Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per l'Economia e l'Impresa

Relazione Finale

Modelli zero-inflated per la sorveglianza di processi ad alto rendimento

Relatore: Prof.re Guido Masarotto
Dipartimento di Scienze Statistiche

Laureando: Remo Marconzini
Matricola: 1149256

Anno Accademico 2019/2020

Indice

Introduzione

Il continuo e intenso progresso tecnologico a cui le imprese sono costantemente sottoposte, ha portato ad avere dei processi produttivi molto ben organizzati per produrre beni di alta qualità; di conseguenza, la sorveglianza dei processi e i controlli di qualità dei prodotti sono diventati un compito molto complesso per i responsabili della qualità. Da questi processi altamente efficienti, i beni prodotti sono per lo più a zero difetti e quindi spesso modellati usando opportune distribuzioni *zero-inflated*.
Considerato che questi particolari processi non possono essere modellati mediante le usuali distribuzioni di Poisson o Binomiale negativa, l'obiettivo di questo studio è quello di introdurre una metodologia ai fini di una efficiente sorveglianza statistica.
Il lavoro è suddiviso in quattro parti: la prima ha lo scopo di introdurre il concetto di qualità, le sue origini ed evoluzioni, gli aspetti importanti da tenere in considerazione nella progettazione di un valido schema di sorveglianza e infine le fondamenta su cui si basano le più note carte di controllo; nella seconda e terza parte si approfondiranno le metodologie per la modellazione di questi tipi di dati e si tratteranno due diverse soluzioni presenti in letteratura: (i) uno schema basato sull'utilizzo di una carta di controllo di Shewhart proposto da *Mahmood*[8] (ii) uno schema basato sull'utilizzo di una carta di controllo EWMA proposto da *Fatahi et al*[3].
Nell'ultima parte si avanzerà una proposta basata sull'utilizzo di una carta di controllo EWMA con limiti di controllo dinamici, seguita da una applicazione su di un dataset riferito ad un processo di verniciatura ad alta precisione.

Capitolo 1

Il controllo statistico della qualità

In questo capitolo presenteremo e approfondiremo il concetto di qualità, le origini e i principali strumenti del controllo statistico della qualità.

1.1 Concetto di qualità

Oggi giorno il termine qualità può essere definito in diversi modi. La qualità non risiede solamente in alcune caratteristiche che un prodotto o servizio deve avere, ma risulta essere un utile punto di partenza per definire i concetti di "qualità" e "miglioramento della qualità". Nel corso degli anni, la qualità è diventata uno dei fattori determinanti nel processo di acquisto del consumatore. Indipendentemente dal fatto che si tratti di un individuo o di una organizzazione (impresa o ente che sia) il processo di acquisto risulta essere un elemento molto complesso da definire e analizzare. Di conseguenza, di fondamentale importanza è comprendere che vi è un sostanziale guadagno in termini di vantaggio competitivo nell'investire in progetti di miglioramento della qualità

La qualità di un prodotto o servizio è noto essere un costrutto mutidimensionale; Garvin[5] nel 1987 propose un elenco di otto possibili componenti della

qualità, quali: (i) Prestazione (ii) Affidabilità (iii) Durata (iv) Manutenibilità (v) Aspetti formali (vi) Funzionalità (vii) Livello di qualità percepito (viii) Conformità alle normative.
Una prima, tradizionale, definizione del termine "qualità" si basa sul presupposto che beni e servizi devono soddisfare le richieste di coloro che le utilizzano, quindi qualità significa "essere appropriata all'uso". Ci sono due aspetti che caratterizzano il concetto di "essere appropriato all'uso": qualità di progetto e conformità alle normative.
Si è riscontrato che tale definizione risulta associata maggiormente agli aspetti di conformità del prodotto rispetto a quelli relativi alla qualità di progetto. Si preferisce quindi una più recente definizione: "La qualità è inversamente proporzionale rispetto alla variabilità". Tale definizione implica che se si diminuisce la variabilità all'interno del processo di produzione di un bene o l'erogazione di un servizio aumenta la qualità del prodotto/servizio stesso. Questo deriva dal fatto che la qualità desiderata deve essere riproducibile. Poiché la variabilità può essere espressa solo in termini statistici, i metodi e gli strumenti statistici risultano quindi essere fondamentali nei progetti di miglioramento della qualità.

1.2 Origini ed evoluzioni

I metodi statistici applicati al miglioramento della qualità affondano le proprie radici negli studi di W. A. Shewhart con la pubblicazione del testo *Economici Control of Quality of Manufactured Product* (1931) ed ebbero un grande sviluppo tra le due guerre mondiali.
Fino agli anni Novanta l'attenzione nei confronti della "qualità" risiedeva principalmente nella capacità di un bene prodotto (o servizio erogato) di soddisfare determinate prestazioni o requisiti, definite a priori, misurate sulla base del tasso di difettosità di un bene, o in riferimento alla riproducibilità della produzione in un determinato periodo di tempo. A questa visione è possibile collegare strumenti statistici quali: controlli campionari in fase di

accettazione delle materie prime, verifica del numero di non conformità del prodotto in fase di produzione ecc... e tramite il contributo della statistica descrittiva e inferenziale ha dato origine a quell'insieme di metodi e tecniche denominato *Controllo statistico della qualità.*
La "qualità" è così diventata punto cardine di tutta l'attività aziendale e congloba sia le scelte di tipo tecnico sia quelle economico-finanziare. Infine, dagli anni Novanta in poi, è fortemente aumentata l'attenzione per la certificazione della "qualità" a tutela sia del produttore che del consumatore, attraverso le certificazioni ISO 9000 e ISO 9001.
I metodi statistici per il miglioramento della qualità si compongono di: (i) Analisi della capacità di un processo produttivo (Capability), (ii) Sorveglianza statistica di processo (SPM) (iii) Disegno degli esperimenti (DOE).

1.3 Teoria e metodi del controllo statistico di processo

Ai fini di ottenere un prodotto che possa soddisfare le esigenze del consumatore questo deve essere il risultato di un processo produttivo stabile e ripetibile.
L'SPM può essere applicata a qualsiasi processo produttivo o erogazione di un servizio. A tal scopo, Deming E. (1982) propose i sette più importanti strumenti statistici di cui l'SPM si avvale: (i) Istogrammi e grafici "rami e foglie" (ii) Fogli di controllo (iii) Grafici di Pareto (iv) Diagrammi causa ed effetto (v) Diagrammi sulla concentrazione dei difetti (vi) Grafici a dispersione (vii) Carte di controllo.
Al fine di comprendere i principi statistici su cui si basa l'SPM è fondamentale introdurre le fonti di variabilità che intervengono in un processo produttivo. Fondamentalmente, in un processo produttivo intervengono due fonti di variabilità: (i) variabilità provocata da fattori causali (o cause comuni di variabilità) (ii) variabilità provocata da fattori specifici (o cause speciali di variabilità). La prima è il risultato dell'effetto di molti piccoli fattori co-

stanti e casuali attribuibili alla variabilità intrinseca del processo e quindi ineliminabili; un processo affetto solo da questa fonte di variabilità verrà considerato in controllo. La seconda è dovuta principalmente a quattro fattori: (i) macchinari non ben funzionanti o tarati (ii) errori dovuti agli operatori (iii) particolari condizioni ambientali (iv) materiali grezzi difettosi. La variabilità prodotta da questi fattori risulta essere molto più evidente di quella naturale e può dar luogo ad una frazione di prodotti non conformi rilevante. Quando in un processo intervengono cause speciali di variabilità, il processo verrà considerato fuori controllo.
L'obiettivo cardine del controllo statistico di processo è quello di individuare il più rapidamente possibile il verificarsi di cause speciali di variabilità. Le carte di controllo introdotte da W. A. Shewhart, che di seguito introdurremmo, sono uno strumento ampiamente usato per questo scopo.

1.4 Fondamenti statistisci delle carte di controllo di Shewhart

Le carte di controllo sono uno strumento utile e assai utilizzato per monitorare una certa caratteristica di qualità che può essere sia unidimensionale che multidimensionale.
L'uso delle carte di controllo coinvolge due fasi distinte, caratterizzate da due diversi obiettivi. Nella fase I viene eseguita una analisi retrospettiva dei dati del processo, ai fine di stabilire se possano o meno rappresentare uno stato in controllo, oppure per individuare eventuali cause speciali di variabilità intervenute nel processo. Le carte di controllo di fase I servono inoltre per valutare e modellare la variabilità naturale ai fini della stima dei parametri del processo quando opera in controllo.
Nella fase II i dati vengono analizzati sequenzialmente (sorveglianza *on-line*) sfruttando le stime dei parametri derivanti dalla fase I (media e variabilità in controllo, ad esempio) con lo scopo di individuare il più rapidamente possibili deviazioni da uno stato soddisfacente.

Carte di controllo di fase I

Le carte di controllo di fase I sono uno strumento grafico per rispondere alla domanda, **retrospettiva**: Il processo à stato stabile (in controllo) durante il periodo considerato? In sostanza combinano un test statistico per verificare il seguente sistema di ipotesi:

$$\begin{cases} \text{H0 : Nel periodo considerato il processo è stato in controllo} \\ \text{H1 : Nel periodo considerato il processo era fuori controllo} \end{cases}$$

con delle indicazioni, utili per capire cosa potrebbe essere successo, quando H1 viene rifiutata.

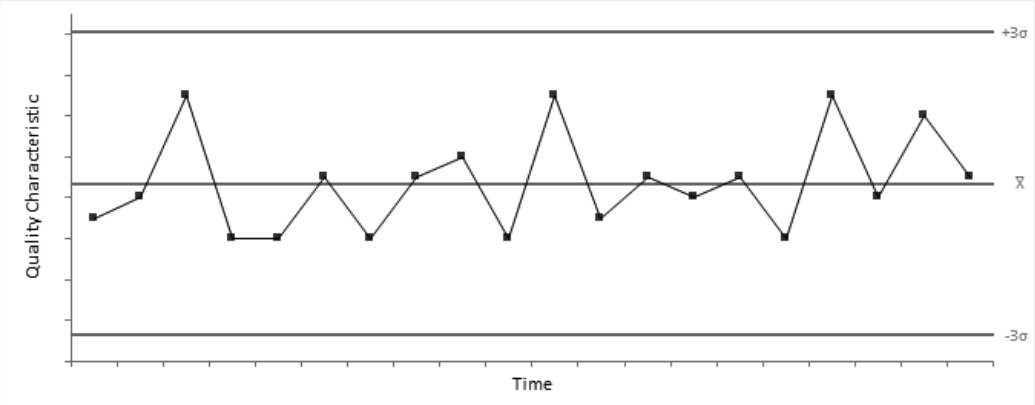

Figura 1.1: Tipica carta di controllo

Formalizzando, i dati disponibili sono:

Osservazioni	1	2	...	j	...	n
Sottogruppo 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,j}$	...	$x_{1,n}$
Sottogruppo 2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,j}$	...	$x_{2,n}$
...	...	...	...	...	...	...
Sottogruppo t	$x_{t,1}$	$x_{t,2}$	...	$x_{t,j}$	...	$x_{t,n}$
...	...	...	...	...	...	...
Sottogruppo m	$x_{m,1}$	$x_{m,2}$	...	$x_{m,j}$	...	$x_{m,n}$

$x_{t,j}$ indica la misura j-esima di una certa caratteristica di qualità (consideriamo il caso in cui sia univariata, senza perdita di generalità n.d.r.) appartenente al sottogruppo t-esimo. Assumiamo che tutte le osservazioni siano indipendenti e che le osservazioni appartenenti allo stesso sottogruppo abbiano la stessa distribuzione di probabilità. Nel caso in cui n sia uguale a uno si parla di misure individuali e non di sottogruppi razionali.
Si usano quindi le osservazioni di ogni sottogruppo per stimare un parametro di una certa caratteristica di qualità che vogliamo sorvegliare, ottenendo cosi W_t statistiche di controllo:

$$x_{t,1}, x_{t,2}, \ldots, x_{t,n} \quad \Longrightarrow \quad W_t = g([x_{t,1}, x_{t,2}, \ldots, x_{t,n}]) \quad \text{per t=1,. . .,m}$$

W_t viene calcolata per ogni sottogruppo e rappresentata graficamente verso il tempo t assieme a dei limiti di controllo: **UCL** (upper control limit), **LCL** (lower control limits) e una linea centrale **CL** (central line). Questi limiti di controllo vengono scelti in modo tale che se il processo è in controllo quasi tutti i valori di W_t cadranno al loro interno; valori di W_t al di fuori dei limiti di controllo sono interpretati come possibili valori affetti da cause speciali di variabilità intervenute nel processo.
La rappresentazione grafica nella sorveglianza statistica di processo ricopre un ruolo fondamentale in quanto permette di verificare tempestivamente se e dove le statistiche di controllo W_t fuoriescono dai limiti di controllo, ma permettono anche di individuare l'eventuale presenza di andamenti sospetti (valori sistematicamente sopra o sotto la *central line*, i.e) che possono indicare la presenza di cause speciali.
Quindi per costruire una carta di controllo dobbiamo:

- Scegliere una opportuna statistica di controllo W_t (dipenderà da ciò che si vuole sorvegliare)
- Determinare degli appropriati limiti di controllo UCL, LCL.

Possiamo notare da quanto detto finora che esiste uno stretto legame tra carte di controllo e test d'ipotesi. Un valore di W_t all'interno dei limiti di

controllo è equivalente al non rifiuto dell'ipotesi che il processo ha operato in controllo mentre un valore di W_t al di fuori dei limiti di controllo è equivalente al rifiuto dell'ipotesi di processo in controllo.
Sebbene esistano comunque delle differenze tra carte di controllo e verifica d'ipotesi la teoria relativa ai test d'ipotesi può comunque risultare utile per valutare **l'efficacia** di una carta di controllo. Consideriamo la probabilità di **falso allarme**, o errore di I° tipo (affermare che il processo è fuori controllo quando in realtà è in controllo) e quella di errore di II° tipo (affermare che il processo è in controllo quando in realtà è fuori controllo).
Valutando l'andamento della probabilità di II° tipo per diversi cambiamenti del parametro di interesse, possiamo avere una indicazione della sensibilità della carta di controllo a individuare più o meno velocemente tali cambiamenti.
La probabilità di falsi allarmi è spesso denominata con l'acronimo *F.A.P.* (fals alarm probability); nell'ambito del controllo statistico della qualità i falsi allarmi sono considerati pericolosi e molto costosi. La condizione quindi che vogliamo imporre nel progettare una carta di controllo è:

$$F.A.P. = Pr_{I.C}(\text{falso allarme}) = 1 - \alpha \tag{1.1}$$

dove $0 < \alpha < 1$ è fissato sulla base dell'applicazione.
I limiti di controllo per una generica carta di Shewhart saranno della forma:

$$\begin{aligned} UCL &= W_t + L\sigma_{W_t} \\ CL &= \mu_{W_t} \\ LCL &= W_t - L\sigma_{W_t} \end{aligned} \tag{1.2}$$

dove σ_{W_t} è la deviazione standard della statistica di controllo W_t e L è un appropriato valore critico che determina l'ampiezza dei limiti di controllo calcolato sulla base di una *F.A.P.* opportuna.
In letteratura vengono proposte diverse soluzioni per il calcolo del valore critico L; la soluzione più generale consiste nel calcolare L via simulazione.
Le carte di controllo di fase I vengono quindi condotte per processi "nuovi" e

spesso ripetute periodicamente ai fini di descrivere, ed eventualmente aggiornare (a causa del deterioramento dei macchinari che compongono il processo produttivo, ad esempio), cosa ci si aspetta "in controllo".

Carte di controllo di fase II

Nelle carte di controllo di fase II i dati vengono analizzati sequenzialmente (man mano che sono raccolti). Lo scopo è quello di costruire uno schema di sorveglianza che ci permetta di identificare rapidamente possibili deviazioni da uno stato soddisfacente.
Le carte di controllo di Shewhart di fase II (e di fase I) sono dette *senza memoria* in quanto dipendono solo dalle osservazioni al tempo t e non dalle osservazioni precedenti a t.
La situazione di riferimento è la seguente:

Tempo	Dati	Stato
1	$x_{1,1}, \ldots, x_{1,n} \sim f_0$	*In Controllo*
2	$x_{2,1}, \ldots, x_{2,n} \sim f_0$	*In Controllo*
...	...	...
$\tau - 1$	$x_{\tau-1,1}, \ldots, x_{\tau-1,n} \sim f_0$	*In Controllo*
τ	$x_{\tau,1}, \ldots, x_{\tau,n} \sim f_1$	*Fuori controllo*
$\tau + 1$	$x_{\tau+1,1}, \ldots, x_{\tau+1,n} \sim f_1$	*Fuori controllo*
...	...	...

- f_0 rappresenta la distribuzione di probabilità *in controllo*, ovvero quando agisce solo la variabilità naturale sul processo.
- f_1 rappresenta la distribuzione di probabilità delle misure dopo che il processo è andato *fuori controllo*, ovvero quando sono intervenuti dei fattori specifici di variabilità.
- τ rappresenta l'instante di tempo (ignoto) dove il processo è andato fuori controllo. τ può essere anche uguale a 1 (il processo parte fuori controllo);

La domanda, **prospettica**, a cui vogliamo rispondere tramite le carte di controllo di fase II è: al tempo t il processo è ancora in controllo? Che è equivalente a saggiare il seguente sistema d'ipotesi:

$$\begin{cases} \text{H0 : il processo è ancora in controllo} \\ \text{H1 : il proceso è gia andato fuori controllo} \end{cases} \iff \begin{cases} \text{H0} : t < \tau \\ \text{H1} : t \geq \tau \end{cases}$$

Si osservi che ad ogni istante temporale viene verificata una ipotesi differente.
La costruzione delle carte di controllo di fase II avviene considerando $W_t = g(x_{t,1}, \ldots, x_{t,n})$, statistica di controllo (la scelta di W_t dipenderà sempre dal parametro che vogliamo sorvegliare), come una statistica test per saggiare il sistema d'ipotesi sopra descritto.
La statistica di controllo W_t sarà scelta, come accade nella teoria della verifica d'ipotesi, tenendo conto delle differenze attese tra f_0 e f_1. I limiti di controllo non saranno altro che le relative regioni di accettazioni (o regioni di non rifiuto n.d.r.) scelte tenendo conto che siamo interessati ad una sequenza di test e non a un singolo test.
La trattazione dei prossimi argomenti avverrà assumendo che la distribuzione dei dati in controllo sia nota. Nella pratica ciò non è verificato in quanto f_0 risulta essere una stima, ottenuta da dati di fase I, e quindi affetta da errori di stima.
Come ormai noto in letteratura, anche piccoli errori di stima possono influenzare di molto le performance di una carta di controllo (in termini, ad esempio, di falsi allarmi). L'effetto della stima dei parametri sulle carte di controllo è ampiamente discusso da *Jensen at al*[14].
Una misura che costituisce la base su cui si valutano le performance di una carta di controllo di fase II è la run length.
La run length è definita come:

$$\text{RL} = \min\{t : W_t \notin [LCL, UCL]\} \tag{1.3}$$

che rappresenta il numero di istanti di tempo tra l'inizio della sorveglianza e il primo allarme. Essendo che i dati (prima di essere osservati) sono variabili

casuali, anche la run length è una variabile causale.
La run length segue una distribuzione geometrica con probabilità di successo pari a $p = Pr(W_t < LCL \cap W_t > UCL)$:

$$\mathbb{E}_{IC}[RL] = \frac{1}{p} \quad sd_{IC}[RL] = \frac{\sqrt{1-p}}{p} \tag{1.4}$$

Quando il processo si trova in condizioni stabili (o in controllo) l'obiettivo è quello di avere una RL alta (il tempo tra due falsi allarmi sia elevato) mentre quando il processo opera in condizioni instabili (o fuori controllo) l'obiettivo è quello di avere una RL bassa (ovvero la carta segnala rapidamente la presenza di una causa speciale).
Spesso la RL viene espressa in termini di:

- *Average run length*: $\mathbb{E}[RL]$
- *Expected detection delay*: $\mathbb{E}[RL - \tau + 1 | RL > \tau]$

L'ARL calcolata per un processo in controllo ci fornisce il tempo medio tra due falsi allarmi; l'ARL calcolata quando il processo parte fuori controllo ci fornisce quanto mediamente la carta di controllo impiega prima di segnalare che il processo è fuori controllo (in questo caso coincide con l'EDD).

1.4.1 Carte di controllo c/u

A seconda della caratteristica della variabile oggetto di studio le carte di controllo si dividono in: (i) carte di controllo per variabili (ii) carte di controllo per attributi. Le carte di controllo per variabili sono utili quando il parametro da sorvegliare è di tipo continuo. In alcuni casi il parametro da sorvegliare è, ad esempio, un conteggio e quindi di natura discreta (si pensi al caso in cui si voglia sorvegliare il numero di non coformità di un prodotto o il numero totale di pezzi/prodotti difettosi).
In questo caso vengono utilizzate delle carte di controllo per attributi che saranno oggetto chiave del nostro studio. Considereremo sia il caso in cui

la dimensione campionaria è costante sia il caso in cui il campione abbia dimensione variabile.
È possibile costruire carte di controllo sia per il numero totale di non conformità per unità prodotta sia per il numero medio di non conformità per unità prodotta. Per tali carte è lecito assumere che le non conformità contenute in un campione seguano una distribuzione di Poisson. La situazione di riferimento è la seguente:
al tempo t si ispezionano n parti e su ognuna è rilevato il numero di non conformità.

$$x_t \sim \begin{cases} Poisson(nc_0) \text{ se t} < \tau \\ Poisson(nc_1) \text{ se t} \geq \tau \end{cases}$$

Tempo	Dati	Stato
1	$x_{1,1} \ldots, x_{1,n} \sim Poisson(nc_0)$	*In Controllo*
2	$x_{2,1} \ldots, x_{2,n} \sim Poisson(nc_0)$	*In Controllo*
...	...	...
$\tau - 1$	$x_{\tau-1,1}, \ldots, x_{\tau-1,n} \sim Poisson(nc_0)$	*In Controllo*
τ	$x_{\tau,1}, \ldots, x_{\tau,n} \sim Poisson(nc_1)$	*Fuori controllo*
$\tau + 1$	$x_{\tau+1,1}, \ldots, x_{\tau+1,n} \sim Poisson(nc_1)$	*Fuori controllo*
...	...	...

Il dato disponibile è quindi:

$$x_t = \text{numero totale di non conformità}$$

con

$$Pr_{IC}(X = x) = \frac{e^{-c_0} c_0^x}{x!} \tag{1.5}$$

dove:

$$\mathbb{E}_{IC}[x] = c_0 \quad \mathbb{V}_{IC}[x] = c_0$$

- n è noto
- c_0 si assume noto o opportunamente stimato da dati di fase I
- τ e c_1 ignoti

Affinché l'assunzione distributiva valga:

- ciascuna delle $x_t \sim Poisson(c_0)$ per $t = 1, \ldots, n$
- il numero di non conformità di ogni pezzo/prodotto ispezionato sia indipendente

A seconda della statistica di controllo considerata si parlerà di:

- Carta di controllo **u** se $W_t = x_t$
- Carta di controllo **c** se $W_t = \hat{c} = \frac{x_t}{n}$

I limiti di controllo per le carte c/u (come per ogni tipo di carta) sono determinati per garantire una ARL in controllo pari a un valore prefissato (ad esempio B) sulla base dell'applicazione. Essendo che x_t segue una distribuzione discreta non sempre è possibile ottenere dei limiti di controllo esatti; in tal caso dovranno comunque soddisfare (almeno approssimativamente) tali equazioni:.

	equazione LCL	equazione UCL
Unilaterali	$LCL = 0$	$Pr_{IC}(W_t > UCL) = \frac{1}{B}$

(1.6)

	equazione LCL	equazione UCL
Bilaterali	$Pr_{IC}(W_t < LCL) = \frac{1}{2B}$	$Pr_{IC}(W_t > UCL) = \frac{1}{2B}$

(1.7)

La struttura tradizionale dei limiti di controllo per il numero totale di non

conformità (bilaterali) saranno del tipo:

$$\begin{aligned} UCL &= c_0 + L\sqrt{c_0} \\ CL &= c_0 \\ LCL &= c_0 - L\sqrt{c_0} \end{aligned} \tag{1.8}$$

In alcune applicazioni risulta difficile estrarre campioni con dimensione costante. Nei casi in cui la dimensione campionaria non è costante possiamo calcolare le statistiche di controllo dai dati disponibili al tempo t:

$$W_t = W(x_{t,1}, \ldots, x_{t,n_t})$$

e rappresentarle assieme a dei limiti di controllo LCL_t e UCL_t che variano nel tempo.
È facile dimostrare che se i limiti di controllo sono determinati a partire dall'equazione 1.6 o 1.7 la distribuzione della run length rimane (approssimativamente o esattamente) una distribuzione geometrica.
Possiamo quindi calcolare per i vari n_t dei limiti di controllo che ci garantiscono la stessa ARL in controllo.
Si noti però che i valori critici L possono essere "pre-calcolati" solo se le numerosità n_t sono note a priori; inoltre è sconsigliato usare lo schema **u** quando n è variabile in quanto il numero totali di non conformi o delle non conformità non sono tra istanti di tempo diversi confrontabili.

Capitolo 2

Modellazione di processi ad alto rendimento

In questa capitolo, presenteremo la distribuzione di Poisson e Binomiale negativa zero-inflated utili nella sorveglianza di processi ad alto rendimento; inoltre verrà proposto un opportuno modello per dati di conteggio basato su queste distribuzioni.

2.1 Distribuzioni zero-inflated

I dati di conteggio sono classificati come interi discreti, non negativi e più comunemente modellati dalle distribuzioni Poisson o binomiale negativa (NB). Una variabile casuale X che segue la consueta distribuzione di *Poisson* con parametro λ, $Poi(\lambda)$, con funzione di densità:

$$Pr(Y = y_i) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i}, \qquad y_i \in \{0, 1, 2, \dots\}, \lambda > 0 \tag{2.1}$$

avente valore atteso e varianza pari a:

$$\mathbb{E}[Y] = \mathbb{V}[Y] = \lambda$$

è nota come distribuzione di equidispersione in quanto media e varianza coincidono.

Una variabile casuale Y distribuita secondo una *Binomiale Negativa* con parametri λ e τ,$Nb(\lambda, \tau)$, con funzione di densità:

$$Pr(Y = y_i) = \left(\frac{\Gamma(y_i + \tau)}{y!\Gamma(\tau)}\right)\left(\frac{\tau}{\lambda + \tau}\right)^{\tau}\left(\frac{\lambda}{\lambda + \tau}\right)^{y} y_i \in \{0, 1, 2, ..\}, \lambda, \tau > 0. \tag{2.2}$$

avente valore atteso e varianza pari a:

$$\mathbb{E}[Y] = \lambda$$
$$\mathbb{V}[Y] = \lambda + \frac{\lambda^2}{\tau}$$

è nota come distribuzione di sovra-dispersione con τ parametro di forma che governa l'intensità dell'eccessiva dispersione. Si noti che quando $\tau \to +\infty$ (no sovra-dispersione) la distribuzione NB si riduce alla distribuzone di Poisson.

Spesso, nel monitoraggio di processi ad alto rendimento, il numero di zeri contenenti nel data-set di conteggio è elevato rispetto agli zeri intrinsecamente consentiti dall'ordinaria distribuzione di Poisson. In tal situazione, un eccesso di zeri nel campione può causare la violazione dell'ipotesi di equidispersione e i modelli ordinari di Poisson forniscono stime distorte e inadeguate. A tal proposito, *Lambert*[6] ha introdotto la distribuzione di Poisson a inflazione zero (ZIP) come alternativa alla normale Poisson. Allo stesso modo, la distribuzione binomiale negativa (ZINB) a inflazione zero è un'alternativa alla distribuzione NB ordinaria (*McCullagh e Nelder*[10]).

2.1.1 La distribuzione ZIP

La distribuzione ZIP è una generalizzazione della distribuzione standard di Poisson e viene spesso utilizzata per modellare dati di conteggio con un numero eccessivo di zeri.

La distribuzione ZIP presuppone che i dati provengano da due processi: (i) il primo processo modella una proporzione $(1-p)e^{-\lambda}$ di zeri provenienti dalla distribuzione di Poisson e dall'inflazione-zero includendo "zeri" con una

proporzione p; (ii) il secondo processo modella i conteggi diversi da zero provenienti da una distribuzione di Poisson troncata a zero.
Pertanto, sia Y una variabile casuale, tale che $Y \sim ZIP(p, \lambda)$, allora:

$$P(Y_i = y_i) = \begin{cases} p + (1-p)e^{-\lambda} & \text{se } y_i = 0 \\ (1-p)\frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \text{se } y_i \in \{1, 2, 3 \dots\} \end{cases} \tag{2.3}$$

dove $0 \leq p \leq 1$ e $\lambda > 0$
Avente valore atteso e varianza:

$$\mathbb{E}[Y] = (1-p)\lambda$$
$$\mathbb{V}[Y] = (1-p)(\lambda + p\lambda^2)$$

Si noti che quando $p \to 0$ il modello ZIP si riduce all'ordinario modello di Poisson; in caso contrario, il modello risulta eccessivamente disperso a causa di una varianza maggiore rispetto alla media, dovuta all'eccessiva presenza di zeri.

2.1.2 La distribuzione ZINB

In genere, quando un fenomeno (di conteggio) presenza un'eccessiva dispersione e eterogeneità tra i dati, viene modellato attraverso una distribuzione *Binomiale Negativa*. Se nel data-set è presente un'eccessiva presenza di zeri, tali fenomeni vengono modellati attraverso la distribuzione Binomiale negativa a *inflazione zero*. Tale distribuzione nasce dall'insieme di una distribuzione Binomiale Negativa e da una distribuzione degenere in zero. Come per il modello ZIP, anche la distribuzione ZINB si compone di due processi: (i) il primo processo modella una proporzione $(1-p)(1+\frac{\lambda}{\tau})^{-\tau}$ di zeri provenienti da una distribuzione NB e dall'inflazione-zero, includendo "zeri" con una proporzione p. (ii) il secondo modella i conteggi diversi da zero provenienti da una distribuzione NB troncata a zero.

Pertanto, sia Y una variabile casuale tale che $Y \sim ZINB(p, \lambda, \tau)$, allora:

$$P(Y_i = y_i) = \begin{cases} p + \big(1-p\big)\left(1+\frac{\lambda}{\tau}\right)^{-\tau} & \text{se } y_i = 0 \\ \big(1-p\big)\dfrac{\Gamma(y_i+\tau)}{y_i!\Gamma(\tau)}\left(1+\frac{\lambda}{\tau}\right)^{-\tau}\left(1+\frac{\tau}{\lambda}\right)^{-y_i} & \text{se } y_i \in \{1, 2\dots\} \end{cases} \tag{2.4}$$

dove $0 \leq p \leq 1, \lambda, \tau > 0$
Avente valore atteso e varianza:

$$\mathbb{E}[Y] = (1-p)\lambda$$
$$\mathbb{V}[Y] = \lambda\big(1-p\big)\left(1 + p\lambda + \frac{\lambda}{\tau}\right)$$

Si noti che la distribuzione ZINB si riduce all'ordinaria distribuzione NB quando $p \to 0$, alla distribzione ZIP quando $\tau \to 0$ e alla distribuzione di Poisson quando $\frac{1}{\tau}$ e $p \approx 0$.

2.2 Modelli di regressione zero-inflated

Nella pratica, capita spesso che oltre alla variabile d'interesse (caratteristica di qualità che vogliamo sorvegliare, n.d.r.) in un processo si osservano anche altre informazioni aggiuntive correlate con la variabile d'interesse, chiamate covariate.
È possibile quindi costruire carte di controllo basate su modelli di regressione; nella letteratura della sorveglianza statistica di processo vengono denominate "carte di controllo model-based". Esiste una vasta letteratura a riguardo, in questo studio ci concentreremo principalmente sui modelli di conteggio quali: *Poisson* e *BinomialeNegativa*. Una panoramica relativa alle carte di controllo per i profili di regressione è fornita *Maleki et al*[9].
Negli studi model-based sopra menzionati, si presuppone che non vi sia inflazione zero nella variabile risposta. Tuttavia, nella maggior parte dei casi reali, la variabile risposta di conteggio subisce l'inflazione zero, e i convenzionali modelli di regressione di Poisson e NB forniscono stime distorte e fuorvianti.

Per sopraffare a ciò, le versioni zero-inflated dei modelli di Poisson (ZIP) e NB ($ZINB$) risultano essere le migliori alternative per ottenere delle stime adeguate.

2.2.1 Il modello ZIP

Per la modellazione di dati di conteggio che presentano un'eccessiva presenza di zeri, il modello di regressione ZIP, come proposto da Lambert[6], tratta i dati come una combinazione di due fattori: (i) zeri costanti con probabilità p_i (ii) conteggi derivanti da una distribuzione di Poisson. Il modello è il seguente:

$$Y_i|X_i, Z_i \sim \begin{cases} 0 & \text{con probabilità } p_i \\ Poisson(\lambda_i) & \text{con probabilità } (1-p_i) \end{cases} \tag{2.5}$$

dove X_i è un vettore di covariate correlate linearmente con Y_i mentre Z_i è un vettore di covariate che definiscono la probabilità di eccesso zero. La distribuzione che ne risulta è quella descritta nella precedente sezione (si veda la 2.3).
Media e varianza condizionata del modello ZIP sono pari a $\mathbb{E}[Y_i|X_i, Z_i] = (1-p_i)\lambda_i$ e $\mathbb{V}[Y_i|X_i, Z_i] = (1-p_i)(\lambda_i + p_i\lambda_i^2)$.
Nel modello ZIP, p_i è modellato usando un modello logit mentre la funzione di legame per λ_i è quella logaritmica:

$$logit(p_i) = Z_i'\gamma \implies p_i = \frac{e^{Z_i'\gamma}}{1+e^{Z_i'\gamma}}$$
$$log(\lambda_i) = X_i'\beta \implies \lambda_i = e^{X_i'\beta}$$

dove $\gamma' = (\gamma_0, \gamma_1, \ldots, \gamma_m)$ e $\beta' = (\beta_0, \beta_1, \ldots, \beta_m)$ sono vettori di parametri non noti.
Si noti che le variabili che influenzano il processo di conteggio possono essere anche le variabili che influenzano il processo degli zeri in eccesso; questo per dire che X_i e Z_i non si escludono a vicenda ma possono assere anche

sovrapposte o uguali.
In questo studio si assume che $y_1, y_2, \ldots, y_n$ siano indipendenti e che p_i sia incorrelato con λ_i (ciò non sempre è verificato nella pratica). Possiamo quindi definire la funzione di verosimiglianza per la variabile risposta come:

$$L(Y, \beta, \gamma) = \prod_{y_i=0} [p_i + [1 - p_i]e^{-\lambda_i}] + \prod_{y_i>0} \left[[1 - p_i] \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \right]$$

e la funzione di log-verosimiglianza è definita come:

$$\begin{aligned} \ell(Y, \beta, \gamma) = & \sum_{y_i=0} ln[exp(Z_i'\gamma) + exp(-exp(X_i'\beta))] + \\ & + \sum_{y_i>0} [Y_i exp(X_i'\beta) - exp(X_i'\beta) - ln(y_i!)] - \\ & - \sum_{i=1}^{n} ln[(1 + exp(Z_i'\gamma))] \end{aligned}$$

Sia I_i una funzione indicatrice definita come segue:

$$I_i = \begin{cases} 1 & y_i = 0 \\ 0 & y_i > 0 \end{cases}$$

La funzione di log-verosimiglianza diventa quindi:

$$\begin{aligned} \ell(Y, \beta, \gamma) = & \sum_{y_i=0} I_i ln[exp(Z_i'\gamma) + exp(-exp(X_i'\beta))] + \\ & + \sum_{y_i>0} (1 - I_i)[Y_i exp(X_i'\beta) - exp(X_i'\beta) - ln(y_i!)] - \\ & - \sum_{i=1}^{n} ln[(1 + exp(Z_i'\gamma))] \end{aligned}$$

Tale funzione di verosimiglianza può essere massimizzata usando l'algoritmo EM (per maggiori dettagli si veda *Søren Asmussen et al*[2]).
L'interpretazione del parametro λ è lo stesso della regressione di Poisson, mentre il parametro γ può essere interpretato attraverso i rapporti dispari (per maggiori dettagli si veda *Lambert*[6]).

Solitamente, nella regressione l'analisi dei residui viene utilizzata per valutare la bontà del modello. In particolare, viene utilizzata per valutare l'omoschedasticità, le deviazioni dall'errore e per rilevare la presenza di eventuali outliers. Per il modello di regressione sono disponibili diversi tipi di residui, ma per il modello zero-inflated, i residui di Pearson risultato essere un utile strumento per la convalida del modello adattato (si veda *Garay et al*[4]).

I residui di Pearson possono essere determinati dalla seguente espressione:

$$PR_i = \frac{y_i - (1-p_i)\lambda_i}{\sqrt{(1-p_i)(\lambda_i + p_i\lambda_i^2)}} \tag{2.6}$$

2.2.2 Il modello ZINB

Per la modellazione di dati di conteggio che presentano una particolare eterogeneità e un'eccessiva presenza di zeri, il modello di regressione ZINB tratta i dati come una combinazione di due fattori: (i) zeri costanti con probabilità p_i (ii) conteggi derivanti da una distribuzione NB con probabilità $1 - p_i$. Il modello è il seguente:

$$Y_i|X_i, Z_i \sim \begin{cases} 0 & \text{con probabilità } p_i \\ NB(\lambda_i, \tau) & \text{con probabilità } (1-p_i) \end{cases} \tag{2.7}$$

dove X_i è un vettore di covariate correlate linearmente con Y_i mentre Z_i è un vettore di covariate che definiscono la probabilità di eccesso zero. La distribuzione che ne risulta è quella descritta nel precedente sezione (si veda la 2.4).

Media e varianza condizionata del modello ZINB sono pari a $\mathbb{E}[Y_i|X_i, Z_i] = (1-p_i)\lambda_i$ e $\mathbb{V}[Y_i|X_i, Z_i] = \lambda_i(1-p_i)(1+p_i\lambda_i+\lambda_i/\tau)$.

In pratica, i parametri λ_i e p_i dipendono rispettivamente dai vettori delle variabili esplicative X_i e Z_i e, come per il modello ZIP, anche per il modello ZINB p_i è modellato usando un modello logit e la funzione di legame per λ_i è quella logaritmica :

$$logit(p_i) = Z_i'\gamma \implies p_i = \frac{e^{Z_i'\gamma}}{1+e^{Z_i'\gamma}}$$
$$log(\lambda_i) = X_i'\beta \implies \lambda_i = e^{X_i'\beta}$$

La funzione di log-verosimgilianza dato un campione di osservazioni, è definita come segue:

$$\begin{aligned}\ell(Y,\beta,\gamma,\tau) = \sum_{i=1}^{n} ln(1+e^{Z_i'\gamma}) - \\ - \sum_{i:y_i=0} ln\left(e^{Z_i'\gamma} + (\frac{e^{X_i'\beta}+\tau}{\tau})^{-\tau}\right) + \\ + \sum_{i:y_i>0} ln\left(\tau\left(\frac{e^{X_i'\beta}+\tau}{\tau}\right) + y_i ln\big(1+e^{X_i'\beta}\tau\big)\right) + \\ + \sum_{i:y_i>0} ln(\Gamma(\tau) + ln(\Gamma(1+y_i)) - ln(\Gamma(\tau+y_i)))\end{aligned}$$

La stima dei parametri del modello ZINB è ottenuta con il metodo BFGS. Come discusso in precedenza, l'analisi dei residui ricopre un ruolo fondamentale nella valutazione della bontà del modello. Pertanto, per il modello ZINB, i residui di Pearson sono ottenuti dalla seguente espressione:

$$PR_i = \frac{y_i - (1-p_i)\lambda_i}{\sqrt{\lambda_i(1-p_i)(1+p_i\lambda_i+\lambda_i/\tau)}} \tag{2.8}$$

2.3 Procedura di selezione del modello

La scelta di un appropriato modello per il data-set è un passaggio delicato nella progettazione di un adeguato schema di sorveglianza, in quanto una scelta errata di un modello può portare a stime distorte e fuorvianti e lo schema di sorveglianza basato su di esso può portare a trarre false conclusioni da parte dei key-user.
In questo studio viene adottato il metodo in due-fasi discusso da *Walter*[12]. Il metodo in due-fasi è noto come metodo LRT-Vuong, la cui procedura prevede appunto due fasi: (i) il test del rapporto di verosimiglianza (LRT) viene utilizzato per testare l'eccessiva dispersione nei dati (ii) il test Vuong viene utilizzato per verificare l'inflazione zero nei dati.
Essendo che il modello di Poisson e un caso particolare del modello NB, verificare se il modello di Poisson è adeguato corrisponde a saggiare il seguente

sistema d'ipotesi:

$$\begin{cases} H_0 : \tau = 0 \\ H_1 : \tau \neq 0 \end{cases}$$

Per un generico modello di regressione NB, il rapporto di verosimiglianza LRT per τ è definito da:

$$LRT_\tau = -2(\ell(Y, \hat{\beta}) - \ell(Y, \hat{\beta}, \hat{\tau}))$$

dove $\ell(Y, \hat{\beta})$ è la log-verosimiglianza massimizzata sotto il modello di Poisson mentre $\ell(Y, \hat{\beta}, \hat{\tau})$ è la log-verosimiglianza massimizzata sotto il modello NB.

Sotto l'ipotesi nulla la statistica LRT si distribuisce come una variabile casuale chi-quadro con una grado di libertà. Il test LRT fornisce evidenza della sovra-dispersione nei dati quando produce risultati significativi (si rifiuta l'ipotesi nulla).

Il test Vuong viene utilizzato per verificare la presenza di inflazione zero nei dati; confronta quindi il modello di Poisson con il modello ZIP, o il modello NB con il modello ZINB.

Sia $P(Y_i = y_i|X_i)$ la probabilità di un conteggio osservato prevista da un modello standard (Poisson, NB) e sia $P(Y_i = y_i|X_i, Z_i)$ la probabilità di un conteggio osservato prevista da un modello a inflazione zero (ZIP, ZINB), la statistica m_i basata sul loro rapporto e definita come segue:

$$m_i = log\left(\frac{P(Y_i = y_i|X_i)}{P(Y_i = y_i|X_i, Z_i)}\right)$$

Quindi, la statistica test Vuong per saggiare il sistema d'ipotesi $H_0 : \mathbb{E}[m_i] = 0$ contro $H_1 : \mathbb{E}[m_i] \neq 0$ è definita come:

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(m_i - \frac{1}{n}\sum_{i=1}^{n} m_i\right)^2}}$$

La statistica Vuong,sotto H_0, si distribuisce asintoticamente come una variabile casuale Normale. Fissato un livello di significatività del 5%, il modello

standard è preferito quando $V > 1.96$, il modello a inflazione zero quando $V < -1.96$ mentre sono considerati equivalenti quando $|V| < 1.96$.
In questa procedura di selezione del modello, il primo step consiste nel valutare se i dati presentano sovra-dispersione o meno attravero il test LRT (Poisson VS NB). Il secondo step consiste nel valutare la presenza di una inflazione zero nei dati attraverso il test Vuong: se al primo step è stato scelto un modello di Poisson, questo viene confrontato con il modello ZIP altrimenti viene confrontato il modello NB con il modello ZINB.

Capitolo 3

Sorveglianza di processi ad alto rendimento

Il capitolo 3 è incentrato sulla presentazione di alcune soluzioni per la sorveglianza di processi ad alto rendimento.

3.1 Carta di controllo di Shewhart basata sui modelli zero-inflated

Le tradizionali carte di controllo c/u (si veda il paragrafo 1.4.1) applicate a un processo ad alte performance producono un elevato tasso di falsi allarmi, anche quando vengono usati i limiti esatti di probabilità. A tal proposito, *Xie e Goh*[15] hanno proposto una alternativa carta di controllo basata sulla distribuzione ZIP, successivamente estesa da *Xie et al*[17]. L'effetto degli errori di stima derivanti dalla fase I nelle carte di controllo di Shewhart ZIP è stato discusso da *He et al.*[16]

In questa sezione verrà sintetizzata la proposta di *Mahmood*[8] che prevede l'estensione a una struttura di regressione della carta di controllo ZIP sopra menzionata, seguendo due approcci: (i) approccio data-based (ii) approccio model-based.

3.1.1 Carta di controllo data-based

Nelle carte di controllo data-based vengono rappresentante direttamente le osservazioni che si presume, in questo contesto, provengano da una delle distribuzioni zero-inflated ampiamente discusse nella sezione 2.1.

Carta di controllo Y-ZIP

Quando Y_i segue una distribuzione ZIP tale che, $Y \sim ZIP(p_i, \lambda_i)$, media e varianza condizionata sono definite come segue:

$$\mu_{ZIP} = \mathbb{E}[Y_i|X_i, Z_i] = (1 - p_i)\lambda_i$$

$$\sigma^2_{ZIP} = \mathbb{V}[Y_i|X_i, Z_i] = (1 - p_i)(\lambda_i + p_i\lambda_i^2)$$

Xie and Goh[15] hanno proposto una carta di controllo di Shewhart basata sulla distribuzione ZIP considerando come limiti di controllo i limiti di probabilità. La struttura dei limiti di controllo proposta da *Mahmood*[8] segue quella tradizionale.
Nella carta di controllo Y-ZIP, le osservazioni vengono rappresentate assieme al limite di controllo superiore, che si ottiene dalla seguente espressione:

$$UCL = \mu_{ZIP} + L_1\sqrt{\sigma^2_{ZIP}}$$

dove L_1 è un appropriato valore critico che determina l'ampiezza del limite di controllo superiore, scelto per garantire un valore di ARL in controllo (ARL_0) fissato sulla base dell'applicazione.
Quando un'osservazione eccede il limite di controllo superiore, la carta segnalerà una possibile situazione di "fuori controllo"; altrimenti, il processo si trova in una situazione di "in controllo".

Carta di controllo Y-ZINB

Quando Y_i segue una distribuzione ZINB tale che, $Y \sim ZINB(p_i, \lambda_i, \tau)$, media e varianza condizionata sono definite come segue:

$$\mu_{ZINB} = \mathbb{E}[Y_i|X_i, Z_i] = (1 - p_i)\lambda_i$$

$$\sigma^2_{ZINB} = \mathbb{V}[Y_i|X_i, Z_i] = \lambda_i(1 - p_i)(1 + p_i\lambda_i + \lambda_i/\tau)$$

Nella carta di controllo Y-ZINB, le osservazioni vengono rappresentate assieme al limite di controllo superiore, che si ottiene dalla seguente espressione:

$$UCL = \mu_{ZINB} + L_2\sqrt{\sigma^2_{ZINB}}$$

dove L_2 è un appropriato valore critico che determina l'ampiezza del limite di controllo superiore, scelto per garantire un valore di ARL in controllo (ARL_0) fissato sulla base dell'applicazione.
Quando un'osservazione eccede il limite di controllo superiore, la carta segnalerà una possibile situazione di "fuori controllo"; altrimenti, il processo si trova in una situazione di "in controllo".

3.1.2 Carta di controllo model-based

Nelle carte di controllo model-based non vengono rappresentate direttamente le osservazioni, bensì i residui derivanti da una opportuna stima di un modello di regressione zero-inflated (si veda 2.2).
Come già accennato in precedenza, i residui giocano un ruolo cruciale nella verifica delle assunzioni sottostanti al modello, nel controllare specifiche errate del modello e/o rilevare la presenza di eventuali outliers.

Carta di controllo PR-ZIP

Nelle carte di controllo PR-ZIP vengono rappresentati i residui di Pearson, determinati sulla base dell'equazione 2.6, assieme a dei limiti di controllo che

si ottengono dalla seguente espressione:

$$UCL = \mu_{PR-ZIP} + L_3\sqrt{\sigma^2_{PR-ZIP}}$$
$$LCL = \mu_{PR-ZIP} - L_3\sqrt{\sigma^2_{PR-ZIP}}$$

dove L_3 è un appropriato valore critico che determina l'ampiezza dei limiti di controllo, scelto per garantire un valore di ARL in controllo (ARL_0) fissato sulla base dell'applicazione.
Quando un residuo eccede i limiti di controllo, la carta segnalerà una possibile situazione di "fuori controllo"; altrimenti, il processo si trova in una situazione di "in controllo".

Carte di controllo PR-ZINB

Nelle carte di controllo PR-ZINB vengono rappresentati i residui di Pearson, determinati sulla base dell'equazione 2.8, assieme a dei limiti di controllo che si ottengono dalla seguente espressione:

$$UCL = \mu_{PR-ZINB} + L_4\sqrt{\sigma^2_{PR-ZINB}}$$
$$LCL = \mu_{PR-ZINB} - L_4\sqrt{\sigma^2_{PR-ZINB}}$$

dove L_4 è un appropriato valore critico che determina l'ampiezza dei limiti di controllo, scelto per garantire un valore di ARL in controllo (ARL_0) fissato sulla base dell'applicazione.
Quando un residuo eccede i limiti di controllo, la carta segnalerà una possibile situazione di "fuori controllo"; altrimenti, il processo si trova in una situazione di "in controllo".

3.1.3 Prestazioni carta di controllo

Nella progettazione di un valido schema di sorveglianza fondamentale è valutare le performance delle carte di controllo considerate. In particolare si valuta il "livello di sensibilità" delle carte quando il processo opera in condizioni instabili.

Nelle successive sub-sezioni andremo a definire le misure di performance utili a confrontare le carte di controllo e successivamente sintetizzeremo le analisi condotte sui modelli in controllo a cui sono stati applicati determinati *shift* ai fini di valutare il profilo della run-length per diversi cambiamenti (per maggiori dettagli si veda *Mahmood T.*[8]).

Misure di performance

In letteratura sono presenti diverse misure per valutare le performance di una carta di controllo. In questo studio le carte di controllo vengono valutate secondo le proprietà della run-length, definita dall'equazione 1.3, come: (i) Average run length (ARL) (ii) Deviazione standard della run-length (SDRL). L'Average run length (ARL) rappresenta il numero medio di campioni prima che la carta segnali un allarme; è classificato in ARL in controllo (ARL_0) e ARL fuori controllo (ARL_1). L'ARL_1 è la misura di performance della carta quando il processo opera in condizioni instabili mentre l'ARL_0 è la misura di performance della carta per un processo che opera in condizione stabili.
Per un fissato valore di ARL_0 verrà considerata migliore la carta di controllo avente ARL_1 minore.

Simulazione dei modelli in controllo

L'analisi di tale performance prevede che vengano generati dati da un modello ZIP e da un modello ZINB, usando le seguenti specifiche.
Per il modello ZIP:

$$Y_i|X_i, Z_i \sim \begin{cases} 0 & \text{se } c_i = 1 \\ Poisson(\lambda_i) & \text{se } c_i = 0 \end{cases} \tag{3.1}$$

dove,

$$c_i \sim Bernoulli(p_i) \quad X_i \sim Normale(\mu_X = 0, \sigma_X^2 = 0.5) \quad Z_i \sim Normale(\mu_Z = 3, \sigma_Z^2 = 0.5)$$

tale che:

$$\lambda_i = e^{\beta_0 + \beta_1 X_i} \quad p_i = \frac{e^{\gamma_0 + \gamma_1 Z_i}}{1 + e^{\gamma_0 + \gamma_1 Z_i}}$$

La stima dei parametri non è oggetto di questo studio, pertanto si è scelto di seguire la soluzione proposta da *Liu et al*[7] che considerano: $\beta_0 = 1.0$,$\beta_1 = 1.5$, $\gamma_0 = 0.5$, $\gamma_1 = -0.5$ una valida stima in termini di bias, errore di I° tipo e inferenza.
Per il modello ZINB:

$$Y_i|X_i, Z_i \sim \begin{cases} 0 & \text{se } c_i = 1 \\ NB(\lambda_i, \tau) & \text{se } c_i = 0 \end{cases} \tag{3.2}$$

dove,

$$c_i \sim Bernoulli(p_i) \quad X_i \sim Normale(\mu_X = 0, \sigma_X^2 = 1) \quad Z_i \sim Normale(\mu_Z = 0, \sigma_Z^2 = 1)$$

tale che:

$$\lambda_i = e^{\beta_0 + \beta_1 X_i} \quad p_i = \frac{e^{\gamma_0 + \gamma_1 Z_i}}{1 + e^{\gamma_0 + \gamma_1 Z_i}}$$

per la stima dei parametri in questo caso viene seguita la proposta di *Williamnson*[13], che ritiene valida la scelta di tali parametri: $\beta_0 = 1.609$, $\beta_1 = 0.25$, $\gamma_0 = -0.406$, $\gamma_1 = -0.65$ e $\tau = 0.2$.
Lo studio comparativo è effettuato generando $n = 1000$ osservazioni rispettivamente dai modelli 3.1 e 3.2, applicando la procedure di selezione del modello (si veda la sezione 2.3) per verificare l'adattamento effettivo delle osservazioni generate e infine si applicano le carte di controllo in questione. Vengono utilizzati dei limiti di controllo statici determinati usando come valori critici (i.e L_1, L_2) dei valori arbitrari e ripetendo tale operazioni un numero elevato di volte (per maggiori dettagli si veda *Mahmood*[8]).

Risultati e discussioni

Le prestazioni delle carte di controllo vengono valutate tenendo conto dei possibili cambiamenti diretti e indiretti in λ_i per il modello ZIP, e in λ_i e τ per il modello ZINB. Non si considerano cambiamenti nel parametro p_i in quanto *Fatahi et al*[3] sostengono che questo parametro governi solo la probabilità che nel processo si verifichi uno shock, e per questo il suo eventuale cambiamento può considerarsi trascurabile. Gli *shifts* considerati sono quindi:

- Per i modelli ZIP e ZINB gli *shitf* diretti in λ_i vengono introdotti cambiando λ_i in $\lambda_i + \delta\sqrt{\sigma^2_{ZIP}}$ o $\lambda_i + \delta\sqrt{\sigma^2_{ZINB}}$ rispettivamente.

- Per i modelli ZIP e ZINB gli *shift* indiretti in λ_i vengono introdotti cambiando β_0 in $\beta_0 + \eta$ oppure cambiando la media delle covariate X_i: da μ_X a $\mu_X + \omega$.

- Per il modello ZINB gli *shift* in τ vengono introdotti cambimando τ in $\tau - \Omega$.

In tutti gli *shifts* considerati si valuteranno cambiamenti di diversa grandezza ai fini di valutare la sensibilità della carta per cambiamenti da piccoli a grandi. Inoltre come valori di ARL_0 si è scelto di considerare due prefissati valori: (i) $ARL_0 = 200$ (ii) $ARL_0 = 500$ così da poter evidenziare le varie performance per valori di ARL in controllo diversi.
Per i cambiamenti diretti in λ_i, per un valore di $ARL_0 = 200$, abbiamo che per il modello ZIP la carta di controllo model-based (i.e. PR-ZIP) è di gran lunga più performante della carta di controllo data-based (i.e. Y-ZIP); per $\delta = 0.5$ abbiamo un valore di $ARL_1 = 47.82$ per la carta PR-ZIP, mentre per la carta Y-ZIP un valore di $ARL_1 = 96.94$. Lo stesso vale per $\delta = 1$: $ARL_1 = 17.58$ per la carta PR-ZIP, $ARL_1 = 54.13$ per la carta Y-ZIP. Il confronto produce analoghi risultati anche per $\delta = 2$.
Per il modello ZINB la carta di controllo model-based anche in questo caso risulta più performante, anche se con minor evidenza: per $\delta = 0.5$ abbiamo un valore di $ARL_1 = 45.63$ per la carta PR-ZINB e un valore di $ARL_1 = 48.82$ per la carta Y-ZINB. Per $\delta = 1$ (ARL_1: $ARL_{PR} = 16.80$, $ARL_Y = 19.88$) e $\delta = 2$ (ARL_1: $ARL_{PR} = 5.42$, $ARL_Y = 6.64$) le due carte raggiungono prestazioni quasi analoghe, permane comunque una leggera sensibilità in più nella carta di controllo PR-ZINB.
Tali risultati vengono riscontrati anche per un fissato valore di $ARL_0 = 500$.
Per i cambiamenti indiretti in λ_i introdotti in β_0 abbiamo che, per un valore di $ARL_0 = 370$, per il modello ZIP la carta più performante è ancora quella model-based, anche se la differenza tra i due valori di ARL_1 è dimi-

nuita per cambiamenti grandi (i.e. $\eta = 1.60$, $ARL_1 = 1.53$ per PR-ZIP, $ARL_1 = 4.22$ per Y-ZIP) mentre rimane ancora evidente per cambiamenti piccoli (i.e. $\eta = 0.20$, $ARL_1 = 81.16$ per PR-ZIP, $ARL_1 = 104.39$ per Y-ZIP).
Per il modello ZINB notiamo invece come i valori di ARL fuori controllo siano molto simili sia per cambiamenti grandi (i.e. $\eta = 1.60$, $ARL_1 = 1.83$ per PR-ZINB, $ARL_1 = 2.00$ per Y-ZINB) che per cambiamenti piccoli (i.e. $\eta = 0.20$, $ARL_1 = 64.50$ per PR-ZINB, $ARL_1 = 69.88$ per Y-ZINB). Analogamente per ARL in controllo pari a 500.
Per quanto riguarda il secondo tipo di cambiamento indiretto in λ_i (da μ_X a $\mu_X + \psi$) notiamo in primis la scarsa performance di entrambe le carte per entrambi i modelli, anche per cambiamenti grandi; rimane comunque leggermente più performante la carta di controllo model-based. L'ultimo *shift* considerato si riferisce al parametro aggiuntivo (i.e τ) del modello ZINB che governa la sovra-dispersione; a tutti i livelli di cambiamenti la carta di controllo che fornisce le migliori prestazioni è quella model-based.
In sostanza, la carta di controllo model-based fornisce sempre prestazioni migliori rispetto a quella data-based anche considerando cambiamenti di grandezza diversa, e diversi valore di ARL in controllo (per maggiori dettagli si veda *Mahmood*[8]). Nel complesso, si nota come le performance delle carte di controllo considerate non siano particolarmente "brillanti" soprattutto quando il cambiamento è piccolo. In parte ciò è dovuto alla struttura della carta di controllo di Shewhart, inadeguata nel rilevare determinati *shift*, in parte è dovuto alla metodologia applicata nello studio , in particolare nella scelta e costruzione dei limiti di controllo.

3.2 Carta di controllo EWMA basata sui modelli zero-inflated

Un'altra soluzione presente in letteratura, frequentemente utilizzata per modellare questi tipo di processi, è stata proposta da *Fatahi et al*[3] che considerano una carta di controllo EWMA basata su una distribuzione ZIP (estendibile anche a modelli ZINB) utilizzando dei limiti di controllo asintotici. A differenza della soluzione proposta da *Mahmmood*[8], *Fatahi et al*[3] non considerano una struttura di regressione.
La carta di controllo presenta tale struttura:
Sia $Y_i \sim ZIP(\lambda, p)$ per $i = 1, \ldots, n$ tale che $\mathbb{E}[Y] = (1-p)\lambda$. Quando il processo opera in condizioni stabili, $\mathbb{E}_{IC}[Y_i] = (1-p_0)\lambda_0$, con λ_0 e p_0 noti o opportunamente stimati da dati di fase I. La statistica di controllo W_t è definita come segue:

$$W_t = \begin{cases} (1-p_0)\lambda_0 & t = 0 \\ wY_t + (1-w)W_{t-1} & t = 1, 2, \ldots \end{cases} \tag{3.3}$$

dove $0 < w < 1$ è la costante di lisciamento che determina il livello di memoria dello schema EWMA; piccoli valori di w attribuiscono maggior peso alle osservazioni passate e quindi adatti per identificare cambiamenti piccoli, valori grandi di w attribuiscono maggior peso alle osservazioni recenti e risultano adatti per identificare cambiamenti grandi.
La statistica di controllo può anche essere scritta come:

$$W_t = w \sum_{i=0}^{t-1} (1-w) Y_{t-i} + (1-w)^t (1-p_0)\lambda_0$$

Avente valore atteso (in controllo) e varianza pari a:

$$\begin{aligned} \mathbb{E}[W_t] &= (1-p_0)\lambda_0 \\ \mathbb{V}[W_t] &= \frac{w}{2-w}\left[1-(1-w)^{2t}\right]\left[(1-p)\lambda + p\lambda^2\right] \end{aligned}$$

I limiti di controllo time-varying dello schema ZIP-EWMA sono definiti come

segue:

$$UCL_t = (1-p_0)\lambda_0 + L\sqrt{\frac{w}{2-w}\left[1-(1-w)^{2t}\right]\left[(1-p)\lambda+p\lambda^2\right]}$$

$$CL = (1-p_0)\lambda_0$$

$$LCL_t = (1-p_0)\lambda_0 - L\sqrt{\frac{w}{2-w}\left[1-(1-w)^{2t}\right]\left[(1-p)\lambda+p\lambda^2\right]}$$

Il limiti di controllo inferiore, come fanno notare *Fatahi et al*[3], è usato solitamente quando il numero di zeri è troppo piccolo. La carta di controllo viene costruita tracciando le statistiche W_t rispetto ai limiti di controllo. Quando una statistica W_t eccede il limite di controllo superiore, la carta segnalerà una possibile situazione di "fuori controllo"; altrimenti, il processo si trova in una situazione di "in controllo".

Capitolo 4

Nuova proposta e applicazione dei risultati

4.1 Motivazioni

Si è scelto di considerare una soluzione diversa da quelle proposte da *Mahmood*[8] e *Fatahi et al*[3] principalmente perché: (i) le carte di controllo di Shehwhart proposte da *Mahmood*[8] presentano una struttura di regressione, risultano essere un buon punto di partenza in quanto sono facilmente calcolabili e interpretabili ma non tengono ne in considerazione la possibilità che la distribuzione in controllo vari nel tempo, ne che la numerosità campionaria possa non essere nota a priori; inoltre le carte di controllo di Shewhart presentano gli ormai noti limiti nell'identificare cambiamenti piccoli (non cumulano le osservazioni) (ii) le carte di controllo proposte da *Fatahi et al*[3] potrebbero risultare una valida scelta anche per la struttura dei limiti che presentano, ma non tengono in considerazione una struttura di regressione nei dati.
A tal proposito nella successiva seziona verrà presentata una alternativa carta di controllo EWMA che includa una struttura di regressione e che consideri dei limiti di controllo dinamici.

4.2 Carta di controllo EWMA model-based con limiti di controllo dinamici

La scelta di una carta di controllo EWMA avviene principalmente per due motivi: (i) performance nel rilevare rapidamente cambiamenti sia piccoli che grandi a seconda della scelta di λ (ii) facilità di interpretazione della carta da parte dei key-user.
Si è scelto di applicare una carta di controllo EWMA model- based unilaterale in quanto è di interesse rilevare rapidamente un aumento del numero di non conformità. La struttura della carta di controllo è la seguente:
Sia $Y_i|X_i, Z_i \sim ZIP(\lambda_i, p_i)$ (o ZINB) per $i = 1, \ldots, n$ tale che $\mathbb{E}[Y_i|X_i, Z_i] = (1 - p_i)\lambda_i$, $\mathbb{V}[Y_i|X_i, Z_i] = (1 - p_i)(\lambda_i + p_i\lambda_i^2)$.
La statistica di controllo W_t è definita come segue:

$$W_t = \begin{cases} 0 & t = 0 \\ \max\big[0, \omega PR_i + (1-\omega)W_{t-1}\big] & t = 1, 2, \ldots \end{cases}$$

dove:

$$PR_i = \frac{y_i - (1 - p_i)\lambda_i}{\sqrt{(1 - p_i)(\lambda_i + p_i\lambda_i^2)}}$$

indicano i residui di Pearson ottenuti dal modello stimato e $0 < \omega < 1$ è la costante di lisciamento che determina il livello di memoria dello schema EWMA; piccoli valori di w attribuiscono maggior peso alle osservazioni passate e quindi adatti per identificare cambiamenti piccoli, valori grandi di ω attribuiscono maggior peso alle osservazioni recenti e risultano adatti per identificare cambiamenti grandi.
La successiva sub-sezione è dedicata alla trattazione dei limiti di controllo che andremo a considerare.

4.2.1 Limiti di controllo di probabilità dinamici

I limiti di controllo fino ad ora considerati vengono calcolati a priori e presumono inoltre che: (i) la distribuzione in controllo non vari nel tempo, (ii) la

numerosità campionaria sia noti a priori. Se consideriamo le carte di controllo model-based menzionate nel capitolo precedente, è lecito aspettarsi che la statistica W_t dipenda dal valore assunto dalle covariate ai vari istanti temporali. Per tale motivo, sconsigliato è assumere che la distribuzione della statistica di controllo rimanga costante, in quanto fare una ipotesi distributiva inappropriata potrebbe portare a prestazioni inaspettate della carta di controllo quali ad esempio, falsi allarmi eccessivi.
Per ovviare a ciò proponiamo dei limiti di controllo di probabilità dinamici calcolati in maniera tale da garantire un valore di ARL in controllo fissato, ad ogni istante temporale.
Per maggiori dettagli riguardo la struttura di tali limiti si veda *Shen X. at al*[11], in questo studio ci limiteremo a dare una idea generali della loro struttura per poi applicarli nel successivo capitolo.
Considerato che la distribuzione di W_t varia nel tempo è lecito usare un limite di controllo variabile con t, ovvero segnalare un allarme quando:

$$W_t > L_t$$

dove L_t è una successione appropriata che dipenderà dai valori delle covariate e dalla dimensione del campione. Per garantire un valore di ARL in controllo desiderato, una possibilità è quella di fissare L_t in maniera tale che:

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

Questo garantisce che la distribuzione della run-length rimanga geometrica con media B.
Quindi L_t può essere calcolato come il quantile $1 - 1/B$ della distribuzione di W_t condizionata a $W_1 \leq L_1$, $W_2 \leq L_2$,...,$W_{t-1} \leq L_{t-1}$, visto che:

$$Pr(RL > t|RL \geq t) = Pr(W_t \leq L_t|W_1 \leq L_1, \ldots, W_{t-1} \leq L_{t-1}) = 1 - \frac{1}{B}$$

Solitamente la distribuzione W_t è difficile da calcolare quindi il calcolo di L_t avverrà via simulazione. Un possibile algoritmo è presentano nella successiva applicazione ad un dataset reale.

4.3 Un caso di studio sul numero di non conformità in un processo di verniciatura

L'ultima sezione ha lo scopo di mostrare come gli strumenti presentati in questo lavoro possano essere di supporto in merito alla sorveglianza di un processo produttivo aventi standard qualitativi elevati.

4.3.1 Il dataset

L'applicazione si concentra su una fase del ciclo di produzione di un bene prodotto dall'azienda "x". Tale fase prevede l'applicazione di uno strato di vernice, attraverso uno strumento ad alta precisione. Recentemente l'azienda ha deciso di rinnovare tale processo di verniciatura, rinunciando quindi alle metodologie adottate fino a quel momento (che garantivano un elevato standard qualitativo) in quanto ritenevano che la qualità potesse aumentare ancora. La questione che si pone l'azienda ora è quella di valutare se il nuovo processo ha effettivamente alzato gli standard qualitativi della verniciatura o meno.
Ovviamente, il numero di difetti nel processo di verniciatura dipende anche da diverse altre variabili, quali ad esempio il livello di umidità nell'ambiente produttivo.
I dati si riferiscono a 877 lotti di produzione, dove i primi 824 sono stati prodotti con il vecchio processo (NEW= 0), i restanti con il nuovo (NEW = 1) ; per i suddetti lotti viene rilevato: (i) il livello di umidità presente nell'ambiente per ogni lotto prodotto (ii) il numero di non conformità della verniciatura.

4.3.2 Analisi preliminare e stima del modello

Una preliminare analisi sul numero di non conformità rilevate mostra che 639 lotti di produzione hanno registrato zero n.c. (non conformità, n.d.r.), 48 lotti una n.c., 49 lotti due n.c, 42 lotti tre n.c. e 99 lotti di produzione

hanno registrato più di tre n.c. L'istogramma del numero di non conformità è riportato in figura 4.1.

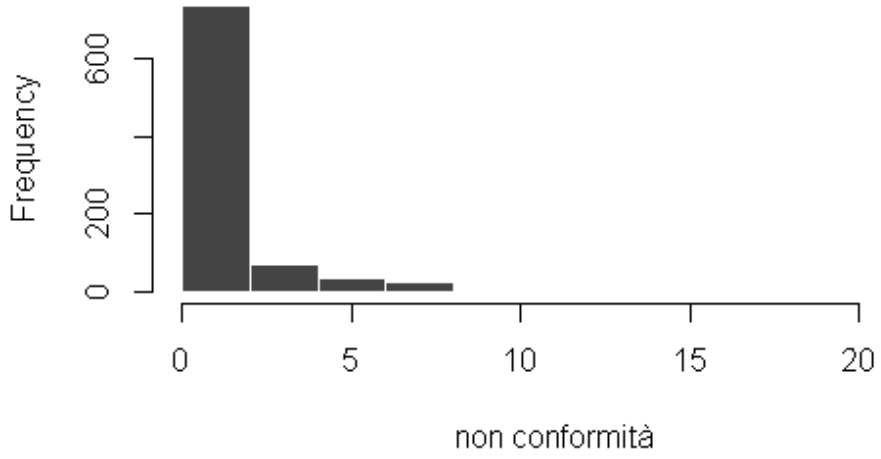

Figura 4.1: Istogramma n° di n.c.

È ragionevole assumere che il numero di non conformità segua una distribuzione di Poisson (o Binomiale Negativa in caso di sovra-dispersione).
A tal proposito, considerando anche le informazioni aggiuntive disponibili, viene condotta una analisi diagnostica sui modelli, partendo dalla stima di un modello di regressione di Poisson confrontandolo con un modello Binomiale Negativo (per valutare una eventuale eccessiva dispersione). La tabella 4.1 riporta la stima dei modelli di regressione sopra menzionati.

Come si può notare per entrambi i modelli i parametri β_1 e β_2, riferiti alle variabili di umidità e alla variabile dicotomica sull'uso del processo nuovo o vecchio, risultano significativi segno di un'evidente impatto sulla variabile risposta (numero di non conformità).
Ai fini di una corretta selezione del modello si considera il test LRT riportato

Tabella 4.1: Stima modello Poisson e NB

	Poisson				**NB**			
	Est	s.e.	z-val.	p-val.	Est	s.e.	z-val.	p-val.
β_0	-2.11	0.33	-6.45	0.00	-2.09	0.84	-2.49	0.01
β_1	0.03	0.005	6.08	0.00	0.03	0.013	2.306	0.02
β_2	1.51	0.07	20.26	0.00	1.50	0.32	4.69	0.00
τ	–	–	–	–	0.185	0.0176	10.51	0.00
AIC	3330.6					2097.2		

Tabella 4.2: LRT test Poisson vs NB

	D.f.	Log-lik	D.f.	chi-value	p-value
Poisson	3	-1162.3			
NB	4	-1044.6	1	1235.4	0.00

nella tabella 4.2.

Il test, che confronta il modello di Poisson (ridotto) con quello Binomiale Negativo (completo), indica che il parametro τ è significativo. Tuttavia l'elevato numero di zeri presenti nei dati possono portare a stime distorte e fuorvianti; ciò suggerisce quindi l'adattamento di un modello zero-inflated, la cui stime vengono riportate nella tabella 4.3.

Per entrambi i modelli c'è una evidente significatività sia dei parametri che modellano i conteggi diversi da zero (β_1 e β_2) sia dei parametri che modellano la parte zero-inflated (γ_1 e γ_2), mentre il parametro di forma τ, che governa la sovra-dispersione, non risulta significativo come confermato dal test LRT riportato nella tabella 4.4

La non significatività del parametro di sovra-dispersione porta alla scelta di

Tabella 4.3: Stima modello zero-inflted

	ZIP				**ZINB**			
	Est	s.e.	z-val.	p-val.	Est	s.e.	z-val.	p-val.
β_0	-4.17	0.37	-11.39	0.00	-4.17	0.37	-11.39	0.00
β_1	0.084	0.0055	15.14	0.00	0.08	0.005	15.13	0.00
β_2	0.83	0.076	10.88	0.00	0.83	0.076	10.88	0.00
γ_0	-4.064	0.86	-4.70	0.00	-4.06	0.86	-4.70	0.00
γ_1	0.070	0.013	5.84	0.00	0.07	0.013	5.84	0.00
γ_1	-1.10	0.28	-3.88	0.00	-1.09	0.28	-3.88	0.00
τ	–	–	–	–	11.09	56.9	0.195	0.84

Tabella 4.4: LRT test ZIP vs ZINB

	D.f.	Log-lik	D.f.	chi-value	p-value
ZIP	6	-921.23			
ZINB	7	-921.24	1	0.00	0.97

Tabella 4.5: Test di Vuong: Poisson vs ZIP

	Vuong statistic		p-value
Raw	−15.027	ZIP > Poisson	0.00
AIC	−14.967	ZIP > Poisson	0.00
BIC	−14.821	ZIP > Poisson	0.00

un modello ZIP. L'effettiva presenza di un numero eccessivo di zeri nei dati viene confermata dal test di Vuong, riportato nella tabella 4.5.

4.3.3 Applicazione carta di controllo

Per poter valutare se il passaggio alle nuove metodologie ha portato o meno un aumento nelle performance del processo di verniciatura viene applicata una carta di controllo EWMA basata sulla stima del modello zero-inflated sopra discusso.

Il cambio nella metodologia di verniciatura avviene dall'817-esimo lotto in poi. Si considerano quindi le prime 600 osservazioni per la stima di un modello zero-inflated, includendo quindi solo la variabile umidità, per valutare se il processo operava in condizione stabili; verrà quindi simulata una fase 2, applicando la carta di controllo sequenzialmente dalla 601-esima osservazione, così da simulare i punti di forza nel caso in cui una schema di sorveglianza fosse già stato attivato prima del cambiamento del processo.

Sulla scelta della costante di lisciamento λ, un ragionevole compromesso consiste nel fissare $\lambda = 0.2$. Un valore di ARL in controllo ritenuto accettabile per questa applicazione è pari a 200.

Essendo che, in questo caso, i limiti di controllo di fase II non possono essere pre-calcolati, in quanto la distribuzione y_t del numero di non conformità dipenderà dal livello di umidità dell'ambiente, rilevato giorno per giorno, verranno applicati dei limiti dinamici di probabilità dinamici (si veda 4.2.1).

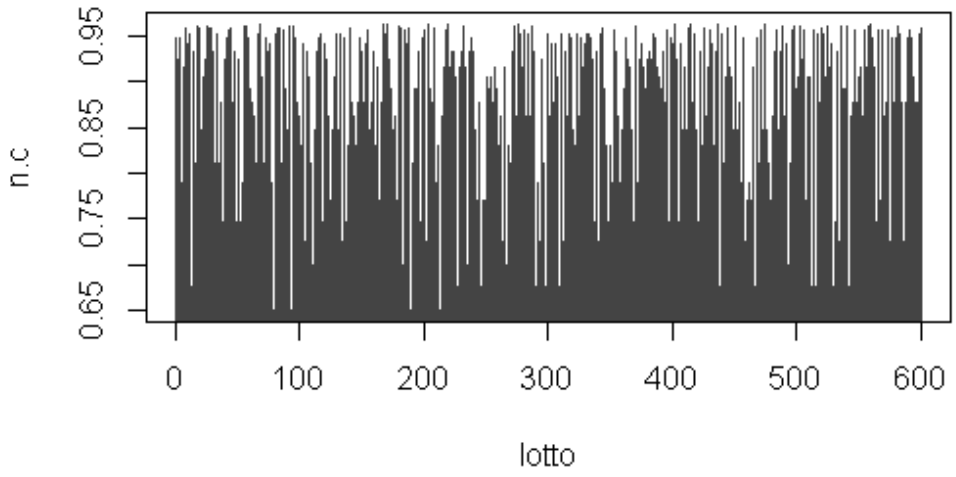

Figura 4.2: Distribuzione in controllo del n° di n.c.

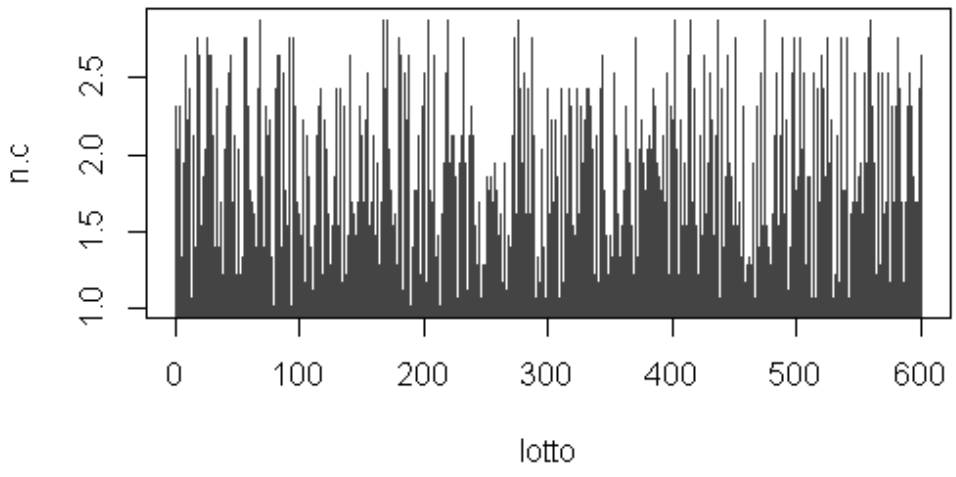

Figura 4.3: Deviazione standard del n° di n.c. in controllo

Considerato che la distribuzione di $W_t|W_1,\ldots,W_{t-1}$ analiticamente risulta difficile da calcolare, una possibile soluzione è quella di determinare i limiti di controllo via simulazione, generando in parallelo alla statistica di controllo W_t che calcoleremo dai dati, anche un numero elevato di statistiche W_t^* generate dalla distribuzione in controllo di W_t.
Un possibile algoritmo è il seguente:
1. *Inizializzazione*: Scegliere un numero $Nsim > 0$ e porre:

$$W_{0,1}^* = \cdots = W_{0,Nsim}^* = 0$$

2. *Al tempo t:* Simulare $Nsim$ determinazioni

$$y_{t,1}^* \ldots y_{t,Nsim}^*$$

dalla distribuzione in controllo di y_t (fig. 4.2).
3. *Calcolare* W_t^*

$$W_{t,i}^* = max\left[0, \lambda\left(\frac{y_{t,1}^* - (1-p_i)\lambda_i}{\sqrt{(1-p_i)(\lambda_i + p_i\lambda_i^2)}}\right) + (1-\lambda)W_{t_1}^*\right]$$

4. *Porre*:

$$L_t = \text{quantile empirico}\left(1 - \frac{1}{B}\right) \quad \text{di} \quad W_{t,1}^* \cdots = W_{t,Nsim}^*$$

5. *Infine* sostituire tutte le W_t^* maggiori di L_t con valori ricampionati casualmente dalle W_t^* minori o uguali a L_t.
N.B: Il passo (5) è necessario per imporre la condizione:

$$W_1 \leq L_1, W_2 \leq L_2, \ldots, W_{t-1} \leq L_{t-1}$$

così facendo tutte le traiettorie generate da W_t^* includeranno solo i valori che soddisfano quella condizione.
La carta di controllo EWMA con limiti di probabilità dinamici è riportata in fig. 4.4.
La carta di controllo, al netto degli allarmi al lotto 92, 93, e 94 (n.b: lotto 1 della carta di controllo corrisponde al 601-esimo lotto dei dati a disposizione)

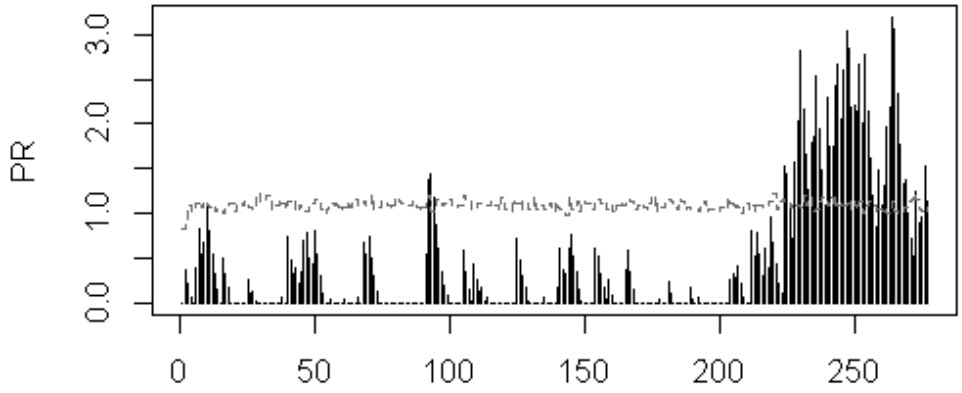

Figura 4.4: EWMA model-based: $\lambda = 0.2$, $ARL_0 = 200$

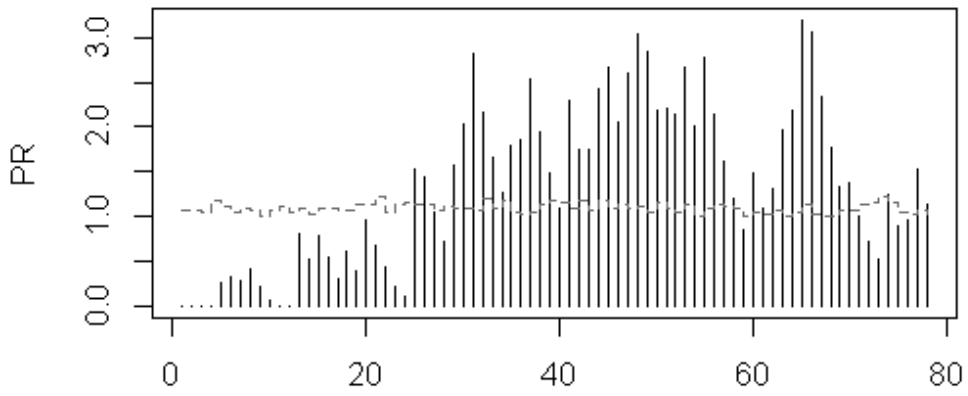

Figura 4.5: EWMA dall'800-esimo lotto in poi

che andrebbero investigati separatamente per valutare se sono stati o meno dei falsi allarmi, segnala il primo allarme all'osservazione 224 (che corrisponde all'824-esimo lotto di produzione).
Essendo che il nuovo processo di verniciatura è stato introdotto dall'818-esimo lotto in poi, se assumiamo $\tau = 818$ (istante incognito dove il processo è andato fuori controllo) il *Detection Delay* $= RL - \tau + 1$ è pari a 7, indice di una buona performance della carta di controllo. Notiamo subito che l'allarme segnalato non è un falso allarme in quanto la maggior parte delle osservazioni successive (fig 4.5) eccedono il limite di controllo superiore.
Concludiamo che la nuova metodologia applicata non ha portato nessun miglioramento del processo, al contrario ha di fatto aumentato il numero di non conformità.
Il vantaggio principale nel caso in cui una carta di controllo fosse già stata attiva è il tempo necessario nell'accorgersi del cambiamento: alla carta di controllo gli sono serviti 7 lotti di produzione per segnalare un allarme, i responsabili hanno bloccato la produzione all'877-esimo lotto, con una differenza di 53 lotti di produzione in più rispetto alla carta di controllo.

Conclusioni e raccomandazioni

Gli attuali processi di produzione sono integrati con sistemi di alta qualità a causa di un continuo progresso tecnologico. Questo continuo cambiamento porta sui mercati prodotti a difetti zero. Tuttavia, il monitoraggio e il controllo dei prodotti senza difetti sono diventati un compito impegnativo per i responsabili della qualità.
In generale, i prodotti con zero difetti sono ben modellati dalle distribuzioni zero-inflated e le carte di controllo basate su queste distribuzioni vengono utilizzate per diagnosticare qualsiasi brusco cambiamento nei processi ad alta qualità. Le distribuzioni ZIP e ZINB sono le distribuzioni più comuni e le carte di controllo basati su queste distribuzioni vengono spesso usate per monitorare tali processi. Solitamente nel progettare tali schemi di sorveglianza vengono rilevate anche alcune covariate; in questo caso la stima dei parametri avviene attraverso un modello lineare generalizzato e quindi vengono costruite carte di controllo model-based. Il capitolo 3 e 4 hanno messo in evidenza le migliori performance delle carte di controllo model-based.
Lo scopo principale del caso di studio considerato nell'ultimo capitolo, era quello di far emergere i vantaggi derivanti dall'utilizzo di un appropriato schema di sorveglianza (al netto che si tratti di un processo ad alto rendimento). Se i manager aziendali avessero fin da subito sviluppato e attivato uno schema di sorveglianza, come visto nell'applicazione, sarebbe emersa di gran lunga in anticipo la scarsa capacità della nuova metodologia di produrre con un numero di non conformità molto basso.
Infine, si tenga presente che lo schema proposto è risultato valido ed efficiente

considerando questo particolare processo produttivo e questi particolari dati. Si sconsiglia di applicare tale procedura senza prima aver condotto specifiche analisi sui dati a disposizione e sulle informazioni da ritenere significative.

Appendice

Verranno riportati in seguito i comandi necessari ad ottenere i risultati precedentemente illustrati.

```
library(pscl)
library(lmtest)
library(MASS)


#modello di Poisson
m.poi <- glm(NC ~ UH+NEW, family = "poisson")

Call:
glm(formula = NC ~ UH + NEW, family = "poisson")

Deviance Residuals:
Min       1Q   Median       3Q      Max
-3.3182  -1.3982  -1.2362   0.2856   6.6112

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.116594   0.328221  -6.449 1.13e-10 ***
UH           0.030792   0.005064   6.080 1.20e-09 ***
NEWTRUE      1.512871   0.074673  20.260  < 2e-16 ***
---
```

```
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2975.6  on 876  degrees of freedom
Residual deviance: 2609.0  on 874  degrees of freedom
AIC: 3330.6


#Modello Binomiale Negativo
m.nb <- glm.nb(NC ~ UH+NEW)

Call:
glm.nb(formula = NC ~ UH + NEW, init.theta = 0.18,
link = log)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.1241  -0.8247  -0.7784   0.1212   1.6914

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.09505    0.84129  -2.490   0.0128 *
UH           0.03048    0.01322   2.306   0.0211 *
NEWTRUE      1.50353    0.32058   4.690 2.73e-06 ***
---

(Dispersion parameter for Negative Binomial(0.185)
family taken to be 1)

Null deviance: 585.61  on 876  degrees of freedom
Residual deviance: 545.64  on 874  degrees of freedom
AIC: 2097.2

Number of Fisher Scoring iterations: 1
```

```
Theta:  0.1850
Std. Err.:  0.0176

2 x log-likelihood:  -2089.1740

#LRT - TEST; POISSON vs NB
lrtest(m.poi, m.nb)

## LRT - TEST
Model 1: NC ~ UH + NEW
Model 2: NC ~ UH + NEW
#Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -1662.3
2   4 -1044.6  1 1235.4  < 2.2e-16 ***

#Modello Zero-inflated Poisson
m.zip.completo <- zeroinfl(NC~UH+NEW|UH+NEW,data=d)

Call:
zeroinfl(formula = NC ~ UH + NEW | UH + NEW, data = d)

Pearson residuals:
Min      1Q  Median      3Q     Max
-1.0853 -0.5585 -0.4503  0.1697  5.6894

Count model coefficients (poisson with log link):
Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.172978   0.366423  -11.39   <2e-16 ***
UH           0.084292   0.005568   15.14   <2e-16 ***
NEWTRUE      0.833135   0.076554   10.88   <2e-16 ***
```

```
Zero-inflation model coefficients (binomial with logit link):
Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.06439    0.86366  -4.706 2.53e-06 ***
UH           0.07979    0.01366   5.841 5.20e-09 ***
NEWTRUE     -1.10000    0.28289  -3.888 0.000101 ***
---

Number of iterations in BFGS optimization: 10
Log-likelihood: -921.2 on 6 Df

# Modello Zero-inflted NB
m.zinb.completo <- zeroinfl(NC~UH+NEW|UH+NEW,data=d,
dist="negbin")

Call:
zeroinfl(formula = NC ~ UH + NEW | UH + NEW, data = d,
dist = "negbin")

Pearson residuals:
Min      1Q  Median      3Q     Max
-1.0853 -0.5585 -0.4503  0.1697  5.6893

Count model coefficients (negbin with log link):
Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.173018   0.366454 -11.388   <2e-16 ***
UH           0.084292   0.005569  15.136   <2e-16 ***
NEWTRUE      0.833139   0.076559  10.882   <2e-16 ***
Log(theta)  11.093266  56.901722   0.195    0.845

Zero-inflation model coefficients (binomial
with logit link):
Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.06437    0.86367  -4.706 2.53e-06 ***
```

```
UH           0.07979    0.01366   5.841 5.20e-09 ***
NEWTRUE     -1.09999    0.28289  -3.888 0.000101 ***
---

Theta = 65727.0521
Number of iterations in BFGS optimization: 10
Log-likelihood: -921.2 on 7 Df

#LRT TEST ZIP vs ZINB

Likelihood ratio test
Model 1: NC ~ UH + NEW | UH + NEW
Model 2: NC ~ UH + NEW | UH + NEW
#Df  LogLik Df Chisq Pr(>Chisq)
1   6 -921.23
2   7 -921.24  1 6e-04     0.9798

## VUONG TEST POI VS ZIP
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
null that the models are indistinguishible)
-------------------------------------------------------------
Vuong z-statistic             H_A    p-value
Raw                  -15.02797 model2 > model1 < 2.22e-16
AIC-corrected        -14.96713 model2 > model1 < 2.22e-16
BIC-corrected        -14.82184 model2 > model1 < 2.22e-16

# CARTA DI CONTROLLO EWMA MODEL-BASED

m.zip <- zeroinfl(NC~UH|UH, data=d[1:600,])

Call:
zeroinfl(formula = NC ~ UH | UH, data = d[1:600, ])
```

```
Pearson residuals:
Min      1Q  Median      3Q     Max
-0.6375 -0.5629 -0.4394  0.1038  5.5497

Count model coefficients (poisson with log link):
Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.114048   0.527581  -7.798 6.29e-15 ***
UH           0.083465   0.008112  10.289  < 2e-16 ***

Zero-inflation model coefficients (binomial with
logit link):
Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.92328    1.10251  -4.466 7.99e-06 ***
UH           0.09318    0.01744   5.342 9.17e-08 ***
---

Number of iterations in BFGS optimization: 8
Log-likelihood: -598.1 on 4 Df

 p.poisson <- coef(m.zip)[1:2]
count_(Intercept)          count_UH
-4.11404837        0.0834647

p.logit <- coef(m.zip)[3:4]
zero_(Intercept)          zero_UH
-4.9232802        0.0931806

p.zero <- 1/(1+exp(-p.logit[1]-p.logit[2]*UH))

l.poisson <- exp(p.poisson[1]+p.poisson[2]*UH)

mt <- (1-p.zero)*l.poisson
```

```
st <- sqrt((1-p.zero)*(l.poisson+p.zero*l.poisson^2))
r <- (NC-mt)/st

lambda <- 0.2
EWMA <- 0
rzip <- function(n, p, l) {
           (runif(n)>p)*rpois(n,l)
         }
 B <- 200
 N <- NROW(r)
 W <- L <- double(N)
 Nsim <- round(20*B)
 Wstar <- double(Nsim)

 #SIMULAZIONE FASE II
for (i in seq_along(r)) {
  EWMA <- max(0, (1-lambda)*EWMA+lambda*r[i])
  Rstar <- (rzip(Nsim, p.zero[i], l.poisson[i])-mt[i])/st[i]
  Wstar <- pmax(0, (1-lambda)*Wstar+lambda*Rstar)
  W[i] <- EWMA
  L[i] <- quantile(Wstar, 1-1/B)
  idx <- which(Wstar > L[i])
  Wstar[idx] <- sample(Wstar[Wstar<=L[i]], length(idx))
 }
matplot(cbind(W, L), type = c("h", "s"), ylab = "PR")
```

Bibliografia

[1] V. Alevizakos e C. Koukouvinos. «Monitoring of zero-inflated Poisson processes with EWMA and DEWMA control charts.» In: *Qual Reliab Engng Int.* 36 (2020), 88:111.

[2] S. Asmussen, O. Nerman e M. Olsson. «Fitting Phase-Type Distributions via the EM Algorithm.» In: *Scandinavian Journal of Statistics* 23:4 (1996), pp. 419–441.

[3] A. Fatahi et al. «Zero inflated Poisson EWMA control chart for monitoring rare health-related». In: *J Mech Med Biol.* 12:04 (2012), pp. 12500651–125006514.

[4] AM. Garay et al. «On estimation and influence diagnostics for zero-inflated negative binomial regression models.» In: *Comput Stat Data* 55:3 (2011), pp. 1304–1318.

[5] D. A. Garvin. «Competing on eight dimensions of quality.» In: *Harvard Business Review* 65:6 (1987), pp. 101–109.

[6] D. Lambert. «Zero-inflated Poisson regression, with an application to defects in manufacturing.» In: *Dent Tech.* 34 (1992), pp. 1–14.

[7] X. Liu et al. «Simulating comparisons of different computing algorithms fitting zero-inflated Poisson models for zero abundant counts.» In: *J Stat Comput Simul.* 87:13 (2017), 2609:2621.

[8] T. Mahmood. «Generalized linear model based monitoring methods for high-yield processes.» In: *Qual Reliab Engng Int.* 36 (2020), pp. 1570–1591.

[9] M. Maleky et al. ≪The effect of parameter estimation on phase II monitoring of poisson regression profiles.≫ In: *Commun Stat-Simul C.* 48 (2019), pp. 1964–1978.

[10] P. McCullagh e J. Nelder. *Generalised linear models.* A cura di Chapman e Hall. 2nd. London, 1983.

[11] X Shen et al. ≪Monitoring Poisson Count Data with Probability Control Limits when Sample Sizes are Time Varying≫. In: *Navel Research Logistics* 60 (2013), 625:636.

[12] GD. Walter. ≪Using Poisson class regression to analyze count data in correctional and forensic psychology: a relatively old solution to a relatively new problem.≫ In: *Crim Justice Behav.* 34:12 (2007), pp. 1659–1674.

[13] JM. Williamson et al. ≪Power calculations for ZIP and ZINB models.≫ In: *Data Sci J.* 5 (2007), pp. 519–534.

[14] A. Jensen Willis et al. ≪Effects of Parameter Estimation on Control Chart Properties:A Literature Review≫. In: *Journal of Quality Technology* 38:4 (2006), pp. 349–364.

[15] M. Xie e T. Goh. ≪Spc of a near zero-defect process subject to random shocks.≫ In: *Qual Reliab Eng Int.* 9:2 (1993), 89:93.

[16] M. Xie, TN. Goh e P. Ranjan. ≪On the estimation error in zero-inflated Poisson model for process control.≫ In: *Int J Reliab Qual Saf Eng.* 10:02 (2003), pp. 159–169.

[17] M. Xie, M. Xie e T. Goh. ≪Control charts for processes subject to random shocks.≫ In: *Qual Reliab Eng Int.* 11:5 (1995), pp. 355–360.