

Università degli Studi di Milano-Bicocca

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di Laurea Magistrale in Data Science

Influenza delle sorgenti locali sul bilancio di polveri minerali in Antartide

Relatore: Prof. Samuel Albani

Tesi di Laurea Magistrale di:

Mario Alessandro Napoli

Matricola 860571

Anno Accademico 2020-2021

Indice

Indice	2
Elenco delle figure	3
1 Introduzione	7
2 Raccolta dati e creazione dei clusters	9
2.1 I Dati	9
2.1.1 Dati stazioni di interesse	9
2.1.2 Dati geografici	10
2.1.3 EPSG	12
2.1.4 Enrichment dati altimetrici	13
2.2 Clustering	16
2.2.1 Introduzione al Clustering	16
2.2.2 Flat Clustering e K-Means	17
2.2.3 Hierarchical Clustering	19
2.2.4 Clustering Evaluation	21
2.2.5 Soluzione Proposta	23
2.3 Analisi dati meteo	30
2.3.1 raccolta dei dati AWS	30
2.3.2 Integrazione e arricchimento	34
3 Generazione delle traiettorie	36
3.1 Introduzioni a HYSPLIT	36
3.1.1 HYSPLIT	36
3.1.2 Air Parcel Trajectory	37

3.1.3	Raccolta dei dati climatici e identificazione dello starting point ideale	39
3.2	Calcolo delle traiettorie	41
3.2.1	Starting points della simulazione e output ottenuto	41
3.2.2	Calcolo dell'influenza dei cluster sulle stazioni	42
3.2.3	Clustering delle traiettorie	48
4	Costruzione di un modello statistico	52
4.0.1	Descrizione dei dati e analisi preliminare	52
4.0.2	Definizione di un modello OLS	56
4.0.3	Risultati modello OLS	62
5	Conclusioni e possibili sviluppi	68
Bibliografia		70
6	Ringraziamenti	73

Elenco delle figure

1.1	(a) Mappa tematica dell'Antartide con le principali stazioni (stelle) e con le aree deglaciate (in marrone). (b) Analisi di campioni in un sito di perforazione con dei nunataks nello sfondo (sommità non coper- te da neve, possibili fonti di polveri). (c) Esempio di dust-particles depositate al suolo, in alto a destra una barra bianca da 5µm per riferimento.	8
2.1	Posizione delle stazioni di carotaggio in Antartide	11
2.2	Copertura dati altimetrici: in grigio scuro l'area non coperta dal modello Antarctic 5-km DEM Model from ERS-1 Altimetry	14

2.3	sinistra: Distribuzione spaziale delle aree deglaciate, il colore identifica la tipologia di materiale di composizione dell'outcrop, se roccioso (Rocks) o no (Unconsolidated). destra: Altitudine dell'outcrop roccioso integrato tramite Modello di Elevazione Digitale REMA	15
2.4	Pseudo-codice algoritmo di Clustering K-Means	18
2.5	Dendogramma di esempio di un Clustering Gerarchico	20
2.6	Clustering output con diverse tipologie di Hierarchical Agglomerative Clustering, Single-Link o Complete-Link	21
2.7	Worst-Case Scenario per K-Means Clustering classico	24
2.8	Coefficiente di Silhouette per clustering con diversi valori di K	26
2.9	Esempio di Clustering k-means++ con valore $k = [20, 25, 30, 35]$	27
2.10	Visualizzazione multi-variata distanza dai cluster vs flusso	29
2.11	Distribuzione delle stazioni AWS in Antartide con relativi enti di appartenenza	31
2.12	Distribuzione delle stazioni AMRC in Antartide di cui si posseggono i dati	33
2.13	Osservazioni meteorologiche disponibili per ogni anno	33
2.14	Esempio visualizzazioni su variabili meteorologiche integrate sui Cluster	35
3.1	Distribuzione della velocità del vento media giornaliera per le stazioni AMRC per ogni mese nell'anno 2011	40
3.2	Visualizzazioni delle retro-traiettorie prodotte per il sito EPICA-DML	43
3.3	Curva logistica scelta per il calcolo dello score di influenza	45
3.4	Influenza dei cluster sulle stazioni	47
3.5	Variazione TSV	50
3.6	Clustering delle traiettorie	51
4.1	Correlazione tra tutte le variabili esplicative disponibili	59
4.2	Scatterplots tra variabili esplicative e variabile target	61
4.3	Possibili trasformazioni della variabile esplicativa	63
4.4	Linea di regressione stimata dal modello OLS per la l'influenza non trasformata, e con il logaritmo naturale	64

4.5	Valore di flusso[5-10 μm] nel modello finale al variare di distanza e influenza dai clusters	65
4.6	Previsione Flusso[5-10 μm]	66

Elenco delle Equazioni

2.1	Distanza Euclidea	17
2.2	Centroide di un Cluster	17
2.3	Misura di similarità per HAC di tipo Single-Link	20
2.4	Misura di similarità per HAC di tipo Complete-Link	20
2.5	Coefficiente di Silhouette	21
2.6	Misura di Purezza per un Cluster	22
2.7	Rand Index	22
2.8	Precision, Recall, F-Beta Measure	23
2.9	Frobenius Norm	25
2.10	K-Means++ Stopping Criterion	25
3.1	Air Parcel first guess position	38
3.2	Air Parcel final position	38
3.3	Maximum transport velocity	39
3.4	Logistic Function	44
3.5	Spatial Variance	48
3.6	Cluster Spatial Variance	49
3.7	Total Spatial Variance	49
4.1	Score influenza totale stazione	54
4.2	Funzione costo del modello OLS	57
4.3	Funzione lineare del modello OLS	57
4.4	Forma vettoriale modello OLS	57
4.5	Relazione matematica tra Flusso[5-10 μm], distanza, e influenza dei clusters.	65

Abstract

Le polveri minerali originate dall'erosione eolica costituiscono il più abbondante aerosol presente in atmosfera, contribuendo ad alterare il clima della Terra. Le carote di ghiaccio antartiche sono uno principali archivi naturali per le ricostruzioni paleoclimatiche, conservando informazioni sulla composizione atmosferica del passato, incluso il contenuto di polveri provenienti dalle zone aride del Sud America e dall'Australia. Il lavoro di tesi utilizzerà sia dataset statici (che descrivono la superficie terrestre) che dinamici (che descrivono l'evoluzione della circolazione atmosferica nel tempo) per costruire degli indicatori che consentano di stimare il possibile contributo di piccole aree deglaciate in Antartide nel budget di polveri misurato dalle principali perforazioni di ghiaccio, confondendo il segnale principale da sorgenti remote.

Introduzione

Il ghiaccio delle calotte polari antartiche rappresenta uno dei più importanti archivi naturali, fonte di conoscenza paleoclimatica; grazie ad analisi chimiche e fisiche del ghiaccio e del materiale particolato e gas intrappolati negli strati via via più profondi, è infatti possibile reperire importanti informazioni anche quantitative sulla composizione dell'atmosfera e su altre variabili di interesse climatico nel passato fino ad un milione di anni fa.

Le polveri minerali provenienti dall'erosione eolica di superfici generalmente aride e sparsamente vegetate raggiungono siti di deposito ad alta elevazione in Antartide dopo aver viaggiato per lunghe distanze, anche oltre i 1000km (Lambert et al., 2008[1], Petit et al., 1999[2]), nella media-alta troposfera. Dopo essersi depositate le polveri sepolte negli anni da neve e molteplici strati di ghiaccio possono essere studiate tramite tecniche quali la stratigrafia, queste analisi possono quindi essere utilizzate per documentare meglio la circolazione delle polveri nell'emisfero sud e la sua variabilità nel tempo. Durante gli ultimi 0.5-1 Milioni di anni si è assistito ad un cambiamento nella concentrazione di polveri nel ghiaccio antartico, la motivazione alla base di quest'evoluzione è da ricercarsi nelle diverse condizioni ambientali che modulano l'intensità di emissione della fonte (Delmonte et al., 2017 [3]), condizioni come il ciclo di vita delle particelle di polvere in atmosfera e il tasso di accumulo delle nevi, così come altri fattori in qualche modo collegati con il ciclo idrologico. Data la grande distanza delle terre di origine, i flussi di polveri depositati in Antartide durante l'Olocene, ossia negli ultimi 11'700 anni, sono tra i più bassi nella terra rispetto agli altri periodi più lunghi e remoti. In diverse zone periferiche del continente, la presenza di aree deglaciate favorisce l'immissione di polveri in atmo-

sfera. Durante l'Olocene gli influssi estremamente bassi di polveri provenienti da aree extra-Antartiche si sono mischiate in proporzioni variabili con le polveri di origine locale; i carotaggi di ghiaccio sono di conseguenza diventati molto sensibili al clima locale e alla circolazione di queste polveri nel corso del tempo, come possibile vedere nei record di ghiaccio prelevati dai siti Talos Dome e Taylor Dome (Albani et al., 2012 [4]; Baccolo et al., 2018 [5]; Delmonte et al., 2010b [6], Aarons et al., 2016 [7]).

Il lavoro di tesi si inserisce in questo contesto per cercare di migliorare le nostre conoscenze sulla circolazione delle polveri e sul budget che le fonti locali rappresentano nel totale dell'aerosol presente in Antartide. Per fare ciò verranno analizzati i dati a nostra disposizione sui siti di raccolta dei record di ghiaccio, così come sulla disposizione ed estensione delle aree deglaciate. Utilizzeremo queste informazioni per costruire un raggruppamento delle aree deglaciate in grado di identificare le potenziali fonti locali di polveri e quindi caratterizzarle tramite l'integrazione di dati climatici. Andremo successivamente a stimare la circolazione atmosferica delle correnti tramite l'utilizzo di un modello di simulazione delle traiettorie di masse d'aria, le informazioni estratte da questo processo verranno utilizzate, infine, per cercare di mettere in relazione l'afflusso di polveri dalle fonte locali fino alle stazioni dove vengono eseguiti i carotaggi.

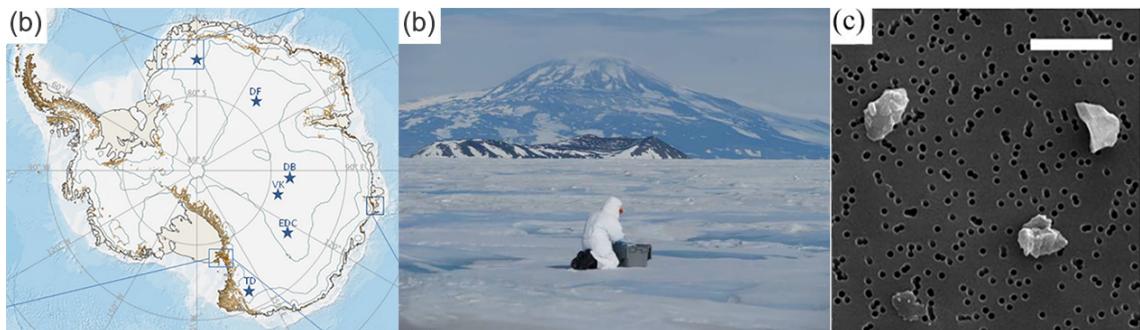


Figura 1.1: (a) Mappa tematica dell'Antartide con le principali stazioni (stelle) e con le aree deglaciate (in marrone). (b) Analisi di campioni in un sito di perforazione con dei nunataks nello sfondo (sommità non coperte da neve, possibili fonti di polveri). (c) Esempio di dust-particles depositate al suolo, in alto a destra una barra bianca da $5\mu\text{m}$ per riferimento.

Raccolta dati e creazione dei clusters

2.1 I Dati

Nella prima fase del lavoro di tesi il focus è quello di raggruppare i dati riguardanti le misure di polveri dalle carote di ghiaccio e integrarli con i dati da carte tematiche digitali che descrivono lo stato degli strati di roccia deglaciate presente in Antartide. Dopo aver raggruppato i dati verrà prodotto un Clustering per le aree deglaciate che andremo ad utilizzare successivamente per integrare i raggruppamenti trovati con le informazioni su intensità e direzione del vento così da trattare le polveri provenienti da questi cluster in base alle variabili climatiche di interesse e estrarre un modello della circolazione di queste polveri, nei successivi step.

2.1.1 Dati stazioni di interesse

Le informazioni necessarie per questa prima fase sono di natura descrittiva, prima di poter iniziare con delle analisi ci servono dei dati che possano descrivere i siti di perforazione. Per ognuno di questi punti ci interessa ottenere le informazioni relative al suo posizionamento (latitudine, longitudine), unitamente alle informazioni di rilevanza sul budget delle polveri; le variabili a disposizione per ogni stazione saranno:

- **Age:** Età delle polveri prese in considerazione.

- **Dust Concentration:** Concentrazione media di polvere per il periodo di tempo indicato (parts per bilion), corredato da deviazione standard.
- **Accumulation rate:** Media velocità di accumulazione espressa in centimetri di acqua equivalente per anno dedotto dalla scala AICC2012 (Antarctic Ice Core Chronology).
- **Dust Flux:** Media flusso di polveri calcolato a partire dalla Dust Concentration e dall'Accumulation Rate in ogni sito, espresso in milligrammi di polvere per m^2 per anno.
- **Altitude:** Altitudine dei diversi siti di raccolta dei dati.

Questi dati sono disponibili per un totale di 12 stazioni, le informazioni riportate di seguito provengono da [8] [9].

ice core	latitude	longitude	age	dust.conc_(ppb)	dust.conc_(ppb).sd	acc.rate	dust.flux	altitude
Dome B	-77.05	94.55	2.0-11.7 kyr BP	14.2	8	2	0.28	3650
VOSTOK-BH7	-78.28	106.48	3.6-9.8 kyr BP	17.8	8	2	0.36	3480
EPICA-DC	-75.06	123.21	2.0-11.7 kyr BP	14.3	8	2.7	0.39	3233
SOLARICE-DC	-75.0631	123.248	3.3-4.5 kyr BP	11.7	5	2.6	0.31	3233
DOME FUJI	-77.1901	39.4212	3.2-11.7 kyr BP	20	11	2.7	0.54	3810
EPICA-DML	-75.00	00.04	2.0-11.7 kyr BP	25	14	6	1.5	2882
TALDICE	-72.49	159.11	2.0-11.7 kyr BP	25.1	10	7	1.76	2315
TAYLOR DOME	-77.4747	158.4326	2.0-11.7 kyr BP	21	14	6	1.18	2365
DC ITASE	-75.06	123.21	AD 1570-1800	8	4	2.5	0.2	3230
D4 ITASE	-75.35	135.49	AD 1420-1700	9	4	2.0	0.19	2795
MDP-A ITASE	-75.32	145.51	AD 1620-1800	14	3	3.6	0.5	2455
TALDICE	-72.49	159.11	AD 1420-1810	8	4	8.7	0.75	2315

Tabella 2.1: Tabella informazioni statistiche su stazioni di carotaggio

2.1.2 Dati geografici

Ottenuti dei dati descrittivi delle stazioni avremo bisogno di recuperare altri dati di tipo spaziale contenenti vari livelli informativi sulle aree deglaciate e sulle loro proprietà fisiche e geografiche, queste informazioni saranno la base per creare un clustering delle aree deglaciate del continente. Nello specifico i Dataset che utilizzeremo sono:

- Shapefile dell'Antartide

- Dati sulle Geological Units
- Dati altimetrici
- Dati sull'Outcrop roccioso delle aree deglaciate

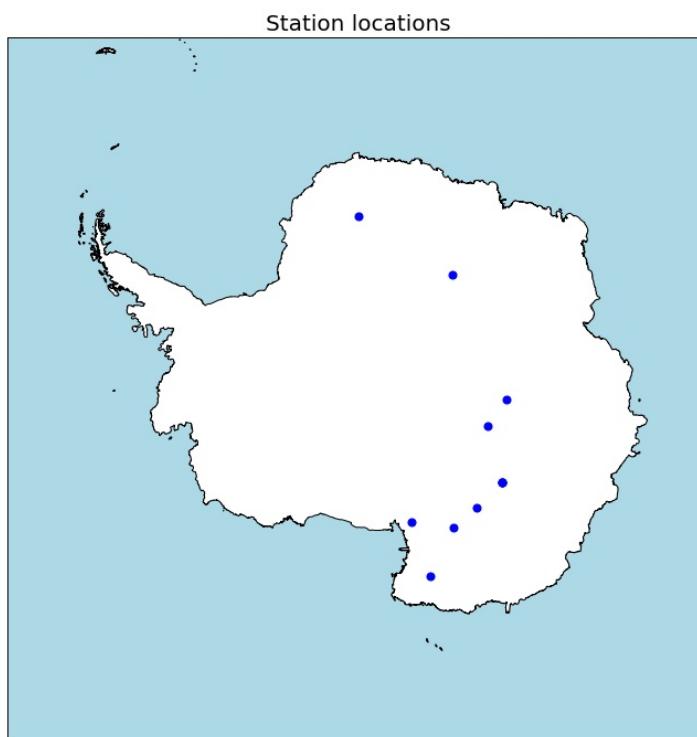


Figura 2.1: Posizione delle stazioni di carotaggio in Antartide

Shapefile: Descrivono la superficie e l'estensione della costa, questi dati saranno necessari per calcolare informazioni come le distanze tra i vari punti di interesse e più in generale a supporto per le molteplici visualizzazioni dei dati e del clustering che ci accompagneranno nelle processi di analisi.

Nello specifico sul sito della **British Antarctic Survey** sono resi disponibili diverse versioni degli Shapefile della costa Antartica a diverse risoluzioni spaziali, questi dati sono ottenuti dal mix di molteplici mapping del territorio, ciò viene fatto al fine di aumentarne la precisione spaziale. La versione utilizzata per questo progetto è la 7.2 rilasciata nel 2020 utilizzata nella sua risoluzione massima, così da non introdurre errori nelle visualizzazioni delle diverse grandezze e nel calcolo delle distanze dalle

coste.

Geological Units: Dati relativi alla datazione degli strati rocciosi distribuiti per ere geologiche, la fonte dei dati: (**SCAR GeoMAP**) contiene informazioni anche sui depositi superficiali, queste informazioni nello specifico verranno utilizzati per integrare le nostre conoscenze sulle aree che presentano Outcrop (affioramento roccioso in superficie) al fine di comprendere se le suddette aree siano formate da materiale roccioso o meno.

Dati Altimetrici: Per meglio descrivere gli elementi in gioco sarà necessario dotarsi di un buon modello altimetrico dell'Antartide, queste informazioni saranno utili per integrare i dati sull'Outcrop roccioso e avere un'ulteriore dimensione di analisi delle zone di interesse.

Rock Outcrop: Dati costituiti dalle informazioni delle aree deglaciate, da questi dati sarà possibile risalire al posizionamento e all'estensione delle aree deglaciate e su queste informazioni sarà eseguito un primo clustering spaziale. Il dataset è stato ottenuto automaticamente tramite elaborazione delle immagini provenienti dal satellite **Landsat 8** [10], il dataset originale è stato successivamente aggiornato per adattarsi all'ultima versione, la 7.2 della coastline Antartica, versione che stiamo utilizzando in questo progetto.

Il Dataset è composto da **366106** data point, ogni rilevazione identifica una precisa area deglacata rappresentata dalle coordinate di un poligono nel sistema di riferimento prescelto dall'autore del dataset. In Figura 2.3 le aree deglaciate integrate con i dati delle Geological Units.

2.1.3 EPSG

Per uniformare le varie fonti dato presenti in questo lavoro si rende necessario avere un unico sistema di riferimento così da poter utilizzare i vari layer informativi, provenienti dalle diverse fonti, nello stesso spazio di coordinate. I sistemi di riferimento più comuni sono quelli definiti dagli standard dell'**European Petroleum Survey**

Group, in breve **EPSG**, un comitato scientifico dedito allo studio della cartografia per l'industria petrolifera che nel 1993 ha reso pubblico un registro di riferimenti standard di ellisoidi e datum geodetici. L'**EPSG** più utilizzato è il sistema **4326** che definisce latitudine e longitudine basandosi sul centro di massa della terra, questo **EPSG** è ottimo per punti geografici che si trovano più vicini all'equatore, ma nel caso dei dati in nostro possesso non sono ottimali. L'EPSG:4326 utilizzato in contesti dove si opera in prossimità dei poli restituisce delle coordinate spesso distanti dal punto di vista numerico, pur considerando due punti vicini geograficamente. In fase di Clustering degli outcrop una rappresentazione delle coordinate spaziali di questo tipo potrebbe condizionare pesantemente la qualità del modello prodotto, nel processo di clustering infatti (che meglio descriveremo nella prossima sezione) vengono spesso utilizzate distanze euclidee per calcolare i raggruppamenti da estrarre. Per questo motivo, prima di utilizzare qualsiasi informazione geografica ci assicureremo che esse siano convertite nello standard **EPSG:3031**, che si basa sullo stesso Ellissoide di riferimento **WGS:84** utilizzato dallo standard **4326**, ma ha le proprie coordinate centrate al polo sud, e definisce ogni punto come distanza in metri dal proprio centro. In questo sistema di riferimento i punti geograficamente prossimi delle aree deglaciate avranno una rappresentazione numerica più rappresentativa della reale distanza, evitando così di distorcere le informazioni utilizzate per le varie analisi che ci attendono.

2.1.4 Enrichment dati altimetrici

Prima di procedere con la fase di Clustering i dati in nostro possesso sugli outcrop rocciosi vengono integrati con i dati altimetrici disponibili per l'Antartide. In questa fase vengono considerati due dataset pubblici come potenziali fonti dato per il nostro progetto. Il primo è Antarctic 5-km Digital Elevation Model from ERS-1 Altimetry, questo **DEM**(Digital Elevation Model) fornisce dati altimetrici ad una scala di 5km, i dati estratti provengono dall'altimetria radar prodotta dal satellite **ERS-1** durante la fase geodetica dal Marzo 1994 a Maggio 1995. I dati forniti da questo modello sono in forma di proiezione stereografica con origine al Polo Sud, referenziati al geoide **OSU91A**, il totale di data point prodotti è di circa 20,000,000. Seppur

la risoluzione dei dati sia accettabile il modello altimetrico in questione copre le aree fino ad una latitudine di 81.5°S, lasciando quindi una grande porzione di dati mancanti per tutti quei punti che si trovano tra latitudine 81.5°S e 90°S, ossia fino al Polo. La suddetta fascia di dati mancanti è evidenziata in Figura 2.2.

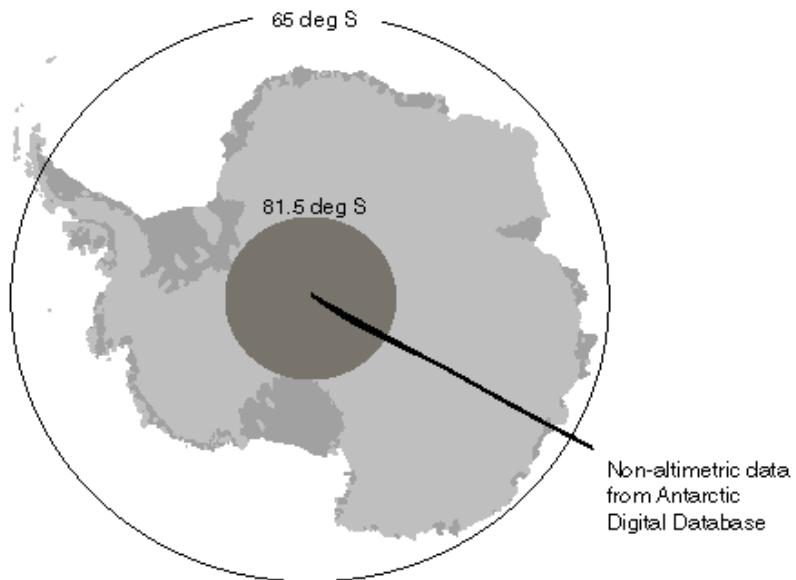


Figura 2.2: Copertura dati altimetrici: in grigio scuro l'area non coperta dal modello Antarctic 5-km DEM Model from ERS-1 Altimetry

Il secondo dataset analizzato è quello del **REMA**(Reference Elevation Model of Antarctica), per la creazione di questi dati vengono presi in considerazione diversi modelli di elevazione digitale. In particolare per costruire il modello REMA è stata implementata una nuova tecnica chiamata **SETSM**(Surface Extraction by TIN-based Search space Minimization)[11]. In questo algoritmo viene utilizzata la stereofotogrammetria automatizzata per la ricostruzione del territorio in ambienti che risultano di difficile interpretazione a causa di fenomeni caratteristici dell'Antartide, come ad esempio presenza di neve e ghiaccio, che creano vaste aree a basso contrasto, oppure montagne e pendii molto ripidi che al contrario creano delle marcate zone d'ombra. I dati prodotti SETSM sono stati successivamente incrociati con quelli provenienti da altri **DEM** preesistenti: WorldView-1, WorldView-2, WorldView-3, GeoEye-1, Cryosat-2, ICESat. Una volta integrati i dati originali con

i diversi DEM si ottiene un unico database con risoluzione fino a 2 metri per le aree di interesse (ad esempio quelle che presentano dell'outcrop), e fino a 8m per il resto del continente. Data la maggior risoluzione e le tecnologie più recenti utilizzate il dataset di riferimento prescelto per il lavoro di tesi è quello del **REMA**, tuttavia utilizzare il dataset alla sua massima risoluzione creerebbe una grandissima quantità di dati non necessari. Per riportare i dati ad una scala consona ai nostri scopi è stata fatta la decisione di utilizzare una versione aggregata del dataset, rappresentata da un'immagine raster dell'Antartide dove ogni pixel rappresenta un area di 1km^2 , i valori dei pixel sono stati ulteriormente aggregati facendo scorrere una griglia 3×3 per creare quindi un dataset con risoluzione finale di 3km , dove ogni pixel rappresenta quindi un'area di 9km^2 . Il dataset così composto presenta informazioni di latitudine e longitudine riportate al sistema di riferimento prescelto **EPSG:3031**, per un totale di circa **900'000** data-point utili ad arricchire le informazioni sull'outcrop roccioso con il dato di altitudine più vicino possibile nella griglia prodotta, in Figura 2.3 una visualizzazione dei dati altimetrici estratti.

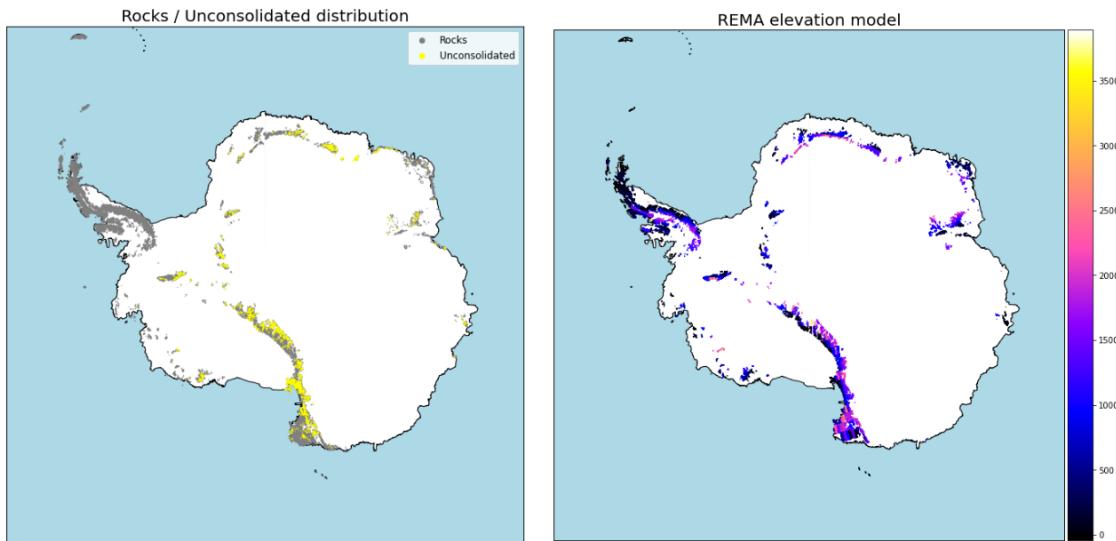


Figura 2.3: sinistra: Distribuzione spaziale delle aree deglacierte, il colore identifica la tipologia di materiale di composizione dell'outcrop, se roccioso (**Rocks**) o no (**Unconsolidated**). destra: Altitudine dell'outcrop roccioso integrato tramite Modello di Elevazione Digitale REMA

2.2 Clustering

Il Clustering è il processo di raggruppare un set di oggetti, in classi di oggetti simili. In genere le entità all'interno di un cluster dovrebbero essere simili, mentre quelle appartenenti a cluster differenti dovrebbero essere dissimili. A differenza della classificazione il task del Clustering di solito non è supervisionato, la composizione dei cluster è quindi inferita direttamente dai dati e dall'algoritmo senza intervento dell'umano. In questa sezione andremo a descrivere le principali metodologie di clustering per poi illustrare la soluzione proposta nel nostro specifico contesto.

2.2.1 Introduzione al Clustering

Il problema principale del Clustering è quello di scegliere come **rappresentare** le entità in uno spazio comune e come definire una nozione di **similarità/distanza** tra gli oggetti, così come la scelta del numero ottimale di clusters. Bisognerebbe fissare un numero di cluster a priori o lasciamo che la decisione su quanti cluster utilizzare sia completamente data-driven? Queste e molte altre domande rendono il clustering una sfida particolare che spesso necessita di soluzioni ad-hoc per ottenere dei buoni risultati.

In generale possiamo avere due tipologie di Clustering:

- **Flat:** crea un set di cluster senza una specifica struttura, quindi di solito il punto di partenza è casuale, e solo dopo si occupa di rifinire i clusters iterativamente, ad esempio con tecniche di **k-means** o utilizzando altri modelli.
- **Hierarchical:** In questo tipo di clustering si crea una gerarchia di cluster, possiamo approcciarlo bottom-up, affrontando un problema **agglomerative**, oppure top-down, quindi operando in modo **divisive**, una volta ottenuta una gerarchia completa dei nostri cluster si decide a che altezza tagliare la struttura per ottenere un numero preciso di cluster.

Bisogna poi specificare anche la differenza tra:

- **Hard Clustering:** dove ogni oggetto appartiene esattamente ad un cluster.

- **Soft Clustering:** dove un'entità può appartenere a più cluster dove l'appartenenza ad ogni cluster viene specificato con un peso, questa metodologia è quella che di solito ha più senso e viene utilizzata in molte applicazioni pratiche.

2.2.2 Flat Clustering e K-Means

Possiamo definire rigorosamente lo scopo dell'**Hard Flat Clustering** come segue: Dato un set di oggetti $O = (o_1, o_2, \dots, o_N)$, un numero k di cluster desiderati e una funzione obiettivo che valuta la qualità di un cluster, vogliamo calcolare un assegnamento: $\gamma : O \rightarrow (\omega_1, \omega_2, \dots, \omega_k)$ che minimizza, o in alcuni casi massimizza, l'objective function. In molti dei casi di Clustering si richiede che γ sia suriettiva, ossia che nessuno dei k cluster sia vuoto, la funzione obiettivo viene definita in termine della distanza tra gli oggetti.

K-Means:

Il K-means è l'algoritmo di Clustering più importante, si assume che gli oggetti siano rappresentati come vettori normalizzati in uno spazio di valori reali, l'obiettivo del k-means è quello di minimizzare la distanza euclidea media degli oggetti dal centro del rispettivo Cluster. La distanza Euclidea tra due vettori \vec{u}, \vec{v} è definita come:

$$d(\vec{u}, \vec{v}) = \|\vec{u} - \vec{v}\| = \sqrt{(u_1 - v_1)^2 + \dots + (u_n - v_n)^2} = \sqrt{\sum_{i=1}^n (u_i - v_i)^2} \quad (2.1)$$

Il centro di un cluster viene definito come **Centroide**, il centroide $\vec{\mu}$ degli oggetti in un cluster w è calcolato come:

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x} \quad (2.2)$$

La riassegnazione delle istanze, ossia degli oggetti, ai cluster nel processo iterativo è basato sulla distanza rispetto ai centroidi. Il primo step del K-Means è quello

di selezionare dei centroidi, di solito vengono scelti k oggetti in maniera casuale chiamati **seeds**: (s_1, s_2, \dots, s_k) , successivamente per ogni cluster ω_j , $s_j = \vec{\mu}(\omega_j)$ e per ogni oggetto o_j viene assegnato o_j al cluster w_j in modo tale che la distanza (o_j, s_j) sia minimizzata. L'algoritmo quindi, dopo aver assegnato tutti gli oggetti in base alle loro distanze, muove i centroidi per minimizzare la distanza rispetto a tutti gli oggetti che sono stati assegnati a quel cluster. Questo procedimento iterativo viene ripetuto finché non viene soddisfatto uno **stopping criterion**, ricapitolando quindi:

1. Riassegno ogni oggetto al Cluster con il Centroide più vicino
2. Ricalcolo ogni Centroide basandomi sugli attuali membri del Cluster

In figura lo pseudo-codice dell'algoritmo:

```

K-MEANS( $\{\vec{x}_1, \dots, \vec{x}_N\}, K$ )
1  $(\vec{s}_1, \vec{s}_2, \dots, \vec{s}_K) \leftarrow \text{SELECTRANDOMSEEDS}(\{\vec{x}_1, \dots, \vec{x}_N\}, K)$ 
2 for  $k \leftarrow 1$  to  $K$ 
3 do  $\vec{\mu}_k \leftarrow \vec{s}_k$ 
4 while stopping criterion has not been met
5 do for  $k \leftarrow 1$  to  $K$ 
6   do  $\omega_k \leftarrow \{\}$ 
7   for  $n \leftarrow 1$  to  $N$ 
8     do  $j \leftarrow \arg \min_{j'} \|\vec{\mu}_{j'} - \vec{x}_n\|$ 
9        $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$  (reassignment of vectors)
10    for  $k \leftarrow 1$  to  $K$ 
11      do  $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$  (recomputation of centroids)
12 return  $\{\vec{\mu}_1, \dots, \vec{\mu}_K\}$ 

```

Figura 2.4: Pseudo-codice algoritmo di Clustering K-Means

Lo **Stopping Criterion** può essere scelto tra:

- Viene raggiunto un certo numero di iterazioni.
- La partizione dei nostri oggetti non è cambiata rispetto al passato.
- La posizione dei Centroidi non è cambiata.
- La distanza tra gli oggetti e i Centroidi ricade all'interno di una certa soglia.

La scelta dei **seeds** può variare anche di molto il risultato del Clustering, esistono quindi delle euristiche per scegliere dei buoni elementi di partenza da utilizzare. Di

solito per inizializzare i seeds di un clustering si usa un set di oggetti con la minor similarità possibile, oppure si utilizzano diversi punti di partenza, si può anche inizializzare i seeds utilizzando i Centroidi di un risultato precedente. Discorso simile può essere fatto per la decisione del numero di Cluster, il parametro k non è conosciuto a priori, possiamo settare un optimization problem che penalizza l'avere un numero di cluster molto alti. Il numero ideale dipende comunque dall'applicazione ed esiste un **Trade-Off** tra l'avere molti cluster con un focus migliore all'interno dei singoli cluster, e l'averne troppi.

2.2.3 Hierarchical Clustering

Il Clustering gerarchico procede secondo una diversa filosofia, in questa tipologia di clustering si crea una gerarchia tra gli elementi. Questa struttura risulta quindi più informativa rispetto al set di clustering non strutturati che abbiamo visto nel caso Flat, in aggiunta a ciò il metodo non richiede che venga specificato un parametro k di cluster che vogliamo ottenere, i vantaggi del Clustering gerarchico vengono al costo di una bassa efficienza computazionale, la complessità dell'algoritmo è almeno quadratica al numero di oggetti presenti nel dataset, molto peggiore quindi rispetto alla complessità lineare del k-means. Gli algoritmi gerarchici possono procedere in un modo **Bottom-Up** o **Top-Down**.

- **Bottom-Up:** All'inizio ogni oggetto è considerato un cluster, successivamente gli n cluster vengono raggruppati a coppie finché non si raggiunge un unico cluster che contiene tutti gli elementi, si parla di Hierarchical Agglomerative Clustering(HAC).
- **Top-Down:** Si inizia con un cluster generico che contiene tutti gli oggetti e si divide di volta in volta fino ad ottenere un cluster per ogni singolo oggetto, si parla di Hierarchical Divisive Clustering(HDC)

Prendiamo ad esempio l'HAC e analizziamolo più nel dettaglio, la storia dei vari merge forma un albero binario chiamato Dendogramma, come in figura:

Partendo dal livello più basso possiamo unire i clusters in coppia, una volta ottenuto il Dendogramma finale, tagliandolo ad una certa altezza possiamo ottenere

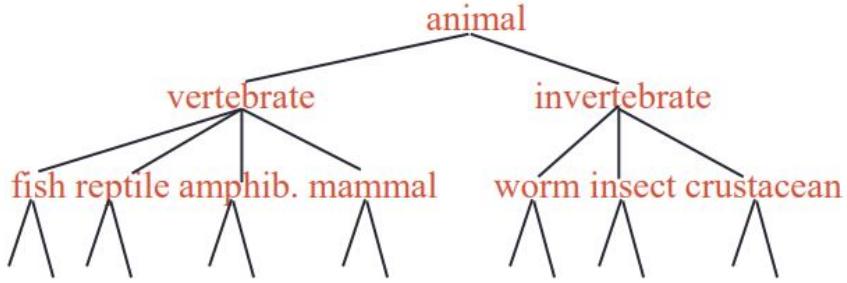


Figura 2.5: Dendogramma di esempio di un Clustering Gerarchico

un clustering con diversi numeri di clusters. Abbiamo a disposizione molti metodi per unire le coppie più simili in base a come definiamo la similarità tra due clusters, questi metodi vengono chiamati **merging criteria**. Vediamoli tutti: Nel **Single Link** la similarità tra due cluster corrisponde alla similarità tra la coppia di oggetti più vicini, quindi:

$$sim(\omega_i, \omega_j) = \max_{x \in \omega_i, y \in \omega_j} sim(x, y) \quad (2.3)$$

Nel **Complete-Link clustering** la similarità tra due cluster è definita come la similarity tra i due elementi più dissimili, quindi al contrario del Single-Link avremo che:

$$sim(\omega_i, \omega_j) = \min_{x \in \omega_i, y \in \omega_j} sim(x, y) \quad (2.4)$$

In figura vediamo le diverse gerarchie create dai due metodi di linkage tra clusters:

Diversamente nell’Hierarchical Divisive Clustering iniziamo con un cluster unico, successivamente dividiamo il cluster utilizzando un metodo di Flat clustering come il K-means, la procedura viene applicata ricorsivamente per ogni cluster finché non si ottengono solo clusters con un solo documento al loro interno, HDC ha il vantaggio di essere più efficiente se non ci serve generare tutta la gerarchia fino ai documenti singoli, possiamo fermarci prima.

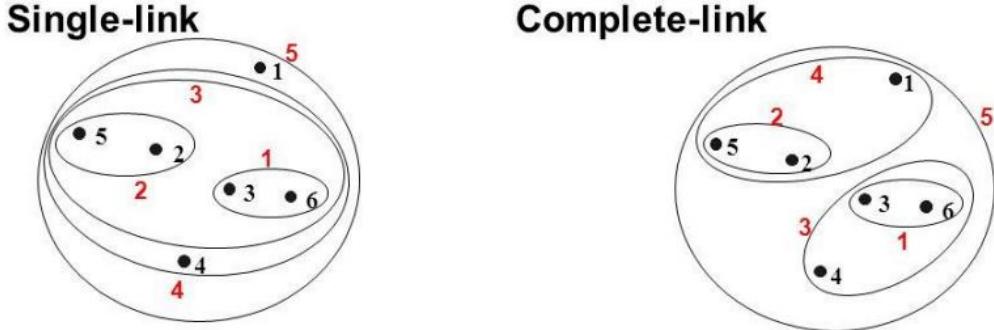


Figura 2.6: Clustering output con diverse tipologie di Hierarchical Agglomerative Clustering, Single-Link o Complete-Link

2.2.4 Clustering Evaluation

In generale un buon clustering è ottenuto quando, considerando un **Internal Criterion** abbiamo una similarità **intra-classe** alta, e similarità **inter-classe** bassa. La misura di qualità di un clustering dipende sia dalla rappresentazione che viene utilizzata per i data-points, sia dalla similarity measure che abbiamo scelto per eseguire il clustering. Uno dei metodi più popolari per le metriche interne, ossia quelle che operano con le informazioni già in nostro possesso, è quella della **Silhouette Analysis**, questo metodo misura la bontà del clustering stimando la distanza media tra i cluster, il Silhouette Coefficient è calcolato utilizzando la distanza media intra-classe $a(i)$ e la distanza media dal cluster più vicino $b(i)$ per ogni sample i

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2.5)$$

Il coefficiente S risiede nel range $[-1, 1]$, per valori vicini a 1 l'oggetto i è ben clusterizzato all'interno di un cluster ben definito, per valori vicini allo 0 il sample potrebbe essere assegnato al cluster più vicino mantenendo la distanza tra i due cluster molto simile, ciò indica regioni di accavallamento tra clusters dove non è chiaro a quale gruppo assegnare un particolare documento, infine nel caso di valori

vicini al -1 il sample non è classificato correttamente e risiede in un punto non definito tra i clusters.

Per quanto riguarda invece le **External criteria** la qualità viene misurata dall'abilità del clustering nello scoprire pattern nascosti o classi latenti conosciute a priori nel dataset. Avendo informazioni su quanti classi C esistono veramente nei nostri dati e sapendo quali elementi appartengono a quali classi posso calcolare delle misure, una di questa è la **purity**, ossia il ratio tra la classe dominante in un cluster ω_i e la dimensione del cluster.

$$\boxed{Purity(\omega_i) = \frac{1}{n_i} \max_j(n_{ij}) \quad j \in C} \quad (2.6)$$

Clustering di bassa qualità avranno purity vicina a 0, ciò indica che nei cluster da noi evidenziati in realtà risiedono oggetti che appartengono a molte classi reali, misure di purity vicine a 1 ci indicano al contrario un clustering migliore dove il numero di oggetti che non appartengono realmente al clustering è basso, questa misura è assimilabile all'accuracy nel task di Classification.

Altra misura esterna è quella del **Rand Index**, questo indice misura la percentuale di decisioni corrette prese dall'algoritmo di clustering. definendo come:

- **TP** numero di decisioni che assegnano due oggetti simili allo stesso cluster.
- **TN** numero di decisioni che assegnano due oggetti dissimili a cluster differenti.
- **FP** numero di decisioni che assegnano due oggetti dissimili allo stesso cluster.
- **FN** numero di decisioni che assegnano due oggetti simili a cluster differenti.

definiamo il Rand Index come:

$$\boxed{RI = \frac{TP + TN}{TP + FP + FN + TN}} \quad (2.7)$$

Il Rand Index assegna pesi equivalenti ai falsi positivi e ai falsi negativi, in molte applicazioni vale il fatto che separare oggetti simili produce effetti peggiori che mettere coppie dissimili nello stesso cluster. Per ovviare a questa situazione possiamo utilizzare la **F-measure** per penalizzare i falsi negativi rispetto ai falsi positivi selezionando un valore di $\beta > 1$, quindi dando più peso alla Recall:

$$P = \frac{TP}{TP + FP} ; R = \frac{TP}{TP + FN} ; F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad (2.8)$$

2.2.5 Soluzione Proposta

Nel nostro caso i primi due problemi relativi a come rappresentare i data-point e come definire una distanza tra di essi sono di facile soluzione. Le aree deglaciate sono rappresentate, nei nostri dati, tramite dei poligoni che ne descrivono la forma, per estrarre quindi un unico **data-point** per ogni poligono possiamo pensare di calcolare il centroide come punto rappresentativo dell'intera posizione spaziale di ogni outcrop. Per la misura di distanza tra gli outcrop invece, trattandosi di punti sullo spazio, utilizzeremo semplicemente la distanza euclidea calcolata sull'EPSG di riferimento, ossia il 3031. L'algoritmo scelto per il lavoro di tesi è **k-means++**, come già visto, nel classico algoritmo K-means esiste il problema di trovare i centroidi che minimizzano la varianza intra-classe ossia la somma delle distanze al quadrato per ogni data point al relativo centroide. Sebbene trovare una soluzione esatta per l'algoritmo è un problema **NP-hard** l'approccio standard per trovare una soluzione approssimativa produce soluzioni spesso soddisfacenti. Tuttavia esistono almeno due problemi a livello teorico:

1. Nel worst case scenario il tempo d'esecuzione dell'algoritmo è più che polinomiale rispetto alla dimensione dell'input.
2. La soluzione trovata può essere arbitrariamente cattiva rispetto al clustering ottimale.

L'algoritmo di K-means++ affronta la seconda problematica specificando un procedimento per inizializzare i centroidi in maniera ottimale prima di procedere con

il classico algoritmo di K-means. Per illustrare quanto il clustering tramite classico k-means può facilmente degradare in performance, basta fare l'esempio di quattro punti disposti a rettangolo in uno spazio R^2 dove la larghezza è molto maggiore rispetto all'altezza. Una volta definito questo rettangolo di punti, se cerchiamo di costruire due cluster inizializzando i centroidi nei punti mediani rispetto ai due vertici che si trovano nella parte superiore e ai due vertici che si trovano nella parte inferiore, l'algoritmo standard di k-means convergerà istantaneamente raggruppando i due vertici superiori nel primo cluster e i due vertici inferiori nel secondo cluster, producendo così una soluzione non ottimale. Il raggruppamento ottimale sarebbe infatti dividere i due punti sulla sinistra con i due punti sulla destra in quanto l'altezza del rettangolo è molto minore rispetto alla larghezza, in figura 2.7 l'esempio appena discusso riportato graficamente.

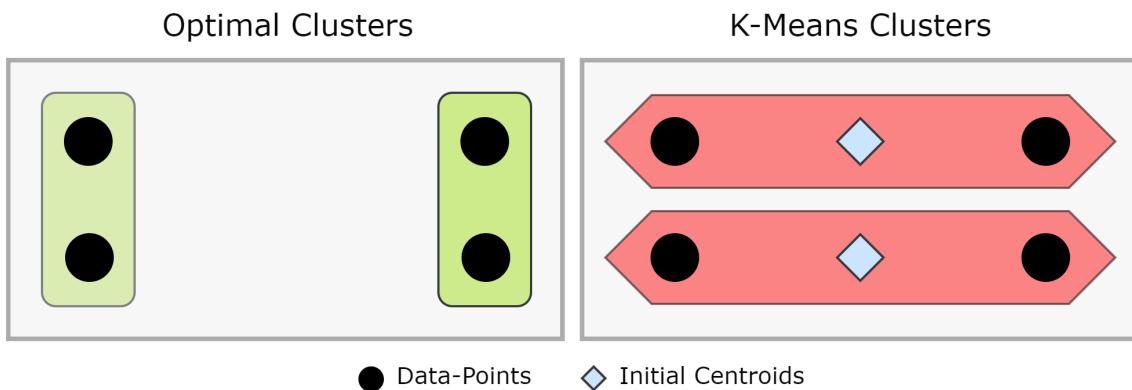


Figura 2.7: Worst-Case Scenario per K-Means Clustering classico

Aumentando la larghezza del rettangolo possiamo far peggiorare di una quantità arbitraria l'output dell'algoritmo k-means rispetto alla soluzione di clustering ottimale. Nella versione migliorata **k-means++** si cerca invece di spargere i k cluster iniziali. Il primo centroide è scelto randomicamente tra tutti i data-points che formano il dataset di partenza. Successivamente ogni altro centroide viene scelto dalle rimanenti osservazioni con una probabilità proporzionale al quadrato della distanza rispetto al centroide esistente più vicino. Con questa metodologia di selezione dei centroidi il miglioramento è considerevole, sebbene scegliere i centroidi porta un tempo di computazione aggiuntivo l'algoritmo converge poi più velocemente e quindi il tempo di computazione totale ne risulta diminuito. La soluzione generata

da k-means++ ha inoltre un rapporto di approssimazione rispetto alla soluzione ottimale direttamente proporzionale al logaritmo di k dove \mathbf{k} è il numero di clusters utilizzati, al contrario dell'algoritmo di k-means classico in cui la soluzione può essere arbitrariamente peggiore rispetto a quella ottimale.

Per garantire un ulteriore convergenza l'algoritmo **k-means++** viene eseguito per **10** volte con differenti centroidi, ogni esecuzione dell'algoritmo gode di un massimo di **300** possibili iterazioni prima di raggiungere una convergenza. Lo stopping criterion viene deciso in base a una threshold di tolleranza rispetto alla **Frobenius Norm** della differenza dei centri in due iterazioni consecutive. La **Frobenius Norm** è definita come la radice quadrata della somma dei quadrati assoluti per ogni elemento della matrice di distanze dei centri. Ad esempio con un numero k di centri, avremo la matrice $k * k$ di distanze per i centri di due iterazioni consecutive, per questa matrice estraiamo il valore:

$$||A||_F \equiv \sqrt{\sum_{i=1}^m \sum_{j=1}^m |a_{ij}|^2} \quad (2.9)$$

Il valore di tolleranza scelto è di $1 * 10^{-4}$, ciò vuol dire l'algoritmo viene stoppato ed il suo stato considerato come soluzione valida se e solo se si raggiunge il numero massimo di iterazioni possibili, o se:

$$(||A||_F)_{[t,t-1]} < 1 * 10^{-4} \quad (2.10)$$

Una volta definito l'algoritmo di clustering ottimale resta da decidere il numero di clusters in cui dividerei nostri punti, ossia gli outcrop rocciosi, la scelta del numero di cluster adatto non è banale. Se avessimo un numero insufficiente di gruppi rischieremmo di accoppare tra loro outcrop rocciosi troppo distanti che condividono poche variabili dal punto di vista geografico. Non dobbiamo infatti dimenticare che il raggruppamento prodotto in questa fase serve per associare tra loro gli outcrop rocciosi che condividono delle precise caratteristiche dal punto di vista climatico,

in particolare ci interessa identificare le proprietà locali di circolazione del vento, la sua direzione e intensità, nonché variabili come la distanza dai punti di perforazione di nostra conoscenza e la distanza dalle stazioni meteo che andremo ad utilizzare successivamente, per finire con la distanza dalla costa. Per tutte queste motivazioni dobbiamo preferire avere molti cluster più piccoli piuttosto che pochi cluster di grandi estensione, senza comunque esagerare. Avere troppi cluster porterebbe alle problematiche opposte, suddividendo troppo il nostro spazio rischieremmo di perdere il quadro generale e i pattern più grandi, l'implementazione soffrirebbe di una complessità maggiore e i risultati rischierebbero di essere poco chiari, è quindi cruciale cercare di trovare il miglior trade-off tra queste due parti. Per tutte le motivazioni sopra citate non possiamo utilizzare le classiche metodologie di **Clustering Evaluation** per cercare di quantificare la bontà dei raggruppamenti, non abbiamo riferimenti esterni che ci possano dire a quali cluster dover assegnare gli outcrop, quindi non è possibile utilizzare le metriche di validazione con il Rand Index o l'F1-score. Basandoci sulle metriche interne che ci indicano la bontà del clustering rischiamo di prendere decisioni poco sagge, l'indice di Silhouette infatti scoraggia fortemente dei clustering con un numero di classi molto ampio.

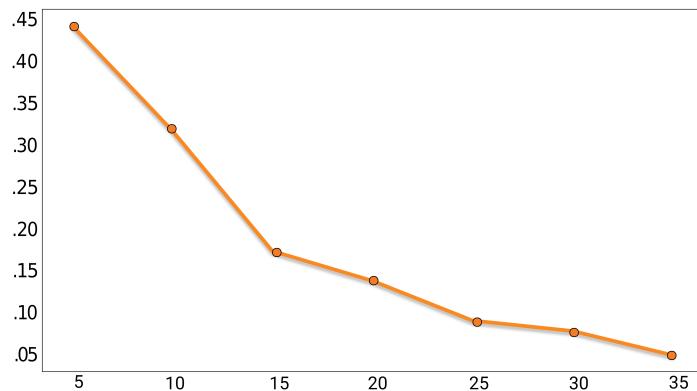


Figura 2.8: Coefficiente di Silhouette per clustering con diversi valori di K

Analizzando il valore della metrica di Silhouette in figura 2.8 per un numero di cluster variabile è infatti possibile notare come il valore di questo indice decresca al crescere del numero di cluster. Come sappiamo utilizzare solo pochi cluster non è la scelta che vogliamo fare per altri motivi specifici al problema che un analisi super-

ficiale come quello dell'indice di Silhouette non può carpire. Vista l'inadeguatezza dell'indice di Silhouette e la mancanza di altre misure di bontà esterne la scelta del numero ideale di clusters verrà fatta esclusivamente su base visiva, cercando un buon compromesso tra il mantenimento di features locali del cluster e una buona suddivisione delle varie aree di nostro interesse.

Dall'analisi visiva dei clustering prodotti scegliendo un parametro $k = 20, 25, 30, 35$ la soluzione con $k = 30$ sembra quella in cui l'estensione media dei cluster rimane più o meno costante senza andare a creare dei cluster troppo grandi che rischierebbero di far perdere contenuto informativo al cluster stesso.

In Figura 2.9 possiamo vedere degli esempi variando il valore di k , i cluster vengono mostrati tramite il loro **Convex-Hull**, ossia il minimo poligono convesso che contiene tutti i data-point del cluster che vogliamo descrivere.

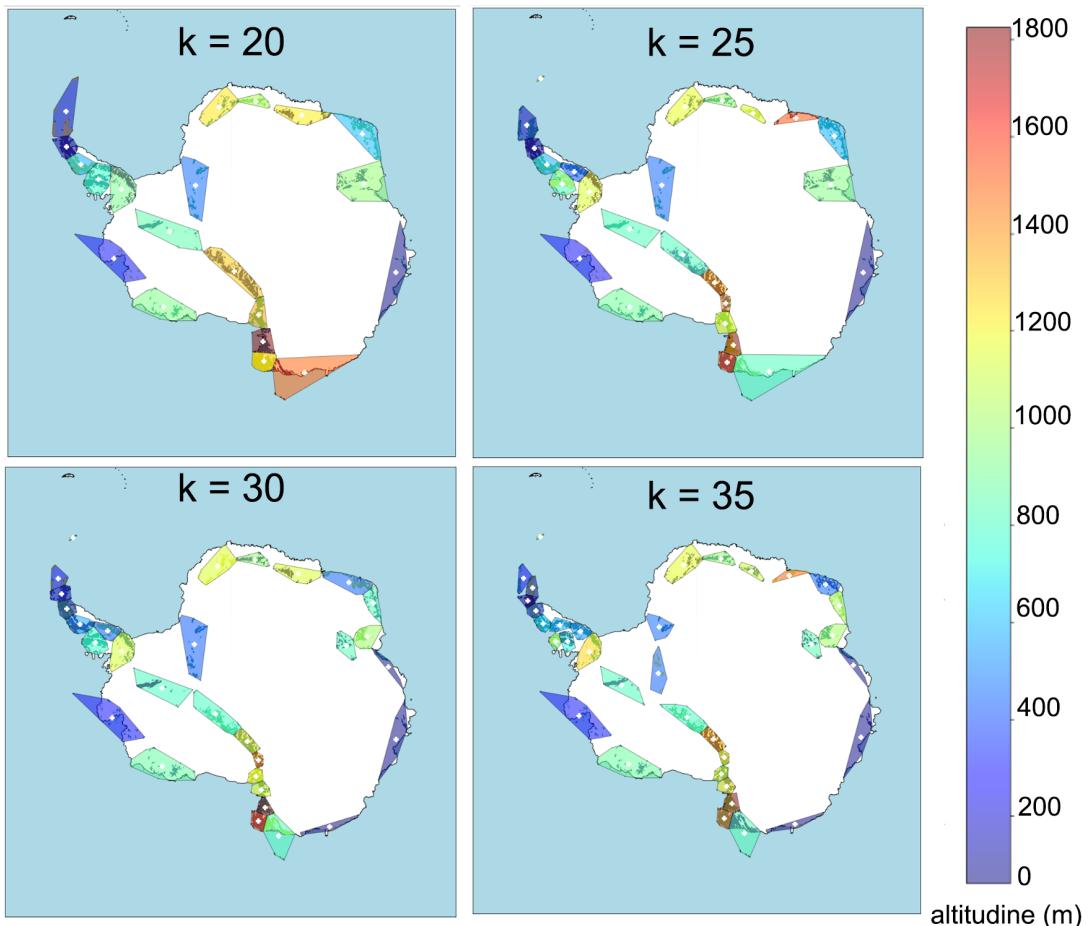


Figura 2.9: Esempio di Clustering k-means++ con valore $k = [20, 25, 30, 35]$

Una volta ottenuto un raggruppamento dei cropout secondo un clustering pos-

siamo calcolare delle statistiche descrittive di interesse per andare ad arricchire le informazioni sui punti di perforazione in nostro possesso. In particolare per ogni sito di perforazione andiamo a calcolare:

- **Distanza dalla costa:** ci interessa per sapere il potenziale di trasporto di particelle di polvere provenienti da oltreoceano, ossia dai continenti dell'emisfero sud, verso i punti più interni del continente Antartico, quanto più un punto si troverà vicino la costa quanto più sarà la possibilità che quel sito abbia forti influenze da più fonti.
- **Cluster\Outcrop più vicino:** È molto importante sapere se nelle prossimità delle stazioni di prelievo delle carote di ghiaccio sono presenti outcrop, in mancanza della superficie erodibile rappresentata dai cluster di roccia esposta infatti non ci sarebbe il potenziale trasporto del materiale erosivo dai cluster ai punti di perforazione ad opera del vento.
- **Arearie di cropout entro un determinato raggio:** Oltre la distanza dal più vicino outcrop roccioso viene calcolata l'area totale di outcrop presente all'interno del raggio di 500 e 1000 chilometri a partire dalle coordinate del punto di perforazione. Quanto più ci sarà presenza di outcrop nei dintorni dei siti di carotaggio quanto più sarà possibile che dei venti di forte intensità possano sollevare e quindi spostare il materiale erosivo e esposto in quota.

Allo stesso modo vengono integrate delle informazioni sui clusters identificati, in particolare:

- **Elevazione e Range:** Viene calcolata l'elevazione media di ogni cluster e il range di quest'ultima, ossia la differenza tra il punto più in alto e quello più in basso sul livello del suolo.
- **Area:** Area totale del cluster in km quadrati.
- **Area rocks/unconsolidated:** Area totale della frazioni di outcrop roccioso e non roccioso.

In tabella 2.2 un esempio dei dati estratti per alcune stazioni di interesse, mentre in tabella 2.3 le informazioni calcolate per alcuni clusters d'esempio.

Stazione	Distance From Coastline (km)	Cropout Area Within 1000km (km^2)	Nearest Cropout (km)
Dome-B	857.75	1216.26	695.40
SOLARICE-DC	864.86	4226.02	885.76
DOME FUJI	773.77	2621.77	547.33
TAYLOR DOME	103.02	15376.44	37.22

Tabella 2.2: Integrazione di variabili spaziali per le stazioni

Cluster	Mean Elevation (m)	Elevation Range (m)	Area (km^2)	Cropout Percentage (%)
10	64	1161	171014	0.35
12	1176	1903	47236	3.34
14	1825	2615	43287	5.29
24	1227	2290	40302	15.83

Tabella 2.3: Integrazione di variabili spaziali per i clusters

A partire da questi dati è possibile costruire delle visualizzazioni che contengano più layer informativi per comprendere meglio come le quantità di polveri di interesse varino in base alla distanza dalla costa, dalla posizione e presenza dei cluster e dagli outcrop più vicini. Nell'esempio riportato di seguito possiamo vedere alcune di queste variabili visualizzate insieme, nella figura 2.10 vediamo la distanza in chilometri dalla costa nelle X , la distanza dal cluster più vicino nelle Y , la quantità di accumulo del flusso viene rappresentata come la dimensione del data-point, ed il colore rappresenta la concentrazione di flusso totale.

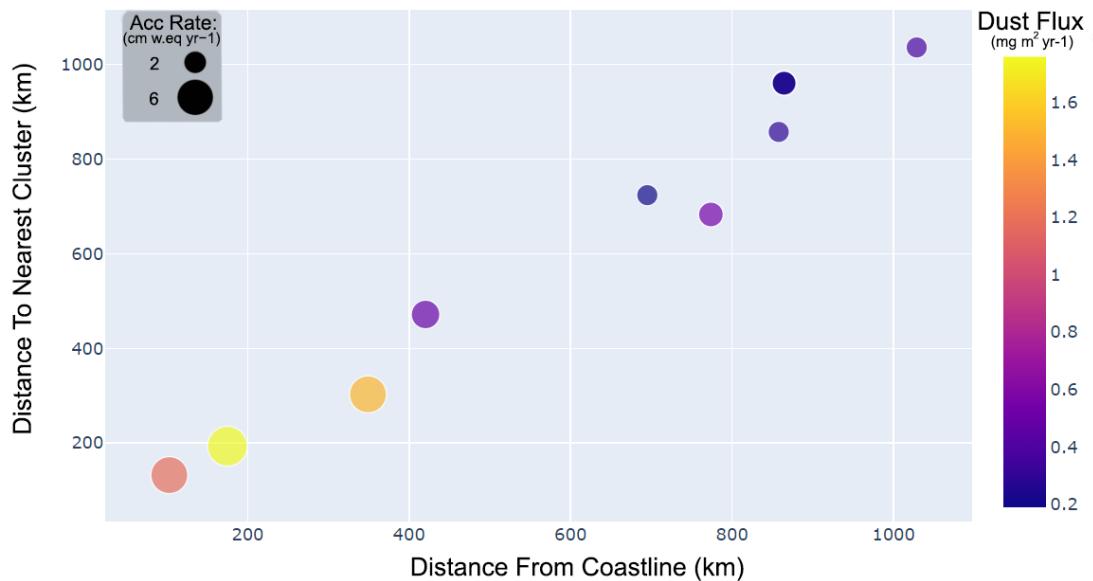


Figura 2.10: Visualizzazione multi-variata distanza dai cluster vs flusso

Come è possibile vedere dal grafico le stazioni più vicine alla costa sono anche quelle più vicine ai clusters, e tendenzialmente sono anche quelle che tendono ad avere la maggior concentrazione di polveri sia in termini totali (come identificato dal colore) sia in termine di rapporto di accumulo (identificato dalla dimensione), ciò conferma le nostre ipotesi su quali siano le aree di maggior interesse per i nostri scopi.

2.3 Analisi dati meteo

Una volta creati i cluster è di vitale importanza integrare queste informazioni con delle variabili climatiche che descrivano il contesto in cui ogni cluster si inserisce, per fare ciò andremo quindi a raccogliere, trasformare ed analizzare alcuni dei dataset esistenti. In primis verranno raccolte informazioni sui venti in quanto di fondamentale importanza per identificare i periodi temporali più interessanti al fine di valutare il trasporto del materiale erodibile dalla fonte alle possibili destinazioni. A corredo di queste informazioni sui venti verranno raccolte anche altri dati secondari che potrebbero comunque rivelarsi utili all'analisi dei pattern meteorologici che vogliamo studiare, informazioni quali la temperatura e l'umidità.

2.3.1 raccolta dei dati AWS

Sfortunatamente le stazioni meteo dislocate nel continente Antartico non fanno riferimento ad un solo ente organizzativo, ogni paese che ha delle missioni di ricerca in Antartide raccoglie e cataloga le informazioni ricavate dalle stazioni meteo con metodologie e tempistiche differenti, non esiste quindi un'unica fonte dato per quello che ci interessa. Per la fase di retrieve sarà quindi necessario mettere insieme le possibili fonti, ma per evitare di spendere troppo di **data ingestion** bisognerà concentrare gli sforzi sulle sorgenti più ricche che forniscono quanti più dati possibili. Questa scelta viene fatta anche in base a quali fonti dato forniscono la copertura del territorio più adatta al nostro scopo, esistono infatti vaste aree di Antartide con scarsa quantità di outcrop roccioso, avere delle informazioni meteo su queste regioni è indubbiamente utile, ma meno rispetto alle aree che presentano una densità di

outcrop roccioso molto più alta come le aree costiere.

Una delle fonti dato più complete per quanto riguarda i dati climatici in Antartide è quella del progetto **AMRC** (Antarctic Meteorological Research Center), il progetto inizia dalle ricerche del Prof. Werner Schwerdtfeger presso l'Università del Wisconsin nel corso degli anni '60, continua poi con le prime installazioni di stazioni meteo in Antartide nel 1979 fino ai giorni nostri grazie ai fondi stanziati dalla **National Science Foundation**. L'Università del Wisconsin conta circa 63 stazioni **AWS** (Automated Weather Station) sparse nell'Antartide che servono come base dati per molteplici applicazioni, tra cui appunto anche il progetto **AMRC** che si è occupato nella costruzione di un sistema meteorologico unificato e integrato da dati di imaging compositi ottenuti dai satelliti. Come è possibile vedere in figura 2.11 il progetto **AMRC** copre buona parte del territorio Antartico, soprattutto le fasce di territorio vicino le coste, che risultano poi quelle più dense dal punto di vista dell'outcrop roccioso, come possibile notare in figura 2.3

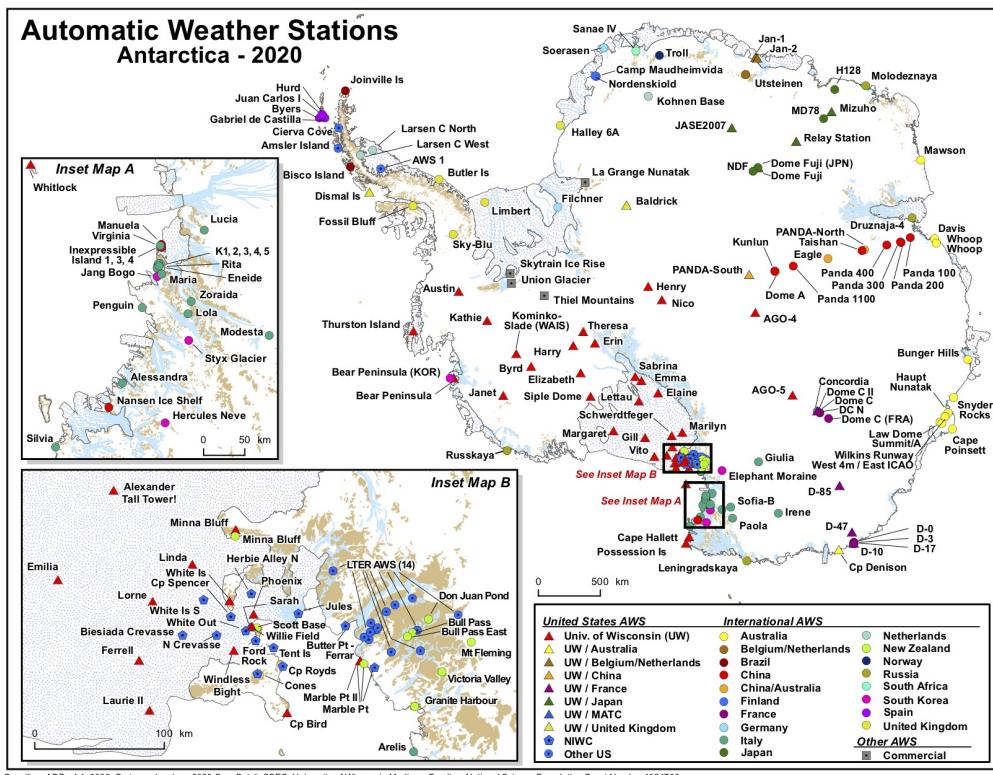


Figura 2.11: Distribuzione delle stazioni AWS in Antartide con relativi enti di appartenenza

I dati del progetto **AMRC** sono resi disponibili tramite una connessione **FTP** (File Transfer Protocol) raggiungibile al link <ftp://amrc.ssec.wisc.edu/>, accedendo alla cartella *pub/aws/antrdr* è possibile trovare i dati delle stazioni meteo suddivisi per anno, per ogni anno sono disponibili i file per le diverse stazioni meteo che il progetto mette a disposizione. Ogni file contiene le osservazioni meteorologiche con una risoluzione temporale di tre ore ed è nominato tramite 7 cifre, le prime tre specificano il codice **ARGOS** ossia un identificativo di un sistema satellitare che si occupa di processare e collezionare dati ambientali da varie fonti, le successive 4 cifre identificano il mese e l'anno del dato che stiamo leggendo. Le variabili ambientali sono organizzati per colonne, nello specifico troviamo per ogni step di tre ore relative al mese e all'anno in esame informazioni più o meno complete su: temperatura (in gradi celsius), pressione (in millibar), velocità del vento (in metri al secondo), direzione del vento, umidità relativa e differenza di temperatura verticale (misurata come la differenza della temperatura ad un'altezza di 3 metri rispetto all'altezza di riferimento a 0.5 metri). Le informazioni sulle ultime due colonne, ossia l'umidità relativa e la differenza di temperatura verticale non sono comunque presenti per tutte le stazioni, difficilmente queste variabili potranno rivelarsi utili all'indagine. Le dimensioni di analisi più interessanti sono invece quelle relativa alla temperatura e alla potenza e direzione del vento. Anche per queste variabili esistono alcuni valori mancanti (indicati dal valore **444.0** all'interno del dataset). La direzione del vento può assumere valori da 0 a 360 gradi e indica l'angolo in senso orario a partire da Nord, un vento proveniente da Sud sarà quindi indicato con un valore di 180, così come un vento da Ovest avrà valore di 270. Gli orari di riferimento del dataset partono dalla mezzanotte e hanno step di tre ore, come già detto, la presenza di un particolare record all'interno del dataset è legato all'esistenza di un'osservazione valida per quella particolare stazione all'interno di un intervallo massimo di 40 minuti.

Una volta finita la fase di estrazione dei **15417** file presenti sul server FTP prima citato si ottiene un dataset con **3'771'532** record.

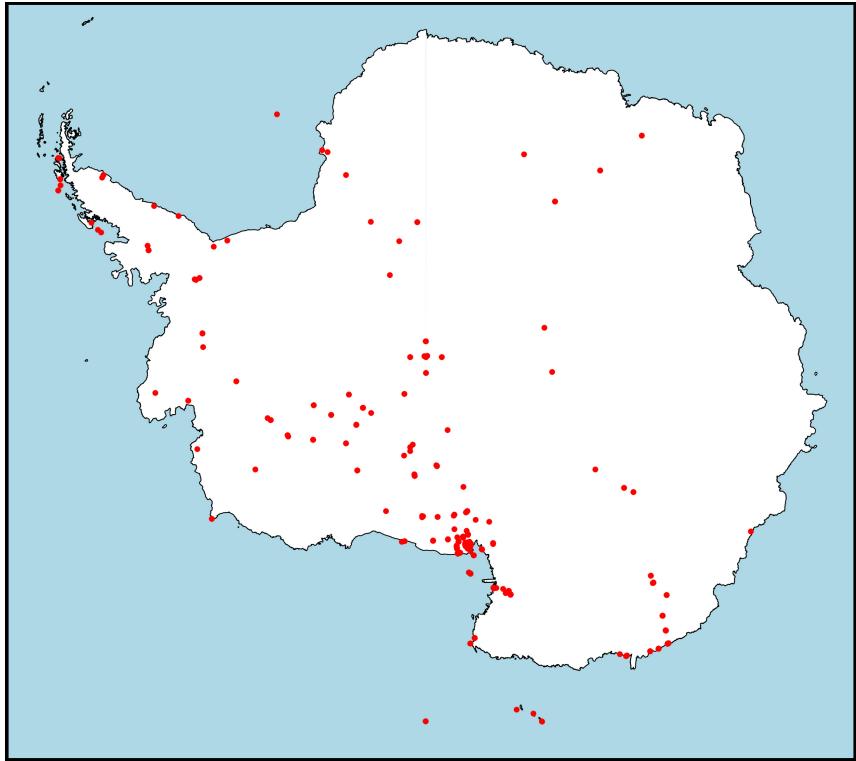


Figura 2.12: Distribuzione delle stazioni **AMRC** in Antartide di cui si posseggono i dati

In figura 2.12 possiamo vedere le posizioni delle stazioni meteo di cui abbiamo dei dati, mentre in figura 2.13 viene rappresentata la distribuzione temporale aggregata dei dati in nostro possesso su tutte le stazioni. Come è possibile vedere dal trend delle osservazioni la disponibilità dei dati aumentano nel tempo, a eccezione degli anni compresi tra il 2003 e il 2009, la massima concentrazione si ha invece negli anni tra il 2010 e il 2012.

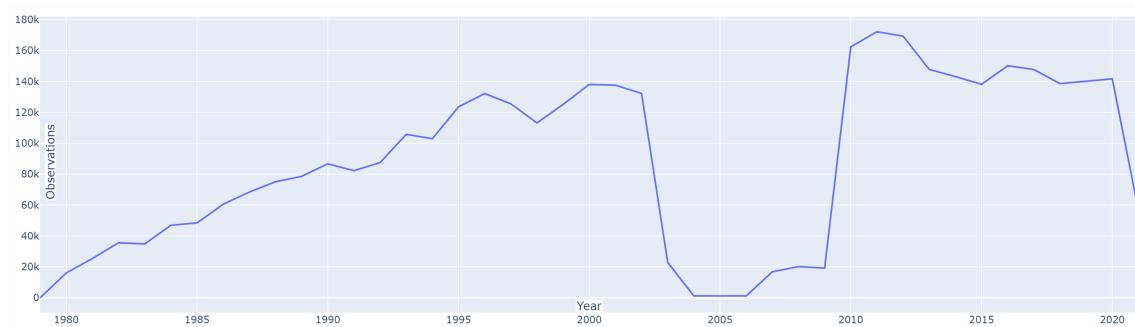


Figura 2.13: Osservazioni meteorologiche disponibili per ogni anno

2.3.2 Integrazione e arricchimento

Una volta estratti i dati grezzi delle stazioni il nostro scopo principale è quello di arricchire le informazioni sui cluster per comprendere quali dei raggruppamenti di outcrop identificati presenta anche delle informazioni climatiche interessanti, per fare ciò utilizzeremo i **Convex-Hull** identificati nella fase di clustering. Intersecando questi poligoni rappresentativi dell'area dei cluster, con le coordinate delle stazioni, possiamo mettere in relazione i clusters identificati nello step precedente con i dati delle stazioni meteo appena estratti, arricchendo ulteriormente la nostra conoscenza sui raggruppamenti di Outcrop. In Figura 2.14 possiamo visualizzare il risultato ottenuto, in rosso i cluster che contengono almeno una stazione meteo al loro interno, in blu i cluster che invece non contengono nessuna stazione del progetto AMRC. Come possibile notare esistono alcune stazioni meteo, nella regione costiera a sud, che seppur non contenute strettamente in nessun cluster ne risiedono poco al di fuori. Potrebbe essere una buona idea considerare come facenti parte di un cluster anche le stazioni meteo all'interno di una data distanza massima dal Convex-Hull che identifica il cluster stesso.

Alla fine di questo processo di integrazione tra le stazioni AWS e i Convex-Hull otterremo una tabella descrittiva completa dei clusters e delle variabili di interesse. Su queste informazioni sarà possibile eseguire una fase esplorativa per evidenziare dei trend nel clima, quindi descrivendo le stagionalità a diverse risoluzioni (annua, mensile, giornaliera). Per supportare queste analisi verrà implementato un piccolo sistema di query che permetta di specificare un set di attributi e ricevere i dati richiesti, in particolare questo sistema di query potrà filtrare i dati finora ottenuti a seconda delle seguenti variabili:

- **Date:** specificata da due campi: **from date**, e **to date** che ci permettono di selezionare un range temporale per i dati che vogliamo estrarre.
- **Variables:** che ci permette di selezionare quali variabili ci interessano tra quelle disponibili (elevation, temperature, pressure, wind speed, wind direction, relative humidity, vertical temperature difference)

- **Stations:** con questo attributo possiamo selezionare un subset di stazioni AWS da cui vogliamo prelevare i dati, tramite questo campo e il lavoro svolto precedentemente potremo quindi selezionare solo le stazioni meteo appartenenti ad un determinato cluster.
- **Wind Above:** con questo campo possiamo specificare una soglia per filtrare solo i record che presentano una velocità del vento superiore al valore indicato, in metri al secondo.

In Figura 2.14 due visualizzazioni di esempio. A sinistra i Cluster che hanno delle stazioni AWS al loro interno o nelle immediate prossimità (rosso) e quelle che invece non ne hanno (blu) con le relative stazioni AWS sulla mappa. A destra la velocità media del vento per le osservazioni appartenenti alle stazioni meteo che caratterizzano i clusters.

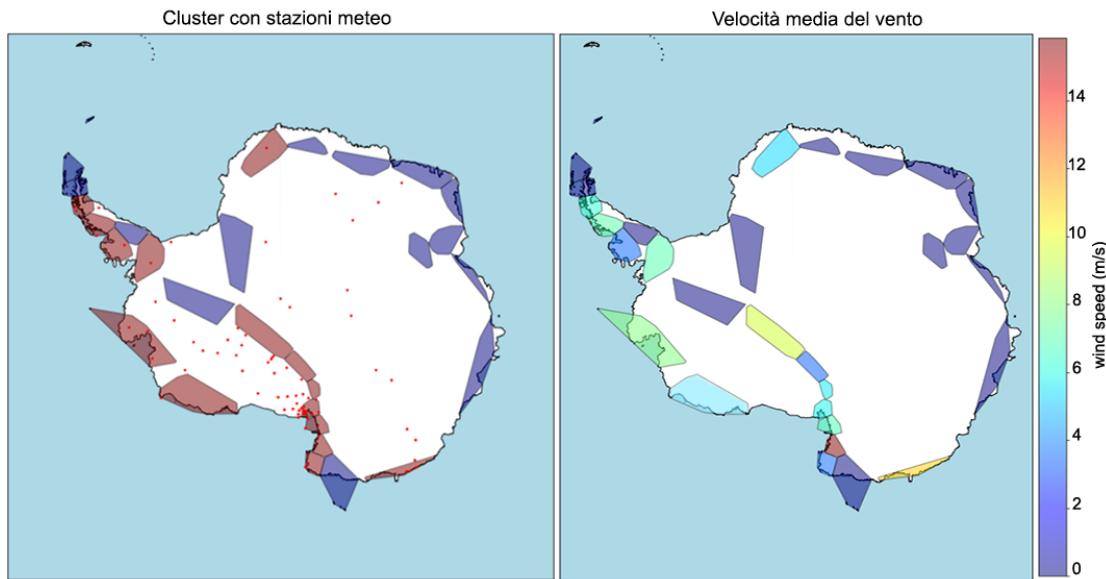


Figura 2.14: Esempio visualizzazioni su variabili meteorologiche integrate sui Cluster

Generazione delle traiettorie

In questo capitolo l’obiettivo sarà quello di collegare tutte le informazioni raccolte finora tramite l’utilizzo di un modello di simulazione per il calcolo delle traiettorie di masse d’aria, in particolare il software utilizzato sarà HYSPLIT. Nelle prime sezioni del capitolo introdurremo HYSPLIT e vedremo come il modello simuli l’evoluzione delle masse d’aria per estrarne delle traiettorie. Successivamente utilizzeremo le informazioni in nostro possesso identificare il periodo migliore dove concentrare le analisi e far partire la simulazione, organizzando i risultati in modo che nel prossimo capitolo sia possibile analizzarli e cercare di quantificare quanto un determinato cluster di outcrop sia in grado di influenzare un carotaggio a centinaia, se non a migliaia di chilometri di distanza.

3.1 Introduzione a HYSPLIT

In questa sezione introdurremo brevemente il modello lagrangiano utilizzato per la generazione delle traiettorie: (**HYSPLIT**) e del tool di cui ci avvarremo per ottimizzare l’esecuzione di molteplici simulazioni (**PYSPPLIT**). Fatto ciò descriveremo il processo di simulazione e come esso faccia evolvere le masse d’aria nel tempo e nello spazio per generare delle traiettorie.

3.1.1 HYSPLIT

L’**Hybrid Single-Particle Lagrangian Integrated Trajectory** model[12] è un modello utilizzato per il forecasting delle traiettorie di particelle di aria, il progetto **HYSPLIT** è stato creato ed è attualmente mantenuto dall’ente governativo

statunitense **NOAA**, che sta per: **National Oceanic and Atmospheric Administration** presso l'**Air Resource Laboratory**, di concerto con l'**Australian Bureau of Meteorology Research Center** nel 1998.

Per produrre le sue previsioni HYSPLIT utilizza un approccio alla fluido-dinamica ibrido tra quello Lagrangiano e quello Euclideo, nel sistema di riferimento Lagrangiano è come se si seguisse un piccolo volume di fluido, in questo caso l'aria, attraverso il tempo e lo spazio. Nel sistema di riferimento Euclideo invece il sistema è definito da una regione di spazio ben delineata dove il fluido evolve e scorre col passare del tempo, rappresentato da dei nodi che riempiono lo spazio che stiamo osservando.

3.1.2 Air Parcel Trajectory

Nel caso del calcolo della traiettoria di una particella (piccolo volume) ideale di aria, viene utilizzato il modello Lagrangiano, si parte quindi da una condizione iniziale per poi far evolvere il sistema al fine di calcolare la nuova posizione del pacchetto di aria che stiamo simulando attraverso lo spazio ed il tempo. Per fare ciò Hysplit utilizza dei dati meteorologici contenenti le principali variabili atmosferiche per ogni punto della griglia di riferimento che viene utilizzata, in base alla risoluzione spaziale disponibile. Le principali variabili contenute in questi file si riferiscono alla descrizione dei dati meteorologici in input ad una quota superficiale, tra di esse troviamo:

- Componenti orizzontali e verticali del vento ad un'altezza di 10 metri
- Temperatura a 2 metri dal suolo e alla superficie
- Pressione atmosferica alla superficie e al livello medio del mare
- Umidità relativa
- Percentuale di cielo coperto da nubi

Questi dati vengono riportati per ogni data-point che costituisce la griglia e a partire da questi viene fatta un'interpolazione per ottenere dei dati in tutte le posizioni intermedie, fatto ciò si otterrà un file completo da poter utilizzare per far evolvere le particelle di aria all'interno di questo spazio. Più nello specifico, una volta interpolati

e convertiti le principali variabili meteo (U , V , W) le tre componenti spaziali della traiettoria vengono calcolate indipendentemente le una dalle altre, e solo dopo le tre componenti vengono mediate per ottenere il vettore velocità risultante per far evolvere la posizione iniziale al tempo t della particella o del pacchetto che stiamo considerando, formalmente si avrà che:

$$P'(t + \Delta t) = P(t) + V(P, t)\Delta t \quad (3.1)$$

dove $P(t)$ rappresenta la posizione iniziale, i vettori di velocità sono quindi interpolati sia nello spazio che nel tempo per ottenere questa prima posizione parziale, per poi calcolare la posizione finale come:

$$P(t + \Delta t) = P(t) + 0.5[V(P, t) + V(P', t + \Delta t)]\Delta t \quad (3.2)$$

Questo metodo di integrazione è molto comune per l'analisi delle traiettorie, metodi più complessi di integrazione spaziale e temporale sono stati presi in considerazione nel corso dello sviluppo del modello di HYSPLIT ma i risultati sperimentali hanno confermato che con dei dati meteorologici interpolati linearmente l'utilizzo di metodologie di integrazione più complesse non risulta in precisioni maggiori, a parità di performance è stata presa la decisione di mantenere il modello più semplice possibile. È importante notare come il delta di integrazione Δt non sia definito ma può cambiare nel corso della simulazione, questo step viene calcolato in base alle necessità e alla risoluzione della griglia. La distanza coperta da una particella d'aria in uno step non può infatti essere maggiore allo spaziamento della griglia di riferimento che stiamo utilizzando. Per questo motivo viene definita una velocità massima di trasporto U_{max} che viene elaborata sulla base della velocità raggiunta dalla massa d'aria nel corso dell'integrazione al time-step precedente, lo step di integrazione può variare da un minuto fino ad un ora e viene limitato dalla seguente relazione:

$$U_{max} \left(\frac{GridUnits}{min} \right) \Delta t < 0.75(GridUnits) \quad (3.3)$$

3.1.3 Raccolta dei dati climatici e identificazione dello starting point ideale

Per restringere il campo di analisi entro cui vogliamo muoverci utilizzeremo le informazioni raccolte durante il secondo step, come già visto in figura 2.13 i dati in nostro possesso identificano degli anni più consoni per il calcolo delle retro-traiettorie rispetto ad altri. In particolare gli anni dal 2010 al 2012 si sono rivelati essere quelli dove abbiamo più informazioni riguardo la velocità del vento nei vari clusters, per questo motivo concentreremo le analisi al centro di questo intervallo, ossia nell'anno **2011**. HYSPLIT, come già descritto, ha bisogno di dati climatici per far evolvere un'air parcel attraverso il tempo e lo spazio, il progetto che ha dato vita al software si occupa anche di catalogare e rendere disponibili questi file in diversi formati tramite un **FTP** server che è possibile raggiungere al link: <ftp://arlftp.arlhq.noaa.gov/archives>, all'interno di questo archivio sarà possibile trovare i files necessari a diverse risoluzioni spaziali. Nel nostro caso utilizzeremo i files presenti nella sottocartella **gdas1**, i files GDAS, che sta per **Global Data Assimilation System** sono creati e mantenuti dal **National Weather Service's National Centers for Environmental Prediction**, e sono file con risoluzione spaziale di 1 grado e risoluzione temporale di 3 ore. I dati sono disponibile nel formato **gdas1.myyy.w#** dove **mmyy** indicano rispettivamente mese e data, e **w#** indica invece la settimana di quel mese a cui si sta facendo riferimento. Per avere i dati che ci servono bisogna avvalersi di tutti i file gdas del 2011 più i dati necessari per coprire traiettorie indietro nel tempo dall'1 Gennaio 2011, quindi i file sulla fine del 2010, e allo stesso modo, i primi file del 2012 nel caso dovessimo andare avanti nel tempo a partire dalla fine di Dicembre 2011. Nel nostro caso la lunghezza del simulazioni che andremo a fare è di 240 ore, ossia 10 giorni, sarà quindi sufficiente scaricare, oltre i dati del 2011, le prime due settimane del 2012 e le ultime 2 del 2010, per un totale di 30.9 GB divisi in 63 files. Per scegliere il periodo temporale migliore per l'analisi bisogna prendere in considerazione anche la componente stagionale Antartica, in questo continente infatti il ciclo delle stagioni è ben diverso dal nostro e le stagioni influenzano direttamente

anche la circolazione dei venti.

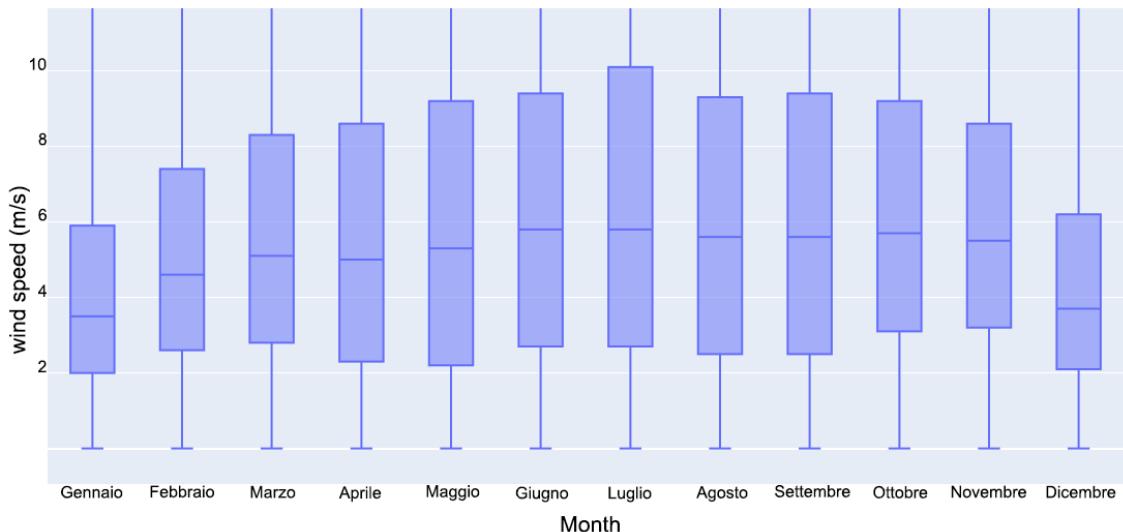


Figura 3.1: Distribuzione della velocità del vento media giornaliera per le stazioni **AMRC** per ogni mese nell’anno 2011

In figura 3.1 vengono riportare le velocità del vento rilevate nelle stazioni AMRC durante l’anno 2011, dall’analisi dei boxplot è possibile vedere come nei mesi centrali dell’anno, quindi durante la stagione Antartica invernale, la componente del vento sia più molto intensa soprattutto nel mese di Luglio. Questa tendenza si inverte durante la stagione estiva, seppur molto breve, di Dicembre Gennaio. Oltre le diverse intensità del vento, le due stagioni Antartiche differiscono anche nelle direzioni che il vento assume. Nei mesi estivi si osservano direzioni che tendono ad essere più dalle coste verso l’entroterra rispetto che nei mesi invernali, dove seppur si abbia la massima intensità la direzione del vento viaggia più spesso dall’entroterra verso la costa. In base alla posizione degli attori in gioco, sappiamo che i clusters di outcrop roccioso sono più presenti nelle zone esterne, invece le stazioni meteorologiche sono più concentrate nell’entroterra, questo crea una situazione peculiare per cui i mesi con le minori raffiche di vento potrebbero comunque rivelarsi i più significativi nel contesto della nostra analisi. Per questi motivi la scelta è stata quella di concentrare il calcolo delle traiettorie da e per le stazioni nei due mesi più rappresentativi della dicotomia appena discussa, quindi nei mesi di Gennaio e di Luglio, così da poter analizzare quanto effettivamente questi fattori contribuiscano allo spostamento delle polveri all’interno del budget totale.

3.2 Calcolo delle traiettorie

In questa sezione vedremo quali traiettorie sono state create descrivendo le condizioni iniziali sia dal punto di vista geografico che temporale. Ottenute tutte le traiettorie simulate vedremo come poter creare uno score di influenza che riassume quanto ogni sito di perforazione sia influenzato dai diversi cluster identificati nel capitolo 2. Ci occuperemo infine di eseguire un clustering delle traiettorie per ogni sito così da facilitare l’analisi visiva dei principali pattern di circolazione per ogni sito considerato.

3.2.1 Starting points della simulazione e output ottenuto

Identificati i periodi migliori utilizziamo **PYSPLIT** per far partire le nostre analisi. **PYSPLIT** è una comoda libreria per Python che si occupa di interporsi tra l’utente e l’installazione locale di **HYSPPLIT** allo scopo di fornire una comoda interfaccia tra i due mondi. Utilizzando PYSPLIT sarà infatti possibile automatizzare l’esecuzione della simulazione così da risparmiare molto tempo, non dovremo, ad esempio, specificare le condizioni di partenza per ogni singolo fascio di traiettorie, oltre l’esecuzione del forecasting PYSPLIT fornisce molte altre funzioni per caricare più fasci di traiettorie, analizzarli, visualizzarli o creare dei clustering sempre appoggiandosi a HYSPLIT. Una volta settate tutte le corrette working-directories tramite PYSPLIT possiamo inizializzare delle simulazioni con i parametri prescelti, grazie a questa metodologia possiamo calcolare in maniera esaustiva il flusso del vento a partire dalle stazioni indietro nel tempo, o dai cluster in avanti, per vedere se esistono delle traiettorie che facciano da ponte tra le regioni ad alta presenza di outcrop e i siti di carotaggio del ghiaccio.

Per generare le traiettorie facciamo partire una simulazione per 10 giorni in avanti o indietro nel tempo, ciò viene fatto per ogni sito di perforazione, per entrambi i mesi considerati, per ogni giorno, per 4 diversi orari (0, 6, 12, 18), con 5 possibili altitudini di partenza dalla stazione (5, 25, 50, 500 e 1000 metri). Per ogni combinazione di questi parametri troveremo quindi un file di output contenente la retro-traiettoria più probabile, calcolata facendo evolvere il sistema con l’equazione di stato discussa

precedentemente. Oltre le back-trajectories a partire dalla stazione generiamo anche delle forward-trajectories a partire dai centri di massa di ogni cluster, per calcolare i centri di massa dei cluster viene effettuata una media pesata dei centroidi dei poligoni che descrivono tutti gli outcrop che costituiscono un cluster, utilizzando l'estensione dell'outcrop come peso per la singola osservazione. A partire da questi starting points vengono quindi inizializzate delle simulazioni con le stesse condizioni di partenza viste per le stazioni, ma con diverse altitudini, ossia: 5, 25, 50, 100, 200. Alla fine di questa fase avremo ottenuto quindi un grande set di traiettorie, sia in avanti dai 30 clusters (**36000**), sia all'indietro dalle 11 stazioni (**12960**), ciò che rimane da fare è andare ad analizzare i percorsi che queste traiettorie hanno seguito e valutare se e quanto i cluster sono riusciti ad arrivare tramite dei pattern di circolazione fino alle stazioni, quantificando questa informazione potremmo avere un'approssimazione di quanto ogni stazione sia effettivamente influenzata dai venti provenienti dai clusters.

3.2.2 Calcolo dell'influenza dei cluster sulle stazioni

I file prodotti da HYSPLIT possono essere interpretati come una serie di coordinate spaziali, ogni coordinata è rappresentata da una tripletta di valori che ci dicono la posizione spaziale a quel determinato step, la distanza tra coordinate adiacenti ci da anche un'idea della velocità del vento. PYHYSPLIT ci permette di definire un oggetto di tipo **TrajectoryGroup** che contiene le informazioni su un gruppo di traiettorie, non ci resta quindi che caricare i TrajectoryGroup per ognuno dei fasci di traiettorie a disposizione e analizzare le coordinate per vedere quando e come esse possano essere causa di uno spostamento significativo di materiale.

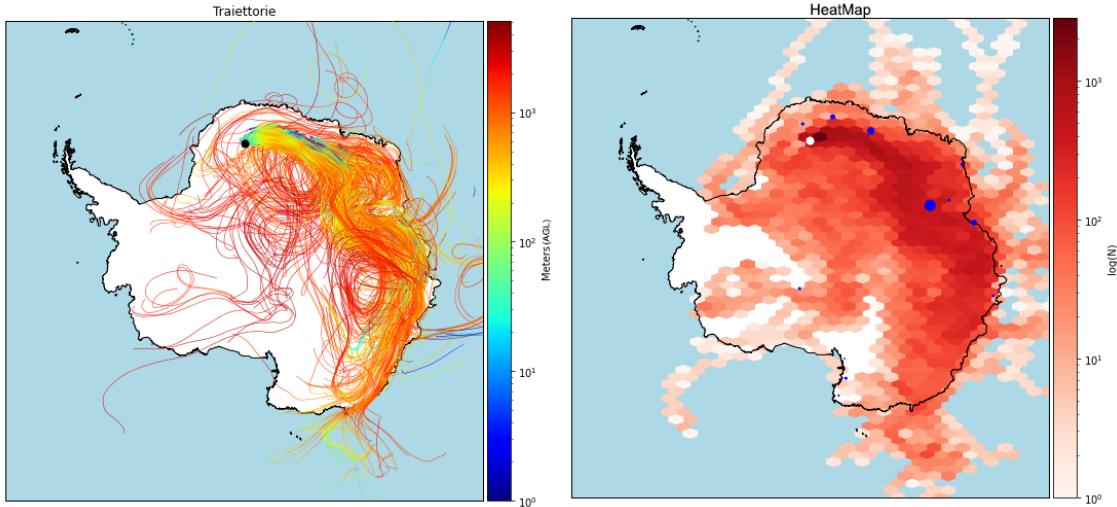


Figura 3.2: Visualizzazioni delle retro-traiettorie prodotte per il sito EPICA-DML

In Figura 3.2 è possibile vedere un esempio di visualizzazione delle retro-traiettorie prodotte per la stazione **EPICA-DML** a partire dalle 5 altitudini in esame per tutti i giorni del mese di Gennaio 2011, quindi in piena stagione estiva. Nella prima figura le retro-traiettorie "grezze" dove il colore indica l'altitudine sopra il livello del terreno per ogni dato punto, in scala logaritmica. Nel grafico di destra viene mostrata un'agglomerazione delle traiettorie su una griglia esagonale, producendo così un'**HeatMap** del passaggio del vento nel suo percorso verso la stazione, ciò ci aiuta a identificare visivamente le zone che più contribuiscono al flusso in ingresso. Questa analisi potrebbe tuttavia non bastare per calcolare i punteggi di influenza dei cluster e anzi potrebbe introdurre delle distorsioni significative, in quanto, può verificarsi il caso in cui seppur molte traiettorie passino per una cella della griglia esagonale definita queste traiettorie potrebbero trovarsi ad alta quota, come vediamo nel primo grafico per i segmenti di traiettorie che tendono più verso il rosso, nel calcolare quindi un indice di influenza bisognerà tenere conto anche dell'altitudine della traiettoria per assicurarsi di contare solo quei percorsi che seppur trovandosi in quota per la maggior parte del tempo passano poi rasenti al suolo in prossimità dei cluster (quando analizziamo le back-trajectories partendo dalle stazioni) o in prossimità delle stazioni (quando analizziamo le forward-trajectories partendo dai clusters).

Per calcolare i punteggi di influenza viene utilizzata una funzione logistica che

dipende da due fattori, vengono presi in considerazione la distanza a cui la traiettoria passa rispetto alla sua destinazione e contemporaneamente, come anticipato, la quota verticale al momento del passaggio. Queste due variabili vengono valutate singolarmente da una funzione logistica personalizzata che va a produrre uno score per entrambe le dimensioni definite. La funzione logistica è una curva molto comune, utilizzata in molti ambiti del Machine Learning, della statistica e della Data Science, sia come funzione di regressione, sia come funzione di attivazione per le reti neurali o anche come funzione per modellare l'evoluzione di una popolazione. Gran parte dell'utilità della funzione logistica proviene dalla sua capacità di essere adattata ai vari use-case in base ai parametri che la funzione mette a disposizione per essere personalizzata, i tre parametri in questione sono:

- \mathbf{x}_0 : Identifica il valore di ascissa che assume la funzione nel suo punto centrale.
- L : Descrive il valore massimo assumibile dalla funzione
- k : Il tasso di crescita della curva.

La funzione logistica nella sua forma generalizzata avrà quindi la seguente formalizzazione:

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}} \quad (3.4)$$

Nel nostro contesto i valori scelti per la funzione logistica sono $L = 1$, $k = -4$ e $\mathbf{x}_0 = 0.5$, avremo quindi un growth-rate abbastanza elevato, ciò ci permette di considerare punti anche moderatamente distanti dalla propria destinazione pur mantenendo il peso di queste osservazioni basso. Questa scelta è stata fatta per la natura del problema, dove le distanze possono essere anche molto elevate e le traiettorie sono più un indicazione dei possibili valori reali in quanto prodotti di varie interpolazioni tra la dimensione della griglia e la risoluzione dei dati meteorologici disponibili. Un'ulteriore modifica è stata apportata alla funzione, la logistica per sua natura assume i due valori estremi $[0, 1]$ in corrispondenza dei due asintoti orizzontali $[-\infty, +\infty]$, per adattarla al nostro use-case bisognerà normalizzare la funzione

per far si che si operi solo nel range $[0, 1]$ delle ascisse, in quanto i due valori di distanza e altitudine che verranno passati saranno anch'essi normalizzati nel range $[0, 1]$ rispetto ai valori soglia che andremo a definire. In figura 3.3 è possibile vedere la funzione logistica descritta dalle variabili scelte nella sua formulazione originale e nella versione normalizzata specificatamente per il nostro use-case.

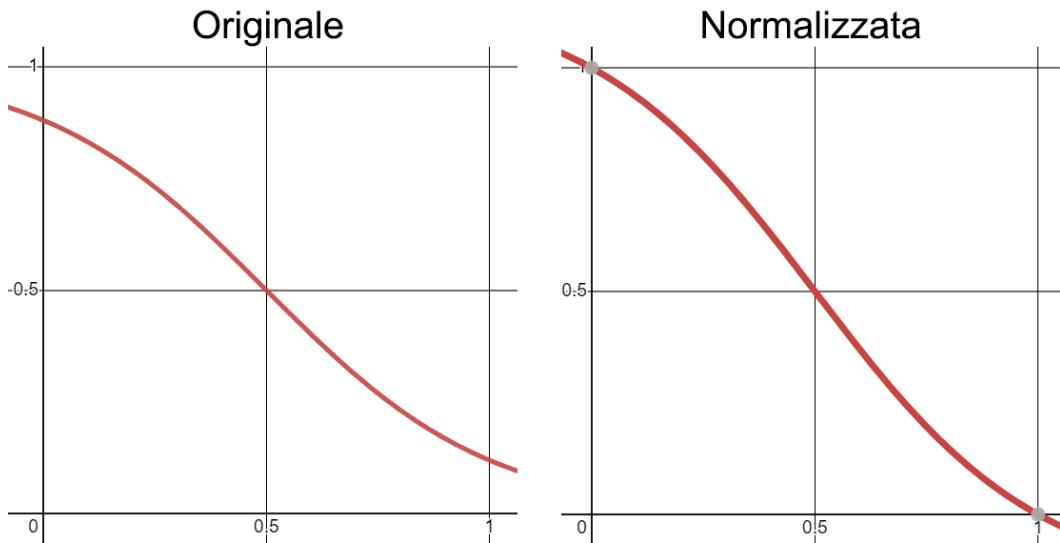


Figura 3.3: Curva logistica scelta per il calcolo dello score di influenza

Una volta calcolato lo **score** ottenuto dalla suddetta curva logistica per entrambe le dimensioni di esame (distanza e altitudine) il punteggio finale sarà la media tra i due score, è importante notare come il punteggio sia additivo e che esso venga calcolato per ogni punto della traiettoria, quindi un passaggio vicino ma breve avrà meno peso di un passaggio altrettanto vicino ma più reiterato nel tempo, ad esempio a causa di correnti circolari o a seconda della natura della traiettoria stessa, in questo modo una traiettoria che ripassa e viaggia più volte nelle prossimità della sua destinazione avrà un peso maggiore rispetto a quelle che si allontaneranno brevemente, con questa scelta si massimizza la possibilità che le traiettorie più influenti siano quelle che hanno avuto più tempo a disposizione per prelevare o depositare del materiale (a seconda della direzione dell'analisi se in avanti dai cluster alle stazioni, o viceversa). Il calcolo procederà nelle due direzioni a disposizione, nel caso degli score di influenza dalle stazioni, indietro nel tempo, fino ai cluster abbiamo delle quote di partenza che sono quelle delle stazioni, ossia $[5, 25, 50, 500, 1000]$, e ci interessa che a partire da queste quote le traiettorie che, possibilmente, trasportano del

materiale vadano a finire nelle prossimità dei centroidi dei cluster, abbastanza vicino al suolo, a questo scopo quindi andiamo a considerare solo i punti che rientrano in un raggio di **100km** dal centroide del cluster e fino ad un'altezza massima di **100m**.

Nel caso opposto, quindi quando consideriamo le traiettorie in avanti nel tempo a partire dai cluster fino alle stazioni, vogliamo che le quote di partenza siano relativamente basse, perché devono passare in prossimità del suolo affinché esse riescano a prendere in carico delle polveri da trasportare, per questo motivo le quote di inizializzazione delle traiettorie per i cluster sono inferiori rispetto a quelle per le stazioni, avremo delle quote di partenza di [5, 25, 50, 100, 200] metri, come già descritto. Tuttavia i limiti di altezza sono più ampi quando consideriamo l'arrivo nei pressi delle destinazioni di queste traiettorie, ossia nei pressi delle stazioni, questo perché anche delle traiettorie che arrivano più in quota possono precipitare al suolo a causa di piogge o altri fenomeni, per questo motivo quindi seppur le quote di partenza sono più basse, andremo ad aumentare la quota massima che consideriamo nella funzione logistica per lo score fino a **1000** metri di altezza, rimane invariata invece il valore di soglia per il raggio di riferimento, ossia a **100km** dalle stazioni. Nella tabella 3.1 degli esempi di punteggio assegnato a diverse combinazioni di distanza e altitudine nel caso dello score delle back-trajectories dalle stazioni.

Distanza (km)	Altitudine (m)	Punteggio
1	10	0.965
10	25	0.718
25	50	0.651
75	50	0.348
90	90	0.064

Tabella 3.1: Esempio per calcolo dello score di influenza al variare dell'input

Una volta calcolati questi score otterremo due matrici, la prima di dimensione **2x11x30** e la seconda di dimensione **2x30x11**, in base alla direzione dell'analisi. Aggregando opportunamente gli score sarà possibile ottenere un punteggio cumulativo dell'influenza dei cluster per ogni stazione e per ogni stagione nelle due direzioni

prese in considerazione. In figura 3.4 è possibile vedere un esempio di visualizzazione sulla mappa per gli score calcolati nel caso delle stazioni di **Epica-DML** e **TALDICE** nei mesi di Gennaio e Luglio.

Nello specifico a partire dalla stazione (in **blu**) l'influenza tra un determinato cluster (in **rosso**) e la stazione è visualizzato come una linea con uno spessore direttamente proporzionale allo score di influenza, come possiamo aspettarci non sempre i cluster più vicini sono quelli che hanno uno score maggiore, le informazioni ottenute dall'analisi delle traiettorie infatti ci permette di scoprire le relazioni tra clusters e stazioni anche se essi sono distanti, grazie all'azione delle correnti d'aria.

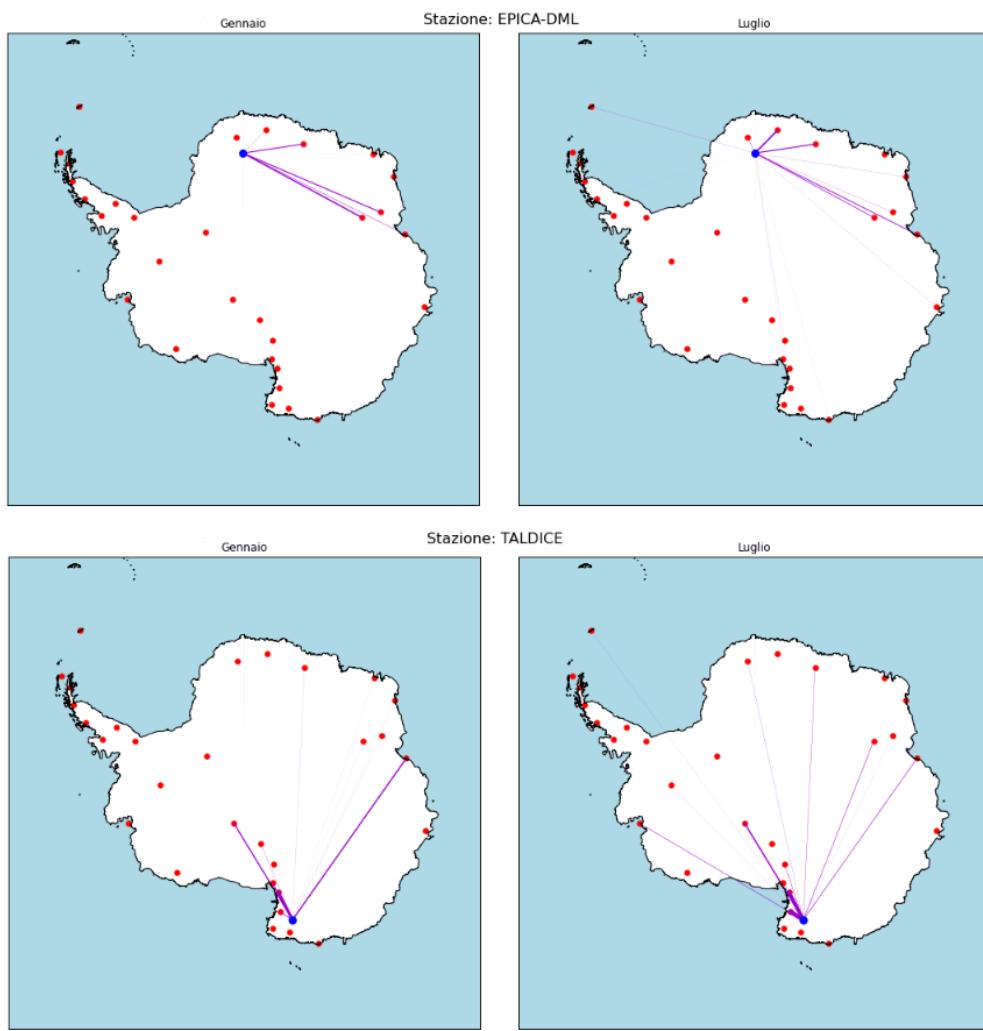


Figura 3.4: Influenza dei cluster sulle stazioni

Nel caso del sito **TALDICE** possiamo ad esempio notare come alcuni cluster riescono ad avere un'influenza non trascurabile anche se molto distanti dalla desti-

nazione, a differenza di altri clusters molto più vicini ma meno influenti a causa della direzione del vento che in quella regione tende quasi sempre a soffiare in direzione della costa e non dell'entroterra.

3.2.3 Clustering delle traiettorie

Come visto in precedenza (figura 3.2) un primo modo di aggregare le informazioni sulle traiettorie è stato quello di costruire delle **HeatMap** che potessero descrivere su quali zone, a livello geografico, le traiettorie tendono a concentrarsi, ottenendo una sorta di mappa di densità dei cammini in ingresso per ogni stazione. Seppur in prima battuta questo modo di calcolare un aggregato spaziale sulle traiettorie ci fornisce delle informazioni utili con un approccio computazionalmente semplice e leggero, un'analisi più approfondita è necessaria per evitare di perdere alcune relazioni tra le traiettorie che possono rivelarsi utili nella scoperta di pattern di circolazione del vento. L'approccio con **HeatMap** infatti rischia di oscurare i tragitti più comuni che molteplici traiettorie potrebbero percorrere, così facendo, dalla loro azione congiunta si verrebbero a creare uno o più 'corridoi' naturali che potrebbero incentivare maggiormente il trasporto delle polveri attraverso questi percorsi. Un approccio più esaustivo che non rischi di tralasciare informazioni significative sarebbe quindi quello di raggruppare le traiettorie che sono molto vicine tra loro e che mostrano un percorso più o meno condiviso, facendo così si andrebbero ad identificare dei gruppi di traiettorie principali in grado di meglio descrivere i vari percorsi in input ad una determinata stazione, ad esempio pesando ogni gruppo con il numero di traiettorie che lo compongono. Andremo quindi a definire dei veri e propri **clusters** di traiettorie dove, per ogni cluster, le differenze tra le traiettorie che lo compongono sia minima, massimizzando al contempo la differenza tra i clusters diversi. Nel processo di clustering la misura principale che ci permetterà di capire quanti cluster avere per ogni sciame di traiettorie è la **Spatial Variance**. Questa misura è calcolata tra ogni punto **k** costituente una traiettorie **j** e il suo cluster **i**.

$$SV_{i,j} = \sum_k (\mathbf{P}_{j,k} - \mathbf{M}_{i,k})^2 \quad (3.5)$$

dove la somma è calcolata sul numero totale di endpoints della traiettoria e \mathbf{P} e \mathbf{M} sono i vettori posizione media calcolata rispettivamente per la traiettoria ed il cluster. A partire dalle singole varianze spaziali delle traiettorie è possibile derivare due importanti quantità, in primis la **Cluster Spatial Variance (CSV)**, che altro non è se non la somma delle varianze spaziali di tutte le traiettorie di un determinato cluster, quindi:

$$CSV_i = \sum_j SV_{i,j} \quad (3.6)$$

ed infine la **Total Spatial Variance** che è la somma dei valori di (CSV) di ogni cluster.

$$TSV = \sum_i CSV_{j,k} \quad (3.7)$$

Il processo di clustering è di tipo gerarchico Bottom-Up, come descritto nella sezione 2.2.3, si inizierà quindi assegnando ogni traiettoria al proprio cluster così che ci siano i clusters con $j = 1$ traiettorie per ogni cluster. Ad ogni iterazione il numero di clusters viene ridotto di 1 a seguito della fusione di due cluster, questo processo continua finché tutti i clusters saranno stati fusi tra loro. Ad ogni iterazione viene calcolata la **TSV** per ogni possibile scelta di fusione di due cluster in uno, avremo quindi un totale di $(i^2 - 1)/2$ possibili valori di TSV, uno per ogni possibile coppia di clusters, il criterio per determinare quale fusione verrà effettuata ad un determinato step è quello di scegliere la combinazione di clusters che incrementa in maniera minore possibile il valore di varianza spaziale totale.

Nello specifico quello che viene calcolato è la variazione percentuale del valore di varianza totale man mano che il processo di fusione dei clusters procede, il modo migliore di decidere dove andare a tagliare il clustering gerarchico e quindi determinare il numero di clusters finale che vogliamo avere è quello di utilizzare questo grafico sull'evoluzione della **TSV** scegliendo il minor numero di cluster possibili prima che

la total spatial variance aumenti in valore percentuale più di una determinata soglia, tipicamente il 20-30%, o comunque prima di aumenti significativi. in Figura 3.5 un esempio dell'evoluzione del valore di **TSV** durante un clustering tipo.

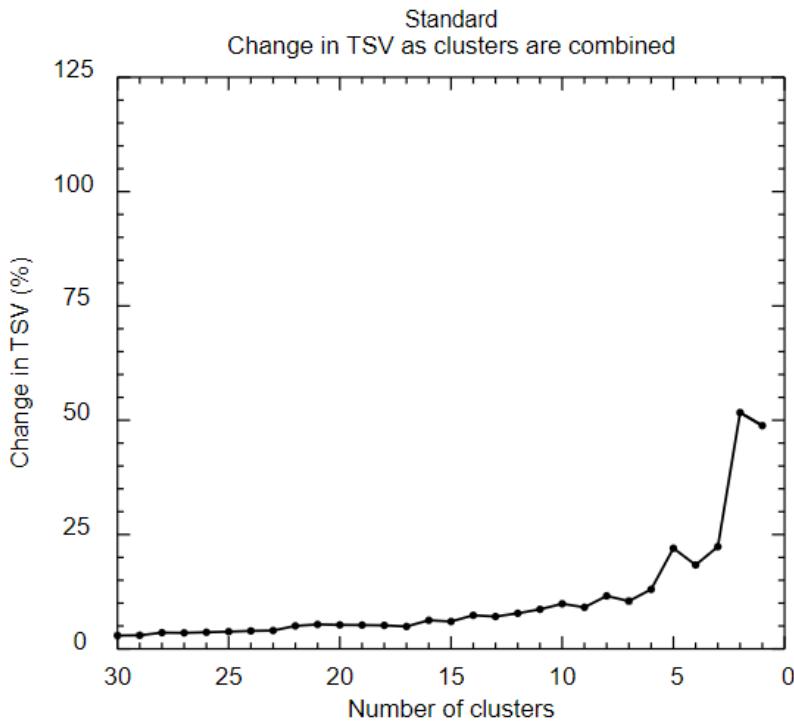


Figura 3.5: Variazione TSV

Quando la variazione percentuale è particolarmente alta, infatti, ci troviamo di fronte al caso in cui stiamo unendo due cluster poco simili, ossia che contengono traiettorie significativamente diverse tra loro, e a causa di ciò la varianza del nuovo cluster appena creato aumenta drasticamente, e di conseguenza anche quella totale. Fermandoci appena prima di questi picchi saremo sicuri di scegliere il minor numero di cluster possibili, per riassumere al meglio le principali componenti che descrivono il budget di correnti in ingresso ad una particolare stazione, senza tuttavia rischiare di unire tra loro clusters fondamentalmente differenti.

Nella pratica il calcolo di questi Cluster è stato effettuato utilizzando **HYSPLIT**, che fornisce un modulo specifico per il Clustering delle traiettorie, HYSPLIT richiede tuttavia che tutti i file delle traiettorie che vogliamo clusterizzare vengano posizionati all'interno di una specifica cartella e richiede inoltre la produzione di un file di listing di tutte le traiettorie chiamato **INFILE**, per sveltire questo processo dato il numero

di cluster da creare sono state scritte delle funzioni apposite in Python. Fatto ciò si procede con il calcolo del clustering ideale tramite il metodo appena descritto, e una volta prodotto l'output desiderato, sempre tramite **PYSPLIT** è possibile leggere i file direttamente dalla cartella di output di HYSPLIT e salvare tutte le informazioni in un file **pickle** utilizzabile in un secondo momento per visualizzare i cluster e/o operare con essi in future analisi, in figura 3.6 vediamo un esempio del clustering prodotto per la stazione **TAYLOR-DOME** nelle due stagioni di riferimento.

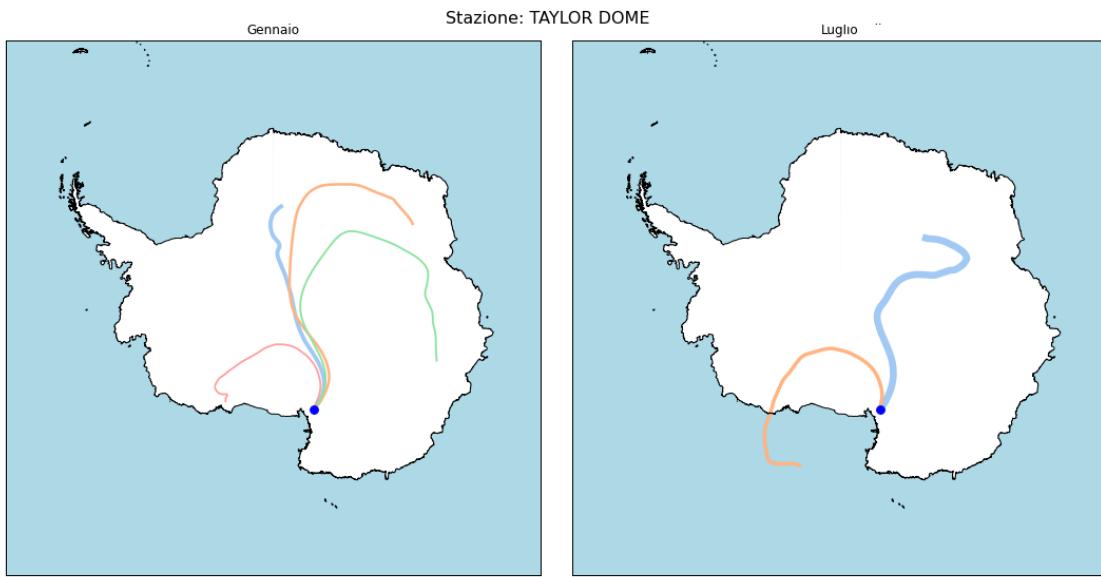


Figura 3.6: Clustering delle traiettorie

Costruzione di un modello statistico

Nel capitolo precedente abbiamo visto come grazie all'utilizzo di HYSPLIT e PY-SPLIT è stato possibile costruire un sistema semi-automatizzato per la creazione di una famiglia di traiettorie utili per stimare il potenziale trasporto nelle diverse zone del continente Antartico. Ottenuti i preziosi dati sulle traiettorie è stato formalizzato un semplice ma efficiente sistema per valutare come le diverse variabili in gioco possano aver contribuito al potenziale trasporto delle polveri dalla fonte alla destinazione. In questo capitolo metteremo insieme quanto appreso e raccolto negli step precedenti, utilizzando le informazioni sull'influenza dei diversi cluster nei confronti delle stazioni di carotaggio vogliamo adesso quantificare quanto le variabili d'interesse siano statisticamente correlate con gli score da noi calcolati. Oltre questa analisi sarà poi possibile utilizzare i molteplici dati a nostra disposizione per creare delle visualizzazioni che possano gettare luce sui complessi fenomeni di circolazione dei venti in relazione alla stagionalità e alla posizione geografica presa in considerazione, così da migliorare la nostra comprensione del fenomeno e guidare la ricerca verso nuove domande e per identificare dove poter agire per migliorare la precisione e la significatività delle nostre conclusioni.

4.0.1 Descrizione dei dati e analisi preliminare

Nel descrivere il budget di polveri presente in ognuna delle stazioni in nostro possesso bisogna analizzare con precisione i dati presenti in letteratura, in [8] e [9] vengono infatti descritti diversi scenari, nella tabella 2.1 vista all'inizio dello step 1 infatti

Site	Conc. [ppb] (d<5 μm)	Conc. [ppb] (5<=d<10 μm)	Flux [mg*m ² /yr] (d<5 μm)	Flux [mg*m ² /yr] (5<=d<10 μm)	Coarse/ tot fraction [flux]
TALDICE	8	4	0,75	0,35	0,32
DC	8	1	0,2	0,03	0,13
MdPt	14	4	0,5	0,14	0,22
D4	9	3	0,19	0,05	0,21
Vostok-BH7	17,8	0	0,36	0	0
EDML			1,64	0,17	0,09

Tabella 4.1: Dati di concentrazione e flusso delle polveri suddivise per diametro delle particelle

vengono indicati i dati relativi alla concentrazione di polveri trovata nei diversi siti di carotaggio misurata in **ppb** (Part Per Bilion) e i dati relativi al flusso di queste polveri, misurata in $\frac{\text{mg}}{\text{m}^2*\text{yr}}$, stiamo quindi misurando quanti milligrammi di materiale si depositano per unità di superficie ogni anno. Queste variabili si riferiscono a quantità nel totale delle polveri, tuttavia la scala dimensionale delle polveri stessa può rivelarci molto sulla provenienza delle stesse. In generale le polveri con grana più fine, quindi con un diametro che si attesta al di sotto dei **5 μm** provengono da luoghi più remoti, in quanto più leggere infatti queste particelle hanno più potenziale di trasporto, ed è più probabile trovarle in alta quota, e di conseguenza queste polveri potrebbero provenire anche dai continenti vicini l’Antartide come il Sud-Africa, l’Australia o il Sud-America. Al contrario le particelle di polvere con diametro maggiore, ossia compreso in un range tra i **5 μm** e i **10 μm** hanno più probabilità di provenire da fonti locali in quanto, essendo più pesanti hanno un potenziale di trasporto inferiore, si può quindi affermare che dai nostri dati ci aspettiamo una correlazione maggiore con le polveri di maggior diametro. I dati sulla distribuzione di concentrazione e flusso al variare del diametro delle particelle sono tuttavia di difficile reperimento, delle stazioni in nostro possesso quindi è stato possibile reperire i dati sulla distribuzione del diametro delle polveri sono in un subset di 6 stazioni, in questo senso oltre che in (Delmonte et al., 2013[8] e Delmonte et. al., 2020[9]) altri preziosi dati sono reperibili su (Wegner et. al., 2015 [13] e Aarons et. al., 2019[14]), in tabella 4.1 le osservazioni disponibili.

Per dare inizio all'analisi vengono calcolate delle medie per capire quanto una stazione sia generalmente influenzata dai cluster, per ottenere questo score aggregato andiamo a sommare i contributi dei singoli cluster verso ogni stazione per ognuna delle due stagioni a nostra disposizione, andando poi a fare una media tra le due stagioni, più formalmente:

$$S_i = \frac{1}{2} \sum_j I_{g(i,j)} + I_{l(i,j)} \quad (4.1)$$

dove con $I_{g(i,j)}$ indichiamo lo score di Gennaio del cluster j verso la stazione i , e con $I_{l(i,j)}$ quelli di Luglio. Lo stesso ragionamento è applicabile con gli score delle forward-trajectories a partire dai clusters, in tabella 4.2 gli score ottenuti per le 6 stazioni di cui abbiamo anche i dati sul flusso:

Stazione	VOSTOK-BH7	EPICA-DML	TALDICE	DC ITASE	D4 ITASE	MDP-A ITASE
Influenza back-traj dalle stazioni	250.47	345.96	635.27	293.11	275.45	351.88
Influenza forward-traj dai clusters	14.58	25.03	486.85	19.77	13.25	19.70

Tabella 4.2: Score di influenza dei clusters per le stazioni oggetto di analisi

Ottenuti questi valori possiamo calcolare un indice di correlazione tra le variabili, il più famoso è indubbiamente l'indice di Pearson, questo indice permette infatti di capire se esiste una relazione di linearità tra due variabili statistiche, tuttavia per poter utilizzare l'indice di Pearson bisogna prima assicurarsi che entrambi i set di dati provengano da una distribuzione normale, questa condizione è infatti richiesta prima di procedere. Per testare la normalità di un set di dati esistono diversi test, i più famosi sono quelli di **Shapiro-Wilk**, il test del **Chi quadro**, il test di **Kolmogorov-Smirnov**, o di **Lilliefors**. Di seguito i risultati dei vari test di normalità per i due set di dati in considerazione, teniamo comunque a mente che in letteratura viene solitamente indicato il test di **Shapiro-Wilk** come il più affidabile quando si lavora con set di dati a bassa numerosità, come nel nostro caso, al contrario di test come **Kolmogorov-Smirnov** più adatto a set ad alta numerosità.

	Shapiro-Wilk	Chisquare	Kolmogorov Smirnov	Lilliefors
Influenze dei clusters(forward)	0	0	0	0
Influenze dei clusters(backward)	0.02	0	0	0.02
Flux [mg*m2/yr] (5<=d<10 µm)	0.33	0.98	0.07	0.53
Flux [mg*m2/yr] (d<5 µm)	0.07	0.78	0.02	0.34

Tabella 4.3: p-values dei test di normalità per le variabili in esame

L'unica variabile che viene accettata da tutti i test come normale è il flusso delle particelle di medio-grosso diametro, per il flusso di polveri particolarmente sottili l'unico test a rifiutare la normalità è Kolmogorov Smirnov, ma sappiamo essere anche il test meno adatto ai campioni con bassa numerosità, quindi potremmo accettare la normalità di questa variabile, seppur con qualche riserva, ma in ogni caso tutti i test rifiutano la normalità degli score di influenza delle stazioni, anche intuitivamente infatti il flusso dei venti favorirà determinate posizioni con influenze molto grandi, come per la stazione **TALDICE** e molti altri punti e/o stazioni potrebbero invece avere score decisamente più bassi, non ci stupisce quindi che questa variabile non si manifesti similmente ad una distribuzione normale. Dal risultato di questo test concludiamo che non è possibile utilizzare il coefficiente di Pearson perché non tutte le variabili sono normali, a questo punto non ci rimane che utilizzare un'altro test, ossia quello di **Spearman**, in questo test invece di cercare la relazione lineare tra due variabili si cerca semplicemente di dimostrare che una variabile è monotona rispetto all'altra, ossia che quando una aumenta, aumenta anche l'altra, seppur la tipologia di relazione può essere di qualsiasi tipo, anche non lineare, anche in questo caso il range del test è $[0, 1]$ e la sua interpretazione è uguale a quanto detto per il coefficiente di Pearson. Il vantaggio della correlazione di Spearman è che essa non presuppone la normalità dei campioni che diamo in input, possiamo quindi utilizzarla anche nel nostro caso particolare, unitamente al coefficiente di relazione di Spearman viene calcolato anche un **p-value** relativo ad un test di ipotesi in cui l'ipotesi nulla è che le due variabili siano non correlate.

	Correlazione di Spearman	p-value
Influence(backward) VS Flux ($5 \leq d \leq 10\mu m$)	0.886	0.018
Influence(forward) VS Flux ($5 \leq d \leq 10\mu m$)	0.771	0.072
Influence(backward) VS Flux ($d < 5\mu m$)	0.657	0.156
Influence(forward) VS Flux ($d < 5\mu m$)	0.714	0.110

Tabella 4.4: Correlazione di Spearman tra variabili dei flussi e influenza tra clusters

Come possiamo vedere nella tabella 4.4 l'influenza dei cluster è correlata con entrambe le tipologie di flusso di polveri, tuttavia seppur la differenza di correlazione tra le polveri sottili e non sembra non essere molta, i p-values ottenuti dai test di ipotesi ci dicono che le correlazioni ottenute con le polveri sottili ([0.714, 0.657]) non è abbastanza per ritenere le due variabili statisticamente correlate, si cade infatti nella regione di accettazione del test con dei p-value di [0.156, 0.110], rispettivamente per le influence di tipo backward e forward. Per la controparte di polveri grossolane si ottengono sia valori di correlazione maggiori ([0.886, 0.771]) e soprattutto p-values di [0.018, 0.072] che cade nella regione di rifiuto data la soglia classica di significatività $\alpha = 0.05$ per le traiettorie backward e al limite della regione per le traiettorie forward, suggerendoci una significanza statistica tra il flusso e le influenze calcolate nella sezione precedente. Supportati da questo risultato ha senso continuare l'analisi cercando di costruire un modello di regressione per verificare quanto l'influenza dei cluster riesce a spiegare la varianza della variabile target che descrive il flusso delle particelle di media dimensione.

4.0.2 Definizione di un modello OLS

Dopo esserci assicurati di avere una variabile Target distribuita secondo una normale possiamo procedere nella costruzione di un modello statistico che cerchi di stimare il flusso di deposito delle stazioni tramite le variabili a nostra disposizione. Data la scarsità delle osservazioni a nostra disposizione e la necessità di avere la massima chiarezza sulle variabili in gioco il modello più semplice è anche quello più adatto, per utilizzare tecniche di Machine Learning infatti sarebbero necessari molti più dati, ed

in generale dovremmo rinunciare al potere esplicativo dei metodi statistici più classici per questo motivo il modello prescelto per il nostro scopo è quello di regressione **OLS** (Ordinary Least Squares). Nel modello **OLS** lo scopo è quello di trovare una relazione matematica tra le variabili in input che riesca a minimizzare gli errori di previsione rispetto alla variabile target, in particolare si sceglie di minimizzare la somma dei quadrati delle distanze tra i dati reali che descrivono il flusso e quelli prodotti dalla relazione matematica formulata.

Più nello specifico sia x_i la serie dei dati in input al modello vogliamo trovare una funzione f tale che la funzione approssimi i dati reali y_i al meglio possibile, per costruire questa funzione come già detto minimizziamo la somma dei quadrati degli errori definita formalmente come:

$$S = \sum_{i=1}^n (y_i - f(x_i))^2 \quad (4.2)$$

Solitamente la funzione f assume una forma parametrica dove date n variabili in input si cerca di stimare dei parametri β per modellare le n variabili in maniera opportuna, stimeremo quindi l'i-esimo valore target y_i come:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + \varepsilon_i \quad (4.3)$$

o, utilizzando la forma vettoriale:

$$y_i = x_i^T \beta + \varepsilon_i \quad (4.4)$$

Per utilizzare un modello **OLS** è inoltre imperativo rispettare determinate assunzioni su cui il modello si basa, se queste ipotesi non venissero rispettate infatti il modello ottenuto sarebbe infatti scorretto e inaffidabile nelle previsioni. Le ipotesi da dover rispettare per il corretto funzionamento del modello **OLS** sono le seguenti:

- **Incorrelazione tra errori e variabili esplicative:** il valore atteso $E(\varepsilon_i|x_i) = 0$ e quindi anche la loro covarianza dev'essere nulla: $Cov(\varepsilon_i, x_i) = 0$ nel caso ideale quindi degli shock sull'errore ε_i generano cambiamenti nella variabile dipendente y_i ma non nelle variabili esplicative x_i , permettendoci di modellare i coefficienti β solo in funzione di ciò che riusciamo a spiegare e non in relazione a cambiamenti casuali, assicurandoci così che i coefficienti $\beta_1, \beta_2 \dots \beta_n$ non siano distorti.
- **Nessuna dipendenza lineare tra i regressori:** Tutti i regressori X devono essere linearmente indipendenti, nel caso di multicollinearità tra le variabili esplicative i coefficienti non saranno ottimizzati e anche se la previsione su nuove osservazioni rimane possibile il risultato potrebbe essere non affidabile.
- **Errori sferici:** Gli errori devono avere la stessa varianza σ^2 in ogni osservazione, quando ciò viene violato si parla di **eteroschedasticità** degli errori e implica che l'errore non è casuale, oltre la condizione di **omoschedasticità** gli errori devono anche essere incorrelati tra di loro, quindi $E[\varepsilon_i \varepsilon_j | X] = 0$ per ogni $i \neq j$
- Normalità degli errori: gli errori si distribuiscono secondo una normale condizionata ai regressori, quindi $\varepsilon | X \sim \mathcal{N}(0, \sigma^2 I_n)$

Vista la maggior correlazione con la variabile target degli score provenienti dalle traiettorie di tipo backward le analisi che seguono verranno riportate solo per questa tipologia di traiettoria, riportando solo alla fine i risultati ottenuti anche con gli altri score di tipo forward. Per rispettare la condizione di multicollinearità analizziamo la correlazione lineare tra le possibili variabili in input per stimare la variabile target, le variabili collezionate finora sono:

- **distance_from_coastline(km):** Quanto la stazione è distante dalla costa in km
- **distance_from_nearest_cluster(km):** Quanto la stazione è distante dal cluster più vicino tra quelli identificati nella prima fase

- **nearest_cropout(km)**: Quanto la stazione è distante dal primo cropout disponibile dai dati in input ai cluster.
- **cropout_area_within_500km**: Area totale dei cropout in un raggio di 500 km
- **cropout_area_within_1000km**: Area totale dei cropout in un raggio di 1000 km
- **jans_influences**: Influenza dei cluster nel mese di Gennaio
- **juls_influences**: Influenza dei cluster nel mese di Luglio
- **influenced_by_clusters**: Media delle due influenze stagionali

Di seguito, in figura 4.1 la matrice di correlazione tra tutte le variabili:

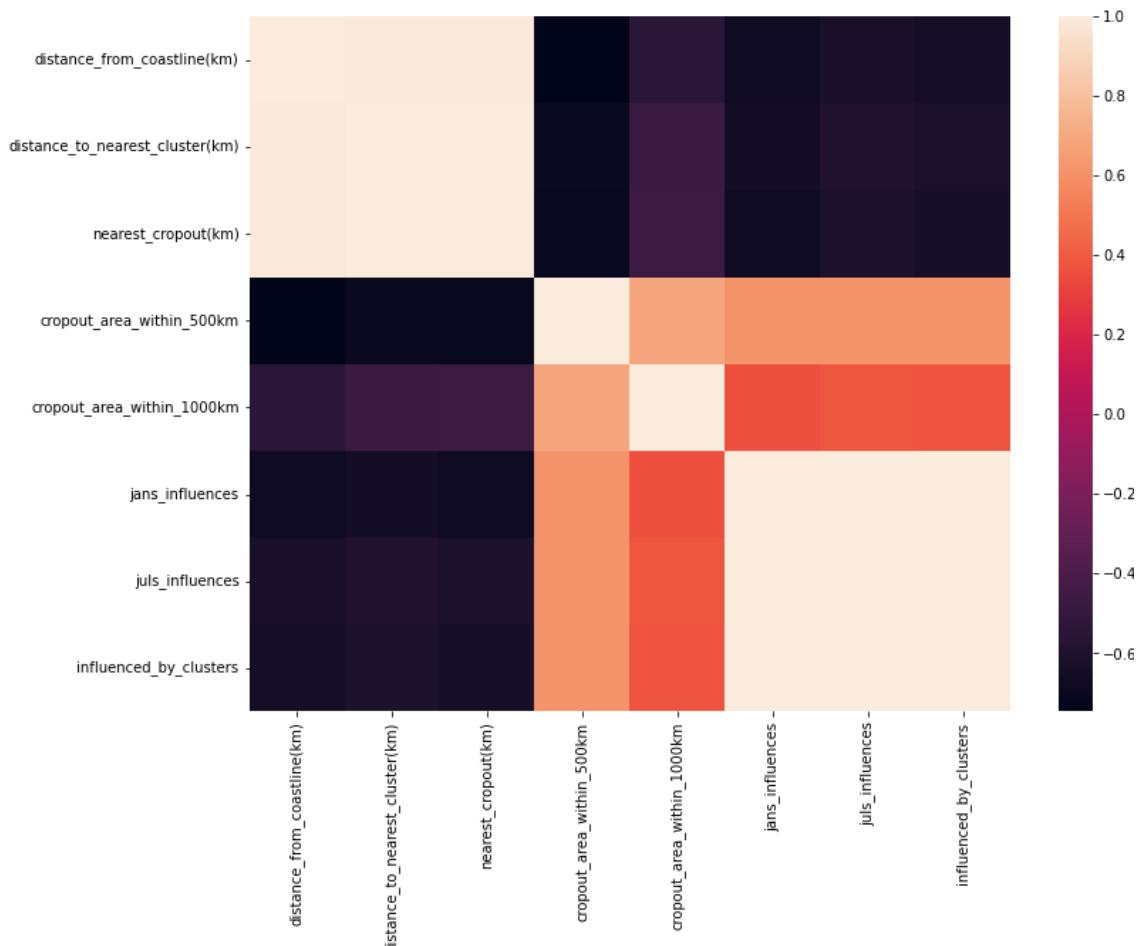


Figura 4.1: Correlazione tra tutte le variabili esplicative disponibili

Le tre variabili sulle distanze dalla costa, dai cluster e dal cropout sono fortemente correlate, così come le influenze stagionali con l'influenza media, i dati sulla quantità di cropout nei raggi di 500 e 1000 km sono altamente correlate seppur non multicollineari, come variabili esplicative si è quindi deciso di prendere una per ogni gruppo, identificando le tre variabili **distance_from_cluster(km)**, **cropout_area_within_1000km** e **influenced_by_clusters** come le tre principali variabili esplicative distillate dalle analisi precedenti, potenzialmente in grado di spiegare la variabile target di nostro interesse. In tabella 4.5 i dati in possesso per le 6 stazioni presenti nel nostro dataset.

Stazione	distance_to_nearest_cluster(km)	cropout_area_within_1000km	influenced_by_clusters
VOSTOK-BH7	1035.97	698.30	250.47
EPICA-DML	302.70	2540.52	345.96
TALDICE	193.16	9906.88	635.27
DC ITASE	960.60	4031.86	293.11
D4 ITASE	723.78	10203.59	275.45
MDP-A ITASE	471.63	11649.42	351.88

Tabella 4.5: Variabili esplicative delle stazioni disponibili

Una volta isolate le potenziali variabili esplicative si procede con una veloce analisi esplorativa costruendo degli scatterplot che confrontino i regressori con la variabile dipendente: **Flux[5-10µm]** che descrive il flusso totale di particelle di media dimensione, ossia la frazione che supponiamo essere interamente controllata dalle arre deglaciate. Ciò viene fatto per analizzare visivamente potenziali correlazioni tra le coppie di variabili alla ricerca di elementi che possano favorire la costruzione del modello.

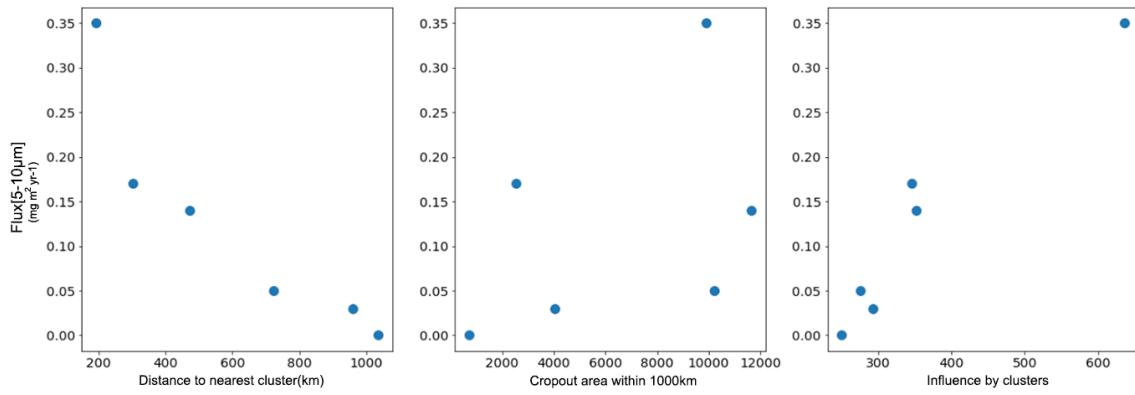


Figura 4.2: Scatterplots tra variabili esplicative e variabile target

Nel primo grafico di figura 4.2 possiamo vedere che la distanza sembra avere una chiara correlazione negativa con la variabile target, intuitivamente quanto più una stazione si trova in prossimità dei clusters quanto più essa ha probabilità di avere una alta concentrazione di particelle grossolane nelle vicinanze, la relazione non sembra comunque di tipo lineare. Nel grafico al centro possiamo vedere lo scatterplot tra la quantità di cropout nel raggio di 1000 km e il flusso[5-10 μ m], in questo caso non sembra esserci nessuna chiara dipendenza tra le due variabili, vediamo infatti i punti distribuirsi senza un preciso ordine formando una nube al centro del grafico che lascia presagire una scarsa significatività con valori di flusso simili che presentano valori di area molto differenti tra loro. Nel terzo ed ultimo scatterplot vediamo l'influenza calcolata dalle retro-traiettorie contro il flusso, in questo caso il valore estremo della stazione **TALDICE** tende ad ammassare gli altri punti sulla sinistra del grafico, seppur teoricamente si tratti di un Outlier la stazione TALDICE in realtà presenta una situazione diversa dalle altre in quanto si trova nell'immediata prossimità di molti clusters, proprio al centro di una regione ad alta densità di outcrop, per questo motivo è impensabile rimuovere questa osservazione, potrebbe essere utile considerare delle trasformazioni che rendano il range di valori più uniformi così da poter effettivamente valutare la relazione dell'influenza con il flusso.

Alla luce di quanto visto dagli scatterplot la variabile che descrive l'area di cropout presente nel raggio delle stazioni viene scartata e l'analisi si concentrerà sulle altre due variabili che descrivono la distanza dai clusters e l'influenza che essi riesco-

no ad avere sulle stazioni, inizieremo utilizzando soltanto l'influenza, per poi valutare le eventuali trasformazioni sui dati in input e l'integrazione di entrambe le variabili in un modello con entrambe le variabili esplicative.

4.0.3 Risultati modello OLS

Come anticipato la prima variabile che vogliamo studiare è quella relativa all'influenza dei cluster, per prima cosa quindi definiamo un modello di regressione lineare OLS con una costante e un regressore per modulare la variabile esplicativa. In tabella 4.6 Vediamo il risultato di questo primo modello, notiamo subito un valore di R-quadro abbastanza elevato di 0.90 circa, segno che la variabile esplicativa scelta è in grado di spiegare buona parte della varianza della variabile target descrivente il Flusso[5-10 μ m], osservando poi i risultati dei t-test sulle variabili possiamo notare che il p-value associato alla cluster influence (0.002) è inferiore alla soglia $\alpha = 0.01$, ciò ci indica che la variabile esplicativa è statisticamente significativa nello spiegare la variabile target, meno valido invece il p-value dell'intercetta a 0.016, in questo non è chiarissimo se l'intercetta serva effettivamente a migliorare le previsioni o no, inoltre i test sui residui ne accettano la normalità.

	coeff	std_err	t	$P > t $		
const	-0.1918	0.048	-4.035	0.016	R2	0.925
cluster_influence	0.0009	0.000	7.045	0.002	R2-adj	0.907

Tabella 4.6: risultati modello OLS di partenza

Come suggerito durante la discussione degli scatterplot possiamo pensare di visualizzare delle potenziali trasformazioni dei dati in input per vedere se la distribuzione dei punti in questo nuovo spazio sembra suggerirci delle correlazioni migliori così da aumentare il fitting del modello.

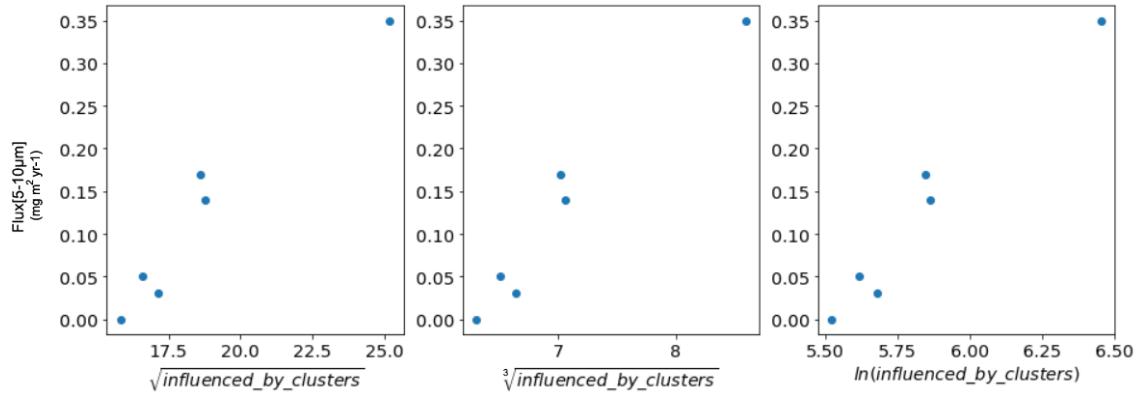


Figura 4.3: Possibili trasformazioni della variabile esplicativa

Nella figura 4.3 possiamo vedere lo scatterplot dell'influenza dei cluster contro il flusso[5-10μm] nelle tre varianti in radice quadrata, cubica e logaritmo naturale della variabile d'origine. Applicare le radici migliora la distribuzione sulle x della variabile esplicativa così come il logaritmo, le trasformazioni di questo tipo tendono infatti a linearizzare la relazione tra i dati in input e il flusso in output, favorendo la convergenza del modello verso valori di R^2 più alti.

In tabella 4.7 i risultati dei modelli OLS dove si considerano le diverse versioni trasformate dell'influenza dei clusters (CI) come regressore, la versione che meglio riesce a spiegare la varianza del flusso sembra essere il logaritmo più intercetta, in questo caso infatti arriviamo ad un valore di R-quadro intorno al 96%, i residui sono normali e nel nuovo grafico di regressione in figura 4.4 nel grafico di destra vediamo come con la radice cubica l'osservazione apparentemente estrema di TAL-DICE in realtà giace quasi perfettamente in linea con la retta tracciata dalle altre 5 osservazioni.

Model	R2	R2-adjusted
$Flux[5 - 10\mu m] \sim \sqrt{CI} + c$	0.946	0.933
$Flux[5 - 10\mu m] \sim \sqrt[3]{CI} + c$	0.952	0.940
$Flux[5 - 10\mu m] \sim \ln(CI) + c$	0.962	0.952

Tabella 4.7: Risultati dei diversi modelli OLS in termini di R^2

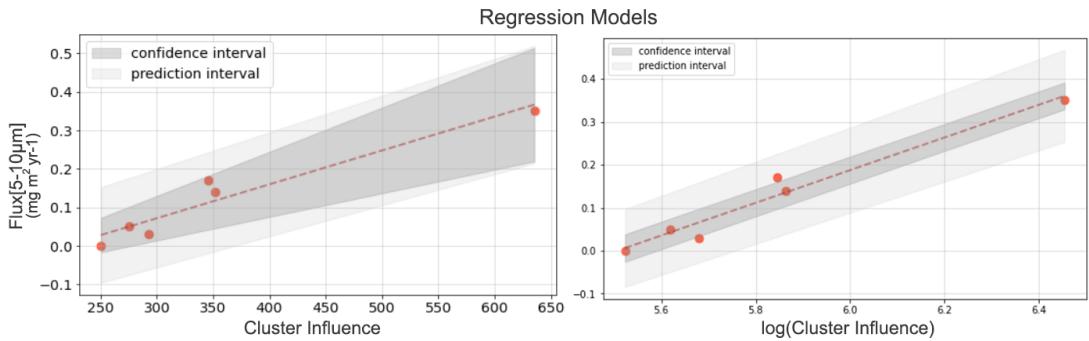


Figura 4.4: Linea di regressione stimata dal modello OLS per la l'influenza non trasformata, e con il logaritmo naturale

Arrivati a questo punto abbiamo già ottenuto un ottimo modello sulla base delle osservazioni a disposizione, per cercare di migliorare ulteriormente la bontà del modello possiamo considerare l'introduzione dell'altra variabile esplicativa sulla distanza delle stazioni dai cluster, questa informazioni infatti unitamente al potenziale di trasporto potrebbe darci un ulteriore insight sul trasporto delle polveri. Andiamo quindi a definire un ulteriore modello OLS dove consideriamo entrambe le variabili più una costante di intercetta, in tabella 4.8 i risultati ottenuti per quest'ultimo modello e di seguito in figura 4.5 il modello finale identificato dove il colore identifica la quantità di flusso[5-10 μ m] per ogni punto del grafico, anche la dimensione dei punti è direttamente proporzionale al flusso.

	coeff	std_err	t	P > t 	R2	
const	-0.6595	0.152	-4.285	0.023	R2	0.999
ln(cluster_influence)	0.2246	0.017	12.939	0.001	R2-adj	0.998
ln(cluster_distance)	-0.0858	0.009	-9.969	0.002		

Tabella 4.8: Risultati modello finale con entrambe le variabili esplicative

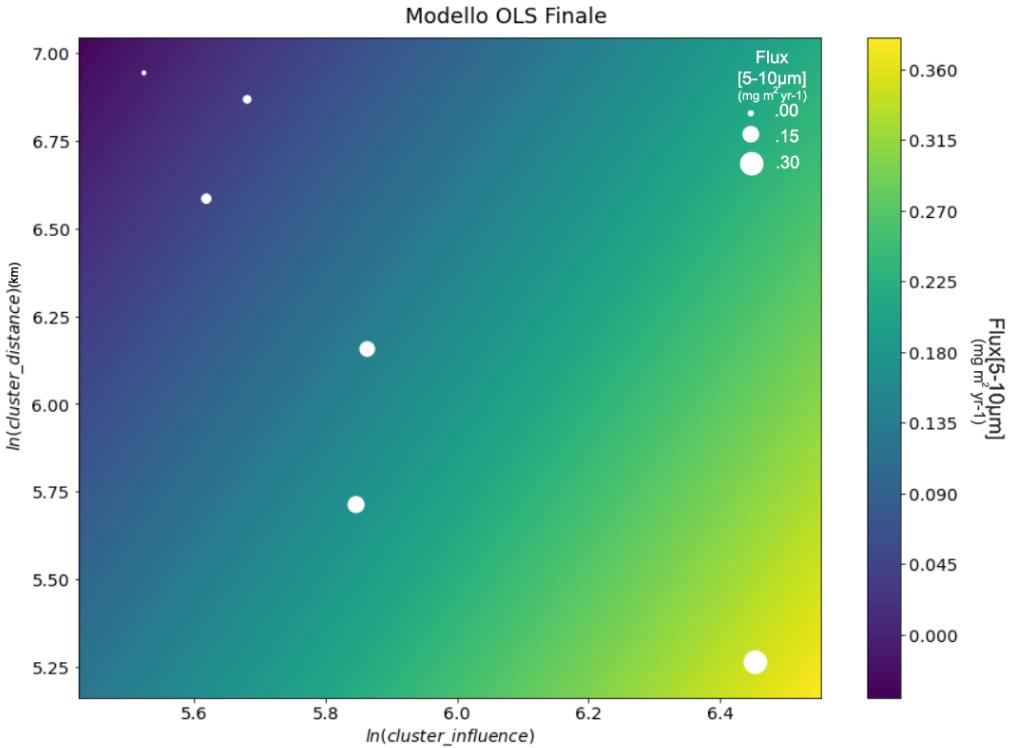


Figura 4.5: Valore di flusso[5-10 μm] nel modello finale al variare di distanza e influenza dai clusters

Il modello ottenuto sembra quindi in grado di spiegare quasi perfettamente il Flusso in input nelle stazioni, almeno nelle 6 osservazioni a nostra disposizione e per la frazione di interesse tra i 5 e i 10 micrometri, utilizzando sia le informazioni sulla distanza dai cluster sia quella sulla loro influenza. Intuitivamente possiamo pensare che seppur l'influenza riesca a dirci molto bene quali stazioni possano avere delle polveri in arrivo dai cluster, essa non tiene in considerazione la lunghezza del tragitto, delle traiettorie provenienti da cluster lontani hanno meno potenziale di trasporto delle particelle rispetto ai tragitti più brevi, seppur in minima parte. La relazione matematica ottenuta dal modello che lega le tre variabili tra loro è la seguente:

$$y = -0.0858 * x_0 + 0.02246 * x_1 - 0.6595 \quad (4.5)$$

dove $x_0 = \ln(Cluster_Distance)$ e $x_1 = \ln(Cluster_Influence)$, utilizzando questa legge sarebbe possibile costruire una mappa tematica dell'Antartide per ricavare la

quantità di flusso in input che ci aspettiamo in base alle variabili qui identificate. Nella pratica calcolare una simile mappa è computazionalmente troppo oneroso, per fare ciò infatti sarebbe necessario calcolare le retro-traiettorie e quindi gli score di influenza per tutti i punti della griglia che andremmo ad utilizzare. Per fornire una previsione del flusso atteso che sia il più precisa possibile pur rimanendo in costi computazionali affrontabili è stato costruito un modello alternativo utilizzando soltanto la distanza dai cluster come variabile esplicativa. Calcolare la distanza minima tra un dato punto e i centroidi dei 30 clusters è un'operazione calcolabile in frazioni di secondo, rendendo fattibile il computo di questo valore per una grande quantità di punti. Il modello alternativo basato solo sulla distanza raggiunge un valore di R^2 del 93.6%, in figura 4.6 il valore di flusso atteso dal modello calcolato per 1'000'000 di punti con una risoluzione di 1000x1000, ottenendo uno step di 5km circa tra le singole previsioni, e interpolando i valori intermedi tra una previsione e l'altra.

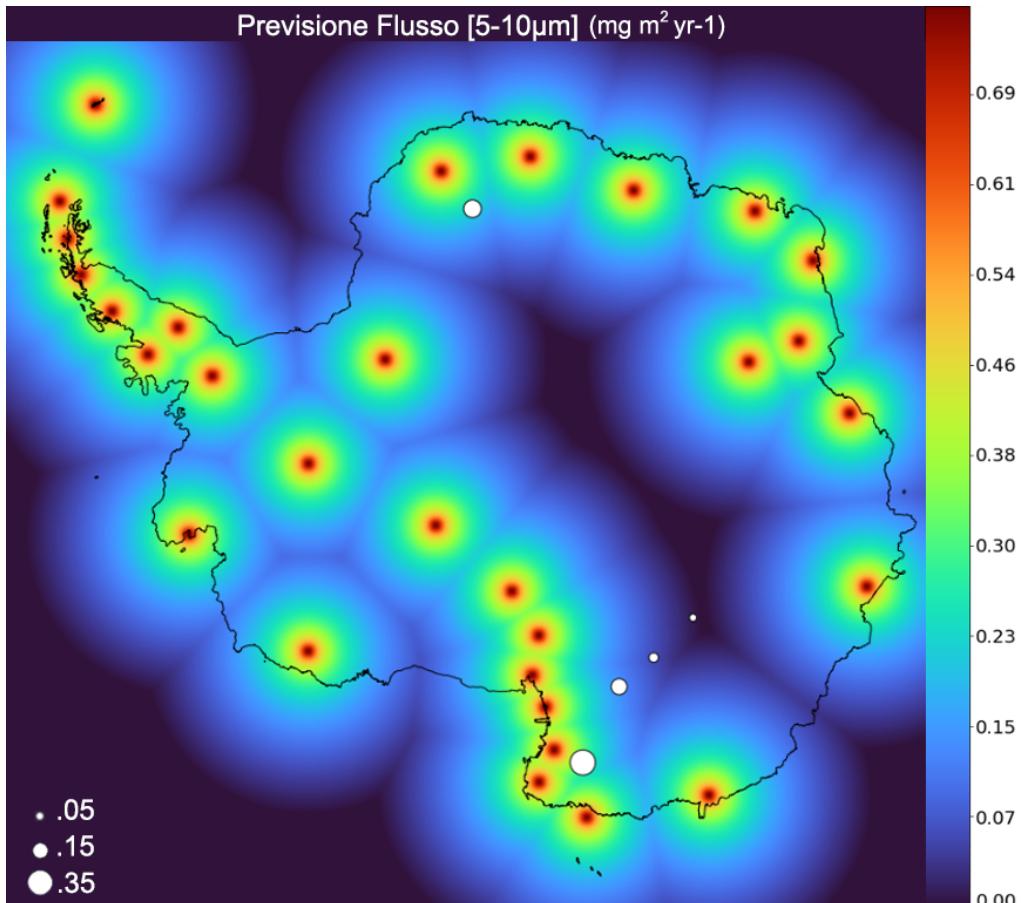


Figura 4.6: Previsione Flusso[5-10μm]

Per mitigare gli effetti di previsioni estremamente alte e potenzialmente fuori scala, come nei casi limite di punti in prossimità dei centroidi dei clusters ($<10/20\text{km}$) è stato applicato un leggero filtro gaussiano all'immagine. Così facendo le vaste aree dove il gradiente del valore di flusso atteso cambia molto lentamente non subiscono alcuna modifica, mentre i picchi in prossimità di punti con condizioni estreme vengono mediati con i valori nelle immediate prossimità. In questo modo ci assicuriamo di restituire un range di previsioni più uniformi.

Oltre i valori previsti per il flusso vengono riportati sulla mappa le osservazioni reali per i 6 siti di raccolta delle carote di ghiaccio utilizzati per la costruzione del modello.

Per gli altri 5 siti di cui non si hanno a disposizione dati sul flusso[$5-10\mu\text{m}$] in input vengono calcolati dei valori attesi utilizzando il modello completo, ossia calcolando la previsione sia con la distanza che con l'influenza dai clusters. Per questi siti infatti esistono i dati di influenza calcolati nel capitolo precedente. In tabella 4.9 le previsioni effettuate.

DOME-B	EPICA-DC	SOLARICE-DC	DOME-FUJI	TAYLOR-DOME
0.002	0.031	0.027	0.125	0.307

Tabella 4.9: Previsioni del valore di Flux[$5-10\mu\text{m}$] per i siti di perforazione rimasti fuori dall'analisi per mancanza di dati

Conclusioni e possibili sviluppi

L'Antartide è una delle fonti di informazioni più importanti per le ricostruzioni paleoclimatiche, grazie ad analisi chimiche e fisiche delle carote di ghiaccio è possibile infatti risalire a preziose informazioni sulla storia del pianeta e di come il clima si sia evoluto nel corso del tempo. Il lavoro di tesi svolto aveva l'obiettivo di analizzare i dati esistenti e creare di nuovi per comprendere come il potenziale spostamento di polveri, provenienti dalle aree deglaciate del continente Antartico, sia in grado di influenzare il flusso deposizionale di polveri (e quindi il record paleoclimatico) nei pressi di siti di perforazione.

Per fare ciò è stato necessario raccogliere i dati sulle aree deglaciate delineando 30 clusters significativi, ciò è stato fatto tramite l'utilizzo di algoritmi dell'algoritmo k-means++, come visto nel capitolo 2. In questo modo sono state individuate quali aree a livello geografico possono rappresentare delle possibili sorgenti di polveri con elevato potenziale di trasporto. Successivamente sono stati analizzati i dati meteorologici del progetto **AMRC** per arricchire la nostra conoscenza sui clusters e sui siti di perforazione con delle variabili descrittive a livello climatico, gli stessi dati sono quindi stati utilizzati per identificare le finestre temporali più adatte per concentrare le nostre analisi sul trasporto delle polveri. A partire da queste informazioni sono stati generati degli sciami di traiettorie che, utilizzando i dati climatici **GDAS** e il tool di generazione di traiettorie **HYSPPLIT**, abbiamo fatto evolvere nel tempo per identificare quali zone con potenziali di trasporto siano effettivamente collegate con i siti di carotaggio del ghiaccio tramite dei pattern di circolazione. Analizzando i percorsi intrapresi dalle circa 50.000 traiettorie prodotte durante le simulazioni è stato possibile, grazie ad una funzione logistica personalizzata, quantificare l'in-

fluenza dei cluster precedentemente identificati verso ogni sito di carotaggio. Grazie alle informazioni distillate da questo processo è stato costruito un modello statistico in grado di mettere in relazione la quantità di polveri in ingresso nei siti, nella frazioni tra i 5 e i 10 micrometri, con la distanza dai cluster ed il loro punteggio di influenza. Il modello proposto riesce a raggiungere un punteggio di R^2 del 99% sui 6 data-point in nostro possesso, dimostrando una significatività statistica tra le variabili frutto delle nostre analisi e il dato Target sul flusso. Il modello statistico identificato è stato utilizzato per produrre delle previsioni sulle 5 stazioni che non presentano misurazioni sulla frazione di polveri oggetto di studio, questi risultati potranno essere confrontati con i valori reali, una volta che le misurazioni per queste stazioni saranno portate a termine, eventualmente confrontando le osservazioni reali con i valori attesi del modello per verificarne la correttezza. È stata infine utilizzata una versione meno potente del modello, basata esclusivamente sulla distanza dai cluster, per produrre una mappa tematica del valore atteso di flusso[5-10 μm] per l'intero continente Antartico. Per confermare le relazioni scoperte tra i clusters e i siti di perforazione sarebbe opportuno riuscire ad ottenere nuove osservazioni così da avere un maggiore supporto statistico per la creazione di un modello più affidabile. Oltre ciò sarebbe vantaggioso estendere le stesse analisi a più archi temporali e con delle simulazioni più esaustive, ciò potrebbe essere particolarmente utile per discernere meglio in quali periodi dell'anno si ha la maggior influenza e secondo quali ciclicità. Oltre il caso riportato nel capitolo 4 dove vengono messi in relazione i clusters con il flusso di polveri in input sarebbe possibile espandere ulteriormente le analisi verso lo studio delle serie storiche prodotte dalle stazioni meteo o sullo sviluppo di nuove aggregazioni ad esempio filtrando i dati sulla velocità del vento in emissione o sull'utilizzo delle componenti del vento ad una risoluzione temporale maggiore permettendoci così di sondare i pattern meno evidenti sulle grandi scale. Più in generale è possibile dire che grazie alle diverse fonti dato integrate e create durante tutto il progetto di tesi, sono state poste le basi per un sistema di analisi ricco e completo, il lavoro svolto non si limita quindi alla costruzione di un unico modello ma alla creazione di un ambiente dove poter sfruttare tutte le informazioni raccolte per fornire gli strumenti in grado di rispondere a quante più domande di ricerca funzionali all'oggetto di studio e finora rimaste inesplorate.

Bibliografia

- [1] F. Lambert, B. Delmonte, J.-R. Petit, M. Bigler, P. R. Kaufmann, M. A. Hutterli, T. F. Stocker, U. Ruth, J. P. Steffensen, and V. Maggi, “Dust record from the EPICA Dome C ice core, Antarctica, covering 0 to 800 kyr BP,” 2008, supplement to: Lambert, F et al. (2008): Dust-climate couplings over the past 800,000 years from the EPICA Dome C ice core. *Nature*, 452, 616-619, <https://doi.org/10.1038/nature06763>. [Online]. Available: <https://doi.org/10.1594/PANGAEA.695995>
- [2] J. R. e. a. Petit, “Climate and atmospheric history of the past 420,000 years from the vostok ice core, antarctica,” *Nature*, vol. 399, no. 6735, pp. 429–436, Jun 1999. [Online]. Available: <https://doi.org/10.1038/20859>
- [3] B. Delmonte, C. I. Paleari, S. Andò, E. Garzanti, P. S. Andersson, J. R. Petit, X. Crosta, B. Narcisi, C. Baroni, M. C. Salvatore, G. Baccolo, and V. Maggi, “Causes of dust size variability in central East Antarctica (Dome B): Atmospheric transport from expanded South American sources during Marine Isotope Stage 2,” *Quaternary Science Reviews*, vol. 168, pp. 55–68, Jul. 2017. [Online]. Available: <https://hal.archives-ouvertes.fr/hal-02105539>
- [4] S. Albani, B. Delmonte, V. Maggi, C. Baroni, J.-R. Petit, B. Stenni, C. Mazzola, and M. Frezzotti, “Interpreting last glacial to holocene dust changes at talos dome (east antarctica): implications for atmospheric variations from regional to hemispheric scales,” *Climate of the Past*, vol. 8, no. 2, pp. 741–750, 2012. [Online]. Available: <https://cp.copernicus.org/articles/8/741/2012/>
- [5] G. Baccolo, B. Delmonte, S. Albani, C. Baroni, G. Cibin, M. Frezzotti, D. Hampai, A. Marcelli, M. Revel, M. C. Salvatore, B. Stenni,

- and V. Maggi, “Atmospheric dust fluxes from the Antarctic ice core Talos Dome.” PANGAEA, 2018, in supplement to: Baccolo, Giovanni; Delmonte, Barbara; Albani, Samuel; Baroni, Carlo; Cibin, Giannantonio; Frezzotti, Massimo; Hampai, Dariush; Marcelli, Augusto; Revel, M; Salvatore, Maria Cristina; Stenni, Barbara; Maggi, Valter (2018): Regionalization of the atmospheric dust cycle on the periphery of the East Antarctic ice sheet since the Last Glacial Maximum. *Geochemistry, Geophysics, Geosystems*, <https://doi.org/10.1029/2018GC007658>. [Online]. Available: <https://doi.org/10.1594/PANGAEA.890896>
- [6] B. Delmonte, C. Baroni, P. Andersson, H. Schoberg, M. Hansson, S. Aciego, J.-R. Petit, S. Albani, C. Mazzola, V. Maggi, M. Frezzotti, C. Andersson, H. Hansson, S. Mazzola, and Aeolian, “Aeolian dust in the talos dome ice core (east antarctica, pacific/ross sea sector): Victoria land versus remote sources over the last two climate cycles,” *J. Quaternary Sci.*, vol. 25, 12 2010.
- [7] S. Aarons, S. Aciego, P. Gabrielli, B. Delmonte, J. Koornneef, A. Wegner, and M. Blakowski, “The impact of glacier retreat from the ross sea on local climate: Characterization of mineral dust in the taylor dome ice core, east antarctica,” *Earth and Planetary Science Letters*, vol. 444, pp. 34–44, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0012821X16301224>
- [8] B. Delmonte, C. Baroni, P. Andersson, B. Narcisi, M. Salvatore, J. Petit, C. Scarchilli, M. Frezzotti, S. Albani, and V. Maggi, “Modern and holocene aeolian dust variability from talos dome (northern victoria land) to the interior of the antarctic ice sheet,” *Quaternary Science Reviews*, vol. 64, pp. 76–89, 2013. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0277379112005355>
- [9] B. Delmonte, H. Winton, M. Baroni, G. Baccolo, M. Hansson, P. Andersson, C. Baroni, M. C. Salvatore, L. Lanci, and V. Maggi, “Holocene dust in east antarctica: Provenance and variability in time and space,”

The Holocene, vol. 30, no. 4, pp. 546–558, 2020. [Online]. Available: <https://doi.org/10.1177/0959683619875188>

- [10] A. Burton-Johnson, M. Black, P. T. Fretwell, and J. Kaluza-Gilbert, “An automated methodology for differentiating rock from snow, clouds and sea in antarctica from landsat 8 imagery: a new rock outcrop map and area estimation for the entire antarctic continent,” *The Cryosphere*, vol. 10, no. 4, pp. 1665–1677, 2016. [Online]. Available: <https://tc.copernicus.org/articles/10/1665/2016/>
- [11] M.-J. Noh and I. M. Howat, “Automated stereo-photogrammetric dem generation at high latitudes: Surface extraction with tin-based search-space minimization (setsm) validation and demonstration over glaciated regions,” *GIScience & Remote Sensing*, vol. 52, no. 2, pp. 198–217, 2015. [Online]. Available: <https://doi.org/10.1080/15481603.2015.1008621>
- [12] A. F. Stein, R. R. Draxler, G. D. Rolph, B. J. B. Stunder, M. D. Cohen, and F. Ngan, “Noaa’s hysplit atmospheric transport and dispersion modeling system,” *Bulletin of the American Meteorological Society*, vol. 96, no. 12, pp. 2059 – 2077, 2015. [Online]. Available: <https://journals.ametsoc.org/view/journals/bams/96/12/bams-d-14-00110.1.xml>
- [13] A. Wegner, H. Fischer, B. Delmonte, J.-R. Petit, T. Erhardt, U. Ruth, A. Svensson, B. Vinther, and H. Miller, “The role of seasonality of mineral dust concentration and size on glacial/interglacial dust changes in the epica dronning maud land ice core,” *Journal of Geophysical Research: Atmospheres*, vol. 120, no. 19, pp. 9916–9931, 2015. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JD023608>
- [14] S. M. Aarons, S. M. Aciego, J. R. McConnell, B. Delmonte, and G. Baccolo, “Dust transport to the taylor glacier, antarctica, during the last interglacial,” *Geophysical Research Letters*, vol. 46, no. 4, pp. 2261–2270, 2019. [Online]. Available: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL081887>

Ringraziamenti

A conclusione di questo elaborato, desidero ringraziare tutte le persone che mi sono state vicine e senza le quali questo lavoro di tesi non sarebbe stato possibile.

Ringrazio il mio relatore, il Professor Albani Samuel per la sua instancabile cortesia e disponibilità nel seguirmi durante tutti i mesi di lavoro in cui mi ha saputo guidare tra le mille difficoltà ed insidie nell'affrontare un campo di ricerca a me sconosciuto ma che ho scoperto, grazie alla sua passione ed umiltà, essere molto interessante e meritevole di attenzione e instancabile ricerca.

Ringrazio di cuore i miei genitori e i miei familiari per avermi sempre sostenuto durante tutto il mio percorso accademico in ogni modo possibile, solo grazie a loro ho avuto il tempo di portare a termine i miei studi e poter approfondire le mie innumerevoli passioni.

Ringrazio i miei amici, i miei colleghi e i miei affetti più stretti con cui ho condiviso questo cammino difficile ma soddisfacente, le innumerevoli sere passate a lavorare e studiare sarebbero state insostenibili senza tutti voi.

Vorrei infine dedicare questo traguardo anche a me stesso, il mio cammino non è stato semplice o lineare, ma sono fiero di poter dire che non avrei voluto seguire nessun'altra carriera se non questa, le molte difficoltà e gli anni che ho lasciato dietro non li riterrò mai persi, perché senza tutte quelle esperienze non sarei mai arrivato dove sono adesso, che questo mio traguardo possa essere l'ennesimo punto di partenza verso i miei obiettivi.