Web Working Papers
by

The Italian Group of Environmental Statistics

Gruppo di Ricerca per le Applicazione della Statistica
ai Problemi Ambientali

www.graspa.org

Quanto adeguata è la distribuzione di Poisson per l'analisi di serie temporali di mortalità ? un confronto tra modelli.

Vito M. R. Muggeo e Lorena Viviano

GRASPA Working paper n.20, Novembre 2003

Quanto adeguata è la distribuzione di Poisson per l'analisi di serie temporali di mortalità? un confronto tra modelli.

Vito M. R. Muggeo
vito.muggeo@giustizia.it

Lorena Viviano
lorena.viviano@tin.it

11 novembre 2003

Abstract

I numerosi studi pubblicati sugli effetti a breve termine dell'inquinamento atmosferico sulla mortalità per mezzo di serie temporali giornaliere non hanno mai fornito indicazioni relative alla bontà di adattamento del modello ai dati empirici. Anche se l'interesse è rivolto principalmente alla stima dell'effetto dell'inquinante, la validità del modello rimane una sorta di garanzia sui risultati ottenuti. In questo lavoro vengono considerate distribuzioni alternative alla Poisson in presenza di dati discreti ed eventuali specificazioni che potrebbero essere utilizzate per migliorare il grado di accostamento del modello ai dati. I risultati condotti sulla serie di Palermo (1997-1999) dimostrano che, sebbene modelli più sofisticati possano garantire un migliore adattamento, la regressione di Poisson fornisce un'adeguata rappresentazione del fenomeno, soprattutto in termini di stima del parametro dell'inquinante. Ciò avvalora maggiormente le analisi condotte attraverso il semplice modello poissoniano.

1 Introduzione

Recentemente sono stati pubblicati numerosi studi basati sugli effetti dell'inquinamento atmosferico sulla salute, ricoverabilità e soprattutto mortalità. Attraverso l'uso di modelli poissoniani impiegati su serie temporali giornaliere è emersa l'esistenza di un'associazione positiva tra inquinamento atmosferico e salute umana (Katsouyanni *et al.*, 1995; European Commission Enviromental Research Programme, 1996; Schwartz *et al.*, 1996; Kelsall *et al.*, 1997). Dopo aver consistentemente confermato tale associazione in diverse città europee e americane, recentemente l'interesse della ricerca si è rivolto ad ulteriori approfondimenti: meta-regressione e meta analisi per aggregare i risultati nelle diverse città e cercare di spiegarne le possibili differenze (Biggeri *et al.*, 2001); stima della relazione dose-risposta (Schwartz e Zanobetti, 2000) e analisi dell'effetto di anticipazione (*haversting*)(Schwartz, 2000, 2001). Comunque in tutte le

analisi poca rilevanza è stata data alla bontà di adattamento dei valori teorici a quelli empirici a cui sono legati tutti gli altri risultati. Caratteristica comune di tali studi è l'assunzione di un modello poissoniano per la serie degli eventi (decessi e/o ricoveri) che è soltanto un possibile modello di probabilità per dati di conteggio (sicuramente il più semplice e probabilmente il più famoso). Il modello in questione può essere formulato in

$$\log \mu_t = \eta_t = f(t, df) + \boldsymbol{\alpha}' \boldsymbol{x}_t + \beta \texttt{INQUI}_t \tag{1}$$

dove Y_t è la v.c. associata al numero di decessi per il giorno t, $E[Y_t] = \mu_t$ è il corrispondente valore atteso stimato come funzione di *i)* un termine non parametrico $f(t, df)$ che modella l'andamento stagionale e di lungo periodo attraverso una funzione 'liscia' (*smooth*) con *df* gradi di libertà (Hastie e Tibshirani, 1990); *ii)* alcuni termini parametrici (lineari) che comprendono variabili quali giorni della settimana ed epidemie di influenza e *iii)* il termine relativo all'inquinante che è il principale interesse dello studio.

I dati analizzati risultano quasi sempre sovradispersi; ad esempio Kunst *et al.* (1993) osserva che le serie di mortalità complessiva, ma non per cause specifiche, presenta una variabilità superiore a quella di Poisson. Conseguentemente per tenere in considerazione tale variazione 'extra-Poisson', l'approccio solitamente seguito per la stima dei modelli poissoniani è quello di 'quasi verosimiglianza', che consente di ottenere stime 'corrette' degli errori standard attraverso un parametro di eterogeneità calcolato sui residui di un classico modello poissoniano (Schwartz *et al.*, 1996). Questo approccio è privo di un modello generatore dei dati, ma poichè l'interesse è solitamente centrato sulla stima del parametro dell'inquinante e non sul modello complessivo, questo potrebbe non essere un reale problema in pratica. Tuttavia la mancanza di una verosimiglianza, oltre ad essere difficile da giustificare teoricamente, di fatto preclude numerose possibilità di analisi, quali la selezione di modelli non annidati con differenti gradi di libertà: ad esempio, potrebbe risultare interessante valutare se l'effetto dell'inquinante possa essere correttamente quantificato attraverso un modello a ritardi distribuiti o considerando una semplice media ai lag 0-3 che costituisce un'approssimazione della struttura dei ritardi (Schwartz, 2000). Per cui per poter procedere ai confronti tra diversi modelli, ad esempio attraverso l'AIC (*Akaike Information Criterion*), è richiesto un modello dotato di una verosimiglianza piena (Lindsey, 1999) relativo ad distribuzioni, alternative a quella di Poisson, che potrebbero essere utilizzate in tale ambito. Inoltre tali distribuzioni con due o più parametri potrebbero rappresentare i dati (sovradispersi) in modo più accurato consentendo, qualora se ne presentasse la necessità, di modellare altri aspetti (ad es., eterogeneità) in funzione di opportune variabili esplicative.

Lo scopo di questo lavoro è quello di investigare altri aspetti della modellazione per poter eventualmente arrivare ad un modello probabilmente più sofisticato ma migliore in termine di accostamento ai dati empirici. Naturalmente non ci aspetteremmo di riuscire a ridurre drasticamente la variabilità residua, anche per la natura stessa dei dati in esame, tuttavia potrebbe risultare interessante verificare se qualche altro modello 'migliore' porti a differenze sostanziali in termini di stima dei parametri.

Il lavoro si articola come segue. Nella Sezione 2 vengono accennate le possibili fonti di sovradispersione attribuibili principalmente ad una non corretta specificazione del modello per il parametro media (*mean model* o *location model*) o del modello per il parametro di forma (*shape model*); vengono quindi passate in rassegna le distribuzioni impiegabili in presenza di dati di conteggio. Nella Sezione 3 vengono presentati i dati utilizzati nel presente lavoro ed i relativi risultati. La Sezione 4 è dedicata a commenti e discussione.

2 Potenziali Cause di Sovradispersione

2.1 Errata specificazione del modello per il parametro media $\boldsymbol{\mu = E[Y]}$

Un non adeguato adattamento del modello ai dati può essere causato da un'errata specificazione del modello per il parametro di locazione $\mu_t = \exp\{\eta_t\}$, ovvero *i)* errata funzione link (in questo caso la funzione logaritmica); *ii)* errata relazione tra variabile e una o più esplicative; *iii)* autocorrelazione tra le osservazioni; *iv)* ed infine mancanza di una qualche variabile esplicativa.

Negli studi degli effetti a breve sulla salute, l'uso della funzione legame logaritmica assicura che i parametri (lineari) nel predittore possano esprimersi come log-rischio relativo per cui altre funzioni link sono escluse a priori; l'approccio nonparametrico è sufficientemente flessibile per valutare ed eventualmente modellare possibili relazioni non lineari (come è il caso del termine $f(t)$); ed infine i dati provengono da rilevazioni di routine e conseguentemente nella costruzione del modello regressivo possono essere contemplate soltanto le variabili per cui esistono misurazioni: temperatura, umidità e recentemente anche pressione barometrica.

Conseguentemente l'unico punto su cui il modello appare migliorabile è il punto *iii)*, cioè la modellazione dell'autocorrelazione. I conteggi giornalieri di mortalità sono verosimilmente non indipendenti, per cui una scelta ragionevole potrebbe essere quella di modellare esplicitamente tale autocorrelazione attraverso un modello quale

$$\log E[Y_t] = \eta_t + \tau_t$$

dove τ_t è una componente che tiene conto delle osservazioni non-iid. Ad esempio, un modello GARMA(p,q) (Generalized Auto Regressive Moving Average)

$$\log E[Y_t] = \eta_t + \sum_k^p \rho_k \{\log y_{t-k} - \beta x_{t-k}\} + \sum_k^q \theta_k \{\log y_{t-k} - \eta_{t-k}\}$$

potrebbe essere utilizzato a tal scopo (Benjamin *et al.*, 2003).

Comunque come evidenziato da una discussione apparsa sull' *American Journal of Epidemiology*, il problema è che poiché anche i dati dell' inquinante sono autocorrelati, inserendo un termine autoregressivo può portare a sottostime nell'effetto dell'inquinante (Saez e Figueiras, 2000; Zeger *et al.*, 2000). Conseguentemente tali modelli sono raramente utilizzati nelle applicazioni e

perciò d'ora in avanti nel presente lavoro si assume che il modello per la media sia del tipo $\log \mu = \eta$, senza alcuna componete ARMA e sia stato correttamente specificato e non sia ulteriormente migliorabile per i motivi discussi sopra.

2.2 Distribuzione di probabilità non adeguata

Nel modellare dati continui, oltre alla classica distribuzione Gaussiana, sono di comune utilizzo fra i ricercatori distribuzioni alternative (ad es., Gaussiana inversa, log-normale, gamma) che talvolta possono meglio interpolare il fenomeno in esame. Anche per dati discreti di conteggio esistono altre distribuzioni oltre a quella di Poisson, sebbene queste risultino poco note e conseguentemente poco utilizzate nelle applicazioni. Di seguito vengono passate in rassegna tali distribuzioni che essendo caratterizzate da 2 parametri potrebbero meglio adattarsi ai dati empirici; per completezza viene inserita anche la distribuzione di Poisson. Anche se relativo a diverse distribuzioni, il secondo parametro viene sempre riportato con 'ϕ' per non appesantire troppo la trattazione e per evidenziarne il medesimo ruolo (quello di tenere conto della sovradispersione). Alcune distribuzioni presentano una costante di normalizzazione $c(\mu, \phi)$ che viene determinata numericamente al momento della stima dei parametri. Infine l'insieme di definizione è chiaramente per $y = 0, 1, 2, \ldots$ che quindi non viene ripetuto ogni volta e dettagli teorici sono omessi perché non di diretto interesse pratico.

Distribuzione di Poisson È il modello maggiormente utilizzato nelle analisi delle serie temporali di mortalità; rappresenta probabilmente il più famoso modello per dati di conteggio ed ha funzione di densità

$$f(y; \mu) = \frac{\mu^y}{y!} e^{-\mu}$$

con $E[Y] = V[Y] = \mu$. Essendo parametrizzato soltanto da μ tale modello potrebbe non essere sufficiente ad interpolare adeguatamente i dati che risultano per loro natura sovradispersi. Assumendo (erroneamente) una distribuzione di Poisson, le stime di massima verosimiglianza sono consistenti, ma gli errori standard non sono corretti (Agresti, 2002). L'introduzione di un parametro di eterogeneità $\phi > 1$ (Lovison, 1997) tale che $\mathbb{V}[Y] = \phi\mu$ può ridimensionare la sovrastima degli errori standard dei parametri ottenuti attraverso il semplice modello poissoniano, ma a tale ipotesi sulla varianza non corrisponde alcun modello di probabilità, come è stato già discusso.

È bene rimarcare che la sovradispersione non si può verificare in presenza di distribuzioni con un parametro aggiuntivo oltre a quello 'medio', come è il caso della distribuzione normale, per esempio. Questo è perché esiste un parametro differente che descrive la variabilità. Per i dati discreti (Binomiale e Poisson) la varianza è funzione della media e il problema della sovradispersone si verifica spesso, per cui distribuzioni con due (o più) parametri potrebbero risultare davvero utili.

Distribuzione Binomiale-Negativa Assumendo che il parametro μ sia esso stesso una v.c. la distribuzione condizionata $(Y|\mu)$ è Poisson, ma quella marginale è il risultato dell'integrale $\int f_{Y|\mu} f_\mu d\mu$. Quando la distribuzione di μ è gamma la distribuzione risultante è binomiale negativa:

$$f(y;\mu,\phi) = \frac{\Gamma(y+\phi)}{\Gamma(\phi)\ \Gamma(y+1)} \left(\frac{\phi}{\mu+\phi}\right)^{\phi} \left(\frac{\mu}{\mu+\phi}\right)^{y}$$

ϕ^{-1} è il parametro di dispersione che quantifica l'eteorgeneità (Agresti, 2002). Infatti si ha che $E[Y] = \mu$ e $V[Y] = \mu + \mu^2/\phi$. Se $\phi^{-1} = 0$ non si ha alcuna sovradispersione rispetto all'assunzione poissoniana per cui $E[Y] = V[Y]$. La distribuzione Binomiale Negativa rientra nell'approccio così detto di 'verosimiglianza piena' caratterizzati dalla possibilità di esprimere in forma chiusa la distribuzione marginale. L'obiettivo è cogliere l'eterogeneità considerando delle forme distribuzionali note. Si osservi che la densità riportata sopra è stata ottenuta assumendo che la v.c. condizionata $Y|\mu$ sia poissoniana con opportuna media, m, diciamo, essendo m una v.c. Gamma con parametro di forma (*shape parameter*) pari a ϕ e parametro di scala (*scale parameter*) pari a ϕ/μ. Assumendo per m una Gamma con parametri $\phi\mu$ e ϕ rispettivamente si ottiene una alternativa formulazione della binomiale negativa tale che $V[Y] = \mu(1 + 1/\phi)$. Questa formulazione viene talvolta riferita come *constant dispersion* perché il parametro di eterogeneità ($(1 + 1/\phi)$) viene assunto 'costante' in contrasto con la prima formulazione riferita come *variable dispersion* dove il parametro di eterogeneità è $1 + \mu/\phi$, variabile in quanto funzione di μ. La formulazione a dispersione costante è molto simile all'assunzione di una Poisson sovradispersa (per le quali risulta $V[Y] = k\mu$ con k reale qualsiasi), per cui non appare molto diffusa nelle applicazioni e non sarà considerata nel presente lavoro. Altre formulazioni della Binomiale Negativa sono possibili, ma la loro trattazione esula dallo scopo del presente scritto.

Modello di Poisson Doppio In Efron (1986) è stata presentata una famiglia di distribuzione 'esponenziale doppia' la cui corrispondente distribuzione di Poisson doppia è:

$$f(y;\mu,\phi) = c(\mu,\phi)\frac{e^{\phi(y-\mu)}\mu^{y(\phi+1)}}{y^{y\phi}y!}$$

Modello di Poisson Moltiplicativo Questo modello è una generalizzazione di quello Poissoniano e la funzione di densità:

$$f(y;\mu,\phi) = c(\mu,\phi)\frac{e^{-\mu}\mu^{y}e^{\phi y^2}}{y!}$$

si riduce a quella classica poissoniana con $\phi = 0$. Tale modello è un estensione al caso poissoniano della generalizzazione della distribuzione binomiale di Altham (1978).

Modello di Conteggio Gamma È noto che se la distribuzione del verificarsi di eventi in intervalli di tempo è Poisson, allora la distribuzione degli intervalli di tempo tra eventi è esponenziale; analogamente se la distribuzione stessa tra eventi è Gamma, allora la distribuzione degli eventi è *gamma count* (Winkelmann, 1997):

$$f(y;\mu,\phi) = F_g(\mu;\phi y,\phi y) - F_g(\mu;\phi(y+1),\phi(y+1))$$

dove $F_g(\mu;\cdot,\cdot)$ è la funzione di ripartizione della distribuzione Gamma nella variabile μ con i due parametri, media e dispersione, uguali.

Modello di Poisson Generalizzato Quest'ultimo modello per dati discreti è presentato in Consul e Famoye (1992) ed ha funzione di densità:

$$f(y;\mu;\phi) = \frac{\mu[\mu + y(\phi-1)]^{y-1} e^{[\mu+y(\phi-1)]/\phi}}{\phi^y y!}$$

con ϕ^2 fattore di eterogeneità.

2.3 Errata specificazione del modello di forma

Per le distribuzioni che possiedono parametri aggiuntivi al parametro di locazione μ, si apre la possibilità di modellare anche tale parametro come funzione di un qualche vettore di esplicative attraverso una funzione $h(\cdot)$ tale che:

$$h(\phi_t) = \boldsymbol{\gamma}\boldsymbol{z}_t$$

dove $\boldsymbol{\gamma}$ è il vettore dei parametri da stimare e $\boldsymbol{z}_t$ un vettore di variabili esplicative non necessariamente incluse nel modello per la media. Smith (1989) discute questa estensione per GLM con dati quantitativi e mostra come stimare simultaneamente un 'sub-modello' per il parametro di dispersione. Nell'articolo di Smith appare evidente come una specificazione di un modello per il parametro di dispersione possa migliorare l'accostamento dei dati empirici a quelli teorici e talvolta modificare anche i risultati del modello per μ a cui si è maggiormente interessati. Inoltre la modellazione del parametro di forma può anche essere visto come una mancanza di conoscenza della media (Smith, 1989).

3 Dati e Risultati

Per confrontare i modelli presentati sopra, vengono utilizzati i dati di Palermo (agosto 1996-dicembre 1999) già descritti e analizzati attraverso classici modelli poissoniani sovradispersi (Muggeo, 2000; Biggeri *et al.*, 2001). Le serie di Palermo utilizzate mostrano tutte una variabilità residua (*Deviance*) superiore ai gradi di libertà, ovvero un non adeguato adattamento e quindi sovradispersione. In questa sede viene considerata soltanto la serie della mortalità complessiva per cause naturali (ICD.IX 1-799) ed il principale interesse è rivolto alla stima del parametro del PM_{10} (polveri con diametro inferiore a 10 micron) ed eventuali confronti tra i diversi modelli anche non-annidati.

Si noti che sebbene esistano diverse procedure per saggiare l'ipotesi di sovradispersione (ad es., Ganio e Schafer, 1992; Dean e Lawless, 1989; Collings e Margolin, 1985; Dean, 1992), negli studi epidemiologici tali procedure vengono ignorate e si procede direttamente alla stima di ϕ attraverso la statistica di Pearson (McCullagh e Nelder, 1989). Comunque un semplice *score* test che è utilizzabile dopo aver stimato il modello poissoniano standard è (Dean e Lawless, 1989; Dean, 1992).

$$T = \frac{\sum \left[(y_t - \hat{\mu}_t)^2 - (1 - \hat{h}_t)\hat{\mu}_t \right]}{(2 \sum \hat{\mu}_t^2)^{1/2}}$$

dove $y_t, \hat{\mu}_t$ sono i valori osservati e quelli attesi e $\hat{h}_t$ è il t-esimo elemento della diagonale della *hat matrix*. Tale statistica test, che ha una distribuzione normale standard sotto l'ipotesi di nessuna sovradispersione, assume come alternativa la funzione di varianza della distribuzione binomiale negativa $\mu + \mu^2 \times \phi^{-1}$. Per le serie di Palermo il valore di tale statistica test è sempre superiore a 4.2, indicando, come atteso, una dispersione empirica superiore a quella nominale assunta dal modello poissoniano.

Per valutare possibili differenze dovute a differenti indicatori dell'inquinante, per ogni distribuzione di probabilità quattro modelli sono stati stimati, ciascuno con un diverso indicatore dell'inquinamento medio giornaliero: lag0, lag0-1, lag0-3 e 'Almon'. Per Almon si intende una formulazione utilizzata per modellare ritardi distribuiti, ovvero un polinomio che contempla l'effetto dell'inquinante per i giorni precedenti $t-1, t-2, \ldots$ (Schwartz, 2000). Tale formulazione è richiesta in quanto l'inserire nel modello direttamente la stessa variabile sfasata a $t, t-1, t-2 \ldots$ porterebbe a problemi di multicollinearità. In breve data una variabile X_t e assunto un polinomio di terzo grado per contemplare fino a 3 lag, la parametrizzazione di Almon porta a considerare nel modello le tre variabili

$$Z_{0t} = \sum_k^3 X_{t-k} \qquad Z_{1t} = \sum_k^3 k X_{t-k} \qquad Z_{2t} = \sum_k^3 k^2 X_{t-k}$$

piuttosto che direttamente $X_t, X_{t-1}, X_{t-2}, X_{t-3}$. I coefficienti delle variabili trasformate Z, $(c_0, c_1, c_2)^\mathsf{T}$, diciamo, sono utilizzati per ottenere quelli relativi alle variabili originali X:

$$\beta_0 = c_0 \quad \beta_1 = c_0 + c_1 + c_2 \quad \beta_2 = c_0 + 2c_1 + 4c_2 \quad \beta_3 = c_0 + 3c_1 + 9c_2$$

dove β_j è il coefficiente per la variabile X_{t-j} $j = 0, 1, 2, 3$ che modella l'effetto dell'inquinante X_t sui decessi del giorno $t+j$. Quindi l'effetto medio complessivo su quattro giorni è dato semplicemente dalla somma dei coefficienti, ovvero $\sum_{j=0}^3 \beta_j$.

L'AIC è stato utilizzato per la ricerca del miglior modello ed è definito da (Lindsey, 1999)

$$AIC = -2 \log L + 2 \times n.param$$

Migliore è il modello, più elevata è la verosimiglianza, e più basso è l'AIC. Su tutte le serie analizzate, relazioni nonlineari, quali stagionalità, temperatura e umidità, sono state modellate in un approccio pienamente parametrico (attraverso le *spline di regressione*) che al contrario di un approccio nonparametrico (*smoothing spline*) risulta fornire stime non distorte per i parametri lineari (Dominici *et al.*, 2002).

La Tabella 1 illustra gli AIC per i differenti modelli di probabilità nel *core model*, ovvero il modello con tutte le variabili esplicative ad eccezione dell'inquinante.

Tabella 1: AICs per i *core model* per i diversi modelli per sovradispersione.

Model	AIC
Poisson	3515.0
Binomiale Negativa	3509.1
Poisson Doppia	3509.5
Poisson Moltiplicativa	3512.0
Conteggio Gamma	3509.4
Poisson Generalizzata	3510.1

Forse sorprendentemente è interessante osservare come gli AIC siano sostanzialmente simili. Come atteso i modelli con due parametri sembrano fornire un migliore adeguamento rispetto al semplice modello poissoniano, ma le differenze sono davvero trascurabili. Anche inserendo gli inquinanti, i risultati sono sostanzialmente immutati: seppure con piccole differenze, la distribuzione binomiale negativa risulta essere il modello con un AIC più basso.

Tabella 2: Stime puntuali (errori standard in parentesi) degli effetti delle polveri per diversi lag e per i differenti modelli di probabilità. I valori si riferiscono ad incrementi di $1000\mu g/m^3$.

	Stime puntuali (errori standard)			
Model	lag0	lag0-1	lag0-3	lag0-3(Almon)
Poisson	0.769(0.453)	1.903(0.553)	2.616(0.666)	2.401(0.686)
Binomiale Negativa	0.768(0.492)	1.905(0.596)	2.619(0.714)	2.408(0.729)
Poisson Doppia	0.768(0.492)	1.905(0.596)	2.618(0.714)	2.408(0.729)
Poisson Moltiplicativa	0.768(0.492)	1.905(0.596)	2.618(0.715)	2.408(0.729)
Conteggio Gamma	0.907(0.485)	2.103(0.591)	2.630(0.715)	2.412(0.733)
Poisson Generalizzata	0.789(0.483)	1.903(0.589)	2.611(0.709)	2.367(0.726)

Coerentemente con i risultati della Tabella 1, una diversa assunzione di probabilità sulle osservazioni non sembra influenzare le stime degli inquinanti. Il modello poissoniano porta ad errori standard sensibilmente più piccoli rispetto alle altre distribuzioni, ma la correzione per sovradispersione attraverso l'approccio di quasi verosimiglianza porta a valori decisamente accostabili a quelli ottenuti con distribuzioni con 2 parametri. Ad esempio considerando il valore dell'inquinante al lag 0-3 la prima riga della Tabella 2 riporta uno errore standard pari a 0.666 che diventa 0.715 quando moltiplicato per la corrispondente

statistica di Pearson. Simili risultati si hanno per le altre misurazioni degli inquinanti.

L'uso di distribuzioni alternative consente di specificare anche sub-modelli di regressione per il parametro aggiuntivo oltre che quello di locazione. È da osservare, comunque, che generalizzando l'approccio di quasi-verosimiglianza è possibile specificare un modello per il parametro di dipersione attraverso la 'quasi verosimiglianza estesa' (EQL, extended quasi likelihood) (Smyth e Verbyla, 1999). Comunque coerentemente con quanto discusso precedentemente un approccio di verosimiglianza piena attraverso le altre distribuzioni sarebbe auspicabile. Ai dati di mortalità di Palermo sono stati adattati tutti i modelli in cui il parametro aggiuntivo risultava espresso come combinazione lineare di stagionalità (mesi), temperatura ed inquinante. Per nessun modello sono emersi miglioramenti o differenze degne di nota, suggerendo la mancanza di alcuna variazione sistematica nella dispersione.

4 Conclusioni

In questo lavoro sono stati discussi i diversi aspetti che riguardano la modellazione delle serie temporali impiegate nell'analisi degli effetti a breve termine dell'inquinamento atmosferico sulla mortalità; in particolare sono stati confrontati alcuni altri modelli per dati di conteggio che potrebbero costituire un'alternativa al semplice modello poissoniano. I modelli di probabilità descritti permettono di tenere conto della sovradispersione attraverso una verosimiglianza piena e quindi consentono l'uso di statistiche, quali l'AIC, per confrontare modelli non-annidati. L'analisi svolta sulla serie di mortalità complessiva naturale per Palermo (agosto 1996-dicembre 1999) ha suggerito che la distribuzione Binomiale Negativa può essere un modello più adatto nel trattamento di tali dati. Ciò appare anche plausibile in quanto si può ipotizzare che ogni individuo nella popolazione sia soggetto al verificarsi di eventi secondo un modello di Poisson, ma contemporaneamente il rispettivo parametro non è omogeneo tra gli individui, ma varia (a causa della diversa età, causa di decesso,...) secondo una distribuzione di qualche variabile casuale. La scelta di una distribuzione Gamma è in realtà un'ipotesi di comodo in quanto essendo tale distribuzione 'coniugata' con la Poisson, permette di risolvere analiticamente l'integrale per il calcolo della distribuzione marginale: il risultato è la distribuzione binomiale negativa del paragrafo 2.2.

Tuttavia in termini di stima dell'effetto dell'inquinante, le diverse assunzioni sulla densità delle osservazioni, unitamente ad una modellazione del parametro 'di forma', non hanno portato ad alcuna differenza sostanziale. I risultati sono stati ottenuti per diverse misurazioni del PM_{10} (lag0, lag0-1, lag0-3, Almon) e per la serie di mortalità complessiva di Palermo, ma non c'è ragione per ritenere che lo stesso valga per altri inquinanti e per le serie di altri centri, sebbene potrebbe essere interessante verificare questa sensazione.

Studi condotti su serie per cause specifiche (ad esempio, respiratorie) o per gruppi selezionati (ad esempio, maschi con età superiore a 75 anni) dovrebbero portare a dati più omogenei e quindi dovrebbero portare a ridurre le già piccole

differenze tra i diversi modelli. A questo proposito è interessante osservare che una eccessiva stratificazione può di fatto portare a serie con un elevato numero di zeri; lo stesso risultato potrebbe verificarsi allorquando lo studio sia condotto in città medio-piccole. In tali casi dovrebbe essere interessante valutare se anche per tali serie 'sparse' le differenze si mantengono trascurabili; in particolare i confronti dovrebbero riguardare anche le distribuzioni che modellano esplicitamente l'eccesso di zeri (ad esempio, la distribuzione ZIP, *zero inflated poisson*).

In definitiva per analisi standard di serie temporali di mortalità, il semplice modello di Poisson sovradisperso, fin qui utilizzato da quasi tutti gli studi pubblicati, risulta essere un adeguato strumento per valutare gli effetti a breve dell'inquinamento sulla salute. Il suo utilizzo, rispetto ad altre distribuzioni, può essere giustificato anche per la semplicità di calcolo: come è noto un GLM poissoniano viene stimato attraverso l'algoritmo dei minimi quadrati iterati pesati, che risulta essere particolarmente adeguato soprattutto in presenza di un grande ammontare di dati (questi studi arrivano ad analizzare serie anche per 15 anni); d'altro canto l'uso di modelli con distribuzioni alternative necessita usualmente di algoritmi meno efficienti che richiedono valori iniziali adeguati che potrebbero causare difficoltà nelle applicazioni.

Riferimenti bibliografici

Agresti A. (2002). *Categorical Data Analysis.* John Wiley & Sons, New York.

Altham P. (1978). Two generalizations of the binomial distribution. *Appl Stat*, **27**, 162–167.

Benjamin M. A., Rigby R. A., e Stasinopoulos M. D. (2003). Generalized autoregressive moving average models. *J Am Statist Assoc*, **98**, 214–223.

Biggeri A., Bellini P., e B. Terracini B E. (2001). Meta-analysis of the italian studies on short-term effects of air pollution. *Epidemiologia & Prevenzione*, **25**, 1–72.

Bissell A. F. (1972). A negative binomial model with varying element sizes. *Biometrika*, **59**(2), 435–441.

Collings B. J. e Margolin B. H. (1985). Testing goodness of fit for the Poisson assunption when observations are not identically distributed. *J Am Statist Assoc*, **80**(390), 411–418.

Consul P. C. e Famoye F. (1992). Generalized Poisson regression model. *Communications in Statistics - Theory and Methods*, **21**, 89–109.

Dean C. e Lawless J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *J Am Statist Assoc*, **84**(406), 467–472.

Dean C. B. (1992). Testing for overdispersion in Poisson and Binomial regression models. *J Am Statist Assoc*, **87**(418), 451–457.

Dominici F., McDermott A., Zeger S. L., e Samet J. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *Am J Epidemiol*, **156**, 193–203.

Efron B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81**, 709–721.

European Commission Enviromental Research Programme (1996). The APHEA project. *J Epidemiol Comm Health*, **50**, Suppl 1.

Ganio L. M. e Schafer D. W. (1992). Diagnostics for overdispersion. *J Am Statist Assoc*, **87**(419), 795–804.

Gelfand A. E. e Dalal S. R. (1990). A note on overdispersed exponential families. *Biometrika*, **77**(1), 55–64.

Hastie T. J. e Tibshirani R. J. (1990). *Generalized Additive Models.* Chapman & Hall, seconda edition.

Katsouyanni K., Zmirou D., Spix C., *et al.* (1995). Short term effects of air pollution on health: a european approach using epidemiologic time series data. *Eur Resp J*, **8**, 1030–1038.

Kelsall J. E., Samet J. M., Zeger S., e Xu J. (1997). Air pollution and moratlity in Philadelphia, 1974-1988. *Am J Epidemiol*, **146**, 750–762.

Kunst A. E., Looman C., e Mackenbach J. P. (1993). Outdoor air temperature and mortality in the Netherlands: a time series analysis. *Am J Epidemiol*, **137**, 331–341.

Lambert D. e Roeder K. (1995). Overdispersion diagnostics for generalized linear models. *J Am Statist Assoc*, **90**, 1225–1235.

Lindsey J. (1999). *Models for repeated measurements, II ed.* Oxford University Press.

Lovison G. (1997). Some notes on binary data with overdispersion. *SA*, **9**(3), 403–418.

McCullagh P. e Nelder J. A. (1989). *Generalized Linear Models.* Chapman & Hall, II ed.

Muggeo V. (2000). *Effetti a breve termine dell'inquinamento atmosferico sulla salute umana: un'analisi sulla città di Palermo.* Tesi di Dottorato di Ricerca, Università di Palermo.

Saez M. e Figueiras A. (2000). Re: Air pollution and moratlity in Philadelphia, 1974-1988 (letter). *Am J Epidemiol*, **151**, 106.

Schwartz J. (2000). The distributed lag between air pollution and daily death. *Epidemiol*, **11**, 320–326.

Schwartz J. (2001). Is there harvesting in the association of airborne particles with daily deaths and hospital admission? **12**, 55–61.

Schwartz J. e Zanobetti A. (2000). Using meta-smoothing to estimate dose-response trends across multiple studies, with application to air pollution and daily death. **11**, 666–672.

Schwartz J., Spix C., Touloumi G., Bacharova L., *et al.* (1996). Methodological issues in studies of air pollution and daily counts of deaths or hospital admission. *J Epidemiol Comm Health*, **50 (Suppl 1)**, S3–S11.

Smith G. K. (1989). Generalized linear models with varying dispersion. *J Roy Statist Soc*, **51**(1), 47–60.

Smyth G. e Verbyla A. (1999). Double generalized linear models: approximate REML and diagnostics. *Proceedings of the 14th International Workshop on Statistical Modelling.*

Winkelmann R. (1997). *Econometric Analysis of count data.* Springer-Verlag.

Zeger S., Samet J. M., e Kelsall J. E. (2000). Re: Re: Air pollution and moratlity in Philadelphia, 1974-1988 (letter). *Am J Epideiol*, **151**, 107.