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

Statistica per l'economia e l'impresa

RELAZIONE FINALE

Carta di controllo CUSUM per dati di conteggio spazio-temporale

Relatore: Prof. Guido Masarotto
Dipartimento di Scienze Statistiche

Candidato: Luca Trevisan
Matricola N°1166995

Anno Accademico 2021/22

1 Introduzione

La carta di controllo come mezzo grafico per applicare i principi di significatività statistica al controllo del processo di produzione fu proposta per la prima volta da Walter Shewhart nel 1924. Le carte di controllo sono uno strumento prevalentemente statistico e/o ingegneristico, esse hanno il compito principale di tenere sotto controllo i vari parametri di un processo per vedere se l'andamento, in un determinato momento subisce delle trasformazioni di qualsiasi tipologia e dunque passa ad uno stato di fuori controllo e di conseguenza dannoso per qualsiasi ambito in cui si sta svolgendo la rilevazione.
Possiamo considerare differenti tipi di carta di controllo, come quelle retrospettive o sequenziali, nella prima tipologia viene esaminato il processo dopo che lo stesso è già avvenuto e che quindi analizza se é stato stabile nel periodo considerato oppure ha subito alterazioni e dunque è andato fuori controllo.
In quella sequenziale invece, è presente un analisi sequenziale dei dati, ovvero man mano che essi sono raccolti vengono esaminati.
Lo scopo principale di tale procedimento è l'individuazione di un andamento anomalo dei dati con conseguente attuazione di un intervento nel più breve tempo possibile volto a ridurre il danno che potrebbe susseguirsi a causa dello sviluppo fuori controllo del processo.
In questo lavoro verrà trattata principalmente la carta CUSUM (Cumulative Sum) e alcune varianti della stessa per il monitoraggio e l'identificazione dei cambiamenti spazio-temporali di tipo epidemiologico su vari scenari territoriali.
Nella prima parte della relazione sarà presente una spiegazione degli elementi fondamentali che formano la CUSUM, con una breve panoramica e illustrazione riguardante come la carta analizza sequenzialmente i dati, infine verrà fornito un esempio per il monitoraggio dei dati su una distribuzione Normale.
Successivamente verranno introdotti i metodi di individuazione del cluster, definendo i vari elementi che compongono le tipologie della Carta CUSUM, come, per esempio, il rapporto di verosimiglianza. Dunque si passerà alla descrizione delle varie tipologie che possono formare la carta CUSUM (Circular CUSUM, Nearest-neighbourhood CUSUM, CSTS CUSUM). Nella seconda parte invece, verrà implementata una funzione che prenderà in considerazione una delle varie tipologie della carta CUSUM attraverso il software R con la quale verranno forniti degli esempi per vedere in che modo la carta lavora cambiando i parametri che la compongono. Verranno poi date alcune conclusioni in merito agli argomenti trattati ed anche a riguardo dell'efficacia della funzione sviluppata.

Indice

2 Carte Cusum

Le carte di controllo CUSUM sono state introdotta da Page (1954) e poi modificate e/o avvalorate da altri esperti, tra i quali Woodall e Adams nel 1993. Le CUSUM fanno parte della categoria di carte che analizzano sequenzialmente i dati così da intervenire repentinamente nel caso che il processo di creazione di qualsiasi tipo di prodotto stesse prendendo un andamento non conforme a quello voluto.
Quindi, si deve innanzitutto definire un punto temporale, non noto, dove il processo cambierà andamento passando da "in controllo" a "fuori controllo", verrà indicato questo momento temporale con la lettera τ.
Le osservazioni $y_1, y_2, ...$ dunque hanno due differenti tipi di funzione di densità $f(.)$ (nota) in base al tempo t in cui ci troviamo:

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

Si deve definire inoltre un valore critico L dove, la carta Cusum segnala un allarme quando la statistica di controllo $W_t > L$. W_t è così definita:

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

Mentre per quanto riguarda il valore critico L, esso è una costante positiva.
Di seguito viene proposto un esempio che riguarda l'utilizzo di una carta Cusum per sorvegliare la media di una distribuzione Normale.
Esempio:

Si hanno, come spiegato precedentemente, delle osservazioni indipendenti $x_{t,1}, x_{t,2}, ..., x_{n,t}$, che seguono due differenti tipi di distribuzione Normale:

$$x_{t,1}, x_{t,2}, ..., x_{n,t} \sim \begin{cases} N(\mu_0, \sigma_0^2) & \text{se t}< \tau \\ \\ N(\mu_0 + \delta\sigma_0, \sigma_0^2) & \text{se t}\geq \tau \end{cases}$$

Si consideri inoltre che μ_0 e σ_0 sono noti. Dunque si può notare che nella seconda distribuzione è presente oltre alla media e la varianza anche δ_0 che viene aggiunto alla media, esso rappresenta il "salto" dalla media originaria

del processo che si verifica quando quest'ultimo subisce una trasformazione dovuta ad un particolare evento,e che è noto ma solo in questo esempio, così da rendere più efficiente l'applicazione di quanto detto nei paragrafi precedenti.
Nella realtà non è risaputo di quanto la media possa "saltare" dopo un evento anomalo nel processo che stiamo esaminando. Infine come detto anche precedentemente, τ, che ricordiamo, rappresenta il momento temporale in cui il processo cambia stato, non è possibile conoscerlo e quindi risulterà per forza di cose, ignoto. La statistica W_t dunque assumerà la forma:

$$W_t = \begin{cases} 0 \quad \text{se t= 0} \\ \\ \max\left(0, W_{t-1} + log\frac{f_1(x_{t,1},x_{t,2},...,x_{t,n})}{f_0(x_{t,1},x_{t,2},...,x_{t,n})}\right) \quad \text{se t> 0} \end{cases}$$

Dove:

$$-f_0(x_{t,1}, x_{t,2}, ..., x_{t,n}) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi\sigma_0^2)}} e^{-\frac{1}{2\sigma_0^2}(x_{t,i}-\mu_0)^2}$$

$$-f_1(x_{t,1}, x_{t,2}, ..., x_{t,n}) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi\sigma_0^2)}} e^{-\frac{1}{2\sigma_0^2}(x_{t,i}-\mu_0-\delta_0\sigma_0^2)^2}$$

Dunque lo sviluppo del log-rapporto tra le due funzioni di densità: $log\left(\frac{f_1(x_{t,1},x_{t,2},...,x_{t,n})}{f_0(x_{t,1},x_{t,2},...,x_{t,n})}\right)$ sarà il seguente:

$$log\left(\frac{f_1(x_{t,1}, x_{t,2}, ..., x_{t,n})}{f_0(x_{t,1}, x_{t,2}, ..., x_{t,n})}\right) = \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}[(x_{t,i}-\mu_0)^2 - (x_{t,i}-\mu_0-\delta_0\sigma_0^2)^2] =$$

$$= \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}[2\delta\sigma_0(x_{t,i}-\mu_0) - \delta^2\sigma_0^2)] =$$

$$= \frac{1}{2\sigma_0^2}[2n\delta\sigma_0(x_{t,i}-\mu_0) - n\delta^2\sigma_0^2)] =$$

$$= \sqrt{n}\delta[\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma_0} - \frac{\sqrt{n}\delta}{2}].$$

Per individuare aumenti e/o diminuzioni della media del processo, è possibile inoltre dividere la statistica test W_t in due differenti tipi, distinguendo i due casi:

-Partendo da $W_0^+ = 0$, si calcola sequenzialmente:

$$W_t^+ = \max\left(0, W_{t-1}^+ \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_0} - \kappa\right)$$

E, dato che vogliamo intercettare aumenti della media, tale statistica deve segnalare un allarme quando $W_t^+ > L$, con L, che come definito precedentemente, è una costante positiva.

-Partendo da $W_0^- = 0$, si calcola sequenzialmente:

$$W_t^- = \min\left(0, W_{t-1}^- \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_0} - \kappa\right)$$

Invece in questo caso W_t^- deve "chiamare" un allarme quando $W_t^- < L$. Per semplicità ho posto $\kappa = \frac{\sqrt{n}\delta}{2}$.

2.1 Come si può valutare una carta Cusum?

Le carte di controllo CUSUM vengono costruite scegliendo opportunamente i valori dei parametri, quali K, che contiene δ e L, questo per ottenere migliori prestazioni in termini di ARL, ma cosa s'intende con questo termine?
Prima di definire cosa specifica il termine ARL, è importante spiegare cosa sia la Run Length. Essa è per definizione:

$$RL = \min\left(t : W_t \notin [LCL_t, UCL_t]\right)$$

Si può tradurre questa definizione come: "La Run Length è definita come il numero di istanti temporali dall'inizio del processo (e dunque della sorveglianza) e la segnalazione del primo allarme".
Dove LCL e UCL sono detti limiti di controllo, che invece, nel caso della

Carta Cusum vengono "sostituiti" dalla costante L.
Dunque, in parole povere la Run Length mette in evidenza la tempistica della Carta di Controllo presa in riferimento nella segnalazione di un allarme (giusto o sbagliato che sia).
Paradossalmente ci si auspicherebbe sempre che:

$$Pr(RL = \tau) = 1,$$

Ovvero che:

-La carta di controllo non segnali mai un errore quando il processo è in controllo, (t $< \tau$).

-Segnali subito, ovvero al tempo t $= \tau$, che il processo è andato fuori controllo.

Noteremo subito come questa condizione appena riportata sia praticamente impossibile, poiché si tratta di variabili casuali, dunque, per fare una breve sintesi:

- **Se processo IN CONTROLLO:**

  RL "grande" così da avere tanto tempo tra i falsi allarmi segnalati.

- **Se processo FUORI CONTROLLO:**

  RL "piccola" così che la carta segnali nel più breve tempo possibile la presenza di un qualsiasi fattore che sta facendo subire una trasformazione non desiderata del processo che stiamo esaminando.

Per concludere, ora si può spiegare cosa si intende con il termine ARL, ovvero Average Run Length, dunque la media della Run Length.
Essa esprime "in media" quanto una carta ci metta per segnalare un allarme, ed è sostanzialmente un indice di performance per valutare quale sistema di controllo sia meglio utilizzare.

2.2 Visualizzazione grafica

Di seguito vedremo graficamente, attraverso l'ausilio del codice R, l'esempio descritto in precedenza dove i dati seguivano una distribuzione Normale.

Nota: per l'esempio verrà utilizzato un dataset di nome "flow1.dat" fornito dal prof. Guido Masarotto.

```
x <- rnorm(../Dati/flow1.dat)
mu0 <- mean(x)
sigma0 <- sd(x)
n <- 5
delta <- 0.2/sigma0
k <- sqrt(n)*delta/2
```

In queste prime righe di codice oltre ad aver caricato un dataset si è attuato il calcolo della media e deviazione standard di x.
Inoltre viene definito n, che è il campione di osservazioni estratto al tempo t e delta, tali elementi servono per la creazione di kx che è definito (come detto in precedenza) il "salto" della media dopo il cambiamento di stato del processo.

```
B <- 2000 #Valore dell'Arl che fissiamo
L <- xcusum.crit(k, B, sided="two") #Valore critico L
salto <- c(0.1, 0.2, 0.3)
```

Si può notare che in "salto" vengono racchiusi 3 differenti tipi di valore del salto della media che si vorrebbe identificare.

```
curve(dnorm(x, mu0, sigma0), 0.5, 2.5, lwd=3, col=2)
grid()
curve(dnorm(x, mu0+salto[1], sigma0), 0.5, 2.5, lwd=3, col=3, add=TRUE)
curve(dnorm(x, mu0+salto[2], sigma0), 0.5, 2.5, lwd=3, col=4, add=TRUE)
curve(dnorm(x, mu0+salto[3], sigma0), 0.5, 2.5, lwd=3, col=5, add=TRUE)
abline(v=1, lty="dotted", lwd=3, col="gray")
abline(v=2, lty="dotted", lwd=3, col="gray")
```

Nb: La funzione *xcusum.crit* consente di calcolare il valore critico L garantendo che l'ARL in controllo sia pari a B. Nel caso si voglia solamente utilizzare una delle due statistiche di controllo W_t^+ o W_t^- basta non scrivere "sided = two".
Tale funzione deriva dalla libreria di R "spc".

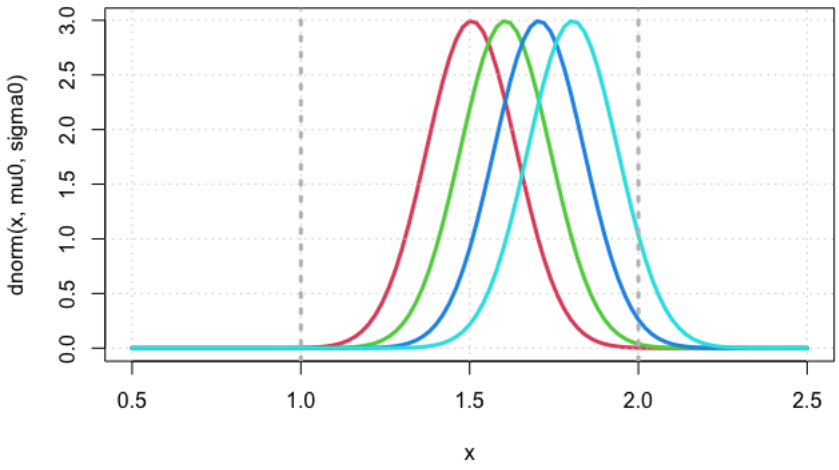


Figura 1: Con il colore rosso viene indicata la distribuzione del processo in controllo, notiamo come entrambe le code stiano all'interno del valore critico L che abbiamo definito e che dunque la proporzione di pezzi difettosi sia pressoché nulla.
Con gli altri colori si possono identificare i 3 differenti stati della distribuzione a seconda dell'ampiezza del salto della media del processo.

3 Metodi di individuazione del cluster

Molti metodi per la prevenzione del quadro epidemiologico sia locale che nazionale sono di tipo retrospettivo, mentre, in questo lavoro verrà esaminato un caso in cui il metodo di analisi del processo è di tipo sequenziale, dunque i dati verranno analizzati in serie.
L'obiettivo è, sostanzialmente, cercare di identificare ed agire nel più breve tempo possibile dopo che il cluster è emerso.
Verrà analizzata una determinata area spaziale chiamata "mappa" e ci sarà un lavoro basato su dati di conteggio regionale, quindi il tipo di clustering che si andrà ad esaminare si discosta dalla casualità spaziale anche se verrà presentata una sezione per l'estensione nel caso di popolazione non omogenea.
Si ponga: $Y = (Y_{(t)}, t = 1, 2, ..)$ e si consideri come il processo sotto sorveglianza e dunque in controllo, dove in qualsiasi momento t, avremo un vettore p-variato , $Y_{(t)} = (Y_1(t), Y_2(t), ..., Y_p(t))^T$, tale vettore contiene il numero di casi per ogni regione della mappa considerata.
In assenza di clustering si può dire che il processo segua una distribuzione Poisson di parametro, che indicheremo con un livello di "intensità", λ_0.
L'emergere del cluster verrà espresso con intensità λ_1 e dunque la proposta di questo metodo che utilizza la carta Cusum è quello di individuare rapidamente questo cambio, ovvero da λ_0 a λ_1.
Verranno indicati i momenti temporali precedenti in cui viene individuato il cambio d'intensità con t, dove: $Y_S = (Y_{(t)}, t \leq s)$.
Verrà utilizzata questa definizione per creare la statistica d'allarme $p(Y_s)$, dunque il sistema di sorveglianza del processo che si sta esaminando farà scattare un allarme quando $p(Y_S)$ andrà al di là del limite di controllo L_S e dunque ci sarà la scoperta del cluster nella mappa che era sotto osservazione.
I metodi di sorveglianza che verranno esaminati si basano tutti sul rapporto di verosimiglianza LRs per un cambio di stato di un processo con distribuzione Poisson, che passa da λ_0^*, al nuovo grado di intensità λ_1^* al tempo $T = \tau$ nell'area A, che è una delle regioni che andremo a controllare.
Possiamo scrivere il rapporto di verosimiglianza (parziale) come:

$$L(s, \tau, \lambda_1^*, \lambda_0^*, A) = \frac{f_{Y\,s}(y_s|T = \tau, \lambda_1^*, A)}{f_{Y\,s}(y_s|T > s, \lambda_0^*)} = \prod_{t=\tau}^{s} \prod_{i \in A} \left(\frac{\lambda_1^*}{\lambda_0^*}\right)^y exp\big(-(\lambda_1^* - \lambda_0^*)\big) \tag{1}$$

Tale rapporto di verosimiglianza può essere usato in differenti modi per co-

struire la statistica d'allarme.
Per l'individuazione di cluster si passerà alla costruzione di metodi statistici basati sia sullo spazio che sul tempo, ecco perché il rapporto di verosimiglianza si basa su due differenti stati del processo. Questo tipo di approccio è motivato anche dal fatto che verranno considerate differenti regioni per la scansione dei vari metodi che portano all'individuazione dei cluster questi ultimi verranno considerati come un aumento di intensità in un area (spaziale) ristretta.
Dunque verrà evidenziato che i metodi di scansione/individuazione spazio temporale si adattano ad una carta Cusum, poi si passerà alla presentazione di 3 tipi di carte, come la "conditional space time-scan" (CSTS), la "nearest-neighbourhood CUSUM" indicata come NN-CUSUM ed anche la "circular CUSUM" nota come C-CUSUM.

3.1 Carta Cusum per individuazione di cluster

Come detto in precedenza, la metodologia che adotta la carta Cusum per la prevenzione della salute pubblica attraverso l'individuazione di cluster è ampiamente utilizzata. Spostando l'attenzione al cambiamento di stato del processo che si vuole analizzare nel caso univariato, in questo caso la statistica d'allarme $p(Y_s)$ che fa sempre riferimento al rapporto di verosimiglianza (LRs) sarà del tipo:

$$p(Y_S) = \max_{[\tau]} L(s, \tau, \lambda_1^*, \lambda_0^*) \tag{2}$$

Si noti che, a differenza della formula scritta in precedenza, manca la componente A poiché qui ci si riferisce al caso univariato. E' possibile riscrivere l'espressione appena usata usando la ricorsività della formula che si basa sulla somma cumulativa del logaritmi del rapporto di verosimiglianza, dunque log-LRs, ovvero $lnL(s, s, \lambda_1^*, \lambda_0^*)$ e quindi p(Y_s) diventerà:

$$p(Y_S) = \max\left(0, p(Y_{S-1}) + lnL(s, s, \lambda_1^*, \lambda_0^*)\right) \tag{3}$$

Un'estensione plausibile della formula (2) e', riportandoci al caso multivariato, l'inclusione della lettera A dove, $A \in \Lambda$ e Λ è un set riguardante uno spazio territoriale con la presenza di cluster.

Dunque l'espressione della statistica d'allarme sarà del tipo:

$$p(Y_S) = \max_{[\tau, A]} \left(L(s, \tau, \lambda_1^*, \lambda_0^*, A) \right) \tag{4}$$

Che può essere riscritta come:

$$p(Y_s) = \max_{[A]}(p_A(Y s)) \tag{5}$$

Dove la formula (5) corrisponde alla specifica statistica d'allarme per il cluster A, con A $\in \Lambda$.
Anche in questo la formula si può riscrivere come fatto precedentemente:

$$p_A(Y_s) = \max \left(0, p_A(Y_{S-1}) + lnL\left(\frac{\lambda_1^*}{\lambda_0^*}\right) \sum_{i \in A} Y_i(S) - r_A(\lambda_1^* - \lambda_0^*) \right) \tag{6}$$

Con il termine r_A viene inteso il numero di regioni che è contenuto in A.
Si può inoltre notare che nell'ultima formulazione è presente anche lo 0 come valore minimo, questo perché per ogni CUSUMs viene assicurato che l'allarme per l'insorgere del cluster non sia troppo lontano dal limite (L) prefissato.

3.2 Scansione spazio temporale condizionata

Per un determinato insieme di punti in una mappa si possono definire dei cerchi di differente grandezza che varia a seconda dell'ampiezza del raggio del cerchio. L'insieme dei cerchi che vengono scansionati sono indicati con Θ mentre uno specifico elemento di Θ viene indicato con Z.
Per i dati di conteggio di tipo regionale, la pratica specifica che viene ampiamente utilizzata consiste nella raccolta di specifici punti centrali delle regioni come possibili punti centrali per i cerchi, se il punto centrale di una regione è incluso in uno dei cerchi di conseguenza viene inclusa l'intera regione.
Cambiando il punto temporale per l'inizio del raggruppamento di tutti i possibili cluster per ogni area circolare della mappa che otteniamo, vediamo che la statistica d'allarme di conseguenza si baserà sul considerare il massimo del LR per un cluster spazio temporale, dove tale cluster partirà al tempo τ e sarà condizionato al totale del numero di eventi che si sono susseguiti durante il periodo [τ,s] e che seguono una distribuzione $N(\tau,\ s)$, dunque il riferimento a questo metodo prenderà il nome di CSTS.
La statistica d'allarme (sotto il modello Poisson) prenderà la forma:

$$p(Y_s) = \max_{[\tau,Z]}[G(s,\tau,Z,N(\tau,s))] \tag{7}$$

dove:

$$\mathrm{G(s,\ \tau,\ Z,\ N(\tau,\ s))} =$$

$$= \left(\frac{\eta_Z(\tau,s)}{\mu_Z(\tau,s)}\right)^{\eta_Z(\tau,s)} \left(\frac{N(\tau,s)-\eta_Z(\tau,s)}{N(\tau,s)-\mu_Z(\tau,s)}\right)^{N(\tau,s)-\eta_Z(\tau,s)} \tag{8}$$

Nella formula (8), μ_Z rappresenta il numero di eventi che ci si aspetta (condizionatamente ad una distribuzione N(τ,s) per uno specifico Z $\in \Theta$ durante il periodo [τ, s] nel caso non vi sia un emergere di cluster. Si noti come la funzione G(s, τ, Z, N(τ, s)) non è altro che il LR massimizzato per un campione di dati fissato dove però, al posto di considerare l'intero dataset, viene considerato il periodo [τ, s].
L'LR massimizzato contiene la stima di massima verosimiglianza del cambio d'intensità ovvero λ_1. Usando il rapporto di verosimiglianza parziale (1) la statistica d'allarme (8) assumerà la forma:

$$p(Y_S) = \max_{[\tau,Z]} \left(\frac{max_{[\lambda_1]} f_{Y_S}(Y_S|T=\tau,\lambda_1,Z,N(\tau,s))}{max_{[\lambda_0]} f_{Y_S}(Y_S|T\leq s,\lambda_0,Z,N(\tau,s))}\right) \tag{9}$$

3.3 "The nearest-neighbourhood" CUSUM

Una strada possibile per poter aggregare le osservazioni è quella di utilizzare la media dei dati di ogni quartiere e dunque di adoperare delle carte CUSUM per ogni quartiere. A questo punto l'allarme s'innescherebbe ogni qualvolta una della Carte CUSUM segni un allarme. Tale procedura può essere vista come un controllo statistico spazio-temporale nel caso in cui ogni carta di ogni quartiere abbia gli stessi limiti di controllo. Ci riferiremo a questo metodo come NN-CUSUM. La differenza sostanziale tra il metodo CSTS (argomentato nel paragrafo precedente) e tale metodo sta nella specificazione dei livelli d'intensità del processo prima e dopo il clustering, poiché nel caso CSTS tali parametri venivano stimati mentre nel caso della NN-CUSUM vengono specificati in anticipo.

3.4 Circular CUSUM

Assembrando i metodi precedenti possiamo creare un nuovo metodo per la carta CUSUM dove i livelli d'intensità del processo saranno conosciuti in anticipo e il metodo di analisi del processo viene eseguito su tutti i cerchi che coprono le regioni come per il metodo usato per la CSTS. Così facendo la statistica d'allarme può essere scritta come nella formula (5) con la sostanziale differenza che $\Lambda = \Theta$. Il riferimento a tale procedura sarà data dal nome "Circular CUSUM" ovvero (C-CUSUM).

3.5 Misure di valutazione delle varie Carte

In una procedura di analisi del processo le decisioni sul come agire vengono riviste svariate volte. Dato che l'andamento del processo può subire vari e continui cambiamenti durante il periodo di sorveglianza, il livello di significatività della statistica test non è una misura di valutazione adeguata. Occorre dunque definire delle misure di valutazione ad hoc per questi tipi di processi. Si deve cercare di creare una sorta di "trade-off" tra falsi allarmi e piccoli ritardi per allarmi motivati.
La distribuzione dei falsi allarmi è spesso segnata dalla media della Run Lenght denominata anche ARL (dove è stata discussa nel paragrafo [2.1]), essa viene così definita: $ARL^0 = E[\tau_A | T = \infty]$, dove t_A rappresenta la prima volta in cui il metodo di sorveglianza segna un allarme. Come detto in precedenza l'abilità del metodo nell'individuazione di un cluster può essere vista dal ritardo con cui viene individuata l'insorgenza. In tale lavoro verrà usato il CED ovvero il "conditional expected delay". La formulazione assunta è la seguente: $CED = E[t_A - \tau | t_A \geq T = \tau]$.

Esso non è altro che il valore atteso tra il momento in cui vi è un cambio d'intensità del processo e l'istante temporale in cui viene segnalato l'allarme.

4 Estensione al caso non omogeneo

Passando ad un caso più reale, possiamo derivare i metodi che compongono la carta C-CUSUM e la NNCUSUM estendendo le loro relative formule (6), che sono basate sulla somma cumulata del log-rapporto di verosimiglianza, andando ad effettuare una modifica fra il caso in cui ci sia la presenza di cluster e il periodo in assenza di clustering. Dunque:

$$lnL\big(s, s, \lambda_1^*(s), \lambda_0^*(s)\big) = \left(\sum_{i \in A} Y_i(s)\, lnL(\frac{\lambda_1^*}{\lambda_0^*}) - \sum_{i \in A} (\lambda_{1i}^*(s) - \lambda_{0i}^*(s)) \right) \quad (10)$$

dove $\lambda_0^*(s) = (\lambda_{01}^*(s), \lambda_{02}^*(s), ..., \lambda_{0p}^*(s))^T$ è un vettore che contiene i valori di riferimento d'intensità per le popolazioni delle varie zone territoriali, mentre λ_1^* è il vettore corrispondente che contiene i valori del cambio d'intensità che fanno aumentare il relativo cluster. Dunque la statistica d'allarme per una specifico cluster A sarà del tipo:

$$p_A(Y_S) = max\left(0, p_A(Y_{S-1}) + \sum_{i \in A} Y_i(s)\, lnL(\frac{\lambda_1^*}{\lambda_0^*}) - \sum_{i \in A} (\lambda_{1i}^*(s) - \lambda_{0i}^*(s))\right) \quad (11)$$

Con $p_A(Y_0) = 0$. Come nell'equazione (5) si avrà che, anche nel caso non omogeneo, la statistica d'allarme per la carta CUSUM sarà del tipo:

$$p(Y_S) = max_{[A]}\ p_A(Y_S) \quad (12)$$

L'importanza nell'analisi dell'estensione nel caso non omogeneo sta nel fatto che fino d'ora si è ipotizzato che la popolazione per ogni zona territoriale rimanesse uguale per ogni valore temporale t e dunque, che essa non crescesse o diminuisse mai, cosa che, è del tutto differente dalla realtà. Si può notare che, per seguire questa non omogeneità e adattare la statistica d'allarme (e non solo) al variare della popolazione, si sono cambiati i parametri che compongono il log-rapporto di verosimiglianza, creando due vettori che rappresentino i valori d'intensità prima e dopo l'emergere del cluster per ogni zona territoriale considerata.

5 I limiti dinamici

Seguendo l'estensione al caso non omogeneo definito nella sezione precedente, i limiti che verranno implementati nella simulazione del metodo C-CUSUM utilizzando il software R, non possono essere calcolati in anticipo, in quanto la distribuzione di p(Y_S) dipende dalla popolazione della zona territoriale che la CCUSUM sta scansionando che, naturalmente, cambierà in base all'istante in cui avviene la scansione. Dunque i limiti andranno calcolati giorno per giorno man mano che le informazioni diventeranno disponibili. Ricordando che la statistica d'allarme che si sta considerando assume la seguente forma:

$$p(Y_S) = max\left(\frac{max_{[\lambda_1]} f_{Y_S}(Y_S|T=\tau,\lambda_1,Z,N(\tau,s))}{max_{[\lambda_0]} f_{Y_S}(Y_S|T\leq s,\lambda_0,Z,N(\tau,s))}\right) \tag{13}$$

Il limite di controllo variabile L_S è una successione che dovrà dipendere dalla popolazione. Una possibilità per calcolare l'ARL che si desidera ottenere consiste nel fissare L_s in modo tale che:

$$Pr(RL > s|RL \geq s) = 1 - \frac{1}{B} \qquad \forall s > 0 \tag{14}$$

Dunque questo ci garantisce che la Run Length sia geometrica con media B. Dove B è il valore dell'ARL in controllo che ci si aspetta di osservare. Quindi notiamo come:

$$\begin{aligned} Pr(RL = s) = Pr(RL > s) * Pr(RL > 2|RL \geq 2) * .. * \\ *Pr(RL > s-1|RL \geq t-1) * [1 - Pr(RL > s|RL \geq s)] = \\ = \left(1 - \frac{1}{B}\right)^{S-1} * \frac{1}{B} \end{aligned} \tag{15}$$

In sostanza, L_S può essere calcolato come il quantile di $1 - \frac{1}{B}$ della distribuzione di $P(Y_s)$ condizionata a $P(Y_s) \leq L_S, ..., P(Y_{S-1}) < L_S$, visto che:

$$Pr(RL > s|RL \geq s) = Pr(P(Y_S) \leq L_S|P(Y_S) \leq L_S, ..., P(Y_S) \leq L_{S-1}). \tag{16}$$

Dunque i limiti che sono appena stati descritti brevemente vengono anche chiamati "false alarm control limits" o "limiti dinamici" e possono essere implementati attraverso simulazione.

6 Implementazione C-CUSUM con R

In questa sezione verranno descritti gli argomenti che compongono la funzione utilizzata per formare il metodo di scansione con la C-CUSUM. Prima di procedere è utile fornire una breve descrizione degli input che verranno passati alla funzione.

- **coord**

  Con il termine "coord" vengono indicate le coordinate delle celle che formano la mappa territoriale che si andrà ad analizzare.

- **raggio**

  Il raggio è l'ampiezza con cui il metodo di scansione analizza le varie aree territoriali.

- **y**

  Con y vengono indicati il numero di casi per ogni area territoriale

- **lambda0 e lambda1**

  Questi due parametri rappresentano i valori d'intensità per ogni singola area, con lambda0 ci si riferisce quando per ogni cella della mappa si ha una situazione in controllo mentre con lambda1 s'intende il cambiamento del processo con un innalzamento del livello d'intensità.

- **arl0**

  Con "arl0" viene inteso il valore dell'Arl in controllo, ovvero il numero di falsi allarmi che si è disposti a tollerare.

- **nsim**

  Nsim rappresenta il numero di traiettorie simulate in controllo che sono state usate per calcolare i limiti dinamici.

Dunque ora si può procedere all'introduzione del codice che forma la funzione.

```
ccusum <- function(coord, raggio, y, n, lambda0, lambda1, arl0, plot = TRUE,
                   nsim = 50 * arl0) {
    ## nn contiene i cluster circolari da usare nel monitoraggio
    ## ncl il loro numero
    ncl <- 0
    nn <- list()
    ## d è la matrice delle distanze tra le varie celle
```

```
d <- as.matrix(dist(coord))
## ricerca di tutti i possibili cluster
npoints <- NROW(d)
for (i in seq.int(npoints)) {
    di <- d[, i]
    for (j in which(di <= raggio)) {
        ## v è un cluster circolare con centro nella cella i-sima
        ## e raggio inferiore o uguale a quello desiderato
        v <- which(di <= di[j])
        ## Se non è già presente, viene incluso nella lista
        ## dei cluster da considerare durante il monitoraggio
        if ((ncl == 0) || all(sapply(nn, function(u) !identical(u, v)))) {
            ncl <- ncl + 1
            nn[[ncl]] <- v
        }
    }
}
## Log verosimiglianza per ogni cluster
clver <- function(y, n) {
    y <- sapply(nn, function(i) sum(y[i]))
    n <- sapply(nn, function(i) sum(n[i]))
    dpois(y, lambda1 * n, log = TRUE) - dpois(y, lambda0 * n, log = TRUE)
}
## Cusum per ogni cluster calcolati dai dati
W <- numeric(ncl)
## Cusum simulati in controllo
star <- seq.int(nsim)
Wstar <- matrix(0, ncl, nsim)
## Cusum e limiti memorizzati e da ritornare
T <- NROW(y)
Ws <- matrix(NA, T, ncl)
Wm <- Ls <- numeric(T)
## ciclo principale
for (i in seq.int(T)) {
    ## Aggiornamento cusum dati
    ni <- n[i, ]
    Ws[i, ] <- W <- pmax(0, W + clver(y[i, ], ni))
    Wm[i] <- max(W)
    ## Aggiornamento cusum simulati in controllo
    for (j in star) {
        Wstar[, j] <- pmax(0, Wstar[, j] + clver(rpois(npoints,
```

```
            lambda0 * ni), ni))
        }
        ## Implementazione limiti dinamici
        Wmm <- apply(Wstar, 2, max)
        Ls[i] <- L <- quantile(Wmm, 1 - 1 / arl0)
        ## Sostituzione traiettorie fuori dai limiti
        idx <- which(Wmm > L)
        Wstar[, idx] <- Wstar[, sample(which(Wmm <= L), length(idx))]
    }
    if (plot) {
        matplot(cbind(Wm, Ls),
            type = c("h", "l"), lty = c("solid", "dotted"), xlab = "t",
            ylab = "CCUSUM"
        )
    }
    invisible(list(nn = nn, Ws = Ws, Wm = Wm, Ls = Ls))
}
```

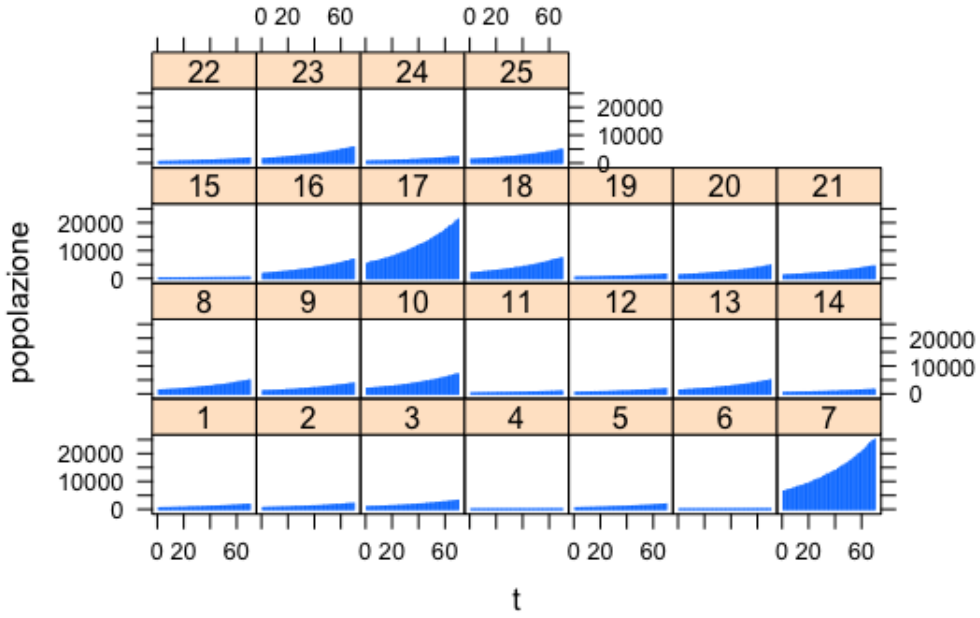
7 Esempi con C-CUSUM

In questa sezione verranno forniti degli esempi per mostrare in che modo la funzione, implementata e descritta nel capitolo precedente, opera con vari cambi d'intensità del processo una volta che esso va fuori controllo ma in un primo esempio questo succede in varie aree territoriali mentre nel secondo tale scenario si verifica solamente in una cella e dunque in una sola zona.

7.1 Esempio con aumento d'intensità di più cluster

```
## Per rendere riproducibile la simulazione
set.seed(12345)

## Una situazione ipotetica griglia nxn
n <- 5
## I punti centrali
x <- rep(1:n, n) - 0.5
y <- rep(1:n, rep(n, n)) - 0.5
## La griglia e numero delle celle
plot(0:n, 0:n, type = "n", xlab = "", ylab = "")
for (i in 0:n) {
    abline(h = i, col = "gray", lty = "dotted")
    abline(v = i, col = "gray", lty = "dotted")
}
for (i in seq_along(x)) text(x[i], y[i], i)

## Istanti di tempo che consideriamo
T <- 70
tm <- seq.int(T)

## Numero di celle
n2 <- n * n

## La popolazione nelle varie regioni è Poisson di media gamma[t,i]
## dove gamma parte da
gamma <- rexp(n2, 1 / 1000)
## E cresce in ogni regione del 2% ogni periodo
pop <- matrix(NA, T, n2)
for (i in tm) {
    pop[i, ] <- rpois(n2, gamma)
    gamma <- 1.02 * gamma
```

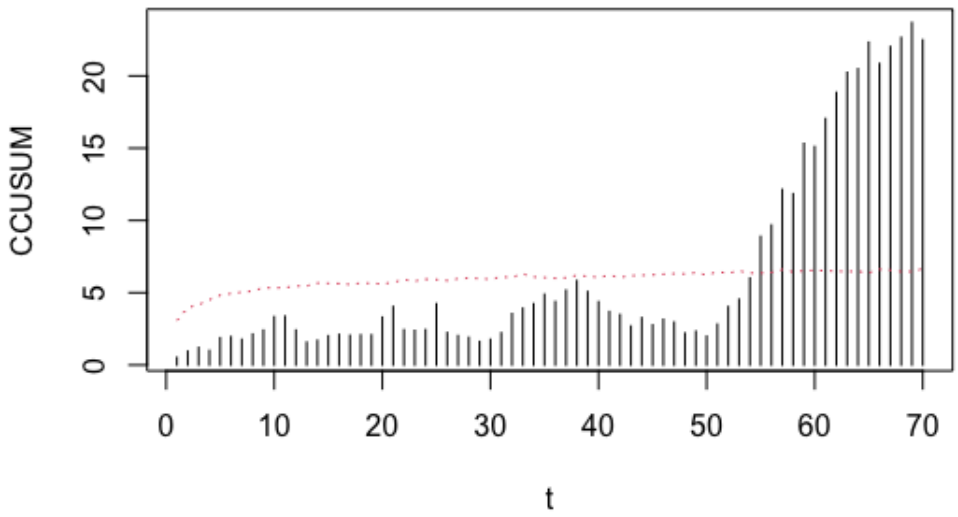
```
}
lattice::xyplot(pop ~ rep(tm, n2) | factor(col(pop)),
type = "h", xlab = "t", ylab = "popolazione")
```



Il grafico che è stato generato rappresenta la popolazione presente per ogni zona della mappa che abbiamo creato e quanto essa cambi nel tempo. Si può notare come il cambiamento sia più marcato in alcune zone rispetto ad altre, ovvero, in termini reali, possono essere presenti dei territori in cui non vi è un grande variazione di persone mentre altri in cui il cambiamento e l'aumento è molto più marcato con il passare del tempo, come per esempio in grandi metropoli o comunque in città con densità di popolazione elevata.

```
## Primo scenario fuori controllo:
## Intesità in controllo
lambda0 <- 2 / 1000
## A partire dal tempo
tau <- 51
## C'è un aumento del 20% dell'intesità
lambda1 <- 1.2 * lambda0
## Nelle celle
fc <- c(3, 7, 8, 9, 13)
```

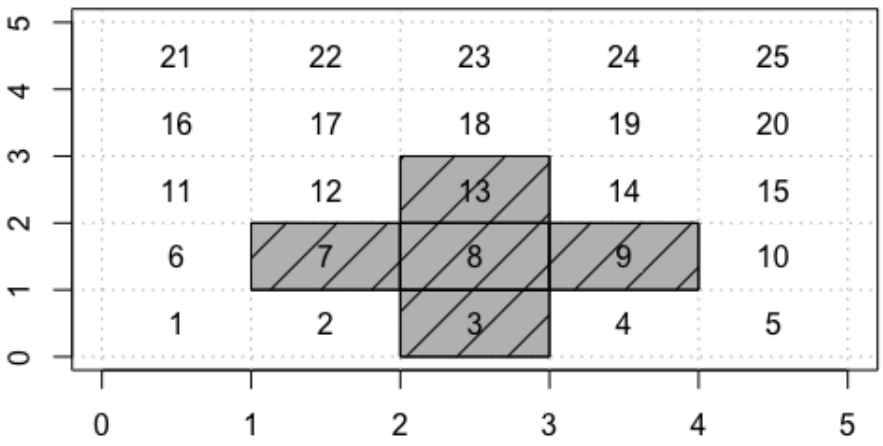
```
## Dati
Lambda <- matrix(lambda0, T, n2)
Lambda[seq.int(tau, T), fc] <- lambda1
casi <- matrix(rpois(n2 * T, Lambda * pop), T)

lattice::xyplot(casi ~ rep(tm, n2) | factor(col(casi)),
type = "h", xlab = "t")
u <- casi / pop Incidenza
lattice::xyplot(u ~ rep(tm, n2) | factor(col(casi)),
type = "h", xlab = "t", ylab = "incidenza")
```



Il grafico mostra l'evoluzione dei limiti dinamici e come essi cambino nel tempo, si può notare come ci sia un superamento del limite di controllo al tempo 55.

```
## Quando è andato fuori controllo?
rl <- min(which(r$Wm > r$Ls))
rl
[1] 55 #Come evidenzia il grafico, al tempo 55 c'è un aumento
```

```
## Quale cluster è responsabile dell'allarme
a <- r$nn[[which.max(r$Ws[rl, ])]]
a
3  7  8  9 13
3  7  8  9 13

## Trattegiate le celle veramente fuori controllo in grigio il cluster
## responsabile dell'allarme
plot(0:n, 0:n, type = "n", xlab = "", ylab = "")
for (i in 0:n) {
    abline(h = i, col = "gray", lty = "dotted")
    abline(v = i, col = "gray", lty = "dotted")
}
for (i in a) {
    rect(x[i] - 0.5, y[i] - 0.5, x[i] + 0.5, y[i] + 0.5, col = "gray")
}
for (i in fc) {
    rect(x[i] - 0.5, y[i] - 0.5, x[i] + 0.5, y[i] + 0.5, density = 5)
}
for (i in seq_along(x)) text(x[i], y[i], i)
```



Il grafico rappresenta la mappa creata in precedenza, dove la parte tratteggiata rappresenta le aree territoriali (definite dalle celle) che sono andate

realmente fuori controllo, dunque dove c'è stato realmente un forte aumento dei cluster nella simulazione che abbiamo effettuato e,invece, le celle grigie rappresentano il cluster responsabile dell'allarme (che era stato definito sopra) ma che non era detto fosse lo stesso definito all'inizio.

```
## Numero di cluster e relativi cusum considerati
length(r$nn$)
## Tutti i cluster che comprendono la cella 8
u <- list()
for (i in seq_along(r$nn)) {
    z <- r$nn[[i]]
    if (any(z == 8)) {
          u[[length(u) + 1]] <- z
      }
}
length(u)
[1] 83
```

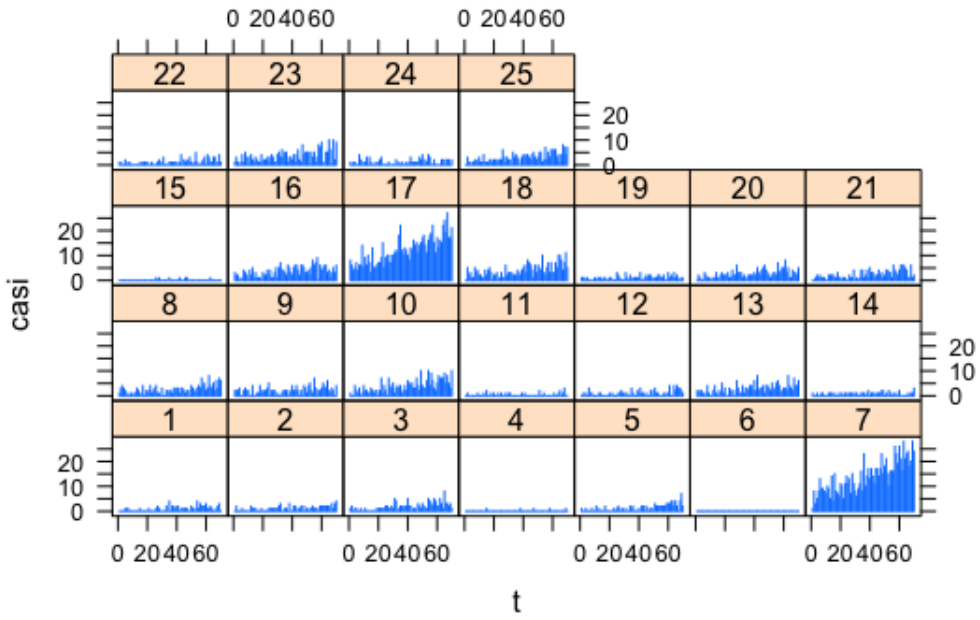
Questo valore "83" sta ad indicare che ci sono state 83 combinazioni di aree territoriali che hanno interessato la zona sulla mappa contrassegnata con il numero 8. Dunque concludendo questo primo esempio può far notare come la carta di controllo riesca ad individuare perfettamente il cluster che abbiamo impostato.

7.2 Esempio con aumento d'intensità di un solo cluster

In questo esempio, come descritto in precedenza sarà presente un cambio d'intensità solamente in una singola area della mappa.

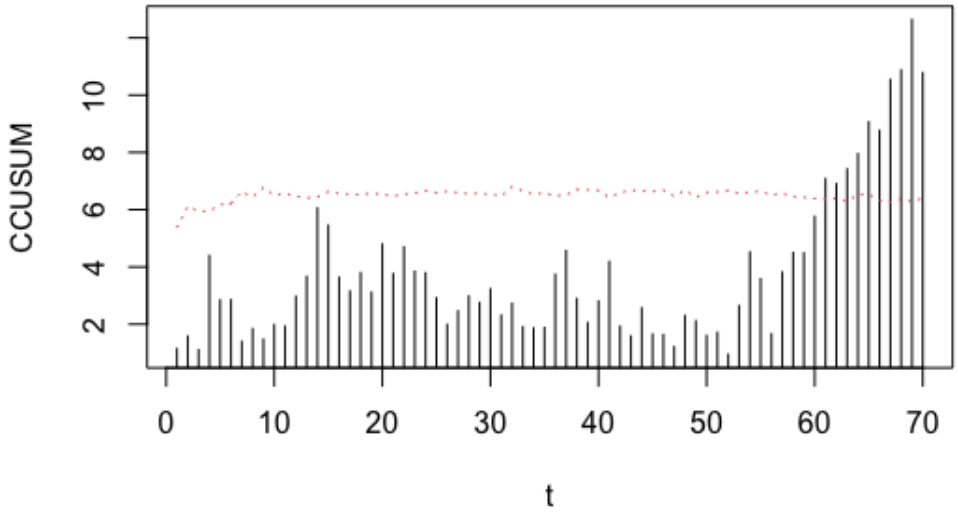
```
## Per rendere riproducibile la simulazione
set.seed(12345)
## Secondo scenario fuori controllo:
## L'intensità in controllo è
lambda0 <- 1 / 1000
## E a partire da
tau <- 51
## L'intensità raddoppia
lambda1 <- 2 * lambda0
## Ma solo nella cella
fc <- 5
```

```r
## Dati
Lambda <- matrix(lambda0, T, n2)
Lambda[seq.int(tau, T), fc] <- lambda1
casi <- matrix(rpois(n2 * T, Lambda * pop), T)

lattice::xyplot(casi ~ rep(tm, n2) | factor(col(casi)),
type = "h", xlab = "t")
u <- casi / pop
lattice::xyplot(u ~ rep(tm, n2) | factor(col(casi)),
type = "h", xlab = "t", ylab = "incidenza")
```



Tale grafico mostra l'andamento dei casi nel tempo per le varie zone territoriali, si può notare come vi siano aree in cui l'aumento è molto marcato, come per esempio la zona 17 o la 7, ed altre in cui l'aumento è pressoché nullo, come per esempio nell'area 4.

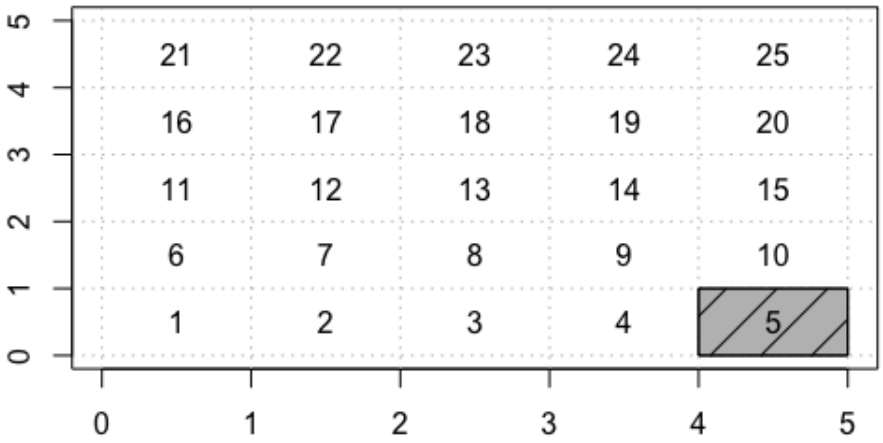
```
##Questo comando serve per mostrare quanto tempo impiega la macchina per
## calcolare i limiti dinamici, produrre un grafico e
## calcolare il resto dei dati
system.time(r <- ccusum(cbind(x, y), 3, casi, pop, lambda0, lambda1, 100))
```



Andando a guardare il grafico relativo ai limiti dinamici si può notare come ci sia un superamento del limite di controllo al tempo 61 e come l'andamento del processo aumenti in maniera sempre maggiore dal tempo 61.

```
## Quando è andato fuori controllo?
rl <- min(which(r$Wm > r$Ls))
rl
[1] 61
## Quale cluster è responsabile dell'allarme
a <- r$nn[[which.max(r$Ws[rl, ])]]
a
5
5
```

```
## Trattegiate le celle veramente fuori controllo in grigio il cluster
## responsabile dell'allarme
plot(0:n, 0:n, type = "n", xlab = "", ylab = "")
for (i in 0:n) {
    abline(h = i, col = "gray", lty = "dotted")
    abline(v = i, col = "gray", lty = "dotted")
}
for (i in a) {
    rect(x[i] - 0.5, y[i] - 0.5, x[i] + 0.5, y[i] + 0.5, col = "gray")
}
for (i in fc) {
    rect(x[i] - 0.5, y[i] - 0.5, x[i] + 0.5, y[i] + 0.5, density = 5)
}
for (i in seq_along(x)) text(x[i], y[i], i)
```



La griglia evidenzia come anche in questo caso la CCUSUM abbia individuato perfettamente l'area territoriale in cui il cluster è aumentato d'intensità nel tempo.

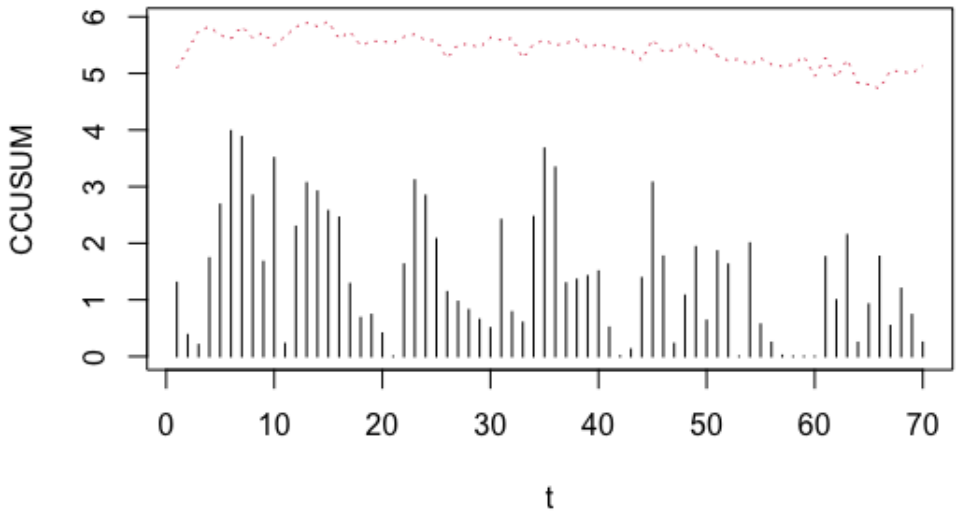
```
## Numero di cluster e relativi cusum considerati
length(r$nn)
[1] 170
## Tutti i cluster che comprendono la cella 5
u <- list()
for (i in seq_along(r$nn)) {
    z <- r$nn[[i]]
    if (any(z == 5)) {
          u[[length(u) + 1]] <- z
      }
}
length($u)
[1] 1750
```

In questo caso sono presenti 1750 combinazioni territoriali che comprendono l'area 5. Un numero significativamente maggiore rispetto a quello riscontrato nell'esempio precedente. Questo aumento può essere dipeso dal fatto che in questo caso si è attuato un aumento d'intensità del cluster relativo ad una sola zona territoriale e non ad un gruppo di più zone limitrofe come nel caso precedente.

7.3 Esempio in cui la Carta non individua i cluster

In tale esempio si andrà ad osservare come la carta non riuscirà ad analizzare i cluster corretti, ovvero quelli che sono aumentati d'intensità e che dunque hanno fatto scattare l'allarme.

```
system.time(r <- ccusum(cbind(x, y), 3, casi, pop, lambda0, 3 *lambda1, 100))
```



Dal grafico sopra riportato si può notare come i limiti dinamici siano molto alti rispetto i differenti valori che assume la statistica d'allarme e dunque si vede chiaramente come la carta non riesca a cogliere variazioni del processo e segnalare allarmi ad un certo tempo t. Quindi, dato che non si riesce ad effettuare nessun tipo di scansione, non si può procedere con l'individuazione di quale (singolo o gruppo) di cluster siano aumentati d'intensità.

8 Conclusioni

In tale lavoro sono stati introdotti in una prima fase i vari elementi che compongono una carta di controllo CUSUM, fornendo anche un esempio. Nelle sezioni successive si è evidenziato come le carte carte di controllo di tipo CUSUM lavorino con i dati di tipo spazio-temporale, si sono analizzati vari aspetti, in particolare come siano presenti diversi tipi della CUSUM solamente cambiando l'area e il modo in cui viene effettuata la scansione. Sono state dunque introdotte la CCUSUM, la NN-CUSUM e la CSTS CUSUM, descrivendone la struttura e il metodo di scansione. Una particolarità che è stata analizzata è l'estensione di tali carte al caso non-omogeneo, ovvero nel caso in cui la popolazione di ogni area della mappa crescesse in maniera assestante, questa estensione è molto utile in quanto rappresenta la realtà, poiché ogni area territoriale ha una popolazione differente che cambia nel tempo ed è soggetta a variazioni.
Ma come si può ampliare quest'analisi? Un primo passo da compiere è quello di analizzare più situazioni reali tramite simulazioni effettuate attraverso R, in questo modo è possibile visualizzare più scenari in cui la carta lavora e cercare di migliorare alcuni aspetti della stessa. Infatti negli esempi che sono stati implementati si è potuto notare come, nel caso in cui λ_1 venga definito molto più grande, la carta faccia fatica a trovare il cluster che è stato effettivamente responsabile del fuori controllo, questo perché, aumentando in maniera considerevole il valore del cambio d'intensità relativo al cluster responsabile dell'innalzamento dei valori, i limiti di controllo dinamici aumentano in maniera di molto superiore ai valori che le statistiche d'allarme assumono e quindi non c'è un superamento della soglia da parte di $P(Y_S)$. Dunque in conclusione, si può affermare che la tipologia di carta presa in considerazione lavori in maniera quasi ottimale per cambiamenti d'intensità bassi, in quanto si è dimostrato nei due esempi come essa riesca ad individuare un gruppo di cluster che hanno avuto un aumento d'intensità e come abbia un'ottima performance anche nel caso di cambio d'intensità di un singolo cluster. Mentre si è osservato che per grandi cambi d'intensità, a prescindere dalla numerosità dei cluster, la carta abbia molte difficoltà e quindi non riesca ad effettuare una scansione che porti a risultati soddisfacenti.

9 Bibliografia

1 **Christian Sonesson**, "A CUSUM framework for detection of space–time disease clusters using scan statistics", Statistic in medicine, Statist. Med. 2007; 26:4770–4789.

2 **O. Arda Vanli, Nour Alawad**, "Space-time surveillance of count data subject to linear trends", Qual Reliab Engng Int. 2021;37:145–164.

3 **Xiaobei Shen, Changliang Zou, Wei Jiang, Fugee Tsung**, "Monitoring Poisson Count Data with Probability Control Limits when Sample Sizes are Time Varying", Wiley Periodicals, Inc.

4 **Guido Masarotto**, materiale fornito per lezioni e laboratori durante l'anno accademico 2020/21.

5 **en.Wikipedia.org**, "CUSUM".