

Dipartimento di Ingegneria   
Elettrica Elettronica e Informatica

Tesi di Laurea Magistrale in Ingegneria Informatica

*Francesco Pandolfo*

Analisi di serie temporali GNSS con tecniche di Machine Learning per la detection di eventi vulcanici

Relatore:

Chiar.ma Prof. Antonella Di Stefano

Correlatore:

Prof. Giovanni Morana

Correlatore esterno:

Dott.ssa Valentina Bruno

Anno Accademico 2023/24

**Abstract**

Max 20 righe (inglese ?)

**Dichiarazione d’Autore**

Il sottoscritto, Francesco Pandolfo, dichiara che questa tesi intitolata, “Analisi di serie temporali GNSS con tecniche di Machine Learning per la detection di eventi vulcanici” ed il lavoro qui presentato è di sua proprietà. Confermo che:

• Questo lavoro è stato svolto interamente o principalmente durante la candidatura per la laurea magistrale presso questa Università.

• Laddove una qualsiasi parte di questa tesi sia stata precedentemente presentata per una laurea o qualsiasi altra qualifica presso questa Università o qualsiasi altra istituzione, ciò è stato chiaramente indicato.

• Laddove abbia consultato il lavoro pubblicato da altri, questo è sempre stato chiaramente attribuito.

• Laddove abbia citato lavori di altri, la fonte è sempre indicata. Ad eccezione di tali citazioni, questa tesi è interamente opera mia.

• Ho riconosciuto tutte le principali fonti di aiuto.

Firma:

Data:

17/10/2024

**INDICE**

[Introduzione 1](#_Toc178870128)

[1 I dati GNSS 3](#_Toc178870129)

[1.1 Data Pipeline 4](#_Toc178870130)

[1.2 Serie temporali GNSS 6](#_Toc178870131)

[1.3 Variazione areale 7](#_Toc178870132)

[2 Analisi dei dati GNSS 11](#_Toc178870133)

[2.1 Periodo 11](#_Toc178870134)

[2.2 Suddivisione area vulcanica 12](#_Toc178870135)

[2.3 Eventi vulcanici 14](#_Toc178870136)

[3 Applicazione del Machine learning 16](#_Toc178870137)

[4 Uso del modello per predizione 17](#_Toc178870138)

[5 Conclusioni 18](#_Toc178870139)

[RINGRAZIAMENTI 19](#_Toc178870140)

[INDICE DELLE FIGURE 20](#_Toc178870141)

[INDICE DEI GRAFICI 21](#_Toc178870142)

[INDICE DELLE TABELLE 22](#_Toc178870143)

[BIBLIOGRAFIA 23](#_Toc178870144)

[SITOGRAFIA 23](#_Toc178870145)

# Introduzione

Studio di eventi vulcanici con tecniche moderne per ridurre i tempi e migliorare la precisione, per predire eventi futuri, sapere con quale probabilità può accadere un certo evento. Al fine del monitoraggio introdurre automatismi che avvisino il personale interessato.

Scopo della tesi: creare un modello di machine learning in grado di riconoscere gli eventi vulcanici a partire dalle serie temporali GNSS, e predire eventi futuri con un minimo di preavviso.

Creeremo a partire dalle serie temporali GNSS delle singole stazioni le aree formate da tre stazioni adiacenti in modo da studiare l’edificio vulcanico come un susseguirsi di aree di triangoli, e ne studieremo la variazione nel tempo.

La prima fase del lavoro svolto prevede la gestione dei dati generati dalle stazioni della rete permanente GNSS gestita dall’Istituto Nazionale di Geofisica e Vulcanologia – Osservatorio Etneo, nello specifico dai segnali appartenenti al vulcano Etna.

Processamento e pulizia dei dati da outliers, salti, etc..

Calcolo della variazione areale in ppm (parti per milione) e creazione delle serie temporali di 20 aree che coprono tutta la superficie del vulcano utile allo studio condotto, soggetto alle deformazioni del suolo.

Applicazione delle tecniche classiche di Machine Learning per creare un modello in grado di fare la detection dell’evento vulcanico (classificatore binario) e comparazione degli score ottenuti.

Scelta del modello che presenta i migliori risultati in termini di accuratezza e precisione (matrice di confusione).

Utilizzare il modello scelto come predittore, aggiungendo una nuova classe, il pre-evento (il giorno che precede l’evento).

Commento sui risultati ottenuti e conclusioni finali; possibilità di miglioramento.

# Il vulcano Etna

PROBLEMA (RISCHIO VULCANICO): I vulcani, oggetto di pericoli e disastri, fenomeni naturali, popolazioni vicine, natural hazard. Durante le eruzioni, cenere vulcanica, aeroporto (blocco trasporti), disagi alla città. Dicco del 24.12.2018, terremoto del 26.12

SOLUZIONE: implementare un sistema per la previsione (early warning) ai fini del monitoraggio, …

Vulcano etna, iniquadramento dell’area, tipo di attività, cosa sono le fontane di lava, eventi passati, i dicchi, scivolamento del fianco orientale,

mappa della sicilia / italia per contestualizzare l’etna

## Monitoraggio geodetico

## Hazard

# I dati GNSS

La geodesia è una disciplina appartenente alle scienze della terra che si occupa della misura e della rappresentazione della Terra, del suo campo gravitazionale e dei fenomeni geodinamici, avvalendosi di misure astronomiche, gravimetriche e trigonometriche. Il monitoraggio geodetico permette di misurare il cambiamento nel tempo della superficie terrestre allo scopo di tenere sotto controllo i rischi annessi ai movimenti e alle deformazioni del suolo. Quest’ultimo è attuato dalla Sezione di Catania dell’Istituto Nazionale di Geofisica e Vulcanologia (INGV CT) sul territorio di competenza, per fini di monitoraggio tettonico e vulcanico, tramite la Rete GNSS (Global Navigation Satellite System). Essa è formata, allo stato attuale, da circa 100 stazioni GNSS permanenti (comprendente cioè sistemi in grado di ricevere segnali satellitari di tipo GPS, Galileo, Glonass, BeiDou) dislocate in Sicilia e Calabria Meridionale che sono collegate in tempo reale con la Sala Acquisizione Dati GPS INGV CT [s2].

Il lavoro di questa tesi si basa sui dati acquisiti dalla suddetta rete prodotti dalle stazioni GNSS presenti all’interno del territorio del vulcano Etna (Figura 1‑1), al fine di costruire un modello di Intelligenza Artificiale che possa associare le deformazioni del suolo ad eventi vulcanici quali eruzioni, intrusioni di magma o fontane di lava. Il punto di partenza è stato quindi l’analisi e la gestione dei dati GNSS



Figura 2‑1 Mappa delle stazioni GNSS della rete permanente dell'Etna gestita dall'Osservatorio Etneo

## Data Pipeline

Un sistema satellitare globale di navigazione (in inglese **G**lobal **N**avigation **S**atellite **S**ystem, acronimo GNSS) è un sistema di geo-radiolocalizzazione e navigazione terrestre, marittima o aerea, che utilizza una rete di satelliti artificiali in orbita. Un tale sistema fornisce un servizio di posizionamento geo-spaziale a copertura globale che consente a piccoli ed appositi ricevitori elettronici di determinare le loro coordinate geografiche (longitudine, latitudine ed altitudine) su un qualunque punto della superficie terrestre o dell'atmosfera, elaborando segnali a radiofrequenza trasmessi in linea di vista da tali satelliti [s1]. Tali segnali vengono acquisiti da un antenna ed elaborati da appositi apparati, i ricevitori, che insieme al sistema di alimentazione e trasmissione costituiscono gli elementi di una tipica stazione GNSS (Figura 1‑2).

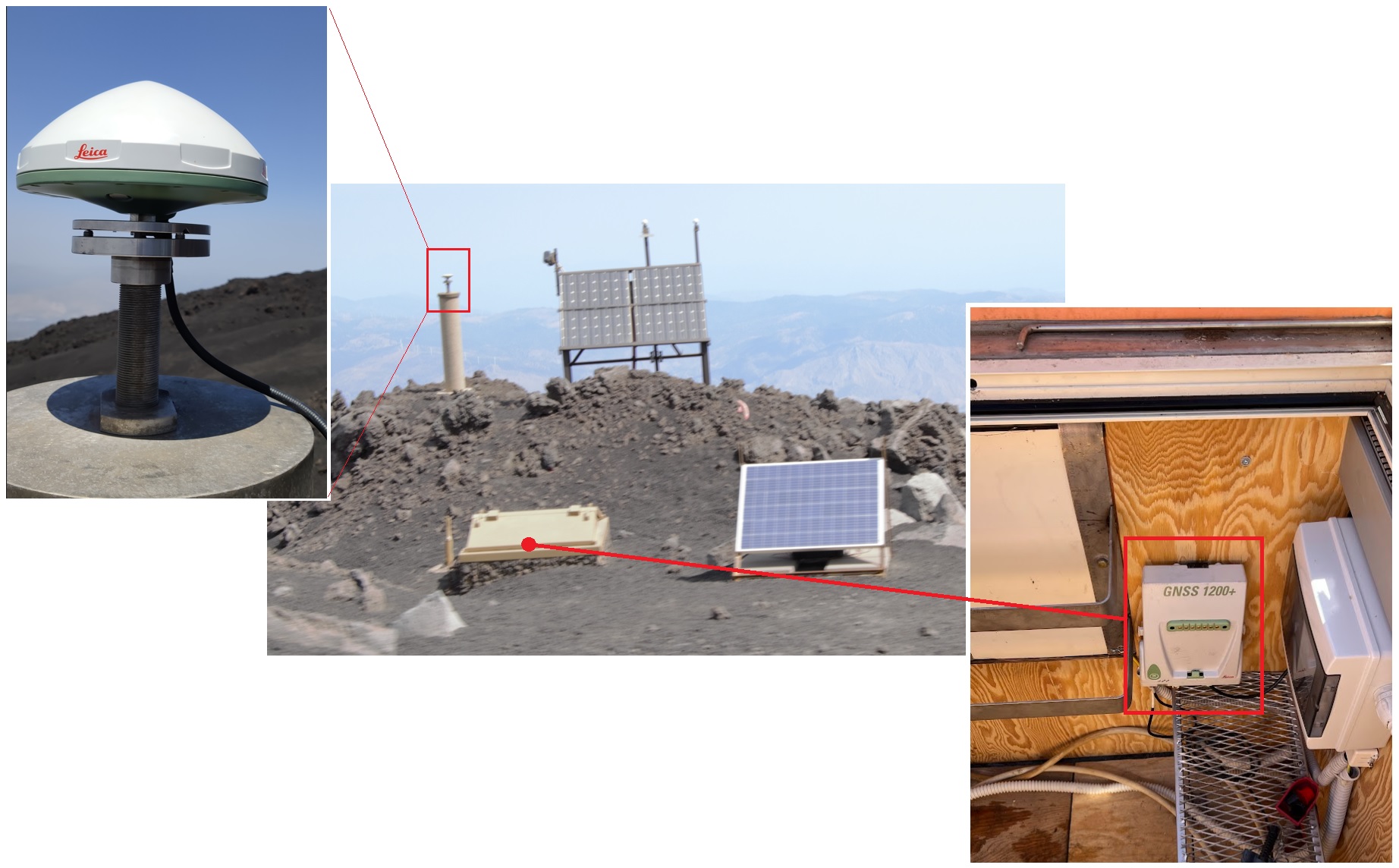


Figura 2‑2 Stazione GNSS presente sull’Etna costituita da antenna (a sinistra), ricevitore (a destra) e altri apparati.

La pipeline dei dati riferiti a tali segnali è di tipo ELT[[1]](#footnote-1): i segnali acquisiti vengono infatti inviati alla sala di acquisizione dati dell’Osservatorio Etneo dove, tramite il software GNSS Spider della Leica®, vengono generati i files nel formato standard RINEX[[2]](#footnote-2). Essi rappresentano il Data Lake, cioè la fonte dei dati da cui vengono generate le serie temporali GNSS.

## Serie temporali GNSS

L’unità funzionale Deformazioni, Geodesia e Geofisica (UFDGG) dell’Osservatorio Etneo si occupa dell’elaborazione dei dati acquisiti in tempo reale dalle stazioni GNSS attraverso l’utilizzo del software Gamit/Globk, sviluppato dal MIT[[3]](#footnote-3), con frequenze di campionamento di 1 ora e di 1 giorno. Il processamento dei dati ha lo scopo di filtrare il segnale ricevuto da tutti quegli elementi di disturbo del segnale stesso (multipath, outliers, rumore). La Figura 1‑3 mostra la serie temporale della posizione di una delle stazioni sommitali dell’Etna, nella sua componente Nord, processata con frequenza di campionamento giornaliera, riferita all’ultimo anno (settembre 2023 – settembre 2024):

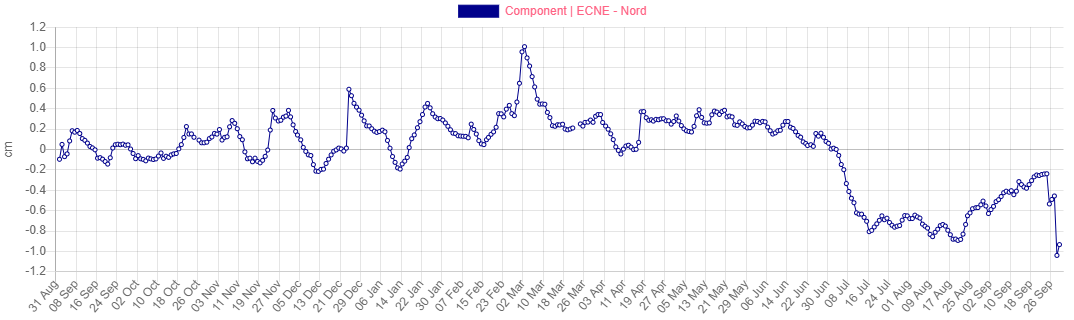


Grafico 2‑1 Serie temporale dell’ultimo anno della componente Nord della stazione ECNE (Cratere di Nord Est - Etna)

Lo studio di tali segnali sta alla base sia del monitoraggio geodetico che di attività di ricerca scientifica, tra cui lo studio delle deformazioni del suolo. Esse rappresentano la base di partenza del lavoro svolto per questa tesi, che ha l’obiettivo di cercare una correlazione tra la deformazione del suolo e il verificarsi di eventi vulcanici, quali eruzione vulcanica, fontane di lava, intrusioni magmatiche. La deformazione viene quindi misurata non solo sulle tre componenti delle singole stazioni (Nord, Est e quota), ma anche come distanze tra coppie di stazioni (Baselines) o variazioni dell’area di triagoli costituiti da stazioni GNSS. L’area geografica oggetto di questo studio è stata pertanto suddivisa in tanti triangoli adiacenti ai cui vertici ci sono le stazioni GNSS gestite dall’Osservatorio Etneo.

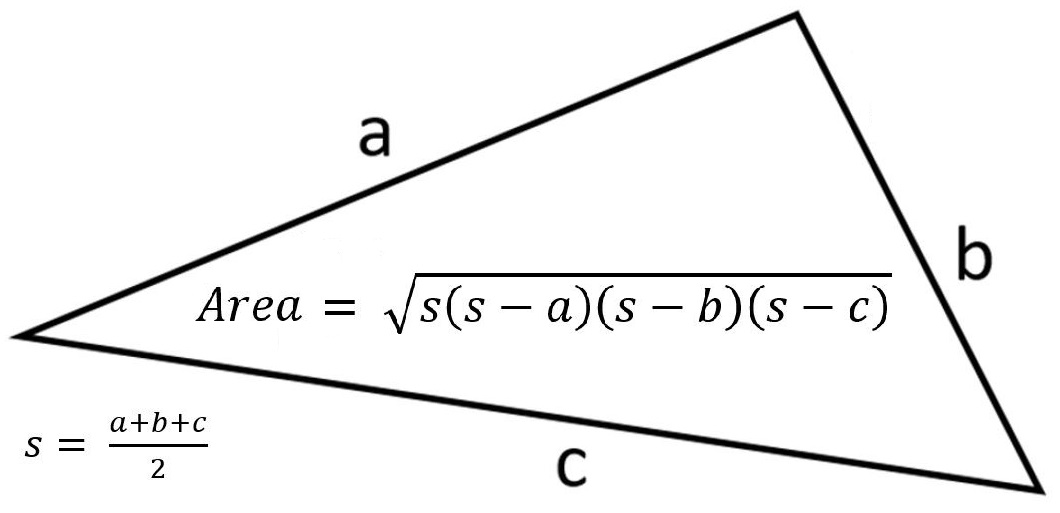
## Variazione areale

A titolo esemplificativo viene mostrata in Figura 1‑4 l’area del triangolo composto dalle stazioni EINT (Intermedia), ESLN (Serra la Nave) e ESPC (Serra Pizzuta Calvarina).



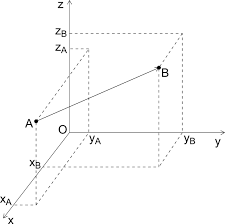
Figura 2‑3 Area del triangolo formato dalle stazioni EINT (Intermedia), ESLN (Serra la Nave) e ESPC (Serra Pizzuta Calvarina)

Per il calcolo dell’area del triangolo è stata utilizzata la formula di Erone, che utilizza il valore di ogni singolo lato del triangolo.



Formula 1‑1 Formula di Erone per il calcolo dell'area di un triangolo

Dato che i vertici del triangolo sono rappresentati dalla posizione delle tre stazioni che lo compongono, e che queste vengono elaborate dal ricevitore come coordinate nello spazio X-Y-Z, è stata adottata la distanza euclidea tri-dimensionale per calcolare la distanza tra due stazioni:



A = () B = ()

Distanza tra A e B =

Formula 1‑2 Distanza euclidea tra due punti nello spazio XYZ

Ritornando all’esempio del triangolo EINT-ESLN-ESPC, applicando la Formula 1‑2 alle coppie di stazioni EINT-ESLN, EINT-ESPC e ESLN-ESPC e quindi applicando la Formula 1‑1 otteniamo l’area del triangolo.

Considerando il fatto che la suddivisione in triangoli di tutta l’area studiata comporta l’avere triangoli di dimensioni diverse, si è deciso di determinare la variazione areale non in valore assoluto (es. cm2) ma in valore relativo, quindi espressa in “parti per milione” (ppm). Infine, se calcoliamo l’area del triangolo per ciascun giorno dell’ultimo anno otteniamo la serie temporale dell’area EINT-ESLN-ESPC, mostrata nel Grafico 1‑2.

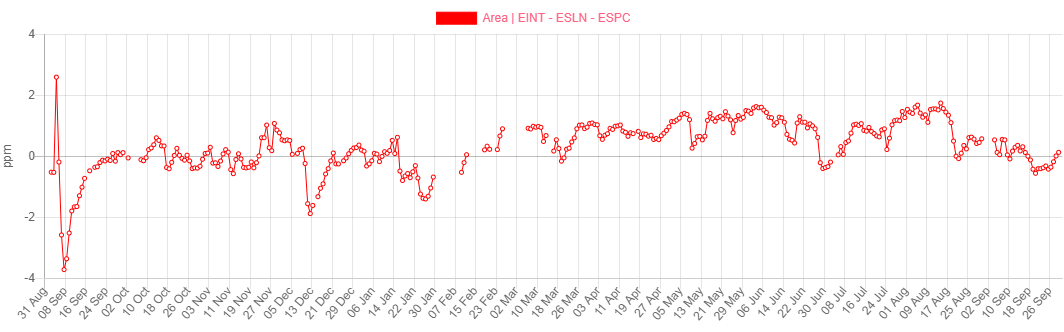


Grafico 2‑2 Serie temporale dell'area del triangolo formato dalle stazioni EINT-ESLN-ESPC

La variazione areale è da considerarsi come il risultato delle variazioni delle tre baseline[[4]](#footnote-4) formate dai tre vertici del triangolo. La serie temporale vista sopra rappresenta quindi la combinazione delle tre serie temporali delle baseline ESLN-ESPC, EINT-ESPC e EINT-ESLN, mostrate nella Figura 1‑8.

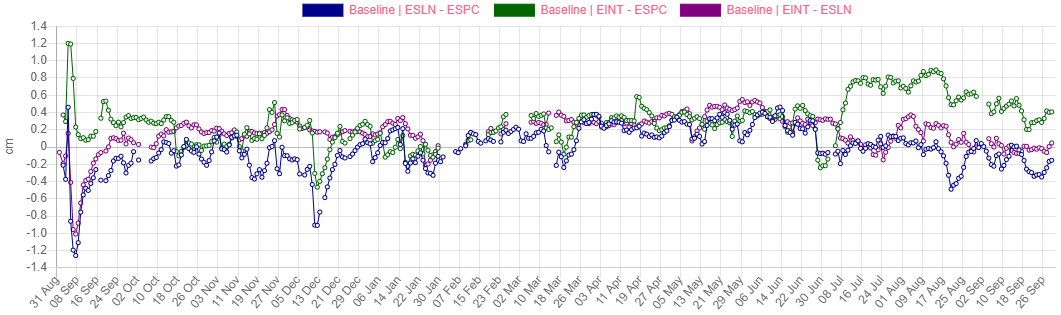


Grafico 2‑3 Serie temporali delle baseline costruite a coppie dei vertici del triangolo ESLN-EINT-ESPC

Trattare dunque una sola serie temporale piuttosto che tre riduce notevolmente la complessità dell’analisi dei dati GNSS.

# Analisi dei dati GNSS

Il periodo di riferimento utilizzato per lo studio e la creazione del modello di Intelligenza Artificiale proposto in questa tesi è stato scelto in funzione dell’attività vulcanica dell’Etna registrata negli ultimi anni dall’Osservatorio Etneo: il periodo selezionato è quello che ha registrato il più intenso processo di inflazione e deflazione al monte Etna degli ultimi 20 anni.

## Periodo

L’analisi dei dati GNSS acquisiti dalle reti di monitoraggio permanente del vulcano Etna ha permesso di studiare approfonditamente le deformazioni del suolo avvenute negli ultimi 20 anni, dal momento cioè in cui è stata creata la rete. La figura Figura 1‑9 mostra la variazione areale del triangolo formato dalle stazioni EDAM (Dammusi), EMEG (Monte Egitto) e EINT (Intermedia) per il periodo 01 gennaio 2019 – 30 aprile 2021, in cui si possono distinguere 4 fasi di inflazione[[5]](#footnote-5) ed una fase finale di deflazione[[6]](#footnote-6), che seguono l’eruzione vulcanica del 24 dicembre 2018 caratterizzata da un’intrusione magmatica sul fianco orientale del vulcano [b1].

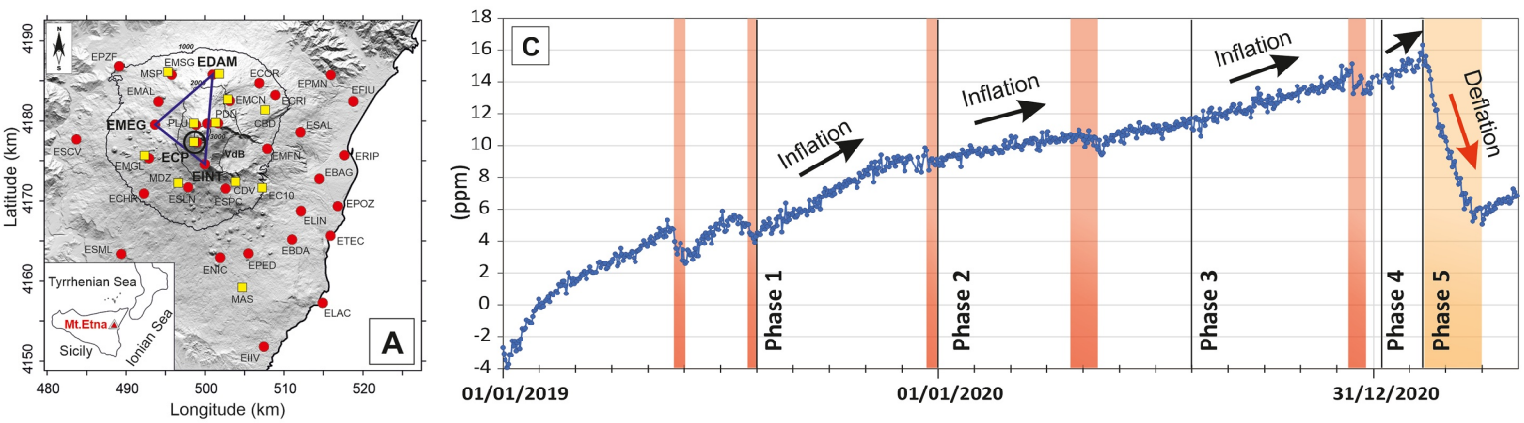


Figura 3‑1 Variazione areale del triangolo EDAM-EMEG-EINT che mostra le fasi di inflazione e deflazione registrate nel periodo 01.01.2019 - 30.04.2021 [b1]

Per il lavoro di questa tesi è stato scelto questo periodo di tempo, nello specifico dal 1° novembre 2018 al 30 aprile 2021 che, come già anticipato, comprende un’intrusione magmatica, decine di eruzioni vulcaniche e ben 17 fontane di lava avvenute nell’ultima fase, quella di deflazione.

## Suddivisione area vulcanica

Individuato il periodo da analizzare si è proceduto con la suddivisione dell’area vulcanica in triangoli adiacenti aventi come vertici le stazioni GNSS della rete permanente INGV-OE. I triangoli ricavati da questa suddivisione sono stati 20, e sono elencati nella Tabella 1

|  |  |
| --- | --- |
| 1) EDAM-ECNE-EPLU  2) EDAM-ECOR-EMCN  3) EDAM-EMSG-EPLU  4) EDAM-EPDN-ECNE  5) EDAM-EPDN-EMCN  6) EINT-ECPN-EMGL  7) EINT-ESLN-EMGL  8) EINT-ESLN-ESPC  9) EMEG-ECPN-EMGL  10) EMFN-ECOR-ECRI | 11) EMFN-ECOR-EMCN  12) EMFN-EMCN-EPDN  13) EPDN-ECNE-ECPN  14) EPDN-EINT-ECPN  15) EPDN-EINT-ESPC  16) EPDN-EMFN-ESPC  17) EPLU-ECNE-ECPN  18) EPLU-EMEG-ECPN  19) EPLU-EMEG-EMAL  20) EPLU-EMSG-EMAL |

Tabella 3‑1 Elenco dei triangoli scelti per la suddivisione dell'area vulcanica analizzata

La Figura 2‑2 mostra la suddivione dell’area vulcanica nei 20 triangoli elencati sopra.

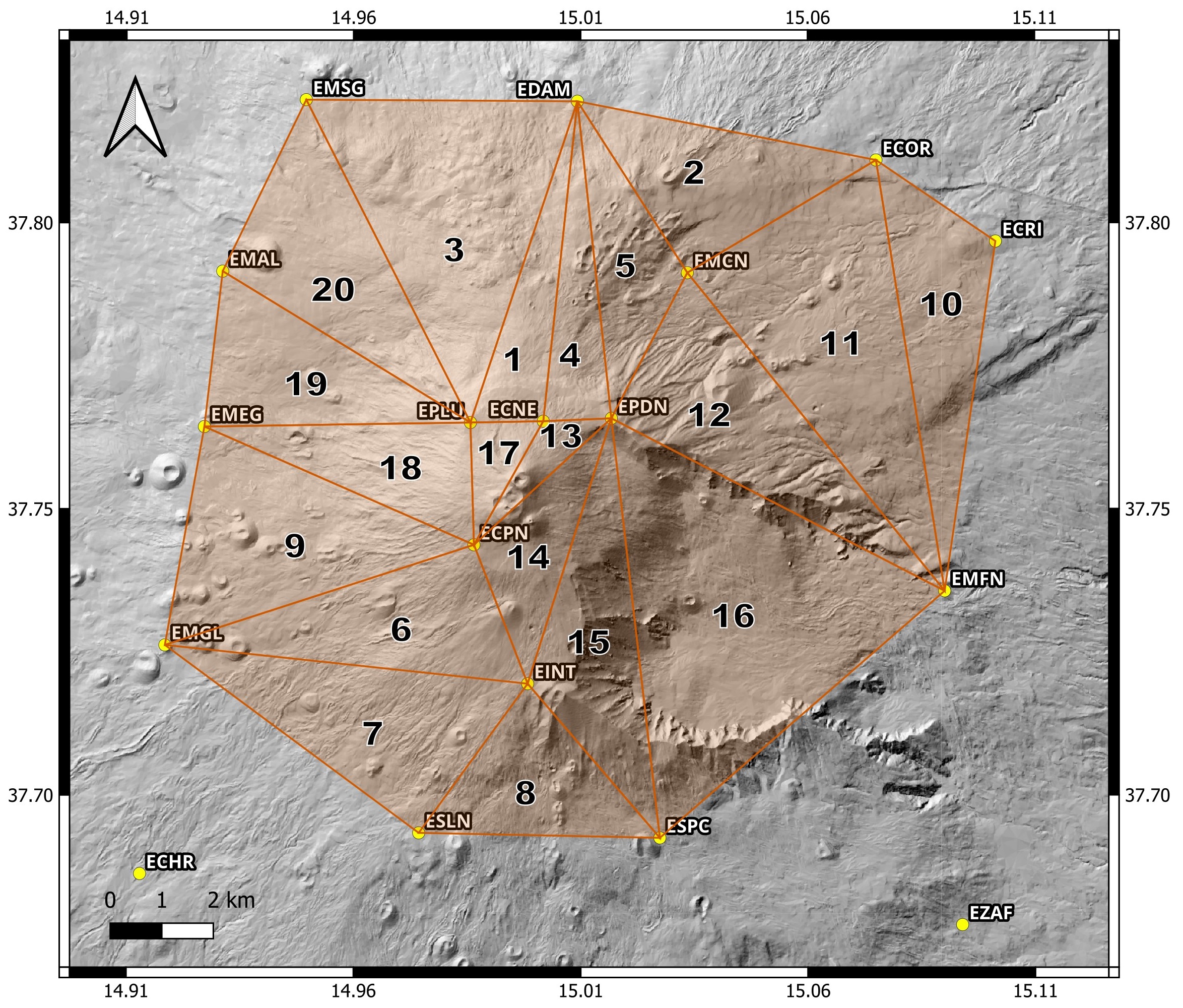


Figura 3‑2 Suddiviosne dell'area vulcanica in triangoli adiacenti.

Definiti i triangoli si è proceduto con il calcolo delle variazioni areali nel periodo scelto, applicando l’interpolazione dei dati per popolare i punti di discontinuità (assenza del dato), generando pertanto le serie temporali per ciascun triangolo, così come riportato nell’esempio descritto precedentemente e rappresentato nel Grafico 1‑2.

Nella Grafico 2‑1 sono mostrate nello stesso grafico le serie temporali di tutti i triangoli individuati.

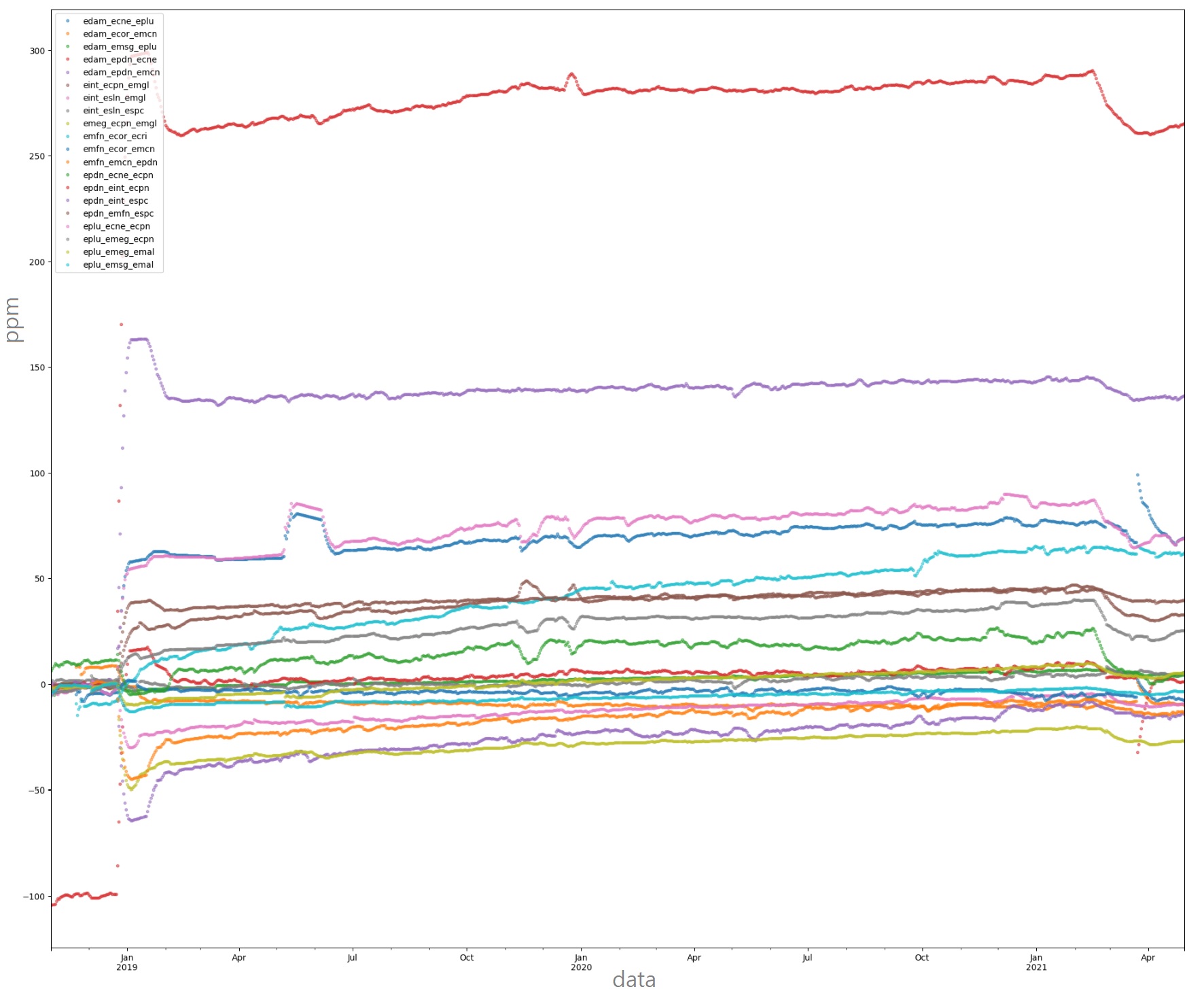


Grafico 3‑1 Serie temporali della variazione areale per i 20 triangoli aventi per vertici le stazioni GNSS dell'Etna

## Eventi vulcanici

Al fine di creare un modello di Intelligenza Artificiale che possa individuare automaticamente l’occorrere di eventi vulcanici si è deciso di utilizzare algoritmi tecnica di apprendimento supervisionato: il modello viene istruito quindi a predire i valori di uscita rispetto agli input (i *sample*s) sulla base di ouput noti, chiamati *target*, che gli vengono forniti. Questa fase del lavoro è stata quindi interessata dall’individuare tutti gli eventi vulcanici per cui vogliamo che il nostro sistema sia in grado di riconoscere automaticamente, sulla base delle ricerche scientifiche condotte dagli operatori dell’Istituto Nazionale di Geofisica e Vulcanologia. Il rusultato di tale ricerca è mostrato nella tabella Tabella 2‑2.

|  |  |  |
| --- | --- | --- |
| Data | Tipo di evento | Riferimento |
| 24/12/2018  21/01-03/02/2019  29/04-05/05/2019  30/05-02/06/2019  18-21/07/2019  27/07-04/08/2019  25-30/12/2019  19/04/2020  07-12/12/2020  13/12/2020  14/12/2020  21/12/2020  22/12/2020  18/01/2021  16/02/2021  18/02/2021  19/02/2021  21/02/2021  23/02/2021  24/02/2021  28/02/2021  02/03/2021  04/03/2021  07/03/2021  10/03/2021  12/03/2021  15/03/2021  17/03/2021  19/03/2021  24/03/2021  01/04/2021 | DI – Dike Intrusion  EA – Eruptive Activity  EA – Eruptive Activity  EA – Eruptive Activity  EA – Eruptive Activity  EA – Eruptive Activity  EA – Eruptive Activity  TA – Transactional Activity  EA – Eruptive Activity  TA – Transactional Activity  EA – Eruptive Activity  TA – Transactional Activity  TA – Transactional Activity  TA – Transactional Activity  LSLF – Large-Scale Lava Fountain  LF – Lava Fountain  LF – Lava Fountain  LF – Lava Fountain  LSLF – Large-Scale Lava Fountain  LF – Lava Fountain  LSLF – Large-Scale Lava Fountain  LF – Lava Fountain  LSLF – Large-Scale Lava Fountain  LSLF – Large-Scale Lava Fountain  LF – Lava Fountain  LSLF – Large-Scale Lava Fountain  LF – Lava Fountain  LF – Lava Fountain  LSLF – Large-Scale Lava Fountain  LF – Lava Fountain  LF – Lava Fountain | [b4]  [b3]  [b3]  [b3]  [b3]  [b3]  [b3]  [b2]  [b3]  [b2]  [b3]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2]  [b2] |

Tabella 3‑2 Elenco degli eventi vulcanici usati come target del modello supervisionato.

Al Grafico 2‑1 aggiungiamo adesso le occorrenze degli eventi vulcanici in modo da sovrapporre le serie temporali della variazione areale al target: il risultato è mostrato nel grafico Grafico 2‑2.

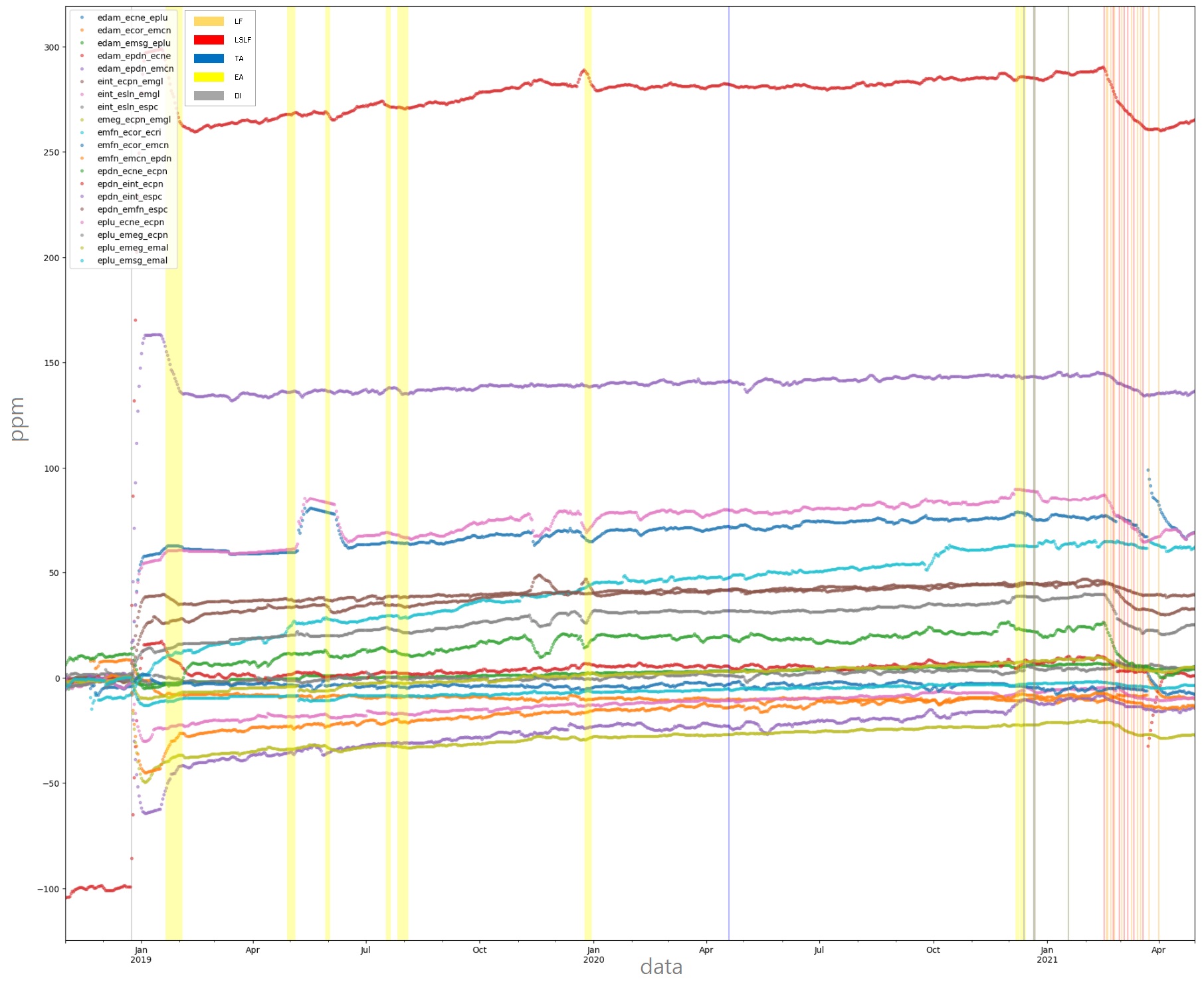


Grafico 3‑2 Serie temporali della variazione areale per i 20 triangoli sovrapposte all'occorrenza degli eventi vulcanici.

## Variazione relativa

L’analisi del Grafico 2‑2 è stata utile per definire il tipo di dato da fornire agli algoritmi di Machine Learning al fine di creare il modello che metta in relazione la deformazione del suolo, vista come variazione areale, e il manifestarsi di eventi vulcanici. Si è infatti ritenuto più efficace cercare la correlazione tra gli eventi vulcanici non con la variazione areale assoluta dei singoli triangoli, ma piuttosto con la loro variazione areale relativa, intesa come variazione positiva o negativa rispetto ai giorni precedenti. Si è quindi proceduto con l’applicazione della *feature engineering*[[7]](#footnote-7) creando nuove serie temporali a partire dalla variazione areale, applicando la seguente formula:

in cui

La variazione areale relativa di un determinato giorno è pertanto la differenza tra la variazione areale di quel giorno rispetto alla media della variazione areale dei due giorni precedenti. Il significato da dare ai valori di VR è il seguente: un valore positivo rappresenta un’inflazione, mentre un valore negativo rappresenta una deflazione.

Vengono riportate, a titolo esemplificativo, le serie temporali della variazione areale assoluta, Grafico 3‑3, e della variazione areale relativa, Grafico 3‑4, entrambe riferite al triangolo EPLU-ECNA-ECPN in modo da poter vedere le differenze tra le due serie temporali.

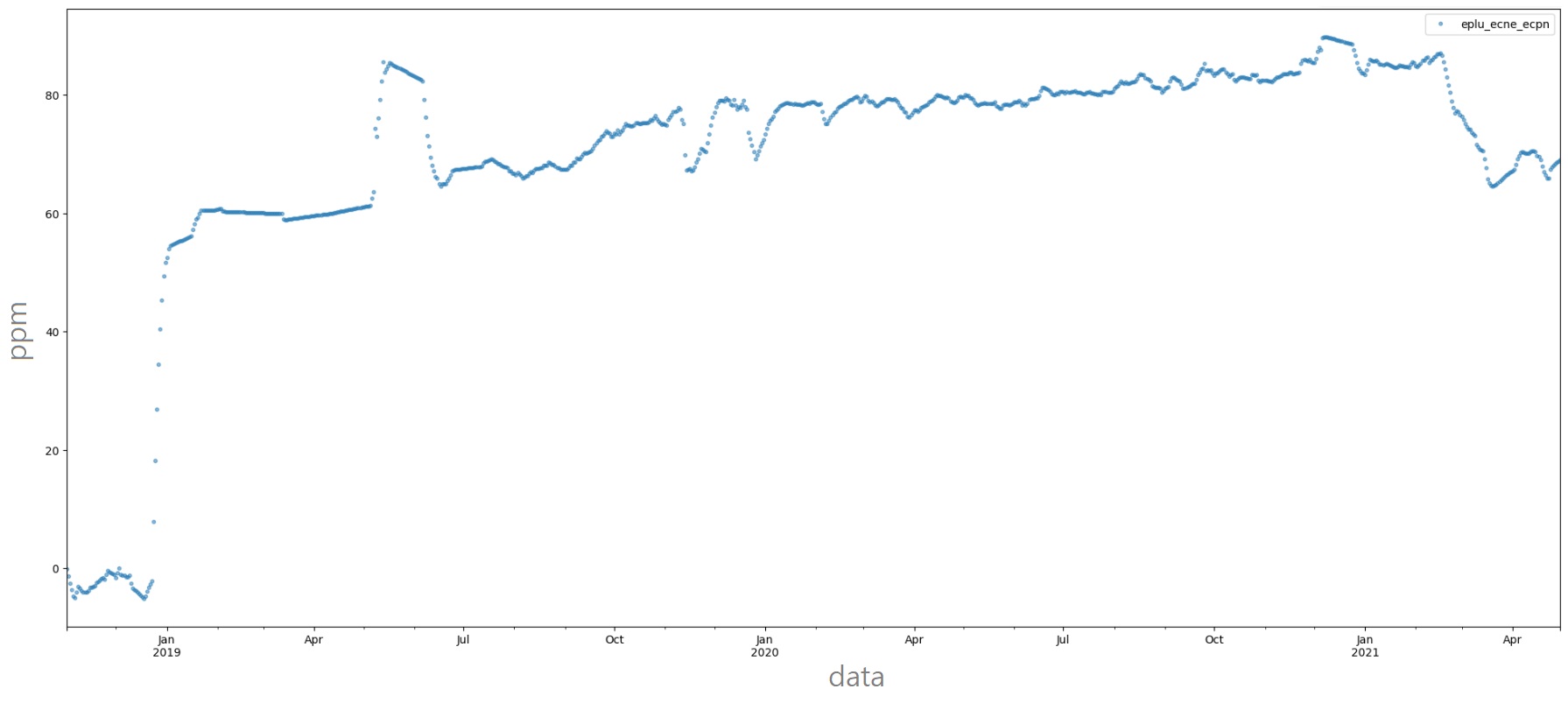


Grafico 3‑3 Serie temporale variazione areale assoluta del triangolo EPLU-ECNE-ECPN

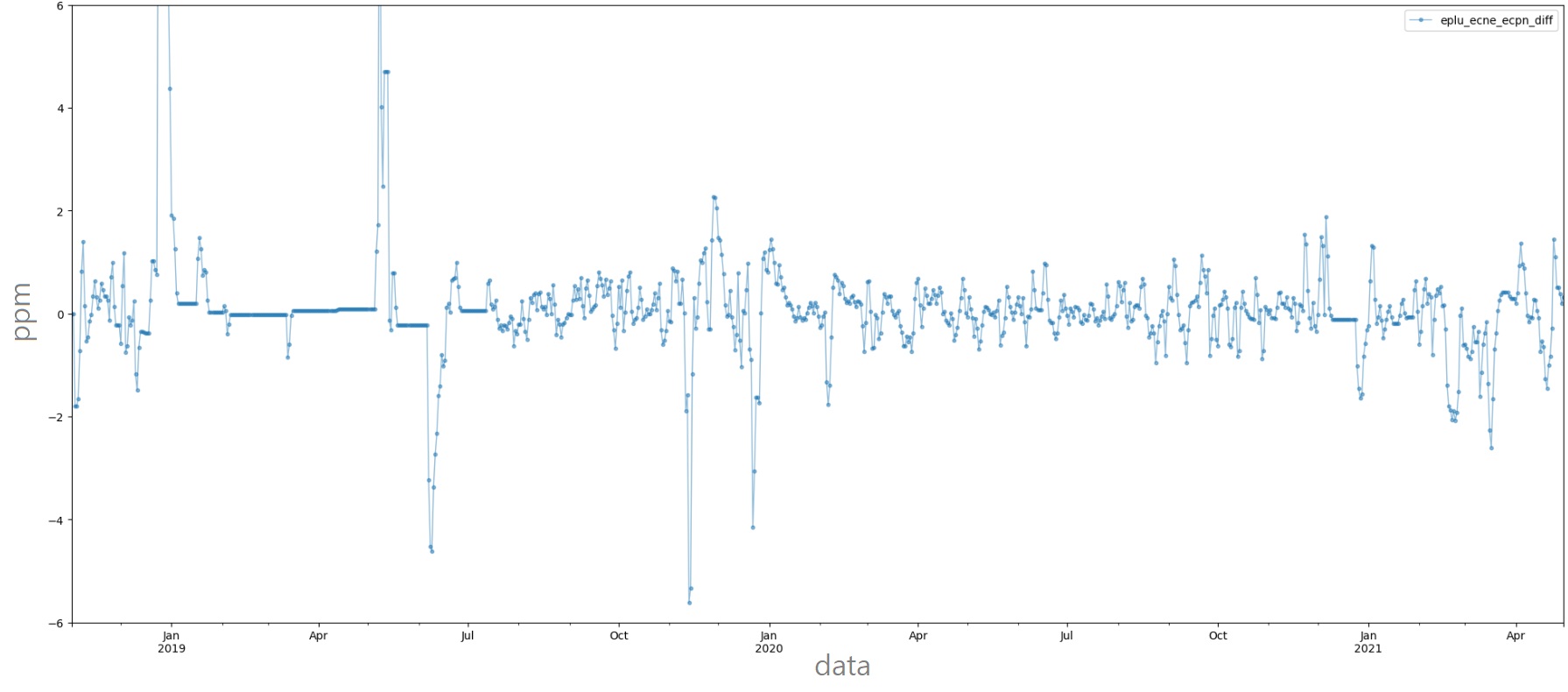


Grafico 3‑4 Serie temporale variazione areale relativa del triangolo EPLU-ECNE-ECPN

La caratteristica importante della serie temporale del Grafico 3‑4 è che è centrata sullo zero, quindi i valori positivi e negativi sono proporzionali alla repentinità della variazione areale, elemento su cui sarà allenato il modello di Machine Learning proposto in questa tesi.

## Il dataset

La scelta del periodo storico da studiare, la suddivisione dell’area vulcanica in triangoli adiacenti e l’applicazione della *feature engineering* hanno permesso di definire il dataset di partenza: esso infatti è costituito da 912 campioni (numero di giorni presenti dal 01/11/2018 al 30/04/2021), e ciascun campione è formato da 20 features (le variazioni areali relative dei 20 triangoli).

L’individuazione degli eventi vulcanici elencati nella Tabella 2‑2 ha permesso di definire il target, che è costituito da 2 classi:

* classe 0: nessun evento vulcanico
* classe 1: evento vulcanico

L’obiettivo del lavoro svolto per questa tesi è quindi quello di definire un classificatore che sia in grado di predire l’occorrenza di eventi vulcanici a partire dalla variazione areale relativa dell’edificio vulcanico.

### Bilanciamento classi – undersampling

La distribuzione delle classi per il dataset di partenza, costituito da 912 campioni, è mostrata nella Tabella 3‑3.

|  |  |  |
| --- | --- | --- |
| Classe | *N° occorrenze* | *% sul totale campioni* |
| 0 | 838 | 91,8% |
| 1 | 74 | 8,2% |

Tabella 3‑3 Distribuzione delle classi per il dataset di partenza

La distribuzione delle classi risulta pertanto essere sbilanciata. La conseguenza derivante dallo sbilanciamento delle classi è la distorsione del modello verso la previsione della classe maggioritaria [s3]. Le soluzioni maggiormente adottate per risolvere questo problema sono:

1. Sovracampionamento della classe minoritaria (oversampling)
2. Sottocampionamento della classe maggioritaria (undersampling)
3. Generare dati artificiali
4. Dare un peso diverso alle classi

Si è pertanto scelto di adottare, come soluzione al problema dello sbilanciamento, il sottocampionamento della classe maggioritaria, scegliendo come giorni appartenenti alla “classe 0” solo quelli antecedenti, in ordine cronologico, all’occorrenza della “classe 1”, nello specifico selezionando solo 8 giorni prima dell’evento vulcanico. Così facendo il modello sarà allenato a predire gli eventi vulcanici sulla base delle deformazioni del suolo che precedono gli eventi stessi.

Il processo di riduzione dei campioni è mostrato nel Grafico 3‑5.

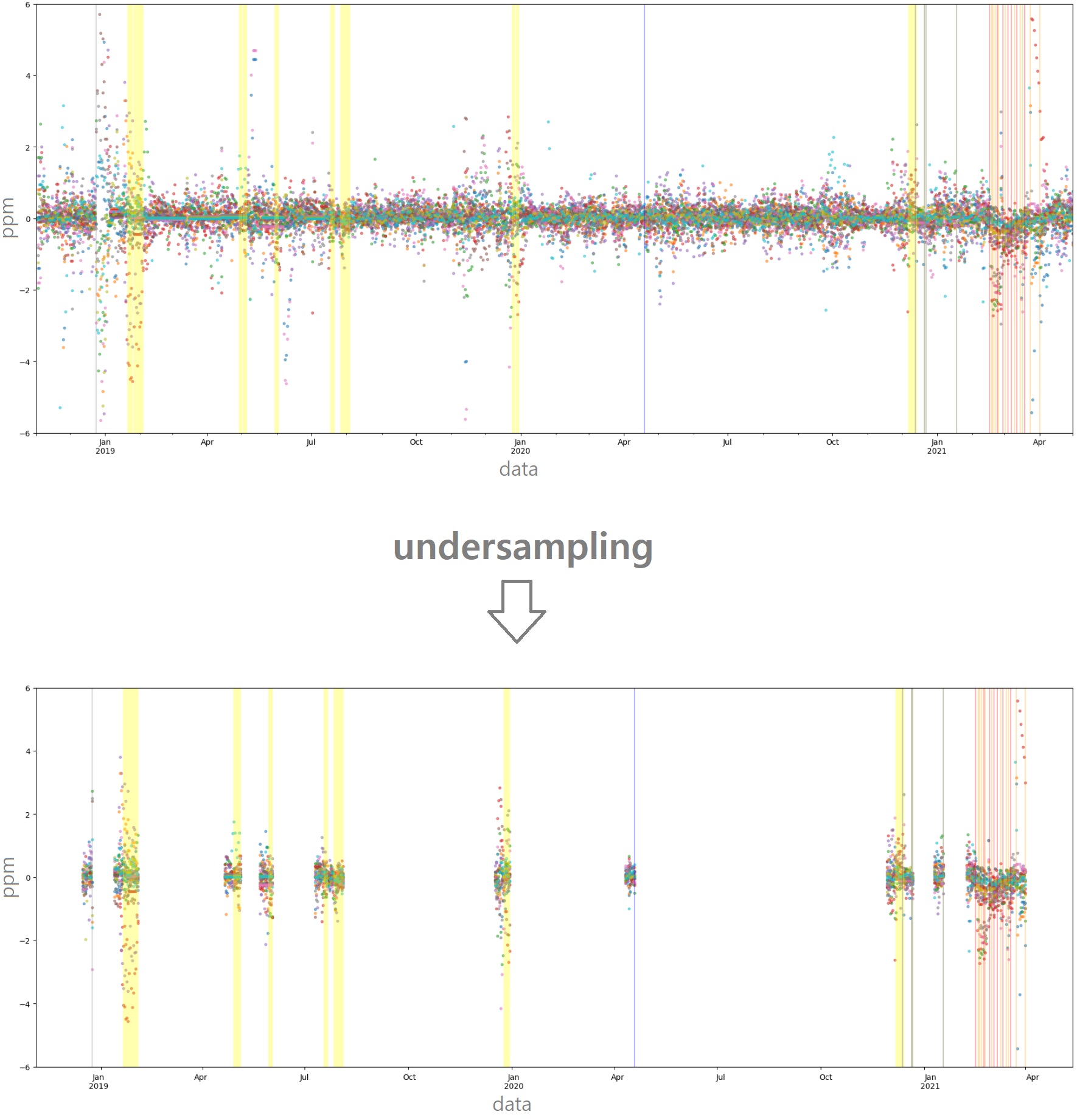


Grafico 3‑5 Processo di riduzione dei campioni per il bilanciamento delle classi: le barre verticali rappresentano l'occorrenza degli eventi vulcanici, i punti colorati rappresentano i campioni diversificati per features.

L*’undersampling* ha ridotto il numero totale di campioni da 912 a 189: essi rappresentano il dataset definitivo, e la distribuzione del target rispetto alle classi è mostrato nella tabella Tabella 3‑4.

|  |  |  |
| --- | --- | --- |
| Classe | *N° occorrenze* | *% sul totale campioni* |
| 0 | 115 | 60,8% |
| 1 | 74 | 39,2% |

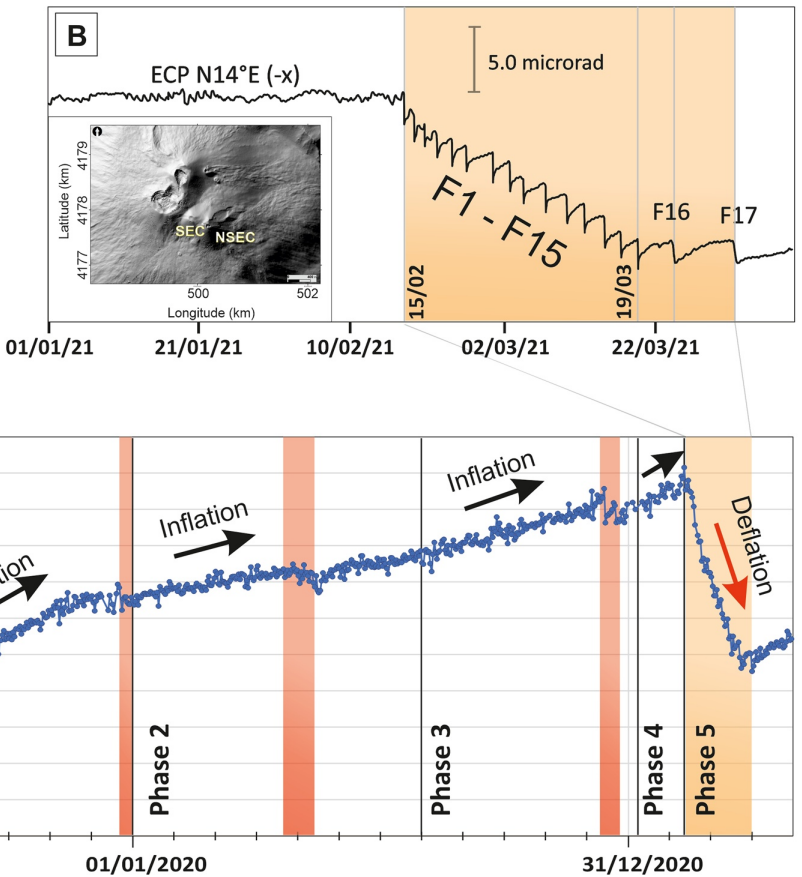
Tabella 3‑4 Distribuzione delle classi per il dataset definitivo

### Training e Test set

L'apprendimento di un modello supervisionato (e.g. un classificatore, un regressore) viene effettuato sulla base di un insieme di addestramento, il **training set**. I modelli supervisionati, che mirano a cercare relazioni empiriche tra esempi dell'insieme di addestramento, possono dar vita al fenomeno del sovradattamento a tale insieme. Ciò significa che essi potrebbero identificare relazioni che valgono nell'insieme di addestramento ma non in generale (*overfitting*), per l'intera popolazione. Pertanto, per verificare se le relazioni empiriche apprese possano avere una validità generale, il modello appreso va valutato su un **test set** disgiunto dall'insieme di addestramento [s4]. Pertanto si è scelto di suddividere il dataset come segue:

* Training set, formato da 126 campioni su 189 (66,7%)
* Test set, formato da 63 campioni su 189 (33,3 %)

Il criterio con cui sono stati creati i due insiemi tiene conto della singolare distribuzione delle tipologie di eventi vulcanici verificate nel periodo storico scelto per questa tesi (vedi paragrafo 3.1). L’ultima parte del periodo scelto, dal 16/02/2021 al 01/04/2021, è caratterizzata infatti da una sequenza di 17 fontane di lava (b1, Figura 3‑3), tipologia di evento manifestata solo in questo periodo.

Figura 3‑3 Sequenza delle 17 fontane manifestate nella fase di deflazione (b1)

E’ stato necessario quindi suddividere il dataset in modo tale che il training set e il test set contenessero entrambi i dati di tale periodo ed in proporzione alla loro percentuale rispetto al totale dei campioni. La Figura 3‑4 mostra la suddivisione del dataset in Training e Test set.



Figura 3‑4 Suddivisione del dataset in training e Test set

### Standardizzazione

Definita la suddivisione del dataset tra training e test set si è proceduto con la standardizzazione dei dati, tecnica usata nel machine learning in fase di preparazione dei dati per garantire una scala e una ponderazione simili per tutti gli attributi (b6). L’effetto della standardizzazione è quello di scalare i dati in modo che abbiano media 0 e deviazione standard pari a 1. La formula applicata è la seguente:

In cui:

Dato che il training set è utilizzazto per allenare il modello di machine learning, la standardizzazione si applica ad esso, calcolandone i valori di media e deviazione standard per ciascuna feature. Nella Tabella 3‑5 sono riportati tali valori. Questi sono gli stessi con cui il test set viene normalizzato.

|  |  |  |
| --- | --- | --- |
| **Feature** |  |  |
| EDAM-ECNE-EPLU  EDAM-ECOR-EMCN  EDAM-EMSG-EPLU  EDAM-EPDN-ECNE  EDAM-EPDN-EMCN  EINT-ECPN-EMGL  EINT-ESLN-EMGL  EINT-ESLN-ESPC  EMEG-ECPN-EMGL  EMFN-ECOR-ECRI  EMFN-ECOR-EMCN  EMFN-EMCN-EPDN  EPDN-ECNE-ECPN  EPDN-EINT-ECPN  EPDN-EINT-ESPC  EPDN-EMFN-ESPC  EPLU-ECNE-ECPN  EPLU-EMEG-ECPN  EPLU-EMEG-EMAL  EPLU-EMSG-EMAL | 0.144713  0.096944  -0.034579  -0.337091  0.001546  -0.119565  -0.047072  0.008210  -0.108813  0.140302  0.034234  -0.055439  -0.166965  -0.180316  -0.069271  -0.066864  -0.191393  -0.063500  0.002059  0.004627 | 3.125811  0.898655  0.159112  3.453219  0.490204  0.585334  0.335151  0.407505  0.304158  0.556443  0.871274  0.362596  0.647669  0.792698  0.418900  0.309505  0.779531  0.549749  0.167340  0.136999 |

Tabella 3‑5 Valori di Media () e Deviazione Standard (σ) calcolati sul training set per ciascuna feature.

# Applicazione del Machine learning

Come già scritto nei capitoli precedenti, l’obiettivo proposto per questa tesi è quello di creare un modello di Machine Learning in grado predire il verificarsi di eventi vulcanici a partire dalla variazione areale dell’edificio vulcanico. Si è pensato quindi di adottare vari algoritmi in modo da poter individuare quello che abbia le migliori performance con il tipo di dato trattato e preparato con i passaggi descritti nel capitolo precedente.

Il linguaggio di programmazione scelto per poter sviluppare il modello proposto è Python versione 3, e l’ambiente di svilippo usato è Colab, la piattaforma di sviluppo Python basata su cloud. Tra le tante motivazioni che hanno portato alla scelta di questo linguaggio di programmazione la più importante è stata la possibilità di accedere ad una vasta gamma di librerie per il supporto al calcolo scientifico come Numpy, Scikit-Learn, Pandas, etc..

## Regressione logistica

Il primo modello statistico utilizzato è stato il modello Logit, noto anche come modello logistico o Regressione Logistica. Esso viene spesso utilizzato per la classificazione e gli *analytics* predittivi. Essendo un classificatore probabilistico, perché utilizza un'ipotesi di modellazione specifica per mappare le *features* a una probabilità di appartenenza alla classe, esso dà come risultato una probabilità, di conseguenza la variabile dipendente è delimitata tra 0 e 1 (s5)(b6).

Nel nostro caso la variabile dipendente è di natura dicotomica, cioè ha solo due possibili esiti ( vedi paragrafo 3.5 ), quindi è stata utilizzata la regressione logistica di tipo binaria, in cui la variabile dipendente (la classe da predire) è ricavata dall’insieme {-1, +1} secondo i passaggi matematici che seguono.

Sia il vettore un vettore di d+1 parametri diversi. L’i-esimo parametro è un coefficiente correlato all’i-esima dimensione x dei dati (le 20 features nel nostro caso) e è il parametro di offset. Quindi per un sample , la probabilità che la variabile di classe C assuma i valori +1 oppure -1, viene modellata con l’uso della funzione logistica.

Nella regressione logistica, i parametri possono essere visti come i coefficienti di un separatore dell’iperpiano tra le due classi (Figura 4‑1).

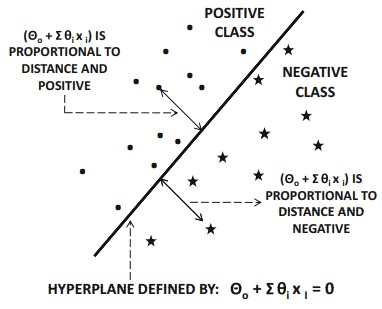


Figura 4‑1 Illustrazione della regressione logistica in termini di separazione lineare (pag 311 di b6)

Il termine , all’interno dell’esponente della funzione logistica, è proporzionale alla distanza tra i punti e il separatore dell’iperpiano.

I valori positivi della distanza assegneranno valori di probabilità maggiori di 0.5 alla classe positiva. I valori negativi della distanza assegneranno valori di probabilità (simmetricamente uguali) maggiori di 0.5 alla classe negativa

Pertanto la funzione logistica converte le distanze in probabilità interpretabili nell’insieme C = (0,1). Infine l’assegnazione alla classe 0 viene fatta per valori di C < 0.5 e alla classe 1 per valori di C > 0.5.

In Python, i modelli di Regressione Logistica utilizzati sono quelli presenti nel framework *scikit-learn*, appartenenti alla libreria *sklearn.linear\_model*, nello specifico le funzioni *LogisticRegression* e *LogisticRegressionCV.*

## Random Forest

Algoritmi usati: pro e contro

* Regressione logistica
* Random forest
* PCA
* XGBoost (?)

Comparazione

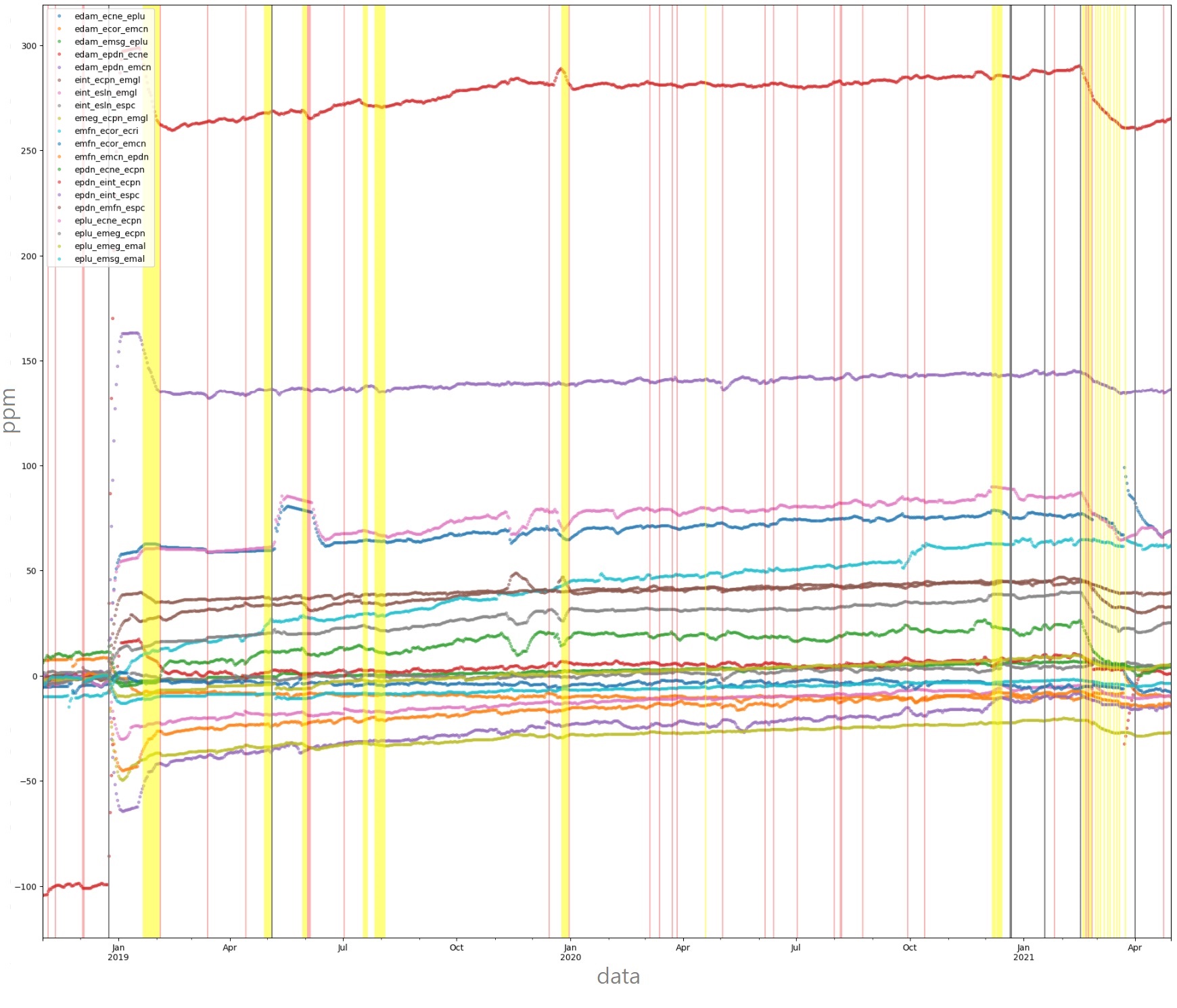
Scelta del modello con migliore accuratezza: grid search

Riscontro delle aree maggiormente correlate col target: spiegazione vulcanologica

# Risultati

## Scelta del modello migliore

## Predizione



# Uso del modello per predizione

Introduzione del pre-evento, 3 classi

# Conclusioni

RINGRAZIAMENTI

INDICE DELLE FIGURE

[Figura 2‑1 Mappa delle stazioni GNSS della rete permanente dell'Etna gestita dall'Osservatorio Etneo 5](#_Toc179540518)

[Figura 2‑2 Stazione GNSS presente sull’Etna costituita da antenna (a sinistra), ricevitore (a destra) e altri apparati. 6](#_Toc179540519)

[Figura 2‑3 Area del triangolo formato dalle stazioni EINT (Intermedia), ESLN (Serra la Nave) e ESPC (Serra Pizzuta Calvarina) 8](#_Toc179540520)

[Figura 3‑1 Variazione areale del triangolo EDAM-EMEG-EINT che mostra le fasi di inflazione e deflazione registrate nel periodo 01.01.2019 - 30.04.2021 [b1] 13](#_Toc179540521)

[Figura 3‑2 Suddiviosne dell'area vulcanica in triangoli adiacenti. 14](#_Toc179540522)

[Figura 3‑3 Sequenza delle 17 fontane manifestate nella fase di deflazione (b1) 25](#_Toc179540523)

[Figura 3‑4 Suddivisione del dataset in training e Test set 26](#_Toc179540524)

INDICE DEI GRAFICI

[Grafico 2‑1 Serie temporale dell’ultimo anno della componente Nord della stazione ECNE (Cratere di Nord Est - Etna) 7](#_Toc179540510)

[Grafico 2‑2 Serie temporale dell'area del triangolo formato dalle stazioni EINT-ESLN-ESPC 10](#_Toc179540511)

[Grafico 2‑3 Serie temporali delle baseline costruite a coppie dei vertici del triangolo ESLN-EINT-ESPC 11](#_Toc179540512)

[Grafico 3‑1 Serie temporali della variazione areale per i 20 triangoli aventi per vertici le stazioni GNSS dell'Etna 15](#_Toc179540513)

[Grafico 3‑2 Serie temporali della variazione areale per i 20 triangoli sovrapposte all'occorrenza degli eventi vulcanici. 18](#_Toc179540514)

[Grafico 3‑3 Serie temporale variazione areale assoluta del triangolo EPLU-ECNE-ECPN 20](#_Toc179540515)

[Grafico 3‑4 Serie temporale variazione areale relativa del triangolo EPLU-ECNE-ECPN 20](#_Toc179540516)

[Grafico 3‑5 Processo di riduzione dei campioni per il bilanciamento delle classi: le barre verticali rappresentano l'occorrenza degli eventi vulcanici, i punti colorati rappresentano i campioni diversificati per features. 23](#_Toc179540517)

INDICE DELLE TABELLE

[Tabella 3‑1 Elenco dei triangoli scelti per la suddivisione dell'area vulcanica analizzata 14](#_Toc179540501)

[Tabella 3‑2 Elenco degli eventi vulcanici usati come target del modello supervisionato. 17](#_Toc179540502)

[Tabella 3‑3 Distribuzione delle classi per il dataset di partenza 22](#_Toc179540503)

[Tabella 3‑4 Distribuzione delle classi per il dataset definitivo 24](#_Toc179540504)

[Tabella 3‑5 Valori di Media () e Deviazione Standard (σ) calcolati sul training set per ciascuna feature. 28](#_Toc179540505)

BIBLIOGRAFIA

[b1] Bruno, V., Aloisi, M., Gambino, S., Mattia, M., Ferlito, C., & Rossi, M. (2022). *The most intense deflation of the last two decades at Mt. Etna: the 2019–2021 evolution of ground deformation and modeled pressure sources.* Geophysical Research Letters, 49(6), e2021GL095195.

[b2] Andronico, D., Cannata, A., Di Grazia, G., & Ferrari, F. (2021). *The 1986–2021 paroxysmal episodes at the summit craters of Mt. Etna: Insights into volcano dynamics and hazard.* Earth-Science Reviews, 220, 103686.

[b3] Calvari, S., & Nunnari, G. (2022). *Etna output rate during the last decade (2011–2022): Insights for hazard assessment.* Remote Sensing, 14(23), 6183.

[b4] Mattia, M., Bruno, V., Montgomery‐Brown, E., Patanè, D., Barberi, G., & Coltelli, M. (2020). *Combined seismic and geodetic analysis before, during, and after the 2018 Mount Etna eruption.* Geochemistry, Geophysics, Geosystems, 21(9), e2020GC009218.

[b5] Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

[b6] Charu Aggarwal (2015). *Data Mining: the Textbook.* Springer

SITOGRAFIA

[s1] wikipedia.org

[s2] ingv.it

[s3] netai.it

[s4] it.wikipedia.org/wiki/Training\_e\_test\_set

[s5] ibm.com/it-it/topics/logistic-regression

1. ELT: Extraction Load and Transform: i dati vengono memorizzati nel loro formato nativo (raw) in repository chiamati Data Lake; la fase di processamento viene eseguita in maniera asincrona ed indipendente. [↑](#footnote-ref-1)
2. Lo standard RINEX permette la memorizzazione e lo scambio dei dati, provenienti da sistemi satellitari di navigazione, indipendentemente dal ricevitore usato per effettuare la misura. [↑](#footnote-ref-2)
3. MIT, Scripps Institution of Oceanography and Harvard University with support from the National Science Foundation [↑](#footnote-ref-3)
4. Baseline: vettore che unisce due punti nello spazio [↑](#footnote-ref-4)
5. Inflazione: rigonfiamento / espansione dell’edificio vulcanico [↑](#footnote-ref-5)
6. Deflazione:sgonfiamento /contrazione dell’edificio vulcanico [↑](#footnote-ref-6)
7. Feature Engineering: l’ingegneria delle funzionalità è una fase di preelaborazione nell'apprendimento automatico supervisionato e nella modellazione statistica che trasforma i dati grezzi in un insieme di input più efficace[s1]. [↑](#footnote-ref-7)