

# Rapporti tecnici

## INGV

**Algoritmi di Change Point Analysis delle  
serie temporali: il tremore vulcanico  
registrato sul vulcano Etna  
come caso di studio**

# 317



## **Direttore Responsabile**

Stefano Gresta

## **Editorial Board**

Luigi Cucci - Editor in Chief (INGV - RM1)

Andrea Tertulliani (INGV - RM1)

Nicola Pagliuca (INGV - RM1)

Umberto Sciacca (INGV - RM2)

Alessandro Settimi (INGV - RM2)

Aldo Winkler (INGV - RM2)

Salvatore Stramondo (INGV - CNT)

Milena Moretti (INGV - CNT)

Gaetano Zonno (INGV - MI)

Viviana Castelli (INGV - BO)

Antonio Guarnieri (INGV - BO)

Mario Castellano (INGV - NA)

Mauro Di Vito (INGV - NA)

Raffaele Azzaro (INGV - CT)

Rosa Anna Corsaro (INGV - CT)

Mario Mattia (INGV - CT)

Marcello Liotta (INGV - PA)

## **Segreteria di Redazione**

Francesca Di Stefano - Referente

Rossella Celi

Tel. +39 06 51860068

redazionecen@ingv.it

in collaborazione con:

Barbara Angioni (RM1)



# Rapporti tecnici INGV

## ALGORITMI DI CHANGE POINT ANALYSIS DELLE SERIE TEMPORALI: IL TREMORE VULCANICO REGISTRATO SUL VULCANO ETNA COME CASO DI STUDIO

Carmelo Cassisi<sup>1</sup>, Salvatore Fiumara<sup>2</sup>, Andrea Cannata<sup>1</sup>, Placido Montalto<sup>1</sup>,  
Marco Aliotta<sup>1</sup>, Marco Scarpa<sup>2</sup>

<sup>1</sup>INGV (Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Catania - Osservatorio Etneo)

<sup>2</sup>Università degli Studi di Messina (D.I.C.I.E.A.M.A.)

# 317



## Indice

Introduzione	7
1. Metodologia e algoritmo	7
1.1 CUSUM	7
1.2 One Pass Concept Change Detection for Data Streams	10
2. Applicazione della CPA alle serie temporali del tremore vulcanico	11
2.1 Criterio di associazione tra tempo di <i>detection</i> e tempo di attività reale	13
2.2 Test CUSUM	14
2.3 Test One Pass Concept Change Detection for Data Streams	16
2.4 Esperimento con campionamento al minuto per l'analisi in real-time	19
3. Discussione e conclusioni	21
Ringraziamenti	23
Bibliografia	23



## Introduzione

Il presente studio ha lo scopo di illustrare l'applicazione della metodologia della *change point analysis* (CPA) alle serie temporali relative al tremore vulcanico registrato sul vulcano Etna, e impiegate per le attività di sorveglianza e monitoraggio.

La CPA consiste nella progettazione di tecniche per il rilevamento automatico (*detection*) dei cambiamenti di stato durante l'osservazione di un processo. L'analisi tiene conto di due caratteristiche fondamentali che sono la tempestività dell'individuazione e la limitata occorrenza di falsi allarmi [Tartakovsky et al., 2014], la cui quantificazione viene definita in base al contesto di applicazione.

La classificazione circa lo stato del processo si basa su delle caratteristiche (misurabili) che nel tempo formano una sequenza di osservazioni (serie temporali, se si tratta di misure quantitative). Nella maggior parte delle applicazioni, i sensori costituiscono un ruolo centrale perché sono gli agenti incaricati alla produzione delle osservazioni temporali. Nel caso della CPA, l'oggetto di maggiore interesse risulta la *detection* del cambiamento di stato da "normale" ad "anomalo". Fintanto che la sequenza di osservazioni suggerisce che il processo risulti in uno stato di normalità, si lascia continuare; se invece si ritiene che qualcosa sia cambiato, l'obiettivo è quello di rilevare il cambiamento in maniera tempestiva, in modo da concedere il tempo adeguato, a chi di competenza, per la valutazione di un'azione, come ad esempio la sospensione del processo o l'innescio di un allarme.

Storicamente, la CPA iniziò ad emergere intorno agli anni '30, nel dopoguerra, in riferimento al controllo della produzione industriale, attraverso lo studio di un tipo di grafico noto come *carta di controllo* o *Carta di Shewhart delle medie* [Shewhart, 1931]. Questo tipo di grafico riporta semplicemente una linea centrale che rappresenta il valore attorno al quale le misurazioni della caratteristica del processo dovrebbero presentarsi nelle condizioni naturali. Successivamente, l'applicazione della CPA si è estesa su diversi campi, nella maggior parte dei casi per lo sviluppo delle reti di sensori e delle relative tecnologie, come ad esempio: nella sorveglianza dell'ambiente [e.g., Ortner and Nehorai, 2007], nell'ambito della meteorologia per l'analisi delle variazioni dei dati climatici [e.g., Tomozeiu et al., 2000; Reeves et al., 2009], nell'analisi dei segnali biomedici [e.g. Gustafson et al., 1978], nell'analisi delle serie temporali dei mercati finanziari [e.g. Andreou and Ghysels, 2002], nel rilevamento delle intrusioni in reti di computer e sistemi di sicurezza [e.g. Tartakovsky et al., 2013], o nel monitoraggio del traffico stradale [e.g. Grossman et al., 2005].

In questa sede si applica la CPA per lo studio e l'identificazione di cambiamenti dello stato del vulcano sulla base di variazioni nelle serie temporali sismo-vulcaniche. Per l'analisi di interesse, le metodologie utilizzate sono il calcolo della *cumulative sum* (CUSUM) descritto da Taylor [2000] e la "*One Pass Concept Change Detection for Data Streams*" di Sakthithasan et al. [2013], di seguito riferita come *SeqDrift*. La prima tecnica si presta meglio per l'analisi storica delle serie temporali, in quanto si basa su informazioni calcolabili solo a posteriori (come ad es. la media della serie); mentre la seconda è stata implementata in modo da analizzare in maniera efficace ed efficiente dati di tipo *stream* ovvero dati acquisiti in *real-time* e *near real-time*, attraverso il concetto di finestra mobile (*sliding window*). Entrambe le tecniche appartengono alla classe dei metodi non-supervisionati.

Di seguito verranno proposti alcuni casi di studio, riguardanti l'analisi di serie temporali di RMS (*Root Mean Square*) calcolate sul segnale sismico allo scopo di individuare i cambiamenti da uno stato di normalità (*steady state*) ad uno stato di anomalia. Sarà prestata particolare attenzione all'analisi sperimentale di *SeqDrift*, in quanto ha fornito risultati più attendibili ai fini del monitoraggio vulcanico.

## 1. Metodologia e Algoritmi

In questo paragrafo sono descritte le tecniche che nei paragrafi successivi saranno applicate sulle serie temporali relative al tremore vulcanico acquisito sul vulcano Etna.

### 1.1 CUSUM

Il primo dei due metodi implementati è quello trattato da Taylor [2000]. La procedura restituisce un valore di attendibilità, specificato da un certo livello di confidenza, a garanzia della vero-somiglianza del *change point* individuato.

Il primo passo eseguito dall'algoritmo per la definizione dei *change point* consiste nel calcolo della CUSUM. Definito un set di  $n$  valori  $x_1, x_2, \dots, x_n$  di una serie temporale  $T$ , il calcolo della CUSUM al passo  $i$ -

esimo,  $S_i$ , è definito come la differenza tra il valore  $x_i$  della serie temporale ed il valore medio  $\bar{X}$  di  $T$ , sommato al valore della CUSUM al passo  $i-1$ :

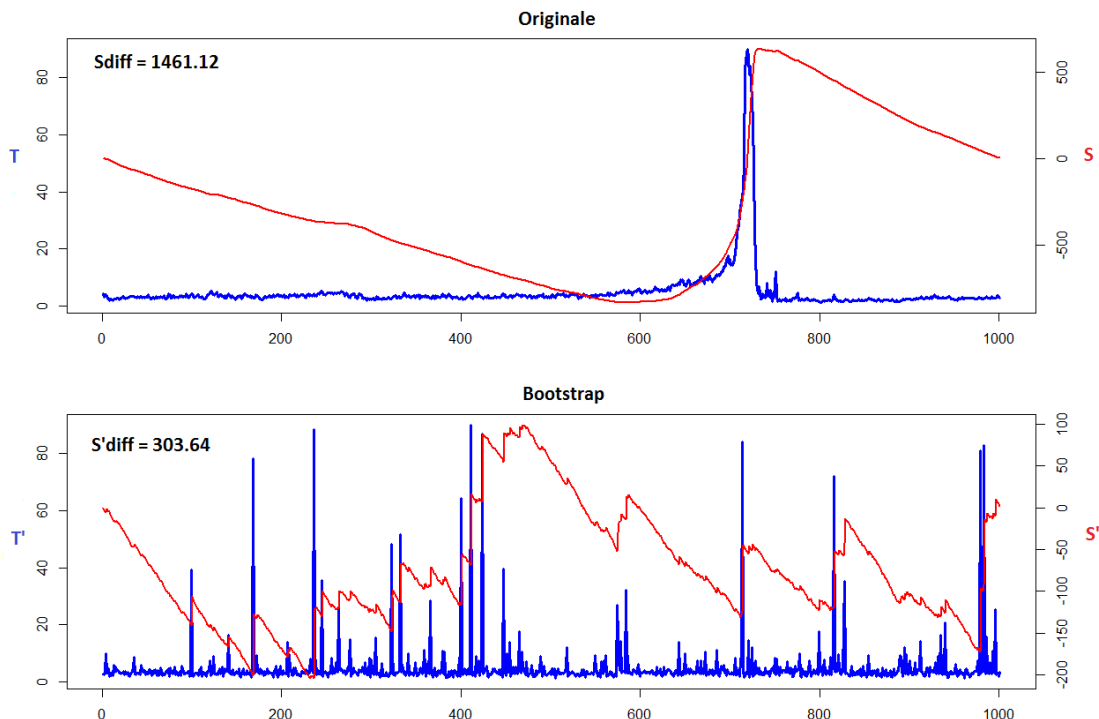
$$S_i = S_{i-1} + (x_i - \bar{X}) \quad (1)$$

dove il primo valore ( $S_0$ ) viene impostato a zero. Questa operazione viene eseguita dall'algorithm in modo iterativo per l'intero intervallo del *dataset* considerato ( $0 < i \leq n$ ).

Il risultato di questa operazione sarà una nuova serie temporale  $S$  che contiene i diversi  $S_i$  calcolati dall'applicazione della CUSUM. Alcune importanti proprietà di questa serie sono:

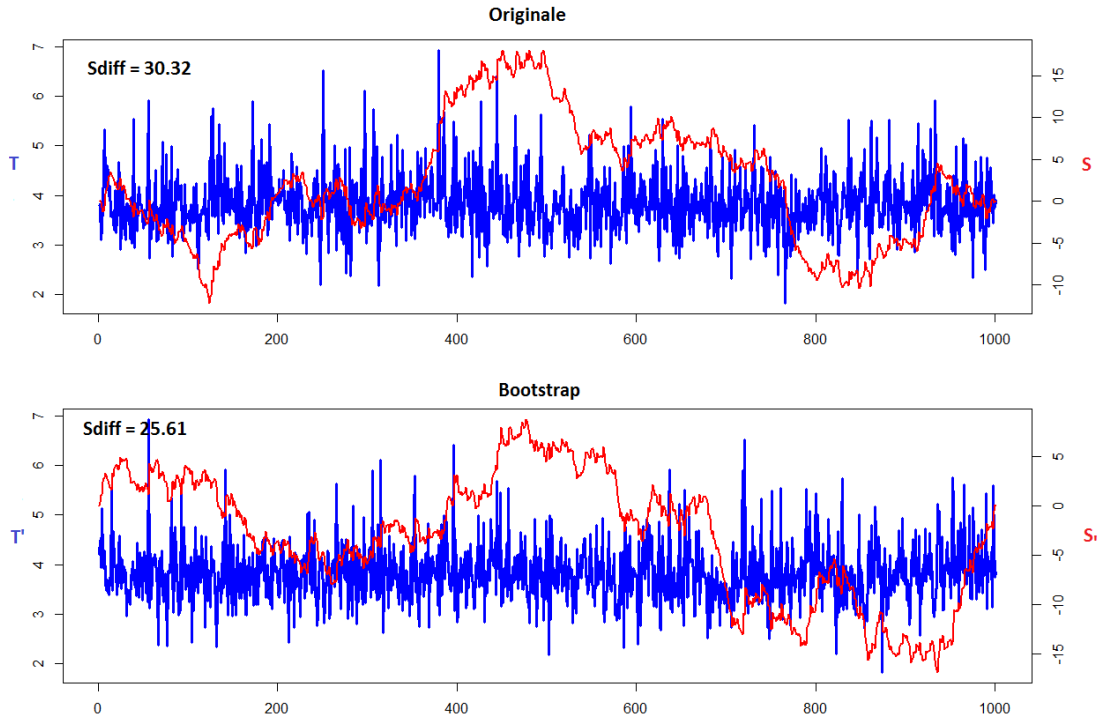
- il valore di  $S_n$  risulta pari a zero come per  $S_0$ ;
- se la pendenza della serie in un definito segmento risulta discendente, significa che i valori all'interno del segmento tendono ad essere al di sotto della media  $\bar{X}$ ;
- se la pendenza della serie in un definito segmento risulta ascendente, significa che i valori all'interno del segmento tendono ad essere al di sopra della media  $\bar{X}$ ;
- eventuali picchi (o cambiamenti di direzione) possono essere interpretati come un cambiamento locale della pendenza e quindi del trend.

Per valutare quanto sia significativo effettuare una ricerca dei *change point* all'interno della serie, l'algorithm prevede l'applicazione di un processo di *bootstrap* iterativo [Efron and Tibshirani, 1993] con il ricalcolo della CUSUM, per ottenere un livello di confidenza. Viene quindi definito un valore  $S_{diff}$  da utilizzare per la determinazione del livello di confidenza.  $S_{diff}$  è ricavato dalla differenza tra il valore massimo  $S_{max}$  e il minimo  $S_{min}$  della  $S$  calcolata sulla serie originale (vedi Figura 1). La probabilità che ci sia un cambiamento significativo del trend nella serie considerata è proporzionale al valore della  $S_{diff}$ . Si noti a questo proposito come si comporta l'andamento della  $S$  nel caso in cui siano presenti cambiamenti significativi all'interno di  $T$  (Figura 1) e quando non sono presenti (Figura 2).



**Figura 1.** Nel grafico in alto è illustrata la serie  $T$  originale in blu (dove è presente un cambiamento di trend in corrispondenza degli indici compresi tra 600 e 800), con la relativa CUSUM  $S$  in rosso. In basso, è mostrato un riordinamento dei valori di  $T$ ,  $T'$  (in blu), con la relativa CUSUM  $S'$  (in rosso). Si può notare come il valore della  $S_{diff}$  calcolata su  $T$  rispetto a  $S'_{diff}$  calcolata su  $T'$  sia molto più grande.





**Figura 2.** Nel grafico in alto è mostrata la serie  $T$  originale in blu (dove non è presente un cambiamento di trend significativo), con la relativa CUSUM  $S$  in rosso. In basso è mostrato un ri-campionamento di  $T$ ,  $T'$  (in blu) con la relativa CUSUM  $S'$  (in rosso). Si può notare come la differenza tra la  $S_{diff}$  calcolata su  $T$  e la  $S'_{diff}$  calcolata su  $T'$  risulti molto minore rispetto alla differenza di questi due valori presenti nell'esempio di Figura 1.

Per ogni permutazione  $T'$  della serie ad ogni iterazione del processo di *bootstrap*, viene calcolata la relativa  $S'_{diff}$ . Se la  $S'_{diff}$  risulta sempre minore della  $S_{diff}$ , i picchi presenti sulla serie originale possono essere considerati significativi. Sia quindi  $B$  il numero di *bootstrap* e  $C$  il numero di volte in cui la  $S'_{diff}$  della permutazione risulta minore della  $S_{diff}$  calcolata sulla serie originale, il livello di confidenza risulta:

$$Confidence\ Level = 100 \frac{C}{B} \% \quad (2)$$

Una volta stabilita una soglia minima per il livello di confidenza, secondo la quale accettare l'ipotesi che esista un *change point*, viene stabilito quale stimatore utilizzare per la ricerca della sua posizione. Gli stimatori considerati sono due: il primo deriva direttamente dal calcolo della CUSUM, mentre il secondo dal MSE (*Mean Square Error*).

Nel primo caso l'indice  $m$  della serie corrispondente al *change point* risulta  $S_m$ , ovvero il punto più lontano dallo zero nella serie  $S$ :

$$|S_m| = \max_{i=0, \dots, n} |S_i| \quad (3)$$

L'MSE invece viene calcolato dividendo la serie temporale in due segmenti. Data una serie di punti  $x_1, x_2, \dots, x_n$ , e un indice  $m$  ( $1 < m < n$ ), viene eseguito uno split in due segmenti che vanno da  $x_1$  a  $x_m$  e da  $x_{m+1}$  a  $x_n$ . Il valore di  $m$  che minimizza l'MSE è il valore che indica il punto del *change point* e  $m+1$  il punto successivo al punto di cambio. Pertanto il valore del MSE è definito come:

$$MSE_{(m)} = \sum_{i=1}^m (x_i - \bar{X}_1)^2 + \sum_{i=m+1}^n (x_i - \bar{X}_2)^2 \quad (4)$$

$$\bar{X}_1 = \frac{\sum_{i=1}^m x_i}{m} ; \bar{X}_2 = \frac{\sum_{i=m+1}^n x_i}{n - m}$$

Metodo alternativo all'MSE è dato dall'applicazione del metodo della regressione [e.g. Montalto et al., 2012]. In questo caso l'algoritmo esegue una combinazione iterativa e calcola il punto di split tra due rette di regressione corrispondente al valore minimo tra gli scarti residui ottenuti dalle varie combinazioni.

Il calcolo per l'individuazione dei *change point* avviene fissando a priori il valore di alcuni parametri: il livello minimo di confidenza per accettare la significatività dei *change point*, il numero di *bootstrap* da eseguire, la scelta dello stimatore da utilizzare, e il numero limite di *change point* da individuare. Riguardo a quest'ultimo parametro, si precisa che la suddetta tecnica restituisce un solo  $m$  relativo al punto candidato come *change point* migliore. Per ottenere più *change point*, dopo aver estratto il primo punto all'indice  $m$ , si applica ricorsivamente la tecnica scelta sui due segmenti della serie che vanno da  $x_1$  a  $x_m$  e da  $x_{m+1}$  a  $x_n$ .

## 1.2 One Pass Concept Change Detection for Data Streams

Il secondo approccio valutato è il metodo *SeqDrift* [Sakthithasan et al., 2013]. Si tratta di un metodo di *Data Stream Mining* per l'analisi delle serie temporali in *real-time*. L'algoritmo esegue il processo di calcolo mettendo a confronto due finestre mobili (o blocchi) corrispondenti a due segmenti consecutivi della serie temporale: il confine (*boundary*) tra le due finestre corrisponde al punto di *change point*.

Il concetto di base è rappresentato dall'applicazione dei test statistici. Siano dati due campioni di valori  $B_1 = x_1, x_2, \dots, x_m$  e  $B_2 = x_{m+1}, x_{m+2}, \dots, x_n$  corrispondenti ai due blocchi analizzati dallo streaming, e siano  $\bar{X}_1$  e  $\bar{X}_2$  le rispettive medie. Il problema della *change point analysis* consiste nel verificare se l'ipotesi nulla  $H_0$  sia  $H_0: \bar{X}_1 = \bar{X}_2$  contro l'alternativa  $H_1: \bar{X}_1 \neq \bar{X}_2$ . Il test serve a verificare se i due campioni provengono dalla stessa distribuzione. In questo caso si accetta l'ipotesi nulla  $H_0$ , nel caso contrario i due campioni provengono da distribuzioni diverse accettando  $H_1$ . In quest'ottica è importante il risultato ottenuto dal test di accettare o meno l'ipotesi nulla: se in modo errato si accetta  $H_0$  quando si verifica un *change point*, allora è stato riscontrato un vero negativo; contrariamente, se si accetta l'ipotesi alternativa nel caso in cui non sia verificato un *change point*, allora il punto riscontrato è un caso di falso positivo. Accettare  $H_1$  significa che si è verificato un cambiamento. Questo evento è rappresentato dalla probabilità:

$$Pr ( (|\bar{X}_1 - \bar{X}_2|) \geq \varepsilon ) > \delta \quad (5)$$

dove  $\delta$  è un parametro definito in un intervallo  $[0,1]$  per la definizione del massimo ammissibile dei falsi positivi, ovvero il livello di confidenza minimo stabilito per il nostro test. Questa metodologia non impone l'appartenenza dei dati trattati ad una particolare distribuzione statistica e permette quindi di non dover affrontare il problema della scelta del valore  $\varepsilon$ ; questo viene invece ottenuto tramite la disuguaglianza di Bernstein [Sakthithasan et al., 2013]:

$$\varepsilon = \frac{2}{3b} \left\{ p + \sqrt{p^2 + 18\sigma_s^2 bp} \right\} \quad (6)$$

dove  $\sigma_s^2$  rappresenta la varianza dei due blocchi della serie analizzati,  $p = \ln\left(\frac{4}{\delta}\right)$  e  $b$  l'ampiezza della finestra (chiamato anche *blockSize*). Quanto appena detto implica che l'algoritmo non ha bisogno di una fase di apprendimento poiché  $\varepsilon$  viene ricalcolato di volta in volta con il passare del tempo, adattandosi allo streaming di dati.

Siano  $B_1$  e  $B_2$  due blocchi di lunghezza fissata al valore *sampleSize* (lunghezza del blocco da prendere dallo streaming inizialmente pari a *blockSize* ma, come vedremo in seguito, può diventare un suo multiplo) numerati in base all'ordine di arrivo dallo streaming di dati. La prima operazione, che esegue l'algoritmo per l'individuazione dei punti di cambio, è il calcolo del valore  $\varepsilon$ . Sono tre i livelli di cambiamento contemplati, il *warning*, il *change* e l'*internal* (sempre relativamente al livello di confidenza scelto come soglia). Essi possono essere intesi come dei livelli di diverso grado di importanza:

1. *Warning*: pre-avvisa l'occorrenza di un possibile *change point*. La media dei due blocchi considerati risulta leggermente diversa rispetto all'andamento corrente della serie.
2. *Change*: individuazione del *change point* (riconoscimento di una anomalia). La media dei due blocchi considerati risulta molto diversa rispetto all'andamento corrente della serie.
3. *Internal*: stesso caso del *change*, ma la media dei valori del blocco  $B_1$  risulta maggiore di quella del blocco  $B_2$ .

Per la definizione di questi livelli vengono stabiliti due diversi valori  $\varepsilon$ -warning e  $\varepsilon$ -change, con  $\varepsilon$ -warning <  $\varepsilon$ -change, da utilizzare al posto di  $\varepsilon$  nell'Eq. 5. L'algoritmo opera come segue. Si supponga che al tempo  $t_2$  si abbiano i due blocchi  $B_1$  e  $B_2$ . Il blocco  $B_1$  è da considerare come il blocco di riferimento per l'individuazione del *change* la cui posizione viene valutata nel confine  $t_1$  tra  $B_1$  e  $B_2$ , con  $t_1 < t_2$  e  $t_2 - t_1 = \text{sampleSize}$ .

Se l'ipotesi  $H_1$  viene rigettata sia per l' $\varepsilon$ -warning, che per l' $\varepsilon$ -change, allora i due blocchi vengono concatenati per formare un nuovo blocco di  $B_{1,2}$  e si reitera il nuovo test utilizzando  $B_{1,2}$  e  $B_3$  (che è il nuovo blocco proveniente dallo streaming sempre di lunghezza *sampleSize* al tempo  $t_3$ ). Questo procedimento verrà eseguito fino al tempo  $t_n$ , quando non si accetta l'ipotesi  $H_1$  rispetto alla  $\varepsilon$ -warning o alla  $\varepsilon$ -change:

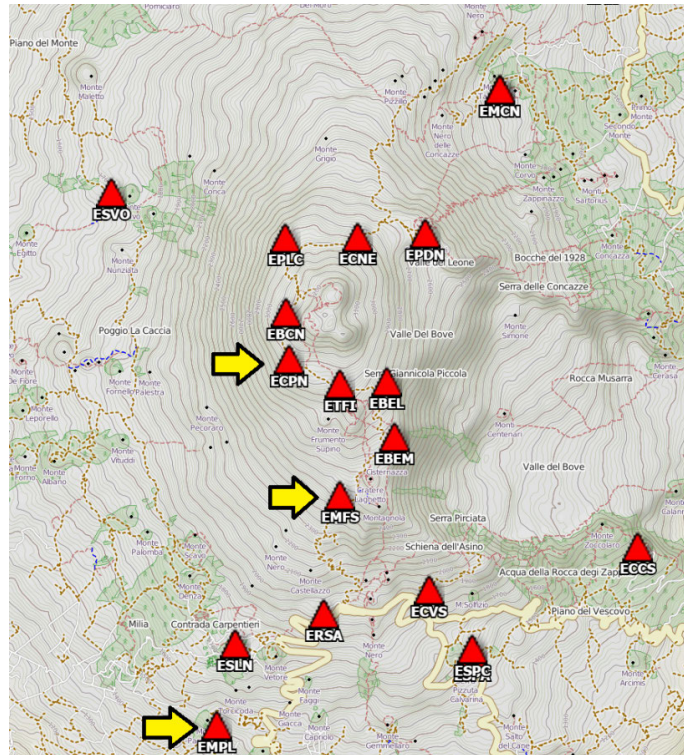
1. Se si accetta l'ipotesi  $H_1$  rispetto alla  $\varepsilon$ -warning, allora il tipo di cambio è un *warning*: in questo caso la finestra  $B_{1,\dots,n-1}$  rimarrà inalterata, mentre sarà  $B_n$  a includere il nuovo blocco proveniente dallo streaming  $B_{n+1}$  al tempo  $t_{n+1}$ , formando il nuovo blocco  $B_{n,n+1}$ . Si raddoppia di conseguenza il valore di *sampleSize* per le iterazioni successive.
2. Se si accetta l'ipotesi  $H_1$  rispetto alla  $\varepsilon$ -change, allora viene individuato un *change* o un *internal*. In questo caso  $B_n$  andrà a sostituire  $B_{1,\dots,n-1}$ , divenendo il nuovo punto di riferimento per l'individuazione del *change point*, e al tempo  $t_{n+1}$  si richiede dallo streaming un nuovo blocco  $B_{n+1}$  di dimensione *blockSize* (si ritorna alle condizioni iniziali).

In tutti i casi in cui uno o entrambi i blocchi utilizzati per il test aumentano di grandezza, il calcolo del valore della media viene effettuato prendendo sempre un numero fissato di campioni casuali da entrambi i blocchi (per l'esattezza pari alla minima lunghezza tra i due). Come riportato in [Sakthithasan et al., 2013] l'uso del campionamento casuale permette di migliorare l'efficienza della computazione pur mantenendone la robustezza. In ogni caso, sempre per motivi di efficienza, viene stabilito un limite massimo della lunghezza dei blocchi.

Come descritto in Sakthithasan et al. [2013], particolare attenzione deve essere rivolta all'impostazione del parametro relativo alla lunghezza iniziale dei blocchi (parametro *blockSize*), in quanto permette di controllare la sensibilità all'individuazione dei cambiamenti e il ritardo che si vuole ammettere per l'individuazione del *change point*.

## 2. Applicazione della CPA alle serie temporali del tremore vulcanico

I dati analizzati in questo lavoro sono relativi al segnale sismico proveniente dalla rete sismica permanente a banda larga, le cui stazioni sono poste ad una distanza che varia tra 1 e 9 km dal centro dell'area sommitale, che acquisisce in *real-time* ad una frequenza di campionamento di 100 Hz. In particolare si è scelto di analizzare la componente verticale del moto del suolo acquisita a tre stazioni collocate a differenti quote (Figura 3 – triangoli segnalati da frecce gialle): 1) ECPN (~3000 m s.l.m.); 2) EMFS (~2500 m s.l.m.); 3) EMPL (~1500 m s.l.m). Per le analisi proposte, si è scelto di utilizzare la serie temporale dell'RMS (*Root Mean Square*), calcolata su finestre mobili di 10 minuti sul segnale filtrato nella banda 0.2–5.5 Hz, nella quale è presente la maggior parte dell'energia del tremore vulcanico dell'Etna [e.g. Cannata et al., 2008].



**Figura 3.** Mappa dell’Etna con la posizione di alcune delle stazione sismiche che fanno parte della rete sismica permanente gestita dall’INGV, Osservatorio Etneo (triangoli rossi). Le frecce gialle indicano le 3 stazioni, i cui segnali sono stati utilizzati in questo lavoro.

Le caratteristiche più interessanti del tremore vulcanico dell’Etna sono: i) la sua persistenza, che lo rende un segnale praticamente continuo [e.g. Cannata et al., 2013a]; ii) la stretta relazione tra le variazioni delle caratteristiche del tremore (energia, caratteristiche spettrali, campo d’onda, posizione sorgente) e i cambiamenti dell’attività vulcanica [e.g. Alparone et al., 2003; Cannata et al., 2008, 2013a]. È stato in particolare evidenziato come l’attività esplosiva sia accompagnata da incrementi dell’ampiezza del tremore vulcanico [e.g. Alparone et al., 2003; Cannata et al., 2008]. Sulla base della caratteristica (i) sopra descritta, le variazioni dell’RMS del segnale sismico, registrato da stazioni installate sull’edificio vulcanico, sono dovute nella maggior parte dei casi a variazioni del tremore vulcanico. Per testare gli algoritmi si è preso come anno di riferimento il 2011, durante il quale si sono verificati 18 episodi di fontana di lava. Obiettivo dell’analisi è stato capire, considerando un *dataset* di variazioni di attività vulcanica, se entrambe le metodologie riescono ad intercettare in maniera automatica le variazioni e con quali tempistiche. Questo tipo di algoritmi viene valutato in base ai seguenti aspetti [Sebastiao and Gama, 2009]:

1. Capacità di individuare ogni cambiamento e con quale tempestività;
2. Numero di cambi non individuati;
3. Tasso di falsi allarmi, ovvero cambiamenti individuati in condizioni di stazionarietà.

Per quanto concerne il *dataset* di variazioni di attività vulcanica, data la relazione esistente tra le variazioni dell’ampiezza del tremore vulcanico e le fenomenologie vulcaniche di tipo esplosivo, si è fatto riferimento ai tempi dei 18 episodi di attività parossistica descritti nei rapporti prodotti dallo staff INGV della Sezione di Catania - Osservatorio Etneo [INGV staff, 2011]. Tali episodi presentano una comune evoluzione, in cui può essere distinta [e.g. Behncke et al., 2014; Viccaro et al., 2014]: i) una fase iniziale Stromboliana, caratterizzata da un graduale aumento nel tasso di accadimento ed intensità delle esplosioni, a cui corrisponde un graduale incremento dell’ampiezza del tremore vulcanico; ii) una fase parossistica, caratterizzata da attività di fontana di lava, durante la quale si raggiungono i valori massimi di energia del tremore; iii) una fase finale, in cui l’attività di fontana gradualmente viene sostituita da attività Stromboliana che termina nella maggior parte dei casi in breve tempo, accompagnata da diminuzioni generalmente veloci dell’ampiezza del tremore. I tempi delle fasi (i) e (ii) dei 18 episodi di fontana di lava del 2011 sono riportati nelle Tabelle 1 e

2, rispettivamente. È possibile notare come le fasi (i) siano caratterizzate da una maggiore variabilità della durata (da ~1 ad oltre 50 ore), rispetto alle fasi (ii) (da ~50 minuti a ~5 ore).

Riguardo al primo aspetto citato sopra (capacità di individuare ogni cambiamento e con quale tempestività), si è calcolato il numero di *detection* rispetto al numero effettivo di attività reali (veri positivi, di seguito indicati con l'acronimo VP) e per ognuno di essi si è calcolato il tempo di anticipo o ritardo. Riguardo al secondo aspetto (numero di cambi non individuati), si è conteggiato il numero di attività reali il cui periodo non contiene nessuna *detection* (veri negativi, di seguito indicati con l'acronimo VN); mentre per il terzo (tasso di falsi allarmi, ovvero cambiamenti individuati in condizioni di stazionarietà) si è calcolato il numero di *detection* non rientranti nei periodi di attività reali (falsi positivi, di seguito indicati con l'acronimo FP).

Il codice utilizzato per gli esperimenti è stato scritto in R [R Development Core Team, 2008], un linguaggio di programmazione, nonché un ambiente software per il calcolo statistico e per la produzione di grafici. Il linguaggio R è largamente utilizzato per lo sviluppo di analisi statistiche e di tecniche di *data mining*.

I risultati ottenuti dall'analisi, sono riportati nelle tabelle (Tabella 3 - 8), indicanti i tempi di *change detection* (nella colonna "*Attività Classificata*"), associati ai tempi di inizio delle fasi di attività di fontana di lava (fase "ii") e delle fasi iniziali di attività Stromboliana (fase "i") ricavati dai report vulcanologici, e riportati in Tabella 1 e 2 (colonne "*Attività Rilevata*"). Sono mostrati inoltre anche i tempi di anticipo e ritardo delle *Attività Rilevate* rispetto ai tempi delle *Attività Classificate*. In fondo alle tabelle, sono riportati i valori medi degli anticipi e dei ritardi. Inoltre, per una maggior chiarezza, nella didascalia di ogni tabella vengono riportati i valori percentuali relativi ai VP, VN e FP.

## 2.1 Criterio di associazione tra tempo di *detection* e tempo di attività reale

Per ogni tempo di attività reale, si associa il tempo di *detection* più vicino ad esso, ma non oltre un certo intervallo massimo temporale. L'associazione viene rappresentata inserendo il tempo di attività reale nella stessa riga del tempo di *detection* più vicino.

Come intervallo massimo temporale si è presa come riferimento la durata massima del tipo di evento considerato, facendo distinzione tra le attività Stromboliane e le attività di fontana (evidenziate in rosso nelle Tabelle 1 e 2 alla colonna "*Durata*").

Consideriamo due esempi prendendo spunto dalla Tabella 3:

1. l'attività Stromboliana rilevata in data **11/01/2011 08:00** viene messa in relazione con la *detection* restituita dall' algoritmo in data **11/01/2011 22:40**, dato che la differenza temporale tra le due (**14:40:15** – hh:mm:ss) è minore della durata massima delle attività stromboliane presenti nella Tabella 1 (**52:07:00** – hh:mm:ss).
2. l'attività di fontanamento in data **12/01/2011 21:53** non viene messa in relazione con la *detection* restituita dall' algoritmo in data **13/01/2011 03:40**, dato che la differenza temporale tra le due (**05:47:15** – hh:mm:ss) è maggiore della massima evidenziata nella Tabella 2 (**05:17:00** – hh:mm:ss).

Nel caso in cui il tempo di *detection* più vicino all'attività rilevata appartenga ad una lista di *detection* consecutive vicine tra loro (la cui differenza del precedente dal successivo sia minore di 24 ore – ciò fa presupporre che si riferiscano ad uno stesso fenomeno), si è proceduto a considerare tale lista come unico blocco. Nelle Tabelle 3 - 8 i blocchi verranno rappresentati come un insieme di righe consecutive di *detection* dello stesso colore (in grigio o in bianco). In questo caso si mette in relazione il tempo reale con la prima data del blocco considerato. Considerando sempre la Tabella 3, l'attività di fontana rilevata in data **18/02/2011 06:50** risulta più vicina alla *detection* in data **18/02/2011 06:00**. Dato che tale *detection* appartiene al blocco che va dal **18/02/2011 03:00** al **18/02/2011 15:20**, l'attività viene messa in relazione al primo tempo presente in questo blocco (**18/02/2011 03:00**).

Inizio	Fine	Durata (hh:mm:ss)
11/01/2011 08:00	12/01/2011 21:53	37:53:00
18/02/2011 01:45	18/02/2011 06:50	05:05:00
08/04/2011 06:00	10/04/2011 10:07	<b>52:07:00</b>
11/05/2011 17:00	12/05/2011 01:27	08:27:00
07/07/2011 20:00	09/07/2011 14:02	42:02:00
18/07/2011 17:00	18/07/2011 23:56	06:56:00
24/07/2011 18:00	25/07/2011 04:01	10:01:00
30/07/2011 07:50	30/07/2011 19:36	11:46:00
05/08/2011 17:00	05/08/2011 21:46	04:46:00
12/08/2011 05:30	12/08/2011 08:31	03:01:00
20/08/2011 02:00	20/08/2011 07:00	05:00:00
28/08/2011 18:00	29/08/2011 04:04	10:04:00
08/09/2011 05:30	08/09/2011 07:22	01:52:00
19/09/2011 02:00	19/09/2011 12:19	10:19:00
28/09/2011 17:30	28/09/2011 19:09	01:39:00
08/10/2011 11:24	08/10/2011 14:26	03:02:00
23/10/2011 17:40	23/10/2011 18:40	01:00:00
15/11/2011 08:00	15/11/2011 11:19	03:19:00

**Tabella 1.** Tempi e durate delle fasi iniziali di attività Stromboliana (fase “i”) (in rosso la massima durata tra gli eventi).

Inizio	Fine	Durata (hh:mm:ss)
12/01/2011 21:53	12/01/2011 23:21	01:28:00
18/02/2011 06:50	18/02/2011 12:07	<b>05:17:00</b>
10/04/2011 10:07	10/04/2011 12:56	02:49:00
12/05/2011 01:27	12/05/2011 03:51	02:24:00
09/07/2011 14:02	09/07/2011 15:15	01:13:00
18/07/2011 23:56	19/07/2011 02:56	03:00:00
25/07/2011 04:01	25/07/2011 06:04	02:03:00
30/07/2011 19:36	30/07/2011 21:15	01:39:00
05/08/2011 21:46	05/08/2011 23:12	01:26:00
12/08/2011 08:31	12/08/2011 10:46	02:15:00
20/08/2011 07:00	20/08/2011 07:53	00:53:00
29/08/2011 04:04	29/08/2011 04:53	00:49:00
08/09/2011 07:22	08/09/2011 08:27	01:05:00
19/09/2011 12:19	19/09/2011 13:19	01:00:00
28/09/2011 19:09	28/09/2011 20:03	00:54:00
08/10/2011 14:26	08/10/2011 15:19	00:53:00
23/10/2011 18:40	23/10/2011 21:06	02:26:00
15/11/2011 11:19	15/11/2011 12:42	01:23:00

**Tabella 2.** Tempi e durate delle fasi di fontana di lava (fase “ii”) (in rosso la massima durata di un evento).

## 2.2 Test CUSUM

Come abbiamo discusso precedentemente, l’applicazione del metodo basato sulla CUSUM è limitata solo all’analisi delle serie storiche. Per questioni di brevità si riportano i risultati ottenuti sul segnale sismico registrato dalla stazione ECPN, in virtù della migliore qualità dei risultati relativamente agli aspetti sopra descritti (capacità di individuare ogni cambiamento e con quale tempestività, numero di cambi non individuati, tasso di falsi allarmi). Si è ritenuto infatti sufficiente confrontare i migliori risultati ottenuti con entrambi i metodi. Di seguito sono riportate le tabelle delle *detection* utilizzando sia lo stimatore basato su  $S_m$  (Tabella 3) che sull’MSE (Tabella 4). Per ogni esperimento sono stati adottati gli stessi valori per i parametri:

1. **Confidence Level** (livello minimo di confidenza per il quale proseguire con la ricerca del *change point*) fissato a 0.97.
2. **B** (numero di *bootstrap*) fissato a 1000, come consigliato anche da Taylor [2000].
3. **maxNumCP** (numero di *change point* da ricercare) fissato a 20, dato che sappiamo a posteriori il numero di attività coinvolte durante il 2011.

CUSUM –ECPN 2011– Stimatore “ $S_m$ ”						
Attività Classificata	Attività di fontanamento			Attività Stromboliana		
	Attività Rilevata	Anticipo	Ritardo	Attività Rilevata	Anticipo	Ritardo
11/01/2011 22:40				11/01/2011 08:00		14:40:15
	12/01/2011 21:53					
13/01/2011 03:40						
11/02/2011 07:30						
18/02/2011 03:00	18/02/2011 06:50	3:49:56		18/02/2011 01:45		1:15:04
18/02/2011 04:50						
18/02/2011 06:00						
18/02/2011 12:30						
18/02/2011 15:20						
09/04/2011 12:50				08/04/2011 06:00		30:50:14
	10/04/2011 10:07					
10/04/2011 22:00						
22/04/2011 04:20						
11/05/2011 13:40	12/05/2011 01:27	11:46:35		11/05/2011 17:00	3:19:35	
11/05/2011 22:20						
12/05/2011 07:20						
30/05/2011 18:40						
20/06/2011 04:20						
01/07/2011 18:30						
04/07/2011 07:30						
06/07/2011 19:40						
07/07/2011 19:40				07/07/2011 20:00	0:19:56	
	09/07/2011 14:02					
12/07/2011 05:50						
19/07/2011 03:10	18/07/2011 23:56		3:14:04	18/07/2011 17:00		10:10:04
24/07/2011 21:10				24/07/2011 18:00		3:10:04
	25/07/2011 04:01					
	30/07/2011 19:36			30/07/2011 07:50		
	05/08/2011 21:46			05/08/2011 17:00		
12/08/2011 14:40				12/08/2011 05:30		9:10:04
	12/08/2011 08:31					
	20/08/2011 07:00			20/08/2011 02:00		
	29/08/2011 04:04			28/08/2011 18:00		
	08/09/2011 07:22			08/09/2011 05:30		
	19/09/2011 12:19			19/09/2011 02:00		
	28/09/2011 19:09			28/09/2011 17:30		
	08/10/2011 14:26			08/10/2011 11:24		
08/10/2011 20:10						
23/10/2011 17:10	23/10/2011 18:40	1:29:55		23/10/2011 17:40	0:29:55	
23/10/2011 20:50						
24/10/2011 00:00						
	15/11/2011 11:19			15/11/2011 08:00		
18/11/2011 12:00						
19/12/2011 04:40						
28/12/2011 21:30						
Media		5:42:09	3:14:04		1:23:09	11:32:37

**Tabella 3.** Risultati ottenuti mediante lo stimatore  $S_m$ . Fasi di fontana di lava: VP=22.2%; VN=77.8%; FP=87.1%. Fasi di attività Stromboliana: VP=50%; VN=50%; FP=71%.

Utilizzando lo stimatore  $S_m$ , per l'attività di fontana i VP sono il 22,2%, con un valore medio di anticipo di 05:42 (hh:mm), mentre i FP ammontano all'87.1%. Per l'attività Stromboliana, la percentuale di VP è del 50%, con un tempo di anticipo medio di 01:23 (hh:mm), e di ritardo pari a 11:32 (hh:mm). Per il metodo del MSE i rapporti percentuali indicano un numero minore di classificazioni per entrambe le attività, pari al 16,7% per l'attività fontana e al 22.2% dei casi classificati per l'attività Stromboliana. Tra i due metodi il più

efficiente risulta lo Stimatore  $S_m$  in quanto fornisce anche un numero maggiore di VP rispetto all'MSE anche se la differenza in rapporto percentuale tra i VP non è particolarmente alta.

CUSUM –ECPN 2011– Stimatore “MSE”						
Attività Classificata	Attività di fontanamento			Attività Stromboliana		
	Attività Rilevata	Anticipo	Ritardo	Attività Rilevata	Anticipo	Ritardo
03/01/2011 22:00						
12/01/2011 21:50	12/01/2011 21:53	0:02:46		11/01/2011 08:00		37:50:14
15/01/2011 23:10						
16/01/2011 09:40						
24/01/2011 07:00						
18/02/2011 12:00	18/02/2011 06:50		5:10:04	18/02/2011 01:45		10:15:04
20/02/2011 00:10						
20/02/2011 00:30						
	10/04/2011 10:07			08/04/2011 06:00		
	12/05/2011 01:27			11/05/2011 17:00		
	09/07/2011 14:02			07/07/2011 20:00		
	18/07/2011 23:56					
19/07/2011 16:10				18/07/2011 17:00		23:10:04
19/07/2011 17:00						
19/07/2011 17:10						
25/07/2011 04:10	25/07/2011 04:01		0:09:04	24/07/2011 18:00		10:10:04
	30/07/2011 19:36			30/07/2011 07:50		
	05/08/2011 21:46			05/08/2011 17:00		
	12/08/2011 08:31			12/08/2011 05:30		
	20/08/2011 07:00			20/08/2011 02:00		
	29/08/2011 04:04			28/08/2011 18:00		
	08/09/2011 07:22			08/09/2011 05:30		
	19/09/2011 12:19			19/09/2011 02:00		
	28/09/2011 19:09			28/09/2011 17:30		
	08/10/2011 14:26			08/10/2011 11:24		
	23/10/2011 18:40			23/10/2011 17:40		
	15/11/2011 11:19			15/11/2011 08:00		
22/12/2011 21:30						
22/12/2011 22:40						
22/12/2011 22:50						
24/12/2011 03:10						
24/12/2011 03:20						
24/12/2011 04:00						
24/12/2011 04:20						
24/12/2011 04:30						
24/12/2011 05:00						
25/12/2011 12:00						
25/12/2011 13:10						
25/12/2011 17:00						
25/12/2011 18:30						
26/12/2011 11:40						
26/12/2011 16:40						
26/12/2011 18:00						
<b>Media</b>		0:02:46	2:39:34			20:21:21

**Tabella 4.** Risultati ottenuti mediante lo stimatore MSE. Fasi di fontana di lava: VP=16.7%; VN=83.3%; FP=89.3%. Fasi di attività Stromboliana: VP=22.2%; VN=77.8%; FP=85.7%.

### 2.3 Test One Pass Concept Change Detection for Data Streams

Di seguito sono riportate le tabelle delle *detection* ricavate dalla simulazione dello *streaming* dei dati del 2011 attraverso la tecnica *SeqDrift*, progettata per l'analisi in tempo reale. Nelle Tabelle 5, 6 e 7 si ripor-



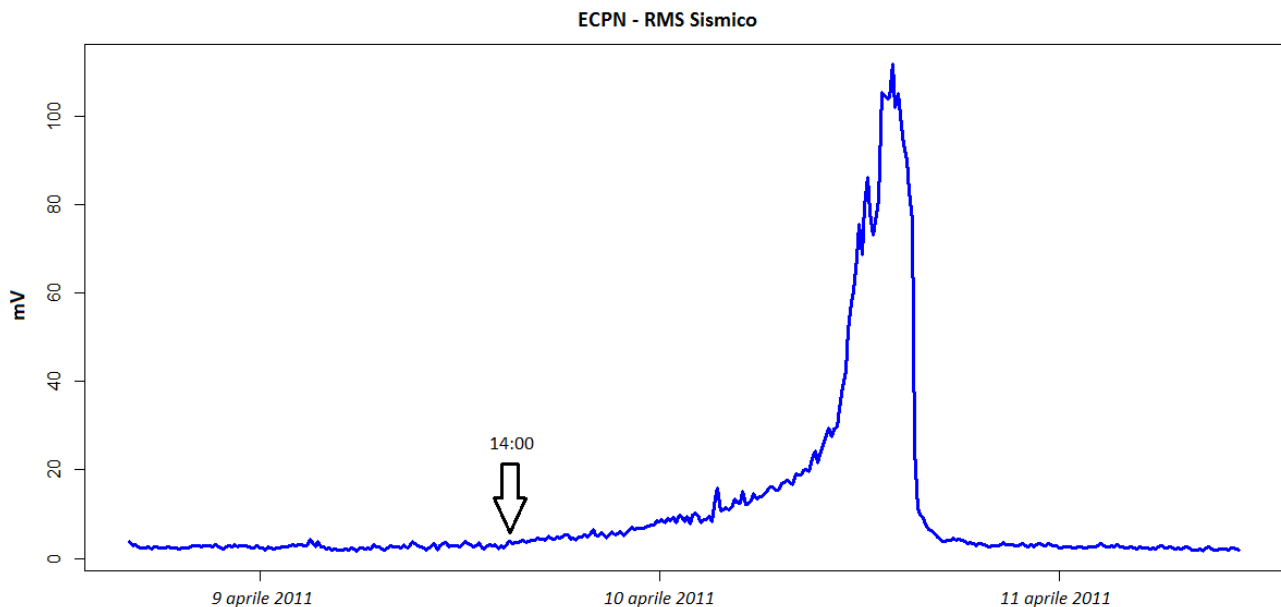
tano i tempi classificati sul segnale sismico acquisito alle stazioni ECPN, EMFS e EMPL. Per tutti gli esperimenti sono stati adottati gli stessi valori per i parametri dell'algoritmo:

1. **blockSize** (lunghezza della finestra, ovvero numero di campioni da considerare all'interno delle finestre mobili) uguale a 6. Dato che si è scelto un campionamento a 10 minuti, i confronti vengono effettuati paragonando le medie orarie dei campioni.
2.  **$\delta$ -warning** (livello di confidenza per accettare l'indicazione di warning) impostato a 0.95.
3.  **$\delta$ -change** (livello di confidenza per accettare l'indicazione di cambio di stato) impostato a 0.97.

Come si può notare, i livelli di confidenza sono stati impostati ad un valore molto alto al fine di ridurre quanto più possibile il numero di falsi positivi (ovvero il numero di *detection* non relative ad attività reali). Si è potuto constatare (e questo è stato anche dimostrato dagli stessi autori dell'algoritmo, [Sakthithasan et al., 2013]) come il numero di *detection* (e di conseguenza dei falsi positivi) sia influenzato in maniera sostanziale anche dal valore di *blockSize* ( $b$  nell'Eq. 6): esso ha maggiore influenza a denominatore e quindi il valore di  $\epsilon$  relativo tenderà a diminuire quanto più aumenta il suo valore. Per effetto di ciò, come si può vedere dall'Eq. 5, minore sarà il valore di  $\epsilon$ , maggiore sarà la probabilità di superare la soglia  $\delta$  (sia del *warning* che del *change*).

Nella Tabella 5 la *detection* alla quale corrisponde uno sfasamento maggiore tra attività rilevata e classificata è quella del 09/04/2011, per la quale si ottiene un anticipo di 15:06 (hh:mm) rispetto all'inizio della fase di fontanamento. Tale dato, se messo in relazione all'attività Stromboliana iniziata l'08/04/2011 06:00, porta un ritardo pari a 37:00 (hh:mm). Il notevole ritardo della classificazione del 09/04/2011 è riscontrabile anche nelle attività stromboliane delle stazioni EMFS e EMPL mostrate nelle Tabelle 6 e 7, con i rispettivi valori di 40:20 (hh:mm) e 45:30 (hh:mm). Tale comportamento risulta giustificabile solo se si analizza il grafico della serie RMS (Figura 4): l'attività Stromboliana comincia ad avere effetti sul trend dell'RMS del segnale sismico a partire solamente dalla seconda metà del 09/04/2011, dove appunto l'algoritmo è in grado di identificare il *change point*.

Di particolare rilevanza sono i valori percentuali dei VP che, nelle tabelle in questione, sono compresi nell'intervallo di valori che va dall'83% (Tabella 7) al 100% (Tabella 5) per l'attività Stromboliana, e per valori compresi tra l'83,3% (Tabella 7) e il 94,4% (Tabella 5) per le fontane.



**Figura 4.** Un esempio di incremento del trend dell'RMS del segnale sismico acquisito dalla stazione ECPN. Nello specifico si riporta il segmento di serie temporale relativa alla fontana del 10 aprile 2011, ore 10:07. Si può notare come l'attività Stromboliana iniziata l'8 aprile 2011 non influisca sul trend dell'RMS fino al giorno 9, ore 14:00, quando inizia un trend con un graduale incremento dell'RMS.

CLASSIFICAZIONE –ECPN 2011– SeqDrift						
Attività Classificata	Attività di fontanamento			Attività Stromboliana		
	Attività Rilevata	Anticipo	Ritardo	Attività Rilevata	Anticipo	Ritardo
03/01/2011 00:00						
12/01/2011 08:00	12/01/2011 21:53	13:52:56		11/01/2011 08:00		24:00:04
12/01/2011 18:00						
12/01/2011 22:00						
18/02/2011 05:00	18/02/2011 06:50	1:49:56		18/02/2011 01:45		3:15:04
09/04/2011 19:00	10/04/2011 10:07	15:06:45		08/04/2011 06:00		37:00:15
10/04/2011 07:00						
11/05/2011 18:10	12/05/2011 01:27	7:16:56		11/05/2011 17:00		1:10:04
11/05/2011 23:10						
02/06/2011 14:10						
07/07/2011 05:10				07/07/2011 20:00	14:49:55	
09/07/2011 12:10	09/07/2011 14:02	1:51:45				
09/07/2011 14:10						
11/07/2011 21:10						
19/07/2011 00:00	18/07/2011 23:56		0:04:14	18/07/2011 17:00		7:00:14
24/07/2011 22:00	25/07/2011 04:01	6:00:45		24/07/2011 18:00		4:00:15
25/07/2011 04:00						
30/07/2011 09:00	30/07/2011 19:36	10:35:56		30/07/2011 07:50		1:10:04
30/07/2011 20:00						
05/08/2011 21:10	05/08/2011 21:46	0:35:56		05/08/2011 17:00		4:10:04
12/08/2011 07:10	12/08/2011 08:31	1:20:46		12/08/2011 05:30		1:40:14
12/08/2011 09:10						
20/08/2011 01:10				20/08/2011 02:00	0:49:56	
	20/08/2011 07:00					
29/08/2011 02:10	29/08/2011 04:04	1:53:55		28/08/2011 18:00		8:10:05
08/09/2011 07:30	08/09/2011 07:22		0:08:24	08/09/2011 05:30		2:00:24
19/09/2011 07:40	19/09/2011 12:19	4:38:56		19/09/2011 02:00		5:40:04
19/09/2011 11:40						
28/09/2011 15:40	28/09/2011 19:09	3:28:55		28/09/2011 17:30	1:49:55	
08/10/2011 12:40	08/10/2011 14:26	1:45:55		08/10/2011 11:24		1:16:05
23/10/2011 18:30	23/10/2011 18:40	0:09:56		23/10/2011 17:40		0:50:04
15/11/2011 09:10	15/11/2011 11:19	2:08:56		15/11/2011 08:00		1:10:04
15/11/2011 11:10						
Media		4:50:33	0:06:19		5:49:55	6:50:12

**Tabella 5.** Risultati ottenuti mediante la tecnica *SeqDrift* applicata all'RMS della stazione ECPN. Fasi di fontana di lava: VP=94.4%; VN=5.6%; FP=46.9%. Fasi di attività Stromboliana: VP=100%; VN=0%; FP=43.7%.

CLASSIFICAZIONE –EMFS 2011 – SeqDrift						
Attività Classificata	Attività di fontanamento			Attività Stromboliana		
	Attività Rilevata	Anticipo	Ritardo	Attività Rilevata	Anticipo	Ritardo
12/01/2011 13:00	12/01/2011 21:53	8:52:55		11/01/2011 08:00		29:00:05
12/01/2011 22:00						
18/02/2011 05:30	18/02/2011 06:50	1:19:55		18/02/2011 01:45		3:45:05
09/04/2011 22:20	10/04/2011 10:07	11:46:55		08/04/2011 06:00		40:20:05
10/04/2011 09:20						
11/05/2011 18:30	12/05/2011 01:27	6:56:46		11/05/2011 17:00		1:30:14
11/05/2011 22:30						
12/05/2011 01:30						
21/05/2011 12:30						
08/07/2011 02:00				07/07/2011 20:00		6:00:35
09/07/2011 11:00	09/07/2011 14:02	3:01:56				
12/07/2011 16:20						
19/07/2011 00:20	18/07/2011 23:56		0:24:05	18/07/2011 17:00		7:20:05
24/07/2011 22:20	25/07/2011 04:01	5:40:55		24/07/2011 18:00		4:20:05

25/07/2011 02:20						
30/07/2011 09:20	30/07/2011 19:36	10:15:55		30/07/2011 07:50		1:30:05
30/07/2011 18:20						
05/08/2011 20:20	05/08/2011 21:46	1:25:56		05/08/2011 17:00		3:20:04
12/08/2011 07:10	12/08/2011 08:31	1:20:56		12/08/2011 05:30		1:40:04
	20/08/2011 07:00			20/08/2011 02:00		
28/08/2011 08:20				28/08/2011 18:00	9:39:56	
29/08/2011 01:20	29/08/2011 04:04	2:43:55				
08/09/2011 07:00	08/09/2011 07:22	0:21:55		08/09/2011 05:30		1:30:05
19/09/2011 07:30	19/09/2011 12:19	4:48:56		19/09/2011 02:00		5:30:04
19/09/2011 11:30						
28/09/2011 15:40	28/09/2011 19:09	3:28:55		28/09/2011 17:30	1:49:55	
08/10/2011 12:40	08/10/2011 14:26	1:45:55		08/10/2011 11:24		1:16:05
23/10/2011 18:20	23/10/2011 18:40	0:19:55		23/10/2011 17:40		0:40:05
15/11/2011 11:10	15/11/2011 11:19	0:08:55		15/11/2011 08:00		3:10:05
<b>Media</b>		4:01:17	0:24:05		5:44:56	7:23:31

**Tabella 6.** Risultati ottenuti mediante la tecnica *SeqDrift* applicata all'RMS della stazione EMFS. Fasi di fontana di lava: VP=94.4%; VN=5.6%; FP=39.3%. Fasi di attività Stromboliana: VP=94.4%; VN=5.6%; FP=39.3%.

CLASSIFICAZIONE - EMPL 2011 – <i>SeqDrift</i>						
Attività Classificata	Attività di fontanamento			Attività Stromboliana		
	Attività Rilevata	Anticipo	Ritardo	Attività Rilevata	Anticipo	Ritardo
12/01/2011 19:00	12/01/2011 21:53	2:52:45		11/01/2011 08:00		35:00:15
18/02/2011 06:30	18/02/2011 06:50	0:19:56		18/02/2011 01:45		4:45:04
10/04/2011 03:30	10/04/2011 10:07	6:36:56		08/04/2011 06:00		45:30:04
10/04/2011 09:30						
11/05/2011 20:20	12/05/2011 01:27	5:06:56		11/05/2011 17:00		3:20:04
09/07/2011 10:20	09/07/2011 14:02	3:41:55		07/07/2011 20:00		
19/07/2011 00:10	18/07/2011 23:56		0:14:04	18/07/2011 17:00		7:10:04
25/07/2011 01:10	25/07/2011 04:01	2:50:56		24/07/2011 18:00		7:10:04
25/07/2011 03:10						
30/07/2011 09:10	30/07/2011 19:36	10:25:56		30/07/2011 07:50		1:20:04
30/07/2011 20:10						
05/08/2011 20:30	05/08/2011 21:46		0:44:05	05/08/2011 17:00		3:30:05
12/08/2011 07:30	12/08/2011 08:31		0:59:04	12/08/2011 05:30		2:00:04
	20/08/2011 07:00			20/08/2011 02:00		
29/08/2011 01:30	29/08/2011 04:04	2:33:56		28/08/2011 18:00		7:30:04
	08/09/2011 07:22			08/09/2011 05:30		
19/09/2011 11:00	19/09/2011 12:19	1:18:56		19/09/2011 02:00		9:00:04
28/09/2011 16:00	28/09/2011 19:09	3:08:55		28/09/2011 17:30	1:29:55	
08/10/2011 12:00	08/10/2011 14:26	2:25:56		08/10/2011 11:24		0:36:04
	23/10/2011 18:40			23/10/2011 17:40		
15/11/2011 11:20	15/11/2011 11:19		0:01:04	15/11/2011 08:00		3:20:04
<b>Media</b>		3:45:44	0:29:34		1:29:55	10:00:56

**Tabella 7.** Risultati ottenuti mediante la tecnica *SeqDrift* applicata all'RMS della stazione EMPL. Fasi di fontana di lava: VP=83.3%; VN=16.7%; FP=16.7%. Fasi di attività Stromboliana: VP=83.3%; VN=16.7%; FP=16.7%.

#### 2.4 Esperimento con campionamento al minuto per l'analisi in real-time

Di seguito verrà analizzato l'algoritmo *SeqDrift*, applicato ad un segnale a più alto tasso di campionamento. Utilizzando un campionamento a 10 minuti non si riescono a sfruttare al meglio le potenzialità dell'algoritmo, visto che è stato progettato per analizzare dati di tipo *stream*.

Per meglio apprezzare la sua efficacia, si è effettuata una simulazione sullo stesso periodo (2011), e con gli stessi parametri precedentemente utilizzati, su uno *streaming* a più alto tasso di campionamento (1

campione al minuto anziché 1 ogni 10). In questo caso il ritardo della *detection* (pari al  $t_2 - t_1 = sampleSize$ , vedi Sezione 1.2), che prima era pari ad almeno 60 minuti (=  $sampleSize \times 10$  minuti) dal tempo di *change point*, in questo caso può essere stimato pari ad almeno 6 minuti (=  $sampleSize \times 1$  minuto).

Si è scelto di continuare a prendere come segnale di riferimento per il confronto, l'RMS del segnale sismico di ECPN (il segnale che si è dimostrato più affidabile per questi esperimenti). I risultati sono mostrati nella Tabella 8.

CLASSIFICAZIONE - ECPN 2011 – SeqDrift						
Attività Classificata	Attività di fontanamento			Attività Stromboliana		
	Attività Rilevata	Anticipo	Ritardo	Attività Rilevata	Anticipo	Ritardo
12/01/2011 14:10	12/01/2011 21:53	7:42:55		11/01/2011 08:00		06:10:05
12/01/2011 20:36						
12/01/2011 22:18						
18/02/2011 04:27	18/02/2011 06:50	2:22:45		18/02/2011 01:45		02:42:15
10/04/2011 01:17	10/04/2011 10:07	8:49:56		08/04/2011 06:00		19:17:04
10/04/2011 08:17						
10/04/2011 11:37						
11/05/2011 19:57	12/05/2011 01:27	5:29:56		11/05/2011 17:00		02:57:04
11/05/2011 22:33						
12/05/2011 01:15						
08/07/2011 03:28				07/07/2011 20:00		07:28:04
09/07/2011 12:27	09/07/2011 14:02	1:34:56				
09/07/2011 14:09						
19/07/2011 00:18	18/07/2011 23:56		00:22:04	18/07/2011 17:00		07:18:04
24/07/2011 22:45	25/07/2011 04:01	5:15:55		24/07/2011 18:00		04:45:05
25/07/2011 02:33						
25/07/2011 04:57						
30/07/2011 08:45	30/07/2011 19:36	10:50:56		30/07/2011 07:50		00:55:04
30/07/2011 19:04						
30/07/2011 19:58						
05/08/2011 20:27	05/08/2011 21:46	1:18:56		05/08/2011 17:00		03:27:04
05/08/2011 22:21						
12/08/2011 07:11	12/08/2011 08:31	1:19:55		12/08/2011 05:30		01:41:05
12/08/2011 07:35						
12/08/2011 09:23						
20/08/2011 05:24	20/08/2011 07:00	1:35:56		20/08/2011 02:00		03:24:04
20/08/2011 07:12						
29/08/2011 01:45	29/08/2011 04:04	2:18:55		28/08/2011 18:00		07:45:05
29/08/2011 04:04						
29/08/2011 04:28						
08/09/2011 06:22	08/09/2011 07:22	0:59:55		08/09/2011 05:30		00:52:05
08/09/2011 07:22						
19/09/2011 09:01	19/09/2011 12:19	3:17:05		19/09/2011 02:00		07:01:55
19/09/2011 11:56						
28/09/2011 16:25	28/09/2011 19:09	2:43:45		28/09/2011 17:30	01:04:45	
28/09/2011 18:55						
28/09/2011 19:43						
08/10/2011 11:53	08/10/2011 14:26	2:32:55		08/10/2011 11:24		00:29:05
08/10/2011 14:41						
08/10/2011 14:59						
23/10/2011 18:15	23/10/2011 18:40	0:24:55		23/10/2011 17:40		00:35:05
15/11/2011 10:04	15/11/2011 11:19	1:14:56		15/11/2011 08:00		02:04:04
15/11/2011 11:04						
<b>Media</b>		<b>3:31:27</b>	<b>00:22:04</b>		<b>01:04:45</b>	<b>07:27:47</b>

**Tabella 8.** Risultati ottenuti mediante la tecnica *SeqDrift* applicata all'RMS della stazione ECPN, considerando finestre temporali di 1 minuto. Fasi di fontana di lava: VP=100%; VN=0%; FP=58.1%. Fasi di attività Stromboliana: VP=100%; VN=0%; FP=58.1%.

Dalla simulazione si evince che in tutti i casi, tranne uno, le fontane sono state rilevate con un anticipo medio di 3 ore e mezza; mentre le attività stromboliane sono state rilevate in media con 7 ore e mezza di ritardo (tranne in un caso) dal loro inizio. Ciò può risultare accettabile in quanto il ritardo non supera mai la fine dell'attività considerata. In ogni caso si è dimostrato come si possa raggiungere la *detection* della totalità delle attività del 2011 con un tasso di falsi positivi (FP) del 58% circa, sia per l'attività di fontanamento che per quella Stromboliana. Da notare che nella Tabella 8, analogamente alle precedenti tabelle (3-7), le percentuali dei falsi positivi sono state calcolate semplicemente mettendo in relazione il numero delle *detection* non allineate con un'attività reale (25) sul numero di *detection* totali (43). Se però le *detection* vengono considerate in blocchi come descritto nella Sezione 2.1, nel rispetto della definizione di falso positivo descritto nella Sezione 2 "il numero di *detection* non rientranti nei periodi di attività reali (falsi positivi)", il numero di *detection* non allineate con un'attività reale scende a 0, portando ad avere una percentuale di falsi positivi che scende a 0%.

### 3. Discussione e conclusioni

Le eruzioni vulcaniche sono sorgente di differenti *hazard*, legati a colate di lava, emissione di cenere, flussi piroclastici, *lahar*, emissioni di gas, etc. Negli ultimi anni, a causa del netto incremento del traffico aereo [e.g. Tilling et al., 2008], l'attenzione della protezione civile e della comunità scientifica si è focalizzata sull'*hazard* da emissione di cenere, che è prodotta da attività vulcanica di tipo esplosivo. Il monitoraggio vulcanico permette di ridurre il rischio legato ai sopra citati *hazard* vulcanici. In particolare, il monitoraggio sismico è considerato come uno dei più affidabili, per i seguenti motivi [e.g. McNutt et al., 2000; Cannata et al., 2013b]: i) a differenza della sorveglianza basata sulle telecamere, è efficace anche in caso di copertura nuvolosa che impedisce di osservare direttamente l'area vulcanica; ii) è meno disturbato dal vento, rispetto al monitoraggio basato sui segnali infrasonici; iii) le attrezzature per il monitoraggio sismico sono relativamente economiche e facili da installare; iv) infine, l'attività sismica è considerata uno dei migliori indicatori a breve e medio termine delle variazioni dell'attività vulcanica.

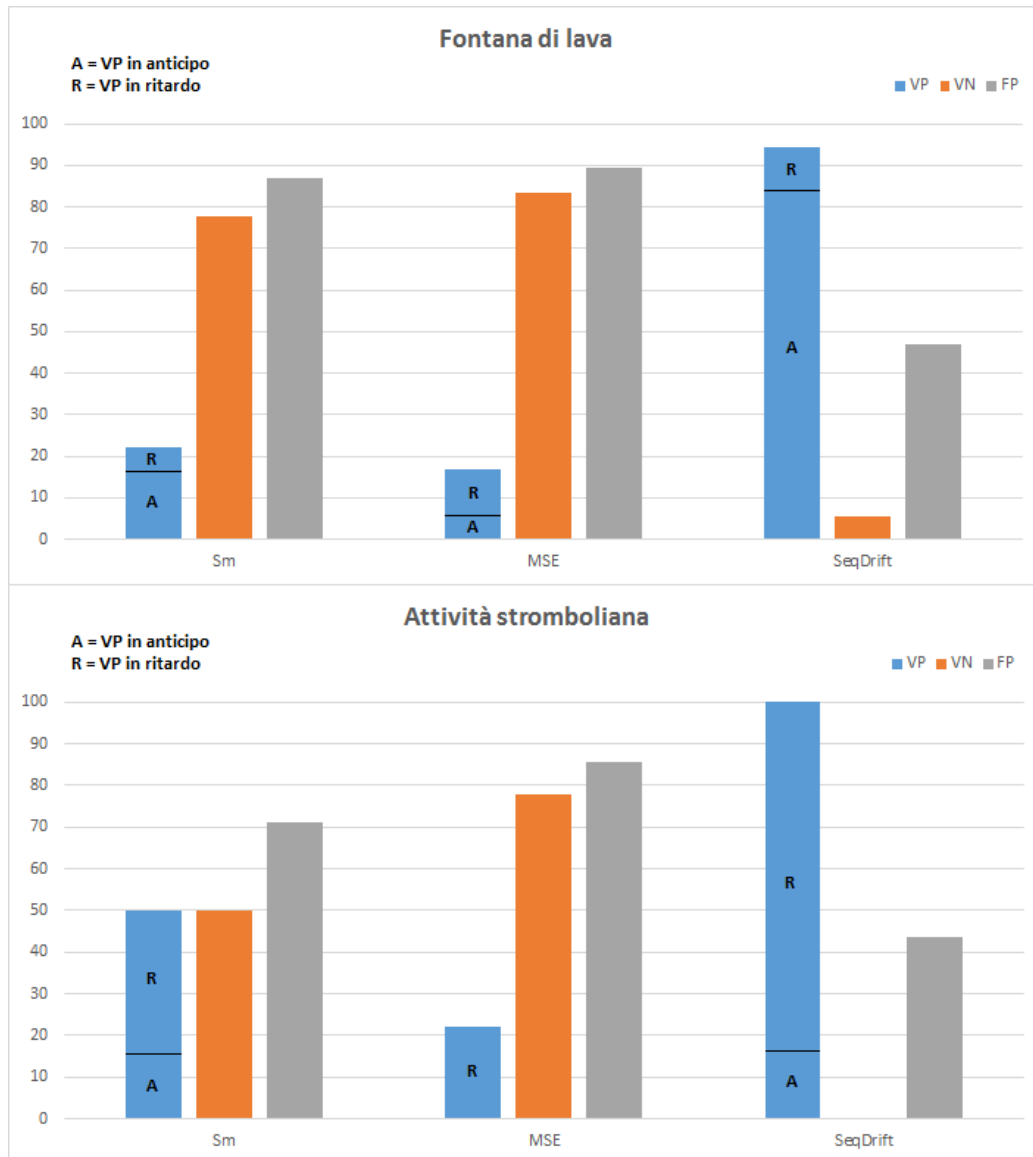
Con lo scopo di migliorare il monitoraggio sismico di attività vulcanica di tipo esplosivo, sono stati applicati diversi metodi di *change point detection* a serie temporali di RMS ricavate dal segnale sismico registrato da stazioni installate sul vulcano Etna. Il monitoraggio dell'attività esplosiva è particolarmente importante in questo vulcano che ha prodotto oltre 50 attività parossistiche dal 2011 al 2015 [e.g. Viccaro et al., 2014; Spampinato et al., 2015; INGV staff, 2011]. Le serie di RMS sismico all'Etna sono principalmente influenzate dal tremore vulcanico, che in tale vulcano è un segnale continuo di ampiezza generalmente sostenuta.

Il primo metodo di *change point detection* applicato alle serie di RMS è stato il CUSUM applicato utilizzando i due stimatori  $S_m$  e MSE. Tra questi il più efficiente è risultato lo stimatore  $S_m$  in quanto fornisce un numero maggiore di VP rispetto all'MSE (vedi Tabelle 3 e 4 e Figura 5). Tuttavia, le tecniche basate su CUSUM non sono adatte per analisi in *near real-time* di *stream* continui di dati.

Per tale scopo è sicuramente più adatto il secondo metodo, il *SeqDrift*. I parametri utilizzati per il metodo *SeqDrift* sono stati scelti in modo da ridurre quanto più possibile la presenza di falsi positivi. Tale scelta ha influenzato principalmente i tempi di individuazione del cambio, riducendo i tempi di anticipo e aumentando di conseguenza i tempi dei ritardi. Si è notato, inoltre, che l'impostazione scelta risulta più efficace per la *detection* dell'attività di tipo fontana ed inefficace riguardo l'attività Stromboliana. Ciò è dovuto al fatto che spesso l'attività Stromboliana, specialmente blanda che caratterizza l'inizio degli episodi parossistici, non influisce sul trend dell'RMS del tremore vulcanico quanto l'attività di fontanamento. I risultati hanno infatti messo in evidenza come la classificazione del metodo non possa prescindere dall'affidabilità del segnale monitorato: quanto più le variazioni sul segnale sismico rispecchiano le variazioni dell'attività vulcanica, tanto più le *detection* possono essere affidabili. Per migliorare la *detection* dell'attività Stromboliana può essere scelta una diversa configurazione dei parametri dell'algoritmo con il rischio però di aumentare notevolmente il numero di falsi positivi.

Sulla base delle configurazioni scelte, si può notare nelle Tabelle 3-7 e in Figura 5 come il metodo *SeqDrift* ottenga risultati migliori rispetto a quelli ottenuti con la CUSUM. Sebbene il valore medio dei tempi di anticipo delle fontane, rispettivamente 04:50, 04:01 e 3:45 (hh:mm), sia minore alla media 05:42 (hh:mm) mostrata in Tabella 3, pochissime sono le corrispondenze riscontrate in ritardo, le cui medie sono rispettivamente di 6 minuti, 24 minuti e 29 minuti. Per l'attività Stromboliana pochi sono i casi di anticipo riscon-

trati; in ogni caso i valori medi di ritardo nelle Tabelle 5, 6 e 7 risultano minori di quello riportato nella Tabella 3. L'alta percentuale di VP ottenuta attraverso *SeqDrift* permette di valutare come tale metodologia possa diventare utile ai fini del monitoraggio, soprattutto per quanto riguarda la *detection* delle fontane. Infatti la *detection* dell'attività Stromboliana avviene nella maggior parte dei casi a posteriori, contrariamente a quanto avviene per le fontane. Da questo punto di vista quindi, l'algoritmo è utile solo nel caso di *warning* relativi all'attività di fontanamento. Le analisi inoltre ci hanno mostrato come tali valori possano migliorare applicando *SeqDrift* su uno *streaming* a più alto tasso di campionamento.



**Figura 5.** Istogrammi che mostrano le performance dei metodi applicati ( $S_m$ , MSE e *SeqDrift*) alle serie RMS, ricavate da finestre mobili di 10 minuti di segnale sismico registrato dalla stazione ECPN, in termini di veri positivi (VP), veri negativi (VN) e falsi positivi (FP) sia nel caso di attività di fontana di lava (istogramma in alto) che nel caso di attività Stromboliana (istogramma in basso). Le barre relative ai VP sono divise in due sezioni di area proporzionale alla % dei VP che anticipano l'evento effettivo (A) contro quelli per i quali la corrispondenza è riscontrata in ritardo (R).

La letteratura indica come l'analisi incrociata di più osservazioni risulti essere la chiave vincente per valutare al meglio le dinamiche del vulcano e monitorare le sue variazioni nel tempo [e.g. McNutt et al., 2000; Tilling, 2008]. A tal proposito, una possibile evoluzione della metodologia potrebbe procedere in due

differenti direzioni, entrambe richiedenti un'esecuzione in parallelo di diverse istanze dell'algoritmo. La prima idea potrebbe consistere nel lanciare l'analisi in tempo reale su  $N$  differenti serie temporali scelte per il monitoraggio e, dato un trigger di *change point* scaturito da una delle  $N$  osservazioni scelte, "convalidare" l'output nel caso in cui all'interno di una finestra temporale di lunghezza prefissata ricadano altri *change point* provenienti dalle altre  $N-1$  osservazioni (usando ad esempio una politica di maggioranza). Dato che è stato evidenziato come dall'impostazione dei parametri dell'algoritmo dipenda la sensibilità all'individuazione dei cambiamenti, la seconda direzione potrebbe consistere nel lanciare diverse istanze dell'algoritmo rispetto ad una singola serie temporale con diversi parametri e seguire la stessa politica di votazione sopra citata per la convalida del *change point*, nonché per definire diversi livelli di allerta. Tale direzione risulterebbe di fondamentale importanza per il miglioramento della qualità dei risultati, specialmente nella riduzione delle percentuali dei falsi positivi. Ovviamente, la combinazione delle due scelte potrebbe portare alla direzione ottimale.

## Ringraziamenti

I risultati di questo report sono stati ottenuti nell'ambito del progetto PON01\_00683 SIGMA (Sistema Integrato di sensori in ambiente cloud per la Gestione Multirischio Avanzata).

## Bibliografia

- Alparone S., Andronico D., Lodato L., SgROI, T., (2003). *Relationship between tremor and volcanic activity during the South-east Crater eruption on Mount Etna in early 2000*. J. Geophys. Res. 108(B5), 2241.
- Andreou E., Ghysels E., (2002). *Detecting multiple breaks in financial market volatility dynamics*. Journal of Applied Econometrics, Wiley Interscience, vol. 17, n. 5, pag. 579-600.
- Behncke B., Branca S., Corsaro R.A., De Beni E., Miraglia L., Proietti C., (2014). *The 2011-2012 summit activity of Mount Etna: Birth, growth and products of the new SE crater*. J. Volcanol. Geotherm. Res., 270, 10-21.
- Cannata A., Catania A., Alparone S., Gresta S., (2008). *Volcanic tremor at Mt Etna: inferences on magma dynamics during effusive and explosive activity*. J. Volcanol. Geotherm. Res. 2, 259-287.
- Cannata A., Di Grazia G., Aliotta M., Cassisi C. Montalto P., Patanè D., (2013a). *Monitoring seismo-volcanic and infrasonic signals at volcanoes: Mt. Etna case study*. Pure and Applied Geophysics, doi:10.1007/s00024-012-0634-x.
- Cannata A., Montalto P., Patanè D., (2013b). *Joint analysis between infrasound and seismic signals by cross wavelet transform: detection of Mt. Etna explosive activity*. Natural Hazards and Earth System Sciences, 13, 1-9, doi:10.5194/nhess-13-1-2013.
- Efron B., Tibshirani R.J., (1993). *An Introduction to the Bootstrap*. Chapman & Hall.
- Grossman R., Sabala M., Aanand A., Eick S., Wilkinson L., Zhang P., Chaves J., Vejcek S., Dillenburg J., Nelson P., Rorem D., Alimohideen J., Leigh J., Papka M., Stevens R., (2005). *Real time change detection and alerts from highway traffic data*. In: Proceedings of the ACM/IEEE Supercomputing Conference (SC'05), Seattle, WA, USA, pages 69-74, Nov. 2005.
- Gustafson D.E., Willsky A.S., Wang J.-Y., Lancaster M.C., Triebwasser J.H., (1978). *ECG/VCG rhythm diagnosis using statistical signal analysis - I. Identification of persistent rhythms*. IEEE Transactions on Biomedical Engineering, 25(4):344-353, July 1978.
- INGV staff, (2011). *Volcanological and geochemical reports*. URL: <http://www.ct.ingv.it/it/rapporti/multidisciplinari.html?view=docman>.
- McNutt S.R., (2000). *Seismic Monitoring, in Encyclopedia of Volcanoes*. Edited by H. Sigurdsson, B. Houghton, S.R. McNutt, H. Rymer, and J. Stix, pp. 1095-1119.
- Montalto P. Aliotta M., Cannata A., Cassisi C., (2012). *Segmentazione delle serie temporali nell'analisi dei dati: un esempio di applicazione a dati sismovulcanici*. Rapporti Tecnici INGV, n. 224.
- Ortner M., Nehorai A., (2007). *A sequential detector for biochemical release in realistic environments*. IEEE Transactions on Signal Processing, 55(7):4173-4182, July 2007.
- R Development Core Team, (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org>.

- Reeves J., Chen J., Wang X.L., Lund R., Lu Q.Q., (2009). *A Review and Comparison of Change-point Detection Techniques for Climate Data*. Journal of Applied Meteorology and Climatology 46:6, 900-915. Online publication date: 1-Jun-2007.
- Sakthithasan S., Pears R., Koh Y.S., (2013). *One Pass Concept Change Detection for Data Streams*. PAKDD (2): 461-472.
- Sebastiao R., Gama J., (2009). *A study on change detection methods*. 4th Portuguese Conf. on Artificial Intelligence, 353-364.
- Shewhart, W.A., (1931). *Economic control of quality of manufactured product*. D. Van Nostrand Company, Inc., New York.
- Spampinato L., Sciotto M., Cannata A., Cannavò F., La Spina A., Palano M., Salerno G.G., Privitera E., Caltabiano T., (2015). *Multiparametric study of the February-April 2013 paroxysmal phase of Mt. Etna New South-East crater*. Geochem. Geophys. Geosyst., doi: 10.1002/2015GC005795.
- Tartakovsky A., Polunchenko A.S., Sokolov G., (2013). *Efficient computer network anomaly detection by change-point detection methods*. IEEE Journal of Selected Topics in Signal Processing, 7(1):4-11, Feb. 2013.
- Tartakovsky A., Nikiforov I., & Basseville M., (2014). *Sequential Analysis: Hypothesis Testing and Change-point Detection*. CRC Press.
- Taylor W.A., (2000). *Change-Point Analysis: A Powerful New Tool For Detecting Changes*, URL: <http://www.variation.com/cpa/tech/change-point.html>.
- Tilling R.I., (2008). *The critical role of volcano monitoring in risk reduction*. Adv. Geosci., 14, 3-11.
- Tomozeiu R., Busuioc A., Marletto V., Zinoni F., Cacciamani C., (2000). *Detection of changes in the summer precipitation time series of the region Emilia-Romagna, Italy*. Theoretical and Applied Climatology, 67 (3-4), 193-200.
- Viccaro M., Garozzo I., Cannata A., Di Grazia G., Gresta S., (2014). *Gas burst vs. gas-rich magma recharge: a multidisciplinary study to reveal factors controlling duration of the recent paroxysmal eruptions at Mt. Etna*. Journal of Volcanology and Geothermal Research, 278-279, 1-13.





# Quaderni di Geofisica

ISSN 1590-2595

<http://istituto.ingv.it/l-ingv/produzione-scientifica/quaderni-di-geofisica/>

I Quaderni di Geofisica coprono tutti i campi disciplinari sviluppati all'interno dell'INGV, dando particolare risalto alla pubblicazione di dati, misure, osservazioni e loro elaborazioni anche preliminari, che per tipologia e dettaglio necessitano di una rapida diffusione nella comunità scientifica nazionale ed internazionale. La pubblicazione on-line fornisce accesso immediato a tutti i possibili utenti. L'Editorial Board multidisciplinare garantisce i requisiti di qualità per la pubblicazione dei contributi.

# Rapporti tecnici INGV

ISSN 2039-7941

<http://istituto.ingv.it/l-ingv/produzione-scientifica/rapporti-tecnici-ingv/>

I Rapporti Tecnici INGV pubblicano contributi, sia in italiano che in inglese, di tipo tecnologico e di rilevante interesse tecnico-scientifico per gli ambiti disciplinari propri dell'INGV. La collana Rapporti Tecnici INGV pubblica esclusivamente on-line per garantire agli autori rapidità di diffusione e agli utenti accesso immediato ai dati pubblicati. L'Editorial Board multidisciplinare garantisce i requisiti di qualità per la pubblicazione dei contributi.

# Miscellanea INGV

ISSN 2039-6651

<http://istituto.ingv.it/l-ingv/produzione-scientifica/miscellanea-ingv/>

La collana Miscellanea INGV nasce con l'intento di favorire la pubblicazione di contributi scientifici riguardanti le attività svolte dall'INGV (sismologia, vulcanologia, geologia, geomagnetismo, geochimica, aeronomia e innovazione tecnologica). In particolare, la collana Miscellanea INGV raccoglie reports di progetti scientifici, proceedings di convegni, manuali, monografie di rilevante interesse, raccolte di articoli ecc..

**Coordinamento editoriale e impaginazione**

Centro Editoriale Nazionale | INGV

**Progetto grafico e redazionale**

Daniela Riposati | Laboratorio Grafica e Immagini | INGV

© 2015 INGV Istituto Nazionale di Geofisica e Vulcanologia

Via di Vigna Murata, 605

00143 Roma

Tel. +39 06518601 Fax +39 065041181

**<http://www.ingv.it>**



**Istituto Nazionale di Geofisica e Vulcanologia**