UNIVERSITA' DEGLI STUDI DI MILANO

Facoltà di Medicina e Chirurgia

Dipartimento di Scienze Cliniche e di Comunità

Dottorato di ricerca in **Statistica Biomedica**

XXVI ciclo

Tesi di dottorato di ricerca

Variazione spaziale del rischio relativo di morte per tumore ematologico in un'area prossima ad una raffineria petrolifera

Dottoranda: Francesca Di Salvo

Tutor: Prof. Adriano Decarli

Co-tutors:

Dott. Andrea Micheli

Dott. Paolo Baili

Dott.ssa Elisabetta Meneghini

Dott.ssa Giada Adelfio

Coordinatore del corso: Prof. Adriano Decarli

Anno Accademico 2012/2013

SOMMARIO

Introduzione e obiettivi

Con il termine epidemiologia spaziale o epidemiologia geografica indichiamo quell'insieme di attività mirate alla descrizione e all'analisi delle variazioni geografiche di eventi connessi alla salute in relazione a fattori di rischio demografici, ambientali, comportamentali, genetici e socio-economici (Elliot, 2000).

Una delle tante interessanti applicazioni dell'epidemiologia geografica è quella della descrizione della variazione del rischio per una malattia o una causa di morte, in una regione di particolare interesse per la presenza di un potenziale fattore di rischio. I risultati ottenuti da studi di questo tipo sono di pubblica utilità per fornire indicazioni eziologiche, dunque sulle possibili cause dell'insorgenza di malattie, legate a fattori di rischio presenti nella zona. Questo tipo di studi ha avuto un grande sviluppo negli ultimi tempi, trovando applicazione soprattutto nell'analisi degli effetti di inquinanti sulla salute umana.

L'obiettivo principale di questo lavoro è stato la descrizione e la rappresentazione grafica del rischio relativo di decesso per tumore del sistema emolinfopoietico, in un'area territoriale prossima ad una raffineria di prodotti petroliferi sita dal 1950 nel comune di Falconara Marittima in provincia di Ancona che, per la natura della produzione svolta, determina esposizione a sostanze irritanti, tossiche, nocive e indicate a potere cancerogeno, come il benzene. Secondo la IARC infatti, il benzene si conferma un agente cancerogeno per l'uomo del Gruppo 1 (IARC 2009).

Il secondo obiettivo è stato quello di valutare l'associazione tra l'eventuale rischio relativo di aumentato decesso e la distanza delle abitazioni principali dei soggetti in studio dalla raffineria petrolifera. A

seguito di uno studio di fattibilità (Baili, 2007) la regione Marche promosse l'effettuazione di un'indagine epidemiologica presso la popolazione residente nei dintorni della raffineria, affidata al Servizio di Epidemiologia Ambientale del Dipartimento di Ancona dell'Agenzia Regionale Protezione Ambientale Marche (ARPAM). La direzione scientifica venne affidata alla struttura complessa di Epidemiologia Descrittiva e Programmazione Sanitaria della Fondazione IRCCS "Istituto Nazionale dei Tumori", presso la quale lavoro ad oggi, ora denominata Epidemiologia Analitia e Impatto Sanitario.

Nella prima parte della tesi, capitolo 1 e 2, descrivo le caratteristiche dell'analisi spaziale e della point pattern analysis, nello specifico descrivo i metodi per un'analisi esplorativa dei dati spaziali, dall'individuazione di eventuali cluster ad un'analisi della variazione del rischio relativo nell'area di studio e la rappresentazione grafica su mappe di rischio. Nel capitolo 3 descrivo i modelli applicati ai dati di processo di punto, per la valutazione di una eventuale associazione tra l'aumento del rischio di morte e la distanza dell'abitazione principale dalla raffineria. Nel capitolo 4 descrivo il tipo di indagine epidemiologica che è stata effettuata a Falconara e dintorni, dal disegno dello studio ai metodi per la raccolta dei dati e mostro i risultati ottenuti dall'applicazione delle tecniche di analisi descritte ai dati.

1 L'analisi spaziale e dei processi di punto

1.1 Introduzione

La statistica spaziale si occupa della raccolta, della descrizione, visualizzazione ed analisi dei dati che posseggono delle coordinate geografiche. Infatti ciò che la distingue dalle altre branche della statistica è l'uso delle coordinate geografiche nella specificazione del modello statistico. Esistono svariati campi di applicazione della statistica spaziale, uno di questi è proprio lo studio del verificarsi di un evento sanitario (incidenza di una malattia, morte per una malattia, recidiva, ect..) in un ambito di popolazione, secondo la sua distribuzione spaziale. I "dati spaziali" possono essere pensati come realizzazione di un processo spaziale, ovvero generati da un processo stocastico dove lo spazio dell'indice è un insieme di dimensione pari a due.

1.2 Processi stocastici spaziali: cenni

Un processo stocastico spaziale è definito come una famiglia di variabili aleatorie Y indicizzate secondo *s*, dove lo spazio S dell'indice *s* è un sottoinsieme di R^2. (Bailey e Gatrell, 1995)

$$S \quad R^2$$

L'insieme S potrà essere finito o infinito, discreto o continuo. Così come le variabili del processo possono essere variabili continue o discrete.

Si dice che un processo spaziale è stazionario (o omogeneo) se le sue proprietà statistiche sono invarianti rispetto a traslazioni, cioè la sua distribuzione è inalterata quando l'origine dell'insieme degli indici viene traslato. Si possono distinguere stazionarietà del primo ordine, riguardante la media e la varianza del processo, da quelle del secondo ordine, riguardante la covarianza (o correlazione). La stazionarietà del

primo ordine implica che media e varianza delle variabili del processo sono costanti nel piano e quindi non dipendono da *s*. La stazionarietà del secondo ordine che la covarianza (e quindi la correlazione) tra due variabili del processo, dipende soltanto dalla loro posizione relativa nello spazio, cioè dalla loro distanza e dalla loro direzione. Se l'invarianza vale anche rispetto a rotazioni intorno all'origine (non ci sono quindi effetti direzionali nel piano), si parla di processo isotropico.

1.3 Tipologia di dati spaziali

I dati spaziali possono essere suddivisi in quattro diverse classi, distinte per la componente che viene considerata stocastica e che quindi deve essere osservata ed analizzata statisticamente mediante un modello.

I dati spaziali possono essere divisi nelle seguenti quattro macroaree:

- dati di processo di punto (point pattern o spatial pattern o spatial point pattern)
- dati di superficie aleatoria (geostatistical data o spatially continuous data)
- dati di area (lattice data o area data)
- dati di interazione spaziale (spatial interaction data)

Nel primo tipo, abbiamo a che fare con dei punti, individuati da coordinate geografiche, dislocati apparentemente in modo aleatorio su un piano (*point pattern*). Il principale interesse sta nel comprendere come questi punti (eventi) si distribuiscono nella ragione di studio. Ad esempio, potrebbe essere importante valutare se gli eventi si siano manifestati vicino ad un punto del piano dove è situata una “sorgente" potenzialmente inquinante, in questo caso l’evento potrebbe essere l’insorgenza di una malattia o un decesso per una malattia.

I dati che abbiamo analizzato per lo studio della variazione del rischio relativo di decessi per tumori del sistema emolinfopoietico, nei pressi della raffineria Api di Falconara Marittima, appartengono a questa categoria.

Nei dati di superificie aleatoria invece, ciò che è considerato aleatorio, è una superficie continua, osservata solo in alcuni punti fissi del piano. L'obiettivo in questo caso è quello di predire la superficie sulla parte di piano non osservata a partire dalle osservazioni. L'inferenza viene fatta condizionatamente alla posizione dei punti dove il processo è stato osservato.

Quando invece abbiamo a che fare con mappe di una regione, partizionata in aree a cui sono associati degli attributi (modalità di una variabile), parliamo di dati spaziali di area. La suddivisione geografica può corrispondere convenientemente ad una suddivisione amministrativa, in modo tale che le informazioni relative a ciascuna area siano di tipo aggregato. Un esempio di questa tipologia di dati riguarda l'analisi di immagini (per esempio immagini da satellite) in cui ogni pixel dell'imagine rappresenta un' area.

Nel caso di dati di interazione spaziale, si considerano dei punti di posizione fissa nel piano (come nel caso dei dati di superficie aleatoria), che possono anche essere i centroidi di aree. Ogni punto, considerato in coppia con un altro, viene considerato origine o destinazione di un flusso. Ciò che è aleatorio, e che quindi occorre osservare, sono proprio i flussi che si stabiliscono tra i due punti origine-destinazione. Esempi sono il flusso migratorio tra due stati, il flusso di telefonate tra due torri di ricezione di segnale, il flusso di passeggeri tra due aeroporti o stazioni ferroviarie. Un modello tradizionale usato in questo tipo di dati è il modello gravitazionale, che considera in un modello di regressione

l'introduzione di una covariata: la distanza fra l'origine e la destinazione, più altre covariate riferite all'origine e alla destinazione (Dreassi, 2008).

1.4 Processo di punto spaziale

Un processo di punto spaziale è un modello stocastico che determina la collocazione di un insieme di n eventi (spatial point pattern) in una certa regione R, definendone le coordinate spaziali $\{s_i\}$, con $s_i = (x_i , y_i) = (s_{i1,} s_{i2})$ le coordinate spaziali dell'evento i-esimo, con i= 1,…..,n.

In generale, un processo di punto è un processo stocastico in cui osserviamo le collocazioni di un certo numero di eventi all'interno di una regione di interesse (Diggle 2003). Da un punto di vista statistico, un processo spaziale di punto può essere pensato in termini di eventi verificatisi in una arbitraria area A contenuta nella regione di interesse R. Generalmente, si considera una misura di conteggio Y su A, Y (A) indica il numero di eventi all'interno della regione A, con A R, e |A| indica la superficie di A. Il processo è quindi una famiglia di variabili {Y (A) : A R}, con Y (A) che rappresenta il numero di eventi presentatisi nell'area A. (Dreassi, 2008)

Nell'epidemiologia spaziale, un problema comune è quello di determinare se i casi di una determinata malattia sono raggruppabili in clusters o meno e questo può essere verificato paragonando la distribuzione spaziale dei casi a quella di un set di controlli selezionati random dalla popolazione.

Il comportamento di un processo di punto spaziale generale può essere caratterizzato in termini di proprietà del primo e del secondo ordine. La distribuzione di eventi nello spazio, entra nella sfera delle *proprietà del primo ordine* e l'esistenza di possibili interazioni tra loro, misurate dalle

proprietà del secondo ordine, che studiano la tendenza degli eventi ad apparire clustered, indipendenti o regolari.

Più formalmente possiamo dire che le proprietà del primo ordine sono descritte in termini di *intensità λ(s)* del processo, che è la media del numero di eventi per unità di area al punto s (Cressie 1993). Questa è definita da:

$$\lambda(s) = \lim_{|ds|\to 0} \frac{\mathrm{E}[Y(ds)]}{|\ ds\ |}. \qquad (1)$$

Dove d**s** è una piccola regione introno al punto **s** e *Y(d**s**)* è il numero di eventi nella piccola regione d**s.** Mentre la funzione di intensità del secondo ordine è definita da:

$$\lambda_2(s_i, s_j) = \lim_{|ds_i|,|ds_j|\to 0} \frac{\mathrm{E}\left[Y(ds_i)Y(ds_j)\right]}{|\ ds_i\ ||\ ds_j\ |} \qquad (2)$$

Possiamo dire che un processo di punto è detto stazionario se l'intensità è costante in tutta la regione R, in maniera tale che $\lambda(s) = \lambda$ e in aggiunta $\lambda(s_i,s_j) = \lambda(s_i - s_j) = \lambda(d)$ implicando che l'intensità del secondo ordine dipende solamente dal vettore differenza *d* (direzione e distanza), tra s_i e s_j e non dalla loro assoluta localizzazione. Il processo è detto isotropico se tale dipendenza è funzione solo del vettore *d*. (Cressie, 1993)

1.5 Processo di Poisson

Supponiamo che la regione che stiamo considerando, ha area A e che all'interno di questa regione ci siano λA punti distribuiti in modo random. L'intensità di questi punti è pari a λA/A=λ per unità di area.

Immaginiamo di suddividere l'intera regione in un grande numero N di sub-regioni di ampiezza N/A. Un pattern di punti random implica che in ogni sub-regione ci dovrebbe essere la stessa probabilità di trovare un certo numero di punti, pari a λN/A. A questo punto ci chiediamo, cosa succede se N tende all'infinito? Siamo nella stessa condizione di derivazione della distribuzione di Poisson come limite della distribuzione binomiale, quindi la probabilità di osservare esattamente r punti in un'unità di area, quando l'intensità per unità di area è pari a λ, sarà data dalla seguente formula:

(3)

Questo spiega perché una distribuzione di punti random nello spazio può essere ricondotta ad un processo di Poisson (Upton and Fingleton 1985).

A questo punto una questione di immediato interesse: E' ragionevole aspettarsi che un pattern di dati reali possa distribuirsi in maniera random?

1.5.1 Processo omogeneo di Poisson

Un processo omogeneo di Poisson è detto tale quando tutti gli eventi sono indipendenti ed uniformemente distribuiti nella regione di interesse A in cui gli eventi sono occorsi. Ciò implica che la localizzazione di un evento non condiziona la probabilità degli altri punti di occorrere nelle vicinanze o non ci sono regioni in cui gli eventi hanno una maggiore probabilità di verificarsi.

Più formalmente:

in una regione A, un processo di Poisson è detto omogeneo se:

1. Il numero di eventi in A, con area |A|, segue una distribuzione di Poisson con media λ|A|, dove λ è l'intensità costante del processo di punto.
2. Dati n eventi osservati nella regione A, essi sono uniformemente distribuiti.

Un processo omogeneo di Poisson è anche stazionario ed isotropico. E' stazionario perché l'intensità è costante e l'intensità del secondo ordine dipende solo dalle direzioni relative di due punti (es: direzione e distanza). Inoltre, è isotropico perché l'intensità del secondo ordine è invariante rispetto alla rotazione. Cioè, il processo ha intensità costante e l'intensità del secondo ordine dipende solo dalla distanza tra due punti. (Bivand 2008)

1.5.2 Processo inomogeneo di Poisson

Quando ci troviamo a che fare con dati reali di popolazione, è quasi irrealistico che questi possano essere considerati come un processo omogeneo di Poisson. E' naturale che questi si distribuiscano per zone abitate, città, regioni etc. Un processo in omogeneo di Poisson è una generalizzazione di un processo omogeneo di Poisson, con intensità non costante. Rimane il principio di indipendenza tra gli eventi, ma è presente una variabilità spaziale dovuta ad eventi che hanno una maggiore probabilità di verificarsi in alcune aree piuttosto che in altre. (Bivand 2008)

1.6 Metodi basati sulle distanze

1.6.1 La funzione F e la funzione G

Il metodo della stima dell'intensità Kernel riguarda l'esplorazione nello spazio della funzione di intensità del primo ordine. Per l'intensità del secondo ordine, possiamo definire due funzioni:

3. la funzione G(w) con W la distanza “evento-evento" (nearest-neighbour)
4. la funzione F(x) con X la distanza “punto-evento" (point-nearest-event)

La distanza W rappresenta la distanza tra un evento ed il suo vicino più prossimo, la distanza X, la distanza tra un punto scelto a caso nel piano e l'evento più vicino che si è verificato.

Un modo per investigare il grado di dipendenza spaziale in un point pattern è quello di esaminare la distribuzione osservata di una di queste distanze, o meglio entrambe.

La stima delle distribuzioni cumulate empiriche sono date da

$$G(w) = \#\,(w_i \leq w)/n \qquad (4)$$

$$F(x) = \#\,(x_i \leq x)/m \qquad (5)$$

con n il numero di eventi nell'area e m il numero di punti del piano considerati.

Le funzioni di ripartizione empiriche possono essere plottate singolarmente oppure plottate una sull'altra contro particolari valori di *w* o *x*. Se gli eventi sono dislocati in modo casuale nel piano, le due funzioni di ripartizione empiriche mostrano lo stesso andamento. Nel caso contrario, in presenza di cluster, le distanze “punto-evento" tenderanno ad essere maggiori di quelle “evento-evento", e quindi *G(w)* starà al di sopra di *F(x),* crescendo più rapidamente. Nel caso di struttura spaziale regolare, la funzione di ripartizione empirica *G(w)* invece crescerà meno rapidamente e sarà sovrastata dalla funzione empirica *F(x).* (Baddeley 2010)

1.6.2 La funzione K di Ripley

Le funzioni *F(x)* ed *G(w*) hanno il difetto di considerare solo le distanze da eventi vicini, quindi danno informazioni solo sull'andamento di piccola scala del fenomeno (effetto locale), l'informazione su grande scala viene ignorata (effetto globale). Affinché si possa parlare di grande scala occorre ipotizzare l'isotropia del processo.

Una caratterizzazione alternativa delle proprietà del secondo ordine di un processo stazionario e isotropico, quindi dell'intensità del secondo ordine, è data dalla funzione K di Ripley (Ripley 1977).

La funzione K di Ripley è definita come:

λK(h)=E(numero di eventi entro una distanza h da un evento arbitrario)

dove *λ* rappresenta l'intensità o il numero medio di eventi per unità.

Una volta determinata *K(h),* questa può essere rappresentata graficamente ed esaminata per avere un' idea della dipendenza spaziale nel processo di punto per diversi valori di *h*.

Nel caso di processi che non presentano né regolarità né cluster, quindi un processo di punto dove la dislocazione degli eventi è casuale, ci si aspetta che $K(h) = \pi h^2$, (πh^2 rappresenta l'area del cerchio di raggio *h*).

Nel caso di regolarità $K(h) < \pi h^2$, nel caso di cluster $K(h) > \pi h^2$.

I picchi su valori positivi indicano un'attrazione spaziale di eventi, o cluster, quelli negativi indicano una regolarità (Fig. 1)

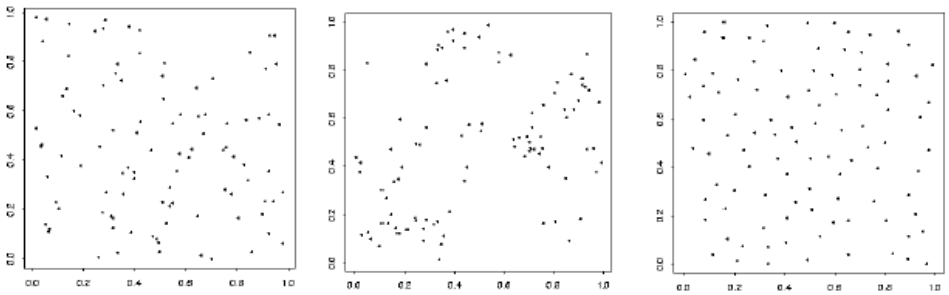

Figura 1 Spatial point pattern: casuale (a sinistra), a cluster (al centro) e regolare (a destra)

Come detto, l'importanza della funzione *K(h)* sta nel fatto di poter fornire informazioni su andamenti su grande scala. Inoltre, la forma teorica della funzione *K(h)* è nota per vari processi di punto, quindi oltre che in fase esplorativa, essa può essere usata per specificare dei modelli statistici.

In un contesto reale, non succede praticamente mai che la K di Ripley dia come risultato un point pattern regolare, anche perché i dati di popolazione sono per natura disomogenei. Un modello che tiene conto di questa disomogeneità è stato ipotizzato da Cuzick ed Edwards. Quando ci troviamo in un contesto di caso-controllo, questi studiosi hanno pensato di stimare una K di Ripley separatamente per i casi e per i controlli ipotizzando un modello basato sulla differenza tra le due funzioni K(.), che tiene quindi conto della naturale disomogeneità della popolazione. In questo modo la funzione K di Ripley può essere utilizzata per vedere se esiste una extra aggregazione di casi rispetto ai controlli. (Cuzick and Edwards, 1990)

2 *Variazione spaziale del rischio relativo in studi caso-controllo*

2.1 Introduzione

Nei paragrafi precedenti abbiamo visto che l'analisi dei processi di punto si occupa di studiare in che modo i punti (gli eventi) sono dislocati nello spazio. Il primo passo da fare per descrivere la dislocazione di questi eventi è quello di calcolarne la densità. Un approccio molto semplice ed intuitivo dal punto di vista metodologico per fare ciò è quello di suddividere la regione in un certo numero di quadrati, più o meno grandi e contare il numero di eventi all'interno di ogni quadrato. Dal punto di vista grafico, una rappresentazione dell'intensità si può ottenere tramite un istogramma bidimensionale, in cui vengono riportate in ordinata le intensità ottenute mediante il rapporto tra il numero di eventi, osservati per ciascun quadrato della griglia, e l'area di questo quadrato. L'istogramma dà un'idea del variare dell'intensità del processo $\lambda(s)$ su R.

Il problema di questo approccio metodologico è che, un istogramma rappresenta sempre una funzione discontinua, avremo quindi una rappresentazione grafica della densità sempre "a scalino". Inoltre, per quanto piccoli possano essere scelti i quadrati della griglia, quando abbiamo a che fare con dati reali di popolazione, è inevitabile che ci siano delle aree a scarsa densità di popolazione. Ciò significa che si ottengono dei quadrati in cui la conta degli eventi è pari a zero, rendendo difficoltoso e di difficile interpretazione il calcolo di rapporti, tassi, indici.

2.2 Lo stimatore di densità Kernel "smoothed"

Per superare questo problema, si è pensato di costruire una funzione di densità generica K( ), da sostituire alla funzione di densità uniforme.

Immaginiamo che la conta degli eventi, all'interno della regione di interesse, avvenga all'interno di una finestra circolare che si muove nella regione in modo continuo. Definito il raggio di questa finestra, che viene mossa su una griglia fine nello spazio, l'intensità, per ogni punto del piano, è stimata attraverso gli eventi che cadono all'interno della finestra circolare, in quel punto. In questo modo si ottiene una stima "smoothed" (lisciata) dell'intensità λ (s) su R. La stima, in questo caso, viene calcolata attraverso lo stimatore Kernel bivariato.

Con s un punto generico della regione R e $s_1,\ldots,s_n$ le coordinate di n eventi osservati, l'intensità λ (s) è definita come segue:

(6)

Dove K() è una densità di probabilità bivariata, simmetrica rispetto all'origine, nota come Kernel e determina la forma delle curve. Il parametro τ >0 rappresenta l'ampiezza di banda (bandwidth), è il raggio della finestra circolare centrata in s entro cui i punti si distribuiscono e determina il grado di "lisciamento" (smoothing) della mappa stimata. $K((s - s_i)/\tau)$ è il contributo che ciascuna osservazione dà alla determinazione della densità λ(s).

La scelta della forma funzionale *k*( ) spesso non è fondamentale, si è visto che per forme diverse della funzione, si ottengono risultati molto vicini (Fig. 2). Il kernel può essere quadratico, gaussiano, triangolare, rettangolare.

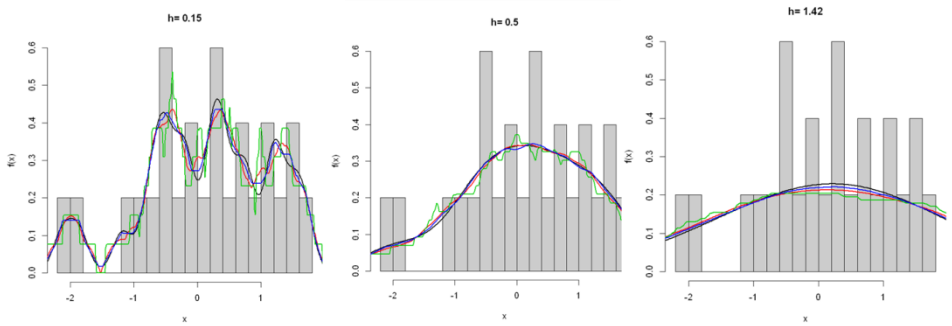

Figura 2 Densità con 4 kernel differenti e con tre diversi valori del bandwith, n=20

Ovviamente pesano di più le osservazioni s_i più vicine a s: ampiezze della banda τ più grandi attutiscono il peso delle osservazioni vicine e diffondono l'influenza di ciascuna s_i osservata su un range più ampio. Valori di τ più piccoli concentrano i pesi nelle immediate vicinanze di ciascun si e quindi forniscono in generale stime di densità meno "lisce" al variare di x. L'influenza di ciascuna si osservata si limita ad un range ridotto. Per avere un esempio numerico, se K( ) è un kernel normale standardizzato, fuori dall'intervallo (s_i - 4τ, s_i + 4τ) la densità è praticamente nulla, il che significa che le osservazioni distanti da s più di 4 volte l'ampiezza di banda τ non influenzeranno la stima di λ(s). (Fig.3)

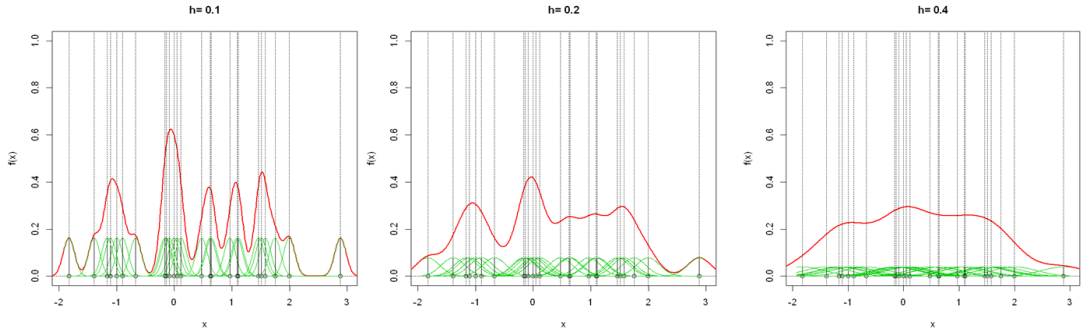

Figura 3 Stima della densità con Kernel normale per diversi valori del bandwidth h. n=20

La determinazione del bandwidth è oggetto di discussione da anni. In R, il package spatstat utilizza la funzione "density" per il calcolo della funzione di intensità Kernel Smoothed (Baddeley 2010) e il valore che specifichiamo sigma, deviazione standard del Kernel Gaussiano isotropico, corrisponde anche ad una funzione che computa un appropriato bandwidth per i dati point pattern.

2.3 Lo stimatore Kernel per il rischio relativo

In questo paragrafo entriamo nel cuore di quello che è stato l'obiettivo del mio lavoro, l'analisi del rischio relativo. In uno studio di tipo caso-controllo, la funzione del rischio relativo, ottenuta dal rapporto tra la densità dei casi e quella dei controlli (Bithell 1990,1991) è un utile strumento per descrivere la variazione spaziale del rischio di malattia.

Nei paragrafi precedenti abbiamo specificato l'utilizzo dello stimatore smoothing Kernel per l'intensità, che costituisce un valido approccio per modellare una distribuzione spaziale altamente eterogenea.

Consideriamo due set di punti x_i, per i=1,…,n_1, come una realizzazione parziale di un processo di Poisson con intensità λ_1 (x) e y_j, per j=1,…,n_2, come una realizzazione indipendente di un processo di Poisson con intensità λ_2(x). L'obiettivo è quello di ottenere una stima non parametrica del rapporto λ_1(x)/ λ_2(x), per studiare la variazione spaziale del rischio relativo all'interno di una regione di interesse. Nello specifico, λ_1(x) e λ_2(x), sono le intensità di casi e controlli, se intendiamo studiare l'eccesso di rischio per una determinata malattia rispetto alla popolazione generale. E' importante quindi trattare i due set di punti simmetricamente.

Stimiamo quindi la funzione:

$$\rho(x) = \log (\lambda_1(x)/ \lambda_2(x)) \qquad (7)$$

che ha la simmetria richiesta. (Kelsall and Diggle 1995).

In letteratura ci sono tanti esempi in cui è stato utilizzato questo approccio, dallo studio del pattern spaziale di motor neurone disease in Finlandia (Sabel, 2000), alla stima del rischio geografico di cirrosi biliare primaria in una regione del nord-est dell'Inghilterra (Prince, 2001), a quella del rischio di leucemia infantile che è stata individuata da Wheeler nello stato americano dell'Ohio (Wheeler, 2007).

Negli ultimi due esempi (e in effetti in quasi tutte le applicazioni ad oggi), è stato preso in considerazione solo il metodo del kernel-fixed bandwidth, cioè con una larghezza di banda fissa, per la determinatione della quantità di smoothing. La popolazione umana tende ad essere altamente disomogenea con una distribuzione geografica naturale che tenda ad aggregarsi in città, comuni, anche a seconda delle caratteristiche geografiche del territorio (fiumi, montagne) che influenzano la densità di casi e controlli, nell'intera regione di studio. Quando viene utilizzato un bandwidth di ampiezza costante (fixed-bandwidth) lo stimatore di densità kernel si comporterà alla stessa maniera sia nelle zone più popolate che in quelle meno, quindi se si sceglie uno smoothing grande, allo scopo di raccogliere più informazione possibile nelle aree meno popolate, potrebbero però perdersi i dettagli più sottili, nelle aree a più alta densità di popolazione. Viceversa, se scegliamo un bandwidth piccolo, potremmo avere un'informazione maggiore nelle aree ad alta densità di popolazione, ma trovare dei picchi di rischio in aree scarsamente popolate, dovute magari all'effetto di pochi singoli casi sporadici, che ci condurrebbero ad un errore di valutazione.

Un approccio più intuitivo è quello di attuare quindi uno smoothing variabile in cui la quantità di smoothing è inversamente proporzionale alla densità di popolazione. È ormai noto che è preferibile utilizzare tale approccio, chiamato adaptive smoothing, quando ci si trova in un contesto di disomogeneità, rispetto ad un fixed smoothing.

L'utilizzo di un parametro fisso (*fixed*) di smoothing (cioè un parametro costante per tutte le osservazioni) è certamente più semplice da implementare ma può risultare poco preciso quando ci troviamo di fronte ad una distribuzione della popolazione altamente eterogenea. In questi casi, risulta più opportuno utilizzare un parametro variabile (*adaptive*) seconodo il metodo di Abramson (Abramson 1982, Hall and Marron 1988, Davies and Hazelton 2010)

Davies e Hazelton hanno implementato il metodo per calcolare la funzione di rischio relativo con parametro *fixed* o *aptive* in R con il package "sparr" (Davies 2011).

2.4 Metodo adaptive: stima del rischio relativo come rapporto di densità kernel

L'idea di base di questo approccio è quella di non fare riferimento ad un'unica ampiezza di banda *h*, ma di usare per ogni osservazione campionaria x_i un valore $h(x_i)$, i = 1,…,n. In questo modo, ad ogni x_i si può associare una differente ampiezza di banda. Questa osservazione porta allo stimatore kernel a banda adattiva o variabile, che, considerate n osservazioni bivariate, indipendenti ed identicamente distribuite, $x_1,...,x_n$ ha la seguente forma funzionale:

$$\hat{f}(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(\mathbf{x}_i)^2} K\left(\frac{\mathbf{z}-\mathbf{x}_i}{h(\mathbf{x}_i)}\right) \qquad (8)$$

Dove K è la funzione Kernel, z è un punto della regione R e h_i è il parametro smoothing o bandwidth per l'i-esima osservazione. (Hazelton 2010)

Naturalmente, il problema è quello di scegliere la forma che deve assumere la funzione $h(x_i)$ nella (8). Intutivamente, come già accennato, vale una considerazione di questo tipo:

$h(x_i)$ dovrebbe essere piccola là dove c'è un'elevata densità di popolazione, ossia dove funzione di densità $f(z)$ assume valori grandi, mentre dovrebbe essere grande là dove i dati sono sparsi, ossia dove la $f(z)$ assume valori piccoli. Pertanto, l'intuizione suggerisce che $h(x_i)$ debba comportarsi, in maniera inversa rispetto a $f(z)$.

Si dimostra (Abramson, 1982) che un modo particolarmente valido di scegliere $h(x_i)$ consiste nell'esprimerla come segue:

$$h_i = h_0 \, f(X_i)^{-1/2} \, \gamma^{-1} \qquad (9)$$

dove h_0 è un moltiplicatore secondario smoothing che chiamiamo *global bandwith* e γ corrisponde alla media geometrica del termine $f(X_i)^{-1/2}$ in modo da limitare la dipendenza del parametro h_i dai dati. Da questa espressione notiamo infatti che il termine smoothing dipende in maniera inversa rispetto ai dati.

In campo applicativo, l'identificazione della banda *adaptive* si basa su una procedura a due stadi, che considera inizialmente una stima pilota della distribuzione ottenuta mediante la banda fissa (*pilot density*). La *pilot density* f_h è una stima della densità Kernel ottenuta con *fixed* bandwith costruita con il *pilot bandwith h*. A partire da questa stima iniziale, si

calcolano i cosiddetti “fattori di aggiustamento” che incorporano la variabilità dei dati all’interno della distribuzione stessa.

Infine sostituiamo la densità sconosciuta dell’espressione precedente con la *pilot density*.

Supponiamo di avere due set di osservazioni bivariate , $X_1;X_2; : : : ;X_{n1}$ e $Y_1;Y_2; : : : ;Y_{n2}$ che rappresentano rispettivamente le coordinate cartesiane degli eventi (casi) e dei controlli. La funzione di rischio relativo può essere espressa come il rapporto tra la densità dei casi e la densità dei controlli, che indichiamo rispettivamente con f( ) e g( ). (Bithell, 1990, 1991)

La funzione di rischio relativo r(z) è quindi definita da:

$$r(z) = f(z)/g(z) \qquad (10)$$

Nel caso nostro, per la stima della funzione di rischio relativo, abbiamo ritenuto opportuno utilizzare la metodologia bandwith *adaptive,* come spiegato nel paagrafo 4.5.

2.4.1 Il problema della scelta del bandwidth

Il problema di selezionare un “optimal bandwidth” per la stima della densità Kernel è stato largamente discusso in letteratura (Danese 2008), mentre il problema del bandwidth più appropriato nella stima del rischio relativo è un argomento meno trattato. Quello che pare essere noto è che, in una stima del rischio relativo con bandwidth fisso, tentare un approccio volto a stimare separatamente un bandwidth ottimo per i casi e uno per i controlli, con opportune tecniche computazionali di “optimal

bandwidth", non porta ad una buona stima del rischio relativo (Kelsall and Diggle, 1995) (Hazelton 2007).

Nello specifico si è visto che porterebbe ad un "undersmoothing" dato dal fatto che solitamente quando andiamo a stimare una densità di probabilità kernel, solo per i casi o per controlli, tendiamo ad utilizzare un bandwidth più piccolo per non rischiare di perdere i "dettagli" ed ottenere quindi una densità eccessivamente "piatta". Nella stima del rischio relativo ρ invece lo stesso bandwidth utilizzato per la stima della densità di casi e controlli porta a problemi computazionali, soprattutto in aree in cui la densità di casi o controlli è prossima allo zero, dovremo quindi utilizzarne uno più grande. (Davies 2009).

Quando ci troviamo in un contesto di stima di *fixed-kernel density*, la letteratura pare essere concorde sul fatto che è meglio utilizzare lo stesso valore di bandwidth per i casi e per i controlli, anche se i metodi di stima automatica sviluppati sono veramente pochi, (Kelsall and Diggle, 1995) (Hazelton 2007). Quando siamo in un contesto di *adaptive-kernel density*, come ampiamente descritto nel paragrafo precedente, dobbiamo scegliere un *pilot* e un *global* bandwidth sia per i casi che per i controlli. Possiamo decidere di utilizzare anche un *global* bandwidth comune e due *pilot* bandwidth separati, come suggerito dagli autori Davies e Hazelton. Questo consentirebbe di non perdere eventuali piccoli cluster che si verificano nei casi, senza per forza dover utilizzare un bandwidth piccolo anche nei controlli.

Un metodo per la scelta dei *pilot* bandwidth, separati per casi e controlli, è quello sviluppato da Bowman e Azzalini, basato sulla tecnica del Least-square Cross-Validation (Bowman 1997). Un altro metodo è quello sviluppato da Terrel (Terrel 1990) basato sul principio del maximal

smoothing. Secondo questo ultimo approccio, il *global* bandwidth h_0 della (9) è pari a:

h= U x OS (11)

dove U è un fattore di scala e OS è un fattore OverSmoothing dato da

$$\mathrm{OS}=\left[\frac{(d+8)^{(d+6)/2}\pi^{d/2}R(K)}{16n^{*}\Gamma[(d+8)/2]d(d+2)}\right]^{2/(d+4)} \quad (12)$$

Dove d è la dimensione dei dati (quindi 2 se ci troviamo in uno studio caso-controllo) e n^* è la media geometrica di n_1 e n_2.

Il razionale di questa specificazione dell'effettivo sample size sta nel fatto che il bandwidth comune per la stima dei casi e dei controlli dovrebbe essere guidato, in prima misura, dalla richiesta di smoothing del dataset più piccolo, solitamente quello dei casi. In pratica il fattore di scala U è pari a

U= IQR/1.34, dove IQR è la media degli interquartile ranges delle coordinate x e y delle nostre osservazioni. (Silverman 1986)

2.4.2 Il problema del Boundary bias

Nella stima della densità di casi e controlli, con lo stimatore kernel smoothing discusso, è fondamentale la regione di studio R, che nella realtà può essere determinata da un comune, una regione, un insieme di comuni, distretti etc, nei quali si osserva la dislocazione dei punti (il verificarsi di eventi).

Il problema del boundary bias nasce quando alcuni eventi si verificano in una zona prossima al confine della regione di studio. Richiamando infatti la formula (6) per la stima della densità kernel, avremo che una porzione del contributo di quell'evento verificatosi in prossimità del confine, cade

fuori dalla regione di interesse. Se non considerato, ciò porta ad un bias della densità e di conseguenza ad un bias del rischio relativo, evidenziando magari dei picchi "ad alto rischio" lungo il confine.

Per correggere questo bias sono stati implementati alcuni metodi, si veda Kelsall e Diggle (1995) per quanto riguarda il fixed smoothing, mentre in un contesto di adaptive smoothing si veda Marshall e Hazelton (2010). Questi ultimi hanno sviluppato un metodo basato su un kernel pesato, che quantifica la porzione di contributo entro i limiti della regione R. Cioè la densità stimata $\lambda(x)$ viene corretta, dividendo ad ogni locazione z, per la seguente quantità:

$$q_{h(\mathbf{z})}(\mathbf{z}) = \int_{\mathscr{R}} \frac{1}{h(\mathbf{z})^2} K\left(\frac{\mathbf{x}-\mathbf{z}}{h(\mathbf{z})}\right) \mathrm{d}\mathbf{x}. \tag{13}$$

Dove *h*(z) è il valore del bandwidth nel punto z. In questo modo si corregge per l'effetto-confne.

2.4.3 Livelli di significatività e tolerance contours

Quando rappresentiamo la variazione spaziale del rischio relativo su una mappa, ci chiediamo se un eventuale picco di incremento di rischio visualizzato dall'intensità del colore, è statisticamente significativo o se è semplicemente il risultato di una variazione random.

Un modo per vedere una significatività sulla mappa è quello di creare dei tolerance contours, basati su valori di p-value calcolati sulla superficie stimata del rischio relativo ρ, secondo l'esempio di Kelsall e Diggle (2009). L'idea è quella di costruire il seguente test di ipotesi:

H0: ρ(z) = 0

H1: ρ(z) > 0

Per ogni punto z appartenente alla regione di studio R. Da questo test otteniamo un insieme di p-values, da cui possiamo computare dei contours che corrispondono ai livelli di significatività. Quindi aree della regione R con valori del p-value inferiori o uguali a 0.05 saranno identificate come significativamente ad alto rischio.

Le critiche a questa metodologia derivano in gran parte dal tipo di test d'ipotesi utilizzato per generare i p-values. Quando ci troviamo a stimare un rischio relativo con un fixed bandwidth, un approccio metodologico molto utilizzato è quello basato sulle permutazioni di Montecarlo (Kelsall & Diggle 1995). Questa tecnica è però molto dispendiosa dal punto di vista computazionale e può portare ad errori nei limiti di significatività soprattutto nelle aree a scarsa densità di popolazione. Un metodo alternativo, più recente, è quello di costruire una statistica z-test basata sulle proprietà asintotiche dello stimatore di densità kernel. Questo metodo, tiene conto del bias relativo all'effetto confine (boundary bias), può anche essere esteso al caso di adaptive bandwidth e i tolerance contours appaiono più stabili nelle aree a scarsa densità di popolazione. (Hazelton and Davies 2010)

3 Modelli per la stima dell'effetto "source pollution"

3.1 Introduzione

Con le tecniche di analisi discusse nei capitoli precedenti, abbiamo visto come individuare delle zone ad alto rischio all'interno dell'area in cui si è deciso di condurre l'indagine ambientale. Vedere quindi in che modo si sono verificati gli eventi (i casi) nello spazio e se si possono individuare dei cluster in una zona piuttosto che in un'altra (proprietà del secondo ordine). A questo punto il passo successivo è quello di studiare quasi sono i fattori che possono aver contribuito a determinare quell'aumento di rischio.

3.2 Il modello di Diggle e Rowlingson basato sulla funzione di decrescita della distanza

In questo paragrafo viene illustrata la metodologia, basata sui processi spaziali di punto, utilizzata per stimare, se esiste, un eccesso di rischio di morte in un'area prossima ad una fonte di inquinamento ambientale. L'obiettivo è quello di utilizzare un modello che tenga conto della distanza di ogni punto (evento morte) rispetto alla sorgente di inquinamento e di eventuali altre variabili confondenti, noti fattori di rischio di malattia.

Ipotizziamo di voler descrivere l'aumento del rischio di mortalità per malattia in una regione geografica R, all'interno della quale sia posizionata una fonte di inquinamento. Dato un insieme di osservazioni bivariate $x=x_1,\ldots,x_n$, si può assumere che la posizione dei casi di morte segua un processo di Poisson non omogeneo con funzione di intensità $\lambda_1(x)$ data da

$$\lambda_1(x) = \rho\, \lambda_0(x)\, f\{d(x);\boldsymbol{\theta}\} \qquad (14)$$

La funzione di intensità del processo è dunque il prodotto di due componenti: la prima $\lambda_0(x)$, che rappresenta la variabilità spaziale dovuta alla non omogenea distribuzione della popolazione a rischio e la seconda $f\{d(x);\boldsymbol{\theta}\}$, che esprime la relazione tra il rischio di malattia e la sorgente di inquinamento ambientale, ρ rappresenta un fattore di scala.

Per $f\{d(x);\boldsymbol{\theta}\}$, secondo Diggle e Rowlingson (1991), si esplicita un opportuno modello parametrico che assume come variabile di esposizione il quadrato della distanza. Il termine esposizione si riferisce a tutte le variabili esplicative che si può ritenere abbiano una potenziale associazione con lo "stato di malattia". Osserviamo infatti, che nella componente parametrica una misura dell'esposizione è ottenuta tramite la distanza euclidea, che fornisce un'adeguata rappresentazione della modalità di diffusione sul territorio dell'agente inquinante.

Nella formalizzazione dell'intensità $\lambda_1(x)$, la funzione $f(d(x);\boldsymbol{\theta})$ è esplicitata nel modo seguente:

$$f(d(x);\boldsymbol{\theta}) = 1+ \alpha \exp(-\beta d^2)\ . \qquad (15)$$

in cui α rappresenta l'aumento del rischio nel punto in cui è posizionata la fonte di inquinamento, mentre β misura la scala spaziale degli effetti: al crescere del parametro β il rischio tende a concentrarsi in prossimità della sorgente di inquinamento ambientale. (Diggle, 1991)

L'analisi della funzione di intensità del processo permette di esaminare la distribuzione dei casi di morte sul territorio e di saggiare l'ipotesi nulla di assenza dell'effetto del punto sorgente, cioè consente di verificare che la distribuzione dei casi sia proporzionale alla densità della popolazione:

$H_0 : \boldsymbol{\theta} = (\alpha, \beta) = 0$

Intuitivamente l'idea sottostante è la seguente: se non c'è relazione tra esposizione alla sorgente e mortalità, allora la distribuzione dei casi dovrebbe essere la stessa della distribuzione dei controlli. La scelta di stimare in modo non parametrico $\lambda_0(x)$ si riflette sulla stima di massima verosimiglianza dei parametri del modello (α, β). (Dreassi, 2008)

Per eliminare l'effetto attribuibile alla stima non parametrica di $\lambda_0(x)$, ossia della componente relativa alla distribuzione della popolazione, Diggle e Rowlingson hanno proposto un approccio che converte il precedente modello, basato sulla formalizzazione dell'intensità di un processo di Poisson non omogeneo, in una regressione binaria. Secondo tale approccio, le localizzazioni dei casi e dei controlli possono essere viste come un insieme di etichette, il cui valore binario è determinato dalla probabilità dipendente dalla localizzazione.

Se si assume che casi e controlli siano le realizzazioni di due processi di Poisson, con intensità rispettivamente pari a $\lambda_1(x)$ e $\lambda_0(x)$, e che la loro sovrapposizione sia ancora un processo di Poisson con intensità $\lambda_0(x) + \lambda_0(x)$, condizionandosi sulle posizioni dei casi e dei controlli, si ricava che le $(n + m)$ etichette, dove n denotano i casi ed m i controlli, sono un insieme di v.c bernoulliane Yi , tali che

Yi=1 se il punto è un caso

$Yi = 0$ se il punto è un controllo

La probabilità che un individuo che risiede nel punto xi sia un caso è

$$P(Y_i = 1 \mid x_i) = p(x_i) = \frac{\lambda_1(x_i)}{\lambda_0(x_i) + \lambda_1(x_i)}. \qquad (16)$$

Si procede stimando $p(x)/[1-p(x)]$ che esprime l'odds, ovvero la probabilità che una persona che si trovi nella localizzazione x sia un caso piuttosto che un controllo; ossia dal rapporto tra la funzione di intensità dei casi e quella dei controlli $\lambda_1(x)/\lambda_0(x)$.

Tornando alla parametrizzazione del rischio legato al punto sorgente si ricava che:

$$p(x) = \frac{\rho\, f(d(x);\boldsymbol{\theta})}{1+\rho\, f(d(x);\boldsymbol{\theta})} \tag{17}$$

$$\log it\,(p(x)) = \log\left(\frac{\lambda_1(x)}{\lambda_0(x)}\right) = \log\rho + \log\left(f(d(x);\boldsymbol{\theta})\right) \tag{18}$$

I termini che figurano nel modello di regressione binaria a cui si è giunti misurano un eccesso di rischio relativo. Questo nuovo approccio evita di dover stimare $\lambda_0(x)$. (Diggle e Rowlingson 1994)

3.3 Modelli con covariate spaziali per processi di punto

L'approccio di Baddeley (Baddeley 2000) a problemi di questo tipo, parte dallo stesso presupposto di Diggle e Rowlingson, cioè che i modelli da adattare ai dati di punto, devono tenere conto della non omogeneità spaziale (trend), data dalla distribuzione dei controlli e da covariate. Baddeley aggiunge anche la dipendenza tra punto (interazione).

I modelli ipotizzati da Baddeley sono modelli di Gibbs, implementati nel package "spatstat" di R. Essi partono da una estensione delle tecniche di analisi dei processi di punto di Berman e Turner (Berman e Turner 1992), in cui si ottiene una approssimazione della pseudoverosimiglianza (Besag

1975) piuttosto che della verosimiglianza, in un processo di punto inomogeneo di Poisson.

Lo stimatore di massima pseudoverosimiglianza è una valida alternativa allo stimatore di massima verosimiglianza, soddisfa le proprietà di non distorsione, consistenza e sotto alcune condizioni è asintoticamente normale. In questa maniera è possibile adattare dei modelli che tengano conto del trend spaziale e delle covariate spaziali e della interazione tra punti (Baddeley 2000)

3.3.1 Formulazione del modello

L'intensità condizionale è una funzione $\lambda(z, x)$ della localizzazione spaziale z e dell'intero point pattern x. In pratica, se consideriamo una regione infinitesimale intorno al punto z dell'area dz, la probabilità condizionale che il processo di punto contenga un punto in quella regione infinitesimale, data la localizzazione di tutti i punti fuori dalla regione è $\lambda(z, x)du$. (Baddeley 2000).

La forma più semplice di un modello in cui utilizziamo l'intensità condizionale è una costante, $\lambda(z, x) = \beta$, che corrisponde ad un omogeneo processo di Poisson ed in molte applicazione costituisce il modello nullo.

Quando l'intensità condizionale $\lambda(z, x)$ dipende solo dalla locazione z, cioè $\lambda(z, x) = \beta(z)$, ci troviamo nel caso di un processo inomogeneo di Poisson, con funzione di intensità $\beta(z)$. In questo caso la forma funzionale di $\beta(z)$ indica il tipo di non omogeneità, ossia il trend spaziale.

La metodologia di Baddeley (Baddeley 2006), prevede di ipotizzare un modello con intensità condizionale nella forma seguente:

$$\lambda(z, x) = \exp(\phi^T B(z) + \upsilon^T C(z, x)) \qquad (19)$$

dove ϕ e υ sono i parametri da stimare.

Il termine B(z) dipende solo dalla localizzazione dei punti nello spazio, quindi rappresenta il trend spaziale o l'effetto delle covariate spaziali. Il termine C(z) rappresenta l'interazione stocastica o la dipendenza tra punti del processo.

Nello specifico problema che intendiamo affrontare, non consideriamo la parte dell'interazione, ma ci fermiamo alla componente B(z) che rappresenta trend e covariate spaziali, rimanendo quindi nell'ambito dei processi di Poisson.

Specifico quindi due modelli con covariate spaziali, definite come immagini pixel.

Modello 1: Un primo modello in cui esprimo il pattern di localizzazioni, ossia le residenze (abitazioni) dei casi di morte per tumore del sistema emolinfopoietico, come un processo di punto con intensità $\lambda_1(z)$ proporzionale alla densità della popolazione a rischio, quindi alla densità dei controlli $\lambda_2(z)$

Stimo dunque la densità dei controlli, utilizzando lo stimatore di intensità Kernel-smoothed, ottenendo una immagine pixels e fitto il modello seguente con log-intensità pari a

$$\log \lambda_1(z) = \alpha + \log \lambda_2(z) \qquad (20)$$

dove α è il parametro sconosciuto, $\lambda_1(z) = \mu \lambda_2(z)$

e

$\mu = e^{\alpha}$ è il solo parametro da stimare.

Il primo modello contiene quindi una covariata spaziale, la densità dei controlli $\lambda_1(z)$ che serve per modellare la distribuzione dei casi.

Modello 2: Il secondo modello tiene conto dell'effetto "source pollution", in cui aggiungo un'altra covariata: la distanza di ogni locazione dalla raffineria petrolifera, che è stata trasformata in una immagine pixel.

Il modello 2 ha la stessa forma funzionale della (14) di Diggle e Rowlingson, solo che la funzione distanza della (15) che descrive una proxi dell'esposizione, modellata dal parametro β, è data dalla distanza euclidea.

$$\lambda(u) = e^{\alpha + \beta x} \rho(u) \qquad (21)$$

e i parametri sconosciuti sono α e β.

Infine confronto i due modelli con un test rapporto delle verosimiglianze.

4 *Il caso di studio: la raffineria petrolifera e l'eccesso di mortalità per tumore del sistema emolinfopoietico*

4.1 Il disegno dello studio

L'Indagine è stata condotta seguendo uno studio analitico di mortalità con la tecnica casi-controlli su base di popolazione. Si tratta di un approccio metodologico basato sulla ricostruzione retrospettiva della storia abitativa dei soggetti in studio, assunta come proxi dell'esposizione individuale. Nell'Indagine, è stato indagato un periodo di 15 anni, antecedente 5 anni la data indice. Lo studio di mortalità ha interessato una popolazione complessiva di 54.994 abitanti (al 2003).

- *area di studio*: comuni di Falconara Marittima, Montemarciano e Chiaravalle (circa 75 km^2)
- *periodo in studio*: dal 1/1/1994 al 31/12/2003

La rilevazione delle informazioni per la costruzione dei database su cui effettuare le nostre analisi statistiche, ha previsto le seguenti fasi:

a. individuazione, dalle schede di mortalità ISTAT, della lista dei casi, cioè delle persone decedute per tumore del sistema emolinfopoietico (ICD-9: 200-208), residenti nell'*area di studio,* nel *periodo di studio;*
b. ricostruzione nel corso del *periodo di studio* delle popolazioni residenti nell'*area di studio;*
c. campionamento dei controlli.

   La procedura di campionamento dei controlli è stata effettuata secondo la procedura denominata "incidence density sampling" tipica degli studio "case-control nested in the cohort" (Rothman 1998). Per ogni caso, è stata individuata la popolazione eligibile (risk set) dalla quale estrarre 2 controlli con metodo casuale; essa era costituita dai soggetti dei tre

comuni aventi le caratteristiche richieste per l'appaiamento (stesso sesso e con età +/- 2.5 anni) in vita al momento della data di decesso del caso (data indice).

- effettuazione di interviste a familiari dei casi e dei controlli, attraverso un questionario finalizzato a ricostruire la storia abitativa e la storia occupazionale dei soggetti in studio e a studiare l'esistenza di eventuali fattori di confondimento;
- rilevazione, mediante georeferenziazione, delle coordinate geografiche delle *abitazioni* dei soggetti in studio. Dato rilevato dai questionari.
- rilevazione, mediante georeferenziazione, delle coordinate geografiche delle *residenze* dei soggetti in studio. Dato rilevato dalle anagrafiche comunali.

Sono stati quindi individuati 531 (177 casi e 354 controlli) soggetti in studio. Di questi, ne vennero rintracciati il 93% pari a 493 soggetti (165 casi e 328 controlli). Tra loro accettarono di rispondere all'intervista i familiari di 376 soggetti (109 casi e 267 controlli).

Dopo appaiamento sono risultati disponibili due differenti database su cui effettuare le analisi:

Database *Residenze*, formato da 509 soggetti in studio (172 casi, 337 controlli), contenente informazioni sulla locazione delle residenze da anagrafe comunale. In questo database mancano tutte le informazioni relative alle abitudini di vita del soggetto.

Database *Abitazioni*, formato da 277 soggetti in studio (101 casi, 176 controlli), contenente informazioni sulla localizzazione delle abitazioni (non sempre coincidenti con le residenze ufficiali) e sulle abitudini di vita o potenziali noti fattori di rischio. Le informazioni sono tutte tratte dal questionario somministrato ai parenti dei soggetti e sono elencate sotto:

- dati anagrafici/stato civile/grado di istruzione;
- esposizione al fumo;
- patologie individuali pregresse e patologie familiari;
- esposizioni professionali;
- storia abitativa nei 40 anni antecedenti l'anno di riferimento, che ha riguardato le caratteristiche delle abitazioni (tipo di riscaldamento, coabitanti fumatori, area urbana) vicinanza dell'abitazione a possibili fonti di inquinamento (elettrodotti, stazioni radio-base, discariche, distributori o depositi carburanti) e a strade interessate a traffico autoveicolare.

4.2 Distribuzione nello spazio delle residenze/abitazioni principali dei soggetti in studio

Per la rappresentazione delle residenze (o abitazioni) principali dei casi e controlli in studio, abbiamo in primo luogo creato una finestra "window", nel nostro caso un poligono che corrisponde approssimativamente all'area dei tre comuni in studio, comprendente una superficie quadrata di 80Km^2. Per convenzione ho fatto in modo che il centroide della raffineria (point source pollution) coincidesse con le coordinate c(0,0).

In Fig.4 vediamo la distribuzione di casi e controlli relativa al database dello studio caso-controllo contenente le coordinate geografiche delle residenze principali di 509 soggetti (172 casi e 337 controlli).

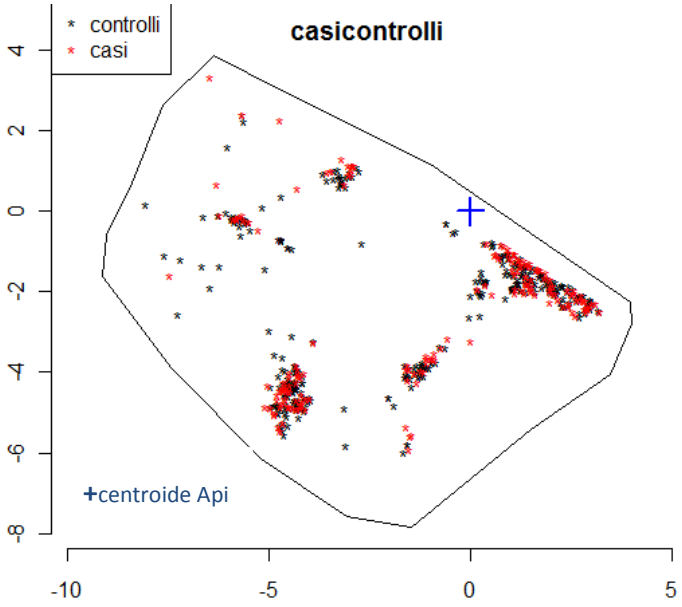

Figura 4 - Distribuzione di Residenze principali di 172 casi e 337 controlli

4.3 Proprietà del primo ordine: la funzione di densità Kernel

Nel paragrafo *2.2* abbiamo introdotto la funzione di densità Kernel e detto che una stima dell' intensità rispettivamente per i casi e per i controlli potrebbe darci una prima valutazione "visiva" di eventuali maggiori aggregazioni di casi di decesso per malattia nella regione considerata. Nei grafici (Fig.5 e Fig.6) seguenti possiamo osservare due picchi di intensità aumentata, rispettivamente per i casi e per i controlli. Il simbolo in rosso indica la posizione del centroide della raffineria.

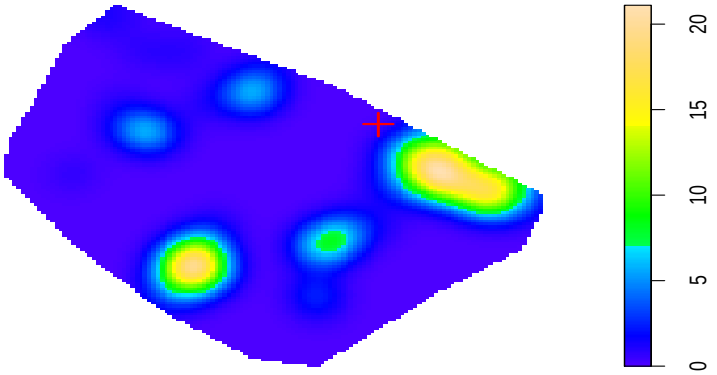

Figura 5 - Densità Kernel. Residenze principali. Casi

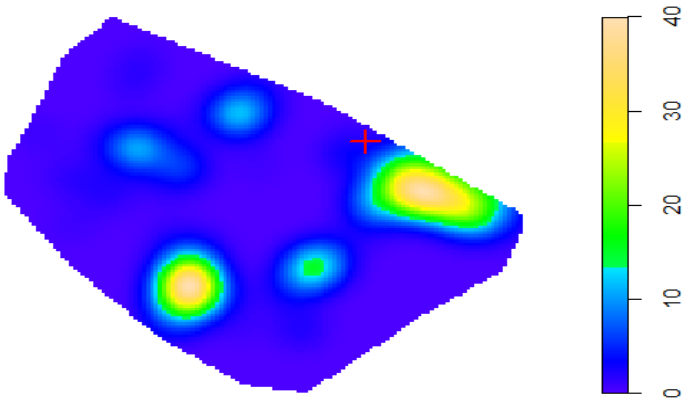

Figura 6 - Densità Kernel. Residenze principali. Controlli

La stima di queste densità non sembra essere di significativo interesse in quanto casi e controlli seguono la naturale distribuzione della popolazione che è concentrata maggiormente nei centri abitati. La zona gialla infatti sembra coincidere con i comuni di Falconara Marittima e Montemarciano.

4.4 Proprietà del secondo ordine: la funzione K di Ripleye la ricerca dei cluster

Dopo aver definito precedentemente la funzione K di Ripley, consideriamo il nostro database contenente 509 soggetti deceduti per tumore del sistema emolinfopoietico (172 casi e 337 controlli)

La funzione K di Ripley viene stimata per i casi [Kca(h)]e per i controlli [Kco(h)] tramite la funzione Kest del package “spatstat” in R. Stimiamo e rappresentiamo graficamente, per una migliore interpretazione, una trasformata della funzione K(h), la funzione L(h) (Cuzik ed Edwards 1990) (Fig.7 e Fig.8).

La linea rossa tratteggiata rappresenta la situazione ipotetica "teorica" di un processo omogeneo di Poisson con cui confrontare i dati osservati. L'area identificata dalle bande grigie rappresenta i limiti di confidenza (stimati tramite tecniche di simulazione di Monte Carlo) entro i quali possiamo affermare che il processo di punto non presenta né regolarità, né cluster, quindi nella situazione in cui $K(h) = \pi h^2$.

Dai nostri grafici si può facilmente intuire che siamo nella situazione di evidenza di cluster in cui $K(h) > \pi h^2$.

Ma questi due grafici considerati singolarmente non fanno che riflettere quanto già evidenziato nelle figure 5 e 6, quindi non ci danno nessuna nuova informazione. E' infatti abbastanza intuibile pensare che casi di cancro presenteranno sempre un certo cluster proprio a causa della distribuzione della popolazione a rischio (Gatrell 1995), addensata più nei centri urbani che nella campagne, mai uniforme.

È dunque più logico pensare di studiare l'evidenza di un cluster di un tipo di eventi in relazione ad un altro tipo, attraverso operazioni sulle funzioni, come descritto nel paragrafo 1.6.2. Sono stati considerati quindi $n_1=172$ "eventi di tipo 1" che sono i casi ed $n_2=337$ "eventi di tipo 2" (controlli), è stata stimata separatamente la funzione K di Ripley per il gruppo dei casi (Kca) e per il gruppo dei controlli (Kco) e calcolata la differenza tra le due funzioni K:

$D(r)= (Kca(r))- (Kco(r))$

Secondo il metodo descritto nel *par 1.6.2* (Cuzik ed Edwards 1990), tale differenza rappresenta una misura di extra-aggregazione di casi osservati sui controlli (Diggle 2003). Per il calcolo delle bande di confidenza è stata utilizzata la randomizzazione di Monte Carlo che permuta in maniera random le localizzazioni di casi e controlli e valuta poi la funzione stimata D(r) per ogni permutazione (Chetwynd e Diggle 2001). Gli andamenti

della funzione D(r), con relative bande di confidenza sono rappresentati nelle figure 9 e 10 e sono stati generati attraverso la funzione “envelope” del package spatstat di R. Valori della funzione D(r) sopra la banda del limite superiore mostrano evidenza di aggregazione, quindi di cluster spaziale.

In definitiva, dal grafico (Fig. 9) che rappresenta la D(r) calcolata per il dataset delle Residenze, possiamo evincere che c’è una tendenza al cluster, in quanto la linea nera dei dati osservati giace prevalentemente al di sopra di quella teorica per 0.5<r<3, ma non è statisticamente significativa, in quanto rimane all’interno delle bande di confidenza.

Se consideriamo il database Abitazioni, la D(r) ci mostra una piccolissima evidenza non significativa al cluster per r compreso tra 0.6 e 2 (Fig.10)

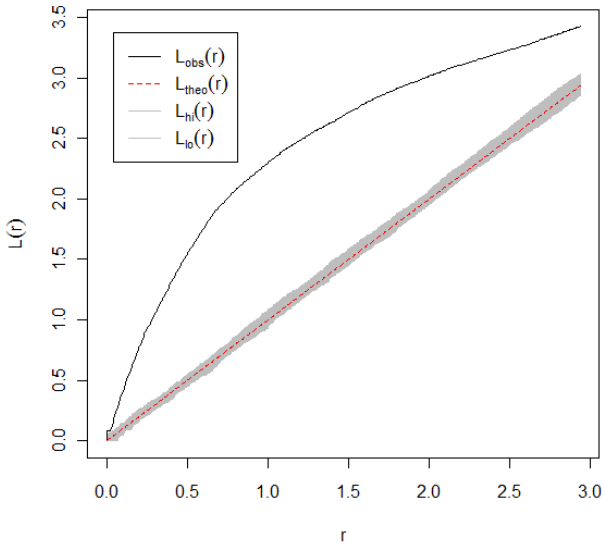

Figura 7 - Trasformata di K di Ripley. Residenze principali, casi

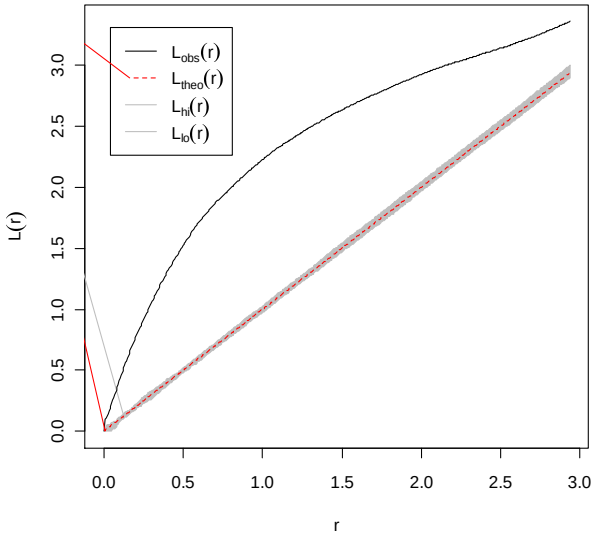

Figura 8 - Trasformata di K di Ripley. Residenze principali, controlli

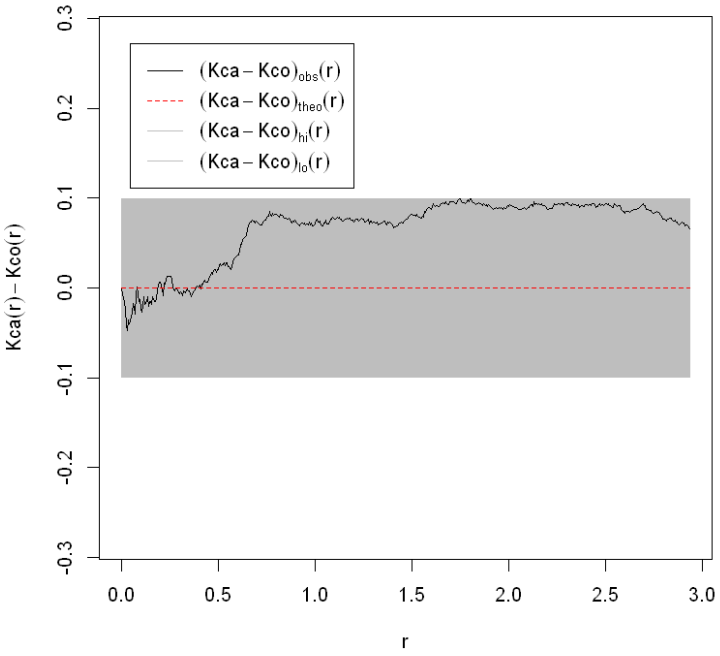

Figura 9 - Funzione differenza K Ripley per casi e controlli. Residenze principali

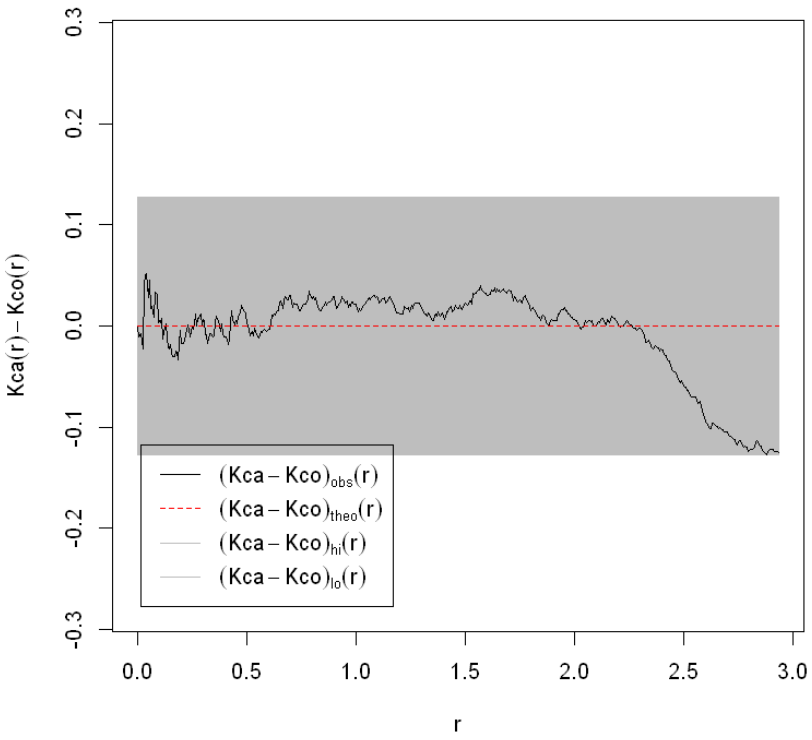

Figura 10 - Funzione differenza K Ripley tra casi e controlli. Abitazioni principali

4.5 Stima della variazione del rischio relativo

In questo paragrafo sono riportati i risultati della stima della variazione del rischio relativo nell'area dei tre comuni in studio, ottenuta, come specificato sopra (*par. 2.4*) dal rapporto tra la densità bivariata Kernel per i casi e per i controlli, utilizzando un *adaptive* bandwith per lo smoothing Kernel.

Un primo grafico (Fig. 11) mette a confronto due mappe: la variazione spaziale del rischio relativo di morte, stimato con la tecnica del fixed bandwidth e quello stimato con la tecnica dell'adaptive bandwidth. Il database è quello delle Residenze, con 172 casi e 337 controlli. Ad occhio ci accorgiamo subito che la tecnica basata su un bandwidth costante per tutti i punti dell'area mette in evidenza un aumento del rischio in una zona in alto con una bassa densità di popolazione. Questo picco del

rischio ci può portare verso errate conclusioni in quanto è molto probabile che esso sia dovuto alla presenza di un caso isolato.

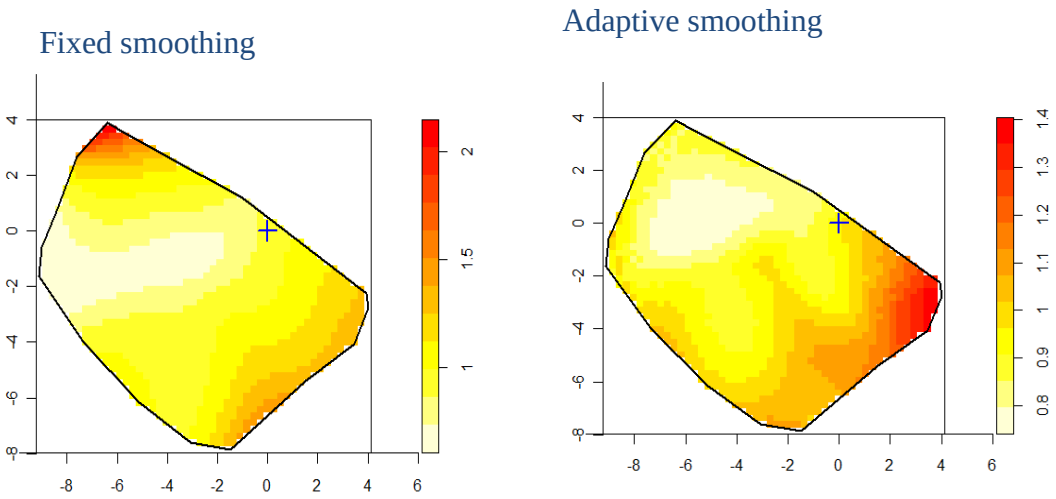

Figura 11 - Variazione spaziale del rischio relativo. Confronto fixed e adptive bandwidth. Residenze principali 172 casi e 337 controlli

I grafici sotto (Figg. 12-15) mostrano come varia il rischio nell'area di studio, considerando i vari sottogruppi dei soggetti in studio.

Nelle figura precedente abbiamo visto una rappresentazione della variazione spaziale del rischio relativo per l'intero di pool di casi e controlli, basati sulle residenze principali (*database Residenze*). Adesso, vediamo cosa cambia se prendiamo in considerazione solo le donne, il campione quindi di 87 casi e 177 controlli, basati sempre sulle residenze principali.(Fig. 12)

Il razionale della scelta di questo sottogruppo nasce dal fatto che le donne, di cui la maggior parte casalinghe o pensionate, abbiano trascorso più tempo nell'abitazione principale e risultino quindi meglio classificate.

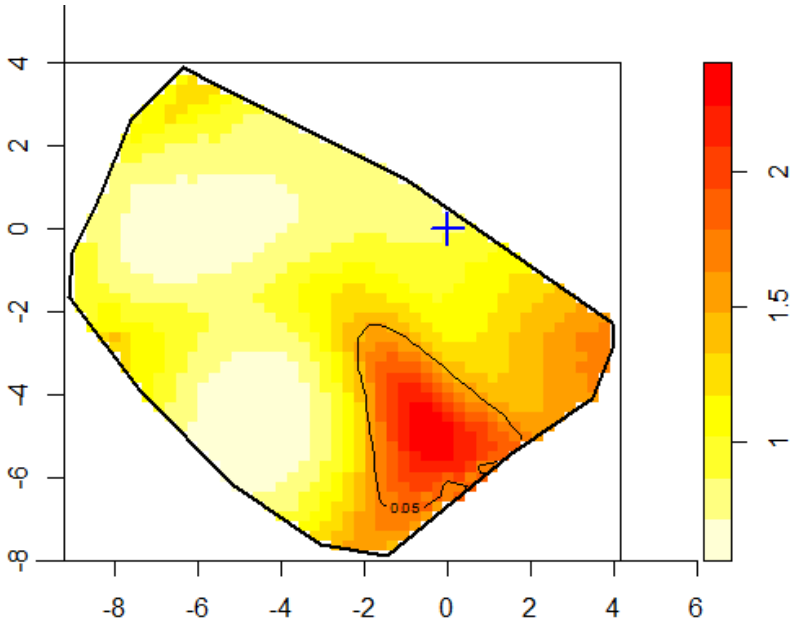

Figura 12 - Variazione del Rischio Relativo. Residenze principali. Donne. 82 casi e 177 controlli

Le figure che seguono (Figg. 13-15) mostrano una rappresentazione della variazione del rischio relativo, considerando il database *Abitazioni* e i seguenti 2 sottogruppi:

Abitazioni donne: Il razionale della scelta di questo sottogruppo riguarda il fatto che le donne, avendo una età media pari a circa 76 anni, erano probabilmente casalinghe e trascorrevano più tempo degli uomini nella residenza/abitazione più o meno prossima alla raffineria.

Abitazioni indoor: questo gruppo, che rappresenta il gruppo di coloro che hanno dichiarato di essere casalinghe, pensionati, inoccupati, per almeno 10 anni nel periodo di studio, è stato preso in considerazione perché si ipotizza che questi soggetti abbiano vissuto nelle loro residenze/abitazioni per più tempo negli anni considerati.

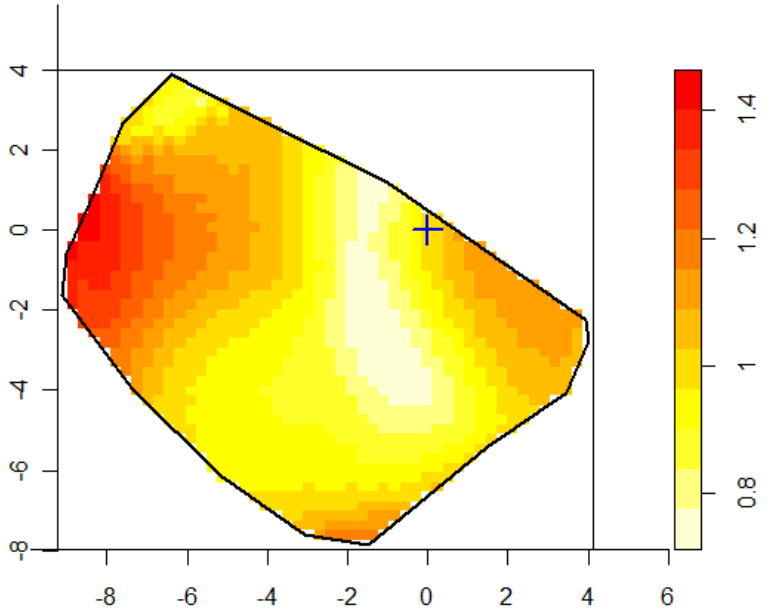

Figura 13 - Variazione del Rischio Relativo. Abitazioni. Uomini e Donne.

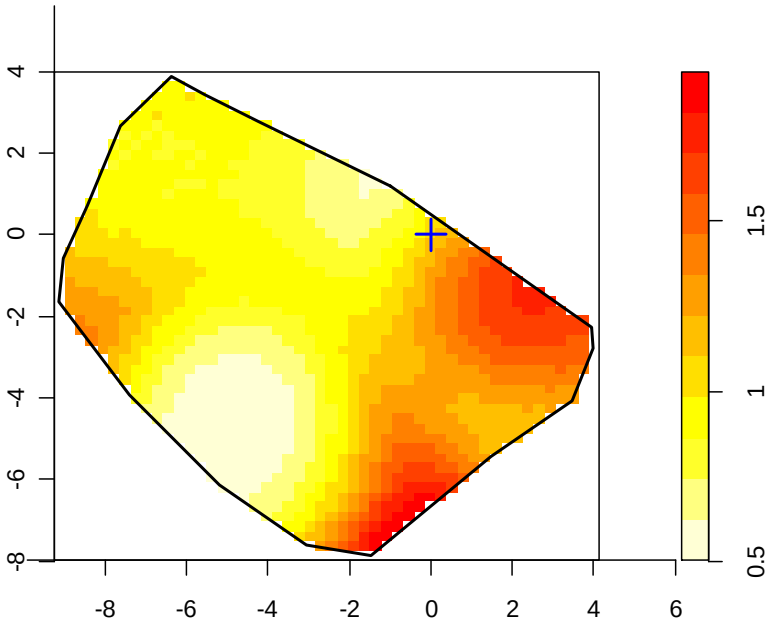

Figura 14 - Variazione del Rischio Relativo. Abitazioni. Indoor

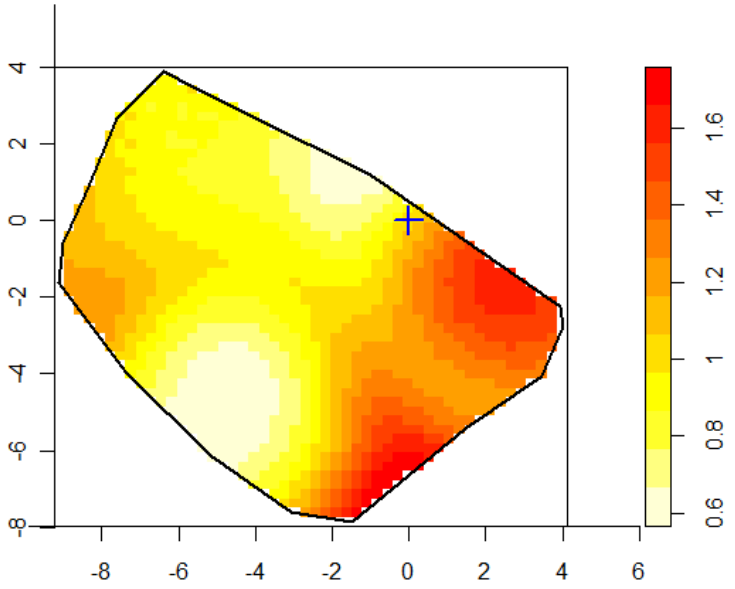

Figura 15 - Variazione del Rischio Relativo. Abitazioni. Donne

Possiamo notare che nel database *Residenze* si evidenzia un lieve aumento del rischio in un'area a circa 3 Km dal centroide Api (coordinate c(-3, 3)) (Fig. 11, destra). Nel sottogruppo delle donne del database *Residenze*, si evidenzia un rischio più elevato in un'area a circa 4-6 Km da Api (coordinate (0, -6)) (Fig. 12). Inoltre questo è l'unico sottogruppo in cui l'area ad alto rischio risulta essere significativa, come mostrato dai contorni di tolleranza (*tolerance countors*), calcolati con la metodologia esposta nel paragrafo 2.4.3. Se consideriamo il database *Abitazioni*, uomini + donne, notiamo un leggerissimo aumento del rischio in un'area (coordinate (-8, -1)) (Fig.13). Nel sottogruppo delle donne e degli *indoor*, cioè pensionati e casalinghe che hanno vissuto nell'area in studio per almeno 10 anni, notiamo due picchi di aumento del rischio: uno in un'area molto vicina alla raffineria (coordinate (2, -2)), un altro più lontano a circa 6 Km verso sud della raffineria (Figg. 14 e 15).

Per la stima delle densità bivariate kernel, che hanno determinato la variazione spaziale del rischio relativo è stato utilizzato un pilot bandwidth comune sia per i casi che per i controlli, determinato con la tecnica basata sul principio del maximal smoothing (Terrel 1990) discussa nel paragrafo 2.4.1. Nella tabella 1 sotto sono riportati i valori del bandwidth utilizzati per ciascun sottogruppo.

Gruppo	N.casi	N.controlli	Bandwidth
Residenze	172	337	1.40 km
Residenze donne	87	167	1.57 km
Abitazioni	101	176	1.54 km
Abitazioni donne	47	87	1.73 km
Abitazioni indoor	54	105	1.80 km

Tabella 1 - Valori del bandwidth selezionato per ciascun gruppo di casi e controlli, ottenuto con la tecnica del "maximal smoothing"

4.6 Modelli con covariate spaziali

Come specificato nel paragrafo 3.3, posso ipotizzare di voler modellare il pattern di localizzazioni residenze (o abitazioni) dei casi di morte per tumore del sistema emolinfopoietico, come un processo di punto con intensità $\lambda_1(x)$ proporzionale alla densità della popolazione a rischio, quindi alla densità dei controlli $\lambda_0(x)$

Stimo dunque la densità dei controlli utilizzando lo stimatore di intensità Kernel-smoothed, ottenendo una immagine pixels (Vedi Figura 6) e fitto i modelli seguenti:

Modello M1: $\log \lambda_1(x) = \alpha + \log \lambda_0(x)$

dove α è il parametro sconosciuto, $\lambda_1(x) = \mu\, \lambda_0(x)$ e $\mu = e^{\alpha}$ è il solo parametro da stimare.

Modello M2: $\log \lambda_1(x) = \alpha + \beta x + \log \lambda_0(x)$

Dove x rappresenta la distanza delle abitazioni dalla raffineria

Il modello è stato stimato tramite la funzione ppp del package "spatstat". (Baddeley 2005)

M1 è il modello che contiene una covariata spaziale, la densità dei controlli $\lambda_0(x)$ che serve per modellare la distribuzione dei casi. Confronto M1 con l'altro modello M2 che tiene conto dell'effetto "source pollution", in cui aggiungo un'altra covariata: la distanza di ogni locazione dalla raffineria petrolifera, trasformata in una immagine pixel (Fig. 16).

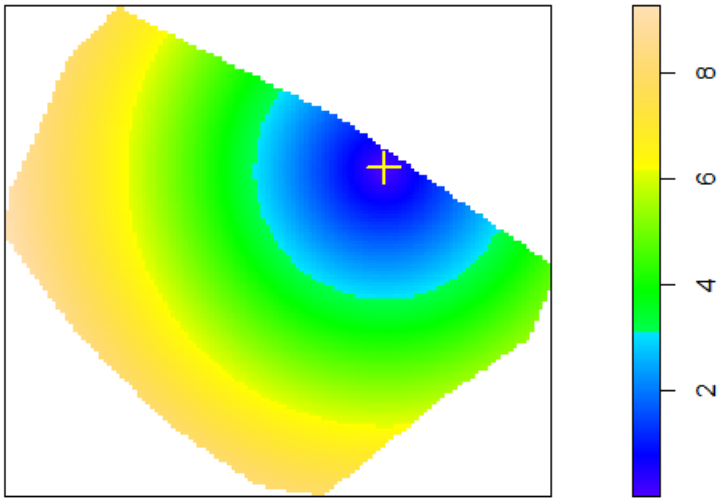

Figura 16 - Covariata spaziale: distanza dalla raffineria in immagine pixel

Nella tabella 2 sono riportate le stime dei coefficient del modello

Gruppi	Modello	Intercetta α	β
Residenze	M1	0.718	
	M2	0.451	0.062
Residenze donne	M1	0.695	
	M2	0.109	**0.137 ****
Abitazioni	M1	0.600	
	M2	0.491	0.024
Abitazioni donne	M1	0.661	
	M2	0.188	0.105
Abitazioni indoor	M1	0.708	
	M2	0.385	0.074
Abitazioni donne indoor	M1	0.726	
	M2	0.205	0.118

Tabella 2 – Stime dei coefficienti del modello

Sembrerebbe che, nel sottogruppo delle donne, ci sia un lieve effetto significativo della distanza dalla raffineria.

Confronto i modelli con il test rapporto verosimiglianze, utilizzando la funzione “anova”. Nelle donne il test “Chi” restituisce un valore di p-value= 0.007 significativo.

5 *Discussione e conclusioni*

L'obiettivo principale di questo studio è stato la descrizione e rappresentazione grafica della variazione del rischio relativo di decesso per tumore ematologico, in un'area territoriale comprendente 3 comuni, prossima alla raffineria petrolifera di Falconara Marittima (AN).

Nel mio percorso di studi del dottorato, ho letto e valutato i vari approcci a questo problema e le tecniche di analisi spaziale utilizzate per la stima della variazione spaziale del rischio relativo. Le tecniche basate sul rapporto di densità smoothing tra casi e controlli, sono ormai abbastanza affermate in letteratura in questo campo. Poiché la distribuzione dei dati è fortemente disomogenea, dopo un confronto tra i due metodi applicato ai nostri dati (vedi figura 11) ho ritenuto opportuno utilizzare il metodo *adaptive* che mostra una stima più verosimile del rischio relativo. Quello che tuttavia è ancora oggetto di studi e dibattiti è la scelta dell'ampiezza di banda, del bandwidth, nella stima del rischio relativo, all'interno di studi di tipo caso-controllo. In alcuni casi, le tecniche computazionali sviluppate di cui abbiamo discusso sopra, possono andar bene per un dataset, meno bene per un altro. Nel nostro caso, la tecnica basata sul principio dell'Oversmoothing porta dei buoni risultati con valori del pilot bandwidth riportati in tabella 1.

Il metodo *adaptive* consente di stimare le densità bivariate di casi e controlli con due differenti valori di bandwidth. Ma per la natura dei dati a mia disposizione, ho deciso di scegliere lo stesso valore del bandwidth per i casi e per i controlli. Ho prima verificato che non ci fossero cluster di casi significativi attraverso la funzione K di Ripley, riportata nelle figure 9 e 10. Questo mi ha portata a concludere che non ci fossero particolari differenze tra casi e controlli in termini di densità di probabilità.

Alcune riflessioni sui modelli utilizzati per studiare l'associazione tra l'eventuale rischio di morte e la vicinanza delle abitazioni dalla raffineria petrolifera. Innanzi tutto, sulla variabile di esposizione. La scelta della distanza, come misura approssimata dell'esposizione, è motivata principalmente dalla carenza di dati ambientali. Infatti l'ottenimento di dati affidabili sull'esposizione è problematico ed in molti casi rappresenta il punto debole di molte applicazioni epidemiologiche. Molto raramente, infatti, è possibile disporre di stime dirette della concentrazione degli inquinanti nell'area di studio.

Dai risultati che abbiamo ottenuto posso concludere che l'analisi esplorativa, effettuata tramite la funzione differenza K di Ripley, mi ha permesso di individuare dei cluster di casi, che però rimangono all'interno delle bande di significatività, maggiormente evidenti quando abbiamo considerato le residenze, anziché le abitazioni, probabilmente per il numero maggiore di soggetti (509 contro 276).

Dalle mappe sulla variazione spaziale del rischio relativo notiamo che nel sottogruppo delle donne pare esserci un picco di aumento del rischio, non significativo, che appare nei pressi della raffineria se consideriamo le abitazioni. Mentre una zona ad alto rischio, significativo, appare a circa 5 Km dalla raffineria se consideriamo le residenze principali.

Un limite importante di queste mappe che mostrano la variazione spaziale del rischio relativo, prodotte con il metodo del rapporto tra le densità, è che non permettono di tener conto dell'appaiamento caso-controllo in particolare per quanto riguarda il periodo di esposizione. E' plausibile pensare infatti che nel tempo, l'esposizione possa essere variata.

Dalle stime dei modelli, possiamo dire che l'unico sottogruppo in cui l'effetto source pollution della raffineria risulta significativo nel modello,

è quello delle donne (peraltro significativamente più anziane degli uomini), in cui abbiamo considerato le residenze principali.

Questo progetto, che è costato anni di lavoro a tante figure professionali, è servito a me da stimolo per iniziare un percorso di studio all'interno del dottorato di ricerca e mi ha consentito di sviluppare competenze nel campo dell'analisi spaziale, applicata all'epidemiologia ambientale. Conto di poter sfruttare queste competenze acquisite in altri studi e indagini della stessa tipologia, che abbiano cioè come obiettivo principale l'individuazione di aree a rischio e l'associazione con fonti di inquinamento ambientale.

Ringrazio il Dott. Andrea Micheli, responsabile del progetto Api-Falconara, la Dott.ssa Elisabetta Meneghini e il Dott. Paolo Baili, che hanno condotto la studio in toto e mi hanno aiutata tantissimo nelle fasi di studio, decisione e di analisi.

6 Bibliografia

- Abramson IS (1982). On Bandwith Estimation in Kernel Estimates – A square root low. The Annals of Statistics. 10 (4): 1217-1223
- Alexander DD, Wagner ME (2010). Benzene exposure and non-Hodgkin lymphoma: a meta-analysis of epidemiologic studies. J Occup Environ Med;52(2):169-89
- Baan R, Grosse Y, Straif K, Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Freeman C, Galichet L, Cogliano V; WHO International Agency for Research on Cancer
- Baddeley A, Turner R. (2000). Practical maximum pseudolikelihood for spatial point pattern (with discussion). Australian and New Zaeland Journal of Statistics. 43(3): 283-322.
- Baddeley A, Turner R (2006). Modellind spatial point patterns in R. In Baddeley et al. editors, Case Studies in Spatial Point Pattern Modelling, number 185 in Lecture Notes in Statistics, pages 27-34. Springer-Verlag. New York.
- Baddeley A, Turner R (2005). Spatstat: An R package for analysing spatial point pattern. Journal of Statistical Software. 12(6): 1-42
- Baili P, Mariottini M, Meneghini E, Micheli A (2007). Feasibility study of launching an epidemiologic survey of the resident population near the API refinery in Falconara Marittima. Epidemiol Prev. 31(1 Suppl 2):48-53
- Bailey TC, Gatrell AC (1995). Interactive Spatial Data Analysis. Longman.
- Bithell JF (1990). An application of Density Estimation to Geographical Epidemiology. Statist Med, 9: 961-701.
- Bithell JF (1991). Estimation od Relative Risk Functions. Statist Med, 10: 1745-1751

- Bivand RS, Gomez-Rubio V, Pebesma EJ (2008). Applied Spatial Data Analysis with R. Springer.
- Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71, 353-360.
- Bowman AW, Azzalini A. (1997). Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-plus Illustration. Oxford University Press Inc.: New York.
- Chetwind AG, Diggle PG (2001). Investigation of spatial lustering from individually matched case-control studies. Biostatistics. 2: 277-293
- Clayton D, Hills M (1998). Statistical Models in Epidemiology. Oxford: Oxford University Press.
- Cocco P, t'Mannetje A, Fadda D, Melis M, Becker N, de Sanjosé S, Foretova L, Mareckova J, Staines A, Kleefeld S, Maynadié M, Nieters A, Brennan P, Boffetta P (2010). Occupational exposure to solvents and risk of lymphoma subtypes: results from the Epilymph case-control study. Occup Environ Med;67(5):341-7.
- Costantini AS, Benvenuti A, Vineis P, Kriebel D, Tumino R, Ramazzotti V, Rodella S, Stagnaro E, Crosignani P, Amadori D, Mirabelli D, Sommani L, Belletti I, Troschel L, Romeo L, Miceli G, Tozzi GA, Mendico I, Maltoni SA, Miligi L (2008). Risk of leukemia and multiple myeloma associated with exposure to benzene and other organic solvents: evidence from the Italian Multicenter Case-control study. Am J Ind Med;51(11):803-11.
- Cressie NAC (1993). Statistics for spatial data. Wiley Series in Probability and Statistics.
- Cuzic J, Edwards R (1990) Spatial clustering for inhomogeneus population. JRSS B 52: 73-104
- Davies TM and Hazelton ML (2010). Adeptive kernel estimation of spatial relative risk. Statist. Med, 29: 2423-2437

- Davies TM, Hazelton ML, Marshall JC (2011). sparr: Analyzing Spatial Relative Risk Using Fixed and Adaptive Kernel Density Estimator in R. Journal of Statistical Software. Vol. 39. Issue 1.
- Danese M et al. (2008). In Kernel Density Estimation Methods for a Geostatistical Approach in seismic Risk Analysis: The case study of Potenza Hilltop Town (Southern Italy). Lecture Notes in Computer Science, vol. 5072. Springer: Berlin; 415-429
- Diggle PG (2003). Statistical analysis of spatial point patterns, 2nd edn. Hodder Arnold
- Diggle PG and Rowlingson BS (1994). A conditional approach to point process modelling of elevate risk. J.R. Statist Soc. 157: 433-440
- Dreassi E (2008). Note del corso di Statistica ambientale. Analisi statistica dei dati spaziali.
- Elliot P, Wakefield JC, Best NG, Brings SE (2000). Spatial epidemiology: methods and applications. In Elliott, P.,Wakefield, J. C., Best, N.G., Brings D. J. editors. *Spatial Epidemiology: Methods and applications*, Oxford: University. Press: 3-14.
- Gatrell AC, Diggle PJ (1995). Spatial point pattern analysis and its application in geographical epidemiology. Trans Inst Br Geogr; 21: 256-274
- Hazelton ML (2007). Bias reduction in kernel binary regression. Computational Statistics and Data Analysis. 51: 4393-4402.
- Hazelton ML and Davies TM (2009). Inference based on kernel estimates of the relative risk function in geographical elidemiology. Biometrical Journal. 51: 98-109.
- Hall P and Marron JS (1987). Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation. Probability Theory and Related Fields, 74, 567-581.

- Hall P and Marron JS (1988). Variable window width kernel density estimates of probability densities. Probability Theory and Related Fields. 80: 37-49.
- IARC Monopraphs on the Evaluation of carcinogenic risk to humans. (2009)
- Infante PF (2006). Benzene exposure and multiple myeloma: a detailed meta-analysis of benzene cohort studies. Ann N Y Acad Sci;1076:90-109.
- Kane EV, Newton R (2010). Benzene and the risk of non-Hodgkin lymphoma: a review and meta-analysis of the literature. Cancer Epidemiol; 34(1):7-12.
- Kelsall JE, Diggle PJ (1995). Non-parametric estimation of spatial variation in relative risk. Statistics in Medicine. 14: 2335-2342
- Kelsall JE, Diggle PJ (1995). Kernel estimation of relative risk. Bernoulli. 1:3-16.
- Kirkeleit J, Riise T, Bratvelt M, Moen B (2008). Increased risk of acute myelogenous leukemia and multiple myeloma in a historical cohort of upstream petroleum workers exposed to crude oil. Cancer Causes Control; 19:13–23.
- Marron, J.S. (1992) Bootstrap Bandwidth Selection. In: LePage, R., L. Billard, (a cura di) Exploring the Limits of Bootstrap. Wiley, New York.
- Parzen, E. (1962). On Estimation of a Probability Density and Mode. The Annals of Mathematical Statistics, 33, 1065-1076.
- Rothman KJ, Greenland S (1998). Modern epidemiology, 2nd edition. Philadelphia: Lippincott Williams & Wilkins
- Sabel CE et al. (2000) Modelling exposure opportunities: estimating relative risk for motor neurone disease in Finland. Social Science & Medicine. 50: 1121-1137.
- Schlesselman J (1982). Case control studies: design, conduct, analysis. Oxford: Oxford University Press.

- Silverman BW (1986). Density Estimation for Statistics and Data Analysis. Chapman & Hall: New York.
- Smith MT, Jones RM, Smith AH (2007).Benzene exposure and risk of non-Hodgkin lymphoma. Cancer Epidemiol Biomarkers; 16(3):385-91.
- Steinmaus C, Smith AH, Jones RM, Smith MT (2008). Meta-analysis of benzene exposure and non-Hodgkin lymphoma: biases could mask an important association. Occup Environ Med; 65(6):371-8.
- Terrel GR (1990). The maximal smoothing principle in density estimation. Journal of the American Statistical Association. 85: 440-447
- Upton G, Fingleton B (1985). Spatial data analysis by example. Point pattern and quantitative data. Vol. 1. Wiley series in probability and mathematical statistics.
- Vlaanderen J, Lan Q, Kromhout H, Rothman N, Vermeulen R (2011). Occupational benzene exposure and the risk of lymphoma subtypes: a meta-analysis of cohort studies incorporating three study quality dimensions. Environ Health Perspect;119(2):159-67.
- Wheeler DC (2007). A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003. International Journal of Health Geographics. 6(13)