



Analisi micro-geografica dei casi di sarcoma nella provincia pistoiese

Giorgia Stoppa, Daniela Nuvolone, Fabio Voller
Osservatorio di Epidemiologia - Agenzia regionale di sanità della Toscana

Documento del 2 Febbraio 2022, rivisto ed aggiornato il 4.5.2023

Introduzione

A seguito della notifica da parte di alcuni medici di famiglia su un numero elevato di tumori del tessuto connettivo – sarcomi - nell'area di Casalguidi, una frazione del comune di Serravalle Pistoiese, l'Azienda AUSL Toscana Centro (ASL TC) ha incaricato l'Agenzia Regionale di Sanità della Toscana (ARS) di condurre un'analisi specifica sulla diffusione dei sarcomi estesa all'intera provincia di Pistoia. L'obiettivo è quello di valutare nel periodo 2013-2018 la distribuzione geografica dei sarcomi e l'eventuale presenza di aggregazioni spaziali anomale o "cluster".

Materiali e metodi

Per l'analisi sono stati utilizzati i casi incidenti di sarcoma occorsi nel periodo 2013-2018 tra i residenti nella provincia di Pistoia, identificati dal Registro Toscano Tumori (RTT) dell'Istituto per lo studio, la prevenzione e la rete oncologica (ISPRO) nel periodo più aggiornato disponibile 2013-2018. Per il periodo 2013-2015 i casi sono stati identificati attraverso i codici ICD-O-3 Topografia e ICD-O-3 Morfologia, mentre per gli anni dal 2016 al 2018, periodo per il quale non sono ancora disponibili i dati definitivi del RTT, sono usati i testi dei referti anatomopatologici con i codici SNOMED (Codici elencati in Appendice).

I casi incidenti totali individuati sono 161, di cui 81 maschi e 80 donne, con età media di 58 anni. In Tabella 1 è riportata la distribuzione per anno dei casi e nella Tabella 2 quella per comune di residenza al momento della diagnosi.

Tabella 1: Distribuzione dei casi di sarcoma in Provincia di Pistoia nel periodo 2013-2018 per anno di incidenza (N=161).

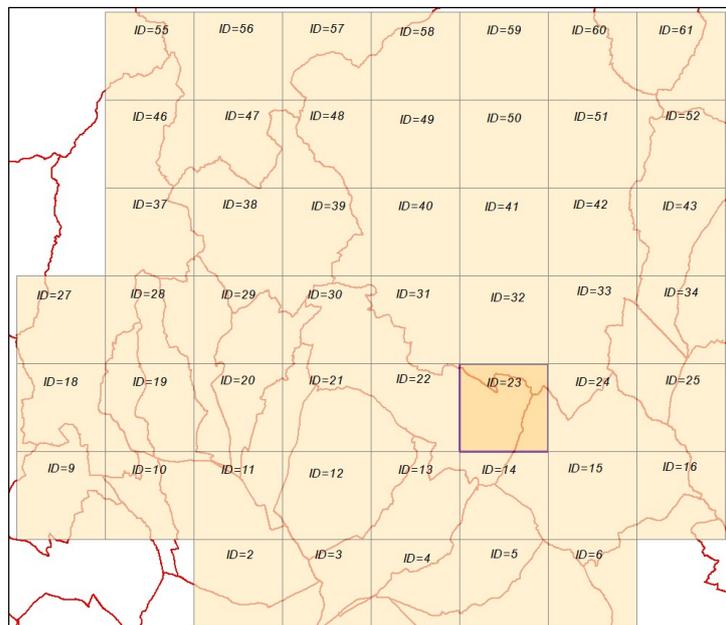
Anno di incidenza	Casi	%
2013	32	19.9
2014	31	19.3
2015	28	17.4
2016	25	15.5
2017	23	14.3
2018	22	13.7
Totale	161	100,0

Tabella 2. Distribuzione dei casi di sarcoma nel periodo 2013-2018 in Provincia di Pistoia per comune di residenza e tassi grezzi di incidenza per 100.000 abitanti (per motivi di privacy, i comuni con meno di 3 osservazioni sono stati aggregati nella voce "altri comuni").

Comune	Casi	%	Tasso grezzo x 100.000 ab.
Pistoia	55	34.2	10.2
Monsummano Terme	16	9.9	12.6
Serravalle Pistoiese	11	6.8	15.7
Montecatini Terme	10	6.2	8.2
Agliana	9	5.6	8.6
Lamporecchio	9	5.6	19.9
Quarrata	9	5.6	5.8
Ponte Buggianese	7	4.3	4.5
Pescia	6	3.7	5.1
Buggiano	5	3.1	9.5
Massa E Cozzile	5	3.1	10.5
Larciano	4	2.5	10.4
Montale	4	2.5	6.2
Pieve a Nievole	4	2.5	7.2
Altri comuni: Uzzano, Chiesina Uzzanese, Marliana, San Marcello-Piteglio	7	4,3	??

In collaborazione con la UFC Epidemiologia dell'ASL TC, i 161 casi sono stati georeferenziati con l'applicativo di Google Maps e sottoposti a verifica per accuratezza e precisione. Per valutare la presenza di aggregazioni spaziali anomale è stata costruita una griglia geografica di 4x4km (Figura 1), composta da 50 celle numerate (ID), di cui quella evidenziata (ID: 23) corrisponde alla frazione di Casalguidi, nel comune di Serravalle Pistoiese. La scelta della dimensione della griglia è basata sulle caratteristiche del territorio e sulla numerosità della popolazione in ciascuna cella. Per evidenziare eventuali eccessi che potevano essere presenti in aree territoriali di dimensioni più limitate rispetto a quelle definite dalla griglia di 4x4Km, l'analisi è stata ripetuta con una griglia di 2x2km. Così facendo, si guadagna in dettaglio geografico, ma si perde in potenza statistica perché le aree sono meno popolate, con minori casi osservati e attesi e maggiore incertezza delle stime. Per limitare questo problema, sono stati utilizzati modelli statistici specifici. In aggiunta questi è stato anche utilizzato un modello statistico specifico per la ricerca delle aggregazioni spaziali di venti.

Figura 1: Griglia 4x4km utilizzata per l'analisi statistica della distribuzione geografica dei casi di sarcomi nel periodo 2013-2018 nella provincia di Pistoia.

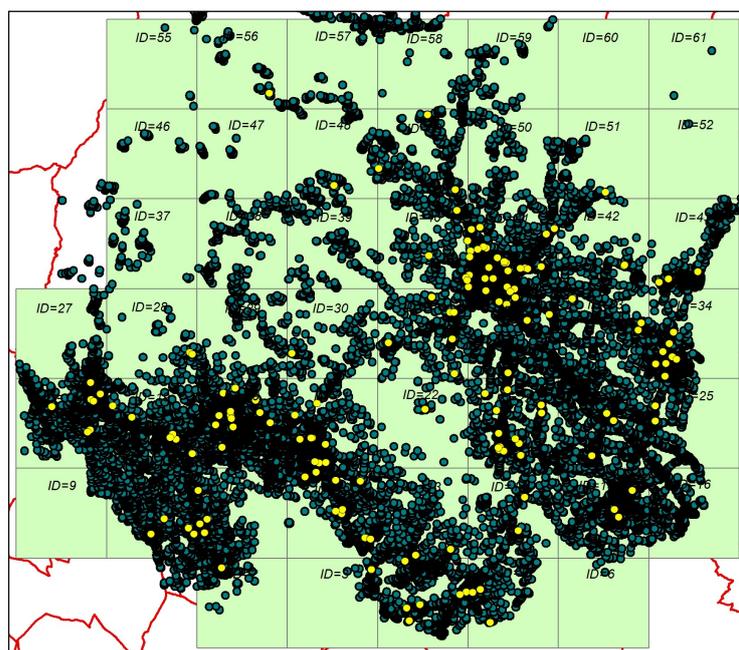


Pertanto, i metodi statistici ed i parametri analizzati per le celle delle grigliature geografiche di 4x4km e 2x2km sono stati:

- rapporti standardizzati di incidenza o Standardized Incidence Ratio (SIR), con intervalli di confidenza (IC) al 90%;
- stime bayesiane dei SIR, corrette con il modello di Poisson-Gamma (PG);
- stime bayesiane dei SIR, corrette con il modello di Besag-York-Mollié (BYM);
- analisi di clustering con Spatial Scan Statistic (SatScan).

Ogni cella della griglia contiene il numero dei residenti e dei casi di sarcomi con le caratteristiche di età e sesso (Figura 2).

Figura 2: Distribuzione della popolazione residente (punti blu) e dei 161 casi di sarcomi diagnosticati nel periodo 2013-2018 (punti gialli) in provincia di Pistoia.



Per quanto riguarda il calcolo dei SIR, in ogni cella della griglia sono stati conteggiati i casi di sarcomi “osservati” nel periodo in studio ed è stato calcolato il numero di casi “attesi” sulla base dei tassi di incidenza regionali. Gli attesi sono calcolati applicando alla popolazione residente in ciascuna cella i tassi regionali di incidenza dei sarcomi, specifici per sesso e classe di età. Il numero dei casi osservati è poi rapportato a quello degli attesi ed il rapporto tra casi osservati (O) e casi attesi (A) è definito Rapporto standardizzato di incidenza (O/A) (Standardized Incidence Ratio-SIR, nell’accezione inglese). Il termine *standardizzato* indica che è stato rimosso l’effetto dovuto alla diversa composizione per età e sesso delle popolazioni a confronto e, pertanto, il valore del SIR di ogni cella non è attribuibile a differenze nella composizione per genere ed età della popolazione residente nella cella e quella regionale. Essendo il SIR il rapporto O/A, se il SIR è maggiore di 1 gli osservati sono maggiori degli attesi e si è, quindi, in presenza di un eccesso di casi. Se il SIR è minore di 1 gli osservati sono minori degli attesi e si è in presenza di un difetto. Se i casi osservati sono 0 il SIR è 0, indipendentemente dagli attesi.

Per tenere conto dell’incertezza delle misure, sono stati calcolati gli intervalli di confidenza (IC) dei SIR, ovvero un range di valori entro cui ci si aspetta, con un certo grado di confidenza, ricada il valore reale. Minore è il numero di osservazioni e maggiore è l’imprecisione della stima e più ampi sono gli IC. In questo studio sono stati calcolati gli IC con approssimazione di Byar (vedi le “Note metodologiche”), con un livello di confidenza del 90%. Ciò significa che si accetta di correre un errore del 10% di identificare una differenza significativa - in eccesso o in difetto - rispetto all’atteso, quando invece la differenza è dovuta al caso. In pratica, per facilitare la lettura dei dati, gli eccessi sono “statisticamente significativi” quando entrambi gli IC – inferiori e superiori – sono maggiori di 1.0 e viceversa i difetti sono “statisticamente significativi” quando entrambi gli IC sono inferiori a 1.0.

Per ovviare alla instabilità statistica legata alla bassa numerosità delle osservazioni nelle aree geografiche, si utilizzano tecniche di aggiustamento dette di “smoothing” (lisciamento), che consentono di correggere le stime dei SIR più instabili. In questo studio sono stati applicati due metodi bayesiani per il “lisciamento” delle stime: il modello Poisson-Gamma (PG) e il modello Besag York Mollié (BYM) (in Appendice i dettagli dei due metodi) (Clayton, 1987; Besag, 1991).

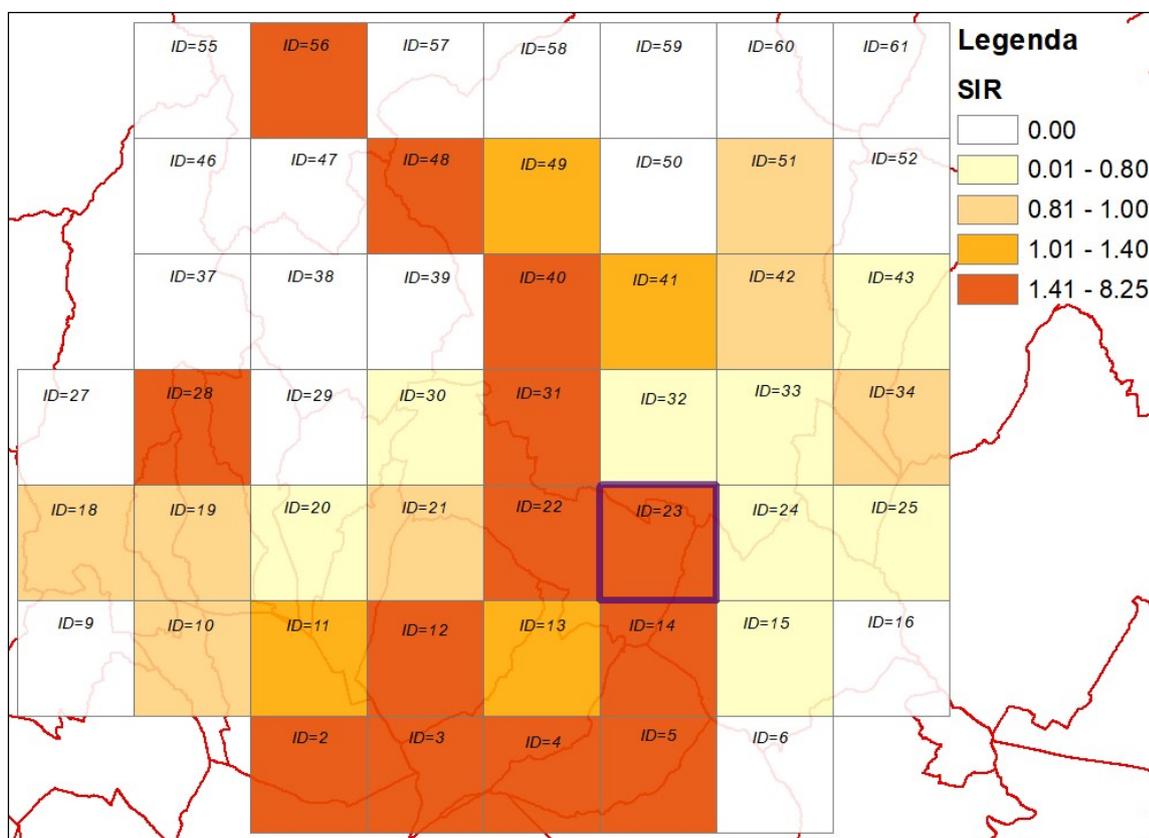
Nel modello PG, oltre agli eventi osservati in ciascuna cella della griglia, si tiene conto anche di informazioni relative alla distribuzione degli indicatori di malattia nell’intera area in studio. Il risultato di questa tecnica è un “lisciamento” dei SIR verso la media generale, applicando un grado di correzione tanto maggiore quanto minore è la numerosità dei casi osservati. Col modello BYM, la variabilità delle stime dei SIR viene attenuata tenendo conto sia della variabilità globale dei SIR nell’area in studio, come nel modello PG, sia della variabilità dei SIR nelle celle immediatamente adiacenti a quella in oggetto. Il modello tiene conto, cioè dell’autocorrelazione spaziale. Per ciascun modello è stata anche prodotta la rappresentazione cartografica della “probabilità a posteriori”, che mostra, per ogni area, la probabilità che il SIR “lisciato” con i modelli PG e BYM sia maggiore di 1. E’ convenzionalmente accettato che valori della probabilità a posteriori superiori a 0,80 identifichino aree a significativo rischio (Richardson 2004).

Infine, è stato applicato anche il modello della “*spatial scan statistic*” (SatScan), proposta da Kulldorff (Kulldorff 1995 e 1997), specifico per l’identificazione di “cluster” spaziali e della loro significatività statistica (in Appendice dettagli sul metodo). Nel caso specifico, sono state create delle finestre circolari di raggio variabile intorno al centroide di ciascuna cella di 4x4km e 2x2km e per ogni finestra è stata valutata l’eventuale presenza di un numero anomalo di casi osservati di sarcomi rispetto a quelli attesi sulla base dell’incidenza regionale.

Risultati

In Figura 3 è riportata la distribuzione geografica dei SIR dei sarcomi del periodo 2013-2018 nelle celle 4x4km della provincia pistoiese. La mappa mostra numerose celle con SIR in eccesso con valori maggiori di 1. In particolare, sono 13 le celle con valori del SIR maggiore di 1,4 e 4 quelle con valori compresi tra 1,0 e 1,4. Comunque, sono 26 quelle con valori inferiori ad 1, ovvero dove i casi osservati sono inferiori a quelli attesi in base all'incidenza media regionale. La rappresentazione grafica va però interpretata alla luce dell'effetto della bassa numerosità dei casi osservati in ogni cella e della conseguente instabilità delle stime statistiche. Il valore della significatività statistica delle stime di ogni cella, espressa dagli Intervalli di Confidenza al 90%, è riportata in forma tabellare (Tabella 1).

Figura 3. Distribuzione dei SIR di sarcomi nella provincia di Pistoia nel periodo 2013-2018 in ciascuna cella numerata progressivamente (ID) della griglia 4x4km.



Nella Tabella 1 per ciascuna cella è riportato il numero dei casi osservati e attesi, la stima dei SIR e gli IC al 90%. Si nota come all'aumentare dei casi osservati diminuisce l'incertezza statistica espressa dalla dimensione degli IC.

La cella 23, che include gran parte della frazione di Casalguidi, mostra un numero di osservati più che doppio degli attesi (SIR: 2.11) e IC sempre maggiori di 1 (da 1.25 a 3.36), ad indicare la presenza di un eccesso di casi di sarcoma "significativamente maggiore" rispetto a quanto atteso in base all'incidenza media regionale.

Nella cella 28, che include l'abitato di Massa e Cozzile, il SIR è ancora più elevato (SIR: 6.99) e statisticamente significativo, ma essendo basato solo su 3 casi osservati, gli IC sono molto ampi (da 1.9 a 18.03) a testimonianza di una forte instabilità della stima.

La cella 12, che include frazioni della zona sud del comune di Monsummano Terme, presenta un SIR di 1.71, quindi un eccesso del 71% rispetto all'atteso, con IC tra 0.9 e 2.9, al limite della significatività statistica.

Per le altre celle, le stime dei SIR seppure molto elevati (celle scure della mappa della Figura 3) o molto bassi (celle chiare), si accompagnano a IC molto ampi e tali da non poter escludere un puro effetto del caso. In nessuna cella si rileva un valore statisticamente significativo inferiore al valore atteso in base al tasso medio regionale.

Anche con la griglia geografica 2x2 (dati non mostrati), si confermano gli eccessi significativi nella cella di Casalguidi (SIR: 2,51; IC: 1.3-4.5) e di Massa e Cozzile (SIR: 6.2; IC: 1.7-16.0) e risulta più evidente l'eccesso nella zona sud di Monsummano Terme (SIR: 2.61; IC: 1.0-5.5), sostenuto da 5 casi osservati a fronte di circa 2 casi attesi. In una cella che interessa il comune di Lamporecchio compare un eccesso significativo (SIR: 4.2; IC 1.1-10.7) che non era visibile nella griglia 4x4Km, anche se con ampia oscillazione degli IC.

Tabella 1. Casi di sarcomi osservati, attesi, SIR e IC al 90% per ogni cella della griglia 4x4km dei 161 casi di sarcoma nella provincia di Pistoia nel 2013-2018.

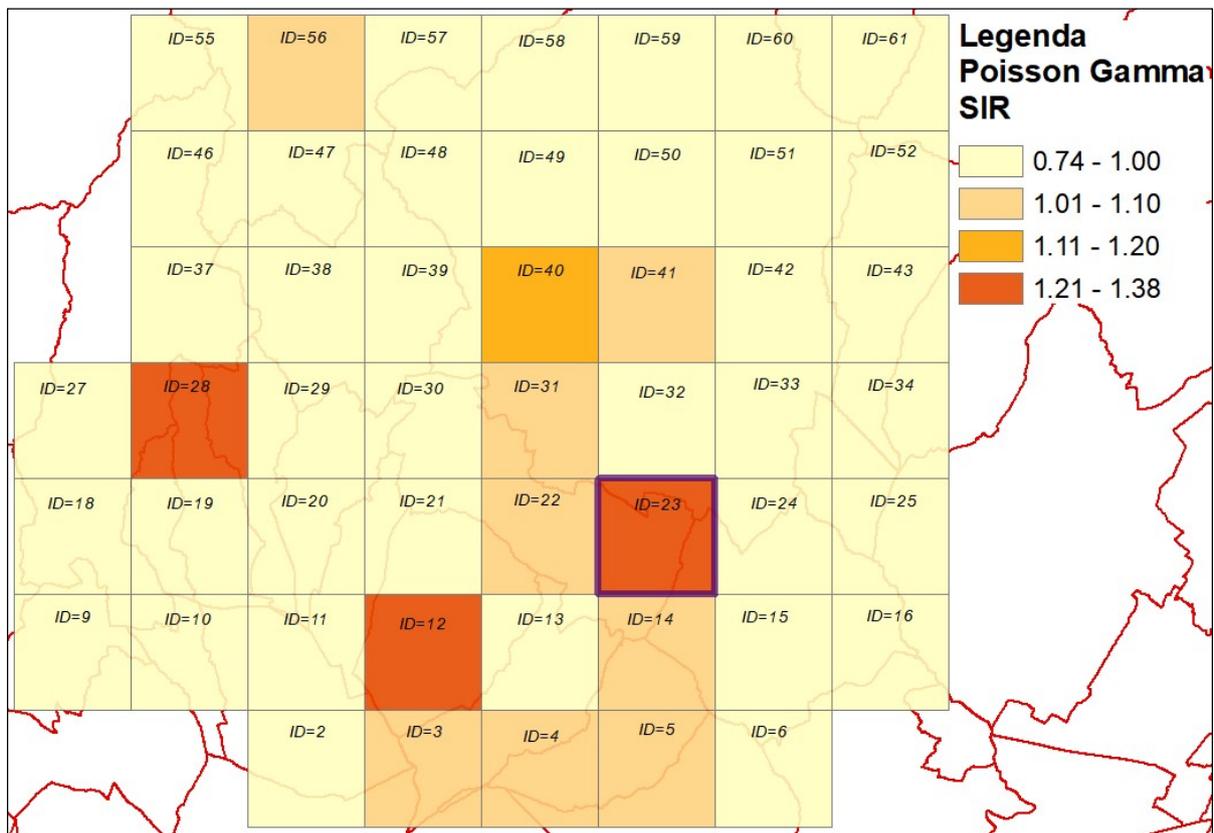
ID cella	Osservati	Attesi	SIR	IC 90% inferiore	IC 90% superiore
2	<3	0.7	1.42	0.07	6.78
3	<3	0.5	2.13	0.11	10.09
4	6	4.3	1.41	0.61	2.78
5	3	1.6	1.90	0.52	4.91
6	0	0.0	0	-	-
9	0	0.2	0	-	-
10	4	4.6	0.88	0.30	2.00
11	4	3.7	1.09	0.37	2.49
12	10	5.8	1.71	0.93	2.90
13	<3	1.8	1.09	0.19	3.42
14	<3	1.3	1.61	0.28	5.04
15	3	6.9	0.43	0.12	1.12
16	0	2.2	0	-	-
18	7	7.4	0.95	0.44	1.78
19	7	7.8	0.90	0.42	1.68
20	12	18.2	0.66	0.38	1.07
21	9	9.3	0.96	0.50	1.68
22	<3	0.1	8.25	0.43	39.53
23	13	6.2	2.11	1.25	3.36
24	3	6.4	0.47	0.13	1.21
25	<3	2.6	0.77	0.14	2.41
27	0	2.7	0	-	-
28	3	0.4	6.99	1.90	18.03
29	0	1.0	0	-	-
30	<3	1.3	0.79	0.04	3.76
31	6	4.1	1.46	0.63	2.87
32	9	14.4	0.63	0.33	1.09
33	3	4.5	0.67	0.18	1.73
34	8	9.1	0.88	0.44	1.59
37	0	0.6	0	-	-
38	0	0.5	0	-	-
39	0	0.9	0	-	-
40	6	3.2	1.89	0.82	3.74
41	24	22.2	1.08	0.74	1.52
42	<3	2.1	0.95	0.17	3.00

43	3	5.3	0.57	0.15	1.47
46	0	0.2	0.00	0.00	10.96
47	0	0.2	0.00	0.00	10.01
48	<3	0.6	1.61	0.08	7.65
49	3	2.7	1.11	0.30	2.87
50	0	1.3	0.00	0.00	1.78
51	<3	1.1	0.93	0.05	4.39
52	0	0.0	0	-	-
55	0	0.0	0	-	-
56	<3	0.4	2.65	0.13	12.48
57	0	0.5	0	-	-
58	0	1.0	0	-	-
59	0	0.4	0	-	-
60	0	0.0	0	-	-
61	0	0.0	0	-	-

In Figura 4 è riportata la mappa dei SIR calcolati applicando il modello bayesiano PG. Come si evince dal confronto delle mappe, il “lisciamento” ha prodotto una contrazione dei valori dei rischi, con i SIR che nella Figura 3 variavano nel range 0-8.35 e in Figura 4 nel range 0.74-1.38. Questo effetto, detto “shrinkage”, è conseguenza del modello bayesiano, che comporta uno smussamento dei valori di SIR più instabili verso la media generale.

La Figura 4 conferma che i valori di SIR più alti si osservano nelle celle 23, 12 e 28, ossia la frazione di Casalguidi, le frazioni del comune di Monsummano Terme e di Massa e Cozzile. I risultati sono confermati anche nell’analisi di PG con la griglia 2x2km (dati non mostrati).

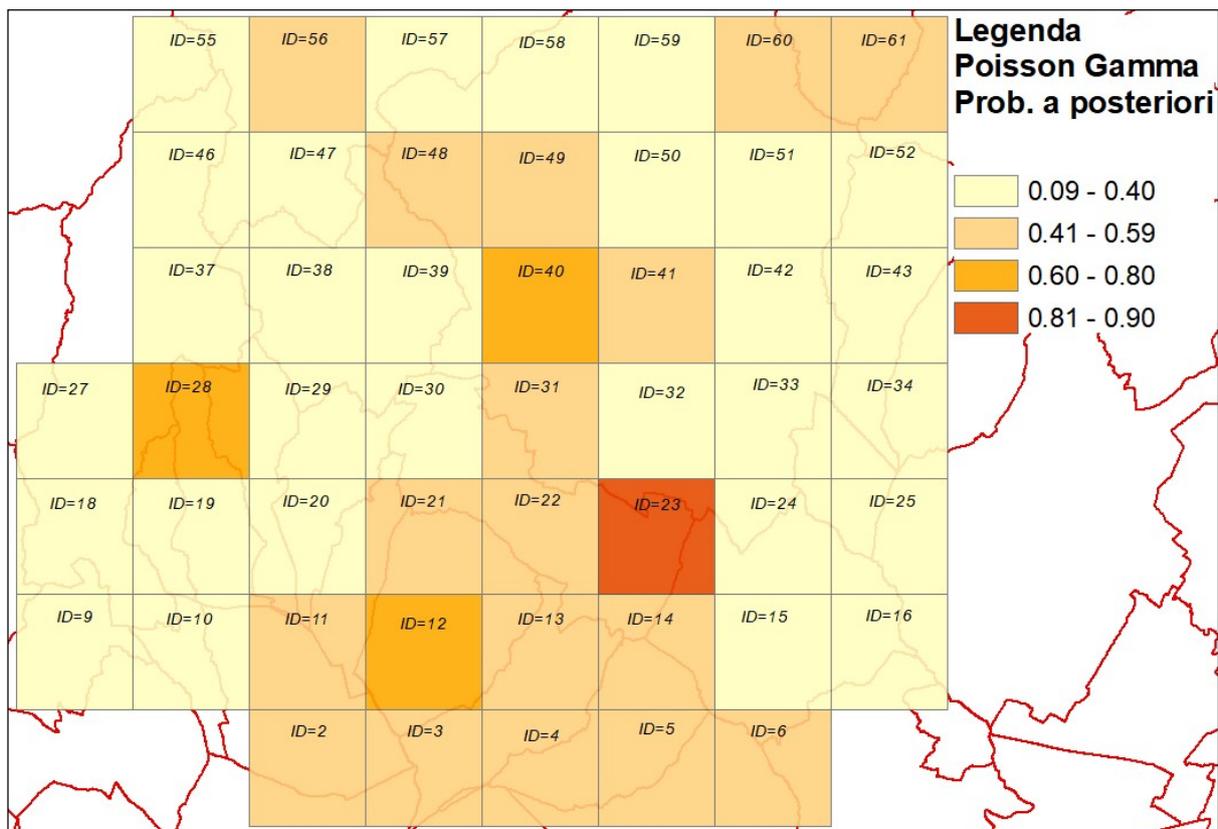
Figura 4: Mappa delle stime bayesiane dei SIR di sarcomi del periodo 2013-2018 ottenuti con il metodo Poisson-Gamma in aree geografiche di 4x4km della provincia di Pistoia.



La Figura 5 mostra la mappa della probabilità a posteriori che in ciascuna cella il SIR sia maggiore di uno in base al modello di PG, ovvero quanto l'eccesso del SIR sia statisticamente solido. Ancora una volta l'area con valore in eccesso di sarcomi più significativo rispetto alla media regionale è la cella di Casalguidi (ID=23; probabilità a posteriori pari a 0.90). Gli eccessi dei SIR delle celle delle zone di Monsummano Terme (ID=12) e Massa e Cozzile (ID=28) sono invece meno significativi (probabilità a posteriori inferiore ad 81).

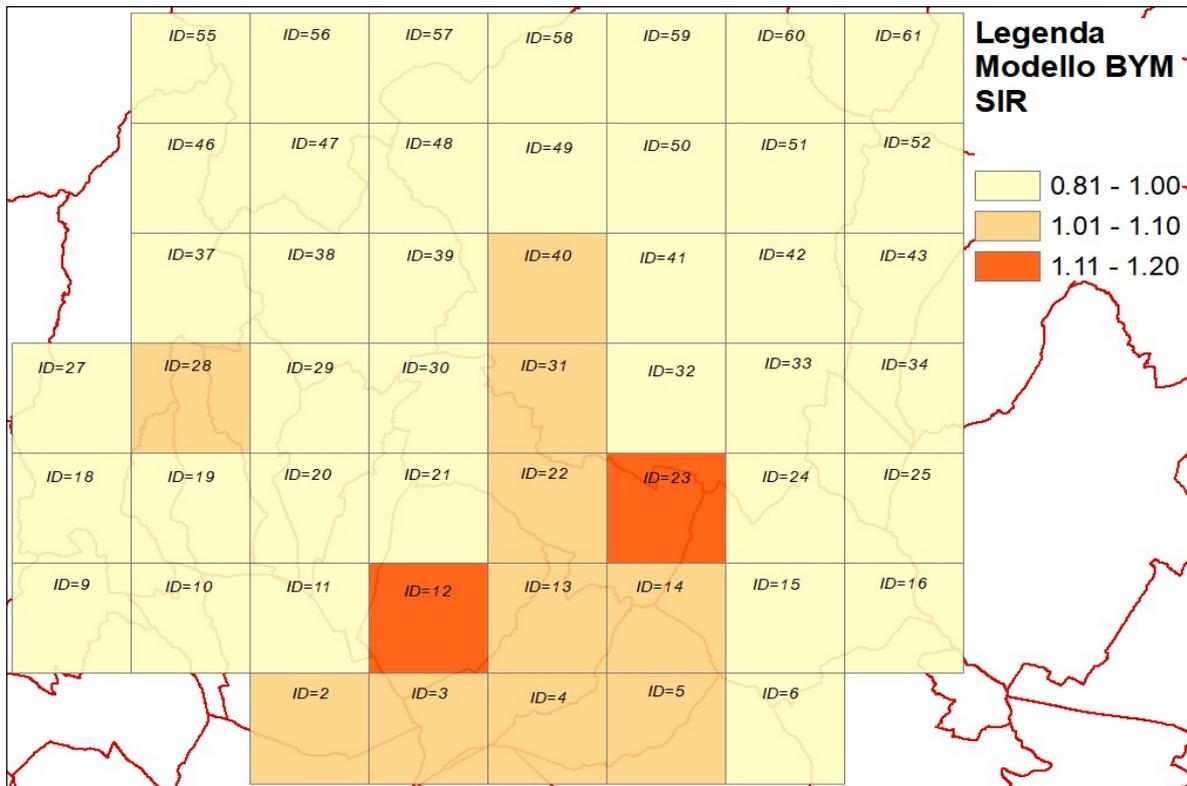
Anche con la griglia 2x2km (dati non mostrati) si conferma il rischio per la cella di Casalguidi, anche se con una probabilità a posteriori meno significativa di quella rilevata nella 4x4km (0,81).

Figura 5: Mappa delle probabilità a posteriori ottenute con il metodo Poisson-Gamma, che in ciascuna cella della griglia 4x4km il SIR per i sarcomi incidenti nella Provincia di Pistoia dal 2013 al 2018 sia maggiore di uno.



La Figura 6 mostra i risultati dei SIR ottenuti con il modello BYM. Si nota come si riduce ulteriormente la variabilità dei SIR, compresi nel range ancora più stretto 0.81-1.20. Anche con questo metodo si conferma il SIR più elevato nella cella di Casalguidi (ID= 23; SIR=1.20) e di Monsummano Terme (ID=12; SIR=1.14), mentre rispetto al modello PG si attenua fortemente il valore di quello di Massa e Cozzile.

Figura 6: Mappa delle stime bayesiane dei SIR di sarcomi del periodo 2013-2018 ottenuti con il modello Besag, York, Mollie in aree geografiche di 4x4km della provincia di Pistoia



Per le due aree con i SIR più elevati di Casalguidi (ID=23) e Monsummano Terme (ID=12), le probabilità a posteriori ottenute con il modello BYM (Figura 7) sono più basse di quelle del modello PG, cioè con rischio meno significativo, con valori rispettivamente di 0.71 per la cella di Casalguidi e 0.63 per la cella delle frazioni di Monsummano Terme e decisamente ancora meno significativa per quella di Massa e Cozzile (inferiore a 0,61).

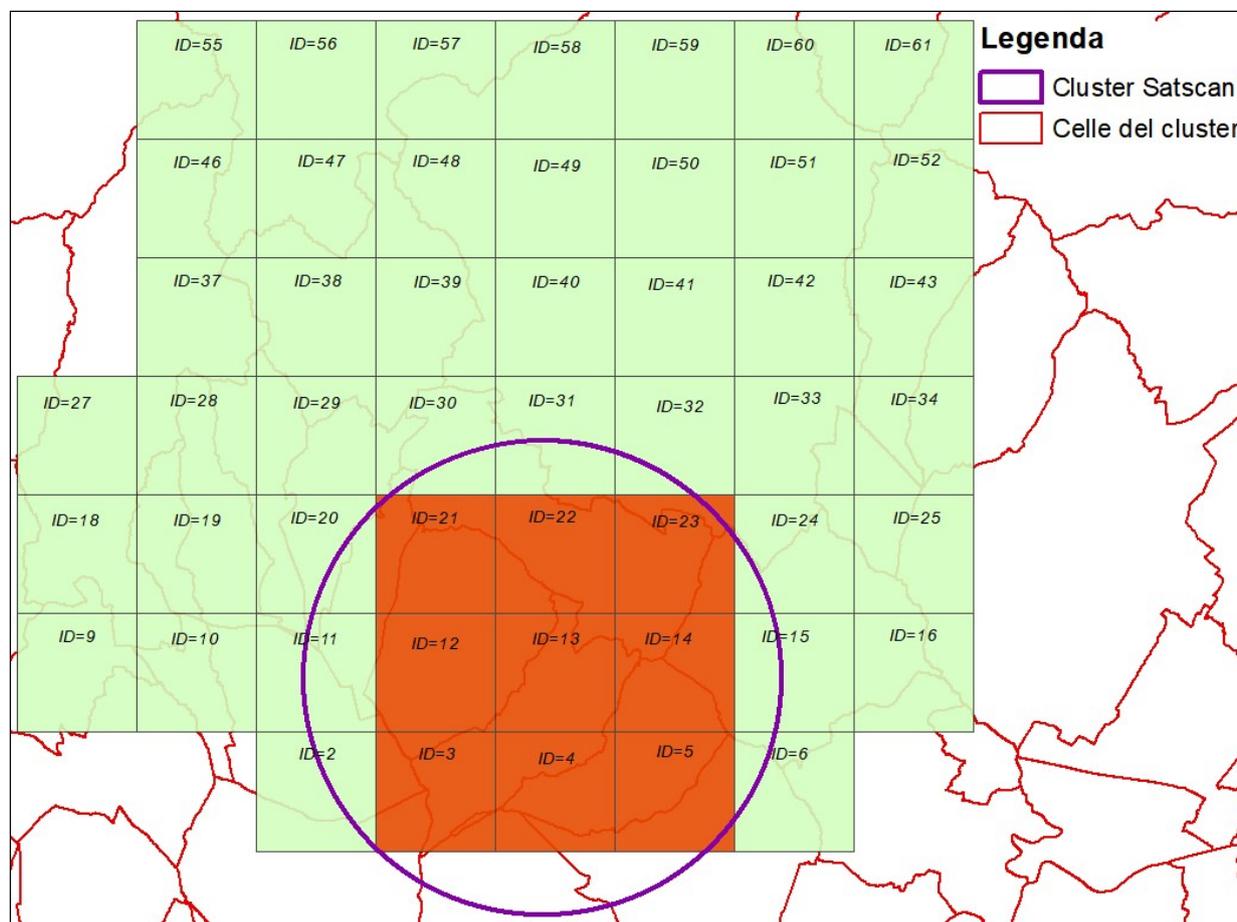
Figura 7: Mappa delle probabilità a posteriori ottenute con il modello Besag, York, Mollie, che in ciascuna cella della griglia 4x4km il SIR stimato sia maggiore di uno.



Risultati analoghi sono stati ottenuti per i SIR con il modello BYM applicato alla griglia geografica 2x2km (dati non mostrati). Da segnalare però che con questa grigliatura compare un SIR più elevato nell'area di Lamporecchio, che non si evidenziava nella mappa 4x4Km, e diminuisce la significatività del SIR dell'area di Casalguidi (probabilità a posteriori: 0.66).

In alternativa al calcolo dei SIR per le celle delle grigliature 4x4km e 2x2km, ed ai modelli bayesiani PG e BYM, è stato anche utilizzato il modello di analisi statistiche per aggregazioni geografiche "spatial scan statistic" (SatScan). La Figura 8 mostra il luogo dove SatScan ha individuato un cluster di sarcomi statisticamente significativo ($p=0.025$). Si tratta di un'area piuttosto ampia, che include 9 delle celle della griglia 4x4km, nell'area sud pistoiese e che include gli abitati della frazione di Casalguidi e dei comuni di Monsummano Terme, Larciano e Lamporecchio.

Figura 8: Cluster individuato dalla *spatial scan statistic* (SatScan) relativo all'incidenza di sarcomi del periodo 2013-2018 nella provincia di Pistoia.



Ripetendo l'analisi SatScan con le celle della griglia 2x2km (immagine non riportata), si conferma il cluster spaziale nella stessa zona sud e con il coinvolgimento degli stessi abitati.

Conclusioni

L'analisi della distribuzione spaziale dei casi di sarcomi nel territorio della provincia pistoiese ha evidenziato le difficoltà e i rischi di interpretazione dei risultati relativi a piccole aree geografiche, con scarsa numerosità delle popolazioni residenti e delle casistiche sanitarie. L'analisi statistica con i metodi classici, come quelli che producono i SIR, senza tenere conto dell'incertezza associata alla bassa numerosità della casistica, può comportare una lettura distorta del fenomeno e portare a conclusioni fuorvianti. In questo studio l'applicazione di metodologie avanzate per lo studio di eventi in piccole aree ha consentito di rimuovere o attenuare il "rumore" generato dall'instabilità statistica collegata alla scarsa numerosità delle osservazioni.

L'analisi tradizionale dei SIR ha evidenziato un eccesso significativo di incidenza dei sarcomi nella cella della griglia 4x4km che corrisponde alla frazione di Casalguidi, con un SIR di 2,1 (IC 90%: 1,25-3,36). L'eccesso è confermato anche con la griglia 2x2km ed anche dopo aggiustamento con i modelli più conservativi di PG e BYM. Anche nell'area sud di Monsummano Terme è emerso un eccesso nella griglia 4x4km, anche se al limite della significatività (SIR: 1,71; IC90%: 0,93-2,90), confermato nella griglia 2x2km e dopo aggiustamento con il modello di BYM e, sebbene con minor evidenza, anche nel modello PG. Una cella di un'area del comune di Massa e Cozzile presenta un SIR in eccesso nella griglia 4x4km (SIR: 7,0; IC 90%: 1,9-18,0) ed anche in quella 2x2km, ma decisamente meno evidente con i modelli PG e BYM. Infine, l'analisi SatScan, sia con grigliatura 4x4km che 2x2km, ha identificato un cluster che include la frazione di Casalguidi e aree di Monsummano Terme, Larciano e Lamporecchio.

In sintesi, i diversi metodi statistici applicati convergono nel confermare un cluster nella frazione di Casalguidi, dove il numero di sarcomi osservati nel periodo 2013-2018 è certamente in eccesso significativo rispetto ai valori regionali. Sebbene meno evidenti, altre aggregazioni anomale emergono nell'area sud di Monsummano Terme, Lamporecchio, Larciano e, con ancora minor evidenza, a Massa e Cozzile, da verificare con la disponibilità di dati aggiornati del Registro Toscano dei Tumori di ISPRO.

APPENDICE

Codici utilizzati per l'identificazione dei casi di sarcomi.

Per il periodo 2013-2015 i casi sono stati identificati in base ai codici:

- ICD-O-3 Topografia: C40.0-41.9
- ICD-O-3 Morfologia: 8710-8711, 8800-8936, 8959, 8963-8964, 8990-8991, 9020, 9040-9044, 9120-9133, 9150, 9170, 9180, 9231, 9240, 9251, 9260, 9364-9365, 9540, 9560-9571, 9581

Per il periodo 2016-2018 sono stati utilizzati i codici SNOMED:

"D1-50460";"M-80333";"M-87103";"M-88003";"M-88013";"M-88023";"M-88033";"M-88043";
"M-88103";"M-88113";"M-88123";"M-88133";"M-88143";"M-88323";"M-88333";"M-88403";
"M-88503";"M-88513";"M-88523";"M-88533";"M-88543";"M-88553";"M-88583";"M-88903";
"M-88913";"M-88943";"M-88953";"M-88963";"M-89003";"M-89013";"M-89023";"M-89103";
"M-89203";"M-89303";"M-89333";"M-89603";"M-89633";"M-89643";"M-89803";"M-89813";
"M-89903";"M-89913";"M-90200";"M-90201";"M-90203";"M-90403";"M-90413";"M-90423";
"M-90443";"M-91203";"M-91243";"M-91303";"M-91403";"M-91703";"M-91803";"M-91813";
"M-91823";"M-91833";"M-91843";"M-91853";"M-91903";"M-92203";"M-92213";"M-92313";
"M-92403";"M-92503";"M-92603";"M-92703";"M-92903";"M-93303";"M-94423";"M-94713";
"M-94803";"M-94813";"M-95303";"M-95403";"M-95603";"M-95813";"M-95923";"M-95933";
"M-96623";"M-96843";"M-97403";"M-98503";"M-99303";"M-89363".

Rapporto Standardizzato di Incidenza

Il Rapporto Standardizzato di Incidenza (SIR) è il rapporto tra il numero di casi osservati e il numero di casi attesi nell'area in studio ($SIR_i = Y_i/E_i$, dove Y_i sono i casi osservati, E_i i casi attesi e i le aree in studio). È una misura sintetica che stima il rischio relativo di occorrenza di una patologia in un'area in studio (Breslow e Day, 1987). Il SIR è il rapporto tra il numero di eventi osservati e il numero di eventi attesi nell'area in studio ed esprime l'eccesso (SIR maggiore di 1) o il difetto (SIR minore di 1) di morbosità/incidenza esistente tra la popolazione dell'area in studio e la popolazione presa come riferimento, nel nostro caso rappresentata dai residenti in Toscana, al netto delle variabili di confondimento considerate. Essendo uno stimatore basato su un rapporto, il SIR assume dei valori estremamente alti in presenza di eventi attesi molto bassi (vicino allo zero), anche con eventi osservati pari a uno o di poco superiori.

Intervallo di confidenza del SIR

L'intervallo di confidenza rappresenta un intervallo di valori entro i quali possiamo essere confidenti che i dati di una certa popolazione contengono il vero valore sconosciuto. L'intervallo di confidenza (IC) indica la forza dell'evidenza. Se gli intervalli sono ampi la stima è meno precisa. In questo studio l'intervallo di confidenza del SIR è calcolato mediante il metodo approssimato di Byar (Rothman 1979).

$$SIR_L = SIR \left(1 - \frac{1}{9O} - \frac{Z_\alpha}{3\sqrt{O}} \right)$$
$$SIR_U = SIR \left(\frac{O+1}{O} \right) \left(1 - \frac{1}{9(O+1)} + \frac{Z_\alpha}{3\sqrt{O+1}} \right)$$

SIR_L = limite inferiore dell'intervallo di confidenza del SIR

SIR_U = limite superiore dell'intervallo di confidenza del SIR

O = Casi osservati

E = Casi attesi

$Z_{\alpha/2}$ = Valore della distribuzione normale per un dato livello di significatività (α). Per gli intervalli al 90% $Z=1.64$

Modello Poisson-Gamma

Il modello di Poisson-Gamma è usato per filtrare la variabilità campionaria e “stabilizzare” il rischio relativo verso la media generale.

$$\begin{aligned} Y_i &\propto \text{Poisson} (\mathcal{G}_i E_i) \\ \mathcal{G}_i &\propto \text{Gamma} (\alpha, \nu) \\ \Theta | Y &\propto \text{Gamma} (E_i + \alpha, Y_i + \nu) \\ \hat{\mathcal{G}}_i &= \frac{Y_i + \nu}{E_i + \alpha} = w_i \frac{Y_i}{E_i} + (1 - w_i) \frac{\nu}{\alpha} \\ w_i &= \frac{E_i}{E_i + \alpha} \end{aligned}$$

Clayton e Kaldor (1987) definiscono un modello bayesiano assumendo che i casi osservati si distribuiscano, condizionatamente ai rischi relativi, come una variabile di Poisson di valore atteso $E_i \theta_i$ per $i=1, \dots, n$. In questo modello si assume che i rischi relativi siano delle variabili indipendenti e identicamente distribuite secondo una distribuzione Gamma, con parametro inverso di scala α e parametro di forma ν . La stima del rischio relativo è data dal valore atteso della distribuzione a posteriori dei rischi relativi. La distribuzione per i rischi relativi, condizionata alle osservazioni, è ancora una volta una Gamma con parametro inverso di scala $E_i + \alpha$ e il parametro di forma $Y_i + \nu$. Le stime bayesiane hanno una dispersione minore rispetto a quelle ottenute tramite l'indice SIR: ogni stima è un compromesso tra quest'ultimo e la stima della media assunta per i rischi relativi, cioè tra Y_i/E_i e le stime di ν/α . La formula del rischio relativo stimato evidenzia come lo stimatore sia una media pesata tra l'indice SIR e la media della distribuzione a priori. Il peso w_i dipende dal numero di casi attesi. Nel caso le stime siano basate su un numero elevato di attesi, esse saranno vicine alla stima SIR dato che il peso sarà prossimo a uno.

Modello Besag, York, Mollie

È un modello di convoluzione Gaussiana che “stabilizza” la stima del rischio relativo vs la media generale e verso una media locale.

$$\begin{aligned} Y_i &\propto \text{Poisson} (\mathcal{G}_i E_i) \\ \log(\mathcal{G}_i) &= \mu_i + u_i + v_i \\ u_i &\propto N(\mu_u, \sigma_u) \\ v_i | v_{j \in \text{adj}_i} &\propto N(\bar{u}_i, \sigma_v / \# \text{adj}_i) \end{aligned}$$

Besag, York e Mollie (Besag 1991) suggeriscono un approccio bayesiano basato su un modello gerarchico a tre stadi. Le stime sono ottenute combinando due tipi di informazione: quella relativa ai casi osservati descritta tramite la verosimiglianza poissoniana e l'informazione data dalla distribuzione a priori sui rischi relativi. Il rischio relativo, quindi, viene modellato come somma di due effetti casuali: uno tiene conto della sovradisersione non spaziale nei dati (u_i : componente di

eterogeneità). Per tale componente viene assunta una distribuzione a priori normale $(0, \sigma_u)$ dove σ_u è il parametro di precisione modellato come una Gamma. Il secondo effetto casuale (clustering) viene modellato attraverso una distribuzione a priori di tipo ICAR (intrinsic conditional autoregressive) facendo in modo che il valore del rischio stimato in ciascuna area sia funzione dei valori delle aree adiacenti. In altre parole, indicando S_i come l'insieme delle aree adiacenti all' i -esima, $v_i | v_{j \in S_i}$ si distribuisce come una Normale dove \bar{v} è la media in termini di aree adiacenti all' i -esima e σ_v è il parametro di precisione che dipende da n_i . Per questo studio si è deciso di assegnare a priori un uguale peso alla componente di eterogeneità e a quella di clustering.

Spatial Scan Statistic (SatScan)

Si consideri un territorio suddiviso in sezioni geografiche minori, chiamate celle, per ognuna delle quali sono note le coordinate geografiche del centroide (baricentro geografico), il numero di casi osservati e la popolazione di riferimento (o gli attesi). La spatial scan statistic (Kulldorff, 1995 e 1997) opera su un insieme di finestre circolari, il cui raggio varia da 0 ad un limite superiore definito a priori ed espresso come proporzione massima della popolazione totale da includere nel cluster finale. 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. 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. Ad ogni variazione del raggio R si definisce una zona z per ciascun centroide. Ogni dimensione, forma e localizzazione geografica di una finestra definisce una zona z che identifica un cluster potenziale del fenomeno. Definito, quindi, 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 $i = 1, \dots, I$ 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 SatScan, l'ipotesi iniziale da valutare è di distribuzione uniforme del fenomeno sul territorio o, equivalentemente, di uguaglianza delle probabilità p e q , $H_0 : p = q$. 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. L'aspetto innovativo della SatScan 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. Oltre alla finestra circolare è possibile applicare il test di "cluster detection" utilizzando una finestra di forma ellittica.

Bibliografia

1. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math*, 1991;43,1-59.
2. Breslow NE, Day NE. *Statistical Methods in Cancer Research. Vol. II, The Design and Analysis of Cohort Studies (IARC Scientific Publication No. 82)*. Lyon, France: International Agency for Research on Cancer, 1987.
3. Clayton, D, Kaldor, J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 1987; 43, 671-81.
4. Kulldorff M, Nagarwalla N. Spatial disease clusters: Detection and Inference. *Statistics in Medicine*, 1995; 14:799-810.
5. Kulldorff M. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 1997; 26:1481-1496.
6. Lance A. Waller, Bradley P. Carlin. Disease mapping. *Chapman Hall CRC Handb Mod Stat Methods*. Author manuscript; available in PMC 2014 Oct 1. Published in final edited form as: *Chapman Hall CRC Handb Mod Stat Methods*. 2010.
7. Richardson S, Thomson A, Best N, Elliott P. Interpreting posterior relative risk estimates in disease mapping studies. *Environ Health Perspect*, 2004; 112: 1016-25.
8. Rothman KJ, Boice JD, Jr. *Epidemiologic Analysis with a Programmable Calculator (NIH Publication 79-1649)*. Washington DC: US Government Printing Office, 1987.