Studi in ricordo di Carlo Cecchi (2015) ISBN: 978-88-88793-67-2
Università degli Studi di Bari Aldo Moro 15-24

Applicazione di un modello bayesiano per la mappatura del rischio relativo in un'area ad elevato rischio ambientale*

Nicola Bartolomeo[1], Monica Carbonara[2], Gabriella Serio[1], Antonella Mincuzzi[3], Aldo Minerba[3], Paolo Trerotoli[1]

[1] *Dipartimento di Scienze Biomediche ed Oncologia Umana, Università di Bari Aldo Moro*
[2] *ISTAT, Ufficio territoriale per la Puglia*
[3] *ASL Taranto*

Riassunto: L'applicazione del modello di Besag-York-Mollié consente di stimare indici di distribuzione geografica delle patologie e addolcirne le variazioni (smoothing) anche quando il numero di casi osservati è esiguo. Gli indici tradizionali di autocorrelazione spaziale hanno messo in luce l'effetto del modello scelto per valutare i cluster di patologia.

Keywords: disease mapping, modelli bayesiani, autocorrelazione spaziale, smoothing spaziale,.

1. Obiettivi

Il comune di Taranto, con i comuni di Statte, Crispiano, Massafra e Montemesola, dal 1987 fa parte di un'area definita dall'Organizzazione Mondiale della sanità (OMS) "a elevato rischio ambientale" ed è da qualche tempo oggetto di studio per

* Corresponding Author: Nicola Bartolomeo, nicola.bartolomeo@uniba.it, Dipartimento di Scienze Biomediche ed Oncologia Umana, Università degli Studi di Bari Aldo Moro.
Attribuzioni del lavoro: Bartolomeo N.: analisi dei dati, pianificazione dell'indagine, stesura dei risultati; Carbonara M.: analisi dei dati sull'indice di deprivazione, contributo alla pianificazione; Serio G.; pianificazione dell'indagine, contributo alla discussione dei risultati; Mincuzzi A.: gestione dei dati della città e provincia di Taranto, contributo alla pianificazione dell'indagine; Minerba A.: pianificazione dell'indagine, gestione dei dati della città e provincia di Taranto; Trerotoli P.: pianificazione dell'indagine, discussione dei risultati, contributo all'analisi dei dati.

la stima del rischio di salute conseguente all'esposizione ambientale, dovuta alla presenza del più grande stabilimento siderurgico a ciclo integrato d'Europa.

Il monitoraggio dell'occorrenza delle patologie è utile alla verifica dell'impatto del rischio ambientale sulla popolazione residente; risulta particolarmente adatta la mappatura della distribuzione geografica delle patologie per valutarne la presenza nelle aree più vicine agli impianti industriali. La mappatura per piccole aree produce stime instabili perché basate su un numero basso di eventi e, inoltre, può accadere che in aree adiacenti si possano registrare un numero di eventi molto differenti. Questa seconda problematica impone la necessità di rendere meno brusche le differenze tra aree adiacenti utilizzando dei metodi di smoothing degli indicatori da utilizzare sulla mappa.

Lo studio si prefigge di valutare gli effetti dell'applicazione di un modello di analisi Bayesiana per la stima dei rischi e realizzare, mediante un sistema informativo Gis, una mappatura degli stessi per sezione di censimento del comune di Taranto, al fine di individuare eventuali cluster di patologia.

2. Materiali e metodi

Le fonti informative utilizzate sono state:

1. Anagrafe sanitaria degli assistiti della provincia di Taranto 2000-2010;
2. Anagrafe comunale e Archivio delle sezioni di censimento del comune di Taranto;
3. I dati del Censimento generale della popolazione e delle abitazioni del 2001 utilizzati per la costruzione dell'indice di deprivazione;
4. Registro delle Cause di Morte dei residenti di Taranto e della Puglia (ReNCaM) del periodo 2006-2008.

Al fine di valutare l'applicazione del modello ai dati disponibili, sono stati esaminati:

- i decessi dei residenti nella città di Taranto con causa principale Tumore del Polmone (codice ICD9 162.*) negli anni 2006-2008, tenendo presente che si tratta di una patologia il cui legame con l'esposizione ambientale è ampiamente documentato in letteratura;
- i decessi dei residenti nella città di Taranto con causa principale Malattie Croniche del Fegato e Cirrosi Epatica (codice ICD9 571.*) negli anni 2006-2008, patologia non strettamente correlata all'esposizione ambientale.

Allo scopo di attribuire sia i decessi sia la popolazione alla propria sezione di censimento, è stata eseguita, per ogni anno, la georeferenziazione di ciascun residente presente nell'anagrafe sanitaria degli assistiti sulla base dell'indirizzo di residenza. L'attribuzione della sezione di censimento ai casi è stata ottenuta con una procedura di "Record Linkage" (RL), misto deterministico e probabilistico, tra l'anagrafe degli assistiti e il ReNCaM.

Nelle analisi di salute pubblica, al fine di evidenziare quanto una popolazione (area) sperimenta l'evento in studio (mortalità) in più (o in meno) rispetto ad una popolazione standard al netto dell'effetto di fattori quali, ad esempio, sesso ed età, viene spesso utilizzato il rapporto standardizzato di mortalità (SMR).

L'SMR, che è una stima del rischio relativo, è calcolato come rapporto tra il numero di casi osservati (O) in una popolazione in esame in un dato arco temporale, ed il numero di casi attesi (E) nella popolazione se questa avesse lo stesso tasso di incidenza, età e sesso specifico, di una popolazione di riferimento usata come standard. Nella nostra analisi gli SMR sono standardizzati utilizzando come popolazione di riferimento quella della regione Puglia.

Poiché la rappresentazione in mappe tematiche dell'SMR presenta alcuni svantaggi, tra cui quello di non riuscire a cogliere la struttura spaziale di un fenomeno, è stato utilizzato il modello bayesiano proposto da Besag, York e Molliè (BYM) che permette di ottenere stime lisciate del rischio relativo. Il modello formulato è il seguente:

$$\log \mu_i = \log E_i + \gamma_0 + \gamma_1 \text{depriv}_i + b_i + h_i$$
$$O_i \sim \text{Poisson}(\mu_i)$$
$$\Gamma_1 \sim N(0; 0.00001)$$

dove:

- O_i è il numero di decessi osservati nella sezione di censimento i-esima;
- E_i è il numero di decessi attesi nella sezione di censimento i-esima, ottenuti da un modello di regressione logistica in cui la variabile dipendente è il logit dell'odd dei casi e quelle indipendenti sono il sesso e la classe d'età;
- depriv_i è l'indice di deprivazione della sezione di censimento i-esima;
- b_i è la componente strutturata degli effetti spaziali random attribuita *a priori* secondo il modello Conditional AutoRegressive (CAR);
  - car.normal(adj[], weights[], num[],τb)
    adj: matrice di adiacenza
    weights: pesi attribuiti alla matrice di adiacenza in base alle sezioni di censimento viciniori

num: numero di comuni adiacenti a ciascun sezione di censimento

τb: è la varianza di $\gamma_1 \sim \Gamma(\alpha; \beta)$

- h_i è una seconda componente non strutturata degli effetti spaziali random per la quale si assume una distribuzione *a priori* Normale (0; τh).

La componente strutturata b_i è stata costruita considerando come "circostanti" all'area i-esima quelle presenti in un raggio di 500 metri dal confine della stessa. Lo scopo è legato alla necessità di rendere meno brusche le variazioni del rischio relativo (RR) tra aree vicine, ipotizzando che il semplice confine amministrativo non possa essere la causa di un'evidente variazione spaziale del rischio di ammalarsi o di morire.

Il rischio relativo è stato così determinato utilizzando la seguente formula:

$$RR = \exp(\gamma_0 + \gamma_1 \, depriv_i + b_i + h_i)$$

La distribuzione geografica del rischio relativo così calcolato è stata posta a confronto con la rappresentazione dell'SMR sia graficamente, sia utilizzando indici globali di autocorrelazione spaziale. La misura globale di autocorrelazione spaziale informa sulla possibilità di una distribuzione casuale di un indice rappresentato geograficamente, all'interno di una regione di interesse (clustering). In questo studio l'autocorrelazione spaziale è stata quantificata attraverso l'Indice di Moran e quello di Geary.

L'indice di deprivazione utilizzato per il modello BYM è quello aggiornato al Censimento della popolazione e delle abitazioni del 2001. Per il calcolo dell'indicatore sono state scelte cinque condizioni che concorrono operativamente a descrivere il concetto multidimensionale della deprivazione: basso livello d'istruzione, disoccupazione, mancato possesso dell'abitazione, famiglia monogenitoriale e alta densità abitativa. Gli indicatori elementari selezionati sono stati prima standardizzati, cioè espressi come scarti dalla media nella popolazione oggetto di misura e in seguito sommati. L'indice è stato costruito come somma di punteggi z che si riferiscono ai cinque indicatori semplici selezionati.

$$ID = \sum_{i=1}^{5} z_i \qquad z_{i=} \frac{x_i - «_{x_i}}{s_{x_i}}$$

dove:

- x_1: % di popolazione con istruzione pari o inferiore alla licenza elementare;
- x_2: % di popolazione attiva disoccupata o in cerca di prima occupazione;

- x_3: % di abitazioni occupate in affitto;
- x_4: % di famiglie monogenitoriali con figli dipendenti conviventi;
- x_5: densità abitativa (numero di occupanti per 100 m^2).

L'indice di deprivazione finale, a scala continua, è stato categorizzato per quintili relativi alla popolazione della provincia di Taranto, individuando cinque categorie: molto agiate, agiate, medie, deprivate, molto deprivate.

I dati sono stati elaborati utilizzando il software SAS 9.3. I RR, aggiustati tenendo conto delle aree circostanti, sono stati determinati usando il software Winbugs14, mentre le mappe sono state realizzate attraverso il «Quantum Gis», software open source per la georeferenziazione e la rappresentazione spaziale di eventi.

3. Risultati

I nodi del modello bayesiano BYM stimati a posteriori per entrambe le patologie dopo 10.000 iterazioni sono riportati nella Tabella 1. L'MC error (Monte Carlo standar error) è una stima della differenza tra la media dei valori campionati e la media a posteriori, pertanto valuta l'accuratezza computazionale delle stime a posteriori dei parametri. Come regola generale l'MC error dovrebbe essere inferiore al 5% della deviazione standard del parametro. Nel nostro caso è di poco superiore al 5% della deviazione standard per le intercette, mentre è inferiore al 5% per i coefficienti relativi all'indice di deprivazione. Media e mediana sono molto simili, indice di una funzione di densità gaussiana per entrambi i parametri ad effetti fissi.

Tabella 1. *Stime a posteriori dei parametri ad effetti fissi del modello BYM.*

Parametri	Media	Dev.St.	MC error	2,5%	Mediana	97,5%
γ_0 - Kpol	-0,168	0,105	0,007	-0,380	-0,168	0,033
γ_1 - Kpol	-0,008	0,021	<,001	-0,048	-0,007	0,032
γ_0 - Lcir	-0,013	0,143	0,009	-0,295	-0,013	0,299
γ_1 - Lcir	-0,006	0,026	<,001	-0,058	-0,005	0,043

Dal momento che non tutte le sezioni di censimento avevano una distribuzione a posteriori del RR di tipo gaussiano, per la loro rappresentazione geografica si è scelto il valore mediano. Nelle figure 1 e 3 sono rappresentati gli SMR, mentre nelle figure 2 e 4 i RR stimati con il modello bayesiano.

Figura 1. *Mortalità per Tumore al Polmone. SMR. Comune di Taranto. Anni 2006-2008.*

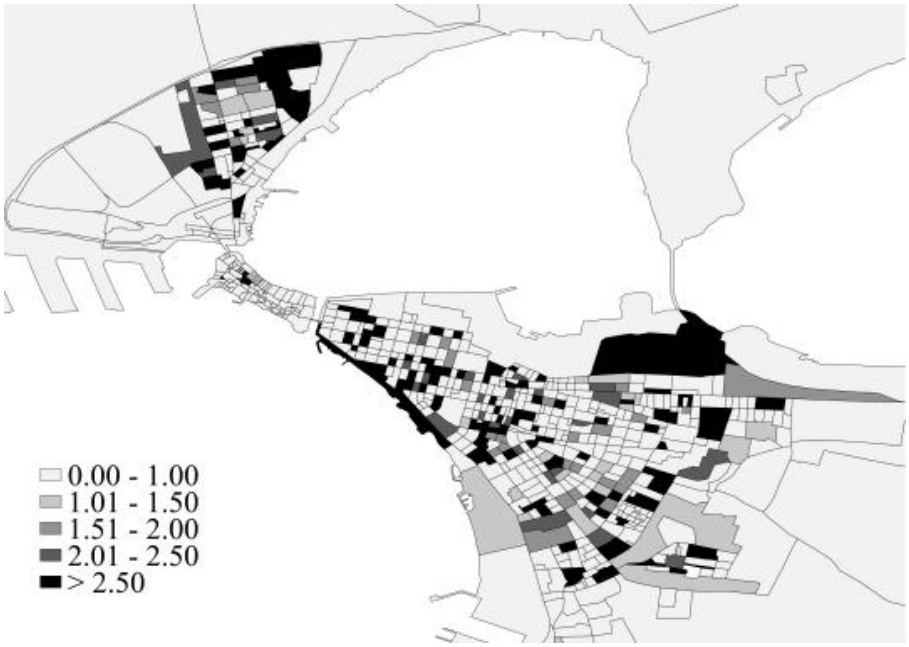
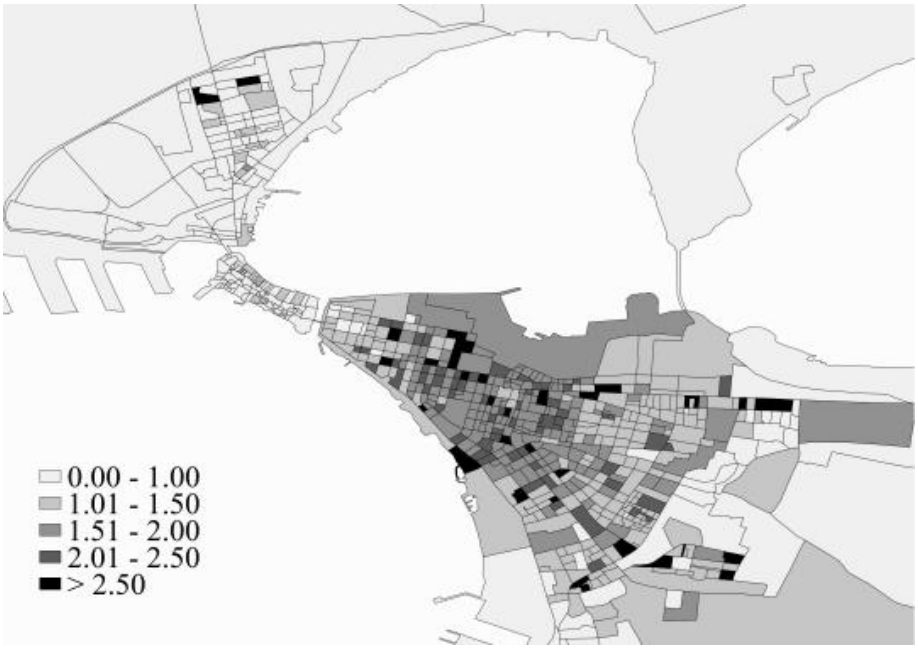


Figura 2. *Mortalità per Tumore al Polmone. RR. Comune di Taranto. Anni 2006-2008.*

Figura 3. *Mortalità per Malattie Croniche del Fegato e Cirrosi Epatica. SMR. Comune di Taranto. Anni 2006-2008.*

Figura 4. *Mortalità per Malattie Croniche del Fegato e Cirrosi Epatica. RR. Comune di Taranto. Anni 2006-2008*



Nelle tabelle 2 e 3 sono riportate le misure globali di autocorrelazione spaziale. Nel caso della rappresentazione della mortalità per Tumore del Polmone (Tab. 2), si può notare come gli indici di autocorrelazione determinati sugli SMR ed il p-value assumano valori che suggeriscono assenza di autocorrelazione, lasciando ipotizzare la non presenza di cluster di patologia. Se, invece, il calcolo degli indici di Moran e Geary viene effettuato sui rischi relativi determinati dopo lo smoothing attraverso il modello bayesiano, si osservano valori che suggeriscono la presenza di autocorrelazione spaziale statisticamente significativa. Lo smoothing ha, quindi, migliorato l'individuazione di aree che possano considerarsi aggregabili rispetto al rischio di morte per neoplasia polmonare.

Nel caso della mortalità per Malattie del Fegato e Cirrosi (Tab. 3), si può osservare che gli indici di Moran e Geary ancora una volta non consentono di ipotizzare aggregazioni di aree. Gli stessi indici, calcolati sui rischi dopo l'applicazione del modello baesiano, assumono valori che suggeriscono presenza di autocorrelazione spaziale e, essendo statisticamente significativi, consentono di concludere positivamente riguardo la presenza di cluster della patologia in oggetto.

Tabella 2. *Indici di autocorrelazione spaziale per il rischio di mortalità per Tumore al Polmone. Taranto. Anni 2006-2008.*

	Indice	Osservato	Dev.St.	Z	Pr>Z
SMR	Moran	<0,001	0,002	1,15	0,251
	Geary	0,367	0,446	-1,42	0,156
RR	Moran	**0,482**	0,006	74,9	**<0,001**
	Geary	**0,362**	0,034	-18,7	**<0,001**

Tabella 3. *Indici di autocorrelazione spaziale per il rischio di mortalità per Cirrosi Epatica. Taranto. Anni 2006-2008.*

	Indice	Osservato	Dev.St.	Z	Pr>Z
SMR	Moran	-0,007	0,007	-0,842	0,400
	Geary	0,,974	0,079	-0,323	0,746
RR	Moran	**0,437**	0,006	68,13	**<0,001**
	Geary	**0,591**	0,047	-8,63	**<0,001**

Considerazioni finali

Le analisi spaziali effettuate su maggiori livelli di dettaglio geografico migliorano l'interpretabilità dei risultati rispetto a studi su larga scala, presentano minore su-

scettibilità al bias ecologico e incrementano la capacità di individuare effetti associati a fattori ambientali (ad es. inquinamento ambientale). D'altro canto, la scelta di un livello di rappresentazione così dettagliato, pone il problema di stima dei rischi in presenza di aree con un ridotto numero di casi. Il modello Bayesiano consente la stima dei rischi per aree con un numero molto basso di eventi per i quali l'assunzione sulle distribuzioni non è sufficientemente supportata dalle frequenze osservate. Tale criticità è superata assumendo distribuzioni a priori sui parametri e, in particolare, sulle varianze. Il modello utilizzato, che comprende la componente non strutturata della variabilità spaziale rispetto al modello CAR senza la componente non strutturata, ha come ulteriore vantaggio quello di catturare la quota di variabilità residua. Questo si ottiene grazie al fatto di poter stabilire la distribuzione di questa componente e, quindi, di poter effettuare le stime relative.

L'inserimento nella componente strutturata di una matrice di pesi che tengano conto dell'influenza delle aree circostanti sul rischio relativo, rende meno brusche le variazioni di rischio tra tali aree, considerando che il semplice confine amministrativo non possa essere la causa di un'evidente variazione spaziale del rischio di ammalarsi o di morire.

Con il modello bayesiano sono, quindi, superati gli svantaggi della rappresentazione, in mappa tematiche, della semplice stima del rischio relativo:

- una potenziale instabilità della mappa che deriva dal fatto che la varianza del rischio è tanto più elevata in un'area quanto più è piccolo il numero di popolazione a rischio e, viceversa, tanto più piccola quanto più la numerosità della popolazione è alta;
- nessuna differenziazione tra le aree in assenza di casi;
- nessun tentativo di cogliere la struttura spaziale del fenomeno.

Nel modello BYM la scelta dei pesi in relazione alla distanza tra le aree è evidentemente un momento critico anche per i risultati da mappare e va valutata sulla base degli obiettivi. Nel nostro caso è stata scelta la distanza di 500 metri tenendo conto del livello di dettaglio geografico da rappresentare, cercando di evitare l'aggiustamento di un'area rispetto a quelle eccessivamente lontane.

Il numero di iterazioni ha un effetto sulla precisione delle stime dei parametri e sulle risorse computazionali necessarie. I valori dell'MC error suggerirebbero la necessità di aumentare il numero di iterazioni ma alcune simulazioni mettono in evidenza come l'effetto finale sulle stime a posteriori dei RR non sarebbe significativo al fine della loro rappresentazione grafica. Comunque il programma Winbugs consente di procedere con ulteriori iterazioni qualora la prima scelta avesse fornito stime largamente imprecise.

La valutazione dell'effetto dell'applicazione del modello tramite gli indici di Moran e Geary ha evidenziato che la rappresentazione degli SMR, sebbene ottenuti dopo procedure di aggiustamento che tengono in giusto conto le caratteristiche della popolazione, descrive un pattern spaziale diffuso, con estrema variabilità tra aree adiacenti, senza che sia possibile individuare aree di aggregazione. I RR derivanti dal modello di BYM generano una descrizione spaziale con minore variazione tra le aree e, quindi, possono anche trasmettere la sensazione di appiattimento del fenomeno in studio. Al contempo, però, consentono di individuare la presenza di pattern di aggregazione spaziale e, quindi, di ipotizzare aree a rischio maggiore per un determinato fenomeno, da approfondire con altre metodiche o con studi specifici.

Bibliografia

Besag J., York J., Mollié A. (1991) *Bayesian image restoration with two applications in spatial statistics. Annals of the Institute of Statistica*l Mathematics, vol. 43. (1-59).

Brooks S.P., Gelman A. (1998) *Alternative methods for monitoring convergence of iterative simulations.* Journal of Computational and Graphical Statistics, vol. 7. (434-455).

Lawson A.B., Biggeri A., Boehning D., et al.(2000) *Disease mapping models: an empirical evaluation*. Stat Med, vol. 19. (2217–2242).

Prieto R.R., García-Pérez J., Pollán M., et al. *Modelling of municipal mortality due to haematological neoplasias in Spain*. J Epidemiol Community Health, vol. 61. (165-171).

Spiegelhalter D., Thomas A., Best N., et al. (2003) *WinBUGS User Manual. Version 1.4*. Cambridge, MRC Biostatistics Unit (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml).

Moran, P.A.P. (1950). "*Notes on Continuous Stochastic Phenomena.*", Biometrika, 37:17–23.

Geary, R. C. (1954). "*The Contiguity Ratio and Statistical Mapping.*", Incorporated Statistician, 5:115–145

Mincuzzi A., Bartolomeo N., Carbonara M., Scarnera C., Trerotoli P., Serio G., Minerba A. (2013). *Relazione sullo stato di salute della popolazione pugliese-Il caso Taranto: risultati dello studio IESIT*. OER PUGLIA, vol. Suppl OER Puglia n1/2013, p. 325-362, ISSN: 2039-7135

Mincuzzi A., Minerba S., Tafuri S., Germinario C., Trerotoli P., Bartolomeo N., Serio G., Serinelli M., Morabito A., Spartera M., Giua R., Assennato G. (2013). *IESIT. Indagine Epidemiologica Sito Inquinato di Taranto*. CLIOEDU Edizioni, ISBN: 8896646-42-7