5

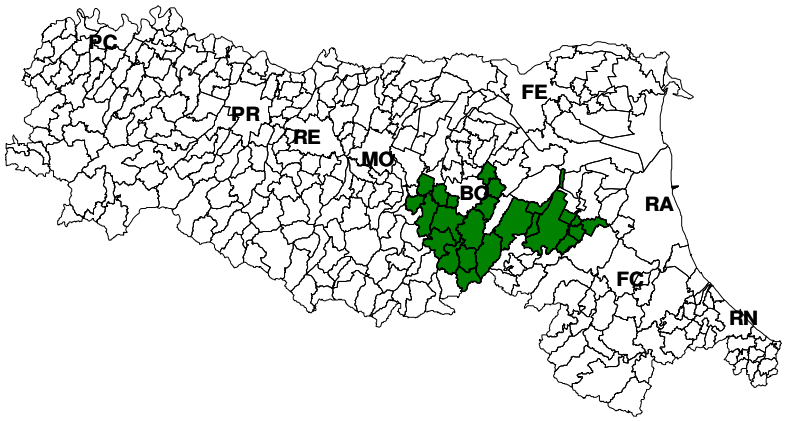


Figura 12 – Mappa CR95+ Femmine GA (5% popolazione). 999 simulazioni. Penalità a=0.5

Prima di analizzare i risultati ottenuti con la metodologia GGS, si riportano alcune osservazioni derivate da un'analisi spaziale condotta su una porzione limitata del territorio regionale. Le ricerche sinora effettuate hanno evidenziato, per entrambi i sessi, la presenza di un *cluster* ad elevata longevità posizionato tra le province di Ravenna, Forlì-Cesena e Bologna. Di contro le aree comunali delle province di Piacenza, Parma e Reggio Emilia non risultano mai inclusi nei *clusters* primari o in quelli secondari (statisticamente significativi). A conferma di ciò, si è quindi deciso di valutare se la presenza di tali *clusters* condiziona l'identificazione di aggregazioni posizionate in queste aree; sono state effettuate nuove ricerche escludendo dal territorio, in un primo momento, le province di Ravenna, Forlì-Cesena e Rimini e, successivamente, anche i comuni dell'area ferrarese e bolognese, limitando l'analisi solo alle province di Piacenza, Parma, Reggio Emilia e Modena. Le elaborazioni sono state effettuate utilizzando sia la metodologia SSS (regolare) che il GA (irregolare). Escludendo le province di Ravenna, Forlì-Cesena e Rimini, i comuni da analizzare risultano 273 con una popolazione maschile a rischio di 81.265 individui e 940 casi osservati ed una popolazione femminile di 86.597 e 4160 casi osservati. Dalla tabella 6, si può notare che, per entrambi i sessi, variando la dimensione massima di popolazione, i risultati appaiono sostanzialmente simili; inoltre, per i *clusters* ellittici l'uso di una penalizzazione non influenza, in modo evidente, le caratteristiche dell'aggregazione finale: l'unica variazione si ottiene nel passaggio da una proporzione massima del 5% ad una proporzione del 10%. I *clusters* finali risultano sempre identificati da comuni della provincia di Bologna per entrambi i sessi ed, in particolare, per gli uomini si identifica un *cluster* secondario non significativo, di dimensioni ridotte, formato da alcune zone della provincia di Piacenza ($p=0.356$) mentre per le donne, invece, si individua un *cluster* secondario non significativo localizzato tra le province di Parma e Reggio Emilia ($p=0.670$).

Tabella 6 - Risultati delle analisi SSS (escluse le province di Ravenna, Forlì-Cesena, Rimini) (analisi ellittica e circolare, solo *cluster* primario)

Finestra	Pop. esaminata	Dim.max pop	Repliche	LLR[1]	Rischio relativo	Penalità[2]	N° comuni cluster	Pop cluster	Casi cluster	p-value
Circolare	Maschi 95+	5%	9999	12.089	1.848	---	19*	4047	83	0.0004
"	"	10%	"	16.935	1.909	---	25*	5175	108	0.0001
"	"	15%	"	16.935	1.909	---	25*	5175	108	"
"	"	20%	"	16.935	1.909	---	25*	5175	108	"
"	Maschi 95+	30%	"	16.935	1.909	---	25*	5175	108	"
"	Femmine 95+	5%	"	22.731	1.545	---	20*	4034	292	"
"	"	10%	"	24.776	1.508	---	25*	5120	360	"
"	"	15%	"	24.776	1.508	---	25*	5120	360	"
"	"	20%	"	24.776	1.508	---	25*	5120	360	"
Circolare	Femmine 95+	30%	"	24.776	1.508	---	25*	5120	360	"
Ellisse	Maschi 95+	15%	"	17.312	1.896	0.0	25*	5463	113	"
"	"	15%	"	16.935	1.909	1.0	25*	5175	108	"
"	"	20%	"	17.312	1.896	0.0	25*	5463	113	"
"	Maschi 95+	20%	"	16.935	1.909	1.0	25*	5175	108	"
"	Femmine 95+	15%	"	29.778	1.517	0.0	28*	6056	426	"
"	"	15%	"	26.965[3]	1.549	1.0	23*	4790	346	"
"	"	20%	"	29.778	1.517	0.0	28*	6056	426	"
Ellisse	Femmine 95+	20%	9999	26.965[4]	1.549	1.0	23*	4790	346	0.0001

1 La funzione di verosimiglianza è stata penalizzata per la penalità imposta; 2 Penalità nulla=0; media=0.5; forte=1
3,4 Il *cluster* riportato è il primo dei *clusters* secondari ad elevata longevità
* I comuni inclusi nel *cluster* appartengono alla provincia di Bologna

I risultati ottenuti con la seconda metodologia (GA) confermano la presenza di *clusters* a maggiore longevità tra le province di Modena e Bologna per entrambi i sessi; solo per una proporzione di popolazione pari al 5% si individua un *cluster* maschile costituito esclusivamente da comuni dell'area bolognese a differenza di quanto avviene in tutti gli altri casi.

Tabella 7 - Risultati delle analisi GA (escluse le province di Ravenna, Forlì-Cesena, Rimini) (tasso mutaz 1%; w=400 crossing-over; G=10 generazioni) (solo *cluster* primario)

Popolazione esaminata	Dim.max pop	Repliche	LLR[1]	Penalità	Tasso interno cluster	N° comuni cluster	Pop. cluster	Casi cluster	p-value
Maschi 95+	5%	999	11.167	1.0	0.025	18*	2755	68	0.001
"	10%	"	15.583	1.0	0.021	37**	8098	168	"
"	15%	"	16.124	1.0	0.020	47**	9897	194	"
"	20%	"	16.198	1.0	0.020	47**	9897	194	"
Maschi 95+	30%	"	16.124	1.0	0.020	47**	9897	194	"
Femmine 95+	5%	"	21.760	1.0	0.079	22**	3704	292	"
"	10%	"	27.326	1.0	0.070	38**	8153	574	"
"	15%	"	28.694	1.0	0.069	45**	9271	636	"
"	20%	"	28.694	1.0	0.069	43**	9108	626	"
Femmine 95+	30%	999	28.569	1.0	0.069	43**	9108	626	0.001

1 La funzione di verosimiglianza è stata penalizzata per la penalità imposta;
* I comuni inclusi nel *cluster* appartengono alla provincia di Bologna
** I comuni inclusi nel *cluster* appartengono alle province di Bologna e Modena

La valutazione dei risultati appena descritti ha indotto a proseguire le indagini restringendo ulteriormente il territorio da esaminare, escludendo anche le province di Ferrara e Bologna. L'area studio si è ridotta a 187 comuni, con una popolazione maschile a rischio di 46.051 e 535 casi osservati e una popolazione femminile di 48.041 e 2397 casi osservati. La caratteristica principale dei risultati è la totale assenza di *clusters* statisticamente significativi sul territorio nonché l'esiguo numero di aree incluse nelle aggregazioni individuate, in particolare per le donne (tabella 8). I parametri caratteristici dei *clusters* risultano uguali, sia nel caso circolare che ellittico, per tutte le dimensioni di popolazione selezionate. La posizione geografica dei *clusters* è ancora differente per le due popolazioni: per gli uomini, i *clusters* si collocano nella provincia di Piacenza e comprendono i comuni di Bobbio, Zerba, Ferriere, Ottone, Corte Brugnatella, Cerignale, Coli, che rappresentano aree comunali di modeste dimensioni demografiche e territoriali posizionate nella parte sud-ovest della fascia montana piacentina, al confine con la Lombardia e la Liguria mentre per le donne, invece, i *clusters* sono formati da comuni della provincia di Modena tra cui Spilamberto e Castelnuovo Rangone.

Tabella 8 - Risultati analisi SSS (analisi ellittica e circolare, solo *cluster* primario)
(escluse le province di Bologna, Ferrara, Ravenna, Forlì-Cesena, Rimini)

Finestra	Popolazione esaminata	Dim.max pop	Repliche	LLR[1]	Rischio relativo	Penalità[2]	N° comuni cluster	Pop. cluster	Casi cluster	p-value
Circolare	Maschi 95+	1%	9999	5.228	3.040	--	5#	345	12	p=0.1708
"	"	5%	"	5.500	2.555	--	7#	584	17	p=0.4747
"	"	10%	"	5.500	2.555	--	7#	584	17	p=0.5120
"	Maschi 95+	20%	"	5.500	2.555	--	7#	584	17	p=0.5120
"	Femmine 95+	1%	"	4.364	1.760	--	2***	378	33	p=0.6802
"	"	5%	"	4.364	1.760	--	2***	378	33	p=0.7831
"	"	10%	"	4.364	1.760	--	2***	378	33	p=0.8741
Circolare	Femmine 95+	15%	"	4.364	1.760	--	2***	378	33	p=0.8741
Ellisse	Maschi 95+	5%	"	5.719	2.699	0.0	6#	520	16	p=0.9208
"	"	5%	"	5.500	2.555	1.0	7#	584	17	p=0.6911
"	"	10%	"	5.719	2.699	0.0	6#	520	16	p=0.9530
"	"	10%	"	5.500	2.555	1.0	7#	584	17	p=0.7210
"	"	20%	"	5.719	2.699	0.0	6#	520	16	p=0.9530
"	Maschi 95+	20%	"	5.500	2.555	1.0	7#	584	17	p=0.7210
"	Femmine 95+	5%	"	4.497[3]	1.409	0.0	7***	1265	88	p=0.9908
"	"	5%	"	4.364[3]	1.760	1.0	2***	378	33	p=0.9440
"	"	10%	"	4.497[3]	1.409	0.0	7***	1265	88	p=0.9930
"	"	10%	"	4.364[3]	1.760	1.0	2***	378	33	p=0.9490
"	"	20%	"	4.497[3]	1.409	0.0	7***	1265	88	p=0.9930
Ellisse	Femmine 95+	20%	9999	4.364[3]	1.760	1.0	2***	378	33	p=0.9490

1 La funzione di verosimiglianza è stata penalizzata per la penalità imposta; 2 Penalità nulla=0; media=0.5; forte=1
3 Il *cluster* riportato è il primo dei *clusters* secondari caratterizzato da elevata longevità
I comuni inclusi nel *cluster* appartengono alla provincia di Piacenza
*** I comuni inclusi nel *cluster* appartengono alla provincia di Modena

Le analisi effettuate con il metodo genetico confermano l'assenza di *clusters* statisticamente significativi del fenomeno (tabella 9). I risultati relativi agli individui maschi sono uguali per tutte le dimensioni di popolazione fissate mentre per la popolazione femminile il *cluster* risulta costituito da 1 solo comune, Spilamberto (Mo), fino ad un limite massimo del 10% di popolazione. La collocazione geografica delle aggregazioni individuate è pressoché simile a quella descritta per la *spatial scan statistic*: i maschi longevi appartengono ai comuni di Zerba, Ferriere, Ottone, Corte Brugnatella e Cerignale. La figura 13 riporta il *cluster* primario relativo alla popolazione maschile identificato dalla SSS (limite popolazione 5%).

Tabella 9 - Risultati analisi GA (escluse province di Bologna, Ferrara, Ravenna, Forlì-Cesena, Rimini) (tasso mutaz 1%; w=400 crossing-over; G=10 generazioni) (solo *cluster* primario)

Popolazione esaminata	Dim.max pop	Repliche	LLR[1]	Penalità[2]	Tasso interno cluster	N° comuni cluster	Pop. cluster	Casi cluster	p-value
Maschi 95+	1%	999	5.021	0.5	0.035	5#	345	12	*p=0.127*
"	1%	"	5.021	1.0	0.035	5#	345	12	*p=0.127*
"	5%	"	5.021	1.0	0.035	5#	345	12	*p=0.201*
"	10%	"	5.021	1.0	0.035	5#	345	12	*p=0.220*
"	15%	"	5.021	1.0	0.035	5#	345	12	*p=0.225*
"	20%	"	5.021	1.0	0.035	5#	345	12	*p=0.400*
Maschi 95+	30%	999	5.021	1.0	0.035	5#	345	12	*p=0.400*
Femmine 95+	1%	999	4.149	0.0	0.099	1***	223	22	*p=0.440*
	1%	"	4.149	1.0	0.099	1***	223	22	*p=0.440*
"	5%	"	4.149	1.0	0.099	1***	223	22	*p=0.520*
"	10%	"	4.149	1.0	0.099	1***	223	22	*p=0.601*
"	15%	"	4.479	1.0	0.060	15§	7104	428	*p=0.610*
"	20%	"	5.688	1.0	0.059	26§	8919	529	*p=0.701*
Femmine 95+	30%	999	5.811	1.0	0.060	34^	10474	621	*p=0.780*

1 La funzione di verosimiglianza è stata penalizzata per la penalità imposta; 2 Penalità nulla=0;media=0.5; forte=1
I comuni inclusi nel *cluster* appartengono alla provincia di Piacenza
§ I comuni inclusi nel *cluster* appartengono alle province di Parma e Reggio Emilia
^ I comuni inclusi nel *cluster* appartengono alle province di Piacenza, Parma e Reggio Emilia
*** Il comune incluso nel *cluster* appartiene alla provincia di Modena

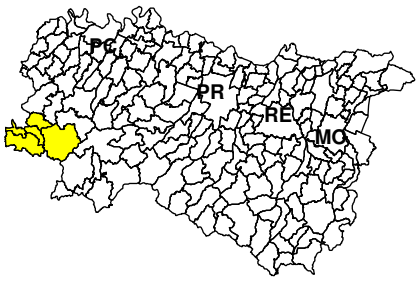


Figura 13 – Mappa CR95+ Maschi SSS (5% dimensione massima popolazione)
(escluse province di Bologna, Ferrara, Ravenna, Forlì-Cesena e Rimini)

L'ultima metodologia utilizzata nello studio è l'algoritmo *Greedy Growth Search* (GGS). Il limite principale della tecnica è la velocità di esecuzione che si rivela particolarmente ridotta anche per analisi con modeste frazioni di popolazione. I due parametri di controllo consentono di intervenire durante la fase di identificazione del *cluster* penalizzando l'irregolarità delle potenziali aggregazioni (parametro di non-connettività, α) e velocizzando il processo di ricerca nei casi in cui si riscontra una variazione minima del LLR (parametro di profondità, u). L'effetto del parametro u si nota già a partire da valori superiori ad 1 in quanto l'estensione geografica del *cluster* aumenta al crescere del valore di u: utilizzando una profondità uguale o inferiore ad 1, la forma dell'aggregazione dipende quasi esclusivamente dal parametro di non-connettività. Dai risultati ottenuti (appendice *app16*) è altresì evidente un carico computazionale maggiore per valori di $u \geq 2$: ad esempio, fissando una profondità u=4 e al crescere della dimensione della popolazione massima, i tempi di elaborazione si aggirano tra 9 a 22 ore ipotizzando che, qualora il numero di aree esaminate risulti superiore a 400, la metodologia possa rivelarsi estremamente lenta. In merito a quanto osservato nel nostro studio è opportuno aggiungere che l'uso di una profondità superiore ad 2 si rivela inutile per la modesta variazione dei risultati finali. Il parametro di non-connettività consente di limitare, invece, l'estensione del *cluster* ed i risultati ottenuti evidenziano che un valore di non-connettività superiore a 2 è già "sufficiente" per la definizione di aggregazioni poco irregolari, anche se la selezione dei due parametri rimane particolarmente difficile per la mancanza di criteri di riferimento.

Tabella 10 - Risultati analisi GGS (solo *cluster* primario)
(u =parametro di profondità; α =parametro di non-connettività)

Pop. esaminata	Dim.max pop	u	α	Repliche	LLR[1]	Rischio relativo	N° comuni cluster	Pop. cluster	Casi cluster	p-value
Maschi 95+	1%	1	1	999	10.316	2.268	7	1352	39	0.007
"	1%	2	2	"	8.909	2.327	7	1081	32	0.001
"	3%	1	1	"	15.193	1.976	20	3103	78	0.010
"	3%	1	2	"	15.550	1.953	18	3098	77	0.010
"	3%	2	1	"	15.193	1.976	20	3103	78	0.010
"	5%	2	2	"	16.191	1.733	33	5333	118	0.090
"	5%	2	1	"	16.295	1.819	27	4494	104	0.095
"	5%	1	1	"	15.670	1.728	33	5276	116	0.095
"	5%	1	2	"	16.191	1.733	33	5333	118	0.080
"	5%	1	3	"	14.665	1.848	21	3828	90	0.095
"	5%	1	4	"	14.400	1.823	22	3968	92	0.035
"	10%	0	0	"	26.439	1.557	22	13230	262	0.001
"	10%	0	1	"	21.555	1.544	14	11610	228	"
"	10%	1	0	"	26.244	1.557	22	13230	262	"
"	10%	1	1	"	24.758	1.540	22	13220	259	"
"	10%	1	4	"	21.555	1.544	14	11610	228	"
"	10%	1	10	"	19.057	1.545	13	10378	204	"
"	10%	1	15	"	19.057	1.545	13	10378	204	0.001
"	10%	2	1	"	27.468	1.659	36	10472	221	0.007
"	10%	2	2	"	25.845	1.673	34	9585	204	0.005
"	10%	4	1	"	27.468	1.659	36	10472	221	0.010
"	15%	0	0	"	26.940	1.555	27	13600	269	0.010
"	15%	0	1	"	26.676	1.556	24	13437	266	0.002
"	15%	0	2	"	24.201	1.559	20	12354	245	0.001
"	15%	0	4	"	23.066	1.473	25	15366	288	"
"	15%	1	1	"	29.868	1.607	52	12867	263	"
"	15%	1	2	"	37.980	1.617	43	15311	315	"
"	15%	1	4	"	23.066	1.473	25	15366	288	"
"	15%	2	1	"	38.991	1.619	42	15589	321	"
"	15%	4	1	"	39.319	1.616	43	15812	325	"
"	15%	30	1	"	---	---	---	---	---	---
"	20%	0	1	"	26.677	1.556	24	13437	266	"
Maschi 95+	20%	1	1	999	46.861	1.576	66	20107	403	0.001
Femmine 95+	1%	1	1	999	21.212	2.111	6	888	95	0.001
"	1%	2	2	"	20.403	2.309	3	641	75	"
"	3%	1	1	"	29.634	1.620	20	3484	286	"
"	3%	1	2	"	28.366	1.591	21	3633	293	"
"	3%	2	1	"	26.661	1.634	21	3344	277	"
"	5%	2	1	"	29.879	1.528	28	4662	361	"
"	5%	1	1	"	29.634	1.620	20	3484	286	"
"	5%	1	5	"	28.474	1.599	21	3567	289	"
"	10%	0	0	"	43.719	1.350	27	13457	921	"
"	10%	1	0	"	43.719	1.350	27	13457	921	"
"	10%	1	1	"	42.808	1.343	25	13728	934	"
"	10%	1	5	"	45.548	1.407	36	10837	772	"
"	10%	2	5	"	45.923	1.399	35	11295	801	"
"	15%	0	0	"	43.632	1.351	26	13392	917	"
"	15%	0	1	"	67.150	1.389	36	16354	1151	"
"	15%	0	2	"	65.500	1.379	37	16685	1166	"
"	15%	1	1	"	67.150	1.389	36	16354	1151	"
"	15%	1	2	"	65.500	1.379	37	16685	1166	"
"	15%	1	4	"	66.587	1.383	45	16620	1165	"
"	15%	1	10	"	51.947	1.341	52	16289	1107	"
"	15%	2	1	"	71.499	1.406	42	16050	1144	"
"	15%	4	1	"	71.499	1.406	42	16050	1144	"
"	20%	0	1	"	68.634	1.382	41	17136	1200	"
Femmine 95+	20%	1	1	999	75.874	1.374	57	19166	1335	0.001

1 La funzione di verosimiglianza è penalizzata per i parametri imposti

Le figure 14 e 15 riportano, rispettivamente, le mappe relative ai *clusters* maschili e femminili limitati da una dimensione massima di popolazione del 5% e con parametri di controllo differenti. Il *cluster* maschile (profondità=2;non-connettività=1) si colloca geograficamente tra le province di Modena e Bologna ed è costituito da 27 comuni di cui 11 comuni dell'area modenese e 16 della provincia bolognese, appartenenti alla fascia di territorio pedemontano, con una popolazione totale a rischio di 4494 e 104 individui con età 95+ (RR=1.819;p=0.095).

In riferimento alla popolazione femminile, il *cluster* riportato in figura 15 (profondità=1;non-connettività=1) è identificato da 20 comuni e si posiziona anch'esso tra la provincia di Modena e Bologna: 6 comuni appartengono al territorio pedemontano modenese (Spilamberto, Vignola, Castelnuovo Rangone, Guiglia, Marano sul Panaro, Savignano sul Panaro) e 14 comuni alla fascia montana bolognese per una popolazione totale a rischio di 3484 e 286 donne con età 95+ (RR=1.620;p=0.001).

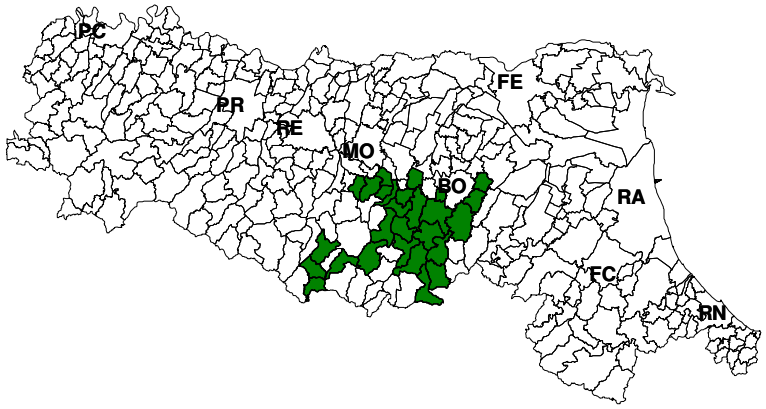


Figura 14 – Mappa CR95+ Maschi GGS
(5% dimensione massima popolazione; profondità=2; non-compattezza=1)
999 simulazioni. *Cluster* primario

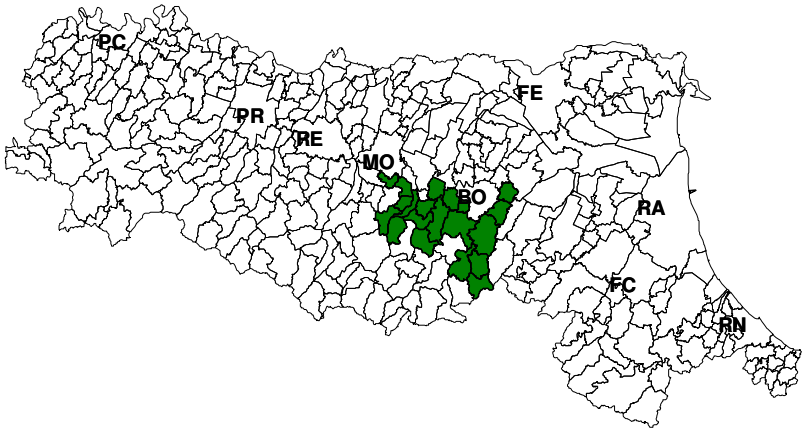


Figura 15 – Mappa CR95+ Femmine GGS
(5% dimensione massima popolazione; profondità=1; non-compattezza=1)
999 simulazioni. *Cluster* primario

Analogamente alla FSC, la procedura GGS fornisce un *output* grafico dei risultati: un esempio è riportato di seguito (figura 16) dove i punti rossi rappresentano i vertici dei comuni inclusi nel *cluster*.

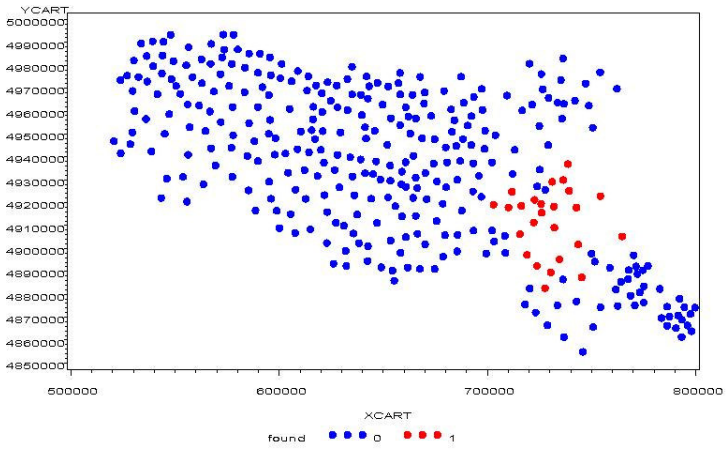


Figura 16 – Esempio di *output* grafico fornito dalla GGS. CR95+ Femmine
(10% dimensione massima popolazione; profondità=1; non-compattezza=1)
999 simulazioni. *Cluster* primario

La valutazione della persistenza di un *cluster* è stata effettuata mediante un confronto tra le matrici di adiacenza delle aggregazioni finali a parità di dimensione di popolazione a rischio. I valori di popolazione massima selezionati per tale operazione sono il 5% e il 3% e per ciascuno di essi sono stati confrontati i *clusters* primari identificati con le quattro metodologie (per la SSS è stato valutato sia il caso circolare che ellittico). L'equiparazione del numero massimo di aree, utilizzato nella tecnica FSC, alla proporzione massima di popolazione, impiegata nelle altre procedure, è stata ottenuta rapportando la popolazione a rischio, corrispondente alla percentuale scelta, al valore medio regionale di popolazione: per una frazione pari al 5%, l'equivalenza si ottiene con 17 comuni mentre, per una frazione del 3%, il numero di aree corrispondenti è 10 per entrambi i sessi.

Un *cluster* è stato definito persistente se risulta costituito da un insieme di collegamenti tra coppie di vertici presenti nelle matrici di adiacenza dei risultati esaminati in una percentuale uguale o superiore all'80%, ovvero da un insieme di collegamenti comuni ad almeno 4 aggregazioni su 5. Nel caso di una frazione di popolazione maschile pari al 5% è stato identificato, come *cluster* persistente, un insieme di 7 comuni di cui 6 appartenenti alla provincia di Bologna (Imola, Monzuno, Pianoro, San Lazzaro di Savena, San Benedetto Val di Sambro e Sasso Marconi) ed 1 appartenente alla provincia di Ravenna (Castel Bolognese) (figura 17). Per la popolazione femminile, le aree persistenti sono rappresentate da 6 comuni appartenenti al territorio bolognese (Castenaso, Imola, Mordano, Pianoro, San Lazzaro di Savena e Sasso Marconi) e 3 comuni della provincia di Ravenna (Bagnara di Romagna, Castel Bolognese e Solarolo) (figura 18).

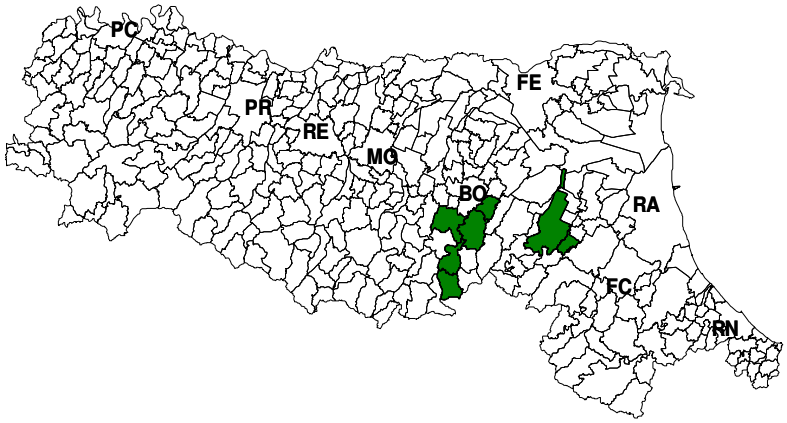


Figura 17 – Mappa CR95+ Maschi. *Cluster* persistente per una popolazione a rischio del 5%

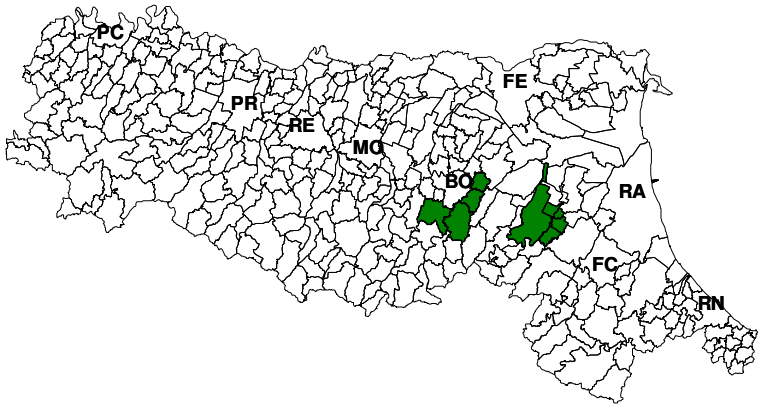


Figura 18 – Mappa CR95+ Femmine. *Cluster* persistente per una popolazione a rischio del 5%

Per una dimensione di popolazione del 3%, si ottiene una differenziazione maggiore delle aree persistenti tra i due sessi: il *cluster* maschile risulta costituito da un numero di comuni superiore a quello femminile. Le aree persistenti maschili sono 11 ed appartengono, prevalentemente, alla fascia montana bolognese (Casalecchio di Reno, Castiglione dei Pepoli, Grizzana Morandi, Marzabotto, Monte San Pietro, Monzuno, Pianoro, San Benedetto Val di Sambro, San Lazzaro di Savena, Sasso Marconi e Vergato) mentre la popolazione femminile longeva risulta persistere in soli 4 comuni anch'essi della

provincia di Bologna (Casalecchio di Reno, Pianoro, San Lazzaro di Savena e Sasso Marconi).
La valutazione della persistenza rappresenta un criterio innovativo per confrontare i risultati ottenuti con le diverse metodologie di *clustering* spaziale ma la sua efficacia rimane comunque vincolata alla soggettività insita nella scelta dei parametri di ricerca da parte del ricercatore, soprattutto nel caso di aggregazioni irregolari.

4.3 - Simulazioni

Le metodologie di *clustering* spaziale impiegate nello studio sono state valutate in un processo di simulazione nel quale sono stati ipotizzati 3 *clusters* di forma geografica diversa: circolare, ad "X" e a forma di "Y" (figura 19, A-B-C). In tutti i casi, i *clusters* sono costituiti da 7 comuni e comprendono una popolazione a rischio di circa 1000 individui (1000 per il primo *cluster*, 1015 per il *cluster* "X" e 1075 per *cluster* "Y"), corrispondente all'1% di quella totale, ed un numero di casi osservati compreso tra 100 e 110. La popolazione utilizzata è quella femminile e il numero di casi osservati, attribuito a ciascun comune, è stato determinato in funzione di quelli attesi, calcolati mediante una standardizzazione indiretta sotto l'ipotesi nulla di distribuzione uniforme. Ai comuni appartenenti al *cluster* è stato assegnato un numero di casi osservati pari al doppio di quelli attesi (in modo da ottenere un RR compreso in un intorno di 2) mentre alle rimanenti zone è stato attribuito un numero di casi osservati uguale a quello degli attesi (determinando un RR prossimo ad 1). La capacità di una procedura di identificare un *cluster* è stata valutata come rapporto tra il numero di comuni individuati correttamente a quelli esistenti realmente (*sens*).
Nella prima ipotesi (*cluster* circolare (A), popolazione=1000, casi 95+=100), la metodologia SSS (finestra circolare) identifica correttamente il *cluster* (*sens*=100%), anche se con l'aggiunta di un comune dovuto alla rigidità geometrica imposta. La procedura GGS (profondità=1;non-connettività=1)

individua perfettamente il *cluster* ipotizzato (*sens*=100%) così come l'algoritmo genetico SA (penalizzazione=0); in quest'ultimo caso, utilizzando una penalizzazione uguale ad 1, si ottiene lo stesso *cluster* identificato dalla SSS. La metodologia FSC riesce ad identificare, invece, solo 6 comuni su 7 (*sens*=86%) utilizzando una dimensione massima del *cluster* pari a 7: occorre fissare una dimensione massima K>7 per poter individuare completamente l'aggregazione ipotizzata. I risultati ottenuti consentono comunque di affermare che le metodologie utilizzate risultano efficaci e idonee nell'individuazione di un *cluster* di forma regolare.

Nella seconda ipotesi (*cluster* ad "X" (B), popolazione=1015, casi 95+=104), come era lecito attendersi, la procedura SSS identifica solo 4 comuni su 7 (*sens*=57%), sia nel caso circolare che ellittico (con tre livelli di penalizzazione: 0, 0.5 ed 1). La metodologia GGS, invece, individua correttamente i 7 comuni del *cluster*, in aggiunta ad un ottavo comune (profondità=1; non-connettività=1) (*sens*=100%); utilizzando un parametro di non-connettività superiore ad 1, si riduce la capacità di identificare il vero *cluster* rendendo l'aggregazione finale più compatta. L'algoritmo genetico GA identifica un *cluster* costituito da 7 comuni, di cui 6 appartenenti a quello ipotizzato (*sens*=86%), con una penalità nulla; al crescere della penalizzazione, il *cluster* tende ad assumere una forma più compatta e regolare allontanandosi dalla geometria del *cluster* simulato. Analogamente al primo caso, la FSC necessita di una dimensione massima superiore a 7 per poter identificare il *cluster*: per una dimensione pari a 7, essa individua correttamente solo 4 comuni su 7 (*sens*=57%).

La terza ipotesi (*cluster* ad "Y" (C), popolazione=1075, casi 95+=110), infine, evidenzia ancora la difficoltà della procedura SSS nell'identificazione di *clusters* irregolari: nel caso circolare, individua solo 4 comuni su 7 (*sens*=57%) mentre nel caso ellittico, indipendentemente dalla penalizzazione, i comuni identificati risultano 5 (*sens*=71%). La procedura GGS definisce un *cluster* costituito da 8 comuni di cui 7 appartenenti a quello simulato (*sens*=100%) anche variando i valori di profondità. La metodologia GA identifica correttamente, invece, un numero di comuni compreso tra 5 e 6 (*sens*>=71%), a seconda della penalizzazione imposta mentre, per la procedura FSC, è necessario utilizzare una dimensione massima superiore ad 8 per individuare 6 comuni su 7; nel caso

di dimensione uguale a 7, essa identifica solo 4 comuni presenti nel *cluster* (*sens*>=57%), similmente a quanto accade per la SSS.

Le procedure GGS e GA si sono rivelate le più idonee nell'identificazione dei *clusters* ipotizzati nello studio in particolare per le aggregazioni di forma irregolare. Il controllo della dimensione finale del *cluster*, invece, si rivela un limite della procedura FSC in presenza di aggregazioni molto irregolari in quanto non consente di identificare *clusters* geograficamente estesi sul territorio.

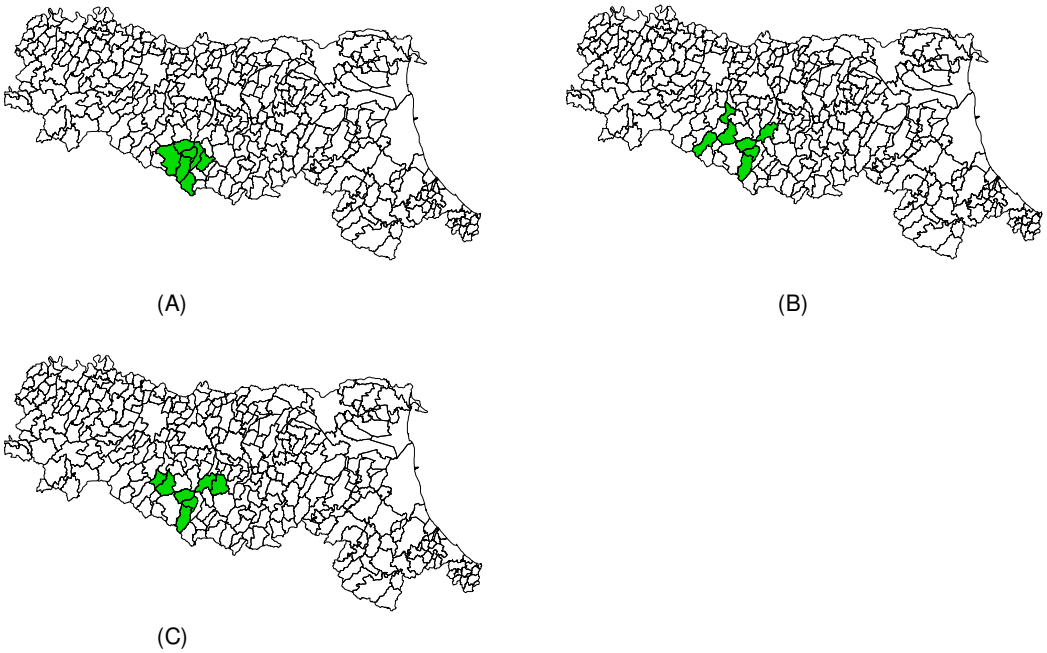


Figura 19 – *Clusters* utilizzati nel processo di simulazione. ((A) circolare; (B) ad "X"; (C) ad "Y")

4.4 - Scelta della dimensione di popolazione a rischio

L'identificazione del valore soglia di popolazione utilizzato per limitare la dimensione del *cluster* finale, in assenza di informazioni a priori del fenomeno, è un'operazione che introduce un grado di soggettività nella ricerca. Al tal

proposito, si è pensato di utilizzare un criterio di scelta basato sui valori caratteristici della distribuzione della misura di sintesi adottata nello studio. L'idea è di determinare, in maniera oggettiva, la dimensione del *cluster* più idonea a rappresentare il fenomeno analizzato mediante l'impiego dei percentili della distribuzione del Centenarian Rate (CR) determinando, per ciascuno di essi, il corrispondente ammontare di popolazione a rischio. La tabella 11 riporta i valori assoluti e le percentuali di popolazione relative alle misure di posizione scelte, separatamente per i due sessi.

Tabella 11 – Percentili del CR e corrispondente popolazione a rischio

	Percentile	**N° comuni**	**Popolazione a rischio (n° e % sul totale)**	
Maschile	75-*esimo*	84	22886	22.3%
"	80-*esimo*	69	14451	14.1%
"	85-*esimo*	51	8726	8.5%
"	90-*esimo*	34	4586	4.5%
"	93-*esimo*	24	2726	2.7%
"	95-*esimo*	17	1929	1.9%
Maschile	99-*esimo*	3	122	0.1%
Femminile	75-*esimo*	86	22461	20.7%
"	80-*esimo*	66	16189	14.9%
"	85-*esimo*	51	7631	7.0%
"	90-*esimo*	34	5460	5.0%
"	93-*esimo*	24	3765	3.5%
"	95-*esimo*	17	2309	2.1%
Femminile	99-*esimo*	3	397	0.4%

Dai risultati ottenuti, si nota che al 90-esimo percentile del CR corrisponde circa il 5% di popolazione a rischio, per entrambi i sessi (4586 M, 5460 F) mentre, al 99-esimo percentile, tale proporzione scende al di sotto dell'1%, rappresentando un limite troppo restrittivo per la dimensione del *cluster*. In considerazione dell'assenza di ipotesi a priori sulla distribuzione del fenomeno e dell'eventualità di effettuare scelte particolarmente drastiche è ipotizzabile utilizzare, come valore di riferimento per il nostro studio, il 90-esimo percentile, a conferma di quanto già ampiamente illustrato nella descrizione dei risultati.

4.5 - Conclusioni e discussione

La propensione alla longevità coinvolge in maniera diversa le aree del territorio dell'Emilia-Romagna. Le province della regione caratterizzate da una maggiore longevità sono Bologna, Ravenna e parte di Forlì-Cesena mentre, di contro, la provincia di Ferrara si distingue per un livello ridotto del fenomeno. La distinzione per sesso non appare netta: gli uomini con età superiore o uguale a 95 anni, numericamente inferiori alle donne, risiedono principalmente nei comuni delle province di Bologna e Ravenna, con qualche estensione nel territorio forlivese, analogamente a quanto accade per la popolazione femminile che mostra, tuttavia, una maggiore prevalenza nei territori di Bologna e Forlì-Cesena, includendo alcune aree del riminese. Le province occidentali della regione, invece, non risultano interessate significativamente da questo fenomeno; la caratterizzazione geografica e demografica di tali comuni non consente di evidenziare in maniera evidente il fenomeno: si tratta di zone con un'estensione territoriale modesta e con un esiguo ammontare di popolazione. Anche un'analisi supplementare condotta su una porzione di territorio limitata alle sole province di Piacenza, Parma, Reggio Emilia e Modena, non ha fornito risultati significativi in merito alla localizzazione di *clusters* di individui longevi in queste aree. Un'analisi della mortalità generale regionale conferma, in linea di massima, quanto osservato nel nostro studio (Regione ER,2007). Nel periodo 1998-2004, le province di Bologna, in particolare il distretto USL di Imola, Ravenna, Forlì-Cesena e parte di quella riminese, si caratterizzano per tassi standardizzati di mortalità generale inferiori al tasso medio regionale, per entrambi i sessi. Viceversa, le aree occidentali della regione si evidenziano per valori dei tassi di standardizzati di mortalità superiori a quello medio regionale.

Le metodologie di *cluster detection* utilizzate nello studio hanno prodotto risultati pressoché simili seppur con criteri di ricerca differenti. La *spatial scan statistic* si conferma una metodologia efficace e veloce ma il vincolo geometrico regolare imposto al *cluster* condiziona il suo utilizzo, rivelando una scarsa adattabilità nell'identificazione di aggregazioni irregolari. L'imposizione di una forma geometrica rigida può altresì condurre alla definizione di *clusters* a rischio elevato comprendenti anche quelle aree caratterizzate da un livello inferiore di

rischio (la considerazione si inverte quando si ricercano *clusters* con un livello ridotto di rischio).

La metodologia FSC ha evidenziato buone capacità di ricerca e velocità di esecuzione, completata da una descrizione chiara e dettagliata dei risultati e dalla possibilità di poter visualizzare graficamente i *clusters* finali, anche se con un livello minimo di dettaglio. Il limite principale della metodologia è la dimensione ridotta del *cluster* finale: l'eccessivo impegno computazionale richiesto dalla procedura induce a fissare il limite massimo al di sotto delle 30 aree, rendendola così utilizzabile solo nelle indagini in cui è ipotizzabile un'estensione limitata del fenomeno sul territorio. Una conseguenza legata alla ridotta dimensione del *cluster* è la scarsa capacità della procedura di identificare aggregazioni spaziali particolarmente irregolari ed estese sul territorio.

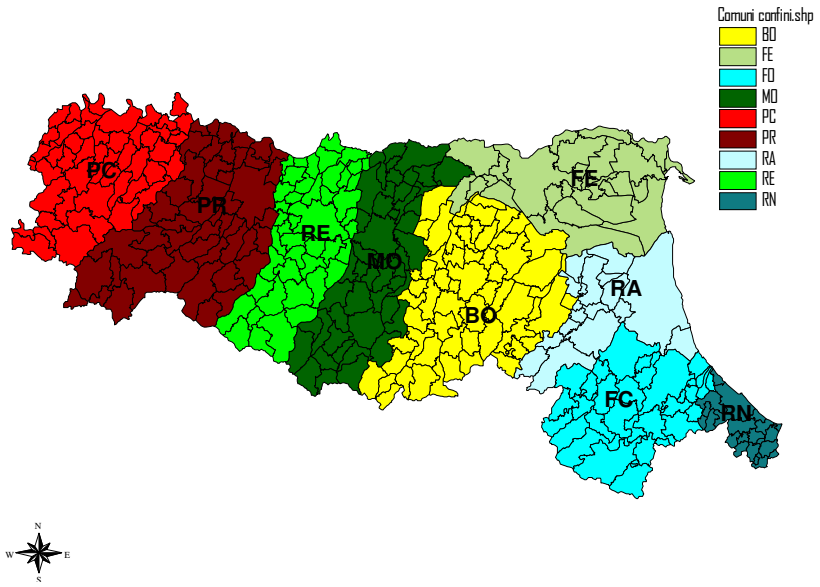
L'algoritmo genetico GA si rivela efficace nell'identificazione di *clusters* di qualsiasi forma ed estensione, seppur con una velocità di esecuzione inferiore rispetto alle procedure finora descritte. Senza un'adeguata selezione dei parametri di ricerca, tuttavia, la procedura può individuare *clusters* molto irregolari ed estesi, consigliando l'uso di penalizzazione non nulla in fase di analisi. La scelta dei parametri di ricerca non è comunque agevole ed immediata e, spesso, è lasciata all'esperienza del ricercatore. Questo modo di procedere, in aggiunta alla mancanza di informazioni a priori sul fenomeno, aumenta il grado di soggettività introdotto nella selezione dei parametri influenzando i risultati finali.

La metodologia GGS richiede un carico computazionale nettamente superiore rispetto a quello necessario per le altre metodologie e l'introduzione di due parametri di controllo favorisce una maggiore arbitrarietà nella selezione dei valori di ricerca adeguati; inoltre, la recente implementazione della procedura e la mancanza di studi su dati reali inducono ad effettuare un numero maggiore di prove durante la fase di ricerca dei *clusters*. Nel nostro studio, è stata ritenuta poco indicativa una profondità superiore a 2 in quanto non modifica, in modo sostanziale, i risultati ottenuti con valori inferiori, così come si è evidenziata l'opportunità di utilizzare una penalizzazione maggiore di zero per limitare l'irregolarità del *cluster.* Tale osservazione rimane, tuttavia, legata alla specificità del territorio e del fenomeno esaminato.

Il punto critico delle metodologie di *clustering* spaziale è il mancato controllo della variabilità extra-poissoniana presente nelle informazioni analizzate. Le ipotesi teoriche alla base del processo di ricerca postulano che la variabilità risulti interamente spiegata dal modello ed è noto, soprattutto per applicazioni in ambito reale, che tale ipotesi risulti spesso violata. L'eterogeneità della popolazione viene controllata utilizzando un processo di Poisson non omogeneo, in cui si ipotizza una variazione del numero di casi attesi tra le regioni esaminate ed un'indipendenza distributiva dei casi osservati sul territorio. Le metodologie utilizzate nello studio appartengono alla classe dei *cluster detection tests* per i quali l'obiettivo è valutare sia la presenza di una qualsiasi forma di *clustering* sul territorio che localizzare geograficamente gli eventuali *clusters*. Tali procedure non consentono e non intendono determinare la natura del processo sottostante al fenomeno analizzato, in quanto lo stesso risultato può derivare da processi differenti, ma bensì intendono verificare se l'ipotesi di rischio uniforme è stata violata in una specifica zona del territorio. In tal caso, un qualsiasi insieme di aree, diverso dalla zona che porta al rifiuto dell'ipotesi ma avente lo stesso numero di casi osservati, non viola l'assunzione di rischio uniforme, ipotizzando che il fenomeno si manifesti in modo simile per le aree esterne al *cluster* identificato. E' opportuno tener presente, tuttavia, che i confini di un *cluster* non coincidono esattamente con quelli del *cluster* reale ma rappresentano un punto di partenza per indagini successive nelle quali potranno eventualmente essere utilizzati approcci statistici differenti che meglio descrivono il fenomeno esaminato consentendo l'inserimento di eventuali covariate.

Appendice e Bibliografia

Fig.01 – Mappa delle province della regione Emilia-Romagna



App01 - Distribuzione dei comuni per provincia di residenza

Provincia	N°Comuni	%	%Cum.
Piacenza	48	14.08	14.08
Parma	47	13.78	27.86
Reggio Emilia	45	13.20	41.06
Modena	47	13.78	54.84
Bologna	60	17.60	72.43
Ferrara	26	7.62	80.06
Ravenna	18	5.28	85.34
Forlì-Cesena	30	8.80	94.13
Rimini	20	5.87	100.00
Totale	341	100.00	

App02 - Distribuzione dei comuni per numero medio di residenti

Comune	Freq.	%	%Cum.
< 999 abitanti	16	4.69	4.69
1.000 - 4.999 abitanti	149	43.70	48.39
5.000 -14.999 abitanti	129	37.83	86.22
15.000-49.999 abitanti	34	9.97	96.19
> 50.000 abitanti	13	3.81	100.00
Totale	341	100.00	

App03 - Distribuzione della popolazione media residente per provincia di residenza e per sesso in ordine decrescente di numerosità (periodo 2000-2004)

	popolazione media		
Provincia	totale	maschi	femmine
Bologna	925764	445512	480252
Modena	638784	312366	326418
Reggio Emilia	462848	227421	235427
Parma	402371	194939	207432
Forlì-Cesena	359844	175355	184489
Ravenna	354937	171933	183004
Ferrara	347540	166221	181319
Rimini	277374	134980	142394
Piacenza	268138	130049	138089
Regione ER	4037600	1958776	2078824

App04 - Distribuzione della popolazione media residente con età superiore o uguale a 95 anni per provincia di residenza e per sesso (periodo 2000-2004)

	popolazione media 95+ anni		
Provincia	totale	maschi	femmine
Bologna	1664	318	1346
Modena	884	170	714
Parma	792	136	656
Ravenna	759	168	591
Reggio Emilia	688	118	570
Forlì-Cesena	588	119	469
Piacenza	568	111	457
Ferrara	504	87	417
Rimini	366	77	289
Totale	6813	1304	5509

App05 - Distribuzione della popolazione media residente con età superiore o uguale a 100 anni per provincia di residenza e per sesso (periodo 2000-2004)

	popolazione media 100+ anni		
Provincia	totale	maschi	femmine
Bologna	142	18	124
Ravenna	80	14	66
Modena	76	12	64
Parma	65	6	59
Piacenza	55	6	49
Reggio Emilia	50	5	45
Forlì-Cesena	47	7	40
Ferrara	38	6	32
Rimini	29	3	26
Regione ER	582	77	505

App06 - Distribuzione della popolazione residente con età 55-59 anni, al censimento del 1961, per provincia di residenza e per sesso

Provincia	popolazione 55-59 anni (censimento 1961) totale	maschi	femmine
Piacenza	18211	8917	9294
Parma	24378	11879	12499
Reggio Emilia	22725	11036	11689
Modena	28778	14219	14559
Bologna	50786	24058	26728
Ferrara	22984	11156	11828
Ravenna	18008	9003	9005
Forlì-Cesena	16103	7809	8294
Rimini	9245	4437	4808
Regione ER	211218	102514	108704

App07 - Distribuzione della popolazione media totale per provincia di residenza e sesso (periodo 2000-2004)

Provincia		min	max	mean	sd	cv	p25	p50	p75
Piacenza	popol.totale	136	98368	5586.208	13966.28	2.500136	1783.5	2946.5	5029.5
	pop.media M	67	46499	2709.354	6598.249	2.435358	881	1472.5	2438
	pop.media F	69	51869	2876.854	7368.312	2.561239	902.5	1478.5	2568
Parma	popol.totale	692	170874	8561.085	24605.73	2.874137	1916	3790	7212
	pop.media M	352	81178	4147.638	11684.11	2.817052	965	1887	3537
	pop.media F	340	89696	4413.447	12921.85	2.927836	926	1885	3697
Reggio Emilia	popol.totale	1003	148070	10285.51	21520.05	2.092269	4160	6221	8949
	pop.media M	501	71873	5053.8	10440.97	2.065965	2081	3037	4392
	pop.media F	498	76197	5231.711	11079.32	2.117724	2072	3184	4581
Modena	popol.totale	745	177637	13591.15	26944.37	1.982494	3034	6776	13223
	pop.media M	387	85351	6646.085	12955.37	1.949324	1538	3249	6505
	pop.media F	358	92286	6945.064	13990.06	2.014389	1537	3390	6718
Bologna	popol.totale	1222	377322	15429.4	48513.12	3.144200	4176.5	6170.5	11873.5
	pop.media M	602	176106	7425.2	22649.99	3.050421	2113.5	2993.5	5776.5
	pop.media F	620	201216	8004.2	25864.65	3.231385	2063	3129	6097
Ferrara	popol.totale	2340	131651	13366.92	25182.3	1.883927	3815	6013	13059
	pop.media M	1115	61614	6393.115	11799.7	1.845688	1824	2916.5	6242
	pop.media F	1187	70037	6973.808	13384.62	1.919270	1978	3096.5	6817
Ravenna	popol.totale	1782	141219	19718.72	32951.62	1.671083	5329	8422	16088
	pop.media M	869	68574	9551.833	15980.16	1.672994	2650	4100	7772
	pop.media F	913	72645	10166.89	16972.13	1.669354	2679	4322	8316
Forlì-Cesena	popol.totale	858	108602	11994.8	24428.22	2.036568	2115	5173.5	9387
	pop.media M	412	51989	5845.167	11755.72	2.011186	1053	2559.5	4622
	pop.media F	441	56613	6149.633	12673.4	2.060839	1092	2614	4779
Rimini	popol.totale	926	132162	13868.7	29059.98	2.095364	1885	4597.5	12732.5
	pop.media M	451	63765	6749	14009.1	2.075730	931.5	2273	6249.5
	pop.media F	475	68397	7119.7	15051.18	2.114019	953.5	2324.5	6483
Regione ER	popol.totale	136	377322	11840.47	29802.35	2.516991	2838	5225	9387
	poptotmmedia	67	176106	5744.211	14116.51	2.457520	1371	2571	4622
	pop.media F	69	201216	6096.258	15690.02	2.573714	1407	2665	4779

App08 - Distribuzione della popolazione media con età superiore o uguale a 95 anni (100 anni) per provincia di residenza e sesso (periodo 2000-2004)

Provincia		min	max	mean	sd	cv	p25	p50	p75
Piacenza	pop.media95+	1	172	11.83333	24.32609	2.055726	4.5	7.5	11.5
	pop.media95+M	0	31	2.3125	4.420245	1.911457	1	1.5	2
	pop.media95+F	1	141	9.520833	20.02338	2.103112	3	6	9
	pop.media100+	0	19	1.145833	2.828349	2.468377	0	.5	1
	pop.media100+M	0	2	.125	.3927535	3.142028	0	0	0
	pop.media100+F	0	18	1.020833	2.653697	2.599540	0	0	1
Parma	pop.media95+	2	298	16.85106	42.87189	2.544165	6	10	13
	pop.media95+M	0	49	2.893617	7.10858	2.456642	1	2	3
	pop.media95+F	2	249	13.95745	35.81229	2.565819	5	8	11
	pop.media100+	0	30	1.382979	4.48453	3.242660	0	0	1
	pop.media100+M	0	3	.1276596	.4941845	3.871112	0	0	0
	pop.media100+F	0	27	1.255319	4.035	3.214322	0	0	1
Reggio Emilia	pop.media95+	2	216	15.28889	31.23054	2.042695	7	9	14
	pop.media95+M	0	38	2.622222	5.597438	2.134616	1	2	2
	pop.media95+F	2	178	12.66667	25.74967	2.032869	5	8	11
	pop.media100+	0	22	1.111111	3.262892	2.936603	0	1	1
	pop.media100+M	0	3	.1111111	.4872102	4.384891	0	0	0
	pop.media100+F	0	19	1	2.828427	2.828427	0	0	1
Modena	pop.media95+	2	237	18.80851	35.11883	1.867178	7	10	17
	pop.media95+M	0	39	3.617021	5.921784	1.637199	1	2	4
	pop.media95+F	1	198	15.19149	29.34772	1.931853	6	8	14
	pop.media100+	0	23	1.617021	3.517368	2.175214	0	1	1
	pop.media100+M	0	3	.2553191	.6067764	2.376541	0	0	0
	pop.media100+F	0	20	1.361702	3.060446	2.247515	0	1	1
Bologna	pop.media95+	2	780	27.73333	100.6123	3.627849	7	9	17.5
	pop.media95+M	0	134	5.3	17.53863	3.309175	1	2	4
	pop.media95+F	2	646	22.43333	83.18641	3.708161	5	7.5	13.5
	pop.media100+	0	71	2.366667	9.268073	3.916087	0	1	1.5
	pop.media100+M	0	10	.3	1.331496	4.438320	0	0	0
	pop.media100+F	0	61	2.066667	7.963618	3.853363	0	1	1
Ferrara	pop.media95+	2	197	19.38462	37.533	1.936226	5	9.5	18
	pop.media95+M	0	28	3.346154	5.677621	1.696760	1	2	3
	pop.media95+F	2	169	16.03846	32.06179	1.999057	4	8	15
	pop.media100+	0	16	1.461538	3.139819	2.148297	0	.5	2
	pop.media100+M	0	2	.2307692	.5144078	2.229100	0	0	0
	pop.media100+F	0	14	1.230769	2.746746	2.231731	0	.5	1
Ravenna	pop.media95+	5	248	42.16667	59.21769	1.404372	15	19.5	40
	pop.media95+M	0	54	9.333333	12.81543	1.373082	3	5	9
	pop.media95+F	4	194	32.83333	46.45333	1.414822	12	15	29
	pop.media100+	0	30	4.444444	7.022839	1.580139	1	2	6
	pop.media100+M	0	6	.7777778	1.437136	1.847746	0	0	1
	pop.media100+F	0	24	3.666667	5.625572	1.534247	1	2	5
Forlì-Cesena	pop.media95+	2	216	19.6	42.57861	2.172378	4	6.5	14
	pop.media95+M	0	44	3.966667	8.755228	2.207200	1	1	3
	pop.media95+F	1	172	15.63333	33.89383	2.168049	3	5.5	12
	pop.media100+	0	15	1.566667	2.955805	1.886684	0	1	2
	pop.media100+M	0	3	.2333333	.6789106	2.909617	0	0	0
	pop.media100+F	0	12	1.333333	2.339073	1.754305	0	1	1
Rimini	pop.media95+	1	195	18.3	43.01542	2.350569	2	4.5	12
	pop.media95+M	0	37	3.85	8.177472	2.124019	.5	1	3.5
	pop.media95+F	1	158	14.45	34.89002	2.414534	2	3.5	9
	pop.media100+	0	17	1.45	3.88621	2.680145	0	0	1
	pop.media100+M	0	1	.15	.3663475	2.442317	0	0	0
	pop.media100+F	0	16	1.3	3.672085	2.824681	0	0	.5
Regione ER	pop.media95+	1	780	19.97947	54.4777	2.726684	6	9	16
	pop.media95+M	0	134	3.824047	9.75817	2.551792	1	2	3
	pop.media95+F	1	646	16.15543	44.85627	2.776545	5	7	13
	pop.media100+	0	71	1.706745	5.211039	3.053203	0	1	1
	pop.media100+M	0	10	.2258065	.7963389	3.526644	0	0	0
	pop.media100+F	0	61	1.480938	4.49808	3.037317	0	1	1

**per la provincia di Ferrara, i valori relativi alla popolazione 100+ sono stati calcolati escludendo tre comuni (Migliaro, Tresigallo e Goro) per i quali non si disponeva del loro ammontare. In merito alla popolazione 95+, i valori mancanti sono stati sostituiti con i valori medi della popolazione (divisa per sesso) dei comuni con un numero di abitanti inferiore a 5000.

App09 - Distribuzione della popolazione con età compresa tra 55 e 59 anni, al censimento del 1961, per provincia di residenza e sesso (periodo 2000-2004)

Provincia		min	max	mean	p50	sd	cv	p25	p75
Piacenza	Pop.55-59aa (1961)	38	5213	379	267	726.8593	1.915834	192	350
	Pop.55-59aa Maschi	18	2372	186	132	329.8298	1.775466	97	176
	Pop.55-59aa Femmine	20	2841	194	129	397.2175	2.051479	92	176
Parma	Pop.55-59aa (1961)	97	8482	519	318	1208.105	2.329188	197	429
	Pop.55-59aa Maschi	47	3972	253	142	564.964	2.235315	103	216
	Pop.55-59aa Femmine	50	4510	266	151	643.2991	2.418998	95	208
Reggio Emilia	Pop.55-59aa (1961)	126	7059	505	319	1017.813	2.015470	243	418
	Pop.55-59aa Maschi	54	3339	245	164	480.9738	1.961202	118	205
	Pop.55-59aa Femmine	61	3720	260	158	537.0023	2.067337	125	215
Modena	Pop.55-59aa (1961)	82	7969	612	311	1170.575	1.911774	233	598
	Pop.55-59aa Maschi	36	3788	302	158	557.5346	1.842895	115	305
	Pop.55-59aa Femmine	46	4181	310	155	613.3302	1.979979	117	289
Bologna	Pop.55-59aa (1961)	103	27284	846	271	3498.592	4.133335	189	407
	Pop.55-59aa Maschi	47	12220	401	142	1567.114	3.908339	97	204
	Pop.55-59aa Femmine	56	15064	445	134	1931.824	4.336631	94	205
Ferrara	Pop.55-59aa (1961)	141	8683	884	455	1639.618	1.854423	272	783
	Pop.55-59aa Maschi	72	4130	429	232	779.399	1.816434	147	382
	Pop.55-59aa Femmine	69	4553	455	222	860.109	1.890669	145	401
Ravenna	Pop.55-59aa (1961)	109	6109	1000	575	1444.944	1.444302	234	1001
	Pop.55-59aa Maschi	58	3035	500	284	718.8837	1.437288	111	476
	Pop.55-59aa Femmine	51	3074	500	275	726.354	1.451901	124	506
Forlì-Cesena	Pop.55-59aa (1961)	72	4849	537	233	1052.679	1.961148	164	394
	Pop.55-59aa Maschi	30	2329	260	111	504.4864	1.938096	79	185
	Pop.55-59aa Femmine	42	2520	276	128	548.3305	1.983351	80	209
Rimini	Pop.55-59aa (1961)	52	4348	462	188	942.9944	2.040010	103	361
	Pop.55-59aa Maschi	28	2046	222	96	442.5886	1.994990	51	189
	Pop.55-59aa Femmine	23	2302	240	87	500.6303	2.082489	52	176
Regione	Pop.55-59aa (1961)	38	27284	619	296	1787.723	2.888232	200	443
	Pop.55-59aa Maschi	18	12220	301	149	817.176	2.718234	101	224
	Pop.55-59aa Femmine	20	15064	319	150	971.4034	3.047253	99	221

App10 - Distribuzione del rapporto Femmine/Maschi della popolazione media residente con età superiore a 95 anni, del rapporto F/M della popolazione totale e del rapporto CR95F/CR95M (periodo 2000-2004)

Provincia		min	max	mean	sd	p25	p50	p75
Piacenza	rapporto CR F/M	.4807692	12.94615	3.972801	2.678664	2.25	3.440601	4.89433
	rapporto pop F/M	.7121212	1.115486	1.006577	.0710665	.9852634	1.023631	1.051238
	rapporto F/M 95+	.5	11	3.875358	2.518284	2.5	3	4.548387
Parma	rapporto CR F/M	1.521739	18.66667	4.988391	3.403375	3	3.9375	6.411765
	rapporto pop F/M	.9353535	1.10493	1.020536	.0391976	.9985097	1.025133	1.043925
	rapporto F/M 95+	2	16	4.892865	2.910843	3	4.0625	5.081633
Reggio Emilia	rapporto CR F/M	1.808279	11.59725	4.559315	2.43232	2.842105	3.647766	6.045454
	rapporto pop F/M	.9254131	1.091049	1.017865	.0327971	1.000954	1.017955	1.04072
	rapporto F/M 95+	1.666667	14	4.778273	2.712257	3	3.666667	6
Modena	rapporto CR F/M	.7164178	11.19718	4.375853	2.560717	2.974359	3.572241	5.424
	rapporto pop F/M	.9250646	1.085565	1.018938	.0401998	.9914894	1.022797	1.055295
	rapporto F/M 95+	1	11	4.35987	2.503465	2.666667	3.555556	5.2
Bologna	rapporto CR F/M	1.319444	9.822785	4.458621	2.288051	2.407815	3.960526	6.795181
	rapporto pop F/M	.9462422	1.147059	1.023497	.0406353	.9964804	1.017257	1.04657
	rapporto F/M 95+	1.5	8	4.380438	2.191699	2.4	4	6
Ferrara	rapporto CR F/M	1.584906	7.654676	4.341758	1.932091	2.721538	3.913203	5.701341
	rapporto pop F/M	1.014362	1.136706	1.068394	.0287755	1.04468	1.064191	1.091557
	rapporto F/M 95+	2	8	4.520346	1.931947	2.958333	4.125	6.017857
Ravenna	rapporto CR F/M	1.645161	5.865854	3.558015	1.037173	3.061594	3.511076	3.727407
	rapporto pop F/M	.9832285	1.104412	1.053142	.0323953	1.0267	1.062529	1.069995
	rapporto F/M 95+	2	6.5	3.54952	1.021909	3	3.473684	4
Forlì Cesena	rapporto CR F/M	.7142857	12.83117	4.022881	2.766576	2.421919	2.995496	5.290578
	rapporto pop F/M	.962963	1.088942	1.022117	.0326139	1.000324	1.017241	1.046733
	rapporto F/M 95+	1	13	4.137348	2.822272	2.166667	3.36	5.625
Rimini	rapporto CR F/M	.8811188	4.168831	2.611259	1.102559	1.6	2.955789	3.658427
	rapporto pop F/M	.9518619	1.087828	1.021151	.0386043	.9897819	1.030627	1.041914
	rapporto F/M 95+	1	4.4	2.634685	1.157098	2	3	3.6
Regione ER	rapporto CR F/M	.4807692	18.66667	4.263145	2.532913	2.51087	3.609022	5.546875
	rapporto pop F/M	.7121212	1.147059	1.024065	.045687	1.000831	1.028078	1.050761
	rapporto F/M 95+	.5	16	4.271198	2.450139	2.5	3.75	5.5

Nota: i valori riportati in tabella possono differire da quelli calcolati direttamente rapportando i totali di provincia in quanto sono presenti errori di approssimazione dovuti al calcolo dei rapporti per ogni singolo comune. I valori medi sono stati calcolati facendo la media dei singoli rapporti comunali e non sui totali assoluti per provincia.

App11 - Distribuzione del Centenarian Rate (CR), per provincia di residenza e sesso, relativo alla popolazione con età superiore o uguale a 95 anni (periodo 2000-2004)

Provincia		min	max	mean	sd	cv	p25	p50	p75
Piacenza	cr95	.0086957	.0588235	.0305044	.012146	.3981736	.0219475	.027687	.0387816
	cr95m	0	.08	.0146121	.0144124	.9863328	.0077879	.0111166	.0164332
	cr95f	.009901	.0846154	.0467386	.0183961	.3935957	.0362948	.0474782	.0570461
Parma	cr95	.0136986	.0618557	.0300963	.0095432	.3170886	.0232558	.0299065	.0358705
	cr95m	0	.0265487	.0100394	.0067707	.6744124	.0047393	.0106383	.0144231
	cr95f	.0217391	.12	.0503756	.0173047	.343513	.0397351	.0472973	.0625
Reggio Emilia	cr95	.0118343	.0518519	.0302447	.0084348	.2788844	.0250784	.0285714	.0364299
	cr95m	0	.0310078	.0102410	.006831	.6670269	.0057803	.0104712	.014881
	cr95f	.0211268	.0803571	.0496108	.0150538	.3034384	.0387597	.0466667	.0588235
Modena	cr95	.0086957	.0613108	.0317255	.0092925	.292902	.0252366	.0321156	.0380952
	cr95m	0	.0316456	.0129378	.0076358	.5901943	.0077519	.0124224	.0181818
	cr95f	.0074627	.0986547	.0506749	.0176711	.3487159	.038835	.0480000	.0636132
Bologna	cr95	.0077519	.0788644	.0361926	.0130871	.3615964	.0262011	.0349244	.04341
	cr95m	0	.0416667	.0147016	.0088715	.6034408	.0089314	.0127518	.0208356
	cr95f	.016	.1418919	.0581624	.0229648	.394839	.0426283	.0547386	.0724694
Ferrara	cr95	.0039841	.0393701	.0210840	.0076556	.3630992	.0164502	.0210897	.023845
	cr95m	0	.0164835	.0073738	.0054926	.7448797	.0041667	.0069048	.0125
	cr95f	.0088496	.0603015	.0343762	.0125037	.3637307	.0272727	.0342242	.040107
Ravenna	cr95	.0229358	.0641026	.0442519	.0107364	.2426201	.0377604	.0443033	.0521008
	cr95m	0	.0392157	.0190027	.0082512	.4342132	.0166667	.0179052	.0236686
	cr95f	.0353982	.1176471	.0699435	.0201155	.287596	.0573123	.0692084	.0754717
Forlì-Cesena	cr95	.0132743	.0628931	.0341040	.0122718	.3598356	.0253807	.0341723	.0432099
	cr95m	0	.0506329	.0156438	.0101545	.6491107	.0100251	.0133353	.0196078
	cr95f	.0178571	.0970464	.0522788	.0212019	.4055543	.0357143	.0522592	.0697674
Rimini	cr95	.009434	.0652174	.0312772	.0140487	.4491675	.020274	.0294813	.039501
	cr95m	0	.0396825	.0159398	.0119079	.7470548	.0057637	.0159086	.0250156
	cr95f	.0181818	.1168831	.0468307	.0230544	.4922918	.0301587	.0437980	.0579259
Regione ER	cr95	.0039841	.0788644	.0319525	.0117936	.3690968	.0240385	.0305992	.0393701
	cr95m	0	.08	.0130385	.0096639	.7411837	.0073529	.0119048	.0172414
	cr95f	.0074627	.1418919	.0509466	.0200846	.394229	.037037	.0480000	.0636132

App12 - Analisi effettuate con la *Spatial Scan Statistic* SSS (finestra circolare)

Finestra	Dim.max Popolazione	Repliche	Analisi	Numeratore (casi)	Denominatore (popolazione)	Tempo elaborazione	High/low level
Circolare	10%	9999	Puramente spaziale	95+ totale	55-59aa (1961)	3 sec	H/L
"	20%	"	"	"	"	35 sec	"
"	25%	"	"	"	"	37 sec	"
"	30%	"	"	"	"	2 min	"
"	50%	9	"	"	"	2 sec	"
"	50%	999	"	"	"	5 sec	"
"	50%	9999	"	"	"	45 sec	"
Circolare	50%	99999	Puramente spaziale	95+ totale	55-59aa (1961)	8 min	H/L
Circolare	1%	9999	Puramente spaziale	95+ Maschi	55-59 M (1961)	21 sec	H/L
"	3%	"	"	"	"	22 sec	"
"	5%	"	"	"	"	21 sec	"
"	10%	"	"	"	"	26 sec	"
"	15%	"	"	"	"	33 sec	"
"	20%	"	"	"	"	35 sec	"
"	25%	"	"	"	"	36 sec	"
"	30%	9999	"	"	"	55 sec	"
"	50%	999	"	"	"	10 sec	"
"	50%	9999	"	"	"	44 sec	"
Circolare	50%	99999	Puramente spaziale	95+ Maschi	55-59 M (1961)	7 min	H/L
Circolare	1%	9999	Puramente spaziale	95+ Femmine	55-59 F (1961)	28 sec	H/L
"	3%	"	"	"	"	27 sec	"
"	5%	"	"	"	"	28 sec	"
"	10%	"	"	"	"	28 sec	"
"	15%	"	"	"	"	25 sec	"
"	20%	"	"	"	"	50 sec	"
"	25%	"	"	"	"	53 sec	"
"	30%	9999	"	"	"	40 sec	"
"	50%	999	"	"	"	5 sec	"
"	50%	9999	"	"	"	45 sec	"
Circolare	50%	99999	Puramente spaziale	95+ Femmine	55-59 F (1961)	9 min	H/L

App13 – Analisi effettuate con la *Spatial Scan Statistic* SSS (finestra ellittica)

Finestra	Penalità	Dim.max popolazione	Repliche	Numeratore (casi)	Denominatore (popolazione)	Tempo elaborazione	High/low level
Ellisse	Media	10%	9999	95+ totale	55-59 (1961)	9 min	H/L
"	Nessuna	15%	"	"	"	9 min	"
"	Media	15%	"	"	"	10 min	"
"	Forte	15%	"	"	"	16 min	"
"	Nessuna	20%	"	"	"	10 min	"
"	Media	20%	"	"	"	16 min	"
"	Forte	20%	"	"	"	16 min	"
"	Nessuna	30%	"	"	"	15 min	"
"	Media	30%	"	"	"	20 min	"
"	Forte	30%	"	"	"	23 min	"
"	Nessuna	50%	"	"	"	14 min	"
"	Media	50%	"	"	"	16 min	"
Ellisse	Forte	50%	9999	95+ totale	55-59 (1961)	21 min	H/L
Ellisse	Nessuna	1%	9999	95+ Maschi	55-59 M (1961)	1 min	H/L
"	Media	1%	"	"	"	2 min	"
"	Forte	1%	"	"	"	2 min	"
"	Nessuna	3%	"	"	"	2 min	"
"	Media	3%	"	"	"	3 min	"
"	Forte	3%	"	"	"	3 min	"
"	Nessuna	5%	"	"	"	4 min	"
"	Media	5%	"	"	"	5 min	"
"	Forte	5%	"	"	"	5 min	"
"	Nessuna	10%	"	"	"	5 min	"
"	Media	10%	"	"	"	5 min	"
"	Forte	10%	"	"	"	5 min	"
"	Nessuna	15%	"	"	"	7 min	"
"	Media	15%	"	"	"	8 min	"
"	Forte	15%	"	"	"	7 min	"
"	Nessuna	20%	"	"	"	8 min	"
"	Media	20%	"	"	"	7 min	"
"	Forte	20%	"	"	"	6 min	"
"	Nessuna	30%	"	"	"	12 min	"
"	Media	30%	"	"	"	9 min	"
"	Forte	30%	"	"	"	10 min	"
"	Nessuna	50%	"	"	"	11 min	"
"	Media	50%	"	"	"	15 min	"
Ellisse	Forte	50%	9999	95+ Maschi	55-59 M (1961)	18 min	H/L
Ellisse	Nessuna	1%	9999	95+ Femmine	55-59 F (1961)	1 min	H/L
"	Media	1%	"	"	"	2 min	"
"	Forte	1%	"	"	"	2 min	"
"	Nessuna	3%	"	"	"	2 min	"
"	Media	3%	"	"	"	4 min	"
"	Forte	3%	"	"	"	4 min	"
"	Nessuna	5%	"	"	"	5 min	"
"	Media	5%	"	"	"	6 min	"
"	Forte	5%	"	"	"	6 min	"
"	Nessuna	10%	"	"	"	6 min	"
"	Media	10%	"	"	"	6 min	"
"	Forte	10%	"	"	"	4 min	"
"	Nessuna	15%	"	"	"	5 min	"
"	Media	15%	"	"	"	5 min	"
"	Forte	15%	"	"	"	7 min	"
"	Nessuna	20%	"	"	"	8 min	"
"	Media	20%	"	"	"	7 min	"
"	Forte	20%	"	"	"	8 min	"
"	Nessuna	30%	"	"	"	13 min	"
"	Media	30%	"	"	"	13 min	"
"	Forte	30%	"	"	"	10 min	"
"	Nessuna	50%	"	"	"	21 min	"
"	Media	50%	"	"	"	22 min	"
Ellisse	Forte	50%	9999	95+ Femmine	55-59 F (1961)	23 min	H/L

App14 – Analisi effettuate con la *Flexible Scan Statistic* FSC
(solo *cluster* primario ad elevata longevità)

N° max aree cluster	Repliche	Numeratore (casi)	Denominatore (popolazione)	Tipo *Cluster*	Tempo elaborazione	High/low level
10	999	95+ totale	55-59aa(1961)	Irregolare	5 sec	High
15	"	"	"	"	1 min	"
20	999	"	"	"	13 min	"
10	9999	"	"	"	10 sec	"
15	"	"	"	"	23 min	"
15	999	"	"	"	48 sec	"
17	"	"	"	"	1 min	"
25	"	"	"	"	5h 46 min	"
30	"	95+ totale	55-59aa(1961)	Irregolare	3 gg (45/341 aree)	High
20	999	95+ totale	55-59aa(1961)	Circolare	11 sec	High
20	9999	"	"	"	11 sec	"
50	999	"	"	"	3 sec	"
50	9999	"	"	"	31 sec	"
75	9999	95+ totale	55-59aa(1961)	Circolare	*errore*	High
1	9999	95+ Maschi	55-59aa M(1961)	Irregolare	2 sec	High
2	"	"	"	"	2 sec	"
3	"	"	"	"	3 sec	"
4	"	"	"	"	3 sec	"
5	9999	"	"	"	4 sec	"
10	999	"	"	"	4 sec	"
15	"	"	"	"	56 sec	"
17	"				14 min	"
20	"	"	"	"	18 min	"
25	999	"	"	"	8h 40 min	"
10	9999	"	"	"	22 sec	"
15	9999	95+ Maschi	55-59aa M(1961)	Irregolare	13 min	High
20	9999	95+ Maschi	55-59aa M(1961)	Circolare	17 sec	High
25	"	"	"	"	8 sec	"
50	9999	"	"	"	12 sec	"
75	999	"	"	"	*errore*	"
75	9999	95+ Maschi	55-59aa M(1961)	Circolare	*errore*	High
1	9999	95+ Femmine	55-59aa F(1961)	Irregolare	2 sec	High
2		"	"	"	2 sec	"
3	"	"	"	"	2 sec	"
4	"	"	"	"	2 sec	"
5	9999	"	"	"	2 sec	"
10	999	"	"	"	2 sec	"
15	"	"	"	"	34 sec	"
20	"	"	"	"	25 min	"
25	999	"	"	"	5 h 50 min	"
10	9999	"	"	"	44 sec	"
15	9999	95+ Femmine	55-59aa F(1961)	Irregolare	27 min	High
20	9999	95+ Femmine	55-59aa F(1961)	Circolare	14 sec	High
25	"	"	"	"	15 sec	"
50	"	"	"	"	23 sec	"
75	9999	"	"	"	*errore*	"
99	999	"	"	"	*errore*	"
100	9999	95+ Femmine	55-59aa F(1961)	Circolare	*errore*	High

App15 – Analisi effettuate con l'Algoritmo Genetico GA
(solo *cluster* primario ad elevata longevità);
(G=n°generazioni;w=*crossing-over;* Pop=dimensione massima di popolazione);
(il tasso di mutazione è fissato all'1% salvo indicazioni contrarie)

Pop	Repliche	Penalità	w	G	Numeratore (casi)	Denominatore (popolazione)	Tempo elaborazione	Level cluster
1%	999	0.0	400	10	95+ Maschi	55-59aa M(1961)	5 min	High
1%	"	0.5	"	"	"	"	6 min	"
1%	"	1.0	"	"	"	"	6 min	"
3%	"	0.0	"	"	"	"	11 min	"
3%	"	0.5	"	"	"	"	13 min	"
3%	"	1.0	"	"	"	"	12 min	"
5%	"	0.0	"	"	"	"	16 min	"
5%	"	0.5	"	"	"	"	18 min	"
5%	"	1.0	"	"	"	"	12 min	"
10%	"	0.0	"	"	"	"	27 min	"
10%	"	0.5	"	"	"	"	27 min	"
10%	"	1.0	"	"	"	"	19 min	"
15%	"	0.0	"	"	"	"	12 min	"
15%	"	0.5	"	"	"	"	29 min	"
15%	"	1.0	"	"	"	"	28 min	"
20%	"	0.0	"	"	"	"	45 min	"
20%	"	0.5	"	"	"	"	39 min	"
20%	"	1.0	"	"	"	"	40 min	"
25%	"	0.0	"	"	"	"	31 min	"
25%	"	0.5	"	"	"	"	33 min	"
25%	"	1.0	"	"	"	"	32 min	"
30%	"	0.0	"	"	"	"	41 min	"
30%	"	0.5	"	"	"	"	39 min	"
30%	"	1.0	400	10	"	"	40 min	"
15%	"	1.0	500	10	"	"	32 min	"
15%	"	1.0	750	10	"	"	1h 15 min	"
15%	"	1.0	1000	10	"	"	1h 40 min	"
15%**	"	0.5	400	10	"	"	30 min	"
15%**	"	1.0	"	"	"	"	28 min	"
20%**	999	1.0	"	"	"	"	39 min	"
15%	9999	1.0	400	10	95+ Maschi	55-59aa M(1961)	4h 50 min	High
1%	999	0.0	400	10	95+ Femmine	55-59aa F(1961)	5 min	High
1%	"	0.5	"	"	"	"	7 min	"
1%	"	1.0	"	"	"	"	6 min	"
3%	"	0.0	"	"	"	"	15 min	"
3%	"	0.5	"	"	"	"	14 min	"
3%	"	1.0	"	"	"	"	16 min	"
5%	"	0.0	"	"	"	"	21 min	"
5%	"	0.5	"	"	"	"	20 min	"
5%	"	1.0	"	"	"	"	18 min	"
10%	"	0.0	"	"	"	"	26 min	"
10%	"	0.5	"	"	"	"	20 min	"
10%	"	1.0	"	"	"	"	19 min	"
15%	"	0.0	"	"	"	"	22 min	"
15%	"	0.5	"	"	"	"	30 min	"
15%	"	1.0	"	"	"	"	32 min	"
20%	"	0.0	"	"	"	"	50 min	"
20%	"	0.5	"	"	"	"	42 min	"
20%	"	1.0	"	"	"	"	38 min	"
25%	"	0.0	"	"	"	"	50 min	"
25%	"	0.5	"	"	"	"	45 min	"
25%	"	1.0	"	"	"	"	40 min	"
30%	"	0.0	"	"	"	"	50 min	"
30%	"	0.5	"	"	"	"	48 min	"
30%	"	1.0	"	"	"	"	47 min	"
15%	"	1.0	400	50	"	"	1h 25 min	"
15%	"	1.0	400	100	"	"	1h 50 min	"
15%	"	1.0	400	200	"	"	2h 20 min	"
15%	9999	1.0	400	10	95+ Femmine	55-59aa F(1961)	5h	High

** Il tasso di mutazione è stato fissato all'1x1000 (0.001)

App16 – Analisi effettuate con la *Greedy Growth Search* GGS
(u =parametro di profondità; α =parametro di non-connettività)
(solo *cluster* primario ad elevata longevità)

Dim.max pop	u	α	Repliche	Numeratore (Casi)	Denominatore (Popolazione)	Tempo elaborazione	Level cluster
1%	1	1	999	95+ Maschi	55-59aa M (1961)	30 min	High
1%	2	2	"	"	"	40 min	"
3%	1	1	"	"	"	50 min	"
3%	1	2	"	"	"	51 min	"
3%	2	1	"	"	"	1h 10 min	"
5%	2	1	"	"	"	2h	"
5%	2	2	"	"	"	3h 50 min	"
5%	1	1	"	"	"	1h	"
5%	1	2	"	"	"	58 min	"
5%	1	3	"	"	"	56 min	"
5%	1	4	"	"	"	55 min	"
10%	0	0	"	"	"	50 min	"
10%	0	1	"	"	"	1h	"
10%	1	0	"	"	"	1h 10 min	"
10%	1	1	"	"	"	2h 20 min	"
10%	1	4	"	"	"	1h 10 min	"
10%	1	10	"	"	"	50 min	"
10%	1	15	"	"	"	49 min	"
10%	2	1	"	"	"	4h	"
10%	2	2	"	"	"	3h 50min	"
10%	4	1	"	"	"	9 h	"
15%	0	0	"	"	"	1h 5min	"
15%	0	1	"	"	"	1h 45 min	"
15%	0	2	"	"	"	32 min	"
15%	0	4	"	"	"	23 min	"
15%	1	1	"	"	"	3h	"
15%	1	2	"	"	"	2h 30 min	"
15%	1	4	"	"	"	1h 15 min	"
15%	2	1	"	"	"	5h 15 min	"
15%	4	1	"	"	"	15h 25 min	"
15%	30	1	"	"	"	*break 68h*	"
20%	0	1	"	"	"	1h 15 min	"
20%	1	1	999	95+ Maschi	55-59aa M (1961)	3h 35min	High
1%	1	1	999	95+ Femmine	55-59aa F (1961)	35 min	High
1%	2	2	"	"	"	45 min	"
3%	1	1	"	"	"	1 h	"
3%	1	2	"	"	"	1h 10 min	"
3%	2	1	"	"	"	1h 20 min	"
5%	2	1	"	"	"	2 h	"
5%	1	1	"	"	"	1h 10 min	"
5%	1	5	"	"	"	55 min	"
10%	0	0	"	"	"	1h	"
10%	1	0	"	"	"	2h 50 min	"
10%	1	1	"	"	"	3 h	"
10%	1	5	"	"	"	1h 15 min	"
10%	2	5	"	"	"	3 h	"
15%	0	0	"	"	"	1h 20 min	"
15%	0	1	"	"	"	1 h	"
15%	0	2	"	"	"	39 min	"
15%	1	1	"	"	"	4h	"
15%	1	2	"	"	"	3h	"
15%	1	4	"	"	"	1h 35 min	"
15%	1	5	"	"	"	1h 10 min	"
15%	1	10	"	"	"	50 min	"
15%	2	1	"	"	"	5h 10 min	"
15%	4	1	"	"	"	22h	"
20%	0	1	"	"	"	1h 30min	"
20%	1	1	999	95+ Femmine	55-59aa F (1961)	4h 20 min	High

Bibliografia

AA.VV.,
Atlante della mortalità in Emilia-Romagna 1998-2004
Regione Emilia-Romagna, ASR, Bologna, (2007)

AA.VV.,
Presidi e servizi socio assistenziali e socio sanitari in Emilia-Romagna
Sistema Informativo per le Politiche Sociali (SIPS) - Regione ER, (2005)

Aarts E.,Korst J.,
Simulated annealing and Boltzmann machines,
Wiley Chichester, (1989)

Bailey T.C.,
Spatial statistical methods in health,
Cad Saude Publica, 17(5):1083-98, (2001)

Barlow R.E.,Bartholomew D.J.,Bremner J.M.,Brunk H.D.,
Statistical inference under order restriction,
New York Wiley

Besag J.,Newell J.,
The detection of clusters in rare disease,
Journal of the Royal Statistical Society, Series A,154:143-155, (1991)

Biggeri A.,Barbone F.,Lagazio C.,Bovenzi M.,Stanta G.,
Air pollution and lung cancer in Trieste: spatial analysis of risk as a function of distance from sources,
*Environmental Health Perspectiv*e, 104:750-754, (1996)

Bonferroni C.E.,
Il calcolo delle assicurazioni su gruppi di teste,
in *Studi in Onore del Professore Salvatore Ortu Carboni*,13–60, (1935)

Caselli G.,
Le migrazioni internazionali,
in Eredità del Novecento: Istituto dell'Enciclopedia Italiana, Roma,622-640, (2001)

Choynowsky M.,
Maps based on probabilities,
*Journal of the Royal Statistical Associa*tion, 54:385-388, (1959)

Diggle P.J,
Overview of statistical methods for disease mapping and its relationship to cluster detection, in *Spatial Epidemiology: Methods and Applications*,
Elliott P.,Wakefield J.,Best N.,Briggs D.J., Oxford University Press,87–103, (2000)

Doll R.,
The epidemiology of cancer,
Cancer 45:2475–2485, (1980)

Duczmal L.,Assunção R.,
A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters,
Computational Statistics and Data Analysis,45:269-286,(2004)

Duczmal L.,Patil G.P.,Tavares R.,Cançado A.L.F*.,*
Detection of spatial clusters in maps equipped with environmentally defined structures*,*
*Environmental and Ecological studi*es,(2006a) *(manuscript)*

Duczmal L.,Kulldorff M.,Huang L.,
Evaluation of spatial scan statistics for irregularly shaped disease clusters,
Journal of Computational and Graphical Statistics,15(2):428-442, (2006b)

Duczmal L.,Cançado A.L.F.,Takahashi R.H.C.,Bessegato L.F.,
A genetic algorithm for irregularly shaped spatial scan statistics,
Computational Statistics & Data Analysis, 52:43-52*,* (2007)

Dykstra R.,Kochar S.,Robertson T.,
Inference for likelihood ratio ordering in the two-sample problem*,*
Journal of The American Statistical Association, 90:1034-1040, (1991)

Elliot P.,Wakefield J.C.,Best N.G.,Briggs D.J.,
Spatial epidemiology: methods and applications,
in *Spatial Epidemiology,* 3-14*,*Oxford University Press, (2000)

Elliot P.,Wartenberg D.,
Spatial epidemiology: current approaches and future challenger,
Environmental Health Perspective, 112:998-1006, (2004)

Huang L.,Kulldorff M.,Gregorio D.,
A spatial scan statistic for survival data,
Biometrics,63(1):109-118, (2007)

Jung I.,Kulldorff M.,Klassen A.,
A spatial scan statistic for ordinal data,
Statistics in medicine,26(7),1594-1607, (2007)

Keys A.,
Seven Countries - A Multivariate Analysis of Death and Coronary Heart Disease,
Boston: Harvard University Press, (1980)

Knjazev D.,
OmeGA: a competent genetic algorithm for solving permutation and scheduling problems,
Kluwer Academic Publishers, Boston, Massachusetts, (2002)

Knox E.G.,
Detection of clusters,
in *Methodology of enquiries into disease clustering*,
(P.Elliot ed.), Small Area Health Statistics Unit, London, (1989)

Kulldorff M.,Nagarwalla N.,
Spatial Disease Clusters: detection and inference,
Statistics in medicine,14:799-810, (1995)

Kulldorff M.,
A spatial scan statistic,
Comm.Statist.Theory Methods, 26(6):1481-1496, (1997)

Kulldorff M.,
Spatial scan statistics models, calculations and applications,
in Glaz J.,Balakrishnan N., (Eds.)*,Scan Statistic and Applications*,
Birkhauser, Boston, 303-322, (1999)

Kulldorff M. and Information Management services, Inc.
SaTScan™ ver.7.0, Software for the spatial and space-time statistics,
http://www.satscan.org/, (2006)

Kulldorff M.,Huang L.,Pickle L.,Duczmal L.,
An elliptic spatial scan statistic,
Statistics in medicine, 25:3929-3943, (2006a)

Kulldorff M.,
Tests of spatial randomness adjusted for an inhomogeneity: a general framework,
Journal of American Statistical Association, 101: 1289-1305, (2006b)

Kulldorff M.,Mostashari F.,Duczmal L.,Yih K.,Kleinman K.,Platt R.,
Multivariate Scan Statistics for disease surveillance,
Statistics in medicine, 26(8):1824-1833,2007

Mignani S., Montanari A.,
Appunti di statistica multivariata,
Editore Esculapio, (1994)

Mitchell M.,
An introduction to genetic algorithms,
MIT Press, Cambridge, (1996)

Montanari A.,
Il metodo *Projection Pursuit,* uno strumento per l'analisi dei dati multidimensionali, CLUEB, (1999)

Naus J.I.,
The distribution of the size of maximum cluster of points on the line,
Journal of the American Statistical Association, 60, 532-538,1965
Clustering of random points in two dimensions, *Biometrika*, 52:263-267, (1965)

Neill D.B.,Moore A.W.,
Efficient scan statistics computations,
in Lawson A. and Kleinman K.
Spatial and Syndromic Surveillance for Public Health, (2005)

Openshaw S.,
The Modifiable Areal Unit Problem,
Norwich*,* UK*:Geo Books*, (1984)

Openshaw S.,Charlton M.,Wymer C.,Craft A.,
A mark 1 analysis machine for the automated analysis of point data sets",
*International Journal of Geographical Information Systems,*1:335-358, (1987);
Lancet, 1:272-273, (1988)

Patil G.P.,Taille C.,
Upper level set scan statistics for detecting arbitrarily shaped hotspots,
Environmental and Ecological Statistics, 11:183-197, (2004)

Robine J.M.,Caselli G.,
An unprecedented increase in the number of centenarians,
Genus LXI (1): 57-82, (2005)

Robine J.M.,Caselli G.,Rasulo D.,Cournil A.,
Differentials in the femininity ratio among centenarians: variations between northern and southern Italy from 1870,
Population Studies, Vol. 60,(1): 99-113, (2006)

Schneider P.J.,Eberly D.H.,
Geometric Tools for Computer Graphics,
Elsevier Science, (2003)

Upton G.,Fingleton B.,
Spatial Data Analysis by Example,
Volume 1: Point Patterns and Quantitative Data,
Editor Wiley, New York*,* (1985)

Takahashi K.,Yokoyama T.,Tango T.,
FleXScan:Software for the flexible spatial scan statistic
National Institute of Public Health,Japan, (2004)

Tango T.,Takahashi K.,
A flexibly shaped spatial scan statistic for detecting clusters,
International Journal of Health Geographics, 4-11, (2005)

Toutain S.,
Les systemes de retraites en Italie: une interminable reforme,
Paris: *L'Harmattan*, (2001)

Turnbull B.,Iwano E.J.,Burnett W.S.,Howe H.L.,Clark L.C.,
Monitoring for clusters of disease: application to leukemia incidence in upstate New York,
American Journal of Epidemiology,132:S136-S143, (1990)

Verkasalo P.J.,
Risk of cancer in Finnish children living close to powerlines,
British Medical Journal, 307:895-899, (1993)

Walter S.D.,
Disease mapping: a historical perspective,
in *Spatial Epidemiology: Methods and Applications*,
Elliott P.,Wakefield J., Best N., Briggs D.J,
Oxford University Press,223–252, (2000)

Whittemore A.S.,Friend N.,Brown B.W.,Holly E.A.,
A test to detect clusters of disease. Biometrika,
Biometrika, 74:631-635, (1987)

Williams G.,
Pleiotropy, natural selection and the evolution of senescence,
Evolution, 11(4),398-411, (1957)

Winkler G.,
Image Analysis, Random Fields and dynamic Monte Carlo methods,
Springer, New York, (1995)

Yiannakoulias N.,Rosychuk R.J.,Hodgson J.,
Adaptations for findings irregularly shaped disease clusters,
International Journal of Health Geographics, 6-28, (2007)