Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Controllo statistico di processo per dati di conteggio con sovradispersione: il caso dell'epidemia di dengue in Ecuador

Relatore: Prof. Guido Masarotto
Dipartimento di Scienze Statistiche

Laureanda: Alice Berto
Matricola n. 2055032

Anno Accademico 2023/2024

Indice

Elenco delle tabelle

Elenco delle figure

Introduzione

Le carte di controllo sono uno strumento cardine dello *statistical process control* (SPC), utili per monitorare ed individuare un cambiamento nel processo sotto sorveglianza. In seguito ad un segnale è importante capire quando è avvenuto il cambiamento, cosa è cambiato e come per individuare al meglio la presenza di possibili cause ed eliminarle.
Nella pratica, quando si ha a che fare con dati di conteggio, è d'abitudine prendere a riferimento il modello di Poisson, anche se non risulta sempre adeguato a causa del fenomeno della *sovradispersione*, situazione in cui la varianza osservata è maggiore di quella teorica assunta dal modello. In questi casi è possibile adottare il modello binomiale negativo, un modello più flessibile grazie all'aggiunta di un parametro che controlla la variabilità. Si veda Fávero et al. (2021) per una panoramica sui dei modelli di regressione per dati di conteggio e del fenomeno di sovradispersione.
Negli ultimi anni è cresciuto l'utilizzo delle carte di controllo oltre che in ambito industriale anche in campo sanitario, in cui il problema della sovradispersione è molto comune. Al fine di trattare questa situazione, in letteratura sono presenti diversi articoli che si occupano di studiare applicazioni ipotizzando la distribuzione binomiale negativa per i dati analizzati, come in Alencar et al. (2017) e in Urbieta et al. (2017), dove si sono monitorate le ospedalizzazioni giornaliere con effetti stagionali di pazienti affetti da malattie respiratorie, mettendo a confronto la performance di carte CUSUM ed EWMA costruite con diverse statistiche di controllo.
In Sparks et al. (2011) è stata proposta una carta EWMA basata sugli errori di previsione per monitorare e segnalare in modo tempestivo la presenza di epidemie stagionali, come malaria o influenza, concentrandosi sull'ipotesi di distribuzione non omogenea, cioè con incidenze medie di casi variabili nel tempo.
In Ali et al. (2020) si affronta in termini di falsi allarmi la conseguenza della stima dei parametri in controllo dalla Fase I, proponendo una carta CUSUM ed EWMA risk-adjusted applicate a casi di pazienti con problemi cardiaci e respiratori.
Infine si può citare anche Albers (2011) per la costruzione di una carta di controllo utile in campo medico in cui si suppone che il processo vada fuori controllo abbastanza raramente e che la sovradispersione sia invece molto comune.

L'obiettivo di questa tesi è dimostrare come la sorveglianza di dati di conteggio attraverso l'assunzione di un modello binomiale negativo sia appropriato in presenza di *sovradispersione*, prendendo a verifica il caso pratico di García-Bustos and Zambrano (2022). Questo articolo tratta la sorveglianza dei casi di infezione dovuti all'epidemia di dengue in Ecuador, un fenomeno attuale non circoscritto, ma che provoca focolai in molti Paesi. Con un occhio già rivolto al caso pratico quindi anche gli argomenti teorici che saranno affrontati in ambito univariato saranno focalizzati alla sorveglianza unilaterale superiore, nel caso di osservazioni singole raccolte nel tempo.
Si veda anche Chen et al. (2020) o Wang and Zwetsloot (2023) per altri esempi di trattazione del problema pratico appena citato.
La struttura dell'elaborato è la seguente:
Nel primo capitolo viene introdotto il controllo statistico di processo, con a seguire un breve richiamo sulle più semplici carte per dati discreti/di conteggio. Uno spazio è dedicato anche alla sovradispersione e alla distribuzione di probabilità binomiale negativa.
Nel secondo capitolo viene presentata una rassegna di proposte presenti in letteratura sulla sorveglianza di dati che presentano sovradispersione, dedicando un piccolo spazio anche alle relative applicazioni pratiche.
Il terzo capitolo è dedicato ad estensioni del precedente con proposte di miglioramento riguardo aspetti poco approfonditi. In particolare verranno proposti i limiti dinamici, limiti variabili con il tempo adatti per la loro applicabilità a diversi contesti e semplicità di calcolo rispetto ad altre alternative. Verrà poi dato spazio alla presentazione delle carte di controllo multivariate per la sorveglianza congiunta di parametri, in particolare analizzando la costruzione della carta MEWMA basata sullo score e le dirette alternative, le carte combinate.
Il quarto e ultimo capitolo è dedicato all'implementazione delle carte per la sorveglianza dei casi di dengue sia rielaborando le proposte proprie di alcuni autori viste nel secondo capitolo sia applicando i miglioramenti suggeriti nel terzo capitolo. Un confronto diretto sarà possibile dall'osservazione delle carte.
Infine si trarranno le conclusioni.
In Appendice è riportato tutto il codice R utilizzato per l'implementazione delle carte nel quarto capitolo.

1 Concetti di base

In questo capitolo introduttivo verranno per prima cosa ricordate le motivazioni e gli strumenti alla base del controllo statistico di processo. Essendo un progetto improntato sulla sorveglianza di dati di conteggio, l'attenzione verrà poi posta unicamente sulle prime e più semplici carte costruite a questo scopo, per continuare poi con l'analisi di un fenomeno che affligge molti casi reali, la sovradispersione. Per concludere verrà trattata dal punto di vista teorico la distribuzione binomiale negativa, adatta al contesto di studio.

1.1 Controllo statistico di processo

Al giorno d'oggi la chiave per essere competitivi è racchiusa nella capacità di soddisfare i bisogni e le aspettative del cliente, fornendo prodotti e servizi di qualità ad un prezzo contenuto e nei tempi prestabiliti.
A questo scopo per prendere le decisioni più adeguate è utile, se non necessario, raccogliere dati e analizzarli. In questo contesto arrivano in aiuto gli strumenti del controllo statistico di processo (SPC), supportando il *decision-maker* nel valutare se il processo sta operando a standard accettabili.
Strumenti alla base del SPC sono le carte di controllo che permettono di monitorare la conformità di un prodotto e di segnalare, il prima possibile, se il processo è andato fuori controllo.
La qualità dei processi è inversamente proporzionale alla variabilità, generata principalmente da due fonti, le *cause comuni* e le *cause speciali*. In caso di allarme è importante individuare tempestivamente ed eliminare successivamente le fonti dell'aumento della variabilità indesiderata, le *cause speciali*, raggruppabili in *uomo, macchina, metodo, ambiente, materiale*.
La variabilità prodotta dalle *cause comuni* è invece intrinseca del processo ed impossibile da eliminare, a meno che non si cambi il processo stesso.
Le fasi chiave nella sorveglianza in ambito SPC sono ben rappresentate nella figura (1.1). Per un approfondimento si veda Madanhire and Mbohwa (2016).
Anche se le carte sono nate in un contesto industriale nulla vieta di applicarle al contesto medico, come sarà ben trattato a seguire. E' utile ricordare anche che l'utilizzo delle carte di controllo coinvolge solitamente due fasi: la fase I,

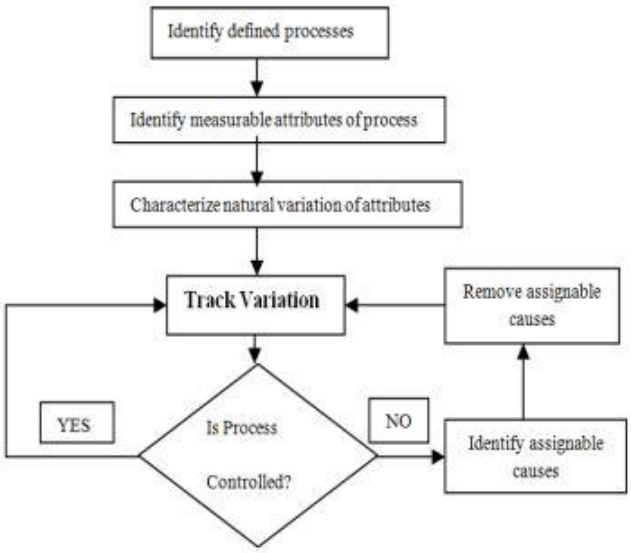

Figura 1.1: Step implementazione SPC. Fonte: Madanhire and Mbohwa (2016)

in cui si attua una sorveglianza di tipo retrospettiva con l'obiettivo principale di verificare se il processo è stato o meno in controllo nel tempo considerato e quindi stimare la distribuzione in controllo, necessaria per l'applicazione della fase II. Durante la fase II la sorveglianza è invece prospettica, i dati sono raccolti in modo sequenziale e si verifica attraverso una serie di test d'ipotesi se il processo è ancora in controllo ai nuovi istanti di tempo o se è già andato fuori controllo.

1.2 Classiche carte di controllo per dati qualitativi e/o di conteggio

Nell'ambito del controllo statistico di processo è usuale parlare di carte di controllo per *variabili* quando la caratteristica di qualità sorvegliata è numerica (diametro, peso...) o per *attributi* quando la caratteristica di qualità osservata è qualitativa (presenza/assenza di un difetto, numero di falsi,...).
Walter Shewhart (1891–1967) è considerato l'iniziatore del controllo della qualità moderno, introducendo a partire dal 1924 le prime carte di controllo per sorvegliare tutte le fasi del processo di produzione e non solo il prodotto finito. Al giorno d'oggi le sue carte sono ancora molto utilizzate, in particolare per dati discreti/di

conteggio sono ben note le seguenti carte:

- `Carta p` che sorveglia la proporzione di prodotti non conformi $p = X/m$, con $X \sim Binomiale(m, \pi)$, nei campioni raccolti nel tempo;
- `Carta mp` che sorveglia direttamente la frequenza dei prodotti non conformi X, con $X \sim Binomiale(m, \pi)$, nei campioni raccolti nel tempo;
- `Carta c` sorveglia il numero c di difetti individuati nell'unità di prodotto raccolta nel tempo, con $c \sim Poisson(\lambda)$;
- `Carta u` sorveglia il numero medio di difetti c per m unità ispezionate nel tempo, cioè $u = c/m$. Si pone $c \sim Poisson(m\lambda)$ per la singola unità;
- `Carta D` che sorveglia nel tempo il numero totale di difetti c_j^* pesati per le differenti gravità che rappresentano, con $c_j^* \sim Poisson(\lambda_j^*)$ e $j = 1, 2, ..., k$ diversi gruppi di gravità.

Per un maggior approfondimento sulla costruzione delle suddette carte si veda Qiu (2014).
Altre carte molto popolari nell'ambito dello SPC sono la carta CUSUM (Page (1954)) ed EWMA (Roberts (1959)), due carte che a differenza della Shewhart cumulano le osservazioni nel tempo, diventando più efficienti a individuare cambiamenti piccoli della quantità d'interesse. Per cambiamenti grandi si può invece dimostrare che le tre carte offrono performance molto simili.
Nel seguito ci concentreremo sulla sorveglianza di dati di conteggio con distribuzione riconducibile alla Poisson, ma argomentando l'inadeguatezza di quest'ultima in presenza di elevata variabilità.

1.3 Modelli per dati di conteggio e sovradispersione

Il modello di regressione Poisson è comunemente applicato per modellare dati di conteggio, dove il comportamento della variabile dipendente può essere spiegato da predittori sia qualitativi che quantitativi. Questo modello assume l'identità tra media e varianza della variabile risposta, una semplificazione che può essere non adeguata per molte applicazioni reali.
Il problema della sovradispersione si manifesta nel caso in cui la varianza della

variabile dipendente, condizionatamente ai predittori, risulta statisticamente maggiore della corrispondente media. Le cause di ciò possono essere la presenza di eterogeneità nella popolazione, mancanza di rilevanti predittori nel modello, la presenza di outliers, correlazione, inflazione di zeri o altre ragioni. Molte volte andando ad agire su questi aspetti la sovradispersione può essere corretta, mentre in altri casi è intrinseca nei dati stessi, facendosi si che non esista nessun rimedio esterno per correggerla se non adeguatamente gestirla.
Come affermano Milanzi and Molenberghs (2012), ignorare la presenza della sovradispersione ha generalmente un effetto debole sulle stime dei parametri β, mentre porta a sottostimare in maniera rilevante gli errori standard degli stimatori $\hat{\beta}$ e quindi alla scorretta identificazione della significatività dei termini di β, che in molti casi è di primo interesse.
Come test per saggiare l'adeguatezza del modello di Poisson contro alternative di sovradispersione si può utilizzare la statistica di Pearson, equivalente asintoticamente alla devianza nulla del modello Poisson dove si confrontano le n frequenze osservate y_i con il valore atteso $\bar{y}$, pari a

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{\bar{y}} \ , \tag{1}$$

con distribuzione nulla approssimata χ^2_{n-1}. La statistica $X^2/(n-1)$ è il rapporto tra la varianza campionaria corretta e la media campionaria, nota come indice di dispersione di Fisher.

Nel seguito verrà introdotto il modello binomiale negativo, una scelta adeguata per incorporare la presenza di sovradispersione.

1.4 Distribuzione binomiale negativa

La distribuzione binomiale negativa è una distribuzione discreta che può essere espressa in due modi, o come la probabilità di avere n prove per ottenere k successi o come la probabilità che si verifichino esattamente y_t fallimenti prima di ottenere un totale di k successi, nello specifico un successo nella prova numero $k + y_t$ ed esattamente y_t fallimenti e $k - 1$ successi nelle prove precedenti. Le due versioni sono equivalenti con i dovuti adattamenti ed è per questo che si è scelto

di concentrarsi nella seconda versione, la cui densità discreta si esprime come

$$P(Y_t = y_t) = \left(\frac{k + y_t - 1}{y_t}\right) p^k (1-p)^{y_t}, \tag{2}$$

con $p \in [0, 1]$, la probabilità di successo.
Una rappresentazione alternativa della distribuzione è quella di considerare Y_t come una mistura di Poisson con parametro λ, quest'ultimo distribuito a sua volta come una variabile Gamma(k, μ). Per dimostrare che la distribuzione binomiale negativa coincide con una distribuzione Gamma-Poisson solo con diversa parametrizzazione si assume che Y_t segua una distribuzione di Poisson(λ_t), dove $\lambda_t > 0$ denota il tasso medio degli eventi nell'unità di tempo. La distribuzione di probabilità di Y_t è una funzione discreta definita per $y_t \in [0, 1, 2, ...]$ data da

$$P(Y_t = y_t) = \frac{e^{-\lambda_t} \lambda_t^{y_t}}{y_t!}, \tag{3}$$

supponendo $\lambda_t = \lambda$ per ogni t, si ha che E[y_t]=Var[y_t]=λ.
Almeno in alcune applicazioni, può essere ragionevole assumere che λ_t sia distribuita come una variabile aleatoria Gamma(k, β), dove $\beta = \frac{\mu}{k}$, con $k > 0$ parametro di forma e β parametro di scala. La Gamma è una variabile continua con densità pari a

$$f(\lambda_t | k, \beta) = \frac{\lambda_t^{k-1} e^{\frac{-\lambda_t}{\beta}}}{\Gamma(k) \beta^k}, \tag{4}$$

con i primi due momenti uguali a E[λ_t]=μ e Var[λ_t]=$\frac{\mu^2}{k}$.

La densità congiunta è quindi data da

$$\begin{aligned} f(y_t, \lambda_t) &= f(y_t | \lambda_t) f(\lambda_t) \\ &= \frac{e^{-\lambda_t} \lambda_t^{(y_t + k - 1)} e^{-\frac{k\lambda_t}{\mu}} (\frac{\mu}{k})^{-k}}{\Gamma(y_t + 1)\Gamma(k)} \end{aligned} \tag{5}$$

e la densità marginale di y_t si ottiene come

$$\begin{aligned} f(y_t) &= \int_0^\infty f(y_t, \lambda_t) d\lambda_t \\ &= \int_0^\infty \frac{e^{-\lambda_t} \lambda_t^{(y_t+k-1)} e^{-\frac{k\lambda_t}{\mu}} (\frac{\mu}{k})^{-k}}{\Gamma(y_t+1)\Gamma(k)} d\lambda_t \\ &= \frac{\Gamma(y_t+k)}{\Gamma(y_t+1)\Gamma(k)} \left(\frac{\mu}{k+\mu}\right)^{y_t} \left(\frac{k}{k+\mu}\right)^{k}. \end{aligned} \tag{6}$$

Dalla distribuzione di probabilità in (2) è quindi facile ricondursi a quella di una mistura Gamma-Poisson ponendo $\left(\frac{y_t+k-1}{y_t}\right) = \frac{\Gamma(y_t+k)}{\Gamma(y_t+1)\Gamma(k)}$ attraverso lo sviluppo del coefficiente binomiale, mentre le altre quantità pari a $(1-p) = \frac{\mu}{k+\mu}$ e $p = \frac{k}{k+\mu}$.
I momenti della variabile Y_t corrispondono a E[Y_t]=μ e Var[Y_t]=$\mu + \frac{\mu^2}{k}$,
con k^{-1} parametro di dispersione.
Si può facilmente notare come la distribuzione di Poisson sia la distribuzione limite quando k tende a infinito, se $k = 1$ si riduce alla distribuzione geometrica e per valori sempre più piccoli aumenta il fenomeno di sovradispersione. A questo punto è possibile delineare la regressione binomiale negativa che, facendo parte dei modelli lineari generalizzati, prevede una funzione di legame tra la media del processo e i predittori. In particolare, utilizzando come funzione di legame il logaritmo si ottiene quindi $\log(\mu) = x^T\beta$. Si dimostra che, soltanto per k fissato, la (6) appartiene alla famiglia di dispersione esponenziale, scrivibile come

$$P(y_t; \theta_t, \phi) = \exp\left\{\frac{\theta_t y_t - b(\theta_t)}{a_t(\phi)} + c(y_t, \phi)\right\}, \tag{7}$$

con $\theta_t \in R$, detto parametro naturale, e $\alpha_t(\phi) > 0$. Spesso $\alpha_t(\phi) = \phi$, con ϕ parametro di dispersione e si può dimostrare che

$$\begin{aligned} \mu(\theta_t) &= E(Y_t) = b'(\theta_t), \\ Var(Y_t) &= a_t(\phi) b''(\theta_t) = a_t(\phi) v(\mu_t), \end{aligned} \tag{8}$$

dove $v(\mu_t) = b''(\theta_t)$ è detta funzione di varianza.
Sia $Y_t \sim Bineg(k, \mu_t)$, allora partendo dalla (6) si ha

$$P(y_t; \theta_t, \phi) = \exp\left\{y_t \log\left(\frac{\mu_t}{k+\mu_t}\right) + k \log\left(\frac{k}{k+\mu_t}\right) + c(y_t, k)\right\} \tag{9}$$

e la corrispondenza con (7) si ottiene ponendo $\theta_t = \log(\mu_t/(k+\mu_t))$, $b(\theta_t) = -k\log(1-e^{\theta})$ e $a_t(\phi) = \phi = 1$.
Si veda Salvan et al. (2020) per una trattazione più approfondita dei modelli lineari generalizzati appartenenti alla famiglia di dispersione esponenziale.
Per quanto riguarda la stima dei parametri d'interesse, le equazioni di verosimiglianza per un modello GLM non ammettono quasi mai soluzione esplicita, quindi è necessario l'utilizzo di uno specifico algoritmo numerico. Quest'ultimo è basato su una particolare versione dell'algoritmo di Newton-Raphson ed è detto dei minimi quadrati pesati iterati (IRLS) per l'analogia esistente con il metodo dei minimi quadrati pesati utilizzato nel modello lineare normale.
Le stime finali vengono ottenute tramite procedure di stima iterata, fino al raggiungimento della convergenza. Si veda Hinde and Demétrio (1998) per più dettagli nei passaggi di convergenza in presenza di sovradispersione.
Autori come Lloyd-Smith (2007) hanno posto particolare attenzione nell'individuare k, il parametro di dispersione, la cui stima è ancora considerata una sfida in dataset con numerosità contenuta. In letteratura diversi studi di simulazione hanno esaminato l'efficacia di stimatori per la stima dei parametri della binomiale negativa, ma la maggior parte di questi si sono focalizzati su $k \geq 1$, quindi non in presenza di elevata sovradispersione.
Nel suddetto articolo si è oltrepassato questo limite approfondendo la stima di massima verosimiglianza (ML) di k in caso di elevata sovradispersione ($k < 1$), analizzando l'accuratezza degli intervalli di confidenza ottenuti da queste stime ed esaminando potenziali *bias* dovuti ai metodi e agli errori nella raccolta dei dati, con applicazioni a dati di natura epidemiologica nello specifico. In sintesi, lo studio dimostra come ci sia un minimo rischio di sottostimare k, quindi sovrastimare il grado di sovradispersione nei dati in caso di campioni limitati, mentre è presente il rischio contrario, quindi di sovrastimare k, con campioni piccoli o quando la classe di zeri è sistematicamente sotto rappresentata.

2 Background teorico

In molti contesti applicativi, a causa della sovradispersione, le carte di controllo brevemente richiamate nella sezione 1.2 non sono adeguate. Ad esempio, questo succede spesso in ambito sanitario quando vengono sorvegliati tassi di infezioni, numero di morti, ecc., per cui le ipotesi di una distribuzione di Poisson non reggono. Nelle sezioni seguenti verranno presentate alcune delle soluzioni teoriche individuate negli ultimi anni per inglobare il fenomeno della sovradispersione nell'operazione di sorveglianza prospettica dei processi d'interesse. Nella sezione finale verranno riportate le applicazioni pratiche degli autori di alcune proposte analizzate.

2.1 CUSUM ed EWMA per variabili di conteggio

Come già accennato, le carte EWMA e CUSUM offrono prestazioni competitive nell'ambito della sorveglianza di processi, differenziandosi principalmente per il fatto che la CUSUM discende da un risultato teorico, cioè l'applicazione del lemma fondamentale di Neyman-Pearson agli schemi di sorveglianza, mentre la carta EWMA risulta più intuitiva, essendo essenzialmente una media mobile ponderata delle osservazioni passate.
Ricordando la costruzione delle carte, nella CUSUM le statistiche di controllo usate per individuare aumenti del parametro d'interesse corrispondono a

$$C_{i,t} = \max(0, C_{i,t-1} + Z_{i,t} - K), \tag{10}$$

con $C_{i,0} = 0$ e $Z_{i,t}$ la i-esima statistica da cumulare al tempo t. I valori di h_i, limite superiore, e K, parametro di controllo, sono scelti per soddisfare un certo ARL in controllo e minimizzare l'ARL fuori controllo, chiamando un allarme se $C_{i,t} > h_i$.
La carta EWMA unilaterale, ipotizzando d'interesse sempre aumenti del parametro in analisi, presenta la statistica di controllo pari a

$$W_{i,t} = \max(\mu_0, \lambda Z_{i,t} + (1 - \lambda) W_{i,t-1}), \tag{11}$$

con $W_{i,0} = \mu_0$, dove μ_0 è il valore in controllo della statistica da cumulare $Z_{i,t}$ che viene posto anche come limite inferiore per far si che la carta reagisca più velocemente.
Focalizzandosi quindi solo su possibili aumenti, la versione estesa del limite di controllo superiore nella carta EWMA si presenta come

$$UCL = \mu_0 + L\sigma_0\sqrt{\left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2t}]}. \tag{12}$$

Nel calcolo del limite, oltre che nella statistica (11), è necessario individuare un valore per λ, costante di lisciamento che pesa adeguatamente le osservazioni passate dove più grande è λ più la statistica reagisce rapidamente al salto, ed L, il valore critico, in maniera tale che l'ARL in controllo sia quello desiderato e l'ARL fuori controllo per individuare il salto scelto sia il più piccolo possibile.
Più in dettaglio si indica con ARL il tempo medio di attesa tra due falsi allarmi in caso di processo in controllo (ARL_0), mentre corrisponde a quanto ci mette lo schema ad accorgersi che il processo è fuori controllo quando parte fuori controllo (ARL_1). D'importanza è anche il valore atteso dell'ARL, la *run length* (RL), cioè il primo momento in cui la carta chiama un allarme. La RL è una variabile casuale geometrica se le osservazioni raccolte nel tempo sono indipendenti, cosa non vera per la carta EWMA. In questo caso l'ARL deve essere calcolata con altri mezzi, per esempio via simulazione.
In (12) compare anche la media μ_0 e la deviazione standard σ_0, quantità della statistica da cumulare solitamente stimata in controllo prima dell'inizio della sorveglianza prospettica. Quando si ha a che fare con campioni di numerosità singola per il calcolo di σ_0 si sfrutta la media dei ranghi mobili, in formula $\hat{\sigma_0} = \frac{\overline{MR}}{1.128}$, dove $MR = |y_t - y_{t-1}|$ è il singolo rango mobile, cioè la differenza dei conteggi in tempi adiacenti. La carta chiama un allarme se $W_{i,t} > UCL_i$.

L'utilizzo di queste carte per il monitoraggio di dati di conteggio Y_t indipendenti con distribuzione binomiale negativa ($Bineg(k, \mu_t)$) è il tema principale di Urbieta et al. (2017), dove gli autori hanno collezionato una serie di proposte presenti in letteratura di $Z_{i,t}$, al fine di ottenere una valutazione delle performance e un confronto tra le due tipologie di carte in termini di ARL_1.

La tabella seguente (Tabella 2.1) riporta schematicamente le statistiche $Z_{i,t}$ e il nome degli autori che le suggeriscono affiancato da una sigla identificativa, per una più facile collocazione nel contesto delle successive considerazioni.

Tabella 2.1: Statistiche monitorate

Metodo	$Z_{i,t}$
Rossi et al. (RS)	$\frac{Y_t - 3\mu_t + 2\sqrt{Y_t \mu_t}}{2\sqrt{\mu_t}}$
Jorgensen (JG)	$\frac{Y_t - \mu_t}{\sqrt{k\pi_t/(1-\pi_t)^2}}$; $\pi_t = \frac{\mu_t}{\mu_t + k}$
McCullagh and Nelder (DR)	$Sign(Y_t - \mu_t)\sqrt{(d_t^2)^*}$
Rogerson e Yamada (RY)	$Y_t - \frac{-k \log((k+\mu_{0,t})/(k+\mu_{1,t}))}{\log(\mu_{1,t}(k+\mu_{0,t})/\mu_{0,t}(k+\mu_{1,t}))}$
Höhle e Paul (LR)	$\log\left(\frac{f_{\mu_{0,t}}(y_t)}{f_{\mu_{1,t}}(y_t)}\right)$

$$\text{con } d_t^{2*} = \begin{cases} 2k\log(1+\mu_t/k) & \text{if } Y_t = 0 \\ 2Y_t \log\left(\frac{Y_t}{\mu_t}\right) - 2k(1+Y_t/k)\log\left(\frac{1+Y_t/k}{1+\mu_t/k}\right) & \text{if } Y_t > 0 \end{cases}$$

In generale, con l'obiettivo di monitorare dati di conteggio che seguono una distribuzione adeguata a seconda della presenza di sovradisperione, è necessario stimare un modello di regressione per monitorare le osservazioni la cui media e varianza dipendono da covariate. Il valore atteso di Y_t date le covariate fino al tempo t è indicato con $\mu_{0,t}$ quando il processo è in controllo e $\mu_{1,t}$ quando il processo è fuori controllo, cioè significa che la media del processo può aver subito un aumento dopo un certo istante temporale. La media di Y_t è chiamata μ_t quando non è necessario esplicitare se il processo è in controllo o meno.
Le principali carte sono state disegnate per monitorare dati gaussiani, perciò individuare statistiche di controllo che propongono una trasformazione delle variabili per avvicinarsi alla normalità, come (RS) o (JG), è molto comune per analizzare dati non gaussiani. Un'altra proposta è l'utilizzo dei residui di devianza ottenuti dal modello GLM (DR), sempre riconducibile alla medesima distribuzione se il modello si adatta bene ai dati.
Le statistiche (RY) ed (LR) sono le uniche a non ricondursi ad una distribuzione normale, ma sfruttano il log-rapporto di verosimiglianza binomiale negativo. In particolare per (RY), riportata con la notazione presente in Hawkins and Olwell

(1998), in (10) il K corrisponde al secondo addendo delle relative $Z_{i,t}$ e quest'ultime possono essere riportate premoltiplicate per un fattore $c_t = h_i/h_{i,t}$, cioè il rapporto tra limiti costanti e variabili, anche se spesso si usa la semplificazione $c_t = 1$. Con (LR) il K in (10) è 0 poiché discende direttamente dal log-rapporto di verosimiglianza. Per una descrizione più dettagliata delle statistiche $Z_{i,t}$ si veda anche Alencar et al. (2017).
Una valutazione delle varie performance delle statistiche appena citate verrà presentata con il caso pratico di sezione 2.4.1 .

Gli autori García-Bustos and Zambrano (2022) hanno ampliato le proposte viste in Tabella 2.1 con l'utilizzo dei residui di un modello GLM binomiale negativo, una scelta comunque non nuova in letteratura, ne sono da esempio i lavori di Park et al. (2018) e Park et al. (2020). I residui di Pearson di un modello GLM binomiale negativo sono definiti come

$$e_t = \frac{y_t - \hat{\mu}_t}{\left(\hat{\mu}_t + \frac{\hat{\mu}_t^2}{k}\right)^{1/2}}, \tag{13}$$

con y_t il conteggio degli eventi, $\hat{\mu}_t$ la media stimata e k il parametro di dispersione. Gli autori hanno inoltre analizzato la versione studentizzata della (13) e dei residui di devianza (DR) della Tabella 2.1, quantità ottenibili dividendo per $\sqrt{1-h_{tt}}$, con h_{tt} il *coefficiente leva* del dato t-esimo.
Il residuo t-esimo è tanto meno variabile quanto più grande è h_{tt}, di conseguenza un elevato h_{tt} è indice di un punto leva, cioè un punto che presenta valori inusuali nei regressori rispetto agli altri dati e che forza il modello a passargli vicino. Più in dettaglio il t-esimo elemento della diagonale della matrice cappello H misura quanto contribuisce y_t nel determinare $\hat{y}_t$, come si può ben vedere dalla seguente formula

$$\hat{y}_t = \sum_{j=1}^{n} h_{tj} y_j = h_{tt} y_t + \sum_{j \neq t} h_{tj} y_j, \quad \text{t=1,2,...,n.} \tag{14}$$

La matrice cappello H è ottenibile quando si ha a che fare con stimatori lineari rappresentabili come $\hat{y} = Hy$, con H dipendente solamente dalle variabili esplicative. Il vettore residuo e si ottiene dunque come $y - \hat{y} = y - Hy$, da cui si ricava $V(e) = (I - H)\sigma^2$, con σ^2 opportunamente stimato e con $0 \leq h_{tt} \leq 1$ per la non

negatività della varianza.
Avendo a che fare con modelli lineari generalizzati (GLR), come in questo caso, si parla di matrice H generalizzata, esprimibile come

$$H_w = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}, \tag{15}$$

con X matrice delle esplicative e W matrice di varianze e covarianze, quantità utilizzata perché è ragionevole assumere che le osservazioni sulla risposta siano eteroschedastiche e incorrelate. Con il modello binomiale negativo in questione la matrice W si presenta come

$$W = diag(w_t), \text{ con } w_t = \frac{1}{(g'(\mu_t))^2 Var(Y_t)}, \quad \text{t=1,...,n} \tag{16}$$

e se il legame è canonico si ha $w_t = \frac{v(\mu_t)}{a_t(\phi)}$, nel caso specifico $w_t = v(\mu_t) = \mu_t(1+\mu_t/k)$ dato che $a_t(\phi) = 1$.

Utilizzando quindi come $Z_{i,t}$ i residui (13), i residui di devianza (DR) o la versione studentizzata si è studiata la carta EWMA con la statistica di controllo come in (11), con valore iniziale di $W_{i,0} = 0$, ipotizzando che la media dei residui sia 0 quando i dati non deviano dal modello stimato.
Il limite di controllo coincide sempre con (12) quando si vuole individuare velocemente solamente aumenti del parametro d'interesse.

2.2 CUSUM per cambiamenti della media: un approfondimento

Ipotizzando una successione di variabili casuali indipendenti $y_1, y_2...$ tali che

$$y_t \sim \begin{cases} f_0(.) & \text{se } t < \tau \\ f_1(.) & \text{se } t \geq \tau \end{cases}$$

dove τ, il momento del fuori controllo, è ignoto e le densità f sono completamente note. La carta di controllo unilaterale segnala un allarme quando la statistica

$C_t > h$ dove

$$C_t \sim \begin{cases} 0 & \text{se } t = 0 \\ \max\left(0, C_{t-1} + \log \frac{f_1(y_t)}{f_0(y_t)}\right) & \text{se } t > 0 \end{cases}$$

con h costante positiva, eventualmente variabile nel tempo, in modo da minimizzare il ritardo atteso tra tutte le carte con ARL in controllo uguale o maggiore a quella dello schema descritto. Cumulare il log-rapporto di verosimiglianza discende quindi direttamente dalla definizione di CUSUM e questo può essere adattato alla distribuzione desiderata, come la binomiale negativa qui studiata.
Per esplicitare il calcolo in questo contesto ci sono due modi, o utilizzando direttamente la distribuzione che descrive il numero di fallimenti precedenti il successo k-esimo in un processo di Bernoulli di parametro p, formula (2), o partendo dalla definizione del modello mistura Gamma-Poisson con k parametro di forma della Gamma, coincidente al parametro di dispersione della binomiale negativa, formula (6). Seguendo la prima versione, per compattezza delle formule, il log-rapporto di verosimiglianza si ottiene come rapporto tra la distribuzione fuori controllo e quella in controllo ottenendo

$$\begin{aligned} &\log\left(\frac{f_{p_{1,t}}(y_t)}{f_{p_{0,t}}(y_t)}\right) \\ =&\log \frac{\left(\frac{y_t+k-1}{y_t}\right) p_{1,t}^k (1-p_{1,t})^{y_t}}{\left(\frac{y_t+k-1}{y_t}\right) p_{0,t}^k (1-p_{0,t})^{y_t}} \\ =&k \log\left(\frac{p_{1,t}}{p_{0,t}}\right) + y_t \log\left(\frac{1-p_{1,t}}{1-p_{0,t}}\right). \end{aligned} \tag{17}$$

Indicando con $Z_{i,t} = y_t \log\left(\frac{1-p_{1,t}}{1-p_{0,t}}\right)$ e con $K = k \log\left(\frac{p_{1,t}}{p_{0,t}}\right)$ si ottiene la statistica CUSUM definita in (10). Da questa versione della distribuzione di probabilità è facile ricondursi all'utilizzo della mistura Gamma-Poisson per i calcoli, grazie ai risultati espressi nel paragrafo 1.4 . Höhle and Paul (2008) si sono occupati proprio di studiare questa versione di sorveglianza, come già introdotto in Tabella 2.1, proponendo il log-rapporto di verosimiglianza in una formulazione 'conveniente' pari a

$$\begin{aligned}&\log\left(\frac{f_{p_{1,t}}(y_t)}{f_{p_{0,t}}(y_t)}\right)\\ =&y_t k + \left(y_t + \frac{1}{\alpha}\right)\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right),\end{aligned} \tag{18}$$

con α il reciproco del parametro di dispersione k e ipotizzando per $\mu_{1,t}$ un cambiamento additivo in scala logaritmica di $\mu_{0,t}$, cioè $\log(\mu_{1,t}) = \log(\mu_{0,t}) + \kappa$. Da questa riscrittura risulta quindi facile isolare le componenti variabili con y_t, le $Z_{i,t}$, da quello che rimane, K, e ricondursi alla formulazione della statistica di controllo presente in (10).
Si veda Höhle and Paul (2008) per un maggior approfondimento.

Seguendo queste tracce teoriche, è naturale pensare allo sviluppo di una carta di controllo CUSUM unilaterale per sorvegliare il parametro di dispersione k e segnalare il caso critico di un aumento di variabilità nei dati dato da una sua diminuzione, considerando la media sempre in controllo. Ad ogni istante di tempo l'ipotesi nulla $H_0 : k_t = k$, corrispondente alla situazione in controllo, viene testata contro $H_1 : k_t = k_a$, la situazione di possibile fuori controllo e una volta stabilito l'ARL_0 e lo shift di k che si vuole individuare velocemente, il log-rapporto di verosimiglianza che si ottiene è pari a

$$\frac{f_{\mu,k_\alpha}(y_t)}{f_{\mu,k}(y_t)} = \frac{\frac{\Gamma(y_t+k_\alpha)}{\Gamma(y_t+1)\Gamma(k_a)}\left(\frac{\mu}{k_a+\mu}\right)^{y_t}\left(\frac{k_a}{k_a+\mu}\right)^{k_a}}{\frac{\Gamma(y_t+k)}{\Gamma(y_t+1)\Gamma(k)}\left(\frac{\mu}{k+\mu}\right)^{y_t}\left(\frac{k}{k+\mu}\right)^{k}}. \tag{19}$$

Per ottenere la statistica CUSUM, applicando il logaritmo a (19) si ottiene

$$C_t = \log\left(\frac{\Gamma(y_t+k_a)\Gamma(k)}{\Gamma(y_t+k)\Gamma(k_a)}\right) + y_t\log\left(\frac{k+\mu}{k_a+\mu}\right) - k\left(\frac{k}{k+\mu}\right) + k_a\left(\frac{k_a}{k_a+\mu}\right) \tag{20}$$

e riscrivendo la precedente equazione nella forma in (10) si ottiene

$$C_t = \log\left(\frac{\Gamma(y_t+k_a)\Gamma(k)}{\Gamma(y_t+k)\Gamma(k_a)}\right) + y_t\log\left(\frac{k+\mu}{k_a+\mu}\right) - k\left(\frac{k}{k+\mu}\right) + k_a\left(\frac{k_a}{k_a+\mu}\right) + C_{t-1}, \tag{21}$$

dove $C_t = 0$ se $t = 0$ e anche se $C_t < 0$, essendo una carta unilaterale. Ad ogni nuova osservazione la carta compara la statistica C_t con il limite di controllo h, chiamando un allarme e giudicando il processo fuori controllo in caso $C_t > h$.

2.3 LR-EWMA per cambiamenti della media: un approfondimento per dati correlati

Nel recente studio di Albarracin et al. (2018) vengono sottolineate l'utilità delle carte EWMA basate sull'utilizzo di un modello lineare generalizzato (GLM) per stimare la media variabile nel tempo di un processo in ambito di sorveglianza sanitaria, grazie all'efficienza nell'individuare velocemente piccoli cambiamenti nei dati di conteggio come i tassi di mortalità. Molto spesso la presenza di correlazione seriale nei dati di conteggio non viene trattata implicando un peggioramento nella performance delle carte, ma può essere affrontata attraverso modelli adeguati, come i modelli autoregressivi a media mobile generalizzati (GARMA). Per un approfondimento sull'utilizzo si veda l'ultimo articolo citato, nel seguito queste carte non verranno prese in considerazione.
Richiamando il log-rapporto di verosimiglianza (LR) per testare cambiamenti nella media di dati di conteggio distribuiti come una binomiale negativa con parametro k fissato, esso si esprime come

$$\begin{aligned} LR_t =& \log\left(\frac{f_{\mu_{1,t}}(y_t)}{f_{\mu_{0,t}}(y_t)}\right) \\ =& y_t \log\left(\frac{\mu_{1,t}}{\mu_{0,t}}\right) + (y_t + k)\log\left(\frac{\mu_{0,t}+k}{\mu_{1,t}+k}\right), \end{aligned} \tag{22}$$

dove $\mu_{1,t} = \delta\mu_{0,t}$.
Si può notare che valori di LR_t in (22) possono essere anche negativi (l'ultimo logaritmo assume valori < 1), soprattutto quando calcolati con grandi valori di y_t. In aggiunta, valori di LR_t non sono indipendenti per $t = 1, ..., n$ (poiché sono funzioni di y_t).
Questi due aspetti comportano che l'implementazione e l'interpretazione di questa statistica nella carta EWMA è spesso infattibile quando una serie storica con alti valori viene monitorata; dopo aver individuato il limite superiore, la carta non segnala un cambiamento nella media se il processo va fuori controllo.

Per cercare di far fronte a questo problema una statistica alternativa è la statistica LRc_t, ottenuta dopo aver centrato y_t, cioè

$$LRc_t = Z_{1,t} = (y_t - \mu_{0,t}) \log\left(\frac{\mu_{1,t}}{\mu_{0,t}}\right) + (y_t - \mu_{0,t} + k) \log\left(\frac{\mu_{0,t} + k}{\mu_{1,t} + k}\right). \tag{23}$$

Una seconda statistica basata sul test LR consiste nel trovare la stima di massima verosimiglianza dello shift k che rappresenta il tasso di cambio di $\mu_{0,t}$ in relazione a $\mu_{1,t}$ per ogni t ($\mu_{1,t} = \mu_{0,t} \exp(k_t)$) basandosi sulle osservazioni $y_1, y_2, ..., y_n$. In questo caso si suppone una crescita additiva della media in scala logaritmica perché computazionalmente vantaggioso rispetto una crescita direttamente sulla media (Höhle and Paul (2008)). La massimizzazione della log-verosimiglianza $l_n = \sum_{t=1}^{n} \log f(y_t)$ come funzione di k in generale può essere svolta attraverso pochi passi di algoritmi iterativi, come visto nel paragrafo 1.4. In questo caso per individuare k_t si può risolvere la derivata $\frac{dl_n}{dk} = 0$ per ogni t, ottenendo

$$Z_{2,t} = \exp(\hat{k_t}). \tag{24}$$

Nonostante queste trasformazioni, si può dimostrare che le due statistiche Z_t appena individuate risultano autocorrelate.
Attraverso studi di simulazione è stato quindi possibile valutare l'impatto nell'utilizzo di dati di conteggio autocorrelati binomiali negativi, in termini di ARL_0 e ARL_1, quando utilizzati per una carta EWMA per dati indipendenti.
I risultati raggiunti possono essere riassunti come segue: quando si prevede assenza di correlazione nei dati l'ARL in controllo è quella desiderata e l'ARL fuori controllo raggiunge i valori più bassi mentre all'aumentare della correlazione aumentano i falsi allarmi con un aumento dei valori dell'ARL fuori controllo.
In Albarracin et al. (2018) è possibile approfondire i confronti delle performance per diversi scenari anche con altre statistiche asintoticamente normali e analizzare più nel dettaglio i risultati ottenuti.

2.4 Casi pratici di alcune proposte

2.4.1 Valutazione delle statistiche analizzate alla sorveglianza di accessi in ospedale per problemi respiratori

Per stimare il numero di accessi di persone sopra i 65 anni per problemi respiratori all'ospedale di San Paolo, Brasile, è risultato adeguato un modello GLM binomiale negativo.
Il modello ottenuto per il periodo in controllo Gennaio 2006-Dicembre 2010 con k stimato pari a 69.99 è il seguente:

$$\log\left(\frac{\mu_{0,t}}{pop_t}100000\right) = \beta_0+\beta_1+\cos\left(\frac{2\pi t}{365}\right)+\beta_2\sin\left(\frac{2\pi t}{365}\right)+\beta_3 Sat_t+\beta_4 Sun_t+\beta_5 Mon_t. \tag{25}$$

All'interno della funzione di legame logaritmica è presente un *offset* pari a g_t=(pop_t/100000) per il calcolo del tasso medio giornaliero di ospedalizzazione, mentre nella specificazione del modello si sono tenuti in considerazione comportamenti stagionali, dati dalla presenza del seno e del coseno, e alcuni giorni della settimana critici, poiché è noto che nel weekend il numero di ingressi cali per poi ricrescere dal lunedì successivo. Dopo aver verificato l'adeguatezza del modello, i valori stimati per il periodo in controllo e per il 2011 rispecchiano l'andamento desiderato.
Attraverso studi di simulazione sono stati calcolati i limiti di controllo, con un ARL_0 pari a 500, e sono stati simulati campioni fuori controllo per valutare la performance delle carte per diversi shift in termini di ARL_1, in particolare per un aumento della media del 25%, 50%, 75% e 100% .
Utilizzando le statistiche di Tabella 2.1 si evince che la carta CUSUM che segnala un vero allarme in modo più tempestivo per tutti gli shift è quella basata sul log-rapporto di verosimiglianza, anche se quest'ultima potrebbe risultare meno intuitiva rispetto le carte che si basano sulla trasformazione normale.
Un'altra comparazione possibile coinvolge le due tipologie di carte, CUSUM ed EWMA. In questo caso la carta EWMA con il parametro di lisciamento pari a 0.2 offre la miglior performance per tutti gli shift e per tutte le statistiche rispetto la CUSUM e la differenza tra le due diventa sempre più piccola con il crescere dello

shift.
La carta EWMA è quindi una buona scelta sia in termini di performance che di facilità di comprensione e calcolo, mentre per la scelta della statistica da utilizzare esse offrono pressoché la stessa prestazione in termini di ARL_1. Le diverse comparazioni e l'analisi più dettagliata di tutti i valori ottenuti si possono trovare in Urbieta et al. (2017).

2.4.2 Applicazione del modello lineare generalizzato alla sorveglianza di casi d'infezione da dengue in Ecuador

La Pan America Health Organization (PAHO) è un'organizzazione internazionale specializzata nella tutela, protezione e miglioramento della salute degli abitanti delle Americhe. Tra le varie iniziative, la PAHO si occupa di raccogliere e condividere dati epidemiologici di malattie infettive trasmesse da zanzare, tra cui i casi di dengue, una delle malattie infettive più diffuse nell'America Latina di origine virale causata da quattro virus molto simili (Den-1, Den-2, Den-3 e Den-4).
Il virus circola nel sangue della persona infetta per 2-7 giorni e in questo periodo la zanzara può prelevarlo e trasmetterlo ad altri causando principalmente febbre molto alta, ma anche vomito, nausea e dolori muscolari che, nelle forme più gravi possono condurre alla morte.
La dengue è conosciuta da oltre due secoli ed è particolarmente presente durante e dopo la stagione delle piogge nelle zone tropicali. Per questo motivo il fattore climatico è fondamentale da tenere in considerazione nella costruzione del modello previsivo del numero di casi di infezioni. Gli autori hanno stimato un modello in controllo binomiale negativo con funzione legame logaritmica in cui l'identificativo della settima analizzata gioca un ruolo determinante nella stima dei casi in controllo, ottenendo:

$$\log(mortiattese) = 4.57 - 0.045 * Settimana + 0.00034 * Settimana^2, \qquad (26)$$

con parametro di dispersione k pari a 19.71. Per la costruzione della carta EWMA per la sorveglianza prospettica è necessario ricondursi alla media stimata attraverso una trasformazione esponenziale della precedente regressione binomiale negativa

ed ottenere i residui con la formula (13), i residui di devianza (DR) o la versione studentizzata. Le quantità appena citate vengono poi utilizzate per costruire le statistiche in (11) con il limite di controllo pari a (12), dove $\mu_0 = 0$ e σ_0 calcolata in controllo per mezzo dei ranghi mobili (si veda sezione 2.1). A questo punto è possibile disegnare diverse carte a parità di ARL_0 in questo caso posto a 370, per diverse combinazioni di L, il valore critico, e λ.
Confrontando la performance per diversi shift in deviazione standard della media dei residui, in García-Bustos and Zambrano (2022) emerge che le prestazioni migliori si hanno considerando piccoli valori di λ, in particolare con 0.05 la carta chiama un allarme più tempestivamente sia per salti della media grandi che piccoli. La medesima performance è offerta indipendentemente dalla statistica usata, cioè segnalando con un vero allarme l'undicesima settima del 2019 per la presenza di El Niño, un evento atmosferico che contribuisce alla proliferazione delle zanzare. Inoltre se l'obiettivo è individuare la statistica più sensibile in termini di ARL_1, usare i residui di devianza studentizzati è la scelta migliore, mentre se si predilige semplicità e interpretazione, i residui di Pearson sono più adatti.

3 Soluzioni teoriche alternative

Nel Capitolo 2 sono state introdotte diverse proposte al fine di modellare dati di conteggio che presentano sovradispersione. Ad esclusione di un piccolo accenno sulla possibile sorveglianza specifica di k, le carte fin qui analizzate si sono concentrate esclusivamente nell'individuare tempestivamente cambiamenti (aumenti) della media del processo. Di seguito verranno quindi analizzati alcuni punti lasciati più in ombra nelle precedenti trattazioni, come per esempio la possibilità di andare a sorvegliare processi in cui i cambiamenti possono coinvolgere sia il parametro di dispersione k che la media μ. Per far questo verrà illustrato una tipologia di carta di controllo multivariata per la sorveglianza congiunta e la diretta alternativa delle carte combinate. Spazio verrà dato anche ai sempre più usati limiti dinamici, limiti cioè variabili con il tempo.
In ambito sanitario, avere una carta che chiama troppo spesso allarmi è molto comune, come una nota ricercatrice (Galit Shmueli) disse "*...most health monitors...learned to ignore alarms triggered by their system. This is due to the excessive false alarm rate that is typical of most systems - there is nearly an alarm every day!*".
Questo succede molto spesso perché i dati sorvegliati non sono adeguatamente trattati e di conseguenza le carte che si costruiscono sono controproducenti, generando molti falsi allarmi. Riconquistare la fiducia degli utilizzatori è quindi indispensabile per dar credito all'utilità degli strumenti del SPC, ma questo può esser fatto solo con le dovute accortezze ed i limiti dinamici ne sono un esempio. Applicazioni pratiche delle proposte saranno presentate nel Capitolo 4.

3.1 Limiti dinamici

Nei precedenti articoli la costruzione dei limiti di controllo è sempre stata omessa nei particolari o appena accennata con l'utilizzo di tecniche di simulazione.
Un metodo che viene utilizzato largamente negli ultimi anni e applicabile universalmente, indipendentemente dalla statistica di controllo e dalla distribuzione dei dati sottostante, è quello di utilizzare limiti di controllo variabili, chiamati limiti dinamici. La caratteristica principale di quest'ultimi è il fatto di non assumere un valore costante nel tempo, ma mutabile, in modo da adattarsi alla distribuzione

delle statistiche di controllo che a sua volta varia molto spesso con il tempo.
Alla base della sorveglianza prospettica si trova l'assunzione di conoscere il parametro in controllo per la distribuzione in esame, anche se molto spesso questo deriva da una stima effettuata in fase I. In alcuni casi i parametri sono ignoti e non possono essere precalcolati perché dipendono da fattori noti solo quando si manifestano in un certo momento t, come gli esempi analizzati in Shen et al. (2016) e Aytaçoğlu and Woodall (2020). In Shen et al. (2016) la proposta riguarda l'uso di limiti dinamici per una carta EWMA al fine di monitorare tassi di Poisson in cui il parametro in controllo dipende dalla popolazione che varia nel tempo e quindi non è possibile il calcolo del limite per un determinato periodo fin quando questa non diventa nota. In Aytaçoğlu and Woodall (2020) viene sorvegliata una distribuzione binomiale in cui la popolazione varia sempre nel tempo e in questo contesto viene implementata una carta CUSUM.
In altri casi, come nell'esempio pratico di sorveglianza dei casi di dengue che verrà sviluppato nel Capitolo 4, la distribuzione in controllo è nota a priori anche se variabile nel tempo, ma la scelta di questi limiti risulta facile rispetto ad altre alternative. In sintesi, una giustificazione per l'uso dei limiti dinamici è il fatto che forniscono una soluzione conveniente e vantaggiosa per diversi scenari di sorveglianza, assicurando protezione ai livelli desiderati. A tale scopo, una possibilità di garantire l'ARL in controllo desiderata pari a B consiste nel fissare L_t, il limite dinamico, in maniera tale che la probabilità condizionata risulti

$$Pr(RL > t|RL \geq t) = 1 - \frac{1}{B} \quad \forall t > 0. \tag{27}$$

Questo garantisce che la distribuzione della run length (RL) sia geometrica con media B in caso di sorveglianza unilaterale, infatti

$$\begin{aligned} Pr(RL = t) =& Pr(RL > 1) \times Pr(RL > 2|RL \geq 2) \times ... \times \\ & ... \times Pr(RL > t-1|RL \geq t-1) \times [1 - Pr(RL > t|RL \geq t)] = \\ =& \left(1 - \frac{1}{B}\right)^{t-1} \frac{1}{B}. \end{aligned} \tag{28}$$

Quindi L_t può essere calcolato come il quantile $1 - \frac{1}{B}$ della distribuzione della

statistica in controllo, chiamata per esempio W_t, condizionata a $W_1 \leq L_1$, $W_2 \leq L_2$,..., $W_{t-1} \leq L_{t-1}$, dato che

$$Pr(RL > t|RL \geq t) = Pr(W_t \leq L_t|W_1 \leq L_1, ..., W_{t-1} \leq L_{t-1}). \tag{29}$$

La distribuzione di $W_t|W_1, ..., W_{t-1}$ è difficile da determinare analiticamente, ma possiamo approssimare i limiti di controllo via simulazione andando a generare in parallelo alla vera statistica di controllo W_t, che calcoliamo dai dati, anche un gran numero di statistiche W_t^* generate dalla distribuzione in controllo di W_t.
I passi da seguire descritti in maniera generale sono i seguenti, varierà la statistica di controllo a seconda della carta utilizzata e la distribuzione di probabilità sottostante i dati, ma il principio è lo stesso:

- Al tempo $t = 1$ si ottiene la prima statistica di controllo. Per il calcolo del limite si generano in controllo M statistiche con la distribuzione sottostante i dati;
- Si ordinano le M statistiche in senso crescente e si prende il quantile $1 - \frac{1}{B}$ che rappresenta il limite L_1;
- Le statistiche simulate eccedenti il limite L_1 vengono sostituite con un campionamento bootstrap da quelle non eccedenti il limite per 'continuare' solo le traiettorie in controllo e quindi soddisfare il condizionamento di (29);
- Se la statistica ottenuta supera il limite L_1 la carta chiama un allarme, altrimenti si prosegue al t=2;
- Al tempo $t = 2$ si ottiene la statistica di controllo. Per il calcolo del limite si generano M statistiche che dipendono dalle statistiche al tempo precedente supposte in controllo grazie alla sostituzione, si ordinano le M statistiche in senso crescente e si prende il quantile $1 - \frac{1}{B}$ come limite L_2;
- Le statistiche simulate eccedenti il limite L_2 vengono sostituite con un campionamento bootstrap da quelle non eccedenti il limite per 'continuare' solo le traiettorie in controllo;
- Se la statistica ottenuta supera il limite L_2 la carta chiama un allarme, altrimenti si prosegue al t=3 e così via.

L'uso dei limiti dinamici permette quindi di superare il problema di avere una performance in controllo inattesa della carta desiderata.

3.2 Carte di controllo combinate

Quando si costruisce una carta di controllo l'obiettivo è molto chiaro, sorvegliare una determinata caratteristica di qualità e chiamare tempestivamente l'allarme se si percepisce un fuori controllo. In molti contesti concentrare la sorveglianza su un singolo aspetto può risultare riduttivo quando si ha a che fare con diverse caratteristiche di qualità e si vorrebbe monitorarle tutte. Una soluzione è quindi utilizzare una procedura combinata, costruendo e monitorando tante carte di controllo quanti sono i parametri d'interesse.
Scenari in cui risulta utile l'utilizzo simultaneo di diverse carte possono essere diversi, come ad esempio:

- combinazione di stesse carte (ad esempio due o più CUSUM) ottimali per diversi shift del parametro da sorvegliare;
- combinazione di carte diverse (ad esempio Shewhart-CUSUM) per individuare velocemente cambiamenti piccoli e grandi del parametro in esame;
- combinazione di carte per individuare velocemente cambiamenti di parametri diversi;

Considerando H carte unilaterali, ognuna è caratterizzata da un proprio limite $L_1, L_2, ..., L_H$ che deve essere individuato in modo tale da rispettare le seguenti condizioni

$$\begin{cases} E_{IC}(\min(RL_1, ..., RL_H)) = B \\ E_{IC}(RL_1) = E_{IC}(RL_2) = ... = E_{IC}(RL_H) \end{cases}$$

dove B è un valore desiderato dell'ARL in controllo.
La prima condizione assicura che lo schema abbia un ARL in controllo complessivo pari a B. Infatti lo schema combinato segnala un allarme al tempo $\min(RL_1, ..., RL_H)$, mentre la seconda condizione assicura una sorta di "bilanciamento" tra gli H schemi: nessuno degli H schemi quando usato da solo tende a

chiamare un falso allarme prima degli altri.

In termini di probabilità la precedente condizione può essere riscritta come

$$\begin{cases} Pr(W_t^1 \geq L_t^1, ..., W_t^H \geq L_t^H) = \frac{1}{B} \\ Pr(W_t^1 \geq L_t^1) = ... = Pr(W_t^H \geq L_t^H) = 1 - \beta \end{cases}$$

con W_t le statistiche di controllo al tempo t dei diversi schemi e L_t i rispettivi limiti che corrispondono allo stesso quantile della distribuzione delle W in controllo, con β la probabilità ignota.
Se si ipotizza l'indipendenza tra gli schemi, la probabilità congiunta della prima condizione può essere riscritta come un prodotto ed è possibile facilmente ricondursi al calcolo di $\beta = 1 - \sqrt[H]{(\frac{1}{B})}$ per risalire al quantile desiderato.
In caso di dipendenza tra gli schemi di sorveglianza il processo per individuare i limiti diventa più complesso e si ricorre spesso alla simulazione, generando in parallelo alla vera statistica di controllo W_t calcolata dai dati anche un gran numero N di statistiche W_t^*, generate dalla distribuzione in controllo di W_t. In questo caso la simulazione delle W_t viene fatta per tutti gli schemi, ottenendo $W_{t(1)}^{1*}, W_{t(2)}^{1*}, ..., W_{t(N)}^{1*}$ per il primo schema, $W_{t(1)}^{2*}, W_{t(2)}^{2*}, ..., W_{t(N)}^{2*}$ per il secondo schema e così via. Una volta ordinati in senso crescente i valori simulati separatamente per ogni carta, bisogna utilizzare una procedura "a tentativi" per i possibili valori assunti da β, in modo tale che il medesimo quantile soddisfi la condizione imposta nelle distribuzioni marginali e contemporaneamente anche nella distribuzione congiunta.
L'utilizzo di carte separate può risultare più oneroso dal punto di vista computazionale quando si hanno tanti parametri da sorvegliare, ma è utile quando si pensa che solo uno o pochi dei parametri possano cambiare ed è d'interesse risalire alla fonte del fuori controllo.
Conviene invece utilizzare un'unica carta multivariata in cui più parametri convergono nel calcolo di una statistica se si pensa possano cambiare tutti, soluzione approfondita nel paragrafo seguente.

3.3 Carta EWMA multivariata basata sullo *score* di verosimiglianza (MEWMA-*score*)

Quando si è interessati alla sorveglianza di più caratteristiche di qualità, l'idea più naturale è quella di utilizzare diversi schemi congiuntamente, come appena trattato, dove ogni carta monitora una singola caratteristica e l'allarme viene dato non appena una di esse segnala un fuori controllo. Molte persone utilizzano questa scelta per la facilità di calcolo ed interpretazione, ma non è sempre la strategia migliore. E' noto dalla letteratura del SPC che testare la presenza di anomalie nel processo in un contesto multivariato risulta spesso più efficiente che monitorare le singole quantità d'interesse, soprattutto grazie alla possibilità di cogliere associazioni tra le componenti di qualità che altrimenti sarebbero ignorate. In questo contesto si sono sviluppate le carte di controllo multivariate, strumenti che cercano di riassumere in un'unica statistica di controllo tutti i parametri d'interesse. Estensioni al caso multivariato delle carte di controllo già incontrate sono per esempio la carta MEWMA (Lowry et al. (1992)) e MCUSUM (Woodall and Ncube (1985)).
Per costruire la carta MEWMA a dati con osservazioni singole, d'interesse in questa trattazione, si è scelto l'utilizzo della statistica di tipo *score*, quantità largamente utilizzata in ambito inferenziale per risolvere problemi di verifica d'ipotesi ed asintoticamente equivalente alle statistiche di tipo Wald e del rapporto di verosimiglianza. Riconducendosi alla distribuzione d'interesse di questa tesi, la binomiale negativa, è quindi necessario individuare lo *score*, cioè la derivata prima della log-verosimiglianza, e l'informazione di Fisher, ricordando $l'(\theta) \sim N_p(0, i(\theta))$, con $p = 2$ in questo caso.
Più nel dettaglio è possibile partire dalla densità in (6) per ottenere la funzione di log-verosimiglianza considerando singole osservazioni (il caso generale con più osservazioni è direttamente ricavabile) e ponendo $\theta = (\mu, k)$ si ottiene

$$l(\theta; y) = y \log \mu + k \log k - (k + y) \log(k + \mu) + G(k), \tag{30}$$

dove $G(k) = \log(\Gamma(y + k)/(\Gamma(y + 1)\Gamma(k)))$.
La funzione punteggio è data da $l'(\theta) = \left(\frac{d}{dk} l(\theta), \frac{d}{d\mu} l(\theta)\right)^T$ con componenti

$$
\begin{aligned}
l_k(\theta) &= \frac{d}{dk} l(\theta) = \log k - \frac{k+y}{k+\mu} - \log(\mu + k) + 1 + D(k), \\
l_\mu(\theta) &= \frac{d}{d\mu} l(\theta) = \frac{y}{\mu} - \frac{k+y}{k+\mu},
\end{aligned}
\tag{31}
$$

dove $D(k) = dG(k)/dk = \varphi(k+y) - \varphi(k)$, con $\varphi(x) = d \log \Gamma(x)/dx$ funzione digamma. Degno di nota è il fatto che si adatta un modello di regressione in cui μ dipende dai coefficienti β dei regressori, per esempio ipotizzando un legame canonico si ha l'uguaglianza $\mu = \exp(\beta^T x)$. Estendendo quindi il ragionamento in termini di β, un ulteriore sviluppo sarebbe calcolare la derivata per ogni β_r, cioè $\frac{dl(\theta)}{d\mu}\frac{d\mu}{d\beta_r}$ che in questo caso equivale a moltiplicare lo *score* già ottenuto per μx_r, con $r = 1, 2, ..., p.$. Alla luce di ciò, il livello di dettaglio di questa tesi si limita ad esplicitare la derivazione fermandosi a μ.

La matrice di informazione osservata e attesa (o di Fisher) corrispondono rispettivamente a

$$
j(\theta) = \begin{pmatrix} j_{kk} & j_{k\mu} \\ j_{\mu k} & j_{\mu\mu} \end{pmatrix} \quad \text{e} \quad i(\theta) = \begin{pmatrix} i_{kk} & i_{k\mu} \\ i_{\mu k} & i_{\mu\mu} \end{pmatrix},
$$

dove

$$
\begin{aligned}
j_{kk} &= -\frac{d}{dk} l_k(\theta) = -\frac{y-\mu}{(k+\mu)^2} + \frac{1}{\mu + k} - \frac{1}{k} - \varphi'(k+y) + \varphi'(k), \\
j_{k\mu} = j_{\mu k} &= -\frac{d}{dk} l_\mu(\theta) = -\frac{d}{d\mu} l_k(\theta) = \frac{y-\mu}{(k+\mu)^2}, \\
j_{\mu\mu} &= -\frac{d}{d\mu} l_\mu(\theta) = \frac{y}{\mu^2} - \frac{k+y}{k+\mu},
\end{aligned}
\tag{32}
$$

con $\varphi'(x) = \frac{d\varphi(x)}{dx}$ la generica funzione trigamma ed infine

$$
\begin{aligned}
i_{kk} &= E[j_{kk}], \\
i_{\mu\mu} &= E[j_{\mu\mu}] = \frac{1}{\mu} - \frac{1}{k+\mu}, \\
i_{k\mu} &= i_{\mu k} = E[j_{\mu k}] = 0.
\end{aligned}
\tag{33}
$$

L'elemento i_{kk} è lasciato al calcolo per mancanza di forma esplicita diretta, mentre $i_{\mu\mu}$ può essere calcolato analiticamente. L'uguaglianza $i_{k\mu} = 0$ ha il significato di ortogonalità dei parametri μ e k garantendo la semplificazione nel calcolo dell'inversa.
Per ottenere la statistica finale, lo *score* per prima cosa è pesato secondo la classica struttura dell'EWMA

$$W_t = \lambda l'(y_t, \mu_{0,t}, k_0) + (1 - \lambda)W_{t-1}, \tag{34}$$

con W_0 vettore di zeri, $\mu_{0,t}$ la media assunta variabile nel tempo e k_0 il parametro di dispersione entrambi stimati in controllo.
Ad ogni nuova osservazione collezionata è anche necessario aggiornare la varianza della statistica Wt, dove entra in gioco l'informazione di Fisher con elementi in (33), poiché dipendente dalla media variabile nel tempo, ottenendo

$$S2_t = \lambda^2 i(\mu_{0,t}, k_0) + (1 - \lambda)^2 S2_{t-1}, \tag{35}$$

con $S2_t$ matrice di zeri al tempo $t = 0$.
Infine la statistica univariata riportata dalla carta che tiene conto dell'eteroschedasticità dei dati calcolabile con le precedenti quantità è pari a

$$U_t = W_t^T S2_t^{-1} W_t. \tag{36}$$

La carta chiama un allarme se $U_t > L$, dove L è un valore critico scelto via simulazione, eventualmente variabile con il tempo, per raggiungere un ARL_0 desiderato. Per un ulteriore esempio di carta MEWMA con statistiche basate sullo *score* si veda Zhang et al. (2016).

4 Sorveglianza dei casi di dengue con alcune carte proposte e soluzioni alternative

In questo capitolo verranno introdotti con maggiore dettaglio le caratteristiche dei dati utilizzati per costruire l'intero processo di sorveglianza, andando ad analizzare tutte le nuove scelte fatte per la modellazione in fase I che andranno poi a ripercuotersi anche in fase II. L'obiettivo del riprodurre le carte proposte in letteratura è quello di verificare con le nuove accortezze la coerenza dei risultati ottenuti con quelli degli autori. Carte per la sorveglianza congiunta dei parametri incontrate finora solo dal punto di vista teorico verranno infine applicate per raffinare ulteriormente il controllo dei parametri, valutando l'affidabilità dei risultati. Tutti i risultati grafici sono riportati nelle sezioni finali con lo scopo di agevolarne il confronto.

4.1 Descrizione dei dati

Riprendendo lo scenario descritto in sezione 2.4.2, il fenomeno dell'epidemia di dengue è sempre più dilagante ed attualmente ci sono 46 Paesi in tutta l'America che lo monitorano, in particolare i Paesi dell'America Latina, in cui il clima tropicale e la presenza di determinati eventi atmosferici, come il noto El Niño, favoriscono il proliferare delle zanzare.
Il lavoro di García-Bustos and Zambrano (2022) si concentra nell'analizzare i casi di dengue nella zona costiera dell'Ecuador, Paese che ha visto un incremento dei casi notevole negli ultimi anni, come si può notare dalla Figura (4.1).
Le analisi successive mirano a verificare l'adeguatezza delle proposte dei suddetti autori sia nella modellazione dei dati sia nell'applicazione delle carte. Per questo motivo verranno introdotte alcune modifiche di miglioramento e proposte alternative.

4.2 Fase I: Stima dei parametri in controllo

L'obiettivo in questa fase è individuare il processo in controllo per la stima dei parametri necessari nella fase di sorveglianza prospettica. I dati utilizzati sono

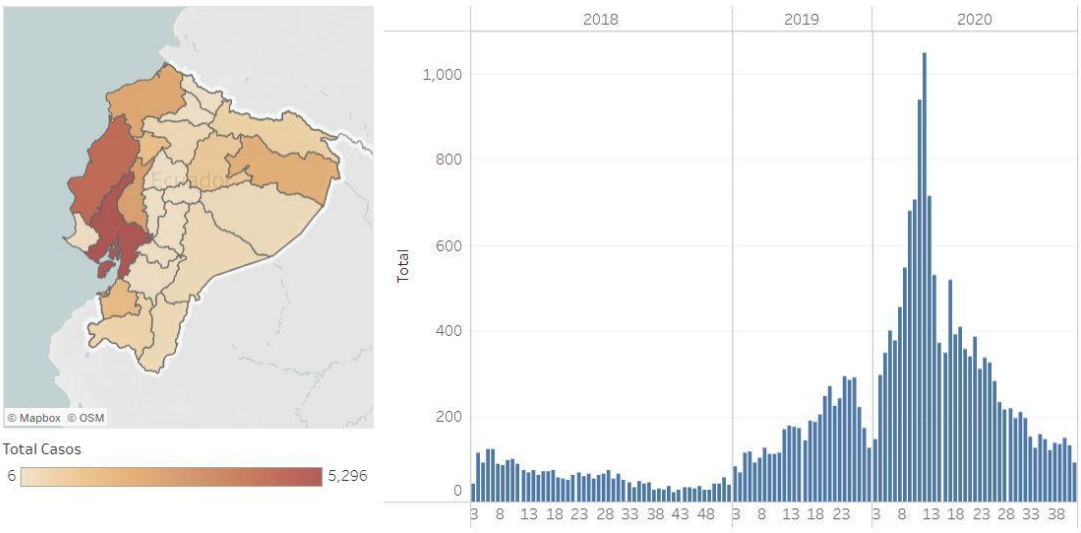

Figura 4.1: A sinistra distribuzione del totale di casi di dengue registrati negli anni 2018, 2019 e 2020 in Ecuador; a destra aumento settimanale dello stesso fenomeno dal punto di vista grafico. Fonte: PLISA, PAHO open data.

quelli forniti da PAHO, che mette a disposizione solo conteggi dei casi a partire dal 2018 per un totale di 121 settimane. Verranno inoltre prese in considerazione tutte le regioni e non solo quelle costiere.
Quando si ha a che fare con dati di conteggio è naturale utilizzare un modello di Poisson per stimare il numero di eventi nell'unità di tempo considerata. In questo caso, con dati epidemiologici riguardanti la trasmissione di un virus per mezzo di zanzare, è ragionevole aspettarsi che la varianza dei casi non coincida con la media, ma sia superiore. Questo fenomeno come già visto si chiama sovradispersione e può essere meglio modellato attraverso un modello binomiale negativo in cui compare k, il parametro di dispersione.

Per verificare l'adeguatezza di un modello binomiale negativo è stato prima stimato un modello di Poisson in controllo, cioè considerando le settimane senza epidemia corrispondenti alla prima fino alla 52esima del 2018, visibili in Figura (4.1). Le variabili dipendenti usate sono la settimana e la stagione, come presente in García-Bustos and Zambrano (2022), calcolando in seguito la statistica di Pearson per verificare la presenza o meno di sovradispersione. La statistica produce risultati coerenti con l'ipotesi di sovradispersione, individuando un *p-value* molto piccolo (si veda Appendice).

In seguito si è quindi adattato ai dati in controllo un modello binomiale negativo come proposto dagli autori in (26), in cui compare come predittore solamente l'identificativo di settimana e lo stesso al quadrato, quantità molto importante perché legata al fattore stagionalità. Ricordando la presenza di due stagioni, quella delle piogge e quella secca, la dengue è particolarmente presente durante e dopo la stagione delle piogge, cioè a inizio anno, per poi calare progressivamente.
Tornando al modello di riferimento individuato, quest'ultimo prevede una diminuzione dei casi all'aumentare della settimana considerata, dato il coefficiente negativo, fornendo una soluzione che potrebbe risultare semplicistica, spingendo la carta a chiamare un allarme solo per la natura della costruzione.
Un miglioramento al modello, ragionevole in questo caso, è considerare in aggiunta la presenza di ciclicità connessa alle settimane dell'anno, utilizzando opportune funzioni matematiche come seno e coseno che permettono di esprime la componente di periodicità. Nella Figura (4.2) è possibile confrontare i due modelli sopra proposti, notando il miglioramento nell'adattamento ai dati conseguenti all'aggiunta degli elementi di periodicità.
Nonostante il netto vantaggio dato dalla nuova proposta nel cogliere l'andamento osservato, in particolare la diminuzione e la risalita dopo circa la 40esima settimana, notiamo come neppure questa specificazione riesca a cogliere adeguatamente l'iniziale aumento dei casi.
Per cercare di incorporare questo aspetto si è scelto l'utilizzo delle splines di regressione (si veda la sezione 4.2.1 per i dettagli teorici).

Nel seguito, per l'adattamento del modello, si è selezionato un modello binomiale negativo con spline di regressione quadratica, cioè con grado $p = 2$, con due nodi scelti in questo caso direttamente sulla base della forma della curva dei conteggi in controllo, nel valore dell'esplicativa (la settimana) pari a 4 e 40. Dalla teoria sulle splines si sa che con queste specifiche si ottiene un modello con cinque predittori, inclusa l'intercetta. Le stime dei coefficienti delle funzioni di base sono riportate in Tabella 4.1 dove si nota che Base2 e Base4 risultano non significativi, ma per le finalità delle analisi vengono comunque mantenuti.
La bontà di adattamento del modello così ottenuto è valutato per prima cosa attraverso l'analisi dei residui di Figura (4.3). Nei grafici in alto a sinistra la dispersione dei punti è informativa sulla possibile omissione di termini quadratici o di errata

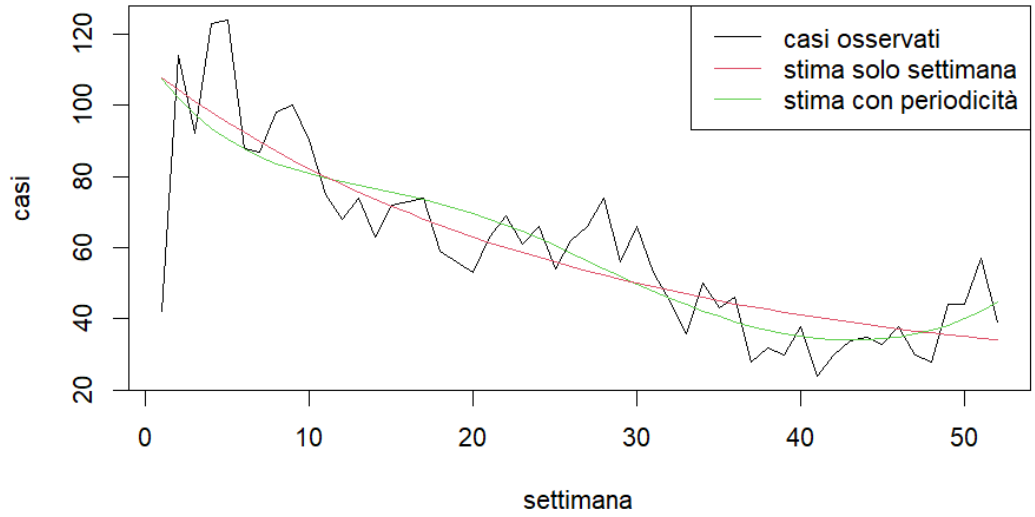

Figura 4.2: Confronto tra i casi di dengue osservati (in nero), stimati con solo la componente settimanale (in rosso) e con l'aggiunta di componenti periodiche (in verde)

specificazione della funzione di legame, mentre il grafico in alto a destra aiuta a valutare se la funzione di varianza è specificata adeguatamente. In entrambi i casi non si notano andamenti "sospetti" dei punti. Il grafico in basso a sinistra mostra che i residui non sono normali ma la loro dimensione è quella sensata, indicando che il modello ha colto bene la media e la varianza. Inoltre può essere valutato in basso a destra il grafico dell'autocorrelazione residua, una struttura di correlazione per ϵ che si presenta specialmente se le osservazioni sono ordinate nel tempo, dove uno dei principali motivi per la presenza di errori autocorrelati è l'esclusione dal modello di variabili esplicative rilevanti. Se, per esempio, la variabile risposta Y_t è legata all'esplicativa X_t, non inclusa nel modello, l'effetto di X_t verrà incluso nel termine d'errore ϵ_t e se X_t è correlata con $X_{t-1}, X_{t-2}, ...$ questa correlazione verrà espressa nel modello da una correlazione tra ϵ_t e $\epsilon_{t-1}, \epsilon_{t-2}, ...$
Alla luce di queste considerazioni il modello scelto può essere considerato soddisfacente.

	Coefficienti	Std. Error	Statistiche Z	p-value
Intercetta	3.990	0.156	25.551	$<$ 2e-16
Base1	0.716	0.181	3.959	7.53e-05
Base2	0.174	0.176	0.989	0.323
Base3	-0.641	0.184	-3.475	0.001
Base4	-0.101	0.193	-0.525	0.599
Devianza nulla: 298.641 con 51 gradi di libertà				
Devianza residua: 50.475 con 47 gradi di libertà				

Tabella 4.1: Risultati dall'adattamento del modello binomiale negativo ai dati in controllo

Il modello binomiale negativo così costruito è quindi rappresentato in Figura (4.4) e riproduce adeguatamente l'andamento dei casi in controllo. Quest'ultimo sarà preso a riferimento per lo sviluppo della fase II.

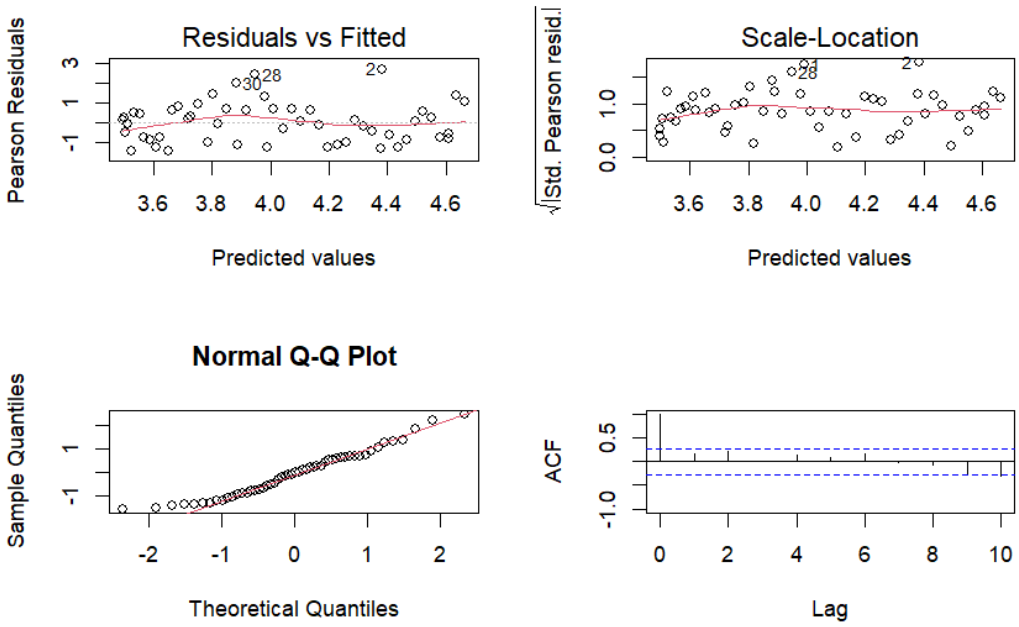

Figura 4.3: Controllo empirico dell'adattamento del modello binomiale negativo ai dati in controllo

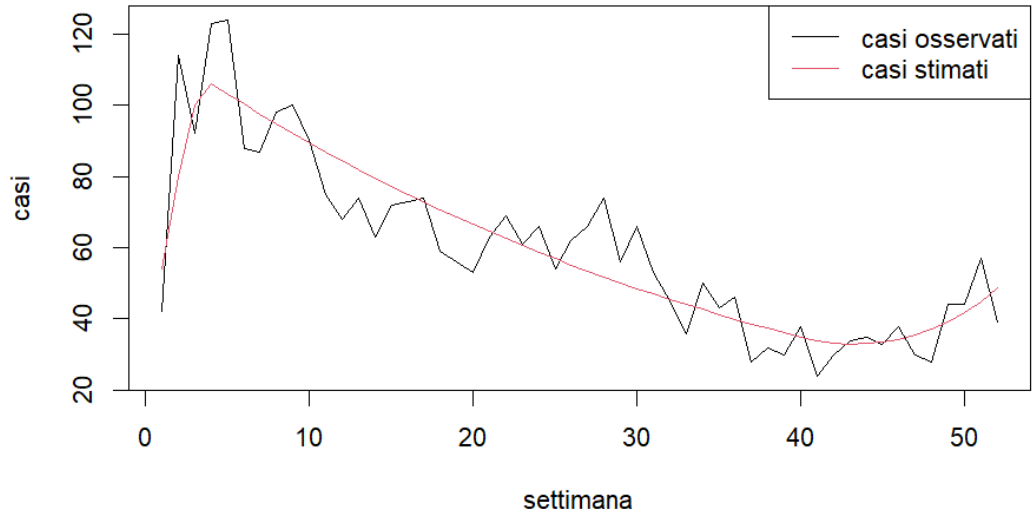

Figura 4.4: Confronto tra i casi di dengue osservati (in nero) e stimati con spline di regressione quadratica (in rosso)

4.2.1 Spline di regressione

In matematica il termine spline è stato utilizzato per la costruzione di funzioni polinomiali a tratti per approssimare funzioni di cui si conosce il valore solo in alcuni punti fissati, chiamati nodi, ovvero si interpola questi punti lasciando più libertà nel resto dello spazio, purché complessivamente presentino un comportamento regolare. Indicando in maniera generale la relazione tra esplicativa e variabile risposta come

$$Y_i = m(x_i) + \epsilon_i, \quad \text{i=1,2,...} \tag{37}$$

un modo per stimare m(x) è dato dall'utilizzo delle funzioni splines. In particolare, siano $t_1 < t_2 < ... < t_Z$ gli Z nodi, una funzione spline polinomiale di grado p è data da

$$s(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + b_1(x - t_1)_+^p + ... + b_Z(x - t_Z)_+^p, \tag{38}$$

dove il generico

$$(x - t_j)_+^p = \begin{cases} 0 & \text{se } x < t_j \\ (x - t_j)^p & \text{se } x \geq t_j \end{cases}$$

Tra due nodi successivi $s(x)$ coincide con un opportuno polinomio e per mantenere una certa regolarità nei punti di giunzione si richiede che la funzione abbia derivate fino al grado $p - 1$ continue in ogni t_j.
Dunque ogni funzione spline $s(x)$ di ordine p con Z nodi può essere rappresentata come

$$s(x) = \sum_{j=1}^{Z+p+1} \beta_j s_j(x), \tag{39}$$

con

$$s_j(x) = \begin{cases} x^{j-1} & \text{se } j = 1, ..., p+1 \\ (x - t_{j-(p+1)})_+^p & \text{se } j = p+2, ..., Z+p+1 \end{cases}$$

Lo spazio delle funzioni splines con K nodi è uno spazio lineare a $Z + p + 1$ dimensioni e le singole funzioni $s_j(x)$ prendono il nome di funzioni di base.
Tornando al modello di regressione (37), si ipotizza di approssimare $m(x)$ con una spline polinomiale $s(x)$, di conseguenza si avrà

$$Y_i = \sum_{j=1}^{Z+p+1} \beta_j s_j(x_i) + \epsilon_i \tag{40}$$

e il metodo delle splines di regressione consiste nel trovare la migliore approssimazione *spline* andando a minimizzare i minimi quadrati per la stima del vettore β:

$$\hat{\beta} = \underset{\beta \in R^{Z+p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{Z+p+1} \beta_j s_j(x_i) \right)^2. \tag{41}$$

Per implementare le splines di regressione occorre scegliere il grado del polinomio p nonché il numero e la posizione dei nodi t_j. Tuttavia, a lato pratico, si propende quasi sempre per la scelta di $p = 3$ poiché l'occhio umano non riesce a cogliere discontinuità dalla derivata terza in poi e si parla quindi di splines cubiche. Per

la scelta del numero di nodi e della posizione di solito i nodi si distribuiscono uniformemente sul range dell'esplicativa x. In generale una strategia è fissare Z, il numero dei nodi, per poi fissarli in corrispondenza di opportuni quantili di x oppure utilizzare strategie in qualche modo più oggettive che tengono conto del compromesso varianza-distorsione, come la *cross-validation*. In questo caso si sceglierà il numero Z di nodi che minimizza una certa funzione di perdita, come l'errore quadratico medio, dopo un'opportuna divisione ed iterazione della procedura sui dati. Per un approfondimento su quest'ultimo tema si veda il paragrafo 3.5.2 di Azzalini and Scarpa (2002) e per le splines il Capitolo 5 di Hastie et al. (2009).

4.3 Fase II: Sorveglianza prospettica

In fase II i dati sono analizzati sequenzialmente, man mano che sono raccolti e lo scopo è individuare il più rapidamente possibile deviazioni da uno stato soddisfacente (o almeno accettabile) verificando attraverso una serie di test d'ipotesi se il processo è ancora in controllo ai nuovi istanti di tempo o se è già andato fuori controllo. Quando la carta di controllo segnala un allarme l'obiettivo è quello di intervenire sul processo cercando di individuare la fonte di variabilità indesiderata ed eliminarla, aspetto cruciale soprattutto in casi in cui si ha a che fare con fenomeni critici che coinvolgono vite umane.
Le carte che verranno applicate in seguito corrispondono alla proposta di García-Bustos and Zambrano (2022) e all'ulteriore proposta di miglioramento, la MEWMA-*score*, valutata insieme alle proposte alternative, cioè le carte combinate EWMA–*score* e CUSUM.
Per individuare il limite di controllo superiore, essendo interessati a individuare velocemente solo aumenti nel numero di casi, si è scelto di utilizzare sempre i limiti di controllo dinamici, come descritti nella sezione 3.1 e di utilizzare un valore di ARL in controllo pari a 520 settimane, che corrisponde ad un falso allarme ogni 10 anni. La costante di lisciamento è scelta pari a 0.05, sapendo da Urbieta et al. (2017) che piccoli valori di λ offrono una miglior performance per salti della media grandi e piccoli. La stima dei parametri in controllo è fornita dall'adattamento del modello binomiale negativo con spline di regressione in fase I; da questo si ricava la media in controllo per ogni settimana, cioè i valori previsti dal modello

per il 2018 che verranno ciclati per le settimane a seguire, il k in controllo pari a 86.2 e la diagonale della matrice cappello H, utile per le quantità studentizzate.

Nel seguito tutte le carte di fase II sono volutamente costruite coinvolgendo anche le osservazioni in controllo, procedura non richiesta quando l'obiettivo è la sorveglianza prospettica. Avendo solo il 2018 come periodo in controllo, questa scelta è stata "obbligata" per mostrare al lettore come le carte non chiamino falsi allarmi, cosa non possibile se si fosse partiti, giustamente, dal 2019.
Di conseguenza anche il valore delle statistiche viene leggermente modificato, ricordando che abbiamo a che fare con quantità che cumulano le osservazioni nel tempo, ma i risultati finali non subiscono cambiamenti significativi.

4.3.1 Carta di controllo EWMA basata sui residui di Pearson

La prima carta di controllo EWMA affrontata da García-Bustos and Zambrano (2022) è la carta che costruisce le statistiche di controllo basandosi sui residui di Pearson, ottenendo lo schema di controllo in Figura 4.5. Dall'osservazione della carta si può notare come la procedura proposta funzioni per i dati in analisi. Più in dettaglio, la carta non chiama un allarme per il 2018, anche se intorno alla 30esima settima le statistiche si discostano visivamente da 0 avvicinandosi al limite, frutto solamente di un aumento dei casi, ma non allarmante. Il primo vero allarme viene chiamato alla 63esima settimana, cioè l' undicesima settimana del 2019, poiché il numero di casi comincia ad aumentare senza scendere come ci si aspetterebbe sulla base del 2018 (si veda Figura 4.1 per l'andamento osservato dei casi). Da quella settimana in poi le statistiche superano il limite e continuano a crescere, indicatore che il fuori controllo permane nel processo.
I limiti dinamici sono ben riconoscibili dall'andamento differente nel tempo, visibili anche nelle carte successive.

4.3.2 Carta di controllo EWMA basata sui residui di devianza

Successivamente la carta EWMA che costruisce le statistiche di controllo sulla base dei residui di devianza di un modello binomiale negativo si presenta come in Figura 4.6.
Dall'osservazione di questa carta si può notare come la procedura proposta fornisca risultati pressoché identici alla precedente carta basata sui residui di Pearson. Anche in questo caso si può notare come la carta non chiami mai falsi allarmi, come le statistiche comincino a crescere senza portare un allarme intorno alla 30esima settimana e come poi il vero allarme venga raggiunto alla 62esima settimana, cioè alla decima settimana del 2019. La carta è quindi leggermente più reattiva rispetto a quella sui residui di Pearson.

4.3.3 Carta di controllo EWMA basata sui residui di Pearson studentizzati

A seguire viene rappresentata la carta EWMA le cui statistiche di controllo cumulano i residui studentizzati, visibile in Figura 4.7.
Da quest'ultima si può notare un andamento delle statistiche molto simile a quello già incontrato nelle due precedenti carte, anche per quanto riguarda le considerazioni. Una differenza che salta agli occhi in questa carta è la reattività delle statistiche man mano che ci si avvicina alla 60esima settimana; nelle carte precedenti l'aumento era più graduale, mentre in questo caso notiamo che le statistiche passano subito ad un livello molto vicino al limite, portando la carta a chiamare un allarme alla 61esima settimana, cioè alla nona del 2019. La carta in questione offre quindi una performance migliore delle precedenti dal punto di vista della reattività.

4.3.4 Carta di controllo EWMA basata sui residui di devianza studentizzata

Come ultima carta riprodotta dalle proposte teoriche analizzate si è costruita la carta EWMA basata sui residui di devianza studentizzati di Figura 4.8. Anche in questo caso l'andamento complessivo della carta mima i risultati visti in precedenza,

cioè non chiama falsi allarmi e chiama un allarme veritiero alla 61esima settimana, cioè alla nona del 2019, come precedentemente visto in Figura 4.7.
Per concludere, usare la forma studentizzata delle statistiche da cumulare sembra portare performance delle carte migliori rispetto alla versione non studentizzata.

4.3.5 Carta di controllo MEWMA-*score*

Nel paragrafo 3.3 si è introdotto dal punto di vista teorico lo strumento delle carte di controllo multivariate, utili per la sorveglianza congiunta di più di un parametro la cui implementazione è riportata in Figura 4.9.
Dall'osservazione della carta in questione si intuisce che cumulare score di verosimiglianza risulta in una performance adeguata e del tutto comparabile agli schemi precedenti, la carta non chiama mai un allarme con valori delle statistiche ben sotto i limiti dinamici e il primo, vero, allarme è chiamato alla 63esima settimana, cioè all'undicesima del 2019. La precedente carta è informativa per la presenza di un fuori controllo nell'andamento dei casi registrati, ma non è possibile capire la natura dell'allarme, cioè se è effettivamente dovuto ad una variazione della media μ o di k. Utilizzando la statistica MEWMA-*score* standardizzata ricavabile dall'implementazione della carta è però possibile analizzarne l'andamento nel tempo, come rappresentato in Figura 4.10, traendo risultati interessanti. Dal grafico superiore sembrerebbe che l'allarme sia dovuto ad un aumento della media del numero di casi, ma allo stesso tempo anche da una diminuzione del parametro k e quindi da un aumento di varianza.

4.3.6 Carte di controllo EWMA-*score* combinate

Alla luce dei risultati precedenti sembra plausibile cercare una soluzione per la sorveglianza congiunta dei parametri per cui sia possibile risalire al singolo autore dell'allarme quando la carta va fuori controllo.
I due parametri di interesse sono stati quindi sorvegliati con due carte di controllo EWMA combinate basate sempre sullo *score* ed entrambe unilaterali per sorvegliare in maniera più efficiente i veri aspetti critici del caso in questione, cioè un aumento di μ e una diminuzione di k. Per la costruzione delle carte si è seguita la procedura teorica del paragrafo 3.2 e si sono ottenute le carte in Figura 4.11. Si veda il

codice riportato in Appendice per il dettaglio nell'implementazione.
Dall'osservazione dei due grafici si nota come effettivamente il fuori controllo sia dovuto ad una variazione ricollegabile ad entrambi i parametri, confermando l'intuizione avuta con l'osservazione del grafico in Figura 4.10. Le statistiche di controllo delle due carte congiunte mimano l'andamento della carta multivariata, chiamando un unico vero allarme alla 63esima settimana, cioè all'undicesima del 2019.

4.3.7 Carte di controllo CUSUM combinate

La stessa proposta di carte combinate può essere sviluppata anche attraverso il disegno di carte CUSUM per la sorveglianza congiunta dei parametri. Il ragionamento logico dell'implementazione segue quello delle carte EWMA viste, con l'adattamento alla natura della carta. In particolare, le due carte CUSUM che cumulano separatamente le variazioni dovute ad aumenti della media μ e quelle dovute alla diminuzione del parametro k si presentano come in Figura 4.12. Queste carte sono costruite per identificare un cambiamento moltiplicativo δ dei parametri, cioè $\mu_{1,t} = \delta\mu_{0,t}$ e di $\sigma^2_{1,t} = \delta\sigma^2_{0,t}$, quest'ultima specificazione utile per risalire al cambiamento d'interesse in termini di k.
Le carte disegnate sono ottimali per un cambiamento pari al 20% del valore originale, cioè $\delta = 1.2$, scelta che comunque può essere modificata. Si veda il codice riportato in Appendice per un maggior dettaglio.
Entrambe le carte della Figura 4.12 confermano quindi i risultati ottenuti con le carte EWMA combinate, chiamando nessun falso allarme e segnalando per la prima volta un cambiamento sia per la media μ che per k alla 63esima settimana, cioè l'undicesima del 2019.

4.3.8 Carte di controllo per la sorveglianza della media μ

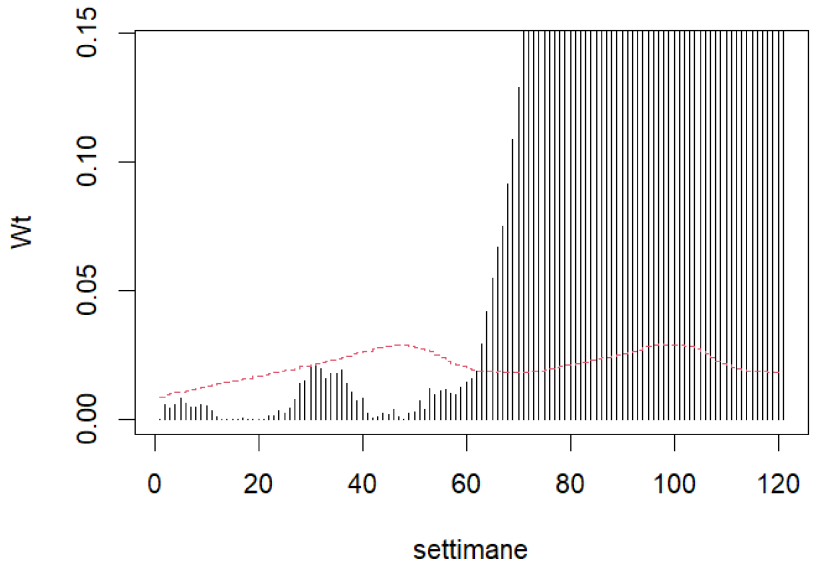

Figura 4.5: Carta di controllo EWMA basata sui residui di Pearson

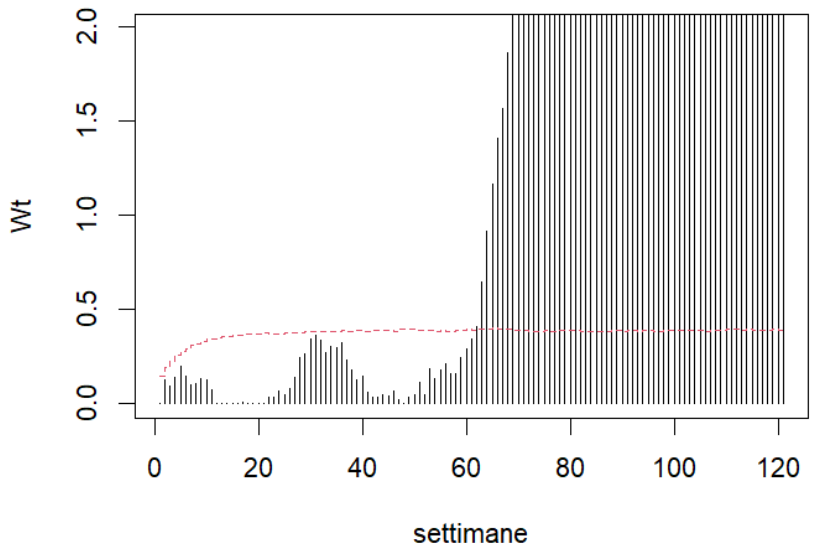

Figura 4.6: Carta di controllo EWMA basata sui residui di devianza

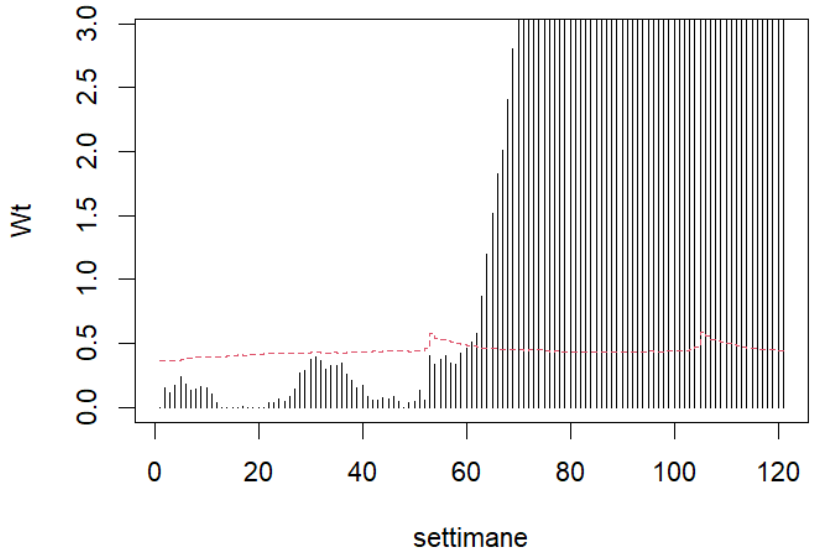

Figura 4.7: Carta di controllo EWMA basata sui residui di Pearson studentizzati

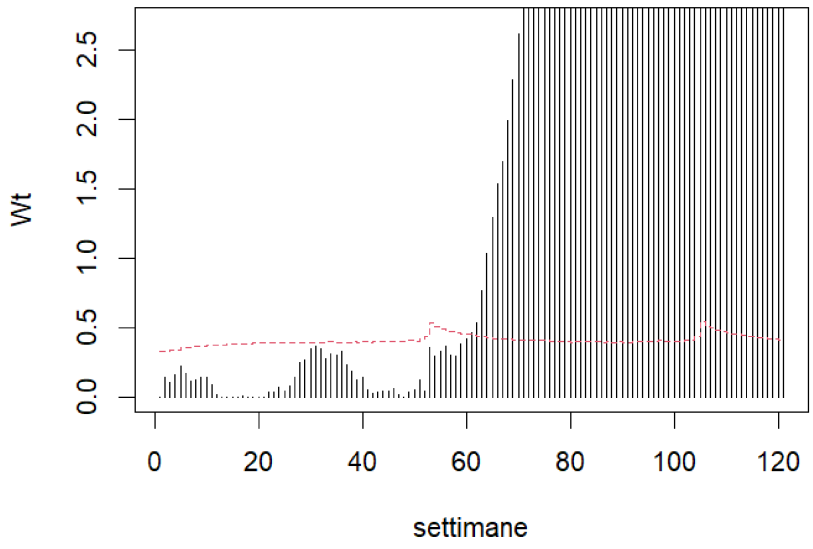

Figura 4.8: Carta di controllo EWMA basata sui residui di devianza studentizzati

4.3.9 Carte di controllo per la sorveglianza congiunta della media μ e del parametro di dispersione k

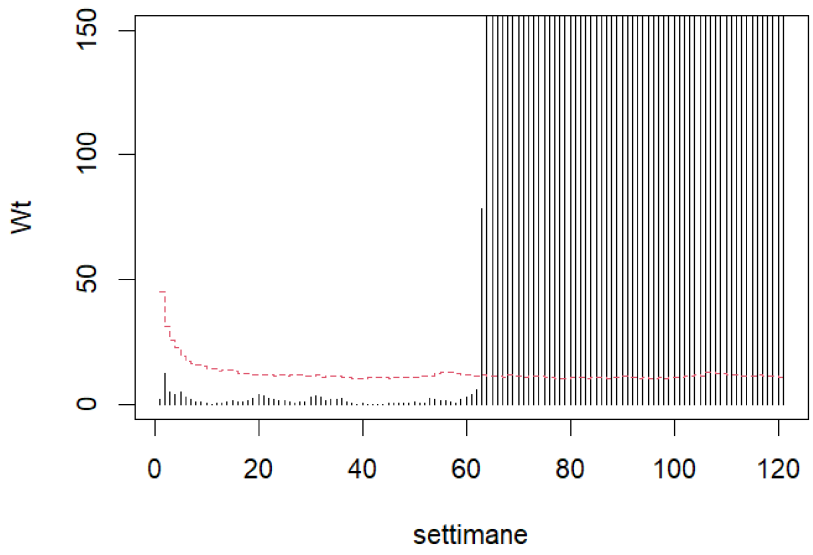

Figura 4.9: Carta di controllo multivariata MEWMA-*score*

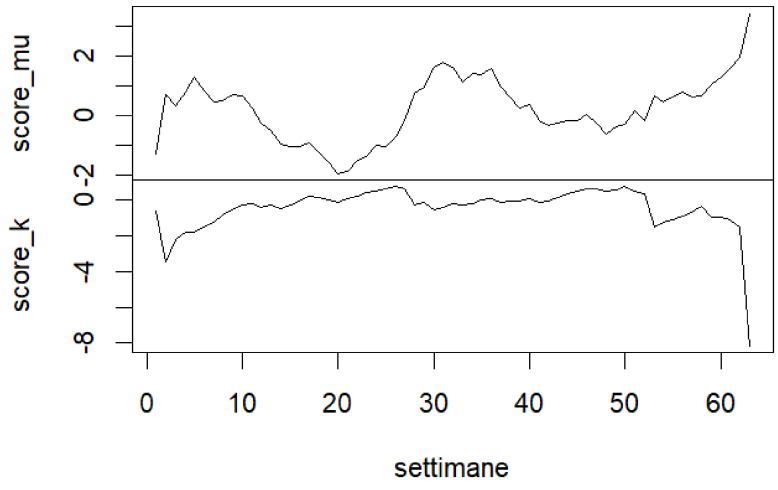

Figura 4.10: Andamento della statistica MEWMA-*score* standardizzata separatamente per μ (quadrante superiore) e k (quadrante inferiore) fino al primo allarme

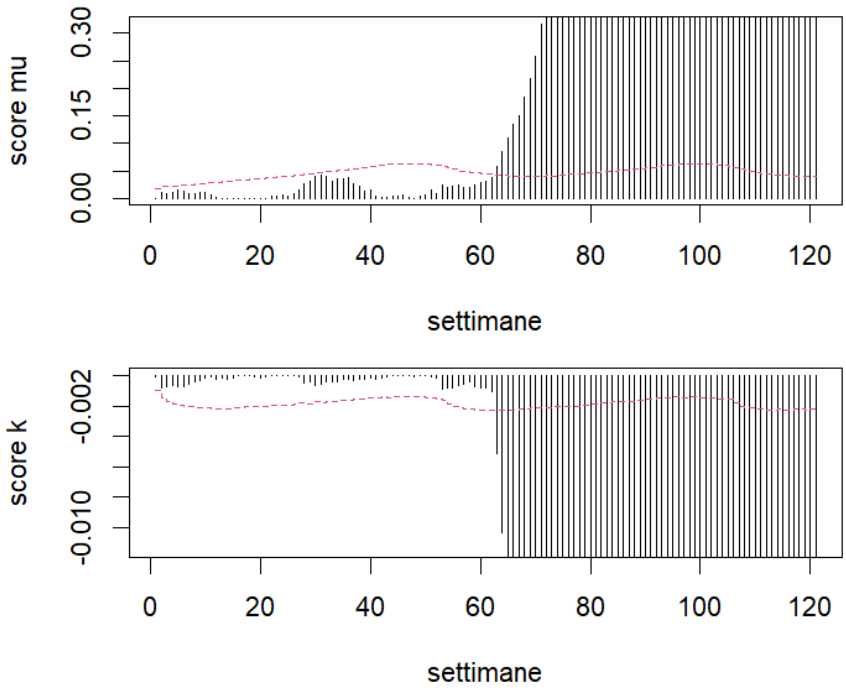

Figura 4.11: Carte di controllo EWMA-*score* combinate

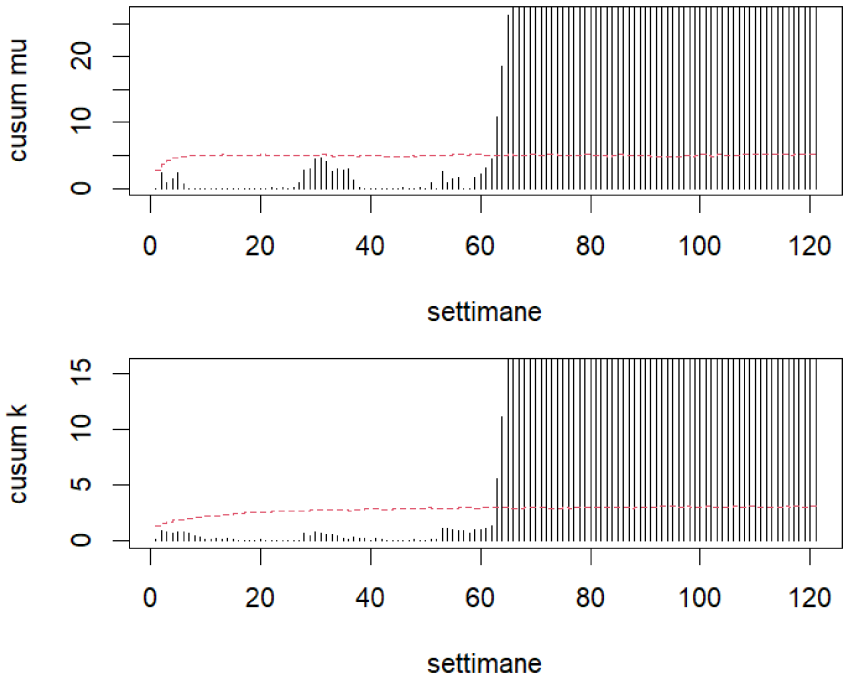

Figura 4.12: Carte di controllo CUSUM combinate

5 Conclusioni e possibili approfondimenti futuri

L'obiettivo di questa tesi è quello di applicare gli strumenti del controllo statistico di processo per monitorare dati di conteggio che presentano sovradispersione, fenomeno molto comune in ambito epidemiologico dove supporre che la media e la variabilità dei casi osservati coincidano risulta alquanto improbabile. Sempre in questo ambito, gli strumenti del SPC possono rappresentare una risorsa preziosa per capire che qualcosa sta cambiando e avere l'occasione di intervenire tempestivamente, aspetto fondamentale come ben conosciamo per cercare di contenere e contrastare i contagi nella popolazione.
Per modellare dati di questa natura, come introdotto nel primo capitolo, si è fatto uso della distribuzione binomiale negativa, una distribuzione più flessibile rispetto alla classica Poisson che permette di tenere in considerazione il fenomeno della sovradispersione grazie all'utilizzo del parametro k. Nel seguito sono state presentate una serie di proposte presenti in letteratura sulla sorveglianza diretta di statistiche basate su dati binomiali negativi o su un loro tentativo di ricondursi alla distribuzione normale, in ogni caso focalizzato su un possibile cambiamento del livello medio nei dati. Esempi pratici sono stati infine riportati.
Gli sviluppi ulteriori mirano quindi a mettere in luce aspetti tralasciati o poco chiari degli schemi teorici finora proposti, in primo luogo considerando anche il parametro di dispersione un aspetto critico da monitorare congiuntamente alla media grazie ad opportune carte.
Un punto molto rilevante quando si implementano gli schemi di sorveglianza riguarda l'utilizzo di limiti adeguati, per garantire le prestazioni desiderate ed evitare troppi falsi allarmi che intaccano la credibilità della carta. A questo scopo sono stati presentati i limiti dinamici, limiti che offrono adattabilità in molti contesti e semplicità di calcolo rispetto ad alternative più classiche.
Un ampio spazio è stato infine dedicato alle applicazioni pratiche, andando a riprodurre le analisi di García-Bustos and Zambrano (2022) per la sorveglianza dei casi di contagio da dengue in Ecuador, ma con personali proposte di miglioramento. Per esempio il modello in controllo è stato rielaborato utilizzando delle splines di regressione invece che una semplice funzione quadratica con l'indicatore di settimana come predittore e la scelta dei limiti è ricaduta su quelli dinamici al

posto di quelli classici. I risultati ottenuti dalla singola sorveglianza della media μ riproduco quelli raggiunti dagli autori, ma con maggior credibilità considerando gli andamenti stagionali fondamentali nel problema trattato.
La sorveglianza congiunta dei parametri della binomiale negativa è stata sviluppata attraverso l'uso di una carta EWMA multivariata basata sullo *score* di verosimiglianza (MEWMA-*score*), uno strumento molto usato in ambito inferenziale, e delle carte EMWA-*score* e CUSUM combinate, cioè la costruzione con alcune accortezze di tante carte quanti i parametri d'interesse. Utilizzando sempre i limiti dinamici, le nuove proposte di sorveglianza congiunta risultano equivalenti tra loro in termini di tempistiche nell'allarme e coerenti rispetto le precedenti carte univariate, confermando quindi la loro efficacia, ma con il vantaggio di una gestione migliore dell'informazione disponibile e della comprensione del fuori controllo.
Concentrandosi infatti solo sulla sorveglianza della media μ si andrebbe a trascurare k, ignorando il suo contributo nel generare il fuori controllo osservato.
Lavori futuri potrebbero essere dedicati a completare l'analisi con studi di simulazione per confrontare la performance delle nuove proposte in diversi scenari.

A Appendice

Codice R utilizzato per l'analisi

Al fine di poter riprodurre l'analisi svolta in questo studio, di seguito ne vengono riportati tutti i passaggi. I grafici della carta di controllo prodotta nei vari casi vengono omessi, in quanto già presentati ai paragrafi 4.3.8 e 4.3.9.

A.1 Funzioni utilizzate per le carte di controllo

```
#stima modelli parametrici in controllo

#modello di Poisson
mod <- glm(casi~settimana+stagione, data=dati[1:52,], family=poisson)

#test di conferma presenza sovradispersione
X2 <- sum(residuals(mod, type="pearson")^2)
pchisq(X2, 49, lower.tail=FALSE)
##4.968196e-11

#primo modello
mod2 <- glm.nb(casi~settimana+settimana2, data=dati[1:52,])

#elementi periodici
t=1:52 #settimane in controllo
x <- sin((2*pi*t)/52)
y <- cos((2*pi*t)/52)

#secondo modello
mod1 <- glm.nb(casi~x+y+settimana+settimana2, data=dati[1:52,])
```

```
#terzo modello con spline di regressione
Bs1 <- splines::bs(1:52, degree = 2, knots = c(4, 40))
sp1 <- MASS::glm.nb(y ~ Bs1)

#ingredienti in controllo che servono per applicare le carte
mu.hat <- fitted(sp1)
k.hat <- sp1$theta
Hdiag <- influence(sp1)$hat


set.seed(123)
lambda <- 0.05
B <- 52 * 10 #ARL in controllo
Nsim <- 100 * B
week <- 1:121


#Ciclo carta EWMA basata sui residui di Pearson
t <- 1
Wt <- 0
Wstar <- rep(0, Nsim)
W <- L <- numeric(121)

for (i in ((week - 1) %% 52) + 1) {

    # residui di Pearson al tempo t
    et <- (dati$casi[t] - mu.hat[i]) / sqrt(mu.hat[i] +
    ((mu.hat[i]^2) / k.hat))

    # aggiornamento statistiche in controllo
    Wt <- max(0, (lambda * as.numeric(et) + (1 - lambda) * Wt))

    # aggiornamento statistiche simulate in controllo
    Wstar <- pmax(0, lambda * ((rnbinom(Nsim, mu = mu.hat[i],
```

```
    size = k.hat) - mu.hat[i]) / sqrt(mu.hat[i]
    + (mu.hat[i]^2 / k.hat))) + (1 - lambda) * Wstar)

    # Limiti
    Lt <- quantile(Wstar, 1 - 1 / B)

    # Sostituzione Wstar>Lt
    idx <- which(Wstar > Lt)
    Wstar[idx] <- sample(Wstar[Wstar <= Lt], length(idx))

    # Memorizzazione dei risultati
    W[t] <- Wt
    L[t] <- Lt
    t <- t + 1
}

#Ciclo carta EWMA basata sui residui di devianza
t <- 1
Wt <- 0
Wstar <- rep(0, Nsim)
W <- L <- numeric(121)

for (i in ((week - 1) %% 52) + 1) {

    # residui di devianza al tempo t
    dt <- sign(dati[t, 1] - mu.hat[i]) *
    sqrt(2 * dati[t, 1] * log(dati[t, 1] / mu.hat[i])
    - 2 * (k.hat + dati[t, 1]) *
    log((k.hat + dati[t, 1]) / (k.hat + mu.hat[i])))

    # aggiornamento statistiche in controllo
    Wt <- max(0, (lambda * as.numeric(dt) + (1 - lambda) * Wt))
```

```
    # aggiornamento statistiche in controllo simulate
    n <- rnbinom(Nsim, mu = mu.hat[i], size = k.hat)
    Wstar <- pmax(0, lambda * (sign(n - mu.hat[i])*
    sqrt(2 * n * log(n / mu.hat[i])- 2 * (k.hat + n) *
    log((k.hat + n)/ (k.hat + mu.hat[i])))) + (1 - lambda) * Wstar)

    # Limiti
    Lt <- quantile(Wstar, 1 - 1 / B)

    # Sostituzione Wstar>Lt
    idx <- which(Wstar > Lt)
    Wstar[idx] <- sample(Wstar[Wstar <= Lt], length(idx))

    # Memorizzazione dei risultati
    W[t] <- Wt
    L[t] <- Lt
    t <- t + 1
}


# Ciclo carta EWMA basata sui residui di Pearson studentizzati
t <- 1
Wt <- 0
Wstar <- rep(0, Nsim)
W <- L <- numeric(121)

for (i in ((week - 1) %% 52) + 1) {

    # resiui di Pearson studentizzati al tempo t
    et <- (dati[t, 1] - mu.hat[i]) / sqrt((mu.hat[i]
    + (mu.hat[i]^2 / k.hat)) * (1 - Hdiag[i]))

    #  aggiornamento statistiche in controllo
    Wt <- max(0, (lambda * as.numeric(et) + (1 - lambda) * Wt))
```

```
    # aggiornamento statistiche in controllo simulate
    Wstar <- pmax(0, lambda * ((rnbinom(Nsim, mu = mu.hat[i],
    size = k.hat) - mu.hat[i]) / sqrt((mu.hat[i]
    + (mu.hat[i]^2 / k.hat)) *
    (1 - Hdiag[i]))) + (1 - lambda) * Wstar)

    # Limiti
    Lt <- quantile(Wstar, 1 - 1 / B)

    # Sostituzione Wstar>Lt
    idx <- which(Wstar > Lt)
    Wstar[idx] <- sample(Wstar[Wstar <= Lt], length(idx))

    # Memorizzazione dei risultati
    W[t] <- Wt
    L[t] <- Lt
    t <- t + 1
}



#Ciclo carta EWMA basata sui residui di devianza studentizzati
t <- 1
Wt <- 0
Wstar <- rep(0, Nsim)
W <- L <- numeric(121)

for (i in ((week - 1) %% 52) + 1) {

    # residui di devianza studentizzati al tempo t
    dt <- sign(dati[t, 1] - mu.hat[i]) * sqrt(2 * dati[t, 1] *
    log(dati[t, 1] / mu.hat[i]) - 2 * (k.hat + dati[t, 1]) *
    log((k.hat + dati[t, 1]) / (k.hat + mu.hat[i])))
```

```
    dts <- dt / sqrt(1 - Hdiag[i])

  # aggiornamento statistiche in controllo
    Wt <- max(0, (lambda * as.numeric(dts) + (1 - lambda) * Wt))

    # aggiornamento statistiche in controllo simulate
    n <- rnbinom(Nsim, mu = mu.hat[i], size = k.hat)
    Wstar <- pmax(0, lambda * ((sign(n - mu.hat[i]) *
    sqrt(2 * n * log(n / mu.hat[i]) - 2 * (k.hat + n) *
    log((k.hat + n) / (k.hat + mu.hat[i])))) /
    sqrt(1 - Hdiag[i])) + (1 - lambda) * Wstar)

    # Limiti
    Lt <- quantile(Wstar, 1 - 1 / B)

    # Sostituzione Wstar>Lt
    idx <- which(Wstar > Lt)
    Wstar[idx] <- sample(Wstar[Wstar <= Lt], length(idx))

    # Memorizzazione dei risultati
    W[t] <- Wt
    L[t] <- Lt
    t <- t + 1
}


#MEWMA-SCORE

#funzione score binomiale negativa per k
kscore <- function(y, mu, k) {
    1 + log(k) - log(k + mu) - (k + y) / (k + mu)
    + digamma(y + k) - digamma(k)
}
```

```
#funzione score binomiale negativa per mu e k
nbscore <- function(y, mu, k) {
    rbind(
        (y / mu) - (k + y) / (k + mu),
        kscore(y, mu, k)
    )
}

#informazione di Fisher
var.nbscore <- function(mu, k) {
    yt <- seq.int(0, qnbinom(1 - 1E-09, mu = mu, size = k))
    c(
        (mu + mu * mu / k) * (k / (mu * (k + mu)))^2,
        sum(kscore(yt, mu, k)^2 * dnbinom(yt, mu = mu, size = k))
    )
}

# Ciclo funzione calcolo MEMWA-score

t <- 1
Wt <- S2t <- Wstar <- 0
W <- L <- numeric(121)
MEWMA <- matrix(NA, nrow = 121, ncol = 2)

for (i in ((week - 1) %% 52) + 1) {

    # dati raccolti al tempo t
    nt <- as.numeric(dati[t, 1])
    mut <- mu.hat[i]

    # aggiornamento MEWMA dati osservati
    Wt <- lambda * nbscore(nt, mut, k.hat) + (1 - lambda) * Wt

    # aggiornamento MEWMA dati simulati in controllo
```

```
    n <- rnbinom(Nsim, mu = mut, size = k.hat)
    Wstar <- lambda * nbscore(n, mut, k.hat) + (1 - lambda) * Wstar

    # aggiornamento varianza
    S2t <- lambda^2 * var.nbscore(mut, k.hat) + (1 - lambda)^2 * S2t


    # Statistiche in controllo
    St <- sum(Wt * Wt / S2t)
    Ss <- colSums(Wstar * Wstar / S2t)

    # Limiti
    Lt <- quantile(Ss, 1 - 1 / B)

    # Sostituzione Wstar>Lt
    idx <- which(Ss > Lt)
    Wstar[, idx] <- Wstar[, sample(which(Ss <= Lt), length(idx))]

    # Memorizzazione dei risultati
    W[t] <- St
    L[t] <- Lt
    MEWMA[t, ] <- Wt / sqrt(S2t)
    t <- t + 1
}


## Carte combinate basate sugli score

#funzione del package dfphase1 per il calcolo dei limiti
dbalance2 <- function(stat, B) {
    dmodq <- function(stat, alpha) {
        eps <- .Machine$double.eps
        if (alpha <= eps) {
            q <- max(stat) + eps
```

```
        } else if (alpha >= 1 - eps) {
            q <- min(stat) - eps
        } else {
            q <- quantile(stat, 1 - alpha)
        }
        q
    }
    obj <- function(beta) {
        q <- apply(stat, 1, dmodq, beta)
        mean((stat[1, ] > q[1]) | (stat[2, ] > q[2])) - 1 / B
    }
    apply(stat, 1, dmodq, uniroot(obj, c(0, 1))$root)
}

#Ciclo costruzione delle carte EWMA-score combinate
t <- 1
Wt <- Wstar <- 0
L <- MEWMA <- matrix(NA, nrow = 121, ncol = 2)


for (i in ((week - 1) %% 52) + 1) {

    # dati raccolti al tempo t
    nt <- as.numeric(dati[t, 1])
    mut <- mu.hat[i]

    # aggiornamento MEWMA dati osservati
    obs <- nbscore(nt, mut, k.hat)
    obs[2] <- -obs[2]
    Wt <- pmax(0, lambda * obs + (1 - lambda) * Wt)

    # aggiornamento MEWMA dati simulati in controllo
    sim <- nbscore(rnbinom(Nsim, mu = mut, size = k.hat), mut, k.hat)
    sim[2, ] <- -sim[2, ]
```

```
    Wstar <- matrix(pmax(0, lambda * sim + (1 - lambda) * Wstar), 2)

    # Limiti
    Lt <- dbalance2(Wstar, B)

    # Sostituzione fuori controllo
    drop <- (Wstar[1, ] > Lt[1]) | (Wstar[2, ] > Lt[2])
    Wstar[, which(drop)] <- Wstar[, sample(which(!drop), sum(drop))]

    # Memorizzazione dei risultati
    L[t, ] <- Lt
    MEWMA[t, ] <- Wt
    t <- t + 1
}



#funzione per il cambiamento di mu e k da cumulare nella carta CUSUM
nblr <- function(y, mu, k, delta) {
    a <- dnbinom(y, mu = mu, size = k, log = TRUE)
    mu1 <- delta * mu
    k1 <- mu * mu / (delta * delta * (mu + mu * mu / k) - mu)
    cat(mu, mu1, k, k1, "\n")
    rbind(
        dnbinom(y, mu = mu1, size = k, log = TRUE) - a,
        dnbinom(y, mu = mu, size = k1, log = TRUE) - a
    )
}

#Ciclo costruzione delle carte CUSUM combinate
t <- 1
Wt <- Wstar <- 0
L <- CUSUM <- matrix(NA, nrow = 121, ncol = 2)
```

```
delta <- 1.2 #salto della media


for (i in ((week - 1) %% 52) + 1) {

    # dati raccolti al tempo t
    nt <- as.numeric(dati[t, 1])
    mut <- mu.hat[i]

    # aggiornamento CUSUM dati osservati
    obs <- nblr(nt, mut, k.hat, delta)
    Wt <- pmax(0, Wt + obs)

    # aggiornamento CUSUM dati simulati in controllo
    sim <- nblr(rnbinom(Nsim, mu = mut, size = k.hat),
    mut, k.hat, delta)
    Wstar <- matrix(pmax(0, Wstar + sim), 2)

    # Limiti
    Lt <- dbalance2(Wstar, B)

    # Sostituzione fuori controllo
    drop <- (Wstar[1, ] > Lt[1]) | (Wstar[2, ] > Lt[2])
    Wstar[, which(drop)] <- Wstar[, sample(which(!drop), sum(drop))]

    # Memorizzazione dei risultati
    L[t, ] <- Lt
    CUSUM[t, ] <- Wt
    t <- t + 1
}
```

Riferimenti bibliografici

Albarracin, O. Y. E., Alencar, A. P. and Lee Ho, L. (2018) Effect of neglecting autocorrelation in regression EWMA charts for monitoring count time series. *Quality and Reliability Engineering International*, **34**, 1752–1762.

Albers, W. (2011) Control charts for health care monitoring under overdispersion. *Metrika*, **74**, 67–83.

Alencar, A. P., Lee Ho, L. and Albarracin, O. Y. E. (2017) CUSUM control charts to monitor series of negative binomial count data. *Statistical Methods in Medical Research*, **26**, 1925–1935.

Ali, S., Altaf, N., Shah, I., Wang, L. and Raza, S. M. M. (2020) On the effect of estimation error for the risk-adjusted charts. *Complexity*, **2020**, 1–21.

Aytaçoğlu, B. and Woodall, W. H. (2020) Dynamic probability control limits for CUSUM charts for monitoring proportions with time-varying sample sizes. *Quality and Reliability Engineering International*, **36**, 592–603.

Azzalini, A. and Scarpa, B. (2002) *Data Analysis and Data Mining Group*. Oxford University Press.

Chen, P., Fu, X., Ma, S., Xu, H. Y., Zhang, W., Xiao, G., Siow Mong Goh, R., Xu, G. and Ching Ng, L. (2020) Early dengue outbreak detection modeling based on dengue incidences in Singapore during 2012 to 2017. *Statistics in Medicine*, **39**, 2101–2114.

Fávero, L. P., de Freitas Souza, R., Belfiore, P., Corrêa, H. L. and Haddad, M. F. (2021) Count data regression analysis: Concepts, overdispersion detection, zero-inflation identification, and applications with R. *Practical Assessment, Research and Evaluation*, **26**, 1–22.

García-Bustos, S. and Zambrano, G. (2022) Control charts for health surveillance based on residuals of negative binomial regression. *Quality and Reliability Engineering International*, **38**, 2521–2532.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. Springer New York, NY.

Hawkins, D. and Olwell, P. (1998) *Cumulative Sum Charts and Charting for Quality Improvement.* Springer Verlag.

Hinde, J. and Demétrio, C. G. (1998) Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, **27**, 151–170.

Höhle, M. and Paul, M. (2008) Count data regression charts for the monitoring of surveillance time series. *Computational Statistics and Data Analysis*, **52**, 4357–4368.

Lloyd-Smith, J. O. (2007) Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS ONE*, **2**, 1–8.

Lowry, C. A., Woodall, W. H., Champ, C. W. and Rigdon, S. E. (1992) A multivariate exponentially weighted moving average control chart. *Technometrics*, **34**, 46–53.

Madanhire, I. and Mbohwa, C. (2016) Application of statistical process control (SPC) in manufacturing industry in a developing Country. *Procedia CIRP*, **40**, 580–583.

Milanzi, E. and Molenberghs, G. (2012) Ignoring overdispersion in hierarchical loglinear models: Possible problems and solutions. *Statistics in medicine.*

Page, E. S. (1954) Continuous inspection schemes. *Biometrika*, **41**, 100–115.

Park, K., Jung, D. and Kim, J. M. (2020) Control charts based on randomized quantile residuals. *Applied Stochastic Models in Business and Industry*, **36**, 716–729.

Park, K., Kim, J. and Jung, D. (2018) GLM-based statistical control r-charts for dispersed count data with multicollinearity between input variables. *Quality and Reliability Engineering International*, **34**, 1103–1109.

Qiu, P. (2014) *Introduction to Statistical Process Control.* Chapman Hall/CRC, Boca Raton.

Roberts, S. W. (1959) Control chart tests based on geometric moving averages. *Technometrics*, **1**, 239–250.

Salvan, A., Sartori, N. and Pace, L. (2020) *Modelli Lineari Generalizzati*. UNITEXT. Springer Milan.

Shen, X., Tsui, K. L., Zou, C. and Woodall, W. H. (2016) Self-starting monitoring scheme for poisson count data with varying population sizes. *Technometrics*, **58**, 460–471.

Sparks, R. S., Keighley, T. and Muscatello, D. (2011) Optimal exponentially weighted moving average (EWMA) plans for detecting seasonal epidemics when faced with non-homogeneous negative binomial counts. *Journal of Applied Statistics*, **38**, 2165–2181.

Urbieta, P., Lee Ho, L. and Alencar, A. (2017) CUSUM and EWMA control charts for negative binomial distribution. *Quality and Reliability Engineering International*, **33**, 793–801.

Wang, Z. and Zwetsloot, I. M. (2023) A transfer learning-based multivariate control chart for dengue surveillance in Hong Kong. *IEEE Access*, **11**, 66415–66427.

Woodall, W. H. and Ncube, M. M. (1985) Multivariate CUSUM quality-control procedures. *Technometrics*, **27**, 285–292.

Zhang, Y., He, Z., Zhang, M. and Wang, Q. (2016) A score-test-based EWMA control chart for detecting prespecified quadratic changes in linear profiles. *Quality and Reliability Engineering International*, **32**, 921–931.