UNIVERSITÀ DEGLI STUDI DI BERGAMO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
E METODI MATEMATICI°

QUADERNI DEL DIPARTIMENTO

Department of Information Technology and Mathematical Methods

Working Paper

Series "*Mathematics and Statistics*"

n. 1/MS – 2008

Modelli di conteggio con eccesso di zeri: due approcci a confronto

by

Lorena CM Viviano

° Viale Marconi. 5, I – 24044 Dalmine (BG), ITALY, Tel. +39-035-2052339; Fax. +39-035-562779

[§] L'accesso alle *Series* è approvato dal Comitato di Redazione. I *Working Papers* della Collana dei Quaderni del Dipartimento di Ingegneria dell'Informazione e Metodi Matematici costituiscono un servizio atto a fornire la tempestiva divulgazione dei risultati dell'attività di ricerca, siano essi in forma provvisoria o definitiva.

Modelli di conteggio con eccesso di zeri: due approcci a confronto

Lorena CM Viviano
lorena.viviano@unibg.it

Abstract

I dati di conteggio che presentano un eccesso di zeri possono essere trattati facendo ricorso ad una metodologia alternativa a quella comunemente impiegata. L'argomento viene affrontato nella modellazione di un insieme di dati reali di tipo ambientale, inquadrati nel contesto di una recente estensione dei modelli lineari generalizzati (GLM) rappresentata dai modelli di mistura GLM. Nel caso di misture a due componenti, se una delle due ha una distribuzione degenere con la massa concentrata a zero, si ottengono i modelli di regressione con eccesso di zeri, che per dati di conteggio prendono il nome di *Zero-inflated Poisson* (ZIP), *Zero-inflated Negative Binomial* (ZINB), *Hurdle Poisson* (HP), *Hurdle Negative Binomial* (HNB). Nel caso di serie storiche di conteggio si può fare ricorso ad una metodologia che consente di tenere conto contemporaneamente della dipendenza temporale e dell'eccesso di zeri (modelli *"hidden markov"* con eccesso di zeri). L'obbiettivo del lavoro è quello di confrontare le caratteristiche dei due approcci e le conclusioni a cui è possibile pervenire.

1 Introduzione

La distribuzione di Poisson è la funzione di probabilità solitamente impiegata per modellare dati di conteggio.

In molte applicazioni reali si può osservare un numero di conteggi nulli maggiore di quello atteso sotto il modello di Poisson. Ovvero la distribuzione di probabilità che modella i conteggi è simile ad una distribuzione di Poisson, ma se ne differenzia poichè in corrispondenza di zero si osserva una frequenza superiore a quella attesa.

In tali situazioni si può fare ricorso a modelli di conteggio che tengono esplicitamente conto della percentuale elevata di conteggi nulli.

Come descritto in Dietz e Böhning (2000), diverse sono le cause che possono portare all'eccesso di zeri. Vi possono essere delle unità campionarie che non sono realizzazioni Poisson oppure ci possono essere delle motivazioni legate al tipo di campionamento scelto che determina un'elevata presenza di conteggi nulli. Prendendo a prestito una terminologia propria dello studio congiunto di variabili categoriali, potremmo parlare nei due casi rispettivamente di 'zeri strutturali' e 'zeri campionari'. Ad esempio, in una indagine sulla pesca viene rilevato il numero di pesci pescati nell'ultimo mese da un campione di individui. Il soggetto intervistato può affermare di non avere pescato alcun pesce

principalmente per due possibili motivi: perchè non ha l'hobby della pesca e quindi non ha mai pescato (zero strutturale) oppure perchè è andato a pesca ma non ha preso alcun pesce (zero campionario). Nel caso in cui il campione presenti uno sbilanciamento verso soggetti che non si dilettano con l'hobby della pesca si osserva una presenza eccessiva di zeri strutturali; quindi la modellazione dell'insieme di dati va affrontata nella direzione dei modelli inflazionati.

I contesti nei quali è più frequente il trattamento di modelli con eccesso di conteggi nulli riguardano l'ambito econometrico, demografico e medico. Tali modelli sono caratterizzati da una struttura parametrica che modella opportunamente la risposta nulla. Zorn (1996) giustifica questo approccio in termini di un processo a 'regime duale' che governa i dati: nel primo stadio, un modello di presenza/assenza (di conteggio nullo) determina se la variabile risposta è pari a zero o è un conteggio diverso da zero; nel secondo stadio, il modello di conteggio governa la variabile di risposta.

L'approccio standard per la modellazione di dati di conteggio con eccesso di zeri può essere inquadrato nel contesto di una recente estensione dei modelli GLM rappresentata dai *modelli di mistura GLM*. In base a questo approccio (Wang, 2004) vengono prese in considerazione misture di un certo numero di variabili casuali y_i avente funzione di probabilità $f(x_i;\boldsymbol{\theta})$ $(i = 1, 2, \ldots, k)$ appartenenti alla classe esponenziale in proporzioni $\pi_1, \pi_2, \ldots, \pi_k$. La $f(\cdot)$ può essere funzione di un certo numero di parametri incogniti $\boldsymbol{\theta}$. La funzione di probabilità della distribuzione marginale è: $f(\boldsymbol{y};\boldsymbol{\theta}) = \sum_{i=1}^{k} \pi_i f(x_i;\boldsymbol{\theta})$, con $\boldsymbol{y} = [y_1, \ldots, y_k]^T$ e $\sum_{i=1}^{k} \pi_k = 1$.

L'impiego di un modello di mistura ha evidenti vantaggi nel caso discreto, quando è possibile ipotizzare che l'eterogeneità presente nei dati può essere riferita a un certo numero di classi latenti. Altre volte, quando questa ipotesi non ha valore, il modello di mistura può essere impiegato come una tecnica parsimoniosa per modellare i dati, per cui ogni singola componente fornisce un'approssimazione locale della vera distribuzione (Cameron e Trivedi, 1998).

Si consideri il generico modello di mistura per due componenti:

$$P(Y = y) = \pi f_1(y) + (1 - \pi) f_2(y) \tag{1}$$

dove y è il conteggio, π è la probabilità di avere un conteggio pari a zero nel modello di presenza/assenza, $f_1(y) = I_{\{0\}}(y)$ corrisponde alla funzione di probabilità prescelta calcolata in corrispondenza del conteggio nullo e $f_2(y)$ è la funzione di probabilità della variabile casuale (di conteggio).

Il trattamento di modelli per dati di conteggio con eccesso di zeri è stato trattato in un lavoro precedente dell'autore nell'analisi di due insiemi di dati reali (Viviano *e altri*, 2005).

Il presente lavoro si articola secondo il seguente schema. Nella Sezione 2 viene illustrato l'approccio standard nel trattamento di dati di conteggio con eccesso di zeri; nella Sezione 3 si descrive l'approccio dinamico. Infine nella Sezione 4 si considera uno studio della balneabilità nella provincia di Rimini impiegato per confrontare i due approcci.

2 Approccio standard per la modellazione di dati di conteggio con eccesso di zeri

In questo paragrafo verranno descritti i modelli che solitamente vengono trattati nella modellazione di insiemi di dati che presentano un eccesso di conteggi pari a zero.

In presenza di dati di conteggio si fa generalmente ricorso alla distribuzione di Poisson, caratterizzata dal valore atteso uguale alla varianza. In molte applicazioni però si osserva che la varianza è maggiore del valore atteso, dando luogo al fenomeno della sovradispersione. In tali situazioni si fa ricorso alla distribuzione Binomiale Negativa.

A partire dall'equazione (1) che illustra un modello di mistura a due componenti, se una delle due ha una distribuzione degenere con la massa concentrata a zero si ottengono i modelli di regressione con eccesso di zeri, che per dati di conteggio prendono il nome di *Zero-inflated Poisson* (ZIP), *Zero-inflated Negative Binomial* (ZINB), *Hurdle Poisson* (HP), *Hurdle Negative Binomial* (HNB).

Tali modelli possono essere ricavati dalla generica formula (1), attraverso un'appropriata scelta di π e $f_2(y)$, ovvero:

$$\text{ZIP: } P(Y=y) = \begin{cases} \pi + (1-\pi)\exp(-\mu) & \text{per } y=0 \\ (1-\pi)\frac{\exp(-\mu)\mu^y}{y!} & \text{per } y>0 \end{cases}$$

$$\text{ZINB: } P(Y=y) = \begin{cases} \pi + (1-\pi)\left(\frac{\phi}{\mu+\phi}\right)^{\phi} & \text{per } y=0 \\ (1-\pi)\frac{\Gamma(y+\phi)}{\Gamma(\phi)\ y!}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y} & \text{per } y>0 \end{cases}$$

$$\text{HP: } P(Y=y) = \begin{cases} \pi & \text{per } y=0 \\ \frac{(1-\pi)\exp(-\mu)\mu^y}{(1-\exp(-\mu))y!} & \text{per } y>0 \end{cases}$$

$$\text{HNB: } P(Y=y) = \begin{cases} \pi & \text{per } y=0 \\ \frac{(1-\pi)\frac{\Gamma(y+\phi)}{\Gamma(\phi)\ y!}(\frac{\phi}{\mu+\phi})^{\phi}(\frac{\mu}{\mu+\phi})^{y}}{1-(\frac{\phi+\mu}{\phi})^{-\phi}} & \text{per } y>0 \end{cases}$$

dove μ è il valore atteso della distribuzione e ϕ^{-1} è il parametro di sovradispersione.

La distinzione tra il modello *zero inflated* e il modello *Hurdle* riguarda il ruolo del conteggio nullo. Ipotizzando l'esistenza di due popolazioni, A_1 costituita da tutte le unità aventi valore zero (zeri strutturali) e A_2 costituita dai conteggi uguali a zero (zeri campionari) o maggiori di zero , è nei modelli di tipo *Hurdle* (in letteratura anche *zero-altered model* o *two part model*) che si realizza la vera partizione, costituita dall'insieme dei soli conteggi nulli e dall'insieme dei conteggi maggiori di zero. Nel modello con eccesso di zeri il conteggio nullo può provenire da A_1 o da A_2 e sopratutto si osserva come il numero di conteggi pari a zero supera significativamente qualsiasi altro conteggio maggiore di zero.

In un contesto più realistico per tutti i modelli inflazionati è opportuno introdurre vettori di variabili esplicative, $\boldsymbol{x}$ e $\boldsymbol{z}$, relazionate a μ e π attraverso delle funzioni legame nello spirito dei Modelli Lineari Generalizzati: $\log(\mu) =$

$\boldsymbol{x}^T\boldsymbol{\beta}$ e $\text{logit}(\pi) = \boldsymbol{z}^T\boldsymbol{\gamma}$. In questo caso si parla di modelli di regressione ZIP (Lambert, 1992).

Va sottolineato inoltre che:

1. usare gli stessi vettori di variabili esplicative ($\boldsymbol{x} = \boldsymbol{z}$) ha come scopo quello di identificare i possibili differenti ruoli della stessa variabile in ogni stadio di cui si compone il modello inflazionato;

2. π e μ potrebbero essere funzione l'uno dell'altro.

Le stime di massima verosimiglianza per $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ si possono ricavare attraverso l'impiego dell'algoritmo EM. Secondo questa tecnica, l'obbiettivo è quello di semplificare l'espressione della funzione di log-verosimiglianza derivante dal modello inflazionato introducendo delle variabili di comodo che valgono 0 o 1 a seconda che lo zero provenga dal modello di presenza/assenza o dal modello di conteggio. L'inferenza riguardante i parametri si basa sulla teoria asintotica degli stimatori di massima verosimiglianza (Lambert, 1992).

3 Approccio dinamico per la modellazione di dati di conteggio con eccesso di zeri

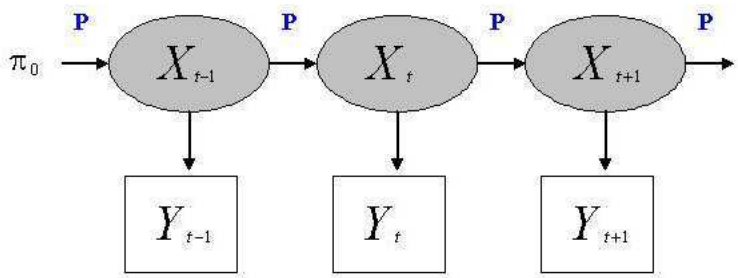
In questo paragrafo verranno descritti i modelli *Hidden* Markov (HMM) dapprima in generale, distinguendo tra HMM in tempo discreto e HMM in tempo continuo, e successivamente si parlerà di HMM con eccesso di zeri.

Nei modelli *Hidden* Markov la distribuzione di probabilità delle osservazioni di una serie storica in ogni istante temporale è unicamente determinata dallo stato (latente) corrente, che è una catena di Markov. Ovvero:

1. $X_1, X_2, \ldots, X_T$ è una catena di Markov di primo ordine in base alla quale la distribuzione della probabilità condizionata di uno stato futuro (X_{T+1}), dati lo stato corrente e quelli passati, dipende unicamente dallo stato corrente;

2. Y_t ($\forall t = 1, 2, \ldots, T$) condizionatamente allo stato latente è indipendente dagli stati passati e dipende unicamente dalla variabile X_t latente.

Una possibile rappresentazione grafica di quanto detto è la seguente:



Nel caso di un processo markoviano latente omogeneo rispetto al tempo o catena di Markov omogenea, le probabilità di transizione tra gli stati sono costanti:

$$p_{ij} = P(X_{t+1} = i|X_t = j) = P(X_t = i|X_{t-1} = j) \quad \forall t, \forall i, j$$

Generalmente il numero degli stati viene deciso a priori, sulla base della tipologia dei dati e delle ipotesi che è possibile formulare a riguardo. In questo lavoro faremo riferimento a un numero di stati finito e perciò discreto, mentre è necessario distinguere HMM in tempo discreto e HMM in tempo continuo.

Relativamente all'insieme di variabili latenti $\{X_t\}$, $t \in T$, quando $T \subseteq \mathbb{N}$ e quindi l'insieme temporale di riferimento è numerabile si parla di HMM in tempo discreto. Si tratta dei modelli *Hidden* Markov classici descritti in MacDonald e Zucchini (1997) e maggiormente diffusi in letteratura. Le prime applicazioni si trovano in ambito ingegneristico e riguardano la teoria dei segnali, mentre più di recente in ambito biomedico sono modelli impiegati per l'identificazione dei geni all'interno delle sequenze di DNA. In questi modelli il tempo non è rilevante o, se presente, la sequenza ordinata sottostante viene convenzionalmente considerata equispaziata.

Un HMM in tempo discreto è definito da:

- il numero degli stati attraverso una variabile (multinomiale) latente X_t;
- la matrice delle probabilità di transizione $\mathbf{P}$, i cui elementi sono $p_{ij} = P(X_t = j|X_{t-1} = i)$ con $p_{ij} \geq 0$ e $\sum_i p_{ij} = 1$;
- le probabilità di transizione iniziali $\pi_0 = P(X_1 = i)$, ovvero la distribuzione iniziale degli stati che dà origine all'intero processo;
- $f(Y_t|X_t; \boldsymbol{\theta})$: distribuzione di probabilità della serie storica osservata, funzione di uno o più parametri incogniti ($\boldsymbol{\theta}$).

Negli HMM in tempo continuo, relativamente all'insieme $\{X_t\}$ $t \in T$ si ha $T \subseteq \mathbb{R}^+$. Va inoltre definita la matrice di intensità di transizione $\mathbf{Q}$, i cui elementi sono i tassi di transizione:

$$q_{ij} = \lim_{\Delta t \to 0} \frac{p_{ij}(t, t + \Delta t)}{\Delta t} \qquad \forall i \neq j$$

Inoltre gli elementi sulla diagonale di $\mathbf{Q}$ sono dati da $q_{ii} = -\sum_{i \neq j} q_{ij}(t)$. Il processo viene detto 'conservativo' poichè $q_{ii} + \sum_{i \neq j} q_{ij}(t) = 0$.

I modelli HMM in tempo continuo si impiegano quando si può avanzare l'ipotesi che il processo latente si evolve in tempo continuo ma è osservato ad intervalli di tempo discreti non equispaziati. Si pensi ad esempio agli studi 'panel' impiegati in medicina. Supponendo di essere interessati allo studio della progressione di una malattia, un paziente viene monitorato in tempi discreti (per ragioni di costi, di scelta del medico, di scelta del paziente stesso) mentre il suo stato di salute si modifica presumibilmente ad istanti che non saranno mai noti. In modelli di questo tipo il tempo occupa un ruolo centrale.

In tale contesto si parla di Processo di Poisson, per il quale un evento si verifica con un tasso $q(t)$, se si verifica un evento nell'intervallo $(t, t + \Delta t)$ con

probabilità pari a $q(t)\Delta(t)$. Se $q(t) = q$, il tempo che intercorre tra due eventi, ovvero il tempo di permanenza in un certo stato, ha distribuzione esponenziale di media $1/q$.

Quindi nell'ottica della catena di Markov di stati latenti, considerando i tempi di permanenza in ogni singolo stato hanno distribuzione esponenziale, mentre si osservano conteggi in corrispondenza della serie storica osservata.

$\mathbf{P}(t)$ si ricava dalle equazioni differenziali di Kolmogorov (Taylor e Karlin, 1994) come:

$$\mathbf{P}(t) = \exp(t\mathbf{Q}) \tag{2}$$

Il valore di $\exp(t\mathbf{Q})$ si ottiene per approssimazione in diversi modi, ad esempio ricorrendo all'espansione in serie di Taylor.

Le singole probabilità di transizione si possono quindi ricavare in funzione dei tassi di transizione come:

$$\begin{aligned} p_{ij}(\Delta t) &= q_{ij}\Delta t + o(\Delta t) \\ p_{ii}(\Delta t) &= 1 + q_{ii}\Delta t + o(\Delta t) \end{aligned}$$

Ovvero le probabilità di transizione sono funzione dei tassi di transizione e di eventuali altri parametri. In presenza di un numero ridotto di stati è possibile ottenere facilmente le p_{ij}.

A volte è realtistico introdurre delle variabili esplicative $\boldsymbol{u}$ costanti nel tempo o dipendenti dal tempo ipotizzando che esse influenzano le intensità di transizione. In questo caso si ha $q_{ij} = \exp(\boldsymbol{\gamma}^T\boldsymbol{u})$. Quest'ultimo caso riguarda gli HMM in tempo continuo e non stazionari, più complicati da trattare da un punto di vista computazionale.

Sui modelli HMM (o in tempo continuo o in tempo discreto) possono essere perseguiti diversi obiettivi inferenziali: inferenza sugli stati non osservati o sulle osservazioni future sulla base della sequenza osservata o inferenza sul modello stesso. Ovvero:

- stima dei parametri incogniti ($\mathbf{P}$ o $\mathbf{Q}, \boldsymbol{\theta}, \pi_0$), vale a dire stima delle probabilità di transizione (o stima delle intensità di transizione per HMM in tempo continuo), dei parametri della funzione di probabilità della serie osservata e della distribuzione iniziale degli stati.

- Inferenza sull'intera sequenza degli stati latenti data la serie storica osservata. Ad esempio si può calcolare la sequenza di stati che massimizza la probabilità $P(x_1, \ldots, x_t, \ldots, x_T|\mathbf{y})$, attraverso l'impiego dell'Algoritmo di Viterbi.

- L'inferenza potrebbe anche essere compiuta su un singolo stato. In particolare, a seconda della sequenza rispetto alla quale si effettua il condizionamento si parlerà di:

  $P(x_t|y_1, \ldots, y_t, \ldots, y_T)$: *smoothing*. In tal caso si calcola la probabilità che all'istante t agisca un certo stato condizionatamente all'intera serie storica osservata.

$P(x_t|y_1, \ldots, y_t)$: *filtering* ($t < T$). In tal caso si calcola la probabilità che all'istante t agisca un certo stato condizionatamente alla serie storica fino al momento t.

$P(x_{T+k}|y_1, \ldots, y_t, \ldots, y_T)$: *prediction* ($k > 0$). In questo caso viene calcolata la probabilità di uno stato al di fuori dell'intervallo di tempo della serie storica osservata.

In questo lavoro vengono trattati modelli *Hidden Markov* per serie storiche di conteggio con eccesso di zeri. Nel caso di serie storiche di conteggio l'inquadramento della problematica secondo un'ottica markoviana o dinamica rappresenta un filone di ricerca piuttosto recente (Wang, 2001). Caratteristica fondamentale della versione dinamica dei modelli con eccesso di zeri è quella di tenere conto contemporaneamente della dipendenza temporale e dell'eccesso di zeri.

Nei suoi due lavori più importanti sull'argomento (Wang (2001),Wang (2003)), Wang parla di modelli MZIP (*Markov zero inflated Poisson*) in tempo discreto e formula le seguenti assunzioni:

1. presenza di due stati sottostanti: perfetto ($X_t = 0$) e imperfetto ($X_t = 1$)[1];
2. condizionatamente allo stato perfetto, l'unica osservazione possibile è quella nulla, mentre condizionatamente allo stato imperfetto si osserva una v.c. Poisson o Binomiale Negativa (il cui v.a. è funzione di variabili esplicative);
3. la probabilità dello stato perfetto vale π mentre quella dello stato imperfetto vale $1 - \pi$.

Le ipotesi che caratterizzano il modello descritto da Wang corrispondono di fatto alle ipotesi della modellazione classica sugli ZIP descritti da Lambert (1992).

La ricerca delle stime di massima verosimiglianza può essere effettuata combinando l'algoritmo EM e l'algoritmo quasi-Newton. Infatti, dal momento che il passo M dell'algoritmo EM dà luogo a delle stime di massima verosimiglianza non ottenibili in forma chiusa, si può ricorrere ad un'approccio di ottimizzazione non lineare (per es. quasi- Newton) per poterle ricavare (Wang, 2001).

Come detto, un modello ZIP con approccio markoviano permette di tenere conto dell'eccesso di conteggi nulli e della correlazione fra le osservazioni tipicamente presente nelle serie storiche.

È quest'ultima caratteristica che distingue l'approccio classico dall'approccio dinamico.

L'approccio classico va impiegato per stimare modelli su un insieme di dati con eccesso di zeri che non costituiscono una serie storica oppure, nel caso di serie storiche, si ricorre ad esso se non è presente correlazione tra le osservazioni. Per questi modelli, l'attenzione è rivolta principalmente al confronto tra modelli

[1] Lambert (1992) parla di perfezione e imperfezione nel contesto dell'industria manifatturiera per cui un processo è perfetto se determina una produzione priva di difetti, viceversa si dirà imperfetto.

hurdle e *zero-inflated* e quindi al diverso modo di considerare il conteggio nullo. Se i dati sono incorrelati la modellazione standard dà luogo a dei risultati più attendibili rispetto all'impiego dell'approccio dinamico. Dalla simulazione di insiemi di dati non correlati e in assenza di sovradispersione, confrontando un modello ZIP con un HMM si osserva generalmente un miglior adattamento del modello con eccesso di zeri e stime di massima verosimiglianza differenti (a volte anche nel segno). Inoltre le righe della matrice $\mathbf{P}$ sono uguali tra loro e il loro valore coincide approssimativamente con le probabilità del modello di mistura (1).

Volendo modellare una serie storica di conteggi che presenta una frequenza eccessiva di conteggi nulli, qualora si decida di impiegare l'approccio standard si potrebbe giungere a delle conclusioni fuorvianti in termini di stima, efficienza, consistenza e significatività di parametri di eventuali variabili esplicative inserite nel modello, se i dati presentano autocorrelazione. Quindi il ricorso a tali modelli determina un presumibile guadagno in termini di valore informativo contenuto nei dati. Sono comunque necessari ulteriori studi di simulazione per verificare le precedenti considerazioni.

Si sottolinea che in questo lavoro l'attenzione viene rivolta a una classe di modelli *Hidden Markov* per serie storiche in tempo continuo in base alla quale la distribuzione condizionata della variabile Y_t dipende dallo stato corrente di una catena di Markov latente non osservabile in tempo continuo. L'approccio metodologico impiegato sfrutta le considerazioni avanzate da Wang nel trattamento di serie storiche di dati di conteggio con eccesso di zeri ma se ne differenzia per il ricorso a HMM in tempo continuo piuttosto che in tempo discreto.

L'impiego di HMM in tempo continuo è stato determinato dalla tipologia dei dati impiegati nel lavoro. Come verrà spiegato nel paragrafo successivo, i dati analizzati riguardano l'inquinamento delle acque. Non è realistico pensare che l'inquinamento si evolva in istanti temporali discreti, mentre sembra più logico pensare che esso si modifichi in un continuum temporale. Le rilevazioni dello stato delle acque vengono comunque effettuate ad istanti temporali discreti e non egualmente spaziati per ragioni economiche, pratiche e di disegno campionario.

La struttura dell'HMM impiegato è identica a quella descritta da Wang. Si ipotizza l'esistenza di due stati $X_t = 0$ e $X_t = 1$, l'uno che determina i conteggi nulli e l'altro che determina il vero e proprio conteggio. La distribuzione della variabile conteggio condizionata allo stato $X_t = 0$ della catena latente attribuisce probabilità uno al valore nullo e la distribuzione condizionata allo stato $X_t = 1$ è una distribuzione Poisson o Binomiale Negativa:

$$\begin{cases} f_0(y_t|\cdot, X_t = 0) = \begin{cases} 1 & \text{se} \quad y_t = 0 \\ 0 & \text{se} \quad y_t > 0 \end{cases} \\ \\ f_1(y_t|\cdot, X_t = 1) \sim \text{Poisson o Binomiale Negativa} \quad \text{se} \quad y_t \geq 0 \end{cases}$$

Gli stimatori di massima verosimiglianza vengono calcolati attraverso l'impiego dell'algoritmo quasi-Newton. Infine non vengono considerate variabili esplicative che agiscono sulle intensità di transizione poichè in questo studio ci si

occupa di HMM stazionari.

Il problema più generale di adattare un HMM in tempo discreto o un HMM in tempo continuo è quello di stabilire se i dati osservati sono determinati da una sequenza di stati latenti che evolve in tempo continuo oppure se la sequenza di stati latenti evolve in tempo discreto e non ha una controparte 'continua'. In particolare la matrice di intensità di transizione **Q** può essere ottenibile dalla matrice delle probabilità di transizione **P** calcolato da un HMM in tempo discreto solo se sono soddisfatte le condizioni di 'embeddability' (Kalbfleish e Lawless, 1985). Tale argomento non riguarda l'obbiettivo principale che ci si è posti ne presente lavoro e verrà approfondito in un lavoro successivo.

4 Uno studio sulla balneabilità

4.1 Tipologia di dati e campionamento

In questo paragrafo vengono confrontati l'approccio classico nella modellazione di dati con eccesso di zeri con il più recente approccio markoviano in uno studio sulla balneabilità della provincia di Rimini negli anni 2004-2005.

Si definiscono **acque balneabili** le acque dolci, le correnti di lago e le acque marine nelle quali la balneazione è espressamente autorizzata ovvero non vietata.

In Italia, il D.P.R. 470/82 indica i valori limite che alcuni 'parametri' (qui indicati come 'variabili') non devono superare affinchè il tratto costiero considerato sia ritenuto balneabile. La direttiva 7/2006/CE è stata introdotta di recente e prevede l'integrazione con la direttiva quadro 2000/60/CE recepita in Italia con il D.Lgs. 152/06 'Testo unico in materia ambientale'. Nei prossimi anni il quadro legislativo di riferimento dovrebbe ulteriormente essere modificato attraverso norme più severe e in accordo ai diversi paesi europei.

Attualmente ogni anno viene emanato un decreto dirigenziale che elenca i tratti di costa da monitorare, generalmente uguali a quelli dell'anno precedente. Per potere monitorare adeguatamente la balneabilità esiste un piano di campionamento adottato dalle Agenzie Regionali per la Protezione Ambiente (ARPA) per cui la raccolta di campioni di acque ha una cadenza prescritta di due campioni al mese per ogni tratto in coincidenza con la stagione balneare, che va dall'1 aprile al 30 settembre di ogni anno. Generalmente i dati vengono raccolti a metà e a fine mese. In pratica, per motivi meteorologici o tecnici, questa cadenza può non essere rispettata, determinando uno slittamento di qualche giorno nel prelievo del campione. I prelievi vengono effettuati ad una profondità di circa 30 cm sotto il pelo libero dell'acqua e ad una distanza dalla battigia tale che il fondale abbia una profondità di 80-120 cm.

I valori limite riguardano le seguenti variabili: coliformi totali, coliformi fecali, streptococchi fecali, colorazione e trasparenza, ossigeno disciolto, pH, oli minerali, fenoli, sostanze tensioattive, più raramente le salmonelle. Le acque vengono giudicate idonee, se i valori registrati nei campioni sono conformi ai valori limite nel 90% dei casi o, in caso di non conformità, quando essi non si discostano più del 50% dai valori limite (80% per coliformi e streptococchi).

Vengono raccolti 5 campioni suppletivi se si registra il superamento dei limiti. Se uno dei 5 campioni conferma i risultati, la zona è vietata alla balneazione.

In questo lavoro l'interesse di studio si è rivolto al **numero di streptococchi fecali** in 100 ml di acqua, una delle tre variabili, insieme ai coliformi fecali e totali, che, se superato, viene considerato un indice di inquinamento. Infatti sono proprio valori elevati di una di queste variabili ad indicare l'esistenza di alterazioni ambientali dovute a precise cause di inquinamento. In particolare, se le acque esaminate non sono inquinate sono assenti le colonie di coliformi e di streptococchi, ma se sono presenti si assiste generalmente ad una loro crescita esponenziale.

Si vuole in particolare modellare la relazione che sussiste tra la variabile risposta di conteggio, numero di streptococchi fecali, ed alcune variabili esplicative che possono essere causa del conteggio registrato. In questo studio verranno considerate le seguenti variabili esplicative: il mese, variabile categoriale corrispondente a uno dei 6 mesi della stagione balneare, la temperatura dell'acqua, l'ossigeno e le condizioni del mare.

4.2 Risultati a confronto

Lo studio in esame è stato condotto sul biennio 2004-2005 nella provincia di Rimini. I conteggi pari a zero di streptococchi fecali sono pari al $\simeq 61\%(n = 2567)$.

Nel primo lavoro (Viviano *e altri*, 2005) i dati della balneabilità sono stati modellati secondo l'approccio descritto nel paragrafo 2. L'attenzione si è focalizzata sul numero di streptococchi fecali indipendentemente dal tempo in cui si era registrato quel valore e l'informazione temporale era stata inserita come una variabile esplicativa di tipo categoriale, ipotizzando che il mese della stagione balneare potesse avere influenza sui conteggi osservati.

In questo lavoro i dati vengono modellati sia secondo l'approccio classico nella stessa ottica del lavoro precedente sia secondo l'approccio dinamico, poichè è presente l'informazione temporale che riguarda due prelievi di acqua nei mesi della stagione balneare in due anni contigui.

I risultati relativi alla metodologia standard sono stati ottenuti attraverso l'impiego della libreria `pscl` di R.

Confrontando tutti e sei i modelli derivanti dall'equazione (1), sulla base del criterio di confronto AIC che tiene conto della verosimiglianza e del numero dei parametri considerati (e quindi della parsimonia), il miglior modello è HNB (Tabella 1), avendo l'AIC notevolmente più basso degli altri modelli.

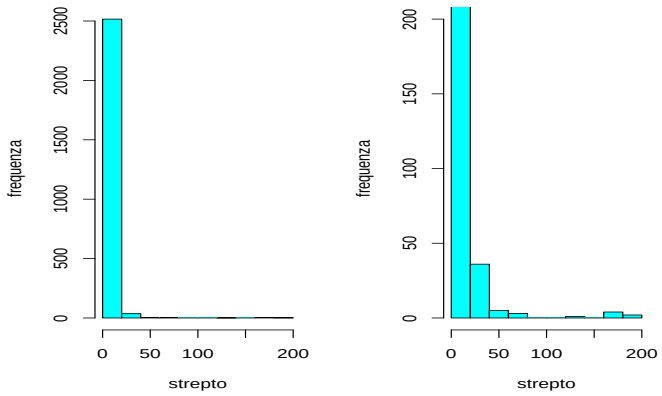
Questo risultato è confermato anche dal Test di Vuong (Vuong, 1989) calcolato come confronto a coppie per tutti i modelli descritti. Il modello HNB sembrerebbe il miglior compromesso per riuscire a cogliere la sovradispersione e l'eccesso di zeri, presenti nei dati e osservabile nella figura (1).

Si osserva infatti un picco in corrispondenza dei valori più bassi del numero di streptococchi determinato proprio dal 61 % delle osservazioni presenti per $x = 0$. Inoltre il range dei conteggi si estende fino a valori di $x = 200$. In particolare, nel grafico di sinistra si osserva l'istogramma con riferimento alle frequenze assolute della variabile in esame. Il grafico di destra riporta lo stesso istogramma

Tabella 1: Criterio di informazione di Akaike per il confronto fra modelli inflazionati

	AIC
Poisson	24866
Binomiale Negativa	8154
ZIP	17166
ZINB	8144
HP	17166
HNB	**8052**

Figura 1: Distribuzione del numero di streptococchi fecali (Rimini 2004-2005)



con l'asse delle ordinate ridotto per potere osservare meglio la distribuzione di tutti i conteggi.

I fenomeni della sovradispersione e dell'eccesso di zeri spesso si presentano contemporaneamente e possono essere la conseguenza di eterogeneità non osservata causata, ad esempio, dall'avere omesso delle variabili esplicative in fase di raccolta dati o dall'effetto di variabili esplicative che varia tra i soggetti campionati. In tali casi, i modelli ZINB e HNB possono rappresentare la migliore soluzione per modellare i dati poichè da una parte catturano il picco in corrispondenza del conteggio nullo e dall'altra riescono a cogliere l'allungamento dei dati sulla coda destra della distribuzione, determinato dalla sovradispersione.

I risultati del modello additivo (in assenza di interazioni) adattato sono riportati nella Tabella 2 dove le variabili `month` sono variabili categoriali corrispondenti ai mesi da Aprile a Settembre; le variabili `sea` si riferiscono alle categorie della 'Condizione del Mare' (rispettivamente 'calmo', 'quasi mosso', 'mosso') e `temp` e `oxy` corrispondono alle variabili continue 'Temperatura dell'acqua' e 'Ossigeno'.

L'inserimento delle variabili esplicative all'interno delle due funzioni di cui si compone il modello inflazionato permette di ragionare secondo la logica dei

Tabella 2: Risultati dell'adattamento del modello HNB

	Componente 'presenza-assenza'		**Componente di conteggio**	
	Stima (s.e)	p-value	Stima (s.e)	p-value
`month2`	-.734(.0.206)	**.002**	.490 (.318)	.123
`month3`	-1.351 (.335)	**.000**	1.630 (.489)	**.000**
`month4`	-1.916 (.425)	**.000**	.815 (.572)	.154
`month5`	-1.710 (.428)	**.000**	.878 (.599)	.143
`month6`	-1.747 (.371)	**.000**	-.154 (.516)	.765
`sea2`	.687 (.122)	**.000**	.135 (.189)	.476
`sea3`	1.580 (.655)	.094	.407 (.652)	.532
`temp`	.124 (.032)	**.005**	-.137 (.043)	**.002**
`oxy`	-.006 (.003)	**.000**	-.001 (.005)	.788

GLM, come spiegato nella sezione 2. Relativamente alla componente del modello inflazionato di presenza-assenza di conteggio nullo (quella che si riferisce ai valori di $y = 0$), la probabilità di avere un conteggio pari a zero rispetto ad un conteggio maggiore di zero (modello logit) si riduce significativamente in corrispondenza di tutte le variabili eccetto la temperatura (che ha segno opposto) e la condizione del mare 'mosso' (che non influenza significativamente i conteggi). Rispetto alla componente di conteggio (modello log-lineare), nel mese di Giugno aumenta significativamente la probabilità di inquinamento se confrontato con il mese di Aprile (secondo la parametrizzazione di tipo 'corner point'), inquinamento che tende a ridursi significativamente all'aumentare della temperatura.

Il modello *Hidden* Markov in tempo continuo è stato stimato attraverso l'impiego della libreria `msm` di R. I risultati del modello con AIC più basso(4774) sono riportati nella Tabella 3. Accanto alla stima viene riportato l'intervallo di confidenza a livello di fiducia del 95%.

Il modello adattato presuppone che lo stato latente $X_t = 0$ determini tutti i conteggi nulli (modellati attraverso la funzione identità pari a zero) e che lo stato latente $X_t = 1$ determini una parte dei conteggi nulli e i conteggi maggiori di zero (modellati attraverso la funzione Binomiale Negativa, poichè i dati si presentano sovradispersi).

Anche in questo caso, sono stati considerati modelli senza interazione. Il miglior modello ha un AIC pari a 4774, che confrontato all'AIC dell'HNB ($AIC = 8052$) è notevolmente inferiore. Come si osserva in Tabella 3, i mesi di Maggio, Luglio e Agosto determinano una riduzione significativa del numero di streptococchi fecali rispetto al mese di Aprile, così come le condizioni del mare 'mosso'. Mentre aumenta significativamente il numero di streptococchi nel mese di Settembre e all'aumentare della quantità di ossigeno presente.

In termini di AIC il miglior modello è l'HMM rispetto all'HNB. Questo risultato riflette la capacità che ha tale modello di cogliere almeno tre caratteristiche presenti nei dati: la percentuale elevata di conteggi nulli, l'autocorrelazione delle

Tabella 3: Risultati dell'adattamento di HMM

	Stima	l95	u95
`disp`	.467	.422	.517
`prob`	.113	.101	.125
`month2`	**-.158**	-.550	.233
`month3`	-.831	-1.443	-.220
`month4`	**-.115**	-.852	.621
`month5`	**-209**	-.971	.552
`month6`	**.519**	-.141	1.180
`oxy`	**.006**	-.0003	.013
`temp`	.070	.014	.125
`sea2`	-.272	-.519	-.025
`sea3`	**-.531**	-1.423	.366

osservazioni e la sovradispersione.

Sempre rispetto all'HMM descritto si ottengono i seguenti risultati in termini di matrice di intensità di transizione stimata:

$$\hat{\mathbf{Q}} = \begin{bmatrix} 0 & 1.266 \\ 1.402 & 0 \end{bmatrix}$$

In un HMM il tempo di permanenza in un determinato stato ha distribuzione esponenziale avente un valore atteso pari al reciproco dell'elemento considerato nella matrice $\mathbf{Q}$.

Mentre la matrice di probabilità di transizione stimata nell'unità di tempo è:

$$\hat{\mathbf{P}}(T=1) = \begin{bmatrix} .558 & .442 \\ .489 & .511 \end{bmatrix}$$

Ogni elemento di $\mathbf{P}$ è calcolato analiticamente in termini di $\mathbf{Q}$. Si può osservare che la probabilità di permanere nello stesso stato è sempre più alta di quella di passare da uno stato all'altro (gli elementi sulla diagonale principale di $\mathbf{P}$ sono più elevati di quelli sulla diagonale secondaria).

5 Conclusioni

Il lavoro ha messo in luce le principali caratteristiche strutturali della modellazione classica e di quella dinamica nel trattamento di dati di conteggio con eccesso di conteggi nulli. I due approcci corrispondono a tipologie di dati differenti. Se le serie storiche di conteggi vengono trattate tralasciando l'informazione temporale, l'unica modellazione possibile è rappresentata dall'approccio classico. Se però i dati presentano correlazione determinata dalla dipendenza temporale del fenomeno, l'approccio dinamico raggiunge il duplice obbiettivo di riuscire a tenere conto della correlazione e dell'eccesso di zeri.

Ulteriori studi su insiemi di dati con simili strutture sono necessari per valutare limiti e potenzialità dei due modelli. Si può anche pensare di introdurre

delle estensioni attraverso l'impiego di ZIP spazio-temporali (Agarwal *e altri*, 2002) o ZIP con effetti casuali (Dobbie e Welsh (2001),Hall e Berenhaut (2002)).

Riferimenti bibliografici

Agarwal D.; Gelfand A.; Citron-Pousty S. (2002). Zero-inflated models with application to spatial count data. *Journal Environmental and Ecological Statistics*, **9**, 341–355.

Cameron A.; Trivedi P. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Dietz E.; Böhning D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics & Data Analysis*, **34**, 441–459.

Dobbie M. J.; Welsh A. (2001). Modelling correlated zero-inflated count data. *Australian & New Zealand Journal of Statistics*, **43**, 431–444.

Hall D.; Berenhaut K. (2002). Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. *The Canadian Journal of Statistics*, **30**, 414–430.

Kalbfleish J.; Lawless J. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, **80**, 863–871.

Lambert D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

MacDonald I.; Zucchini W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman & Hall.

Taylor H.; Karlin S. (1994). *An Introduction to Sthocastic Modeling*. Academic Press.

Viviano L.; Muggeo V.; Lovison G. (2005). Zero-inflated models to analyze environmental data sets with many zeroes. *Atti della Riunione Intermedia SIS 2005*.

Vuong Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307–333.

Wang L. (2004). *Parameter estimation for mixtures of Generalized Linear Mixed-Effects models*. Tesi di Dottorato di Ricerca, Università della Georgia.

Wang P. (2001). Markov zero-inflated Poisson regression models for a time series of counts with excess zeroes. *Journal of Applied Statistics*, **28**, 623–632.

Wang P. (2003). A bivariate zero-inflated negative binomial regression model for count data with excess zeros. *Economics Letters*, **78**, 373–378.

Zorn C. (1996). Evaluating zero-inflated and Hurdle Poisson specifications. *Midwest Political Science Association*, **18-20 april**, 1–16.