Alma Mater Studiorum — Università di Bologna

Dipartimento di Fisica e Astronomia "Augusto Righi"
Laurea Magistrale in Fisica del Sistema Terra

Modelli di machine learning per la stima della quantità di precipitazione da satellite

Presentata da:
Deniel Pavone

Relatore:
Prof. Federico Porcù

Anno Accademico 2023 - 2024
Appello II

Sommario

La tesi si occupa della realizzazione di un modello di machine learning per la stima della quantità di precipitazione da satellite. Lo strumento satellitare utilizzato è SEVIRI, equipaggiato su MSG (EUMETSAT), e misura su 11 canali spettrali fornendo delle riflettanze (%), per i 3 canali nel visibile, e delle temperature di brillanza (K), per gli 8 canali nell'infrarosso. Essendo la relazione, tra le misure del satellite e la precipitazione, debole, si utilizza come metodo di indagine algoritmi machine learning ad alberi decisionali, nello specifico Random Forest e Gradient Boosting. Si realizza quindi un modello di machine learning a due livelli, una classificazione in intervalli di precipitazione tramite un algoritmo Random Forest e una regressione per ogni classe tramite un algoritmo Gradient Boosting, e lo si addestra sulla precipitazione del DPR, radar satellitare equipaggiato da GPM CO (NASA). Effettuato il test del modello, lo si valida sui valori di precipitazione per l'Italia, della rete pluviometrica gestita dal Dipartimento della Protezione Civile, per un evento di perturbazione nei giorni 22-23 settembre 2019, ottenendo risultati compatibili con le attese.

Indice

Introduzione

La pioggia, la neve, la grandine sono solo alcuni dei fenomeni di trasferimento di acqua, allo stato liquido o solido, dall'atmosfera al suolo. La **precipitazione** comprende tutti questi fenomeni, come risultato di complesse interazioni che coinvolgono una semplice particella d'aria con l'ambiente, dando vita ad un ciclo idrologico caratterizzato da avvenimenti che attraversano scale spaziali e temporali di vari ordini di grandezza. Dalla nucleazione di una goccia, dalla risalita di aria frenetica attraverso urti e trasformazioni al limite della saturazione che consentono la crescita della goccia e la sua modifica in forme sempre più articolate e organizzate in strutture, le nubi, si arriva ad una caduta altrettanto incontrollata, sfidando le forze di galleggiamento e delle correnti, che culmina con la precipitazione delle idrometeore sotto forma di eventi più o meno strutturati, più o meno intensi.

E la precipitazione non è solo un fenomeno che affascina e permette la vita sulla Terra, ma è impattante, risultando sia un'opportunità che una minaccia per l'uomo e le sue innumerevoli attività, soprattutto in un contesto di cambiamento climatico che determina l'aumento di fenomeni estremi. Da qui l'importanza della **stima quantitativa della precipitazione** (QPE), obiettivo di questo lavoro di tesi.

Esistono diversi strumenti di misura che permettono di stimare la quantità e il tipo della precipitazione e la sua intensità in mm/h, tra cui tradizionalmente i più diffusi sono i pluviometri e i radar meteorologici. Tuttavia non è sempre possibile valutare la precipitazione in luoghi remoti o morfologicamente complessi come possono essere oceani, territori montani o aree non popolate. Per tentare di risolvere questo problema si utilizza un punto di vista ampio e privilegiato, dotando di strumenti opportuni i **satelliti** che orbitano intorno alla Terra, in grado così di osservare dall'alto i fenomeni che coinvolgono la precipitazione.

É in attività una costellazione di satelliti dotati di strumenti. Ad esempio strumenti radar a doppia frequenza, come **DPR** sui satelliti GPM della NASA, o ottici, in grado di fornire immagini nello spettro del visibile o dell'infrarosso, come **SEVIRI** sui satelliti MSG dell'EUMETSAT, tutti con risoluzioni spaziali nell'ordine dei 3-5km e temporali di 15-90 minuti.

La QPE da satelliti meteorologici non è un compito facile, e lo è ancor meno da strumenti ottici, come SEVIRI, in quanto le relazioni tra le caratteristiche atmosferiche osservate e la precipitazione, sono deboli. É possibile tuttavia utilizzare **algoritmi in machine learning**, che permettono di addestrare grandi quantità di dati al fine di individuarne dei legami con i fenomeni di precipitazione e la loro intensità.

Questa tesi di laurea si pone come obiettivo la stima della precipitazione da satellite addestrando un **modello**, basato su algoritmi in machine learning ad alberi decisionali, sulle misure in riflettanza e temperatura di brillanza degli 11 canali di SEVIRI e sul prodotto di precipitazione fornito da DPR. Il modello è a doppio livello, con una classificazione in intervalli di intensità di precipitazione tramite un algoritmo di **Random Forest** e una regressione interna ad ogni classe, tramite un algoritmo di **Gradient Boosting**.

Nel Capitolo 1 si descrive la fenomenologia delle nubi e della microfisica dietro i fenomeni di precipitazione, introducendo gli strumenti di misura. Nel Capitolo 2 si descrivono le caratteristiche dei satelliti meteorologici e il principio di funzionamento, oltre che il tipo di dato fornito, degli strumenti satellitari SEVIRI e DPR. Nel Capitolo 3 si introducono le tecniche di machine learning, indagando gli algortimi ad alberi decisionali per problemi in classificazione e regressione.

Nel Capitolo 4 si analizza il dataset di coincidenza tra le misure di SEVIRI e la precipitazione del DPR, fornito dal CNR ISAC di Roma, per l'intero anno 2017, realizzandone una statistica qualitativa dei dati. Nel Capitolo **??** si preparano le features, scegliendo delle classi di intervalli di precipitazione con cui bilanciare il dataset e valutando le performance degli algoritmi nella risoluzione dei problemi in classificazione e regressione.

Infine si descrive il modello completo, evidenziandone i risultati, nel Capitolo 6 per poi validarlo, nel Capitolo 7 tramite la stima su di un evento di pioggia sul suolo italiano avvenuto il 22-23 settembre 2019, confrontata con le misurazioni della rete pluviometrica gestita dalla Protezione Civile.

1 QPE: stima quantitativa della precipitazione

1.1 L'atmosfera terrestre

1.1.1 L'atmosfera

L'**atmosfera terrestre** [1] è un involucro stratificato di gas che avvolge il pianeta Terra. É composta principalmente da azoto N_2 (78%), ossigeno O_2 (21%), argon Ar (0,93%), vapore acqueo e altri gas in concentrazione minore.

Ha una **struttura in livelli** (Fig.1.1), ciascuno con specifiche caratteristiche di gradiente di temperatura e composizione chimica, il cui studio è essenziale per comprendere i fenomeni meteorologici e climatici che evolvono in essa e per lo sviluppo di tecniche di osservazione da satellite o da terra.

In particolare il primo strato, la **troposfera**, si estende dalla superficie fino a circa 8-15 km di altezza a seconda della latitudine e della stagione, e contiene la quasi totalità del vapore acqueo atmosferico, elemento essenziale per la formazione delle nubi e della precipitazione.

Lo **spettro di assorbimento dell'atmosfera**, in Fig.1.2, è determinante per comprendere come i diversi gas influenzano la trasmissione della radiazione elettromagnetica solare, cioè ad onda corta o Short Wave (SW) e quella emessa dalla superficie terrestre, ad onda lunga o Long Wave (LW). Tra i principali gas che dominano l'assorbimento atmosferico, di cui è importante tenerne conto per lo sviluppo di tecniche osservative da satellite, vi sono il vapore acqueo (H_2O), il biossido di carbonio (CO_2), l'ozono (O_3), ed altri gas come l'ossigeno (O_2) e il metano (CH_4).

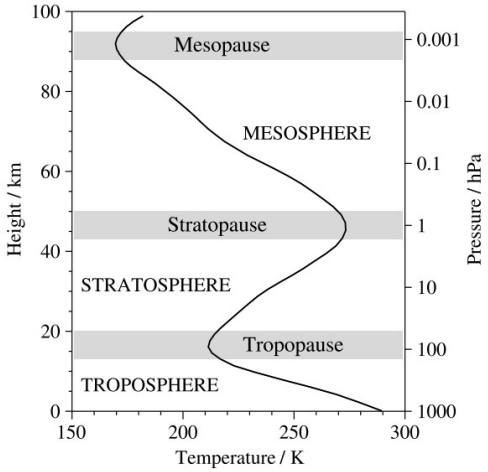


Figura 1.1: Struttura dell'atmosfera: divisione in strati e variazione della temperatura in funzione dell'altitudine

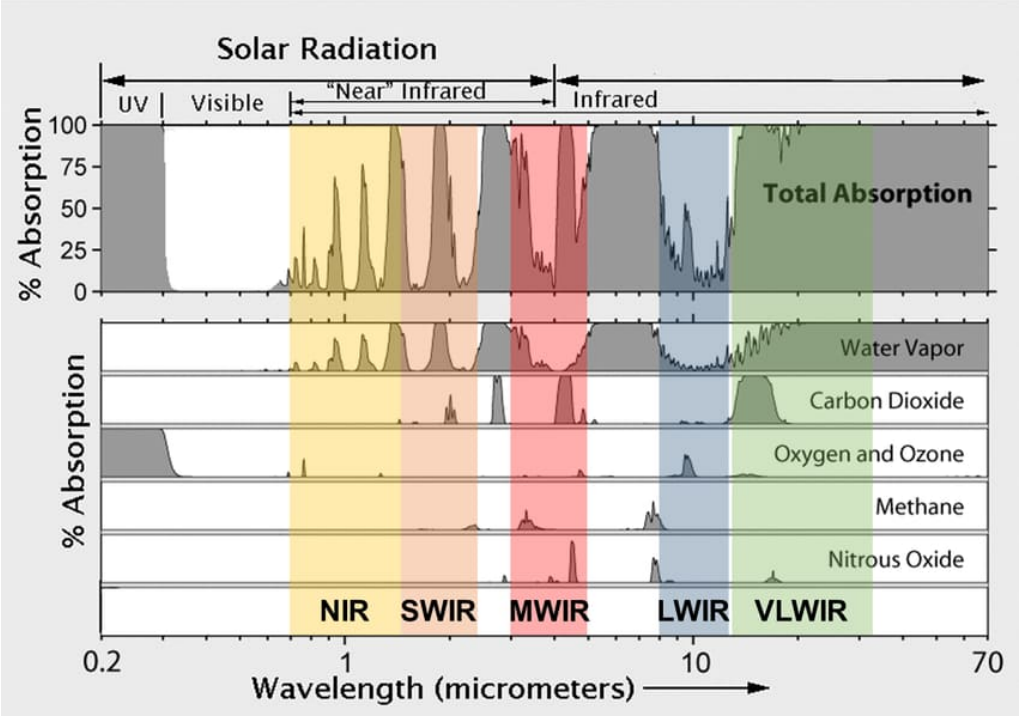


Figura 1.2: Spettro di assorbimento dell'atmosfera e i suoi principali gas, nelle frequenze dell'UV, del visibile e dell'infrarosso [2]

1.1.2 Il vapore acqueo

Il **vapore acqueo** [3, pp. 1-2] ha una concentrazione assoluta in atmosfera piuttosto bassa, nonostante risulti comunque il quarto gas più abbondante. La sua diffusione è altamente variabile a seconda della posizione geografica e dell'altitudine: ha una concentrazione massima vicino alla superficie terrestre e diminuisce con l'altitudine, mentre le zone tropicali contengono più vapore acqueo in confronto con le medie ed alte latitudini.

Dal punto di vista radiativo, il vapore acqueo è caratterizzato da bande di assorbimento nel vicino (NIR) e medio infrarosso (MIR), assumendo un ruolo importante nello spettro di frequenze tra 5 μm e 8 μm, tipiche di alcuni canali di osservazione satellitari.

1.1.3 Le nubi: classificazione

L'acqua, presente abbondantemente come vapore in troposfera, può presentarsi inoltre in fase liquida, sotto forma di gocce, o solida, come cristalli e aggregati di ghiaccio, e costituire quindi le nubi.

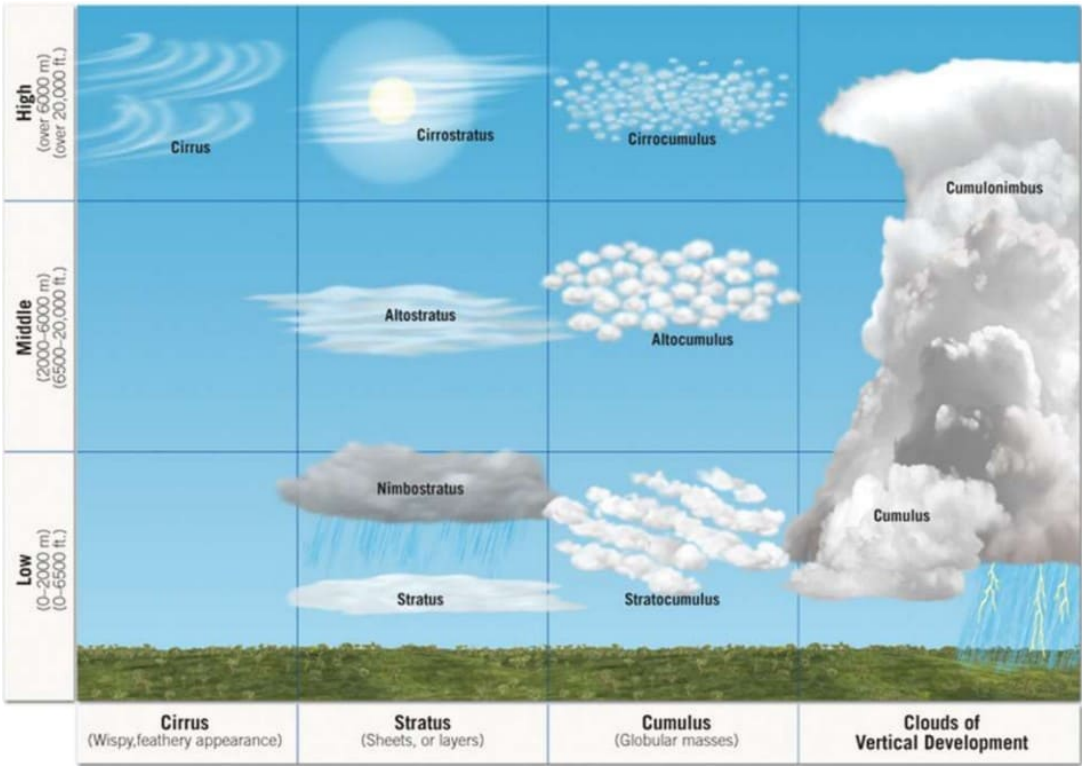


Figura 1.3: Classificazione delle nubi

Le nubi vengono classificate (Fig.1.3) a seconda della loro altezza e della loro forma in [3, pp. 3-14]:

- **Basse**: fino a circa 2 km.
- **Medie**: tra 2 e 6 km.
- **Alte**: sopra i 6 km.
- **Nubi a sviluppo verticale**: con estensione attraverso più livelli di altitudine.

Le nubi alte sono composte quasi interamente da cristalli di ghiaccio e suddivise in tre tipologie principali: cirri (Ci), cirrostrati (Cs) e cirrocumuli (Cc).

Le nubi medie possono essere composte da gocce d'acqua o una combinazione di gocce d'acqua e particelle di ghiaccio e sono di due tipi: altostrati (As) e altocumuli (Ac).

Le nubi basse comprendono: strati (St), stratocumuli (Sc) e nembostrati (Ns). Sono generalmente spesse, scure e possono essere precipitanti. In particolare, gli stratocumuli sono molto comuni sopra gli oceani e coprono aree molto estese, giocando un ruolo radiativo importante nel sistema Terra-atmosfera.

Infine, le nubi a sviluppo verticale sono caratterizzate dalla genesi in ambienti con correnti ascensionali relativamente forti e sono classificate in due tipi, entrambe nubi precipitanti: **cumuli (Cu)** e **cumulonembi (Cb)**.

1.2 Le nubi precipitanti

Le **nubi precipitanti** sono un componente fondamentale del sistema terrestre regolando il ciclo idrologico, contribuendo al bilancio energetico e producendo vari fenomeni meteorologici. La formazione e l'evoluzione delle nubi sono governate da una serie di **processi fisici**, che saranno descritti nel corso di questo paragrafo.

1.2.1 Instabilità atmosferica e correnti ascensionali

L'instabilità atmosferica è un concetto fondamentale nella fisica dell'atmosfera poiché è determinante nella formazione delle nubi e nella possibilità di precipitazioni. Un'atmosfera instabile è caratterizzata dalla tendenza dell'aria a muoversi verticalmente, generando correnti ascensionali. Queste correnti determinano il trasporto di umidità verso i livelli superiori dell'atmosfera, favorendo la condensazione del vapore acqueo e la formazione di nubi.

Instabilità atmosferica

L'instabilità atmosferica è legata alla relazione tra la temperatura dell'aria e la sua densità: una massa d'aria, durante il sollevamento, si espande e si raffredda a causa della diminuzione della pressione atmosferica con l'altezza. Il tasso di raffreddamento di aria secca è il **gradiente adiabatico secco** (Γ_d), ovvero senza scambio di calore con l'ambiente circostante e senza condensazione del vapore, ed è approssimativamente pari a:

$$\Gamma_d = \frac{g}{c_p} \approx 9.8\,^{\circ}\mathrm{C/km}$$

dove Γ_d è il gradiente adiabatico secco, g è l'accelerazione di gravità pari a circa 9.8 m/s^2 e c_p è il calore specifico a pressione costante dell'aria, circa 1004 J/(kg·K).

Il **gradiente adiabatico umido** (Γ_s) è, invece, il tasso di variazione della temperatura di una massa d'aria umida, ossia una miscela di aria secca e vapore acqueo, che si solleva o si abbassa adiabaticamente, tenendo conto della condensazione del vapore acqueo. Poiché la condensazione rilascia calore latente, il gradiente adiabatico umido è inferiore a quello secco e dipende dalla temperatura e dall'umidità dell'aria:

$$\Gamma_s \approx \text{ tra } 4\,^{\circ}\mathrm{C/km} \text{ e } 7\,^{\circ}\mathrm{C/km}$$

Se la massa d'aria sollevata è più calda e meno densa dell'aria circostante, continuerà a salire, generando instabilità. Questa è formalmente descritta dalla **condizione di instabilità atmosferica**:

$$-\frac{\partial T}{dz} = \Gamma < \Gamma_d$$

dove Γ è il gradiente termico verticale e Γ_d è il gradiente adiabatico secco.

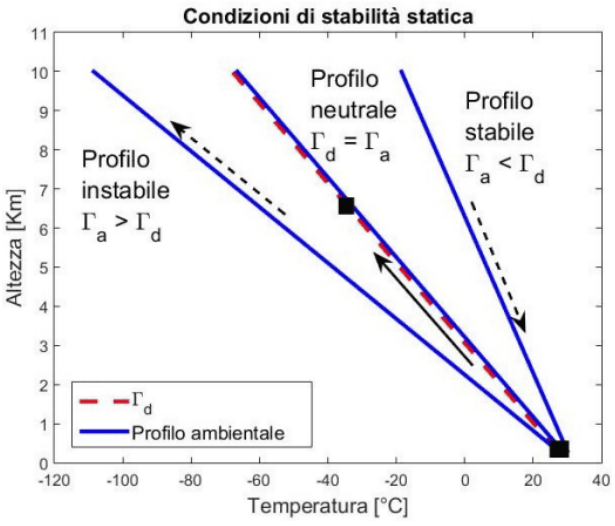


Figura 1.4: Condizioni di stabilità atmosferica. Il moto di una particella ideale (quadrato nero) segue l'adiabatica secca, le tre condizioni di stabilità dipendono dal profilo ambientale

In Fig.1.4 sono riassunte le varie condizioni di stabilità atmosferica, considerando il moto di una particella che segue l'adiabatica secca Γ_d e il profilo di temperatura ambientale Γ_a.

Nel caso di aria satura deve essere considerato il gradiente adiabatico umido e, poiché la condensazione rilascia calore latente, il tasso di raffreddamento è ridotto.

Emagrammi

Per l'analisi dell'instabilità atmosferica, uno strumento fondamentale è il grafico termodinamico emagramma. Gli **emagrammi** forniscono una rappresentazione del profilo verticale della temperatura e dell'umidità dell'atmosfera: sull'asse orizzontale è riportata la temperatura, mentre la pressione, che diminuisce con l'altezza, è riportata sull'asse verticale.

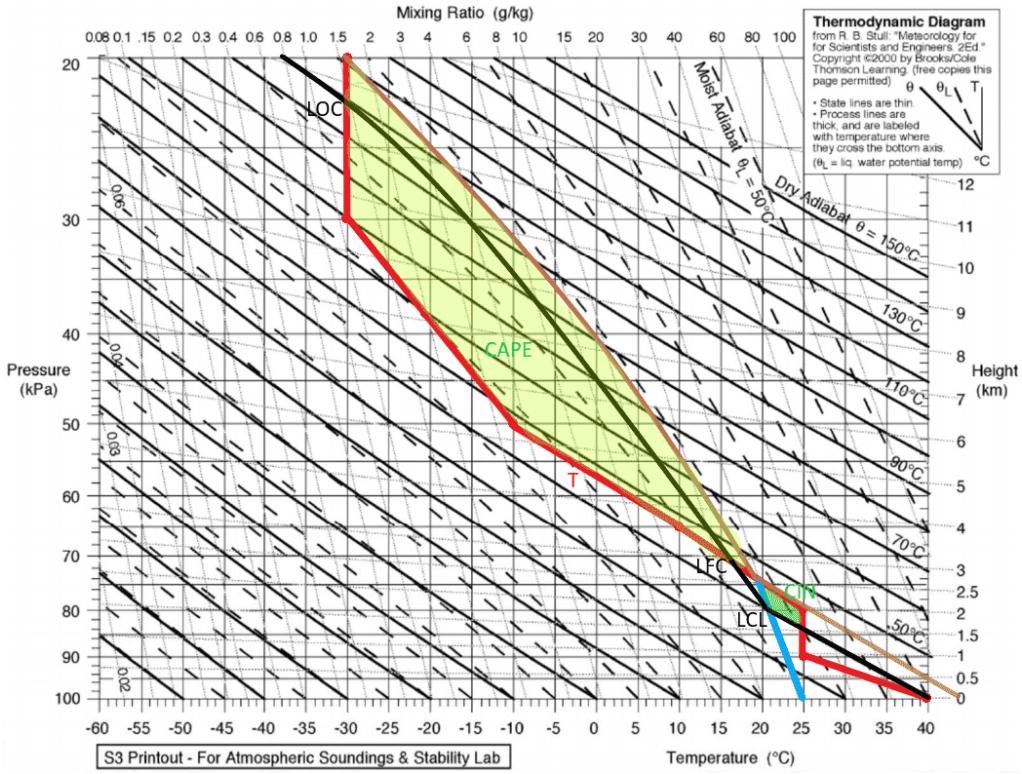


Figura 1.5: Esempio di grafico termodinamico emagramma, con indicazione dei parametri convettivi

In Fig.1.5 è riportato un esempio di emagramma, nel quale si possono osservare le seguenti linee:

- **Isoterme**: linee verticali a temperatura costante.
- **Isobare**: linee orizzontali a pressione costante.
- **Isoumide**: linee diagonali con rapporto di mescolamento $w_s \simeq \epsilon \frac{e_s}{p}$, ovvero la massa di vapore nella massa d'aria alla saturazione, costante.
- **Adiabatica secca**: linee curve che rappresentano il gradiente adiabatico secco.
- **Adiabatica satura**: linee curve che rappresentano il gradiente adiabatico umido.

Sull'emagramma possono essere descritti i profili di temperatura (T) e di temperatura di rugiada (T_d) in funzione dell'altezza e, valutando queste tracce rispetto alle linee adiabatiche secche e sature, è possibile determinare la stabilità atmosferica e prevedere la formazione di nubi e probabilità di temporali.

In particolare si riescono a riconoscere i seguenti parametri fondamentali per la comprensione dei fenomeni convettivi:

- **Livello di condensazione forzata (LCL)**: l'altezza alla quale una massa d'aria umida deve essere sollevata adiabaticamente per raggiungere la saturazione e iniziare la condensazione. È calcolato approssimativamente utilizzando la seguente formula:

$$LCL = T_d - \frac{T_d - T_s}{5.5}$$

dove T_d è la temperatura del punto di rugiada dell'aria a livello del mare e T_s è la temperatura dell'aria a livello del mare.

- **Livello di convezione libera (LFC)**: l'altezza alla quale l'aria, sollevata adiabaticamente fino al LCL, diventa più calda rispetto all'ambiente circostante, ed è indicativo dell'innesco di convezione libera, ovvero di moti verticali dell'aria in atmosfera sostenuti dal gradiente di temperatura e di densità.
- **Energia di inibizione convettiva (CIN)**: rappresenta l'energia che impedisce l'innesco della convezione (rappresentato graficamente in Fig.1.6). È la quantità di energia necessaria per sollevare l'aria al LFC. Se il CIN è alto, l'aria è stabile e la convezione si sviluppa con più difficoltà.
- **Convective Available Potential Energy (CAPE)**: rappresenta l'energia disponibile per il movimento convettivo dell'aria nel caso in cui è sollevata adiabaticamente fino al LFC (in Fig.1.6). È calcolato come l'area tra la traiettoria del profilo di temperatura della massa d'aria (parcel) sollevata e del profilo di temperatura dell'ambiente (env) nella parte superiore dell'emagramma, secondo la formula:

$$\mathrm{CAPE} = \int_{z_{\mathrm{LFC}}}^{z_{\mathrm{LOC}}} \left(\frac{T_v(\mathrm{parcel}) - T_v(\mathrm{env})}{T_v(\mathrm{env})} \right) g \, dz$$

dove T_v(parcel) è la temperatura virtuale, ovvero la temperatura a cui andrebbe portata una massa d'aria secca affinché, alla stessa pressione, abbia la medesima densità di una massa d'aria umida della particella d'aria, T_v(env) è la temperatura virtuale dell'ambiente, g è l'accelerazione di gravità, z_{LFC} è il livello di libera convezione, e z_{LOC} è il livello di equilibrio (EL) che rappresenta l'altezza al limite della convezione.

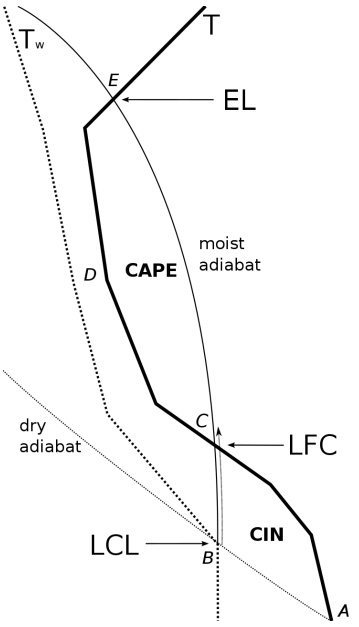


Figura 1.6: Dettaglio di un emagramma sul CIN e CAPE [4]

1.2.2 Transizione in fase dell'acqua

L'acqua si può trovare in tre stati di aggregazione della materia: solido, liquido e gassoso. La **transizione di fase dell'acqua** (in Fig.1.7), che include i processi di evaporazione, condensazione, fusione e solidificazione, è un fenomeno chiave nella formazione delle nubi e delle precipitazioni, e avviene attraverso il **trasferimento di energia** sotto forma di **calore latente**. Ad esempio, durante l'evaporazione, l'acqua assorbe calore dall'ambiente per passare dallo stato liquido a quello gassoso e viceversa, durante la condensazione, il vapore acqueo rilascia calore passando allo stato liquido.

Un importante strumento matematico per descrivere questi cambiamenti di fase è l'equazione di Clausius-Clapeyron.

Equazione di Clausius-Clapeyron

L'**equazione di Clausius-Clapeyron** descrive il cambiamento della pressione di vapore saturo (e_s), ovvero la massima pressione permessa per un vapore a una data temperatura, con la temperatura (T). Questa relazione, data una transizione di fase dalla fase generica 1 alla 2, è espressa dalla formula:

$$\frac{de_s}{dT} = \frac{L_{1\to2}}{(\alpha_2 - \alpha_1)T}$$

dove α è il volume specifico dato dal rapporto tra il volume e la massa di un gas, L è il calore latente di transizione di fase.

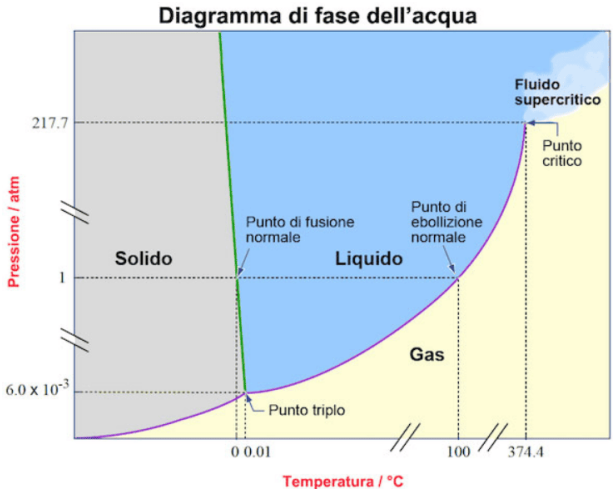


Figura 1.7: Diagramma di fase dell'acqua

Considerando la **transizione da acqua a vapore**, con la condizione che $\alpha_v \gg \alpha_w$, e $e_s\alpha_v = R_vT$ per la legge ideale dei gas, si ottiene:

$$\frac{de_s}{dT} = \frac{L_v e_s}{R_v T^2}$$

dove L_v è il calore latente di vaporizzazione, R_v è la costante specifica del vapore acqueo, e T è la temperatura assoluta.

Questa equazione esprime l'aumento della pressione di vapore saturo esponenzialmente con la temperatura e, integrata, può essere scritta come

$$e_s(T) = e_0 \exp\left(\frac{L_v}{R_v}\left(\frac{1}{T_0} - \frac{1}{T}\right)\right)$$

dove e_0 è la pressione di vapore saturo a una temperatura di riferimento T_0.

L'equazione di Clausius-Clapeyron illustra la capacità di contenere vapore acqueo da parte di una particella d'aria che sale e si raffredda, portando alla condensazione del vapore in gocce d'acqua o alla sublimazione in cristalli di ghiaccio. Questo processo è cruciale per il ciclo dell'acqua e il bilancio energetico terrestre in quanto il rilascio di calore latente durante le transizioni di fase influenzano la dinamica atmosferica, contribuendo alla formazione di sistemi convettivi.

1.2.3 Nubi convettive

In presenza di correnti ascensionali, in accordo con l'equazione di Clausius-Clapeyron, è possibile la formazione di **nubi convettive**. Esistono diverse tipologie di nubi convettive, che variano per dimensioni, struttura e proprietà. In questo paragrafo si analizzano le caratteristiche delle celle singole, delle multicelle, delle supercelle e dei sistemi convettivi alla mesoscala (MCS).

Celle singole

Le **celle singole** sono le unità di base di un sistema convettivo. Una cella convettiva ha una durata di vita tipicamente di 20-30 minuti e attraversa tre fasi principali: la fase di crescita, la fase matura e la fase di dissipazione (in Fig.1.8).

- **Fase di crescita**: durante questa fase iniziale, l'aria calda e umida viene sollevata, raffreddandosi adiabaticamente e raggiungendo il livello di condensazione. Qui il vapore acqueo condensa in gocce, formando una nube cumuliforme.
- **Fase matura**: la nube raggiunge il suo massimo sviluppo verticale. Le correnti ascensionali e discendenti coesistono, e la precipitazione inizia a cadere dalla base della nube. In questa fase, il calore latente rilasciato dalla condensazione dell'acqua favorisce l'ulteriore sollevamento dell'aria.
- **Fase di dissipazione**: le correnti discendenti prevalgono, la fonte di aria calda e umida viene tagliata e la nube inizia a dissiparsi.

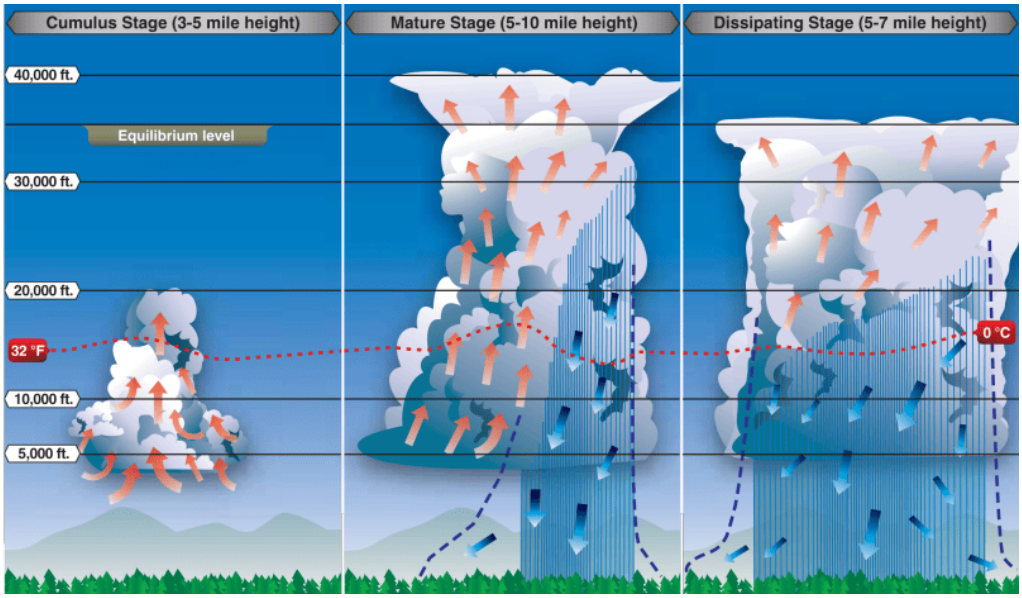


Figura 1.8: Rappresentazione pittorica delle fasi evolutive di un sistema convettivo a cella

Multicelle

Le **multicelle** sono gruppi di celle convettive organizzate in modo che nuove celle si formino in prossimità delle celle mature o in fase di dissipazione. Questo tipo di organizzazione può durare diverse ore, generare precipitazioni abbondanti e temporali. Le multicelle sono caratterizzate da una struttura complessa, con più nuclei di correnti ascensionali e discendenti che interagiscono tra loro, come rappresentato in Fig.1.9.

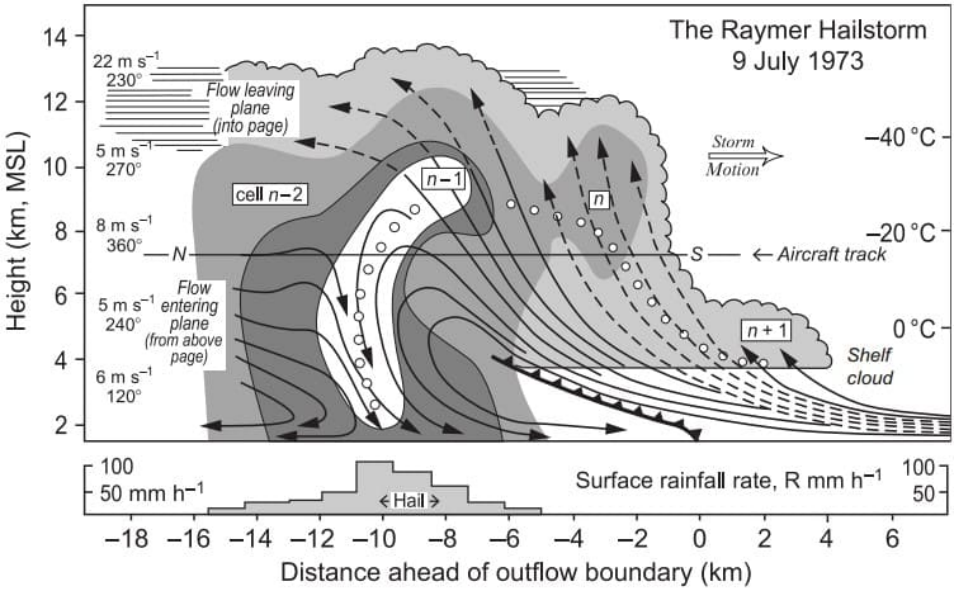


Figura 1.9: Sistema a multi cella [5]

Supercelle

Le **supercelle** sono un tipo particolarmente intenso di nube convettiva. Si caratterizzano per la presenza di un updraft, ovvero una corrente ascensionale, rotante (detta anche mesociclone) che conferisce alla supercella una struttura organizzata e duratura. Una supercella può generare fenomeni meteorologici estremi come grandine di grandi dimensioni, forti raffiche di vento e tornado.

Una supercella presenta una struttura ben definita caratterizzata da:

- **Mesociclone**: la corrente ascensionale rotante al centro della supercella. La rotazione è il risultato dei cambiamenti di velocità del vento verticale (wind shear) che conferisce vorticità inclinando la corrente ascensionale. In particolare, il **wind shear** verticale, che permette il sostentamento e l'intensificazione delle supercelle, è definito come la variazione del vento con l'altezza:

$$\vec{S} = \frac{\partial \vec{V}}{\partial z}$$

  dove $\vec{V}$ è la velocità del vento e z è l'altezza.

- **Area di inflow**: la regione ove l'aria calda e umida viene risucchiata nella supercella.
- **Area di downdraft**: la discesa di aria fredda, spesso accompagnata da precipitazioni e venti intensi.

Sistemi convettivi alla mesoscala (MCS)

I **sistemi convettivi alla mesoscala (MCS)** sono grandi complessi di nubi convettive che possono coprire aree estese e perdurare molte ore, generando grandi quantità di precipitazione. Gli MCS possono assumere diverse forme:

- **Linee di groppo**: strutture lineari di temporali che si formano lungo i fronti freddi o altre discontinuità atmosferiche. La formazione avviene quando l'aria calda e umida viene sollevata bruscamente da un fronte freddo.
- **Cluster di temporali**: gruppi di temporali che si organizzano in modo casuale, ma cooperano per formare un sistema più grande.
- **Complessi di mesoscala convettivi**: sistemi altamente organizzati che possono coprire centinaia di chilometri e durare fino a 12 ore o più.

1.2.4 Nubi stratificate

Le **nubi stratificate** sono un tipo di nubi che si formano in condizioni atmosferiche stabili e sono caratterizzate da una struttura orizzontalmente uniforme. Queste nubi si sviluppano in modo predominante in orizzontale anziché verticalmente, a differenza delle nubi convettive. Le due tipologie più comuni di nubi stratificate sono i nimbostrati e le nubi frontali.

I **nimbostrati** sono nubi stratificate grigie e dense che coprono ampie aree del cielo e si formano in presenza di un'ampia massa d'aria umida che viene sollevata gradualmente, tipicamente in presenza di un fronte, ovvero una discontinuità tra masse d'aria con temperatura, pressione e umidità differenti. A causa della stabilità atmosferica, l'aria si solleva in modo lento e uniforme, favorendo la formazione di uno strato continuo e omogeneo.

La formazione dei nimbostrati può essere descritta utilizzando la legge della diffusione:

$$\frac{\partial \rho_v}{\partial t} + \nabla \cdot (\rho_v \vec{V}) = \frac{1}{r} \frac{\partial}{\partial r} \left(rD \frac{\partial \rho_v}{\partial r} \right)$$

dove ρ_v è la densità di vapore acqueo, $\vec{V}$ è la velocità dell'aria, r è il raggio della goccia, e D è il coefficiente di diffusione.

Il termine di diffusione, $\frac{1}{r} \frac{\partial}{\partial r} \left(rD \frac{\partial \rho_v}{\partial r} \right)$, descrive come il vapore acqueo si diffonde e si condensa sulla superficie delle gocce, permettendone la crescita. Mentre il termine $\nabla \cdot (\rho_v \vec{V})$ rappresenta il flusso di massa del vapore acqueo dovuto al movimento dell'aria e indica come il vapore acqueo viene trasportato dall'aria in movimento, influenzandone la distribuzione spaziale e contribuendo alla formazione e crescita delle nubi.

Le **nubi frontali** si formano lungo i fronti atmosferici, spesso associate a condizioni atmosferiche di instabilità.
I tipi di nubi associate alle nubi frontali sono:

- **Nubi cirrostrati**: sottili strati di nubi di ghiaccio che si formano ad altitudini elevate lungo i fronti.
- **Nubi altostrati**: strati di nubi grigie e dense ad altitudini medie che spesso precedono un fronte caldo.

- **Nubi stratocumuli**: piccole nubi di forma cumuliforme.

1.3 Formazione e struttura microfisica della precipitazione

La **precipitazione** è il risultato di una serie di processi complessi che avvengono all'interno delle nubi, per i quali la **struttura microfisica** delle nubi gioca un ruolo cruciale nella determinazione del tipo e dell'intensità delle precipitazioni.

Uno degli aspetti chiave della formazione delle precipitazioni è la **crescita delle idrometeore**, le particelle che compongono la precipitazione, attraverso processi di urto come la coalescenza, l'aggregazione e il riming. Questi processi caratterizzano e permettono di descrivere la distribuzione della grandezza delle idrometeore (DSD, drop size distribution), fondamentale per classificare il tipo di precipitazione.

1.3.1 Struttura microfisica delle nubi

Le nubi possono essere classificate in tre categorie principali in base alla loro composizione: nubi d'acqua, nubi di ghiaccio e nubi miste. Ciascuna di queste tipologie presenta **caratteristiche microfisiche** specifiche che influenzano la dinamica della nube e i processi di formazione delle idrometeore.

Nubi d'acqua

Le **nubi d'acqua** sono composte principalmente da gocce di acqua, la cui dimensione varia tipicamente da pochi μm fino a circa 50 μm. La concentrazione delle gocce di nube (N_d) è generalmente dell'ordine di 10^8 a 10^9 gocce per m^3.

Le gocce d'acqua nelle nubi si formano attraverso il processo di **nucleazione**, in cui il vapore acqueo condensa su nuclei di condensazione (CCN, Cloud Condensation Nuclei). Questi **nuclei di condensazione** possono essere costituiti da aerosol marini, polveri minerali, particelle di fuliggine, pollini e spore, di origine sia naturale che antropica.

La **crescita delle gocce** avviene inizialmente per diffusione del vapore acqueo e successivamente per coalescenza, quando le gocce più piccole si uniscono per formare gocce più grandi.

Nubi di ghiaccio

Le **nubi di ghiaccio** sono composte da cristalli di ghiaccio. I cristalli possono assumere varie forme, come aghi, colonne, piastrine, dendriti o altre combinazioni complesse, e la loro formazione avviene con la deposizione di vapore acqueo su nuclei di ghiaccio (IN, Ice Nuclei) o attraverso il congelamento delle gocce di acqua sopraffusa, nel processo di riming.

La velocità di crescita dei cristalli di ghiaccio per deposizione del vapore può essere descritta dall'equazione di crescita per diffusione di Maxwell:

$$\frac{dm}{dt} = S_i - 1 \left(\frac{4\pi r \rho_i}{\left(\frac{L}{R_v T} - 1\right) \frac{D_v}{k} + \frac{L^2 \rho_i}{K R_v T^2}} \right)$$

dove m è la massa del cristallo, r è il raggio, ρ_i è la densità del ghiaccio, S_i è la sovrasaturazione rispetto al ghiaccio, L è il calore latente di sublimazione, R_v è la costante dei gas per il vapore acqueo, T è la temperatura, D_v è il coefficiente di diffusione del vapore acqueo nell'aria, k è la conducibilità termica dell'aria e K è la costante di Boltzmann. L'equazione rappresenta la capacità di crescita del cristallo, in risposta alla sovrasaturazione disponibile $S_i - 1$.

Nubi miste

Le **nubi miste** si trovano tipicamente in condizioni di temperatura comprese tra 0°C e -40°C e contengono sia gocce di acqua che cristalli di ghiaccio. La coesistenza di acqua liquida e ghiaccio nelle nubi miste è di particolare importanza per i processi microfisici, poiché i cristalli di ghiaccio possono crescere a spese delle gocce d'acqua attraverso il processo di Bergeron-Findeisen, poichè il vapore acqueo ha una pressione di saturazione più bassa sopra il ghiaccio rispetto che all'acqua liquida, portando alla deposizione del vapore sui cristalli di ghiaccio e all'evaporazione delle gocce d'acqua.

Il tasso di crescita dei cristalli di ghiaccio in una nube mista può essere descritto dall'equazione di deposizione di Pruppacher e Klett:

$$\frac{dr_i}{dt} = \frac{\alpha \rho_w r_l}{\rho_i (r_i + r_l)}$$

dove r_i è il raggio del cristallo di ghiaccio, r_l è il raggio della goccia d'acqua, α è un fattore di forma che dipende dalla geometria del cristallo, ρ_w è la densità dell'acqua e ρ_i è la densità del ghiaccio.

La dinamica delle nubi miste è complessa e coinvolge le interazioni tra gocce d'acqua e cristalli di ghiaccio, inclusi i processi di coalescenza e aggregazione e di riming (descritti nel paragrafo 1.3.2), che portano ad un aumento di massa e della dimensione delle particelle di ghiaccio.

Parametri microfisici delle nubi

La **caratterizzazione microfisica delle nubi** richiede la misurazione di diversi parametri, cruciali per la comprensione dei processi di formazione e sviluppo delle nubi, tra cui la concentrazione di particelle N (m^{-3}), la distribuzione dimensionale delle particelle DSD ($m^{-3}mm^{-1}$), il contenuto d'acqua liquida LWC (g/m^3) e il contenuto di ghiaccio IWC (g/m^3).

La **concentrazione delle particelle in una nube (N)**, come le gocce d'acqua o di ghiaccio, può essere espressa dalla seguente formula:

$$N = \frac{n_d}{V}$$

dove N è la concentrazione delle gocce di nube in numero di gocce per unità di volume, n_d è il numero totale di gocce di nube, e V è il volume della nube.

Il **contenuto d'acqua liquida (LWC)** è dato da:

$$LWC = \int_0^\infty \frac{4}{3} \pi r^3 \rho_w n(r) dr$$

dove $n(r)$ è la distribuzione dimensionale delle gocce e ρ_w è la densità dell'acqua.

Analogamente, il **contenuto di ghiaccio (IWC)** può essere espresso come:

$$IWC = \int_0^\infty m_i n_i(r) dr$$

dove m_i è la massa del cristallo di ghiaccio di raggio r e $n_i(r)$ è la distribuzione dimensionale dei cristalli di ghiaccio.

1.3.2 Crescita di idrometeore per urto

La **crescita delle idrometeore**, ovvero le particelle che compongono le precipitazioni, avviene attraverso diversi meccanismi di urto: coalescenza, aggregazione e riming (in Fig.1.10). Questi processi sono cruciali per determinare la dimensione finale delle particelle di precipitazione e influenzano significativamente la tipologia e l'intensità delle precipitazioni.

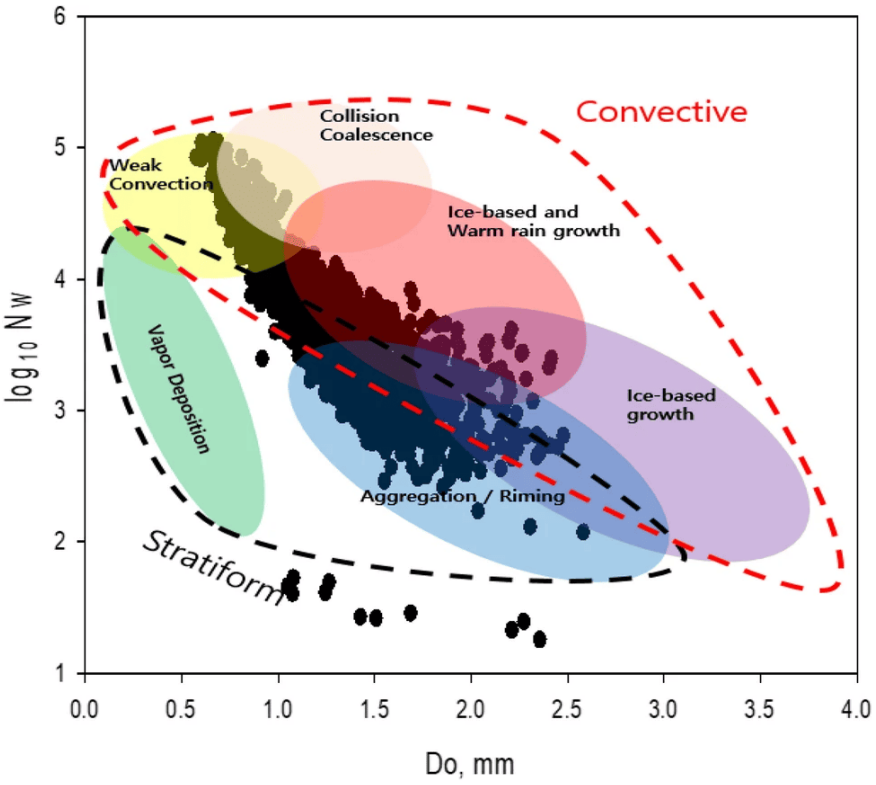


Figura 1.10: Stima dei processi microfisici dominanti per l'interazione tra le particelle di una nube; D_o è il diametro mediano della particella e N_w è un parametro legato alla probabilità di urto tra le particelle. [6]

Coalescenza

La **coalescenza** è un processo predominante nelle nubi d'acqua, dove le gocce di acqua si scontrano e si uniscono per formare gocce più grandi. Questo processo è governato dalle interazioni fluidodinamiche e dalle forze di superficie tra le gocce, e dipende dalla velocità di caduta relativa tra le gocce e dalle loro dimensioni.

La velocità terminale, ovvero la velocità costante di caduta che una corpo raggiunge a seguito del bilancio tra la forza di gravità e la resistenza dell'aria, di una goccia d'acqua (v_t) in aria può essere espressa come:

$$v_t = \sqrt{\frac{8rg(\rho_w - \rho_a)}{3\rho_a C_d}}$$

dove r è il raggio della goccia, g è l'accelerazione di gravità, ρ_w è la densità dell'acqua, ρ_a è la densità dell'aria, e C_d è il coefficiente di drag.

Durante la caduta, le gocce più grandi acquisiscono una velocità terminale maggiore rispetto alle gocce più piccole, aumentando la probabilità di collisione e coalescenza.

Il tasso di crescita per coalescenza può essere descritto dall'**equazione di continuità** per la distribuzione delle dimensioni delle gocce:

$$\frac{\partial n(r,t)}{\partial t} + \frac{\partial}{\partial r}\left(G(r)n(r,t)\right) = S(r,t)$$

dove $n(r,t)$ è la distribuzione delle dimensioni delle gocce, $G(r)$ è la velocità di crescita delle gocce per coalescenza e $S(r,t)$ rappresenta le sorgenti del processo.

Aggregazione

L'**aggregazione** è un processo fondamentale nelle nubi di ghiaccio, in cui i cristalli di ghiaccio si scontrano e si uniscono per formare fiocchi di neve più grandi. La forma e la dimensione dei cristalli di ghiaccio influenzano la loro capacità di aggregarsi: cristalli con superfici irregolari o con bracci estesi, come i dendriti, tendono ad aggregarsi più facilmente rispetto ai cristalli con superfici lisce.

La velocità di aggregazione dipende dalla concentrazione di cristalli di ghiaccio e dalle condizioni termodinamiche della nube. La **frequenza di collisione** tra cristalli di ghiaccio può essere espressa come:

$$C_{ij} = E_{ij}\pi(r_i + r_j)^2 \left|v_i - v_j\right| n_i n_j$$

dove C_{ij} è la frequenza di collisione tra cristalli di ghiaccio di dimensioni i e j, E_{ij} è l'efficienza di collisione, r_i e r_j sono i raggi dei cristalli, v_i e v_j sono le loro velocità terminali, e n_i e n_j sono le loro concentrazioni.

L'aggregazione porta alla formazione di fiocchi di neve che, aumentando di dimensione, hanno una maggiore probabilità di raggiungere il suolo come precipitazione.

Riming

Il **riming** è il processo mediante il quale le gocce d'acqua sopraffuse, ovvero gocce allo stato liquido ad una temperatura inferiore al suo punto di congelamento normale di 0°C, si congelano al contatto con cristalli di ghiaccio o fiocchi di neve. Questo processo è predominante nelle nubi miste, dove coesistono gocce d'acqua liquide e cristalli di ghiaccio.

Il riming aumenta la massa e la dimensione delle particelle di ghiaccio, trasformandole in graupel o, in casi estremi, in grandine. Il **tasso di crescita per riming** può essere descritto dall'equazione:

$$\frac{dm_i}{dt} = \pi r_i^2 \rho_w V (S_w - 1)$$

dove m_i è la massa del cristallo di ghiaccio, r_i è il raggio del cristallo, ρ_w è la densità dell'acqua, V è la velocità relativa tra il cristallo di ghiaccio e le gocce d'acqua, e S_w è la sovrasaturazione rispetto all'acqua.

Il riming non solo modifica la dimensione delle particelle di ghiaccio, ma può anche alterare le loro proprietà fisiche, come la densità e la forma, rendendo le particelle più compatte e sferiche. Questo processo è particolarmente importante per la formazione della grandine, dove il ripetuto ciclo di crescita e congelamento delle gocce porta alla formazione di agglomerati di ghiaccio di grandi dimensioni.

1.3.3 DSD della precipitazione

La **Drop Size Distributione (DSD)** descrive la distribuzione delle dimensioni delle idrometeore all'interno di un volume d'aria e fornisce informazioni essenziali sulla quantità, intensità e tipologia della precipitazione.

Un esempio di DSD relativa alle gocce di pioggia, con le distribuzioni aventi come parametro l'intensità di precipitazione, è riportato in Fig.1.11:

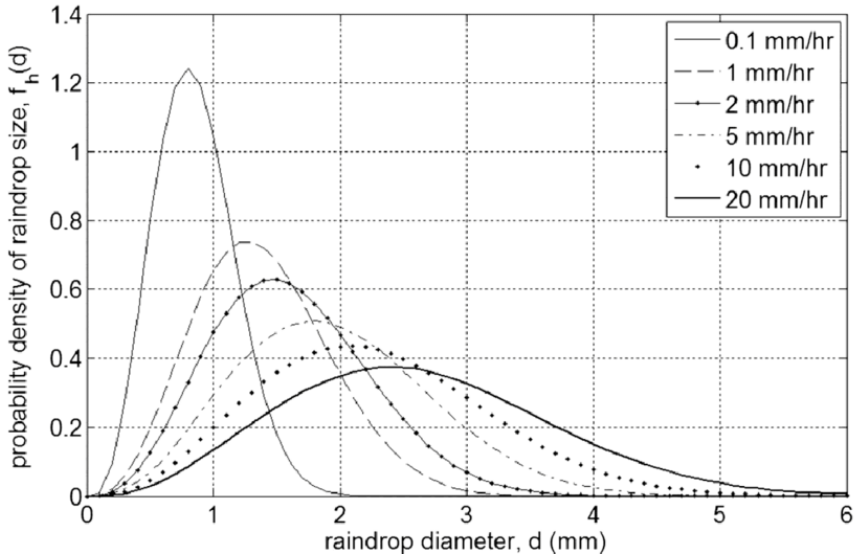


Figura 1.11: DSD delle gocce di pioggia in funzione dell'intensità di precipitazione. [7]

Descrizione della DSD

Una delle rappresentazioni più comuni della DSD è la **distribuzione esponenziale**, proposta da Marshall e Palmer (1948), che può essere espressa come:

$$N(D) = N_0 \exp(-\Lambda D)$$

dove $N(D)$ è il numero di gocce con diametro D per unità di volume per unità di intervallo di diametro, N_0 è un parametro di scala e Λ è il parametro di decadimento che dipende dall'intensità della precipitazione.

Il parametro Λ è inversamente proporzionale alla dimensione media delle gocce, con valori più piccoli di Λ che indicano una presenza maggiore di gocce di grandi dimensioni.

Un'altra rappresentazione comune della DSD è la **distribuzione gamma generalizzata**, che offre maggiore flessibilità rispetto alla distribuzione esponenziale:

$$N(D) = N_0 D^{\mu} \exp(-\Lambda D)$$

dove μ è un parametro che descrive la forma della distribuzione. Quando $\mu = 0$, questa equazione si riduce alla distribuzione esponenziale di Marshall-Palmer.

I parametri della DSD, come N_0 e Λ, possono variare notevolmente in funzione dell'intensità e del tipo di precipitazione. La relazione tra l'intensità della pioggia (R) e il parametro Λ è spesso espressa come:

$$\Lambda = aR^{-b}$$

dove a e b sono costanti empiriche che dipendono dalle condizioni atmosferiche e dal tipo di nube.

Tipologie di DSD in funzione dell'idrometeora

La DSD varia significativamente a seconda del tipo di idrometeora: per le gocce di pioggia, le distribuzioni esponenziali e gamma sono le più comuni. mentre per i cristalli di ghiaccio la DSD può essere descritta da una **distribuzione gamma modificata**:

$$N(D) = N_0 D^{\mu} \exp\left(-\left(\frac{D}{D_m}\right)^{\nu}\right)$$

dove D_m è un diametro medio e ν è un parametro di forma che varia con il tipo di cristallo di ghiaccio.

Per la grandine, la DSD tende ad avere una maggiore concentrazione di particelle di grandi dimensioni e può essere rappresentata da una **distribuzione log-normale**:

$$N(D) = \frac{N_0}{D\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln D - \ln D_g)^2}{2\sigma^2}\right)$$

dove D_g è il diametro geometrico medio e σ è la deviazione standard logaritmica.

Influenza dei processi microfisici sulla DSD

La DSD è influenzata da vari processi microfisici all'interno delle nubi, tra cui la coalescenza, la frammentazione, l'evaporazione e la condensazione.

La **coalescenza** tende ad aumentare la dimensione media delle gocce, spostando la DSD verso dimensioni maggiori. Il tasso di crescita per coalescenza può essere descritto tramite l'equazione di continuità per la distribuzione delle dimensioni delle gocce:

$$\frac{\partial n(D,t)}{\partial t} + \frac{\partial}{\partial D}\left(G(D)n(D,t)\right) = S(D,t)$$

dove $n(D,t)$ è la distribuzione delle dimensioni delle gocce, $G(D)$ è la velocità di crescita delle gocce per coalescenza e $S(D,t)$ rappresenta le sorgenti del processo.

La **frammentazione** di gocce di pioggia in seguito a collisioni, aumenta il numero di gocce di dimensioni più piccole, modificando la DSD. Il tasso di frammentazione può essere descritto dall'equazione:

$$\frac{\partial n(D,t)}{\partial t} = -\beta(D)n(D,t) + \int_D^{\infty} \beta(D')f(D,D')n(D',t)dD'$$

dove $\beta(D)$ è il tasso di frammentazione per gocce di diametro D e $f(D,D')$ è la funzione di distribuzione delle dimensioni delle gocce risultanti dalla frammentazione. I due termini nel membro di destra dell'equazione rappresentano rispettivamente il tasso di perdita e di crescita della popolazione n(D,t) di idrometeore.

L'**evaporazione** delle gocce durante la caduta può ridurre la dimensione delle gocce, soprattutto nelle condizioni di aria secca, a causa della velocità del vento e della concentrazione di vapore acqueo nell'aria circostante. La **condensazione**, al contrario, aggiunge massa alle gocce, aumentando la loro dimensione.

La crescita delle gocce d'acqua può essere quindi descritta dal bilancio tra il tasso di condensazione e il tasso di evaporazione:

$$\frac{dD}{dt} = -\frac{2\rho_a D_v}{\rho_w D}\left(e_s(T) - e\right)$$

dove ρ_a è la densità dell'aria, D_v è il coefficiente di diffusione del vapore acqueo, ρ_w è la densità dell'acqua, $e_s(T)$ è la pressione di saturazione del vapore acqueo alla temperatura T ed e è la pressione parziale del vapore acqueo.

Il termine $e_s(T) - e$ rappresenta la differenza tra la pressione di vapore di saturazione alla temperatura T e la pressione di vapore effettiva nell'aria. Quando $e_s(T) > e$, c'è una tendenza alla condensazione delle gocce d'acqua, quindi $\frac{dD}{dt}$ sarà positivo, indicando crescita delle gocce.

Il termine $-\frac{2\rho_a D_v}{\rho_w D}$ indica il tasso di evaporazione delle gocce d'acqua. Maggiore è la velocità del vento ρ_a, maggiore è il tasso di evaporazione. Viceversa, maggiore è il raggio delle gocce D, minore è il tasso di evaporazione.

1.4 Strumenti di misura: pluviometri, radar e satelliti

I metodi tradizionali per la stima quantitativa delle precipitazioni (QPE) [8] includono l'uso di pluviometri, radar e satelliti. Negli ultimi anni, è aumentato l'interesse nel combinare più metodologie al fine di migliorare l'accuratezza e la affidabilità della QPE, valutando quindi sia tecniche di misurazione da terra che dallo spazio.

1.4.1 Pluviometri

I **pluviometri** sono strumenti basati su metodi di misura meccanici (peso, volume) semplici e diffusi per la valutazione puntuale delle precipitazioni, associati a stazioni meteorologiche da terra. L'utilizzo dei pluviometri offre diversi vantaggi: l'accuratezza (utile per la calibrazione di altri stumenti di stima della precipitazione), la relativa economicità e la facilità di manutenzione. Tuttavia, presentano alcune limitazioni, in particolare la limitata rappresentatività spaziale delle misurazioni puntuali e la possibile incidenza di errori dovuti a fattori ambientali come il vento e l'evaporazione.

1.4.2 Radar

Il **radar meteorologico** (Fig.1.12) è uno strumento in grado di individuare le idrometeore a distanza tramite l'emissione e la ricezione di onde radio, permettendo una copertura spaziale su ampi territori. Il radar è composto da un trasmettitore che genera l'impulso elettromagnetico, un ricevitore che elabora il segnale di ritorno e da un'antenna parabolica che costituisce l'interfaccia tra gli apparati e l'atmosfera, con il compito di focalizzare il fascio trasmesso e di intercettare quello ricevuto a seguito della retrodiffusione delle onde con le idrometeore bersaglio.

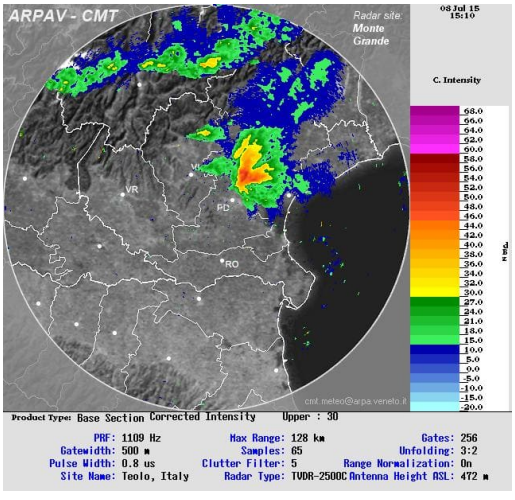


Figura 1.12: Esempio di osservazione radar, rete ARPAV [9]

La riflettività radar (Z) è una misura della potenza del segnale radar riflesso dalle particelle di pioggia ed è correlata alla distribuzione delle dimensioni delle particelle attraverso l'equazione:

$$Z = \int_0^\infty N(D) D^6 dD$$

dove N(D) rappresenta il numero di particelle con diametro D. La riflettività è quindi una funzione della dimensione delle particelle e della loro concentrazione.

L'intensità delle precipitazioni (R) può essere ottenuta dalla riflettività utilizzando relazioni empiriche come la relazione di Marshall e Palmer:

$$R = aZ^b$$

dove a e b sono coefficienti che possono variare in base alla posizione geografica e alla stagione.

Anche il radar presenta delle limitazioni, tra cui la possibilità di errori dovuti alla variabilità della riflettività atmosferica e all'attenuazione del segnale radar stesso. I fattori che possono influenzare l'accuratezza delle misurazioni radar includono la presenza di eco di suolo, la presenza di ostacoli nel percorso delle onde del radar, e l'ambiguità di velocità causata dal movimento delle particelle di pioggia e dei sistemi convettivi.

1.4.3 Satelliti

I **satelliti** offrono un notevole vantaggio nella stima quantitativa delle precipitazioni grazie alla loro copertura globale e una buona frequenza di osservazione, tramite l'utilizzo di sensori passivi, che catturano la radiazione terrestre, che attivi, che registrano la riflessione di un segnale inviato sulla Terra.

È possibile ricavare dati sulle precipitazioni tramite satelliti meteorologici (in Fig.1.13 un esempio di alcuni dei satelliti operativi nel 2015). Massimizzare la disponibilità dei dati comporta l'utilizzo di satelliti geostazionari (GEO) con sensori visibili (VIS) e infrarossi (IR), e sonde a microonde (MW) su satelliti in orbita terrestre bassa (LEO).

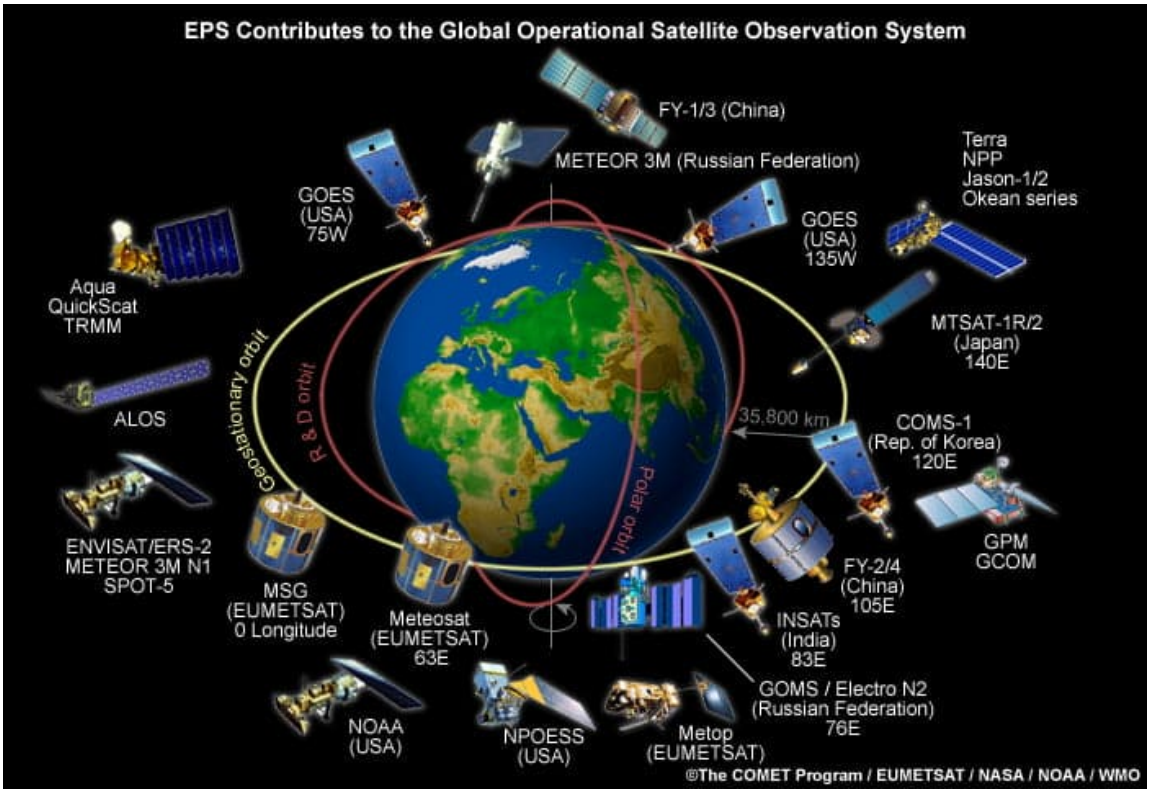


Figura 1.13: Satelliti in orbita nel 2015 [10]

La QPE basata sui satelliti permette una copertura globale e un monitoraggio frequente delle precipitazioni. Tuttavia, non è priva di limitazioni a causa di una difficile calibrazione degli strumenti e della debole relazione fisica tra la precipitazione e le quantità osservabili da satellite.

2 SEVIRI (MSG) e DPR (GPM)

2.1 Satelliti meteorologici

2.1.1 Caratteristiche orbitali e di misura dei satelliti

I **satelliti meteorologici** [11] [12] sono strumenti essenziali per l'osservazione delle condizioni atmosferiche.

Si suddividono principalmente in due categorie: satelliti geostazionari (GEO) e satelliti in bassa orbita terrestre (LEO):

- **Satelliti geostazionari (GEO)**: orbitano a un'altitudine di circa 35.786 km sopra l'equatore e mantengono una posizione fissa rispetto alla superficie terrestre, caratteristica che li rende ideali per il monitoraggio continuo di vaste aree geografiche. La tipica velocità angolare è di circa 3,1 km/s e seguono orbite equatoriali. I satelliti geostazionari geosincroni hanno inoltre un periodo orbitale che coincide con il giorno siderale della Terra.
- **Satelliti in bassa orbita terrestre (LEO)**: orbitano a un'altitudine tipica tra i 600 e i 900 km (da un minimo di 160 fino un massimo di 2000 km) e passano in prossimità dei poli della Terra con un'orbita eliosincrona, con velocità orbitale di circa 7,8 km/s, riuscendo a completare un'orbita in circa 90-120 minuti. Essi forniscono una copertura globale poiché ogni orbita successiva copre una sezione diversa della superficie terrestre, con alta risoluzione spaziale.

In Fig.2.1 si può avere un riassunto delle orbite satellitare divise per inclinazione, forma e altitudine:

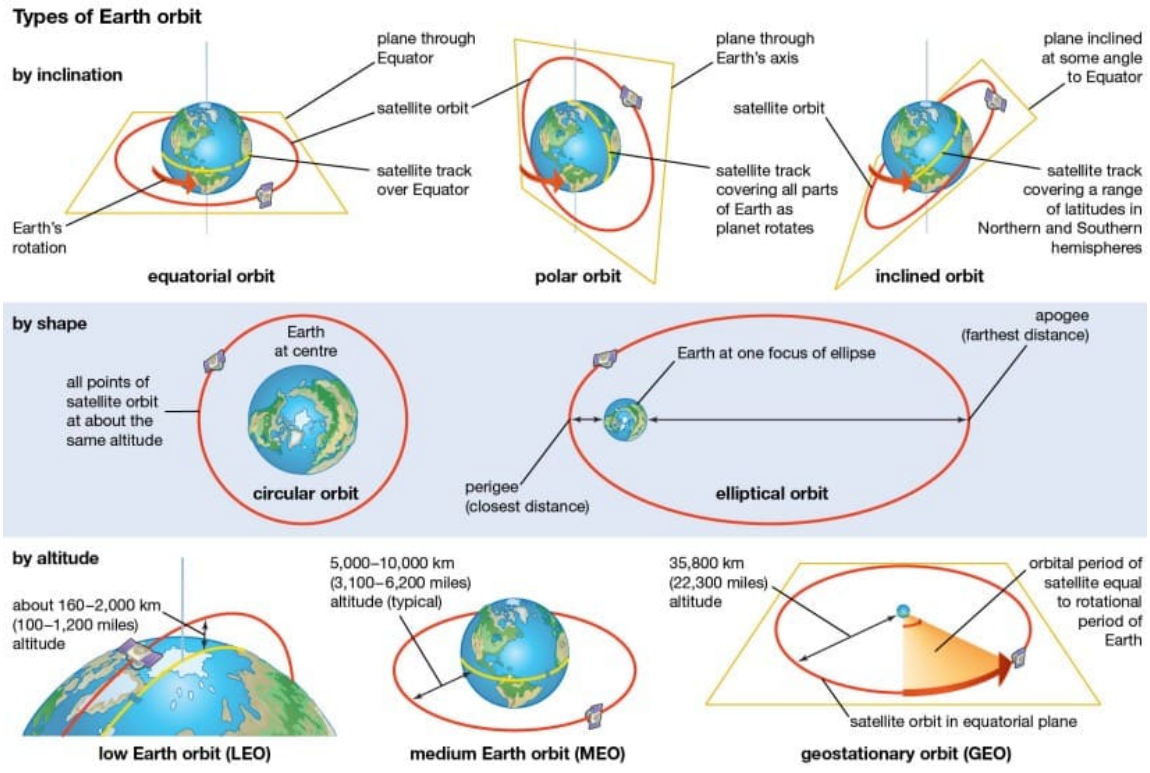


Figura 2.1: Tipologie di orbite dei satelliti [13]

I satelliti utilizzano diverse bande spettrali per monitorare vari parametri atmosferici:

- **Visibile (VIS)**: la radiazione visibile (400-700 nm) è utile per osservare la superficie terrestre e la copertura nuvolosa durante il giorno e i vari canali in cui è misurata la radiazione consentono di valutare le diverse caratteristiche di nubi e superfici.
- **Infrarosso (IR)**: la radiazione infrarossa (700 nm - 1 mm) è utilizzata per misurare la temperatura della superficie terrestre e delle nuvole, sia di giorno che di notte. La radiazione infrarossa termica è emessa dagli oggetti in base alla loro temperatura secondo la legge di Planck e le loro caratteristiche emissive.

- **Microonde (MW)**: le osservazioni nelle microonde (1 mm - 1 m) sono fondamentali per penetrare le regioni meno dense delle nubi e raccogliere dati sulla temperatura, l'umidità e le precipitazioni. I sensori a microonde passivi misurano l'energia emessa, mentre i sensori attivi, ad esempio i radar, inviano e valutano i segnali riflessi per ottenere informazioni sulla struttura e l'intensità delle precipitazioni.

Altre caratteristiche peculiari dei satelliti meteorologici sono legate alla loro capacità di visione e gli errori introdotti nell'osservazione:

- **Field of View (FOV)**: l'angolo solido entro il quale il sensore può raccogliere dati.
- **Dwell Time**: il tempo durante il quale il sensore osserva un singolo punto sulla superficie terrestre. Maggiore è il dwell time, migliore è la qualità dei dati raccolti.
- **Rumori del Segnale**:
  - Rumore Termico, causato dal calore generato dai componenti elettronici del sensore.
  - Rumore di Quantizzazione, dovuto alla digitalizzazione del segnale analogico.
  - Rumore di Background, provocato dalle emissioni ambientali non correlate al target osservato.

2.1.2 Riflettanza

La **riflettanza** $\mathcal{R}$ è una misura utilizzata per descrivere la frazione della radiazione incidente che viene riflessa da una superficie. Nel contesto delle osservazioni satellitari nei canali del visibile, la riflettanza viene calcolata come il rapporto tra la radianza riflessa dalla superficie e l'irradianza incidente sulla superficie.

Formalmente, la riflettanza (Fig.2.2) è definita come:

$$\mathcal{R}(\theta_i, \phi_i, \theta_r, \phi_r) = \frac{L_r(\theta_r, \phi_r)}{E_i(\theta_i, \phi_i)}$$

dove:

- $\mathcal{R}(\theta_i, \phi_i, \theta_r, \phi_r)$ è la riflettanza bidirezionale,
- $L_r(\theta_r, \phi_r)$ è la radianza riflessa dalla superficie nella direzione (θ_r, ϕ_r),
- $E_i(\theta_i, \phi_i)$ è l'irradianza incidente sulla superficie dalla direzione (θ_i, ϕ_i).

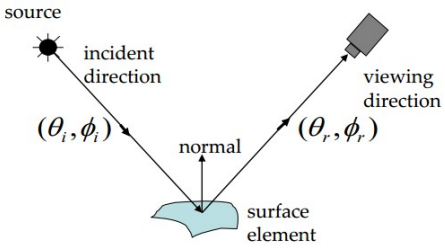


Figura 2.2: Misura di riflettanza

La funzione di distribuzione bidirezionale della riflettanza f_r (BRDF, Bidirectional Reflectance Distribution Function) descrive come la riflettanza varia con l'angolo di incidenza e l'angolo di osservazione. Essa è definita come:

$$f_r(\theta_i, \phi_i, \theta_r, \phi_r) = \frac{dL_r(\theta_r, \phi_r)}{dE_i(\theta_i, \phi_i)}$$

La BRDF esprime la radianza riflessa dL_r per unità di irradianza incidente dE_i, per ciascuna coppia di angoli di incidenza e riflessione.

Per una superficie Lambertiana, ossia che riflette la radiazione in maniera isotropa, la riflettanza è costante in tutte le direzioni e si può esprimere come:

$$\mathcal{R}_{\text{Lambertiana}} = \frac{1}{\pi}$$

L'albedo sferico α rappresenta il rapporto tra il flusso di energia riflessa dall'intero pianeta e l'energia incidente. È dato da:

$$\alpha = \frac{1}{\pi} \int_0^{2\pi} \int_0^{\pi/2} \mathcal{R}(\theta, \phi) \cos\theta \sin\theta \, d\theta \, d\phi$$

La riflettanza è quindi una misura fondamentale per l'analisi delle proprietà ottiche delle superfici terrestri e la BRDF è uno strumento cruciale per modellare le caratteristiche direzionali di riflessione.

2.1.3 Temperatura di Brillanza (BT)

La **temperatura di brillanza (BT)** è la temperatura che un corpo nero avrebbe se emettesse, a una data lunghezza d'onda, la stessa intensità di radiazione per unità di superficie e per unità di angolo solido osservata da una sorgente. Essa può essere calcolata utilizzando la legge di Planck per la radiazione elettromagnetica:

$$B(\lambda, T) = \frac{2hc^2}{\lambda^5} \cdot \frac{1}{e^{\frac{hc}{\lambda kT}} - 1}$$

dove:

- $B(\lambda, T)$ è la radianza spettrale emessa dalla superficie per lunghezza d'onda (λ) e temperatura (T).
- h è la costante di Planck.
- c è la velocità della luce nel vuoto.
- k è la costante di Boltzmann.
- T è la temperatura della superficie.

La temperatura di brillanza può essere ottenuta invertendo la legge di Planck risolvendola per T:

$$T = \frac{hc}{k \cdot \lambda \cdot \ln\left(\frac{2hc^2}{B(\lambda,T)\cdot\lambda^5} + 1\right)}$$

La temperatura di brillanza fornisce quindi una misura della temperatura di un oggetto basata sull'intensità della radiazione elettromagnetica emessa a una specifica frequenza o lunghezza d'onda, assumendo che si comporti come un corpo nero ideale. Viene utilizzata principalmente in astronomia per caratterizzare stelle, pianeti e altre sorgenti celesti, nei satelliti meteorologici per monitorare la temperatura della superficie terrestre e delle nuvole, e in telecomunicazioni e radar per descrivere la radiazione di fondo. Differisce dalla temperatura effettiva, che considera l'emissione su tutte le frequenze, e dalla temperatura cinetica, che misura l'energia cinetica media delle particelle di un gas senza considerare la radiazione emessa.

2.2 SEVIRI

2.2.1 Descrizione del satellite EUMETSAT - MSG

Il **Meteosat Second Generation (MSG)** (Fig.2.3) rappresenta una serie avanzata di satelliti meteorologici sviluppati congiuntamente dall'Agenzia Spaziale Europea (ESA) e dall'Organizzazione Europea per l'Esercizio dei Satelliti Meteorologici (EUMETSAT) [14]. Questo programma è nato con l'obiettivo di migliorare le capacità di osservazione meteorologica rispetto ai satelliti Meteosat di prima generazione, fornendo dati più dettagliati spazialmente e a maggior frequenza temporale. La serie MSG comprende tre satelliti identici: MSG-1, MSG-2 e MSG-3. La progettazione di questi satelliti prevede un periodo operativo di 7 anni ciascuno con il lancio di MSG-1 avvenuto nel 2002, seguito da MSG-2 circa 18 mesi dopo.

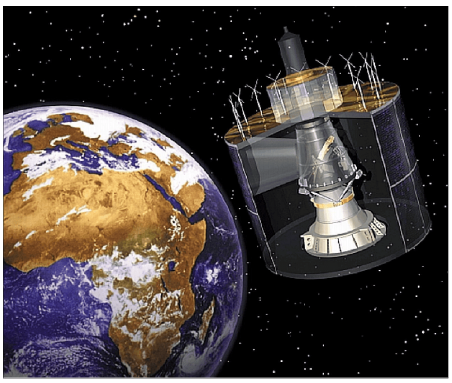


Figura 2.3: MSG in orbita equipaggiato con SEVIRI

I satelliti MSG operano in un'**orbita geostazionaria**, mantenendo una posizione fissa rispetto alla superficie terrestre, centrata su o° di longitudine. Questa caratteristica è essenziale per l'osservazione continua delle condizioni meteorologiche su vaste aree, in particolare sull'Europa e sull'Africa (Fig.2.4).

Un aspetto tecnico importante dei satelliti MSG è il loro sistema di **stabilizzazione tramite rotazione** attorno al proprio asse (spin-stabilized). La rotazione del satellite consente inoltre di scansionare continuamente la Terra, catturando immagini di alta qualità e risoluzione ogni **15 minuti**, offrendo un miglioramento rispetto ai 30 minuti dei satelliti Meteosat di prima generazione.

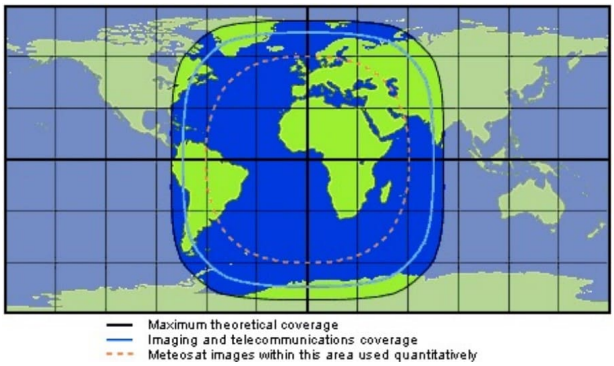


Figura 2.4: Field Of View (FOW) di SEVIRI MSG

I satelliti MSG offrono 12 canali spettrali invece dei 3 di misurazione della prima generazione. Ogni canale è ottimizzato per rilevare specifiche bande di frequenza della luce visibile e dell'infrarosso, facilitando così l'analisi di diverse caratteristiche atmosferiche e terrestri. La quantizzazione delle immagini è stata migliorata da 8 a 10 bit per pixel, consentendo una maggiore precisione nella rappresentazione dei dati raccolti.

In termini di **risoluzione spaziale**, i satelliti MSG offrono distanze di campionamento di **3 km** al nadir per tutti i canali, con un canale aggiuntivo nel visibile ad alta risoluzione (HRV) che raggiunge 1 km. Questo rappresenta un significativo miglioramento rispetto ai 5 km e 2,5 km rispettivamente dei satelliti di prima generazione.

2.2.2 Principio di funzionamento SEVIRI

Il cuore tecnologico dei satelliti MSG è rappresentato dallo strumento **SEVIRI (Spinning Enhanced Visible and Infrared Imager)** [15][16][17]. SEVIRI è progettato per fornire immagini ad alta risoluzione attraverso **12 canali spettrali** (Fig.2.5) che coprono una vasta gamma di frequenze, dal visibile all'infrarosso termico.

Channel no.		Characteristics of spectral band (μm)			Main gaseous absorber or window
		λ_{cen}	λ_{min}	λ_{max}	
1	VIS0.6	0.635	0.56	0.71	Window
2	VIS0.8	0.81	0.74	0.88	Window
3	NIR1.6	1.64	1.50	1.78	Window
4	IR3.9	3.90	3.48	4.36	Window
5	WV6.2	6.25	5.35	7.15	Water vapor
6	WV7.3	7.35	6.85	7.85	Water vapor
7	IR8.7	8.70	8.30	9.10	Window
8	IR9.7	9.66	9.38	9.94	Ozone
9	IR10.8	10.80	9.80	11.80	Window
10	IR12.0	12.00	11.00	13.00	Window
11	IR13.4	13.40	12.40	14.40	Carbon dioxide
12	HRV	Broadband (about 0.4 – 1.1)			Window/water vapor

Figura 2.5: Caratteristiche spettrali dei canali di SEVIRI: lunghezza d'onda centrale, minima e massima dei canali e caratteristica del canale in assorbimento o finestra [14]

Questa ampia gamma di canali permette a SEVIRI di raccogliere dati dettagliati su diverse proprietà atmosferiche, tra cui la composizione e la struttura delle nuvole, il contenuto di vapore acqueo e la temperatura superficiale. I canali nel visibile e nel vicino infrarosso sono particolarmente utili per l'osservazione della copertura nuvolosa e delle caratteristiche della superficie terrestre, mentre i canali nell'infrarosso termico sono essenziali per misurare la temperatura delle nuvole e della superficie, nonché per rilevare la presenza di vapore acqueo nell'atmosfera. Nella seguente Fig.2.6, un esempio della visualizzazione full disk di SEVIRI nei vari canali:

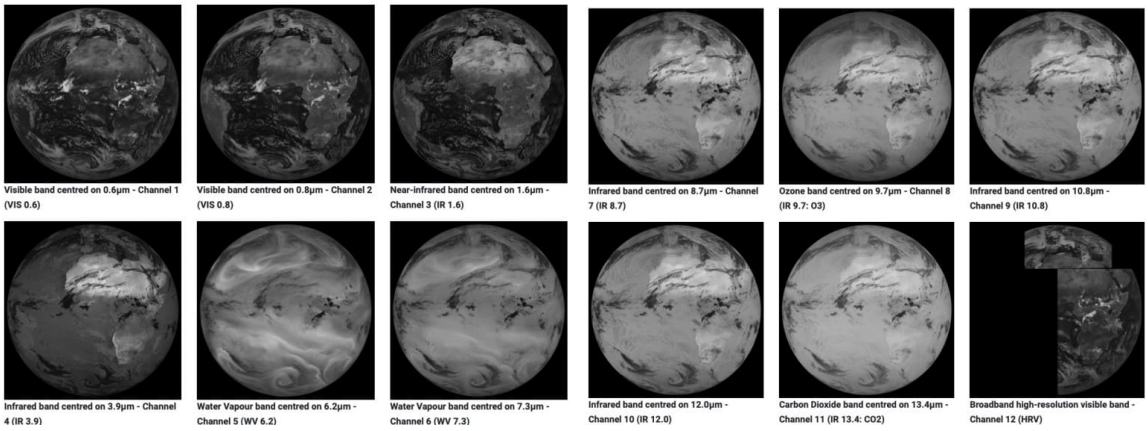


Figura 2.6: Esempio di visualizzazione full disk dei dati di SEVIRI nei suoi canali di osservazione, in scala di grigi dove il chiaro corrisponde ad alta riflettanza e bassa temperatura di brillanza, rappresentando quindi le nubi

La frequenza di osservazione di 15 minuti di SEVIRI permette di monitorare in modo continuo e dettagliato i cambiamenti atmosferici, facilitando l'identificazione tempestiva di fenomeni meteorologici estremi come tempeste e uragani. La **scansione terrestre** viene infatti ottenuta da una scansione bidimensionale della Terra combinando la rotazione del satellite e la rotazione dello specchio di scansione (Fig.2.7), secondo due modalità:

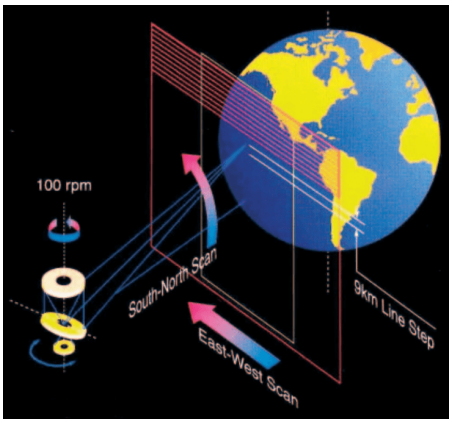


Figura 2.7: Principio di scansione di SEVIRI

- La **scansione rapida** (scansione lineare) viene eseguita da est a ovest grazie alla rotazione del satellite attorno all'asse di rotazione (frequenza di rotazione pari a 100 rotazioni al minuto). L'asse di rotazione è perpendicolare al piano orbitale ed è orientato lungo la direzione nord-sud.
- La **scansione ordinaria** viene eseguita da sud a nord mediante un meccanismo di scansione, che ruota lo specchio di scansione con passi di 125,8 μrad. Viene considerata una gamma totale di scansione di +/-5,5 gradi (corrispondente a 1527 linee di scansione) per coprire l'intervallo esteso di imaging terrestre di 22 gradi nella direzione sud-nord, rispettivamente 1249 linee di scansione per coprire l'intera Terra nel ciclo di ripetizione di base.

L'acquisizione dell'immagine completa del disco terrestre richiede circa 12 minuti. Successivamente, lo specchio di scansione ritorna alla posizione iniziale e si attiva un meccanismo di posizionamento del corpo nero nel percorso ottico ai fini della calibrazione. In seguito alla rimozione del corpo nero dalla posizione di calibrazione, si procede con l'osservazione della Terra, completando così il ciclo in 15 minuti.

2.3 DPR

2.3.1 Descrizione del satellite NASA - GPM

La NASA **Global Precipitation Measurement Mission (GPM)** è una costellazione internazionale di satelliti dedicata all'osservazione globale delle precipitazioni. Il suo scopo principale è misurare pioggia e neve con precisione, migliorando la comprensione dei cicli idrici ed energetici della Terra, perfezionando le previsioni degli eventi meteorologici estremi e fornendo dati vitali per una serie di applicazioni sociali e ambientali [18].

La missione, avviata in collaborazione con l'Agenzia di esplorazione aerospaziale del Giappone (JAXA) e altre agenzie spaziali internazionali, è stata lanciata il 27 febbraio 2014 dal Tanegashima Space Center, Giappone e nasce dall'esperienza della Tropical Rainfall Measuring Mission (TRMM), lanciata nel 1997, che ha evidenziato l'importanza di un'orbita non geosincrona per un monitoraggio più accurato degli uragani e delle piogge tropicali. Rispetto a TRMM, con GPM si introduce una migliore capacità di misurare piogge leggere ($< 0{,}5$ mm/h), precipitazioni solide e le proprietà microfisiche delle particelle precipitanti.

Il Satellite GPM Core Observatory è equipaggiato con il primo radar spaziale di precipitazione a **doppia frequenza Ku/Ka-band (DPR)** e un Radiometro a Microonde Multicanale GPM (GMI). Il DPR, costituito da un radar di precipitazione a banda Ka (KaPR) operante a 35,5 GHz e un radar di precipitazione a banda Ku (KuPR) a 13,6 GHz, offre misurazioni tridimensionali (Fig.2.8) della struttura delle precipitazioni e delle loro caratteristiche, misurando su una banda spaziale di larghezza pari a 245 km.

Parallelamente, il **radiometro a microonde multicanale GMI**, scansiona conicamente una fascia di 885 km con tredici canali che variano da 10 GHz a 183 GHz. Questo strumento sfrutta un insieme di frequenze ottimizzato nel corso degli ultimi due decenni per rilevare precipitazioni intense, moderata e leggere, utilizzando la differenza di polarizzazione su ogni canale come indicatore dello spessore ottico e del contenuto d'acqua dei sistemi di precipitazioni, fornendo dati accurati per la comprensione dei fenomeni meteorologici.

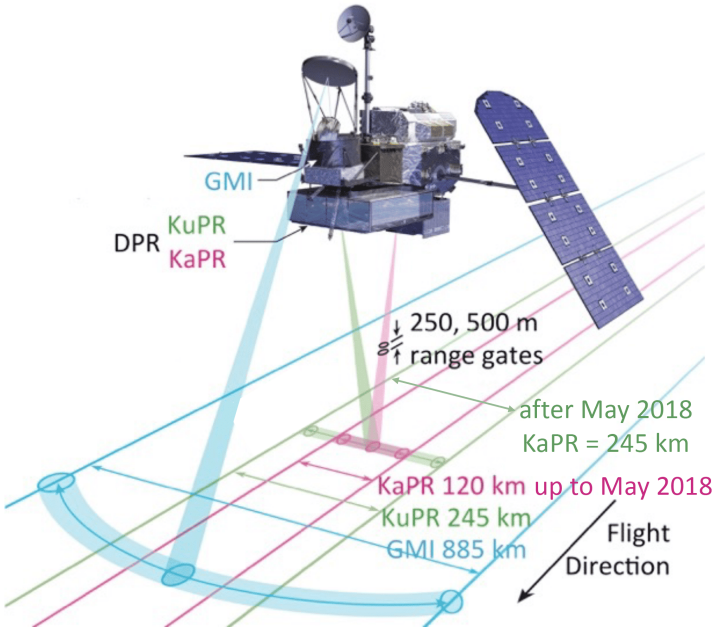


Figura 2.8: Scanning del satellite GPM equipaggiato con il DPR e il GMI

Le misurazioni precise e la copertura estesa di GPM offre quindi molteplici benefici scientifici e applicativi, che possono essere riassunti in questi punti:

- **Conoscenza del ciclo dell'acqua**: migliora la comprensione del ciclo dell'acqua terrestre e del suo legame con i cambiamenti climatici.
- **Strutture delle tempeste**: fornisce nuove intuizioni sulle strutture delle tempeste e sui processi atmosferici su larga scala.
- **Microfisica delle precipitazioni**: approfondisce la conoscenza delle proprietà microfisiche delle precipitazioni.
- **Previsioni meteorologiche**: potenzia le previsioni di eventi meteorologici estremi come cicloni tropicali, inondazioni, siccità e frane.
- **Gestione delle risorse e attivitò**: supporta le attività di gestione delle colture agricole, il monitoraggio delle risorse idriche e la sicurezza alimentare.

2.3.2 Principio di funzionamento DPR

Il **radar Dual-frequency Precipitation Radar (DPR)** [19] è uno degli strumenti principali a bordo del GPM-CO. Il DPR è composto da un radar di precipitazione a banda Ku (KuPR) e un radar di precipitazione a banda Ka (KaPR), e i dati raccolti forniscono osservazioni tridimensionali della pioggia e una stima accurata del tasso di precipitazione. Si riporta l'immagine di esempio di un ciclone extra-tropicale osservato sulle coste in Giappone il 10 marzo 2014, nella seguente Fig.2.9:

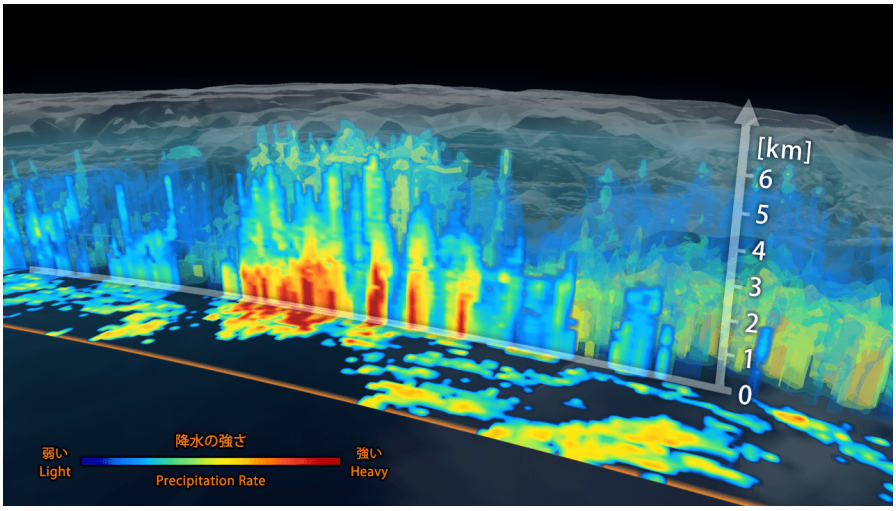


Figura 2.9: Osservazione 3D con il DPR di GPM del cicloce extra-tropicale del 10 marzo 2014 in Giappone

Item	KuPR	KaPR
Swath Width	245 km	120 km
Range Resolution	250 m	250/500 m
Spatial Resolution	5.2 km (Nadir at the height of 407 km)	
Beam Width	0.71° (Center Beam)	
Transmitter	128 Solid State Amplifiers	
Peak Transmit Power[1]	1012.0 W	146.5 W
Pulse Repetition Freq.[2]	4000 to 4500 Hz	
Pulse Width	two 1.6 μs pulses	two 1.6 μs pulses (matched beams) two 3.2 μs pulses (interlaced scans)
Beam Number	49	49 (25 in matched beams and 24 in interlaced scans)
Min. measurable rain rate	0.5 mm/h	0.2 mm/h
Beam matching error	Less than 1000 m	
Observable range	19 km to Surface (to −5 km near nadir)	
Dynamic range	From −5 dB below the system noise level to +5 dB above the nominal maximum surface echo level	
Receiver power accuracy	±1 dB	
Scan Angle	±17° Cross Track	±8.5° Cross Track
Frequencies	13.597 and 13.603 GHz	35.547 and 35.553 GHz
Bandwidth	14 MHz	
Max. Mass	472 kg	336 kg
Power (max)	446 W (orbit average)	344 W (orbit average)
Science Data Rate (max)	109 kbps	81 kbps
	(The Total of KuPR and KaPR is 190 kbps)	
Housekeeping Data Rate[3]	1 kbps (nominal)	

Figura 2.10: Scheda tecnica del DPR

I dati raccolti dal DPR nelle due bande Ku-Ka, tramite l'uso di algoritmi, permettono di migliorare la stima delle precipitazioni sia sulla terraferma che sull'oceano, sia di giorno che di notte. Ad esempio è possibile distinguere tra pioggia e neve utilizzando l'attenuazione differenziale tra le frequenze delle bande Ku e Ka, o

ottenere una maggiore sensibilità pari a 0.2 mm/h tramite la tecnica di frequenza di ripetizione dell'impulso variabile (VPRF) che consente l'aumento del numero di campioni su ciascun campo visivo istantaneo (IFOV). Si riporta in Fig.2.10 la scheda tecnica del DPR:

Il DPR misura [20] l'eco ricevuto dalle gocce di pioggia, con la **potenza ricevuta** $P_r(r)$ proporzionale al fattore di riflettività radar apparente $Z_{m0}(r)$:

$$P_r(r) = \frac{C|K|^2}{r^2} Z_{m0}(r)$$

dove C è la costante radar e K è una costante definita in funzione dell'indice di rifrazione complesso m delle particelle di scattering.

La correzione dell'attenuazione è fondamentale per ottenere il **fattore di riflettività radar effettivo** Z_e da Z_{m0}. Questo processo è descritto dall'equazione:

$$Z_e(r) = Z_m(r) \left[1 - 0.2 \ln(10) \beta \int_0^r \alpha(s) Z_m^\beta(s) ds\right]^{1/\beta}$$

dove

$Z_e(r)$ è il fattore di riflettività radar effettivo al raggio r. Questo è il valore corretto del fattore di riflettività radar Z_m dopo la compensazione dell'attenuazione.

$\alpha(s)$ è il coefficiente di attenuazione specifica del segnale radar per unità di lunghezza, in funzione delle proprietà del mezzo atmosferico.

β è un coefficiente empirico che descrive la relazione tra l'attenuazione e il fattore di riflettività radar.

s è la variabile di integrazione e rappresenta la distanza lungo il percorso del raggio radar.

$\int_0^r \alpha(s) Z_m^\beta(s) ds$ è l'integrale dell'attenuazione accumulata del segnale radar lungo il percorso dal punto di partenza (0) fino al raggio r.

Utilizzando i dati a doppia frequenza, il DPR può stimare meglio la distribuzione delle dimensioni delle gocce (DSD, Drop Size Distribution) e, di conseguenza, il **rate o intensità di precipitazione** R. La formula per il rate di precipitazione R è data da:

$$R = \int V(D) v(D) N(D)\, dD$$

dove $V(D)$ è il volume di una singola goccia con diametro D, $v(D)$ è la velocità di caduta della goccia con diametro D, e $N(D)$ è la funzione di distribuzione delle dimensioni delle gocce, che indica il numero di gocce per unità di volume e per unità di intervallo di diametro.

3 Tecniche di machine learning (ML)

3.1 Il machine learning

Il **Machine Learning (ML)** è una disciplina dell'**intelligenza artificiale (AI)** che si occupa della costruzione e dell'uso di modelli matematici e statistici per apprendere da un campione di dati e ottenere previsioni o decisioni automatizzate. A differenza degli algoritmi tradizionali, che richiedono istruzioni esplicite per trasformare un input in output, il ML utilizza tecniche di apprendimento automatico dai dati stessi [21].

L'algoritmo machine learning elabora i **dati di addestramento** per imparare le relazioni tra input e output, al fine di generalizzare ciò che è stato appreso a nuovi dati non etichettati. Con l'avanzamento delle tecnologie di memorizzazione e calcolo, è diventato possibile gestire e processare efficacemente **grandi volumi di dati**: questo aspetto è fondamentale per il machine learning, dato che la prestazione dei modelli è strettamente legata alla quantità e qualità dei dati a disposizione.

Il processo fondamentale alla base degli algoritmi implica l'ottimizzazione di un **modello parametrico**, di tipo statistico, utilizzando dati di addestramento, minimizzando una funzione di perdita che misura l'errore tra le previsioni del modello e i dati osservati, permettendo quindi di fare inferenze dai dati. I modelli possono essere predittivi, per ottenere previsioni, o descrittivi, per estrarre conoscenze dai dati.

Nell' **apprendimento supervisionato** si cerca di individuare una mappatura dall'input a un output i cui valori corretti sono forniti da un supervisore, mentre nell' **apprendimento non supervisionato** non c'è un supervisore ma si analizzano i soli dati di input senza etichettamento.

3.2 Alberi Decisionali (DTs)

Gli **alberi decisionali** [22] rappresentano uno dei metodi più utilizzati nell'apprendimento supervisionato, sia per la classificazione che per la regressione [23]. Il loro obiettivo è creare un modello in grado di prevedere il valore di una variabile target apprendendo regole decisionali semplici dalle caratteristiche dei dati.

Un modello ad alberi decisionali è costituito da una struttura di decisioni ad albero dove ogni nodo interno rappresenta una decisione, ogni ramo rappresenta l'esito di quella decisione e ogni foglia rappresenta un risultato finale o una classe. In dettaglio:

- **Nodo (Node)**: È un punto in cui l'albero prende una decisione su quale percorso seguire basandosi su un attributo specifico. Ogni nodo interno delinea una condizione (ad esempio, una soglia su una caratteristica) che divide i dati in sottoinsiemi.
- **Foglia (Leaf)**: È un nodo terminale che fornisce una previsione finale o una classificazione. Una foglia non ha ulteriori divisioni; rappresenta una delle possibili categorie o valori di output del modello.
- **Albero (Tree)**: È una struttura composta da nodi e foglie, partendo da un nodo radice (root node) che rappresenta l'intero dataset. L'albero viene costruito iterativamente dividendo il dataset in sottoinsiemi sempre più piccoli, fino a raggiungere i nodi foglia.

I **vantaggi** degli alberi decisionali sono molteplici:

- **Semplicità di comprensione e interpretazione**: gli alberi decisionali sono facili da comprendere e interpretare, in quanto la struttura a diagramma di flusso segue una logica facilmente accessibile, anche senza competenze specifiche.
- **Visualizzazione**: gli alberi possono essere visualizzati, permettendo l'osservazione e la comprensione dei percorsi decisionali.
- **Preparazione ridotta dei dati**: è richiesta poca preparazione dei dati rispetto ad altre tecniche in quanto non è necessario normalizzare i dati, creare variabili dummy (ovvero variabili numeriche fittizie che sostituiscono una variabile qualitativa) per la rappresentazione di categorie di dati, e alcuni algoritmi possono gestire direttamente i valori mancanti.
- **Efficienza nella predizione**: il costo dell'uso dell'albero per le predizioni è logaritmico rispetto al numero di dati utilizzati per addestrare l'albero, rendendolo efficiente anche per grandi dataset.

- **Gestione di diversi tipi di dati**: gli alberi decisionali possono gestire sia dati numerici che categorici, rendendoli versatili.
- **Capacità multi-output**: si possono gestire problemi a multi-output, dove devono essere predette più variabili di output.
- **Modello white box**: gli alberi decisionali forniscono trasparenza, poichè il processo decisionale è chiaro e può essere spiegato utilizzando la logica booleana. Questo contrasta con i modelli black box, come le reti neurali, dove il funzionamento interno non è facilmente interpretabile.
- **Validazione del modello**: è possibile validare un modello ad albero decisionale utilizzando test statistici [24], il che aiuta a valutare l'affidabilità e la significatività del modello.
- **Robustezza rispetto alle assunzioni del modello**: gli alberi decisionali funzionano bene anche se le loro assunzioni sono parzialmente violate dal processo generativo dei dati.

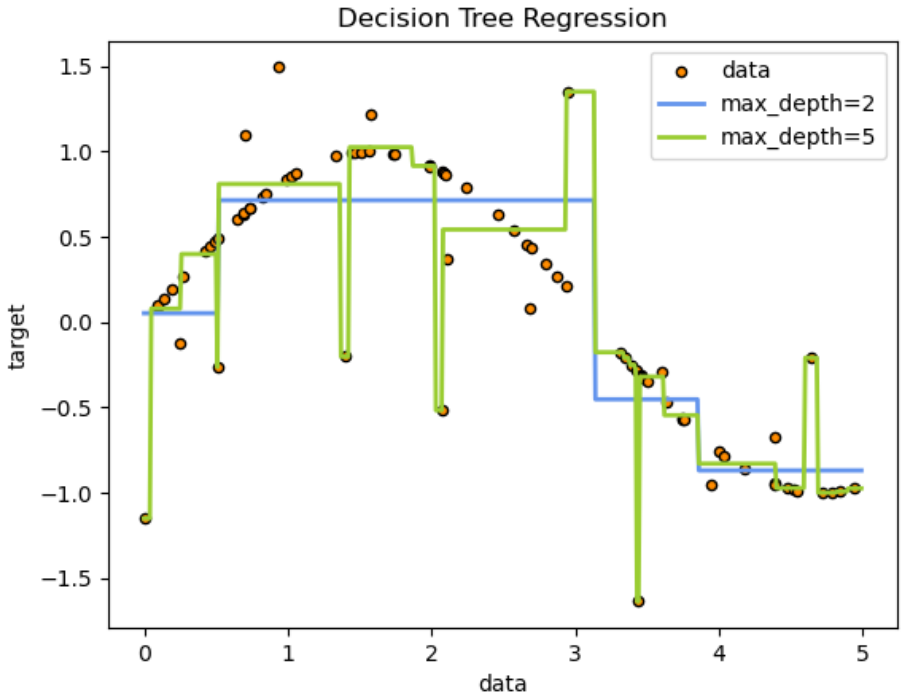


Figura 3.1: Esempio di Regressione con Alberi Decisionali: le predizioni non sono continue ma costanti a tratti

Tuttavia, gli alberi decisionali sono soggetti ad alcune **limitazioni**:

- **Eccessiva complessità (Overfitting)**: gli alberi decisionali possono creare alberi eccessivamente complessi che non generalizzano bene i dati: questo fenomeno è noto come overfitting. Per evitare il problema, sono necessari meccanismi come la potatura, la definizione del numero minimo di campioni richiesti in un nodo foglia, che rappresenta la decisione finale del processo, o la definizione della profondità massima dell'albero.
- **Instabilità**: si può avere instabilità poiché piccole variazioni nei dati possono portare alla generazione di alberi completamente diversi. Questo problema può tuttavia essere mitigato utilizzando alberi decisionali all'interno di un ensemble.
- **Predizioni non continue**: le predizioni non sono continue ma approssimazioni costanti a tratti come mostrato in Fig.3.1.
- **Complessità computazionale**: il problema di apprendere un albero decisionale ottimale è noto per essere NP-completo, ovvero di alta difficoltà computazionale. Di conseguenza, l'apprendimento si basa su algoritmi euristici come l'algoritmo greedy, dove decisioni localmente ottimali vengono prese a ciascun nodo. Tali algoritmi non possono garantire di restituire l'albero decisionale globalmente ottimale e per questo si effettua un ensemble learner tra più alberi, campionati casualmente.

- **Difficoltà nell'apprendimento di alcuni concetti**: ci sono concetti difficili da apprendere perché gli alberi decisionali non li esprimono facilmente, come ad esempio i problemi di classificazione non lineare XOR.
- **Bias verso classi dominanti**: gli alberi sono distorti sulle classi dominanti: è raccomandato pertanto di bilanciare il dataset prima dell'addestramento.

3.2.1 Problemi di classificazione

La **classificazione** rappresenta uno dei principali usi dell'analisi dei dati con machine learning, in quanto con l'aumento della disponibilità di dati e della potenza computazionale, le tecniche di classificazione sono diventate sempre più centrali in una vasta gamma di settori.

La classificazione è un'attività di apprendimento supervisionato in cui l'obiettivo è **assegnare una classe** o una categoria a un insieme di dati in base alle caratteristiche osservate.

La classificazione offre una serie di **vantaggi** significativi:

- **Interpretabilità**: molti algoritmi di classificazione producono modelli che possono essere interpretati facilmente, facendo uso di regole decisionali semplici.
- **Adattabilità**: gli algoritmi di classificazione possono essere adattati a una vasta gamma di problemi, dai semplici problemi binari ai complessi problemi multiclasse.
- **Efficienza computazionale**: molte tecniche di classificazione sono computazionalmente efficienti e possono essere applicate anche a grandi set di dati.
- **Versatilità**: la classificazione può essere utilizzata per una varietà di compiti, tra cui riconoscimento di pattern, clustering e previsione.

Tuttavia, ci sono anche alcune sfide associate alla classificazione nell'apprendimento automatico:

- **Overfitting**: gli algoritmi di classificazione possono adattarsi eccessivamente ai dati di addestramento, producendo modelli che non generalizzano bene ai nuovi dati.
- **Sensibilità ai dati di addestramento**: i risultati della classificazione possono variare significativamente in base ai dati di addestramento utilizzati e piccole variazioni nei dati di addestramento possono portare a risultati molto diversi.
- **Gestione delle classi sbilanciate**: nei casi in cui le classi di dati sono sbilanciate, con un numero significativamente maggiore di campioni in una classe rispetto alle altre: questo può portare a modelli di classificazione parziali o distorti.
- **Interpretabilità limitata**: alcuni algoritmi di classificazione producono modelli complessi difficili da interpretare, come le reti neurali profonde.

3.2.2 Problemi di regressione

Gli alberi decisionali possono essere applicati anche a problemi di regressione. La regressione con alberi decisionali è utile nei casi in cui si desidera prevedere valori continui basati su delle caratteristiche osservate: viene diviso ricorsivamente lo spazio delle caratteristiche in regioni, assegnando a ciascuna regione un valore di output che è la media dei valori target nei campioni di addestramento in quella regione.

Gli alberi decisionali di regressione condividono molte delle stesse caratteristiche e dei vantaggi dei loro omologhi di classificazione, come l'interpretabilità e la capacità di gestire dati misti di natura sia numerica che categorica. Possono inoltre soffrire degli stessi problemi di overfitting e sensibilità ai dati di addestramento e quindi scelta del giusto albero decisionale di regressione e la messa a punto dei suoi parametri sono cruciali per ottenere modelli accurati e generalizzabili.

3.2.3 Formulazione matematica dei modelli ad Alberi Decisionali

Dati i vettori di addestramento $x_i \in R^n$, con $i = 1, \ldots, l$ e un vettore di etichette $y \in R^l$, un albero decisionale partiziona ricorsivamente lo spazio delle caratteristiche in modo tale che i campioni con le stesse etichette o valori target simili siano raggruppati insieme.

Sia Q_m il set di dati al nodo m con n_m campioni. Per ogni candidato di divisione $\theta = (j, t_m)$ composto da una caratteristica j e una soglia t_m, si partizionano i dati in sottoinsiemi $Q_{m_{left}}(\theta)$ e $Q_{m_{right}}(\theta)$ come segue, dove $Q_{m_{right}}(\theta)$ è l'insieme complementare di $Q_{m_{left}}(\theta)$:

$$Q_{m_{left}}(\theta) = \{(x, y) | x_j \leq t_m\}$$
$$Q_{m_{right}}(\theta) = Q_m \setminus Q_{m_{left}}(\theta)$$

La formula per valutare la qualità di una suddivisione θ di un insieme Q_m in due sottoinsiemi è:

$$G(Q_m, \theta) = \frac{n_{m_{left}}}{n_m} H(Q_{m_{left}}(\theta)) + \frac{n_{m_{right}}}{n_m} H(Q_{m_{right}}(\theta))$$

dove $G(Q_m, \theta)$ è la misura dell'impurità combinata delle due partizioni $Q_{m_{left}}(\theta)$ e $Q_{m_{right}}(\theta)$ dell'insieme Q_m dopo la suddivisione basata su θ, n_m è il numero totale di elementi nell'insieme Q_m, $n_{m_{left}}$ è il numero di elementi nel sottoinsieme $Q_{m_{left}}(\theta)$, $n_{m_{right}}$ è il numero di elementi nel sottoinsieme $Q_{m_{right}}(\theta)$, $H(Q_{m_{left}}(\theta))$ è la misura dell'impurità del sottoinsieme $Q_{m_{left}}(\theta)$, $H(Q_{m_{right}}(\theta))$ è la misura dell'impurità del sottoinsieme $Q_{m_{right}}(\theta)$.

La formula $G(Q_m, \theta)$ calcola l'impurità media ponderata delle due partizioni $Q_{m_{left}}(\theta)$ e $Q_{m_{right}}(\theta)$. Ogni sottoinsieme contribuisce all'impurità totale in proporzione alla sua dimensione rispetto all'insieme originale Q_m, permettendo di valutare se una particolare suddivisione θ è utile per ridurre l'impurità complessiva, un passo chiave nella costruzione di alberi decisionali.

Si selezionano quindi i parametri che minimizzano l'impurità:

$$\theta^* = \arg\min_{\theta} G(Q_m, \theta)$$

Si ripete per i sottoinsiemi $Q_{m_{left}}(\theta^*)$ e $Q_{m_{right}}(\theta^*)$ fino a quando non viene raggiunta la profondità massima consentita, $n_m <$ min_samples o $n_m = 1$.

Se un target è un risultato di **classificazione** che assume valori 0,1,...,K-1, per il nodo m, si definisce:

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k)$$

come la proporzione delle osservazioni di classe k nel nodo m. Se m è un nodo terminale, la funzione di predizione Scikit-Learn *predict_proba* per questa regione viene impostato su p_{mk}. Le misure comuni di impurità sono le seguenti:

Gini:

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

Perdita Logaritmica:

$$H(Q_m) = -\sum_k p_{mk} \log(p_{mk})$$

Se il target è un valore continuo e quindi un problema di **regressione**, allora per il nodo m, i criteri comuni da minimizzare per determinare le posizioni per le divisioni future sono l'Errore Quadratico Medio (MSE o errore L2), la devianza di Poisson e la media.

3.3 Random Forest (RF)

Una **Random Forest** [25] è un metodo di apprendimento ensemble costituito da alberi decisionali indipendenti. Ogni albero viene addestrato su un sottoinsieme casuale del set di dati di addestramento, utilizzando tecniche di **bagging** e **attribute sampling** per garantire l'indipendenza tra gli alberi e migliorare la qualità del modello complessivo.

Il **bagging** (bootstrap aggregating) consiste nell'addestrare ogni albero decisionale su un sottoinsieme casuale degli esempi nel set di addestramento. In altre parole, ogni albero nella Random Forest è addestrato su un sottoinsieme diverso di esempi, come illustrato nella Tabella 3.1.

Esempi di Addestramento	#1	#2	#3	#4	#5	#6
Set originale	1	1	1	1	1	1
Albero 1	1	1	0	2	1	1
Albero 2	3	0	1	0	2	0
Albero 3	0	1	3	1	0	1

Tabella 3.1: Distribuzione di sei esempi di addestramento tra tre alberi decisionali.

In generale, ogni albero decisionale è addestrato su un numero totale di esempi pari al numero di esempi nel set di addestramento originale, ma con sostituzione. Questo significa che alcuni esempi possono essere ripetuti più volte all'interno dello stesso albero.

L'**attribute sampling** implica che, invece di cercare la condizione migliore su tutte le caratteristiche disponibili, solo un sottoinsieme casuale di caratteristiche viene testato ad ogni nodo. La Fig.3.2 illustra questo concetto:

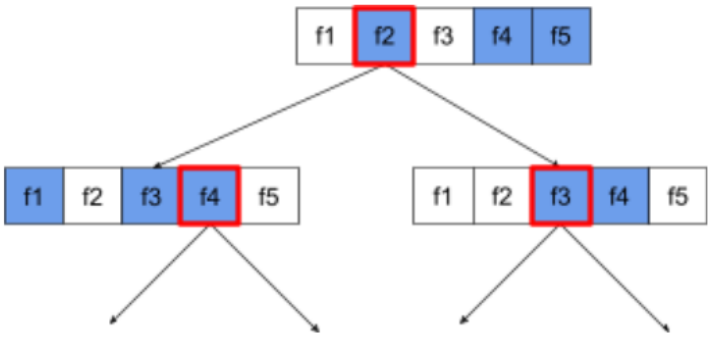


Figura 3.2: Esempio di attribute sampling in un albero decisionale

Il rapporto di campionamento delle caratteristiche è un iperparametro di regolarizzazione importante. Nella maggior parte delle implementazioni, il valore predefinito è 1/3 delle caratteristiche per la regressione e la radice quadrata del numero totale di caratteristiche per la classificazione.

Gli alberi decisionali in una Random Forest sono addestrati **senza potatura**, il che produce alberi molto complessi con scarsa qualità predittiva individuale. Tuttavia, l'ensemble di questi alberi migliora la qualità predittiva complessiva del modello. L'accuratezza di addestramento di una Random Forest è generalmente molto alta e questo non indica un sovradattamento (**overfitting**).

Il rumore casuale migliora la qualità di una Random Forest grazie all'indipendenza relativa tra gli alberi decisionali. La Fig.3.3 mostra un confronto tra le previsioni di un singolo albero e una Random Forest su un problema bidimensionale con un pattern ellittico.

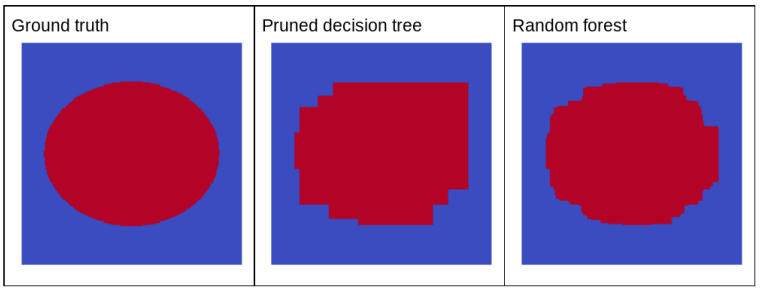


Figura 3.3: Previsioni pattern ellittico di un singolo albero vs. Random Forest

In conclusione, le Random Forests sono potenti modelli di apprendimento automatico che sfruttano la randomizzazione per migliorare la qualità delle previsioni. Le tecniche di bagging e attribute sampling garantiscono l'indipendenza degli alberi decisionali, correggendo il sovradattamento degli alberi individuali. L'aggiunta di più alberi migliora quasi sempre la qualità del modello, senza rischio di overfitting.

3.4 Gradient Boosting (GB)

3.4.1 Algoritmo GB

Il **gradient boosting** [26] è una metodologia applicata sopra un altro algoritmo di machine learning, coinvolgendo due tipi di modelli:

- Un modello di machine learning "debole", definito da un unico albero decisionale.
- Un modello di machine learning "forte", composto da un insieme di alberi decisionali e quindi multipli di modelli deboli.

Nel gradient boosting, ad ogni passo, un nuovo modello debole è addestrato per **prevedere l'errore del modello forte corrente**. In generale, l'errore previsto è la differenza tra la previsione e un'etichetta regressiva: il modello debole, associato all'errore, è quindi aggiunto al modello forte con un segno negativo per ridurre l'errore del modello forte.

Il gradient boosting è iterativo, secondo la seguente formula:

$$F_m(x) = F_{m-1}(x) - \gamma \cdot f_m(x)$$

dove $F_m(x)$ è il modello forte al passo m, $f_m(x)$ è il modello debole al passo m e γ è il tasso di apprendimento.

Questa operazione si ripete fino a quando non viene soddisfatto un criterio di arresto, come un numero massimo di iterazioni, oppure se il modello forte va incontro ad overfitting.

3.4.2 Esempio

Per comprendere il metodo implicito nell'algoritmo di Gradient Boosting si analizza il seguente **esempio**, per il quale si utilizza il dataset rappresentato in Fig.3.4.

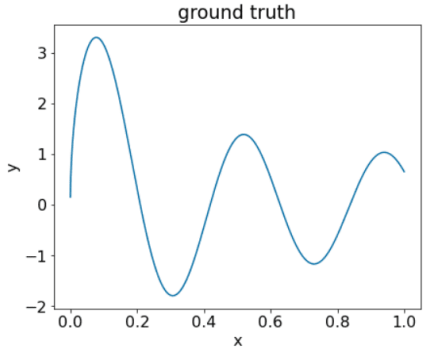


Figura 3.4: Dataset di esempio [27]

Si esegue un'iterazione dell'algoritmo di gradient boosting, utilizzando alberi decisionali come modelli deboli, ottenendo i plot in Fig.3.5:

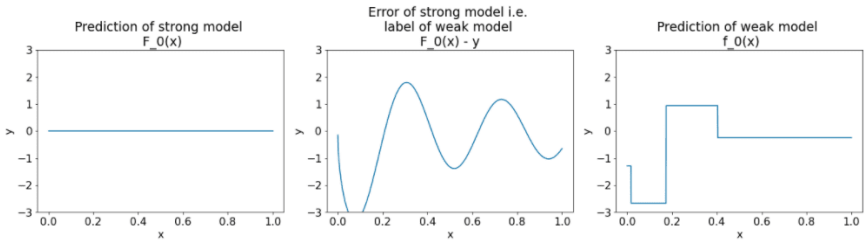


Figura 3.5: Plot modelli forti e deboli dopo la prima iterazione [27]

Il grafico sulla sinistra mostra la predizione iniziale del modello forte $F_0(x)$ come una linea orizzontale costante a $y = 0$. Il grafico centrale rappresenta gli errori del modello forte rispetto ai valori target reali, evidenziando i residui oscillanti. Il grafico sulla destra illustra la predizione del modello debole $f_0(x)$, che cerca di correggere questi residui con una funzione a pezzi costanti.

Il primo modello debole sta apprendendo una rappresentazione approssimativa dell'etichetta e si concentra principalmente sulla parte sinistra dello spazio delle caratteristiche, la parte con la maggior variazione e quindi con l'errore più alto.

Effettuando una seconda iterazione, in Fig.3.6:

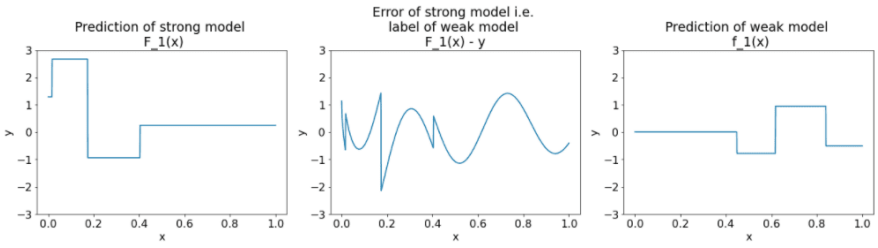


Figura 3.6: Plot modelli forti e deboli dopo la seconda iterazione [27]

Si nota che il modello forte ora contiene la previsione del modello debole dell'iterazione precedente e il nuovo errore del modello forte è leggermente più piccolo. Infine la nuova previsione del modello debole si concentra ora sulla parte destra dello spazio delle caratteristiche.

Si esegue infine l'algoritmo per altre 8 iterazioni (Fig.3.7):

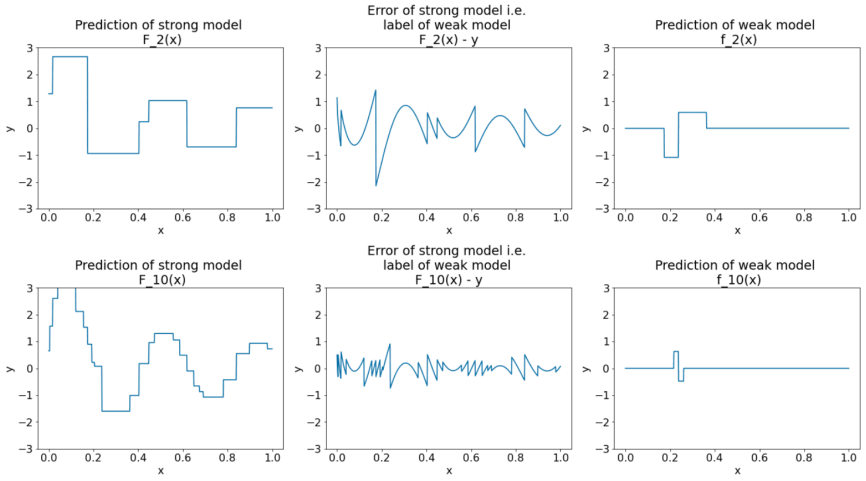


Figura 3.7: Plot modelli forti e deboli dopo la terza e decima iterazione [27]

I grafici dell'esempio mostrano quindi il funzionamento dell'algoritmo di Gradient Boosting e evidenziano come il modello forte gradualmente si avvicini al dataset iniziale ad ogni iterazione successiva, mentre la previsione del modello debole diventa gradualmente meno importante.

3.4.3 Pro e Contro degli alberi con Gradient Boosting

L'uso di algoritmi Gradient Boosting offre diversi vantaggi:

- essendo basato su alberi decisionali, supporta funzionalità numeriche nativamente senza richiedere necessariamente una normalizzazione o una gestione dei dati mancanti nel dataset di input.
- gli alberi con Gradient Boosting hanno iperparametri predefiniti che spesso forniscono ottimi risultati. Tuttavia, ottimizzare questi iperparametri può migliorare notevolmente il modello.
- i modelli ad albero con Gradient Boosting sono generalmente compatti, occupando minor numero di nodi e memoria, e veloci da eseguire.

Ed alcuni svantaggi da considerare:

- gli alberi decisionali devono essere addestrati in sequenza, il che può rallentare notevolmente l'addestramento. Tuttavia, il rallentamento è in parte compensato dalle dimensioni ridotte degli alberi decisionali.
- gli alberi con gradient boosting non possono imparare e riutilizzare le rappresentazioni interne. Ogni albero decisionale deve quindi imparare nuovamente i pattern del set di dati, portando a minori prestazioni con alcuni dataset specifici in imput.

3.5 Scikit-Learn

Scikit-learn [28] è una libreria open-source di machine learning per il linguaggio di programmazione Python. È progettata per offrire un'ampia gamma di algoritmi di apprendimento automatico, tecniche di pre-elaborazione dei dati e valutazione delle prestazioni dei modelli, rendendo più accessibile e conveniente l'implementazione di soluzioni di machine learning.

Gli algoritmi di Random Forest e Gradient Boosting sono ampiamente utilizzati in scikit-learn per problemi di classificazione e regressione.

Per utilizzare il Random Forest, si può importare la classe *RandomForestClassifier* o *RandomForestRegressor*, mentre per il Gradient Boosting si usano *GradientBoostingClassifier* o *GradientBoostingRegressor*. Dopo aver importato le rispettive classi, è sufficiente creare un'istanza dell'algoritmo, adattarlo ai dati con il metodo *fit* e fare previsioni con *predict*.

Di seguito sono elencati i principali **parametri di regolazione** per entrambi gli algoritmi, sia di Random Forest che di Gradient Boosting:

- **random_state**: Controlla la casualità del bootstrapping, ovvero il ricampionamento dei dati utilizzati per costruire gli alberi. Impostare questo parametro garantisce la riproducibilità dei risultati.
- **n_jobs**: Permette di indicare il numero di CPU per l'esecuzione in parallelo.
- **n_estimators**: Il numero di alberi nella foresta. Aumentare il numero di alberi può migliorare le prestazioni del modello, ma aumenta anche il costo computazionale.
- **max_depth**: La profondità massima di ciascun albero. Limitare la profondità dell'albero aiuta a prevenire l'overfitting, ovvero la perdita della capacità di generalizzare su dati nuovi a causa di un apprendimento troppo profondo di dettagli e rumore del dataset.
- **min_samples_leaf**: Il numero minimo di campioni richiesti per essere presenti in una foglia. Questo parametro aiuta a regolare l'albero e prevenire l'overfitting.
- **max_features**: La frazione di features da considerare quando si cerca la migliore suddivisione. Utilizzare una frazione delle caratteristiche migliora la stabilità del modello.
- **max_samples**: La frazione del campione totale da utilizzare per creare ogni albero.

4 Analisi del dataset: statistica e preparazione al ML

4.1 Obiettivo del modello machine learning

L'obiettivo del lavoro consiste nella **stima di intensità di precipitazione da satellite**, utilizzando i dati di radianza misurati da SEVIRI nei suoi 11 canali di osservazione e l'intensità di precipitazione ottenuta da DPR.

Ottenere una stima affidabile dell'intensità di precipitazione tramite strumenti satellitari [29] è fondamentale per **aumentare la capacità di descrivere quantitativamente i fenomeni di precipitazione** in aree, come oceani o territori ad orografia complessa [30] [31], che non possono essere adeguatamente coperte da servizi di stazioni pluviometriche da terra o da una rete radar, o nelle quali non è ancora presente un sistema di osservazione. Risulterebbe inoltre una soluzione, oltre che più versatile, anche più economica rispetto ad altri servizi di osservazione, compresi sistemi radar dedicati sempre da satellite.

Lo svantaggio di tale tecnica è sicuramente una minor accuratezza nelle stime, in quanto non c'è una relazione lineare e diretta tra le misure di radianza satellitari e l'intensità di precipitazione. Infatti è possibile riconoscere, tra le osservazioni nei vari canali satellitari e i fenomeni di precipitazione, soltanto deboli relazioni non lineari [32] e difficilmente interpretabili con un approccio puramente fisico, risultando così particolarmente indicato utilizzare **algoritmi in machine learning** [33], su un campione di dati sufficientemente grande e significativo.

In questo contesto di utilizzo si è deciso di utilizzare algoritmi basati su **alberi decisionali**, che consentono una relativa semplicità di implementazione e robustezza degli algoritmi, facendo quindi affidamento su metodi come le Random Forest, ottimo come classificatore [34], oppure, valutandone la validità in alternativa alla Random Forest, il Gradient Boosting [35].

Il lavoro segue le fasi rappresentate in Fig.4.1:

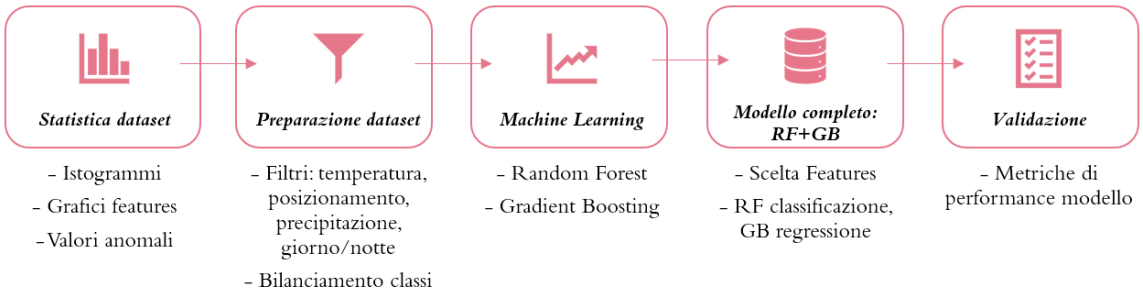


Figura 4.1: Schema del lavoro, suddiviso nelle sue fasi dall'elaborazione del dataset alla validazione del modello

Si comincia con un'analisi statistica dei dataset di SEVIRI o del DPR, li si preparano attraverso dei filtri opportuni e la realizzazione di features che saranno utilizzate per addestrare il modello in machine learning, con gli algoritmi di Random Forest e Gradient Boosting. Si conclude con dei test statistici e la validazione del modello.

4.2 File sorgente dei dataset

L'intero **dataset utilizzato** è stato fornito pre-elaborato dall'ISAC CNR di Roma. Esso è costituito da un file Matlab con estensione .mat, per ogni intervallo di 15 minuti, per l'intero anno 2017, contenenti due variabili: MSGdata_UNET e DPRdata_UNET. La caratteristica del dataset è la **coincidenza temporale e spaziale** tra le osservazioni di SEVIRI e quelle di DPR.

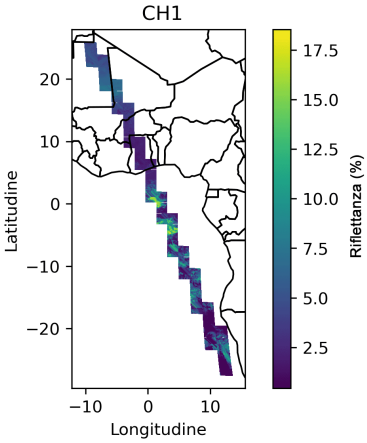
Per ogni file .mat, si hanno N immagini 64x64 pixel della superficie osservata in coincidenza dal DPR sul satellite GPM Core Observatory e da SEVIRI su MSG. La variabile **MSGdata_UNET** è una matrice di dimensione (64x64x16xN) che contiene informazioni relative a 16 valori per gli undici canali di SEVIRI (espressi in riflettanza nel visibile e in temperatura di brillanza nell'infrarosso), latitudine e longitudine, tipo di superficie (se oceano o terra), altezza del top della nube da un prodotto di MSG e orario di osservazione. La variabile **DPRdata_UNET** contiene una matrice di dimensione (64x64x1xN) relativa alle intensità di precipitazione, in accordo con il prodotto di precipitazione 2B-CMB GPM V06 sviluppato dalla NASA per il DPR [36].

Si verificano i valori numerici del DPR tenendo conto che valori negativi, associati ad un errore di stima dell'algoritmo, risultano non validi e sono da escludere dal dataset insieme a tutti i valori ad essi associati in coincidenza misurati da SEVIRI.

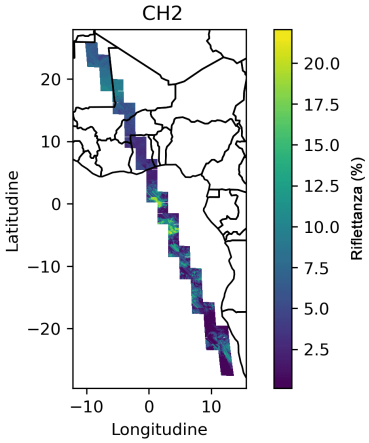
Le immagini satellitari contenute nei file sono dati geospaziali e per la loro lettura e analisi in Python, è possibile utilizzare librerie specifiche come osgeo [37] e satpy [38]. Osgeo (Open Source Geospatial Foundation) offre strumenti utili alla gestione di dati geospaziali. In particolare la libreria gdal al suo interno permette di leggere e scrivere una vasta gamma di formati di file georeferenziati. Satpy è specializzata nell'analisi di dati satellitari meteorologici fornendo funzionalità per caricare, manipolare e visualizzare i dati satellitari nei vari formati.

La prima fase per la gestione del dataset consiste nella conversione dal formato .mat gestibile attraverso Matlab in matrici numpy gestibili tramite script Python. Si è scelto di **utilizzare Python** in quanto è un linguaggio open source ampiamente utilizzato per la sua semplicità e versatilità. Data la divisione in file per ogni intervallo temporale di 15 minuti, si procede nella creazione di matrici uniche per ogni variabile, ovvero ogni canali di SEVIRI e il prodotto di precipitazione del DPR, per tutto l'anno 2017, al fine di poter gestire la fase di preparazione al modello in maniera più agevole.

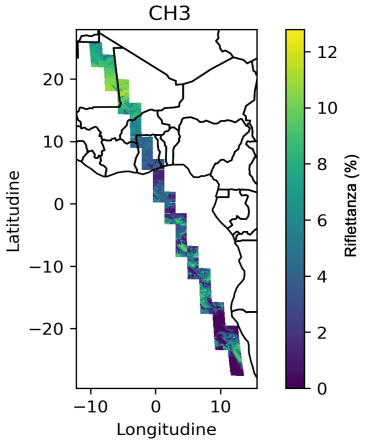
Si riportano di seguito le immagini (Fig.4.2) relative agli 11 canali di SEVIRI, con le prime 3 nel visibile (da Fig.4.2 (a) a Fig.4.2 (c)) e le restanti 8 nell'infrarosso (da Fig.4.2 (d) a Fig.4.2 (k)), e della precipitazione del DPR, per un intervallo temporale di 15 minuti di esempio, con conclusione il giorno 2 Gennaio 2017 alle ore 13:15 UTC. Si nota come il grafico di osservazione complessivo sia la composizione di diverse immagini 64x64 pixel di cui è costituito il dataset.



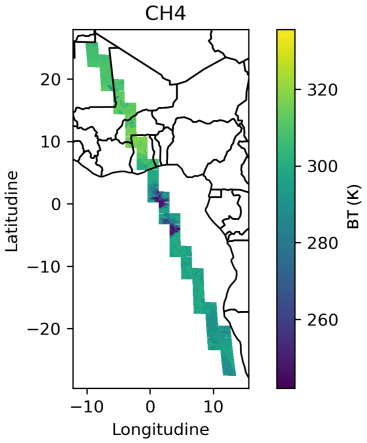
(a) SEVIRI: VIS 0.6



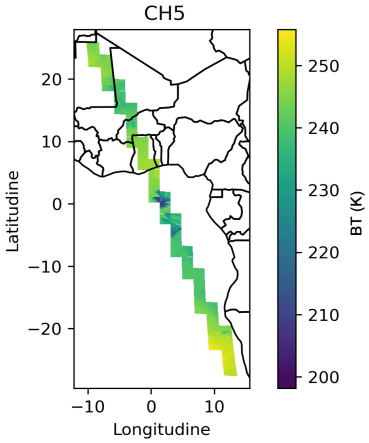
(b) SEVIRI: VIS 0.8



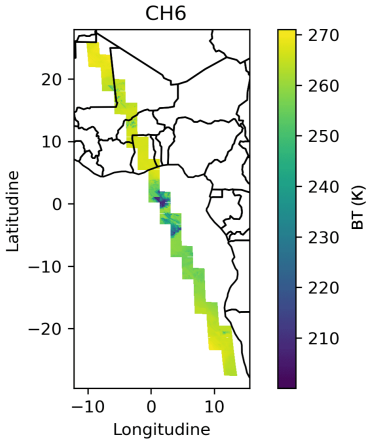
(c) SEVIRI: NIR 1.6



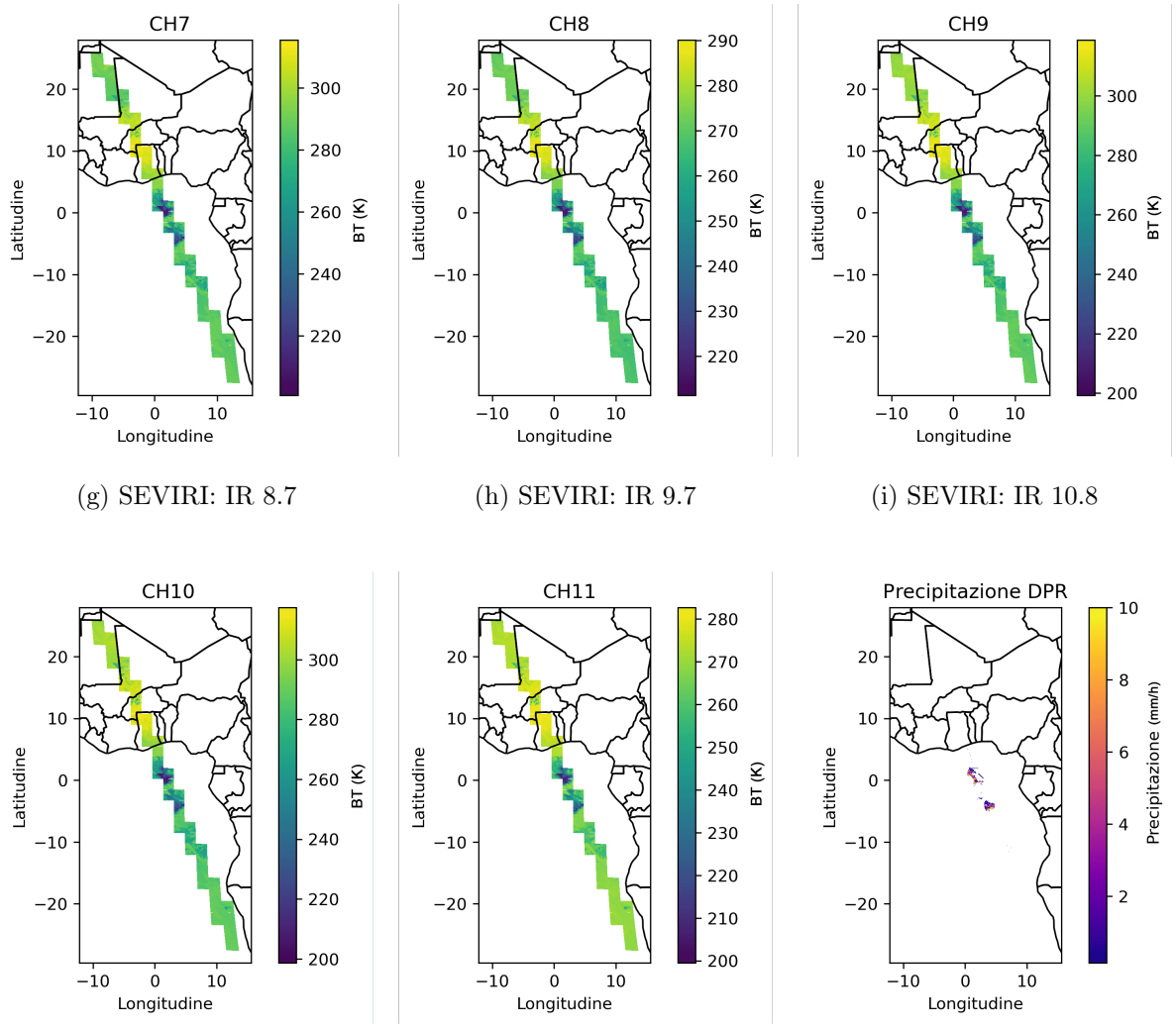
(d) SEVIRI: IR 3.9



(e) SEVIRI: IR 6.2



(f) SEVIRI: IR 7.3



(g) SEVIRI: IR 8.7 (h) SEVIRI: IR 9.7 (i) SEVIRI: IR 10.8

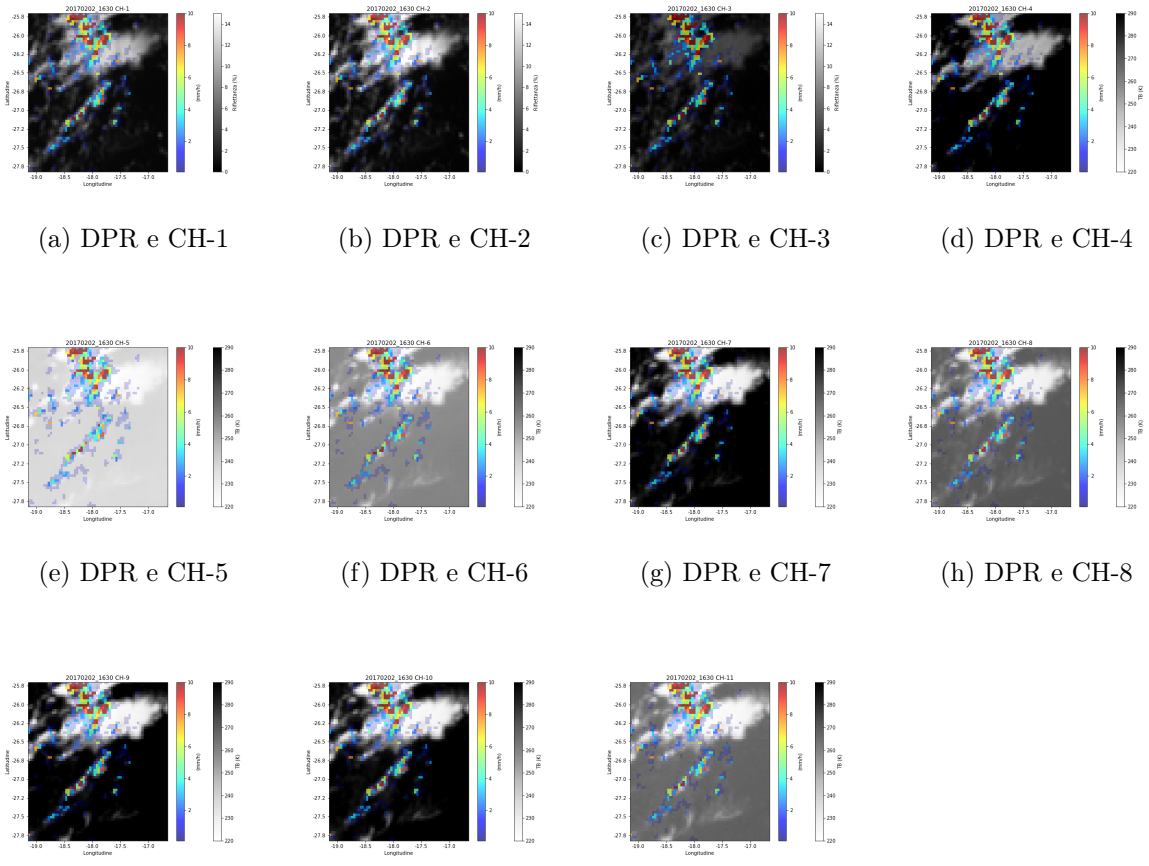
(j) SEVIRI: IR 12.0 (k) SEVIRI: IR 13.4 (l) DPR

Figura 4.2: Esempio dei valori in riflettanza e temperatura di brillanza per i canali di osservazione di SEVIRI e l'intensità di precipitazione del DPR, in scala di colori, 02/01/2017 ore 13:15 UTC

É possibile riassumere le caratteristiche principali di ogni canale di SEVIRI, andando a evidenziare gli utilizzi principali per una data lunghezza d'onda [39]:

- I due canali nel visibile **VIS 0.6** e **VIS 0.8** forniscono le immagini di nubi e superficie terrestre durante il giorno. Queste lunghezze d'onda aiutano a discriminare le superfici con vegetazione dalle nuvole in diversi periodi dell'anno e nella determinazione dell'indice di vegetazione e del quantitativo di aerosol.
- Il canale **NIR 1.6** è utilizzato per discriminare le nubi dalla neve e le nubi d'acqua dalle nubi di ghiaccio. In combinazione con i due canali visibili VIS 0.6 e VIS 0.8, migliora l'osservazione degli aerosol, dell'umidità del suolo e dell'indice di vegetazione.
- I canali **IR 6.2** e **IR 7.3** sono utilizzati per determinare la distribuzione del vapore acqueo in due strati distinti dell'atmosfera. Sono inoltre utilizzati in combinazione con altri canali a lunghezza d'onda maggiore nell'IR per determinare la temperatura delle nubi sottili, che possono apparire più calde a causa del fondo terrestre, e la determinazione del vento nelle aree prive di nubi.
- I quattro canali **IR 3.8**, **IR 8.7**, **IR 10.8** e **IR 12.0** forniscono un'osservazione continua delle nubi ed una stima della temperatura delle nubi, delle superfici terrestri e marine. Il canale IR 3.8 è inoltre particolarmente utilizzato di notte per rilevare la nebbia e le nubi molto basse.
- I canali **IR 9.7** e **IR 13.4** sono utilizzati per l'analisi delle masse d'aria e migliorare le capacità complessive di osservazione superficiale e del moto delle nubi. Il canale IR 9.7 appartiene alla banda di assorbimento dell'ozono ed è utilizzato per il monitoraggio dell'alta atmosfera e dei venti stratosferici. IR 13.4 si trova nella banda di assorbimento della CO_2 ed è utilizzato per il riconoscimento dei cirri, la valutazione della pressione alla sommità delle nuvole e il tracciamento delle nubi.

Nella seconda fase, prima di procedere con la preparazione del dataset, si valuta l'effettiva **coincidenza tra i valori di SEVIRI e del DPR** tramite uno script Python dedicato alla realizzazione dei grafici georeferenziati per singola immagine per i vari canali di MSG, in scala di grigio, e DPR e la loro sovrapposizione. Nel caso specifico in Fig.4.3, si combinano i canali di SEVIRI con i dati radar per i quali l'intensità di precipitazione viene rappresentata tra un minimo di 0.1 mm/h e un fondoscala di 10 mm/h, optando inoltre per una visualizzazione grafica diversa per valori di intensità di precipitazione inferiore a 2 mm/h, con una trasparenza maggiore, così da individuare meglio i confini delle nubi sottostanti. L'intervallo temporale scelto è il 2 febbraio 2017 alle ore 16:30:



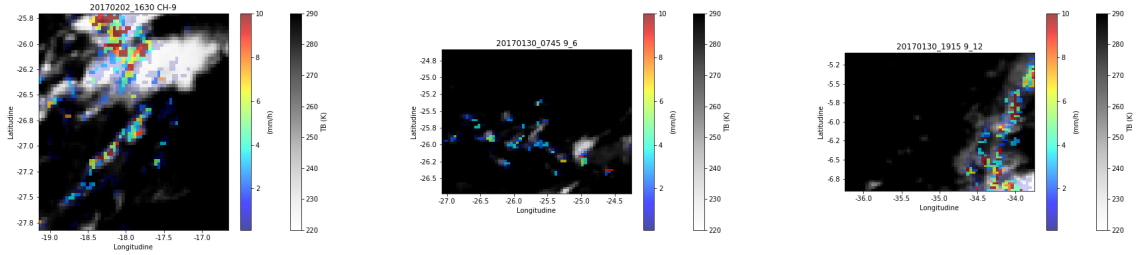
(a) DPR e CH-1 (b) DPR e CH-2 (c) DPR e CH-3 (d) DPR e CH-4

(e) DPR e CH-5 (f) DPR e CH-6 (g) DPR e CH-7 (h) DPR e CH-8

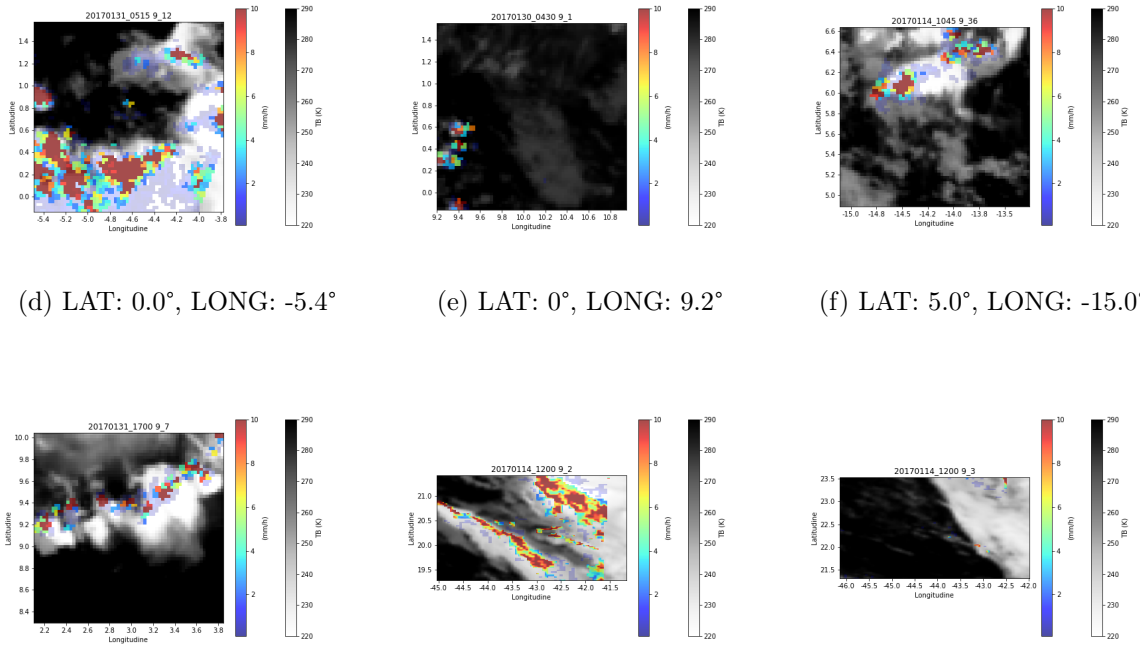
(i) DPR e CH-9 (j) DPR e CH-10 (k) DPR e CH-11

Figura 4.3: Coincidenza tra SEVIRI e DPR, per tutti i canali in scala di grigi, 02/02/17 ore 16:30 UTC

Si riportano, per latitudine minima crescente, inoltre alcune immagini di esempio che mostrano come la coincidenza sia effettivamente rispettata, per i grafici del CH-9 che è rappresentativo per la temperatura al top della nube, a qualsiasi latitudine e in situazioni di nubi dal pattern più articolato (Fig.4.4):



(a) LAT: -27.8°, LONG: -19.0° (b) LAT: -26.5°, LONG: -27.0° (c) LAT: -6.8°, LONG: -36.0°



(d) LAT: 0.0°, LONG: -5.4° (e) LAT: 0°, LONG: 9.2° (f) LAT: 5.0°, LONG: -15.0°

(g) LAT: 8.4°, LONG: 2.2° (h) LAT: 19.5°, LONG: -45.0° (i) LAT: 21.5°, LONG: -46.0°

Figura 4.4: Coincidenza tra SEVIRI (CH-9) e DPR per alcune configurazioni particolari a diverse latitudini

Le operazioni di studio sui file sorgente si completano quindi con la comprensione del tipo di formato dei file, la loro gestione tramite script Python, la verifica della coincidenza spaziale e temporale tra i valori georeferenziati dei due satelliti e la descrizione delle possibilità di utilizzo degli output di ogni canale.

4.3 Preparazione dataset

Costruiti i dataset con le matrici per i valori completi con facile accesso a tutti i valori annuali, si procede alla **preparazione dei dati di input** del modello.

Il modello in machine learning accetta in input delle variabili che posseggono le caratteristiche opportune per ottenere delle predizioni utili alla risoluzione del problema. Queste variabili sono chiamate **features**.

Per preparare il dataset è stato applicato il seguente filtraggio:

- **Temperatura CH-9**: si escludono i valori di temperatura di brillanza per il CH-9 superiori a 305K, questo perchè tali valori sono sicuramente associati a tempo stabile, privo di nubi e privo di precipitazioni. Ovviamente vengono esclusi dal dataset anche tutti i valori relativi degli altri canali e del DPR.
- **Latitudine e longitudine**: tenendo conto della regione di visione di SEVIRI e della possibile distorsione a latitudini e longitudini sui bordi del FOV e dell'errore del parallasse, si considera un filtro spaziale che soddisfa la seguente relazione, in modo da tenere l'area spaziale di osservazione meno soggetta a errori e deformazioni. Vengono quindi escluse dal database tutte le immagini la cui latitudine e longitudine media, espresse in gradi, non rispettano la seguente relazione:

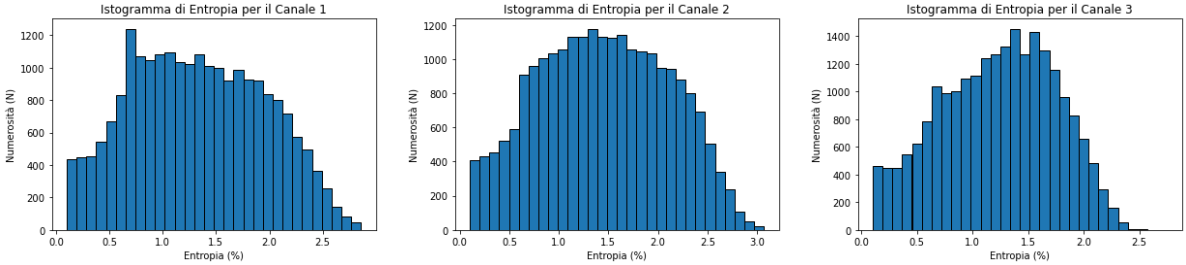
$$|latitudine| + |longitudine| \leq 55$$

- **Valori errati CH-4**: sono filtrati i valori negativi del canale CH-4, in quanto dovuti a probabili errori del sensore e quindi privi di significato fisico.
- **Intensità massima DPR**: nell'uso del modello si effettuano sia misure su tutto il range del prodotto di precipitazione del DPR, che limitandone le intensità di precipitazione a 30 mm/h, valore convenzionalmente scelto come soglia per la pioggia estrema.

- **Distinzione notte/giorno**: è necessario distinguere il giorno dalla notte, poichè durante la condizione notturna è necessario escludere i canali di SEVIRI nel visibile (dal CH-1 al CH-3), privi di informazione. Inoltre, si escludono anche le situazioni di transizione tra la condizione di buio e luce piena [40]. Si prova a realizzare due tipologie di filtro per la classificazione giorno/notte, basate su:
  - **Entropia**: in questo caso viene calcolata l'entropia media di ogni singola immagine in scala di grigi tramite una funzione nel codice Python. L'entropia permette di misurare l'informazione relativa alla luminosità complessiva dell'immagine: valori più alti indicano una condizione diurna, mentre valori nulli una condizione notturna . Per il calcolo la funzione crea una distribuzione di probabilità per i valori dei pixel e viene applicata la formula

$$H = -\sum_i p_i \log(p_i)$$

    dove p_i rappresenta le probabilità dei valori dei pixel, per determinare l'entropia.
    Si calcola l'entropia ai canali del visibile, con l'obiettivo di trovare delle soglie di definizione per il giorno, la notte e di conseguenza le fasi intermedie da escludere. Si riportano quindi in Figura 4.5 gli istogrammi relativi al valore di entropia per i 3 canali nel visibile e la sua numerosità di occorrenza nel dataset, relativamente a una selezione sui primi tre mesi del 2017: è ragionevole attendersi una crescita nella frequenza con l'aumentare del valore di entropia durante le fasi di transizione (alba e tramonto) e una situazione più definita per le condizioni di giorno e notte.



(a) CH-1, primo trimestre 2017 (b) CH-2, primo trimestre 2017 (c) CH-3, primo trimestre 2017

Figura 4.5: Istogrammi entropia nei canali del visibile sui dati combinati del primo trimestre del 2017

    Si sceglie quindi di definire le seguenti soglie:
    - **Notte**: i valori [0, 0.1] % di entropia, cioè inferiori o uguali a 0.1 % sono attribuibili alla situazione in assenza di luce visibile;
    - **Giorno**: i valori [0.9, max] % di entropia, cioè superiori o uguali a 0.9 % sono attribuibili alla situazione in presenza di luce visibile;
    - **Transizione**: i valori (0.1, 0.9) % di entropia, cioè compresi tra le due casistiche precedentemente individuate, sono esclusi dal filtraggio e le relative immagini satellitari non utilizzate come dati di input del modello.
  - **Fuso orario**: in questo secondo algoritmo si procede all'individuazione del caso corretto tramite il fuso orario e la stagionalità della specifica posizione dell'immagine satellitare.
    In un nuovo script Python, per ogni immagine, si assegnano la sua latitudine e longitudine media e l'orario UTC di raccolta dell'informazione. Poiché il dataset è annuale e distribuito su gran parte del globo, bisogna tener conto della diversa stagionalità di ogni punto per stabilire se è effettivamente giorno oppure notte. Si procede allora in questo modo: si converte l'orario UTC in orario locale a seconda del fuso orario dato dal posizionamento geografico dell'immagine, e tramite la libreria *ephem*, adatta per il calcolo preciso di applicazioni astronomiche, si individua l'orario esatto di alba e tramonto per una precisa posizione e giorno dell'anno. Infine si considera notte il periodo compreso tra due ore dopo il tramonto e due ore prima dell'alba del giorno successivo, mentre giorno tra due ore dopo l'alba e due ore prima del tramonto. Le fasce orarie intermedie sono associate alla transizione giorno/notte e vengono quindi esclusi i dati che vi rientrano.

  Tra questi due metodi, entrambi validi ed affidabili, si è deciso di **utilizzare il filtro legato all'entropia** invece del fuso orario. Questo perchè, andando ad applicare una selezione direttamente sull'output di SEVIRI nei canali del visibile, si gestiscono meglio le situazioni di luce limite legate al soleggiamento effettivo osservato dalla posizione relativa del satellite.

4.4 Statistica sui dati

In questo paragrafo si realizza un'analisi statistica preliminare sul dataset precedentemente preparato e pronto per il processamento con algoritmi in machine learning, ottenendo in particolare grafici sulle distribuzioni dei valori e l'analisi di alcuni valori potenzialmente anomali riscontrati.

Visualizzazione immagini

In primis si realizzano i grafici per le immagini nei vari canali, così da valutarne la coerenza dei valori di output, rappresentati in scala di grigio fissando il bianco ad alta riflettività per le osservazioni nel visibile e per basse temperature nell'infrarosso, in modo tale da avere una visione coerente della presenza delle nubi per tutti i canali di osservazione. Per il DPR si è individuata una scala che permetta una buona distinzione dei fenomeni di precipitazione in particolare di debole intensità, fissando quindi come fondoscala precipitazioni di 10 mm/h. Questa nuova visualizzazione va a migliorare quella realizzata in Fig.4.2 e verrà implementata nel resto dell'analisi e si può già notare in Fig.4.3.

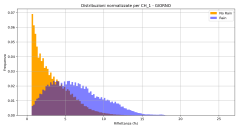
Distribuzioni

Viene effettuata una statistica di occorrenza dei valori dei vari canali, realizzando una serie di istogrammi di distribuzione. Si divide preliminarmente il dataset nelle due classi secco (assenza di precipitazione) e di pioggia, in riferimento ai valori del DPR e fissando la soglia limite a 0.1 mm/h. In seguito con uno script Python si individuano bin di ampiezza 0.2 % nel caso della riflettanza e 2 K nel caso della temperatura di brillanza. Chiamiamo f_i il numero di valori che rientrano nell'i-esimo bin e N il numero totale di valori; distinguiamo con gli apici R e NR rispettivamente per l'istogramma relativo ai valori con pioggia (R) e senza pioggia (NR).

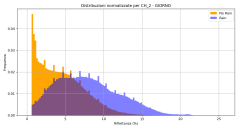
Si valutano tre normalizzazioni:

- **Sulla numerosità totale**: si normalizza rispetto alla numerosità totale, ottenendo per i bin di pioggia $\frac{f_i^R}{N^R+N^{NR}}$ e per quello di assenza di pioggia $\frac{f_i^{NR}}{N^R+N^{NR}}$. Con questa normalizzazione, essendo la numerosità dei dati di tempo non perturbato molto maggiore, non si riesce tuttavia a distinguere bene le due casistiche.
- **Sulla numerosità delle classi secco/pioggia**: per ovviare al problema riscontrato l'istogramma viene realizzato normalizzando le frequenze sul totale dei valori relativi alla stessa classe, in modo che per l'istogramma di pioggia si ha $\frac{f_i^R}{N^R}$ e per quello di secco si considera $\frac{f_i^{NR}}{N^{NR}}$. Ne risulta che ora i due istogrammi hanno la stessa area e sono quindi confrontabili.

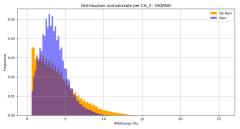
  Si riportano in Fig.4.6 le distribuzioni normalizzate per gli 11 canali nel caso giorno:



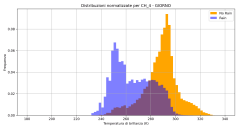
(a) CH-1, bin 0.2 %



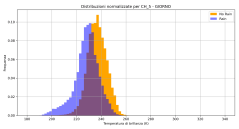
(b) CH-2, bin 0.2 %



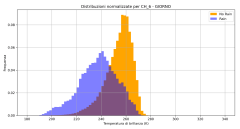
(c) CH-3, bin 0.2 %



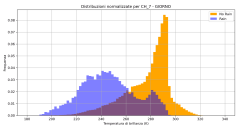
(d) CH-4, bin 2 K



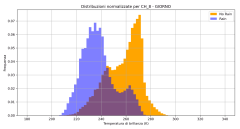
(e) CH-5, bin 2 K



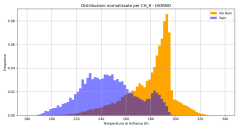
(f) CH-6, bin 2 K



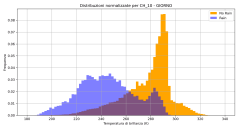
(g) CH-7, bin 2 K



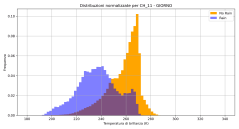
(h) CH-8, bin 2 K



(i) CH-9, bin 2K



(j) CH-10, bin 2K

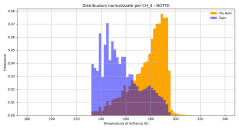


(k) CH-11, bin 2K

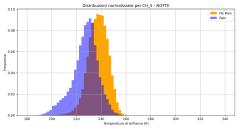
Figura 4.6: Distribuzioni normalizzate per gli 11 canali di SEVIRI, secco (in arancione) vs pioggia (in blu), caso giorno

Si individua chiaramente una differenza tra i valori di massima occorrenza per gli istogrammi secco/-pioggia con una tendenza ad avere riflettività maggiore e temperature di brillanza minori in caso di pioggia.

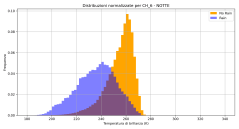
Si verifica inoltre le distribuzioni normalizzate in Fig.4.7 anche per il caso notte:



(a) CH-4, bin 2 K

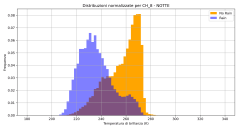


(b) CH-5, bin 2 K



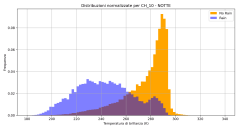
(c) CH-6, bin 2 K

(d) CH-7, bin 2 K

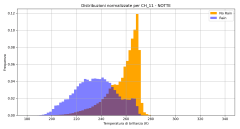


(e) CH-8, bin 2 K

(f) CH-9, bin 2 K



(g) CH-10, bin 2 K



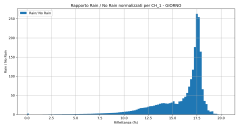
(h) CH-11, bin 2 K

Figura 4.7: Distribuzioni normalizzate per gli 8 canali di SEVIRI, secco (in arancione) vs pioggia (in blu), caso notte

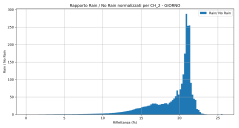
Anche nel caso notte viene confermata la tendenza ad avere temperature di brillanza inferiori in presenza di pioggia.

- **Rapporto secco/pioggia**: in quest'ultima normalizzazione si effettua il rapporto pioggia su secco per le frequenze, in modo tale che l'istogramma evidenzia la maggior frequenza di pioggia a temperature inferiori. A seguito di questa osservazione, si attende che questa distribuzione abbia un massimo a basse temperatura e decresca con continuità aumentando la temperatura, per le quali è meno probabile che piova.

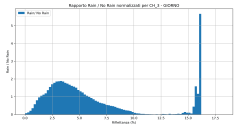
Si riportano i grafici per il caso giorno in Fig.4.8, in quanto per il caso notte i risultati sono generalmente simili anche se meno marcati con istogrammi dalla decrescita più rapida.



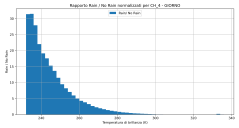
(a) CH-1



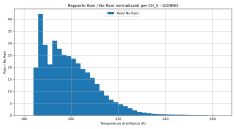
(b) CH-2



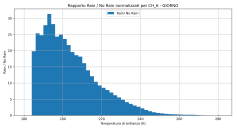
(c) CH-3



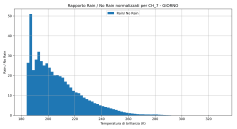
(d) CH-4



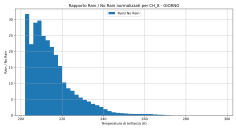
(e) CH-5



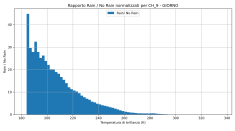
(f) CH-6



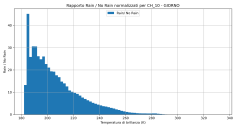
(g) CH-7



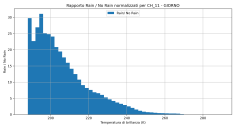
(h) CH-8



(i) CH-9



(j) CH-10



(k) CH-11

Figura 4.8: Distribuzioni del rapporto tra pioggia/secco, nel caso giorno

Il rapporto nel caso delle temperature conferma l'ipotesi, così come per i canali del visibile per i quali la precipitazione si verifica a valori di riflettanza maggiori.

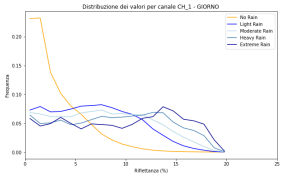
Lo studio di queste distribuzioni è importante in quanto trasmettono l'informazione che esiste una relazione, seppur debole, tra i valori misurati da SEVIRI nei vari canali e la presenza di precipitazione: nello specifico è confermato come fenomeni di precipitazione sono correlati a temperature di brillanza più bassa e a riflettanza maggiori.

Inoltre nel caso giorno si posseggono più informazioni, dai canali del visibile, ma anche una distinzione più netta tra le temperature nel caso di pioggia o assenza di precipitazione: per questo è lecito attendersi una miglior efficienza di un modello di stima di giorno piuttosto che di notte.

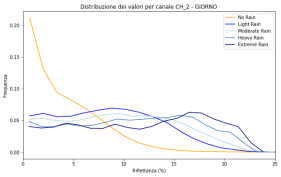
Al fine di valutare se c'è una relazione tra le distribuzioni appena osservate e l'intensità di precipitazione, si divide il campione in ulteriori classi oltre la divisione binaria pioggia/secco, secondo il seguente schema:

- **Secco**: $[0 - 0.1)$ mm/h
- **Pioggia leggera**: $[0.1 - 1)$ mm/h
- **Pioggia moderata**: $[1 - 5)$ mm/h
- **Pioggia intensa**: $[5 - 30)$ mm/h
- **Pioggia estrema**: > 30 mm/h

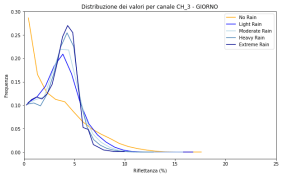
Si ottengono, in Fig.4.9 per il caso giorno, le distribuzioni divise per classi, dove gli istogrammi sono stati interpolati in una curva in modo che segua la frequenza di ogni bin:



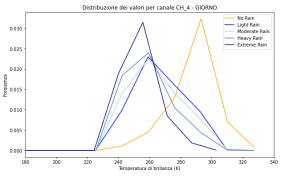
(a) CH-1



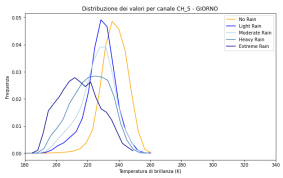
(b) CH-2



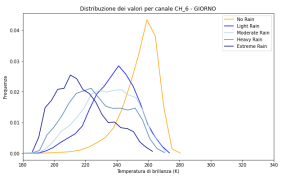
(c) CH-3



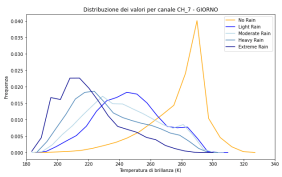
(d) CH-4



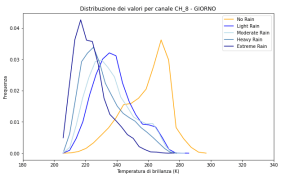
(e) CH-5



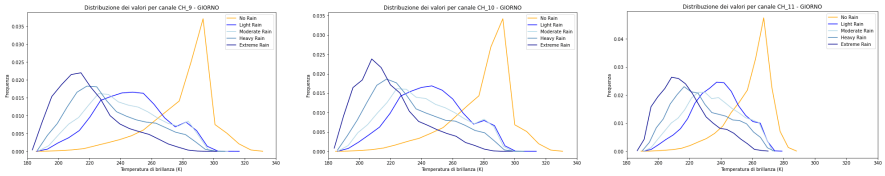
(f) CH-6



(g) CH-7



(h) CH-8



(i) CH-9 (j) CH-10 (k) CH-11

Figura 4.9: Distribuzioni divise in classi di intensità di precipitazioni, caso giorno

Il risultato è importante in quanto oltre ad una relazione che permetta di distinguere il caso di pioggia da non pioggia, si ha statisticamente una distinzione anche tra le classi di differente intensità di precipitazione. I grafici delle distribuzioni mostrano come i picchi relativi agli istogrammi siano ordinati e si abbia **maggior riflettanza** e **minor temperatura di brillanza** con l'**aumentare dell'intensità di precipitazione**.

Tuttavia le **aree di intersezione** tra gli istogrammi per le varie classi sono significative, evidenziando una difficoltà intrinseca nel distinguere l'appartenenza del dato alla classificazione. In particolare, i canali di osservazione a lunghezza d'onda minore, dal CH-1 al CH-5, mostrano una distinzione più evidente tra il caso di assenza di precipitazione e di pioggia ma molto debole tra le classi di intensità di precipitazione. Per i canali a lunghezze d'onda maggiori, dal CH-6 al CH-11, si ha una divisione maggiore tra i picchi e una minor area comune degli istogrammi a diverse classi di precipitazione, suggerendo che questi canali possano avere un contributo più importante nel riconoscimento della precipitazione.

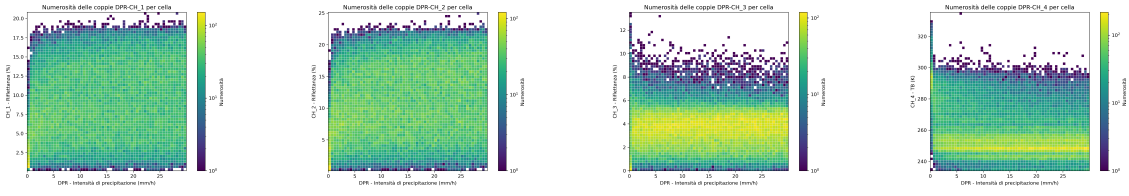
La generale assenza di una netta distinzione tra le classi è il sintomo di una **relazione debole**, tra queste grandezze e la stima di intensità di precipitazione, ed è la motivazione che conduce all'uso di modelli in machine learning per il riconoscimento al posto di relazioni fisiche dirette.

Relazione tra le features SEVIRI e la precipitazione DPR

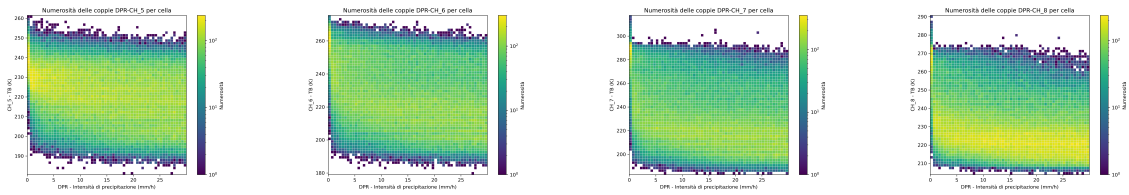
Per completare l'analisi statistica del dataset, si decide di mettere direttamente in relazione l'intensità di precipitazione del DPR con i valori dei canali, al fine di individuare graficamente eventuali relazioni tra queste.

Si procede dividendo il dataset in 60x60 celle quadrate, aventi ampiezza pari a 0.5 mm/h. Limitando la precipitazione a 30 mm/h, si assegna ogni coppia DPR-CH alla corrispondente cella nello spazio del grafico, contandone la numerosità di occorenza, in modo da simulare una densità. Infine si realizza il plot con il DPR sull'asse x e il canale di SEVIRI sull'asse y, evidenziando la densità di celle con una scala di colori adeguata.

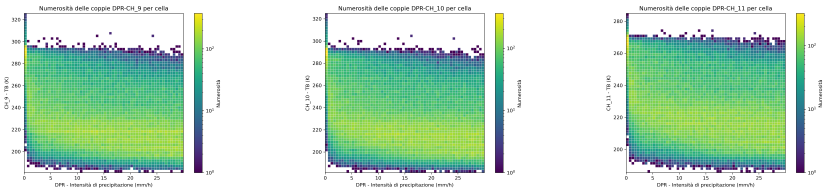
Si riportano nella Fig.4.10 i grafici nel caso giorno per i primi quattro canali, e nel caso giorno e notte insieme per i restanti canali nell'infrarosso:



(a) CH-1 (b) CH-2 (c) CH-3 (d) CH-4



(e) CH-5 (f) CH-6 (g) CH-7 (h) CH-8



(i) CH-9 (j) CH-10 (k) CH-11

Figura 4.10: Densità di occorrenza delle coppie DPR-CH, divisione in celle, caso giorno per i primi 4 canali e giorno + notte per gli altri

Questi grafici sono in accordo con quanto ipotizzato sulla difficoltà di individuare relazioni fisiche chiare tra le osservazioni di SEVIRI e l'intensità di precipitazione, ed in particolare si può osservare come:

- i **primi due canali** nel visibile si evidenzia una diversa distribuzione solamente tra la prima e la seconda classe, mentre sul range di intensità di precipitazione le celle sono equamente distribuite.
- il **terzo e il quarto canale** evidenziano una fascia di maggior densità, rispettivamente a riflettanza più alta e temperatura più bassa in caso di precipitazione che però non caratterizza nessuna relazione con l'intensità di precipitazione.
- per i restanti **canali nell'infrarosso** si individuano deboli relazioni che si manifestano con una curva di maggior densità che decresce gradualmente con l'aumentare dell'intensità di precipitazione. Tra questi il canale dove risulta più evidente questa caratteristica sono i canali a maggior lunghezza d'onda, in accordo con l'osservazione fatta nel paragrafo precedente sulle distribuzioni divise in classi.

  In particolare il canale 8, l'IR 9.7, mostra una maggior densità delle celle tra i 260 K e i 270 K nel caso di assenza di precipitazioni e una curva di densità che decresce di temperatura con l'aumentare dei valori di intensità di precipitazione. Si osserva quindi una discontinuità tra l'assenza di precipitazione e l'inizio dei fenomeni, una rapida decrescita tra 1 mm/h e 5 mm/h e una banda di densità nel resto dello spettro eccessivamente larga per poter individuare relazioni fisiche chiare con un approccio tradizionale. Si proverà comunque a sfruttare questa caratteristica nella realizzazione di un filtro descritto nel paragrafo 5.1.

Valori sospetti

Si segnalano, nell'analisi dei file per diversi intervalli di tempo, alcuni casi che potrebbero risultare anomalie di misura. Osservando le distribuzioni per il CH-9, sia di giorno che di notte, si nota una crescita repentina dei valori sopra i 280 K in presenza di pioggia. Inoltre nel solo caso giorno è presente un gruppo di valori superiori ai 300 K, fino anche a oltre 320 K. Si prova a dare una spiegazione a queste due situazioni sospette:

- **>280 K al top della nube con precipitazione**: sembra anomalo che vi siano nubi precipitanti la cui temperatura al top sia ben sopra lo zero celsius con valori anche oltre i 295 K. Si procede allora selezionando le immagini per le quali vi sia precipitazione e sovrapponendo in scala di colori la temperatura del CH-9. Come si osserva in Fig.4.11, questa casistica di valori è spiegabile con piogge in specifiche aree tropicali.

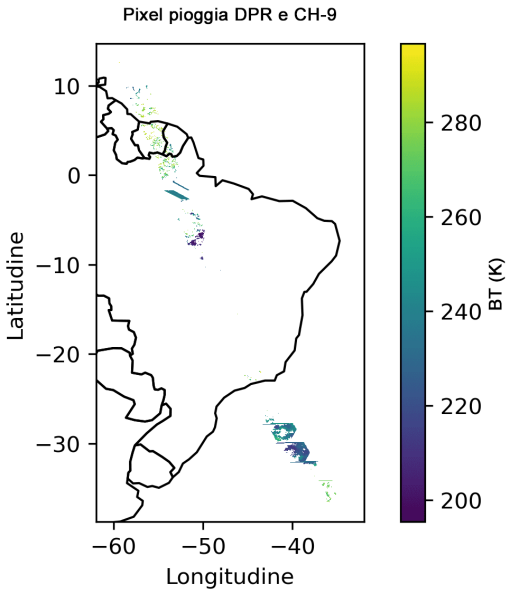


Figura 4.11: Esempio di DPR in area tropicale su scala di colori CH-9, valori del top della nube precipitante sopra 280 K

- **>300 K di giorno**: la distribuzione evidenzia anche una crescita dei valori sopra i 300 K, fino a oltre 325 K, in situazione diurna. Individuando le singole immagini in cui ciò accade, si possono associare questi valori al riscaldamento delle aree desertiche in situazioni di tempo stabile, come si può osservare dall'esempio in Fig.4.12

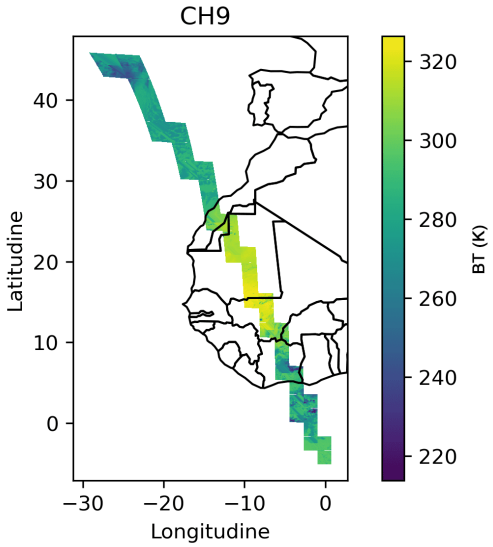


Figura 4.12: Esempio di CH-9 su aree desertiche, con temperature sopra i 325 K

5 Modello ML: stima di precipitazione da SEVIRI con addestramento su DPR

Il modello machine learning scelto per questo lavoro è basato sui due algoritmi di **Random Forest** e **Gradient Boosting**.

Il dataset preparato viene **classificato** con il Random Forest, poiché è un algortimo particolarmente adatto ed efficiente alla classificazione in più classi [41], in modo da identificare le classi di assenza di precipitazione e più livelli di intensità di precipitazione.

Si realizza inoltre una **regressione** con entrambi gli algoritmi, valutando quale possa risultare il migliore sulle caratteristiche dei dati a disposizione. La tecnica di regressione ha lo scopo di fornire una stima numerica sull'intensità di precipitazione, rendendo la predizione di precipitazione un valore continuo e direttamente confrontabile con altri prodotti numerici di stima e misura della precipitazione.

Per la valutazione degli algoritmi si utilizzano una serie di **indicatori statistici** che saranno descritti in dettaglio nella sezione 5.3.

5.1 Scelta e bilanciamento delle classi

La prima operazione è la **scelta delle classi**, ovvero dei sottogruppi aventi una caratteristica comune, basandosi su intervalli di intensità di precipitazione del DPR.

Essendoci un presenza maggiore di valori con precipitazione nulla, poichè l'assenza di precipitazione è lo stato più frequente, le classi avranno un numero di campioni differente, con l'intervallo di precipitazione di maggiore intensità sicuramente meno popolato di quelli a minor intensità o di condizioni secche. É quindi consigliato nel machine learning di **bilanciare le classi** per l'addestramento del modello [42]. L'operazione di bilanciamento permette di rendere omogeneo il dataset e quindi numericamente significative tutte le caratteristiche che esso rappresenta, evitando il problema dell'apprendimento prioritario sui dati più frequenti. Nel nostro caso questa operazione assume un ruolo importante in quanto vorremmo apprendere al meglio le caratteristiche di precipitazione, che risultano però essere le classi minoritarie.

Si effettuano diverse scelte di classificazione, al fine di verificare quale possa essere la miglior soluzione da fornire in input al modello, con relativo bilanciamento. Nello specifico è stato scelto un campione a:

- **2 classi**: Pioggia e secco, con valore soglia fissato a 0.1 mm/h. Questa classificazione è scelta, in combinazione con la limitazione di intensità a 30 mm/h, per valutare quale algoritmo di regressione, tra Random Forest o Gradient Boosting, possa essere più adatto ai dati, come in analisi nel paragrafo 5.5.
- **4 classi**: Secco ([0, 0.1) mm/h), pioggia leggera ([0.1, 1) mm/h), pioggia moderata ([1, 5) mm/h), pioggia intensa ([5, 30) mm/h). Questa scelta delle classi è utilizzata nel test del modello Random Forest di classificazione, descritto nel paragrafo 5.4.
- **5 classi**: Secco ([0, 0.1) mm/h), pioggia leggera ([0.1, 1) mm/h), pioggia moderata ([1, 5) mm/h), pioggia intensa ([5, 30) mm/h) e pioggia estrema ([30, 150) mm/h). Questa scelta delle classi è utilizzata nel modello completo nel capitolo 6.

Il bilanciamento viene effettuato valutando la **numerosità di ogni classe**, riportando la grandezza di ogni classe al numero di elementi contenuto nella classe meno popolata. Questa operazione comporta una perdita di dati e la selezione viene effettuata casualmente. Un'altra opzione di bilanciamento consiste nella generazione casuale di valori al fine di eguagliare le classi sulla più numerosa; tuttavia questa operazione viene scartata per evitare di dover gestire dataset eccessivamente grandi ed eccessivamente onerosi come costo computazionale.

Filtro basato sul CH-8

Si valuta inoltre di filtrare le classi basandosi sui valori del canale 8 di SEVIRI.

Come visto nel paragrafo della distribuzione a celle, il canale 8 mostra più nitidamente una **relazione nella densità** della sua distribuzione rispetto alla variazione di intensità della precipitazione. Tra le celle di distribuzione, si selezionano per ogni step orizzontale, relativo al DPR, le tre celle (chiamate "top celle" nella

legenda di Fig.5.1) con una frequenza di occorrenza dei valori di CH-8, definita numerosità, maggiore per ogni intervallo lungo l'asse x. Su queste si effettua un fit logaritmico del tipo:

$$a\dot{\log}(bx + c) + d$$

i cui parametri interpolati a, b, c, d e le informazioni statistiche sono riportate in Fig.5.1, insieme alla rappresentazione grafica della curva interpolante le celle prima individuate:

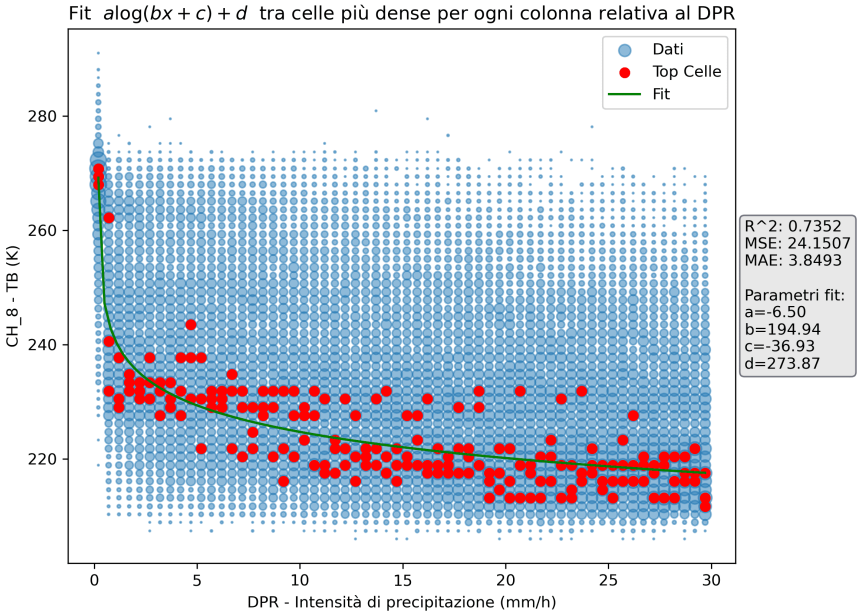


Figura 5.1: Fit sulle celle più densamente popolate del CH-8 in funzione del DPR. A destra vi sono i parametri di interpolazione e gli indici statistici

Con la curva di fit si valuta un intervallo di valori, ad esempio in un range tra la curva di fit ± MAE, l'errore medio assoluto, oppure in un range proporzionale con l'aumentare dell'intensità, come mostrato ad esempio nella Fig.5.2

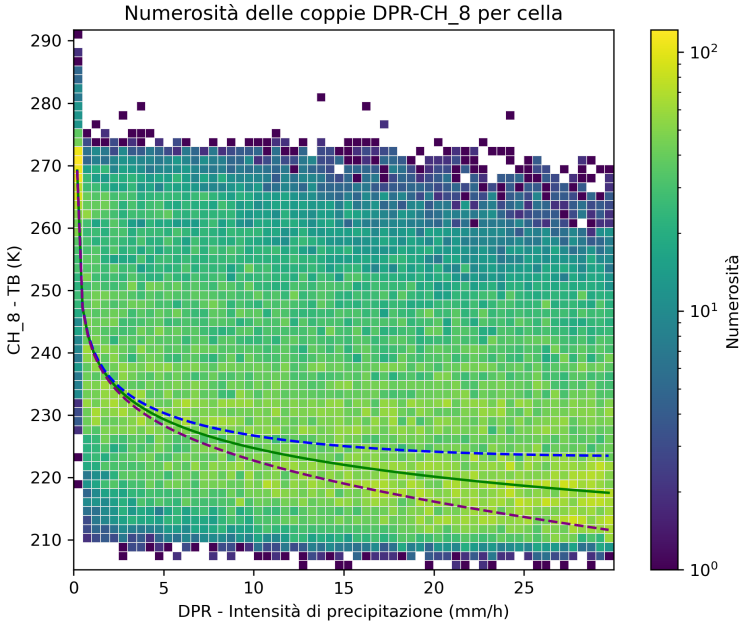


Figura 5.2: Range indicativo dei valori in un range intorno al fit del CH-8, su cui basare un filtro del dataset

Con il range di valori selezionato è possibile bilanciare le classi pescando i valori in maniera casuale nell'intervallo individuato. Si anticipa che i risultati del modello in regressione sono molto buoni e la stima sul dataset di test discrimina con ottima correlazione ($R^2 \approx 0.9$) le intensità di precipitazione. Tuttavia questo approccio verrà escluso in quanto, con un dataset di validazione, si perde la capacità di fare previsioni senza filtrare i dati del CH-8 preventivamente: l'assunzione è quindi troppo limitante e non applicabile con successo nelle condizioni standard di un dataset completo.

5.2 Selezione delle features

La scelta delle variabili di dati da assegnare in **input al modello** è fondamentale per permettere a quest'ultimo di apprendere le relazioni rappresentative sui dati. Nel modello si considerano le seguenti features:

- i **dati** degli 11 canali di SEVIRI (da CH-1 a CH-11), in riflettanza (%) (da CH-1 a CH-3) e temperatura di brillanza (da CH-4 a CH-11),
- le **differenze** tra le varie combinazioni dei canali ad esclusione dei canali nel visibile (ad esempio CH5-CH6 oppure CH10-CH11), che possono fornire informazione sulle variazioni delle proprietà fisiche in atmosfera [43],
- le **medie mobili** su una finestra di 5x5 di valori per ogni canale di SEVIRI (mean_CH). La media è utilizzata per ridurre il rumore nei dati,
- le **deviazioni standard mobili** sempre su una finestra di 5x5 valori per ogni canale (std_CH). La deviazione standard misura la dispersione dei dati intorno alla media e il calcolo per una finestra permette di valutare quanto variano i dati localmente.

5.3 Indici di valutazione statistica

La valutazione della **bontà e dell'affidabilità statistica** del modello viene effettuata tramite una serie di indici, distinguendo nel caso di una classificazione o una regressione.

Nel **caso di classificazione**, si valutano:

- **Matrice di confusione**: Un modello in classificazione effettua una previsione sulla classe positiva, che è quella che si cerca di identificare, e una classe negativa, quella che si cerca di escludere. Una matrice di confusione mostra il numero di predizioni effettuate dal modello per ciascuna classe nei casi di: previsione positiva corretta (True Positive, TP), previsione positiva errata (False Positive, FP), previsione negativa corretta (True Negative, TN) e previsione negativa sbagliata (False Negativ, FN):

$$\begin{bmatrix} \mathrm{TP} & \mathrm{FP} \\ \mathrm{FN} & \mathrm{TN} \end{bmatrix}$$

- **Probability of Detection (POD)**: Questa metrica indica la proporzione di eventi positivi che sono stati correttamente identificati dal modello (TP) rispetto al totale degli eventi positivi reali, che comprendono anche gli eventi positivi non carrettamente identificati (FP), secondo la formula:

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

  L'intervallo dei valori per il POD va da 0 a 1, dove 1 indica un rilevamento perfetto di tutti gli eventi positivi.

- **False Alarm Ratio (FAR)**: Questa metrica rappresenta il rapporto tra il numero di allarmi falsi emessi dal modello (eventi erroneamente identificati come positivi) e il numero totale di allarmi emessi (veri positivi più falsi positivi):

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}$$

  L'intervallo dei valori per il FAR va da 0 a 1, dove 0 indica che non ci sono falsi allarmi.

- **Multiplicative Bias (BIAS)**: Questo indice misura la tendenza del modello a sovrastimare o sottostimare i risultati. Viene calcolato come il rapporto tra la somma dei casi positivi predetti (true positive + false positive) e la somma dei casi positivi reali (true positive + false negative FN):

$$\text{BIAS} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FN}}$$

  Un BIAS pari a 1 indica una previsione bilanciata, mentre valori superiori a 1 indicano una sovrastima e valori inferiori a 1 indicano una sottostima.

- **Critical Success Index (CSI)**: Questo indice valuta la capacità del modello di predire correttamente sia gli eventi positivi che quelli negativi. Viene calcolato come il rapporto tra i veri positivi e la somma dei veri positivi, falsi positivi e falsi negativi:

$$\text{CSI} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

  L'intervallo dei valori per il CSI va da 0 a 1, dove 1 indica una previsione perfetta.

- **Equitable Threat Score (ETS)**: Questo indice tiene conto della frequenza degli eventi previsti rispetto agli eventi osservati, considerando sia le predizioni corrette che quelle incorrette:

$$\text{ETS} = \frac{\text{TP} - \text{E}}{\text{TP} + \text{FP} + \text{FN} - \text{E}}$$

  dove

$$\text{E} = \frac{(\text{TP} + \text{FN})(\text{TP} + \text{FP})}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

  L'intervallo dei valori per l'ETS va da -∞ a 1, dove 1 indica una previsione perfetta e valori pari o inferiori a 0 indicano una previsione non migliore di una casuale.

- **Heidke Skill Score (HSS)**: Questo indice misura l'accuratezza delle previsioni corrette rispetto a ciò che ci si aspetterebbe da un evento casuale, secondo la formula [44]:

$$\text{HSS} = \frac{2(\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN})}{(\text{TP} + \text{FN})(\text{FN} + \text{TN}) + (\text{TP} + \text{FP})(\text{FP} + \text{TN})}$$

  L'intervallo dei valori per l'HSS va da -1 a 1, dove 1 indica una previsione perfetta e valori pari o inferiori a 0 indicano una previsione non migliore di una casuale.

Inoltre, specificatamente per la Random Forest in classificazione si è anche valutato:

- **Accuracy**: Percentuale di predizioni corrette rispetto al totale delle predizioni:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

  L'intervallo dei valori per l'Accuracy va da 0 a 1, dove 1 indica una previsione perfetta.

- **Precision**: Percentuale di veri positivi tra tutte le predizioni positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

  L'intervallo dei valori per la Precision va da 0 a 1, dove 1 indica una previsione perfetta.

- **Recall**: Percentuale di veri positivi che sono stati identificati correttamente rispetto a tutti i veri positivi:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  L'intervallo dei valori per il Recall va da 0 a 1, dove 1 indica una previsione perfetta.

- **F1 Score**: Media armonica tra precision e recall, utile quando c'è uno sbilanciamento tra le classi:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

  L'intervallo dei valori per l'F1 Score va da 0 a 1, dove 1 indica una previsione perfetta.

- **Receiver Operating Characteristic (ROC)**: è una curva che rappresenta la relazione tra il True Positive Rate (TPR) e il False Positive Rate (FPR), al variare della soglia di decisione del modello. Il TPR misura la proporzione di veri positivi rispetto i casi positivi reali, calcolato come TPR = TP/(TP+FN). Il FPR indica la proporzione di falsi positivi rispetto i casi negativi reali, calcolato come FPR = FP/(FP+TN).
- L'**Area Under the Curve (AUC)** è una misura derivata dalla curva ROC che quantifica la capacità discriminativa del modello. L'AUC varia tra 0 e 1, dove un valore maggiore indica una migliore capacità del modello di classificare correttamente le istanze positive come positive e le istanze negative come negative, indipendentemente dalla soglia di decisione scelta. Un AUC pari a 0.5 suggerisce una performance del modello equivalente al caso casuale.

Nel **caso di regressione**, i coefficienti di valutazione del modello utilizzati includono:

- **Mean Squared Error (MSE)**: Questo indice rappresenta la media dei quadrati degli errori tra i valori predetti dal modello e i valori effettivi dei dati. È una misura della dispersione dei dati e un MSE più basso indica una migliore adattabilità del modello ai dati:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  dove y_i e $\hat{y}_i$ sono rispettivamente i valori effettivi e predetti per l'osservazione i, e n è il numero totale di osservazioni nel dataset.

  L'intervallo dei valori per l'MSE va da 0 a ∞, dove 0 indica una previsione perfetta.

- **Mean Absolute Error (MAE)**: L'errore medio assoluto è la media degli errori assoluti tra le previsioni del modello e i valori effettivi. È meno sensibile agli outlier, valore anomalo, rispetto all'MSE ed è utile per capire quanto sono grandi gli errori in termini assoluti:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  L'intervallo dei valori per il MAE va da 0 a ∞, dove 0 indica una previsione perfetta.

- **Coefficient of Determination ($\mathbf{R^2}$)** [45]: Questo indice rappresenta la proporzione di varianza nei dati di output che è spiegata dal modello:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  dove y_i rappresenta il valore effettivo dell'osservazione i nel dataset, $\hat{y}_i$ rappresenta il valore predetto dal modello per l'osservazione i, $\bar{y}$ è la media dei valori effettivi y_i nel dataset, e n è il numero totale di osservazioni nel dataset.

  Con questa formalizzazione, utilizzata di frequente in contesti di data science e implementata nella libreria Scikit-Learn, il valore di R^2 varia da -1 a 1. Un valore di $R^2 = 1$ indica che il modello predice perfettamente i dati, mentre $R^2 = 0$ indica che il modello non migliora le previsioni rispetto a una semplice media dei valori effettivi. Valori di R^2 inferiori a 0 indicano che il modello è meno accurato di una semplice media dei valori effettivi e potrebbe predire nella direzione opposta alla variazione dei dati.

- **Correlation Coefficient (CC)**: Il coefficiente di correlazione misura la relazione lineare tra le previsioni del modello e i valori effettivi:

$$\text{CC} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

  L'intervallo dei valori per il CC va da -1 a 1, dove 1 indica una correlazione perfetta positiva.

- **Coefficient of Variation (CV)**: Il coefficiente di variazione è una misura relativa della dispersione dei dati rispetto alla media. È calcolato come il rapporto tra il valore medio e la deviazione standard e viene utilizzato per confrontare la variabilità dei dati in contesti diversi.

$$\text{CV} = \frac{\sigma}{\mu}$$

  dove σ è la deviazione standard e μ è la media. L'intervallo dei valori per il CV va da 0 a ∞, dove un valore più basso indica meno variabilità rispetto alla media.

- **Mean Error (ME)**: L'errore medio è la media degli errori tra le previsioni e i valori effettivi, considerando il segno. È una misura della tendenza del modello a sovrastimare o sottostimare i valori:

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

  L'intervallo dei valori per il ME può essere positivo o negativo, dove un valore prossimo a 0 indica che il modello non ha bias sistematico.

- **Normalized Mean Absolute Error (NMAE)**: Questo è l'MAE diviso per la differenza massima tra i valori effettivi e la loro media. Questa normalizzazione aiuta a comparare le prestazioni del modello su diverse scale di dati:

$$\text{NMAE} = \frac{\text{MAE}}{y_{\max} - y_{\min}}$$

  dove $y_{\max}$ e $y_{\min}$ sono rispettivamente il valore massimo e minimo dei dati effettivi. L'intervallo dei valori per il NMAE va da 0 a 1, dove 0 indica una previsione perfetta.

Per ogni classificazione o regressione, si valuta l'**importanza delle features** calcolando la riduzione media dell'impurità fornita da ciascuna features negli alberi decisionali (Mean Decrease in Impurity, MDI, calcolata dalla libreria Python Scikit-learn) di una Random Forest o di un Gradient Boosting:

$$\text{Importanza della feature}(i) = \sum_{t \in T} \frac{N_t}{N} \Delta I(i,t)$$

dove:

- T è l'insieme di tutti gli alberi nella Random Forest.
- N_t è il numero di nodi che utilizzano la feature i nell'albero t.
- N è il numero totale di nodi in tutti gli alberi.
- $\Delta I(i,t)$ è la riduzione di impurità apportata dalla feature i nell'albero t.

5.4 Classificazione Random Forest: test

Nella analisi di test della classificazione con Random Forest, si utilizza il dataset filtrato e bilanciato come descritto nella fase di preparazione e si realizza il setup del modello in classificazione.

Il dataset viene suddiviso in un **gruppo di training** e un **gruppo di test** sul quale effettuare la verifica statistica. Il set di training corrisponde al primo 80% di valori in ordine temporale, mentre il set di test l'ultimo 20%, ottenendo due gruppi tra loro indipendenti. Nel caso specifico per l'intero anno 2017, il dataset di training comprende circa i primi 10 mesi dell'anno, da gennaio 2017 a ottobre 2017, mentre il dataset di test gli ultimi due mesi.

La profondità e le caratteristiche di nodi, foglie e alberi vengono regolate tramite gli **iperparametri** del modello [46]. Di seguito sono riportati gli iperparametri Random Forest utilizzati, descritti nel paragrafo 3.5, e i valori specifici utilizzati per l'addestramento, scelti in accordo con altri lavori simili [47]:

- **random_state**: 42
- **n_jobs**: 8
- **n_estimators**: 100
- **max_depth**: 50
- **min_samples_leaf**: 3
- **max_features**: 0.5
- **max_samples**: 0.5

Realizzato il setup del modello è possibile procedere con il test dell'algoritmo Random Forest, utilizzando **4 classi** per l'intensità di precipitazione del DPR, limitata a 30 mm/h e valutando gli indici statistici rispetto al dataset di test.

Si riportano i parametri di valutazione statistica, mediati tra le classi, in Tab. 7.5, la matrice di confusione in Tab. 7.4 e un report degli indici diviso in classi in Tab. 5.3:

Parametro	**Valore medio**	**Parametro**	**Valore medio**
Accuracy	0.60	POD	0.97
Precision	0.59	FAR	0.016
Recall	0.60	BIAS	0.98
F1-Score	0.59	CSI	0.96
ETS	0.95	HSS	0.39

Tabella 5.1: Parametri di valutazione statistica medi tra le classi: classificazione RF, 4 classi, giorno

	Stima			
Misura	**secco**	**leggera**	**moderata**	**intensa**
secco	105992	1670	13275	3614
leggera	4364	180197	18095	34804
moderata	14789	30565	83608	44858
intensa	8982	67958	46911	73420

Tabella 5.2

Classe	**Precision**	**Recall**	**F1-Score**	**Campioni**
secco	0.79	0.85	0.82	124551
leggera	0.64	0.76	0.70	237460
moderata	0.52	0.48	0.50	173820
intensa	0.47	0.37	0.41	197271

Tabella 5.3: Report per classi: classificazione RF, 4 classi, giorno

Si realizzano due grafici. Il primo è una rappresentazione della matrice di confusione in Fig.5.3:

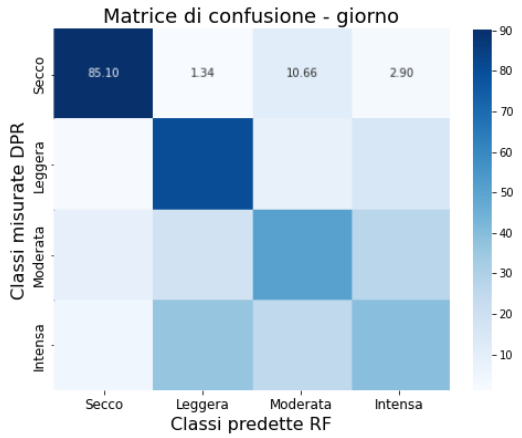


Figura 5.3: Matrice di confusione: classificazione RF, 4 classi, giorno

La ROC e i valori AUC, in Fig.5.4, forniscono una visione delle capacità discriminative del modello Random Forest per ciascuna classe. Un AUC più alto indica una migliore capacità del modello di distinguere tra la

classe di interesse e le altre classi, inoltre più è vicino ad 1 più il suo grafico avrà una forma a gradino. Nello specifico:

- **Classe secco (AUC = 0.97)**: La curva blu mostra che il classificatore è eccellente nel distinguere tra i giorni secchi e le altre classi, con un AUC molto vicino a 1.
- **Classe leggera (AUC = 0.86)**: La curva arancione indica che il classificatore ha buone prestazioni nel distinguere tra pioggia leggera e altre classi, con un AUC alto.
- **Classe moderata (AUC = 0.79)**: La curva verde mostra che il classificatore è moderatamente buono nel distinguere tra pioggia moderata e altre classi.
- **Classe intensa (AUC = 0.73)**: La curva rossa indica che il classificatore ha prestazioni accettabili ma non eccellenti nel distinguere tra pioggia intensa e altre classi.

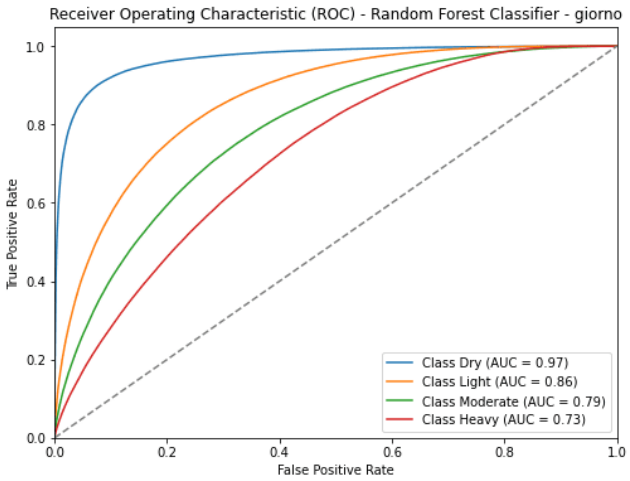
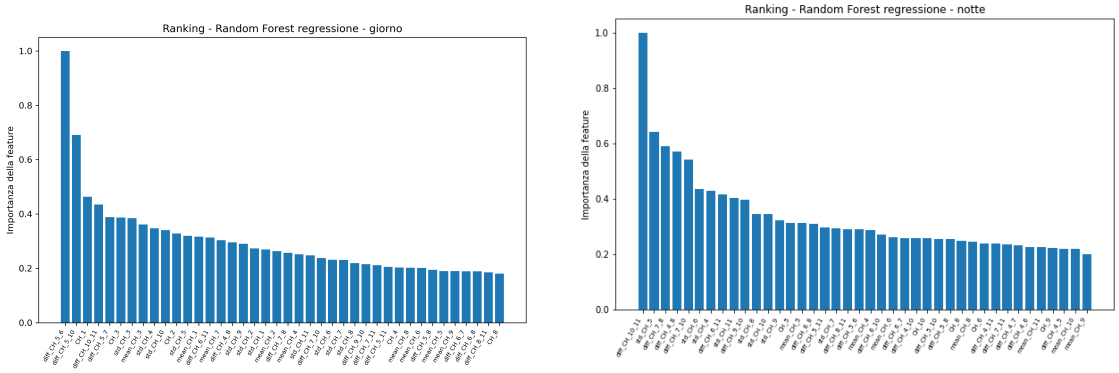


Figura 5.4: ROC: classificazione RF, 4 classi, giorno

Si evidenzia come la classificazione Random Forest è buona nel distinguere le classi di assenza di precipitazione e precipitazione leggera, mentre peggiora la sua abilità per fenomeni intensi. Nello specifico è molto buona la distinzione tra pioggia e non pioggia.

Per questa prima analisi non vengono riportati i risultati nel **caso notte**, che saranno discussi in dettaglio con i risultati del modello completo. In generale si può affermare che offrono una classificazione simile al caso giorno anche se con gli indici statistici generalmente inferiori.

Questa constatazione è supportata anche dal grafico di importanza delle features, riportate in Fig.5.5 per il caso giorno e notte:



(a) Giorno (b) Notte

Figura 5.5: Importanza delle features: regressione RF, 4 classi, giorno e notte

Nel caso giorno la **feature più importante** risulta essere la differenza tra i canali 5 e 6, mentre nel caso notte è la differenza tra i canali 10 e 11.

5.5 Regressione: Random Forest e Gradient Boosting

Per la **regressione** si valutano i due algoritmi di **Random Forest** e **Gradient Boosting**, in quanto il secondo, come già evidenziato, potrebbe fornire prestazioni migliori nel caso di dati con maggior dispersione. Si addestra il modello con un dataset diviso in 2 classi e la precipitazione massima a 30 mm/h.

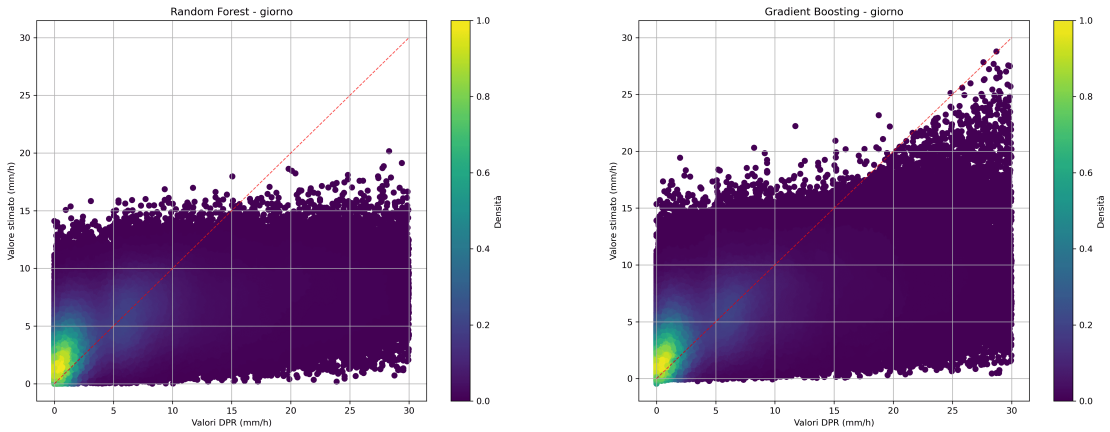
La regolazione degli iperparametri del modello in regressione è la stessa del caso in classificazione. Si riportano quindi i test di regressione nel caso giorno, per entrambi gli algoritmi, completi di tabella con i parametri di valutazione statistica e i grafici di scatter tra i valori di precipitazione misurati dal DPR e quelli stimati dall'algoritmo, con scala di colori che ne indica la densità.

La seguente Tab. 5.4 contiene i parametri statistici di valutazione per i due algoritmi, limitatamente all'uso del filtro giorno:

Indice	**Random Forest**	**Gradient Boosting**
MSE $((mm/h)^2)$	16.37	15.35
MAE (mm/h)	2.67	2.56
R^2	0.37	0.41
CC	0.61	0.64
CV	0.71	0.78
ME (mm/h)	0.12	0.18
MAE Normalizzato	0.69	0.66

Tabella 5.4: Indici di valutazione statistica per Random Forest e Gradient Boosting in regressione, giorno

In Fig.5.6 vengono mostrati i relativi grafici di scatterplot. É importante notare che la densità dei punti è calcolata mediante una densità normalizzata dei punti utilizzando una stima kernel di densità KDE di tipo gaussiano [48], che permette di distinguere maggiormente le differenze di densità.



(a) Random Forest (b) Gradient Boosting

Figura 5.6: Scatterplot DPR-Valore stimato: confronto tra la regressione Random Forest e Gradient Boosting, giorno

Analizzando i risultati, la regressione Gradient Boosting risulta leggermente **più accurata** della Random Forest: i parametri di errore, R^2 e gli altri indici statistici risultano essere più vicini all'ottimale, con le stime aventi un errore MSE e MAE inferiore. Dal grafico si nota come le coppie nel caso del Gradient Boosting tendano a distribuirsi, soprattutto per i valori vicini al limite superiore del dominio, più vicini alla diagonale, confermato da un CC più alto sintomo di una relazione più vicina ad essere di tipo lineare, e compensando parzialmente il problema della Random Forest di fornire output in una regione ristretta del dominio.

Entrambi gli algoritmi hanno un comportamento discreto nella stima dell'intensità di precipitazione, valutata sul dataset di test. Tuttavia, per la miglior accuratezza mostrata si sceglie di utilizzare il Gradient Boosting per la regressione nel modello completo descritto nella seguente Sezione 6.

6 Modello completo: Random Forest e Gradient Boosting

Si procede ora alla descrizione del modello finale, **completo a doppio livello** [49], nel quale i dati di input vengono prima classificati tramite una Random Forest in classi di intensità di precipitazione e in seguito si effettuano le regressioni Gradient Boosting all'interno di ogni singola classe per fornire una stima numerica in mm/h.

Lo schema seguito è il seguente in Fig.6.1:

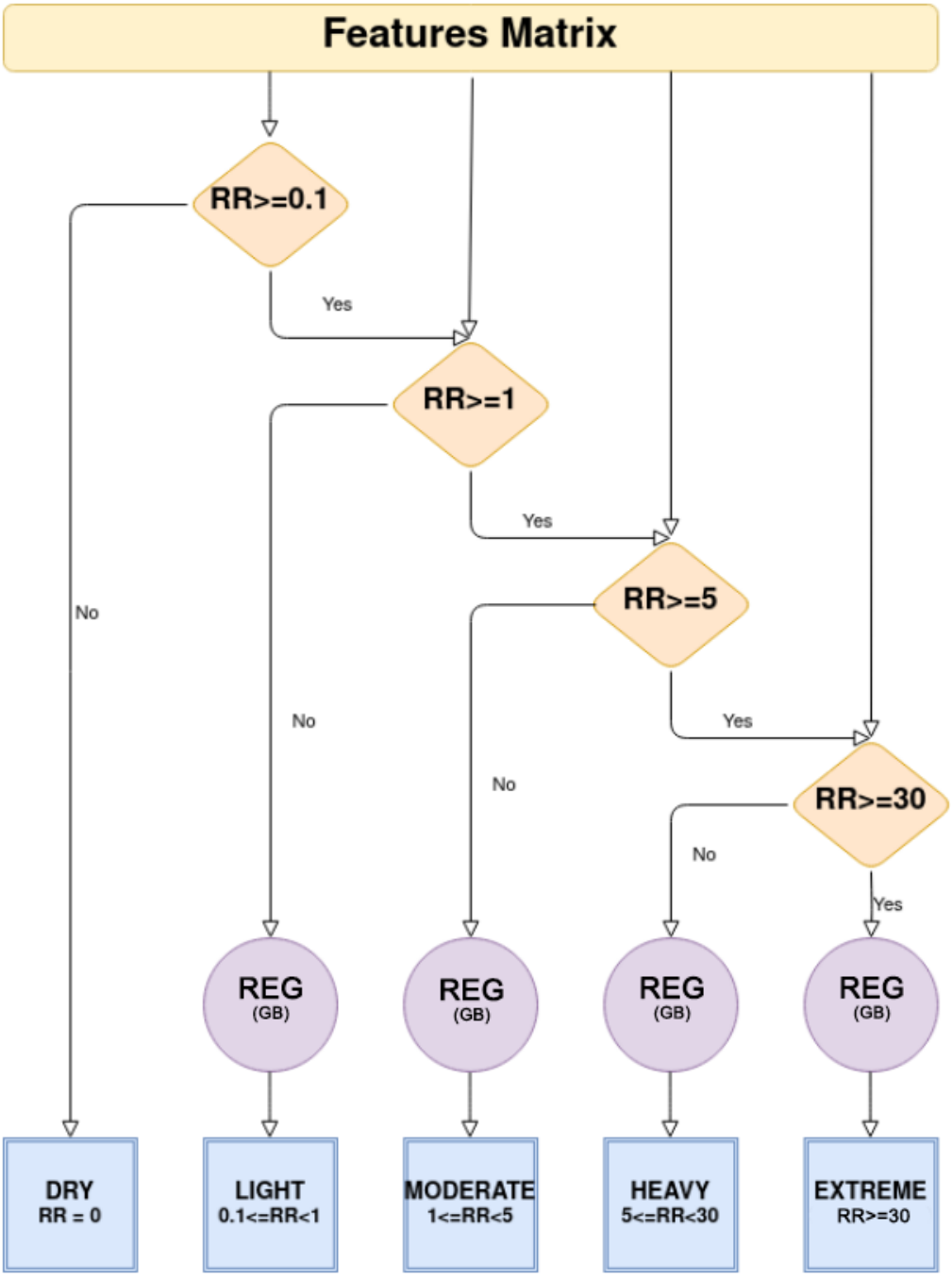


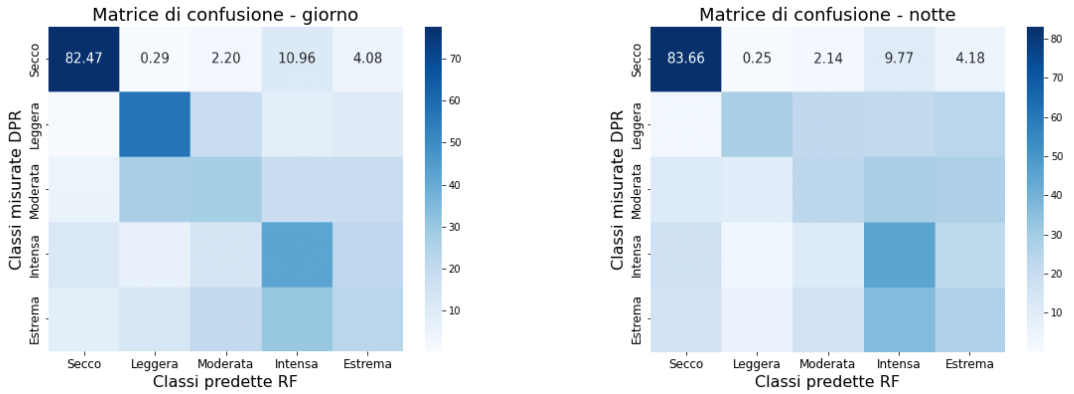
Figura 6.1: Schema modello completo a doppio livello: classificazione con Random Forest e regressione con Gradient Boosting per classe

Il **dataset** è composto dalle matrici di dati relativi all'intero **anno 2017**, filtrati secondo le caratteristiche già descritte. Viene effettuata una divisione in due set indipendenti di training e di test come già descritto per la fase di valutazione del modello nel paragrafo 5.4.

Si procede con il bilanciamento delle 5 classi, sia nel caso giorno che nel caso notte, e se ne discutono i risultati nei seguenti paragrafi.

6.1 Risultati modello completo: classificazione RF

Si applica l'algoritmo di classificazione Random Forest ottenendo le seguenti matrici di confusione in Fig.6.2:



(a) Giorno (b) Notte

Figura 6.2: Matrice di confusione: modello completo, classificazione RF. Anno 2017, 5 classi, giorno e notte.

Si riportano nelle seguenti tabelle, per i casi giorno e notte: i valori per le matrici di confusione in Tab.6.1 e un report generale per le varie classi in Tab.6.2:

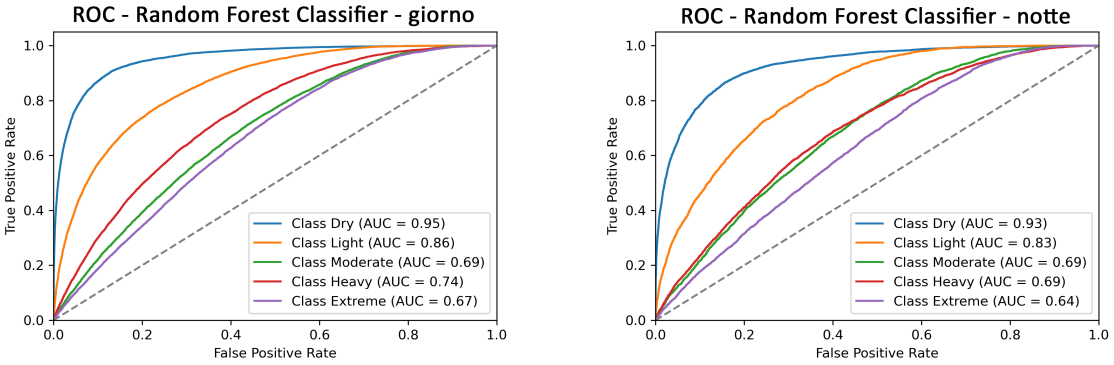
	RF: secco	**RF: leggera**	**RF: moderata**	**RF: intensa**	**RF: estrema**
Giorno					
DPR: secco	8887	31	237	1181	440
DPR: leggera	91	6314	2021	928	1054
DPR: moderata	654	3229	3341	2109	2211
DPR: intensa	1436	752	1661	5235	2654
DPR: estrema	1034	1577	2517	3792	2844
Notte					
DPR: secco	4924	15	126	575	246
DPR: leggera	70	724	577	565	614
DPR: moderata	422	350	896	1107	1055
DPR: intensa	894	149	572	2355	1192
DPR: estrema	716	251	751	1753	1240

Tabella 6.1: Matrice di confusione: modello completo, classificazione RF. Anno 2017, 5 classi, giorno e notte.

Classe	Precision	Recall	F1-Score
Giorno			
secco	0.79	0.85	0.82
leggera	0.64	0.76	0.70
moderata	0.52	0.48	0.50
intensa	0.47	0.37	0.41
estrema	0.44	0.33	0.34
Notte			
secco	0.73	0.84	0.78
leggera	0.57	0.69	0.61
moderata	0.49	0.48	0.45
intensa	0.41	0.37	0.39
estrema	0.39	0.32	0.33

Tabella 6.2: Report per classi: modello completo, classificazione RF. Anno 2017, 5 classi, giorno e notte.

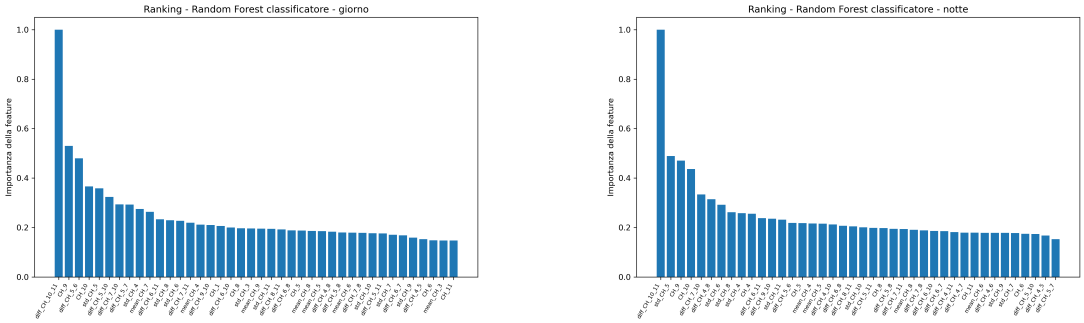
Si realizza una rappresentazione grafica per la ROC delle classi in Fig.6.3:



(a) Giorno (b) Notte

Figura 6.3: ROC: modello completo, classificazione RF. Anno 2017, 5 classi, giorno e notte.

Infine è riportata una classifica per importanza normalizzata delle prime 40 features, in Fig.6.4:



(a) Giorno (b) Notte

Figura 6.4: Importanza delle features: modello completo, classificazione RF. Anno 2017, 5 classi, giorno e notte.

Il modello completo sul dataset relativo all'anno 2017, sembra comportasi discretamente in classificazione, con una Accuracy media di 0.47 per valori diurni e 0.46 per valori notturni. La matrice di confusione evidenzia il buon riconoscimento della classe secco, avendo indici di Precision, Recall e F1-Score superiori alle altre classi, e un AUC più alto, mentre la classificazione peggiora per le classi di intensità di precipitazione crescente, mantenendo comunque una capacità di classificazione statisticamente significativa, avendo un valore HSS multiclasse di 0.27 nel caso diurno e 0.26 nel caso notturno.

Le features che hanno contribuito maggiormente alla classificazione si confermano essere le differenze dei canali 10 e 11, con un'importanza elevata per entrambi i casi diurno e notturno, e dei canali 5 e 6 per il solo caso diurno.

6.2 Risultati modello completo: regressione GB

Seguendo lo schema di funzionamento del modello, in seguito ala classificazione si realizza una regressione per ognuna delle classi con precipitazione.

I parametri statistici sull'algoritmo di Gradient Boosting per le 4 classi di precipitazione sono riportati di seguito. Nel caso giorno, in Tab. 6.3:

Indice	**leggera**	**moderata**	**intensa**	**estrema**
MSE $((mm/h)^2)$	0.0512	1.0487	23.8423	449.7091
MAE (mm/h)	0.1912	0.8448	3.6857	15.4780
R^2	0.0158	0.0358	0.0327	-0.0017
CC	0.1306	0.1927	0.1842	0.1099
CV	0.0426	0.0734	0.0792	0.0939
ME (mm/h)	0.0003	0.0017	0.0989	0.8546
MAE Normalizzato	0.3751	0.3807	0.3811	0.3171

Tabella 6.3: Indici statistici: modello completo, regressione GB. Anno 2017, 5 classi, giorno

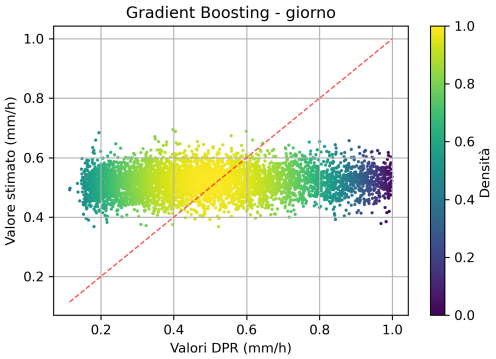
Nel caso notte, in Tab. 6.4:

Indice	**leggera**	**moderata**	**intensa**	**estrema**
MSE $((mm/h)^2)$	0.0501	1.0407	21.0289	351.5369
MAE (mm/h)	0.1887	0.8209	3.3764	13.4315
R^2	-0.0085	0.0065	-0.0133	-0.0315
CC	0.0842	0.1193	0.0800	0.0668
CV	0.0859	0.1001	0.1075	0.0985
ME (mm/h)	0.0108	-0.0129	0.1134	1.0219
MAE Normalizzato	0.3720	0.3874	0.3700	0.2952

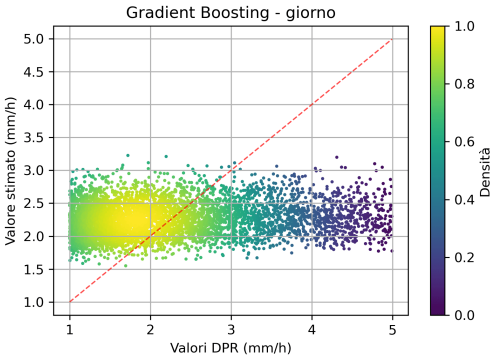
Tabella 6.4: Indici statistici: modello completo, regressione GB. Anno 2017, 5 classi, notte

La regressione Gradient Boosting sul modello completo, diviso in 5 classi, non è buona: i coefficienti R^2, secondo la formula descritta nel paragrafo 5.3, mostrano una debole correlazione lineare assumendo talvolta valori negativi. Questo è consentito nella definizione di R^2 adottata e fornisce l'informazione che il modello è peggiore della semplice media della classe. Gli altri indici di valutazione non mostrano segnali di una buona regressione rispetto i dati di test e gli errori sono crescenti con l'aumentare dell'intensità della precipitazione. Nel caso notturno la regressione è peggiore che nel caso diurno.

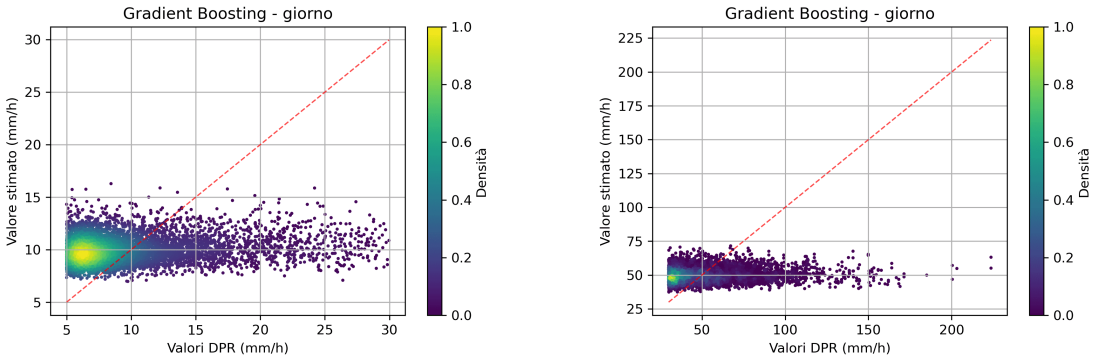
Si riportano i grafici di scatter, sia in scala logaritmica che lineare per le 4 regressioni effettuate.



(a) Pioggia leggera



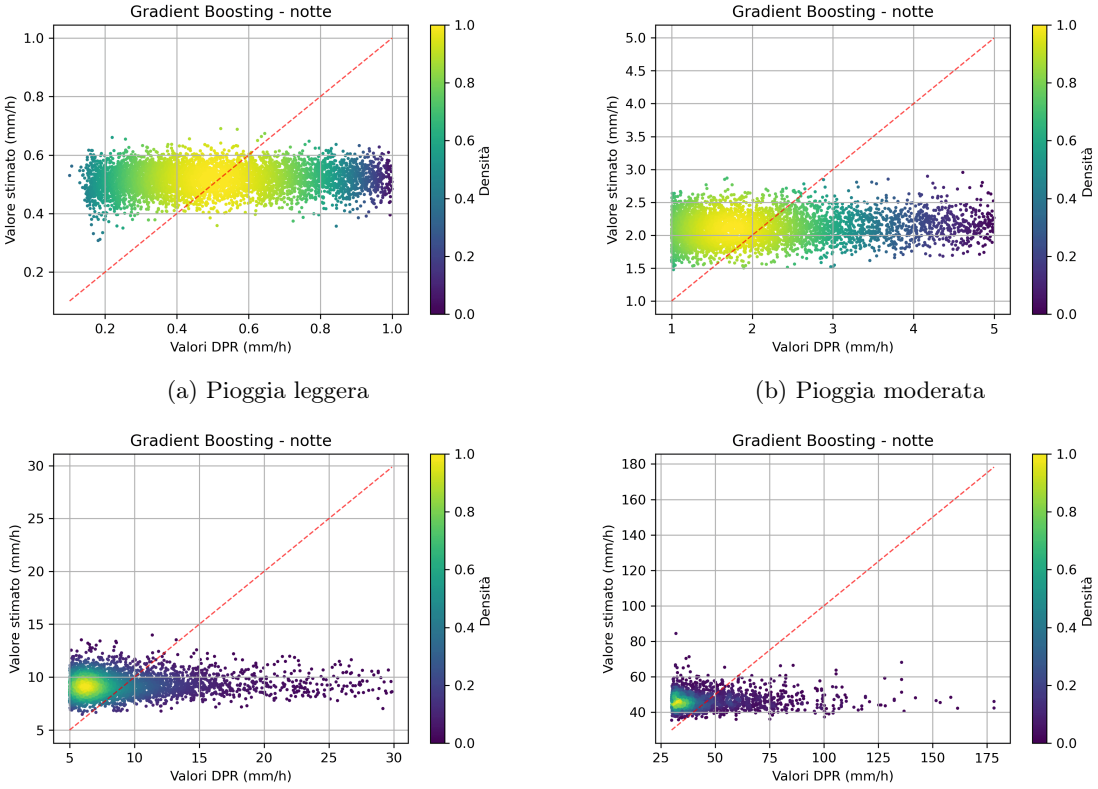
(b) Pioggia moderata



(a) Pioggia intensa (b) Pioggia estrema

Figura 6.6: Scatterplot: modello completo, regressione GB. Anno 2017, 5 classi, giorno.

Nel caso giorno in Fig.6.6, mentre nel caso notte in Fig.6.7;



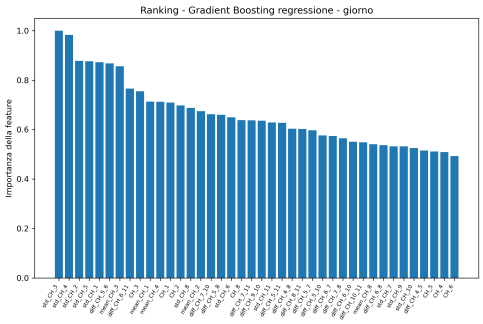
(a) Pioggia leggera (b) Pioggia moderata

(c) Pioggia intensa (d) Pioggia estrema

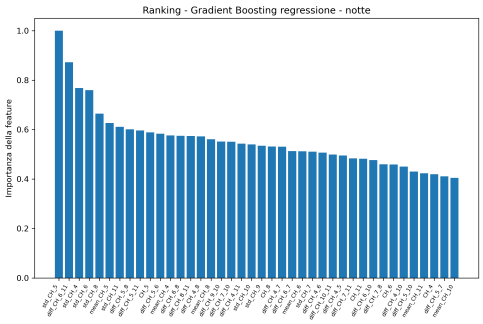
Figura 6.7: Scatterplot: modello completo, regressione GB. Anno 2017, 5 classi, notte.

I grafici di scatterplot evidenziano le problematiche riscontrate in ogni situazione, con i valori stimati che tendono a concentrarsi intorno al valore medio per le classi di precipitazione leggera e moderata. Al contrario, per le classi di precipitazione intensa e estrema si ottiene una evidente sottostima, poichè i valori del modello forniscono stime vicine al limite inferiore della classe.

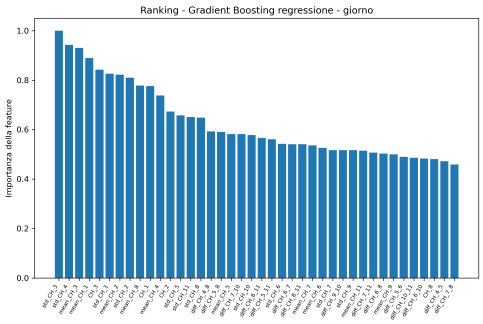
Si riportano infine l'importanza delle features, in Fig.6.9, per entrambi il caso giorno e notte;



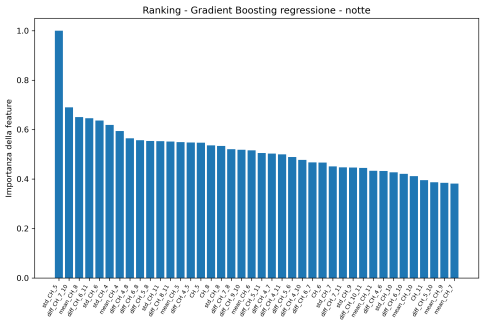
(e) Pioggia intensa, giorno



(f) Pioggia intensa, notte

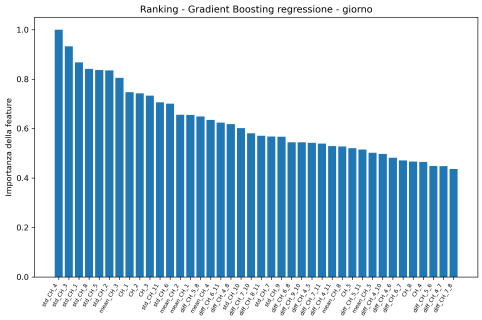


(g) Pioggia estrema, giorno

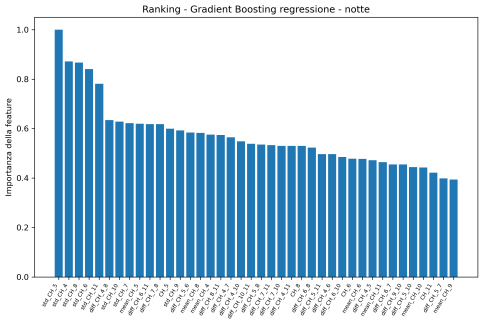


(h) Pioggia estrema, notte

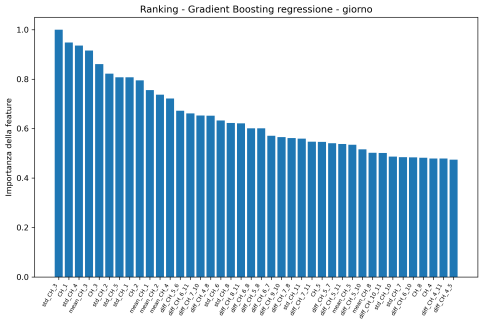
Figura 6.9: Importanza delle features: modello completo, regressione GB. Anno 2017, 5 classi, giorno e notte.



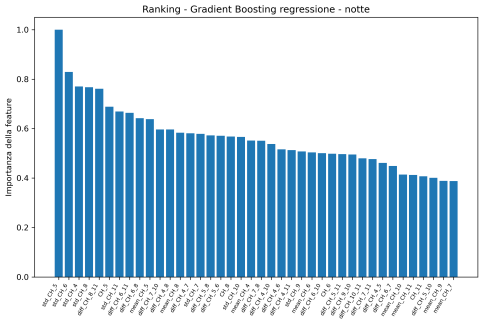
(a) Pioggia leggera, giorno



(b) Pioggia leggera, notte



(c) Pioggia moderata, giorno



(d) Pioggia moderata, notte

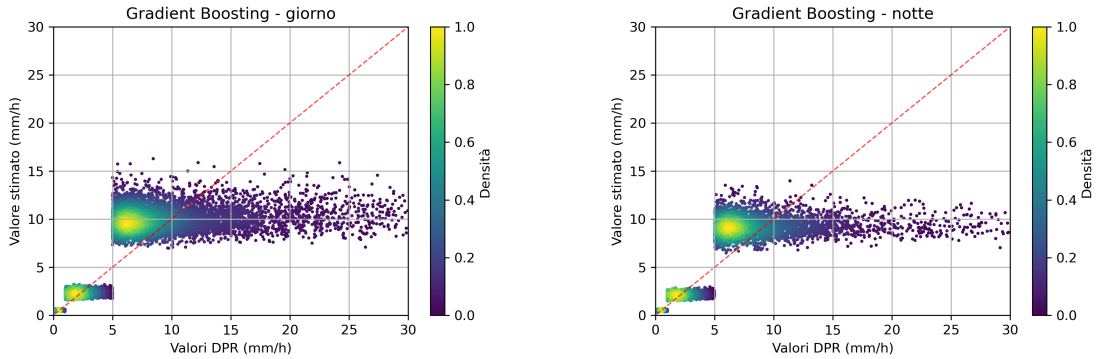
L'importanza delle features forniscono indicazioni utili per comprendere il motivo delle scarse capacità del modello nella regressione per singola classe: in tutte le situazioni hanno perso importanza tutte le features di differenza tra i canali, che potevano contenere maggior informazione fisica, a discapito delle features di media e deviazione standard mobile.

6.3 Osservazioni sui risultati del modello completo

Il modello completo risulta avere una buona abilità di classificazione, in particolare per le classi di precipitazione leggera e assenza di precipitazione. Gli indici associati vanno dal buono al discreto passando a intensità di precipitazioni estreme, peggiorando di poco di notte.

Il secondo livello delle regressioni, nonstante l'utilizzo del Gradient Boosting, presenta risultati non soddisfacenti con le stime del modello che si concentrano su fasce limitate per ogni classe.

Si riassume la regressione classificata dal modello completo per valori inferiori a 30 mm/h, per giorno e notte, nella seguente Fig.6.10:



(a) Giorno (b) Notte

Figura 6.10: Stima per valori $<$ 30 mm/h: modello completo, regressione GB. Anno 2017, 5 classi, giorno e notte.

In definitiva, il modello completo è **discreto nel classificare** gli eventi tra le 5 classi di precipitazione, ma è **inaccurato nel fornire una stima per regressione** all'interno di ogni classe.

Nonostante i buoni indici di classificazione per l'assenza di precipitazione, il modello classifica una parte valori della classe secca come classe di precipitazione, facendo ipotizzare una **sovrastima dell'area di precipitazione**. Inoltre la tendenza per classi di precipitazione intensa e estrema a fornire stime vicine al limite inferiore della classe, può far ipotizzare una **sottostima dell'intensità di precipitazione** più elevata.

7 Validazione del modello

La validazione è un processo critico per la valutazione delle **performance di un modello** in machine learning e la sua abilità di generalizzare su dati non visti durante la fase di addestramento, permettendo quindi di valutare se il modello soffre di under o overfitting, ovvero se è allenato a sufficienza e robusto nel processare nuovi dati.

7.1 Dataset di validazione

Per la validazione si utilizzano i dati raccolti da SEVIRI e le **misure pluviometriche** della rete di Protezione Civile italiana, costituite da una rete di circa 3000 pluviometri provenienti dalle Agenzie Regionali predisposte alla tutela dell'ambiente sul territorio. Si sceglie l'**Italia** come regione spaziale di verifica e il mese di **settembre 2019** come periodo temporale.

I dati di **SEVIRI** sono a libero accesso dal sito dell'EUMETAST Data Store, nella sezione "High Rate SEVIRI Level 1.5 Image Data - MSG - 0 degree" [50]: sono file in formato nativo .nat, per ognuno degli intervalli temporali di 15 minuti di misurazione di SEVIRI, su tutti gli 11 canali di osservazione nel visibile e nell'infrarosso. I dati forniti dall'EUMETSAT coprono tutto il FOV dello strumento, sarà quindi necessario applicare una maschera per limitare le misure all'area geografica di interesse, nello specifico per l'Italia.

Le misure della **rete di pluviometri** forniti dalla Protezione Civile, sono già elaborati e interpolati su una griglia spaziale: sono forniti valori orari in mm/h e una maschera geografica sul suolo italiano, coprendo l'intero mese di settembre 2019.

Analisi sinottica settembre 2019

Durante il mese di settembre 2019, l'Italia è stata interessata di frequente dall'espansione dell'Anticiclone Azzorriano verso il Mediterraneo con condizioni di prevalente stabilità, come mostra la reanalisi NCEP NOAA [51] sull'altezza media del geopotenziale a 850 hPa, in Fig.7.1.

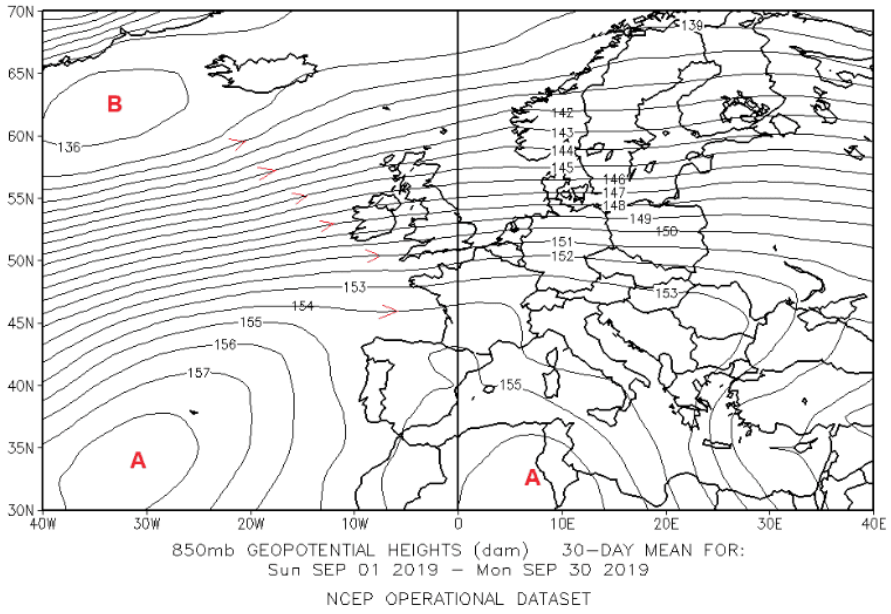


Figura 7.1: Settembre 2019: topografia media mensile dell'altezza di geopotenziale a 850 hPa.

Nonostante la circolazione media sia stata di stampo anticiclonico, la penisola è stata comunque raggiunta da 5 modeste perturbazioni, riassunte in Tab. 7.1:

Periodo	Fenomeno
1-5 settembre	Mentre l'arco alpino è influenzato da un promontorio anticiclonico ad ovest, sul centro del Mediterraneo si mantengono condizioni di instabilità per effetto di una saccatura associata ad un minimo anatolico (perturbazione n.1).
6-9 settembre	Sul golfo ligure si sviluppa una depressione isolata che dal giorno 7 viene riassorbita da una saccatura atlantica in espansione da Nord verso il Mediterraneo e in successivo moto verso est-nordest (perturbazione n.2).
10-11 settembre	Una nuova saccatura atlantica fa il suo ingresso sul Mediterraneo isolando un minimo di cut-off sulle Baleari che il giorno 11 interessa marginalmente la Sardegna (perturbazione n.3).
12-16 settembre	Sull'Italia si afferma un promontorio anticiclonico da Ovest.
17 settembre	L'arretramento dell'anticiclone espone l'Italia a un debole regime di correnti da nordovest.
18-19 settembre	Mentre l'arco alpino è influenzato da un promontorio anticiclonico da ovest l'areale italiano è interessato da una saccatura da est–sudest (perturbazione n.4).
20-24 settembre	L'Italia è inizialmente interessata da un promontorio mobile da sud in progressivo cedimento sotto la spinta di una saccatura atlantica in successivo rapido transito (perturbazione n.5).
25-30 settembre	Sull'Italia regime di correnti occidentali con condizioni di variabilità.

Tabella 7.1: Sintesi delle strutture circolatorie del mese di settembre 2019, con evidenziate le fasi caratterizzate da perturbazioni.

Alla luce di questa analisi meteorologica per il mese in esame, si valuta di validare il modello sui giorni **22 e 23 settembre 2019**, caratterizzati da diffusa instabilità su tutto il territorio italiano.

7.2 Preparazione alla validazione

Per ottenere, con l'algoritmo precedentemente addestrato, delle stime di intensità di precipitazione confrontabili con il campione pluviometrico, è necessario eseguire una **procedura di preparazione al modello**, che coinvolge la preparazione dei dati di SEVIRI e dei pluviometri, seguendo i seguenti passaggi:

1. **Maschera geografica sull'Italia**

   Ottenuti i dataset di SEVIRI e dei pluviometri per le giornate del 22 e 23 settembre 2019, è necessario individuare i confini territoriali italiani per limitare i valori di SEVIRI, che si estende su tutta la sua area di osservazione, e renderli così confrontabili con le misurazioni della rete di pluviometri.

   Insieme ai dati pluviometrici è stata fornita una griglia di coordinate e una maschera sul suolo italiano: si estrapolano da questa le informazioni per ritagliare i punti all'interno della maschera per entrambi i dataset. Si ottengono quindi due dataset di valori su griglie simili di circa $5x5km^2$ che coinvolgono i valori misurati esclusivamente sul suolo italiano.

   Per un confronto sui prodotti del modello con la precipitazione sarà necessario associare i punti di griglia della stima derivante da SEVIRI con l'intensità interpolata dalla rete pluviometrica: si utilizzera una tecnica di Nearest Neighbour, descritta nel passo 6 della procedura.

2. **Visualizzazione dei canali SEVIRI**

   Si procede alla visualizzazione alle immagini nei canali di SEVIRI, limitatamente ad un'area centrata sull'Italia. Avere una rappresentazione grafica dell'osservazione satellitare è importante per l'analisi preliminare dei risultati del prodotto in precipitazione del modello e il loro confronto con l'effettivo posizionamento delle nubi.

   Si riporta come esempio in Fig.7.2 un'immagine per il canale CH-9, a $10.8\mu m$, per il quarto d'ora antecedente le 13 UTC del 23 settembre 2019:

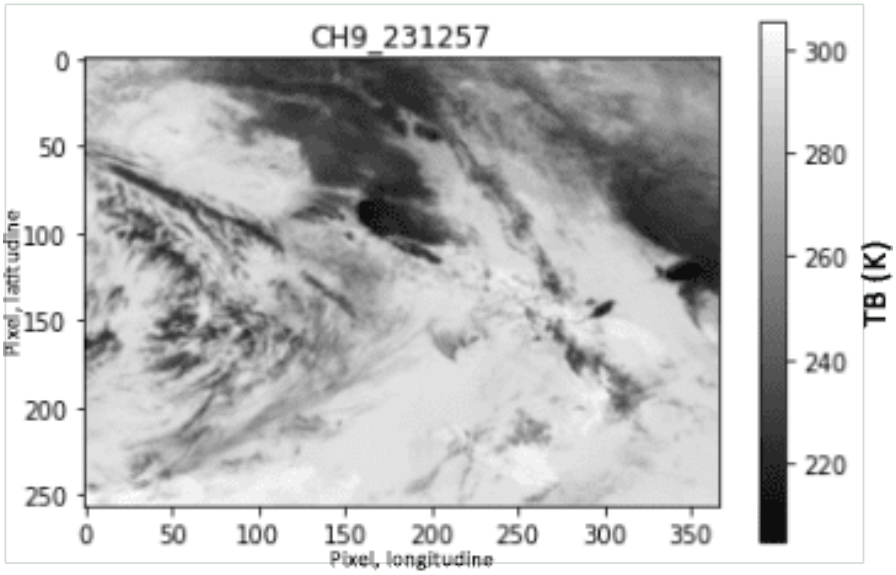


Figura 7.2: Esempio di immagine satellitare di SEVIRI, canale CH9 a 10.8μm, 23/09/2019 ore 12:57

3. **Creazione dataframe features e filtro giorno-notte su SEVIRI**

   Si organizzano i valori dei vari canali di SEVIRI e si realizzano le features utilizzate per l'addestramento del modello, ovvero la media mobile e la deviazione standard mobile su una finestra 5x5 di tutti i canali e le combinazioni di differenze tra i canali nell'infrarosso, preparando così i dataframe da fornire in input al modello.

   Si valuta un filtro per separare le osservazioni diurne da quelle notturne: essendo un periodo temporalmente limitato, si decide di utilizzare un orario fisso dipendente dalle condizioni di soleggiamento tipiche per l'Italia nel mese di settembre. Si considerano in condizione diurna tutti i campioni di dati raccolti in un orario compreso tra le 5 e le 17 UTC, notturna altrimenti.

4. **Stima di intensità di precipitazione ogni 15 minuti**

   La stima di intensità di precipitazione, in mm/h, viene fornita dal modello attraverso lo schema logico descritto nella Sezione 6: si procede ad una prima classificazione con la Random Forest nelle classi secco, leggera, moderata, intensa e estrema di precipitazione già definite e in seguito si stima un valore di precipitazione attraverso una regressione Gradient Boosting specifica per ogni classe di intensità.

   Per evitare risultati in output non bilanciati tra le varie classi, si è corretto il modello in classificazione, che era caratterizzato dall'addestramento su classi bilanciate, tenendo conto di un peso per ogni classe pari al numero di campioni in essa contenuta diviso il totale. In Tab. 7.2 sono riportati i numeri di campioni del dataset di addestramento del DPR prima del bilanciamento.

Categoria	secco	leggera	moderata	intensa	estrema
# Campione	569195125	1318929	10116698	5717230	779708

Tabella 7.2: Numerosità dei campioni per le diverse classi di intensità di precipitazione del dataset DPR utilizzato nel modello machine learning

Si ottengono così le stime ogni 15 minuti per entrambi i giorni scelti come caso studio. In Fig.7.3, ad esempio, si ha la stima per lo stesso intervallo temporale del quarto d'ora antecedente le 13 UTC del 23 settembre 2019:

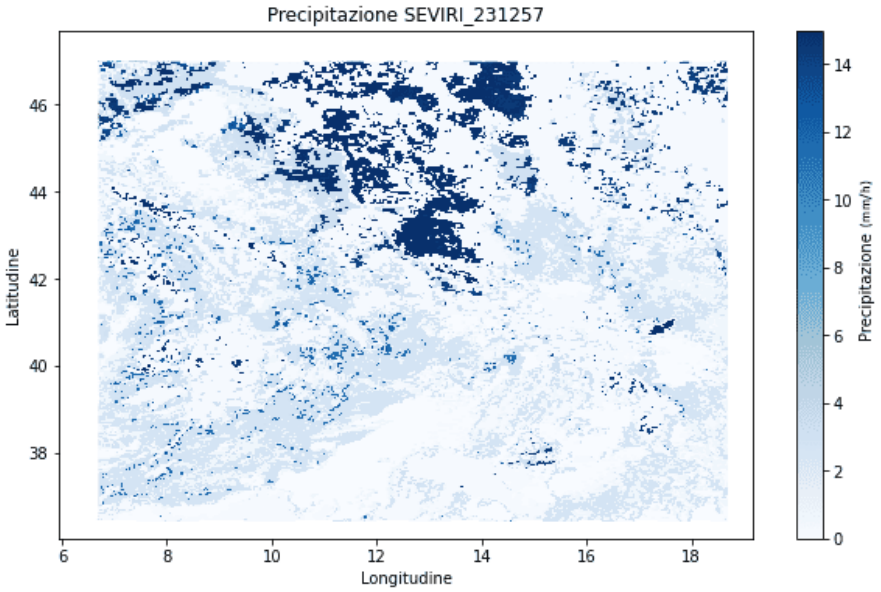


Figura 7.3: Esempio di stima di intensità di precipitazione del modello, 23/09/2019 ore 12:57

Si nota, confrontando tra loro la Fig.7.3 e la Fig.7.2, come le precipitazioni seguano la disposizione delle nubi e in particolare le precipitazione stimate più intense sono in corrispondenza delle nubi più fredde, e quindi più sviluppate verticalmente. Questo è verosimile, tuttavia sembra evidente che l'area di precipitazione è eccessivamente estesa, essendo stimata su gran parte dell'area occupata dalle nubi, mostrando quindi una possibile sovrastima della diffusione spaziale della precipitazione.

5. **Stima di intensità di precipitazione oraria**

   Essendo le misure dei pluviometri orarie, si rende necessario ricavare una stima di tipo orario anche con i risultati del modello. Per raggiungere questo obiettivo si mediano le quattro stime di 15 minuti presenti nella medesima ora, ottenendo un valore di intensità di precipitazione in mm/h.

   In Fig.7.4 si riporta un esempio di stima aggregata per l'ora antecedente le 13 UTC nel giorno 23 settembre 2019:

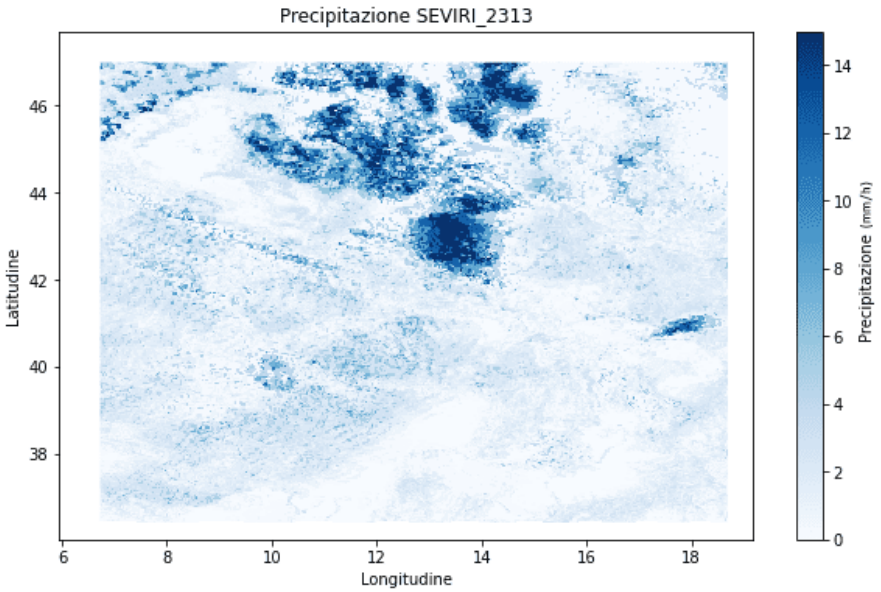


Figura 7.4: Esempio di stima di intensità di precipitazione oraria del modello, un'ora antecedente le 13 UTC del 23/09/2019

6. **Nearest Neighbour tra le griglie della stima e della rete di pluviometri**

Avendo la stima per l'intensità di precipitazione del modello e i valori misurati dalla rete pluviometrica, per ogni ora, è necessario confrontare i rispettivi valori spazialmente corrispondenti. Avendo e due griglie spaziature confrontabili, pari a 5x5 km per i pluviometri e circa 4.5x5 per SEVIRI, si valuta di utilizzare la tecnica del Nearest Neighbour per il confronto.

Il metodo del **Nearest Neighbour** è una tecnica di confronto che viene utilizzata per riportare i dati da una griglia ad un'altra, basandosi sull'assunzione che punti vicini nello spazio mantengano una continuità e che la risoluzione spaziale delle due griglie sia confrontabile. Per determinare il valore di una osservazione in un nuovo punto di griglia, si considerano i valori delle osservazioni già note che sono più vicine ad esso, secondo una misura di distanza euclidea.

Formalmente, il metodo del Nearest Neighbour può essere espresso come segue: siano $P = \{p_1, p_2, \ldots, p_n\}$ un insieme di punti sulla griglia dei valori stimati dal modello, e $Q = \{q_1, q_2, \ldots, q_m\}$ un insieme di punti sulla griglia della rete pluviometrica, per ogni punto $q_i \in Q$, si trova il punto più vicino $p_j \in P$ utilizzando una misura di distanza euclidea. Questa distanza è ottenuta a partire dalle coordinate terrestri dei punti di griglia, convertite in una proiezione che rappresenta distanze sulla superficie terrestre usando la libreria "pyproj" e la proeizione di Mercatore, tramite una funzione "cKDTree" della libreria SciPy. Il valore interpolato $\hat{v}(q_i)$ in q_i è dato quindi da:

$$\hat{v}(q_i) = v(p_j) \quad \text{dove} \quad j = \arg\min_k \|q_i - p_k\|$$

dove $\|q_i - p_k\|$ rappresenta la distanza euclidea tra i punti q_i e p_k, e $v(p_j)$ è il valore del dato stimato nel punto p_j.

7. **Interpretazione statistica dei risultati**

   L'ultima fase della procedura riguarda l'interpretazione statistica dei risultati e la realizzazione di grafici esplicativi, descritti nella seguente sezione Risultati. In Fig.7.5, è riportato, a titolo di esempio, un confronto tra le misure della rete pluviometrica e la stima del modello per l'ora antecedente le 13 UTC nel giorno 23 settembre 2019:

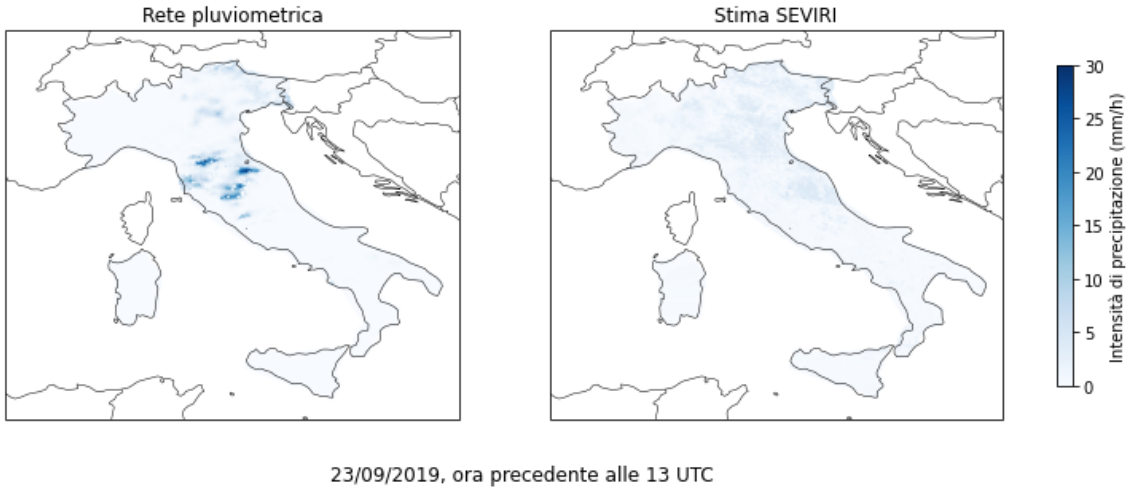


Figura 7.5: Esempio di confronto tra la precipitazione misurata dalla rete pluviometrica e la stima del modello basata sulle misure di SEVIRI, un'ora antecedente le 13 UTC del 23/09/2019

Ad una prima analisi, risulta evidente il problema già evidenziato della sovrastima dell'area di precipitazione. Tuttavia le aree con precipitazione stimata maggiore sembrano coincidere tra le due immagini.

Inoltre si riporta in Fig.7.6, la stima per lo stesso evento realizzata da IMERG, ottenunda tramite l'applicativo "Giovanni" della NASA. [52] IMERG, acronimo di Integrated Multi-satellitE Retrievals for Global Precipitation Measurement (Recupero Integrato dei Dati di Precipitazione Multi-satellitari per la Misurazione Globale delle Precipitazioni), è un prodotto sviluppato dalla NASA che fornisce stime globali di precipitazioni a alta risoluzione spaziale e temporale, utilizzando dati da diverse fonti satellitari, incluso il Dual-frequency Precipitation Radar (DPR) a bordo del satellite GPM.

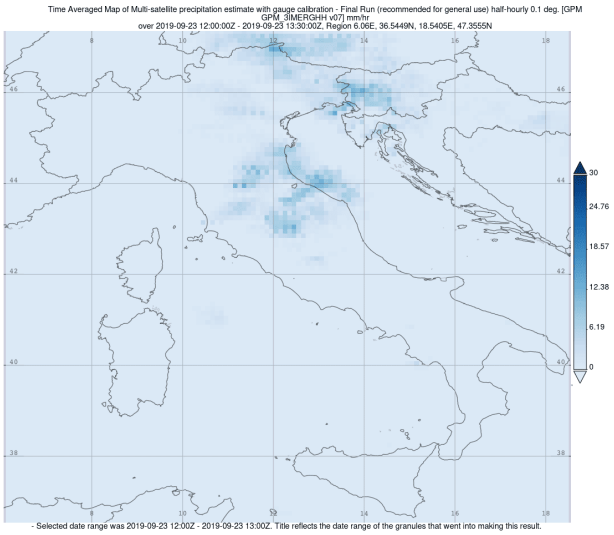


Figura 7.6: Stima della precipitazione IMERG (DPR) sull'Italia, intorno alle 13 UTC del 23/09/2019

Il prodotto IMERG, che tiene conto dei prodotti in precipitazione del DPR, evidenzia la sovrastima del modello machine learning basato su SEVIRI, addestrato proprio sul DPR.

7.3 Risultati

Il modello viene fatto operare su tutti gli intervalli orari del 22 e 23 settembre 2019, ottenendo 48 immagini di confronto e relative matrici sul prodotto di intensità di precipitazione. É necessario ora valutare tramite indicatori statistici la bontà del modello. Per fare questo si fornisce una descrizione prima su un orario di esempio, quello del 23 settembre 2019 ore 13 UTC, e in seguito cercando di ottenere una statistica complessiva su entrambi i giorni del caso di studio.

Statistica 23/09/2019 ore 13 UTC

Continuando l'esempio del 23 settembre 2019, riguardo le precipitazioni per l'ora precedente alle 13 UTC, si realizza un grafico di Bland-Altman e si riportano alcuni indici statistici già descritti nel capitolo per il test del modello.

Il test di **Bland-Altman** è un metodo statistico utilizzato per valutare la concordanza tra due diverse tecniche o strumenti di misura che intendono valutare uno stesso fenomeno. Il grafico di Bland-Altman rappresenta graficamente le differenze tra le misure ottenute dai due metodi contro la media delle due misure: sull'asse delle ordinate vengono riportate le differenze tra i valori misurati dai due metodi ($d_i = M1_i - M2_i$) e sull'asse delle ascisse viene riportata la media delle misure ottenute dai due metodi ($\overline{M}_i = \frac{M1_i + M2_i}{2}$). Questo grafico permette di visualizzare eventuali bias tra i metodi e di identificare la presenza di eventuali errori sistematici o di proporzionalità.

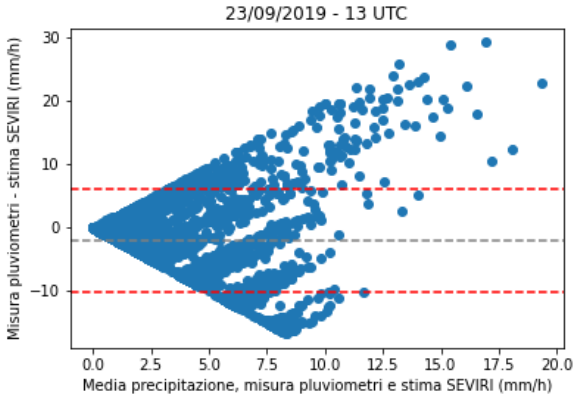


Figura 7.7: Grafico di Bland-Altman per la valutazione della concordanza tra le misure della rete pluviometrica e le stime del modello basato su SEVIRI, 23/09/2019 ore 13 UTC

In Fig.7.7 si riporta il grafico di Bland-Altman per la concordanza tra le misure di precipitazione dei pluviometri e le stime del modello basato su SEVIRI. La linea grigia orizzontale rappresenta il bias sistematico mentre le linee tratteggiate rosse rappresentano i limiti entro i quali ci si aspetta che cadano il 95% delle differenze.

Il grafico mette quindi in evidenza che, in media, le stime del modello sono poco superiori rispetto alle misure della rete di pluviometri, a causa di una sovrastima delle misure a intensità di precipitazione leggera. Tuttavia per precipitazioni moderate e intense il modello tende a sottostima la sua stima.

É riportata la matrice di confusione per la classificazione delle stime del modello basato su SEVIRI e le misure della rete pluviometrica, con la percentuale di stima corretta, in Tab. 7.3:

	Modello SEVIRI				
Misure	secco	leggera	moderata	intensa	estrema
secco	1209	4337	2336	1125	0
leggera	9	370	332	561	0
moderata	6	307	390	545	0
intensa	0	121	203	142	0
estrema	0	0	1	1	0

Classe	Correttezza
secco	1209 su 9007 (13%)
leggera	370 su 1272 (29%)
moderata	390 su 1248 (31%)
intensa	142 su 466 (31%)
estrema	0 su 2 (0%)

Tabella 7.3: Matrice di confusione e percentuale di correttezza per ciascuna classe, 23 settembre 2019 ore 13 UTC

L'accuratezza (Accuracy) totale del modello è 0.10 e in generale mostra una scarsa capacità di previsione del fenomeno di precipitazione: il modello tende ad assegnare valori falsi positivi per le classi di pioggia leggera, moderata e intensa, confermando la sovrastima dell'area di precipitazione.

Statistica complessiva dei giorni 22 e 23 settembre 2019

Si riassumono tutti i risultati delle due giornate del 22 e 23 settembre 2019, realizzando gli istogrammi (in Fig.7.8), con larghezza dell'intervallo di 1 mm/h, di distribuzione normalizzata dei valori stimati e misurati:

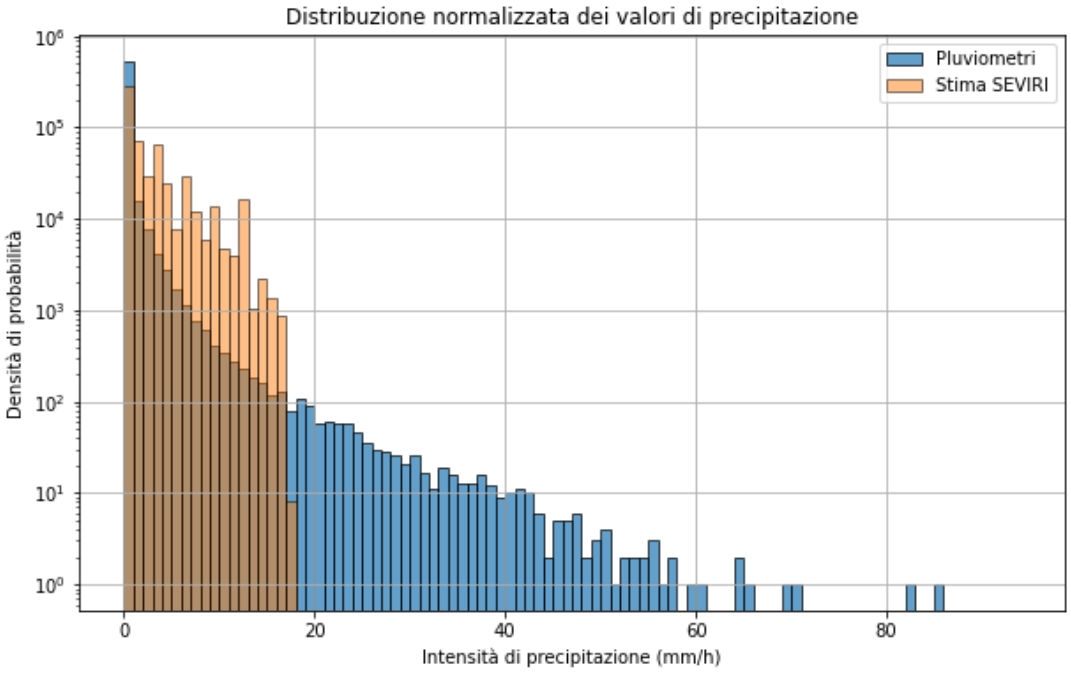


Figura 7.8: Istogramma di distribuzione per le stime del modello e le misure della rete pluviometrica sull'Italia il 22-23 settembre 2019

Anche nell'analisi complessiva dei due giorni emergono i due problemi: sovrastima dell'area di precipitazione e sottostima dell'intensità per classi di pioggia intense e estreme.

Si realizza la matrice di confusione complessiva, data dalla somma di tutti i valori previsti per le 5 classi nei due giorni del caso studio, e vari indici statistici di valutazione, escludendo la classe di pioggia estrema che il modello non è mai riuscito a stimare. La matrice di confusione totale, che riassume le predizioni del modello rispetto alle misure della rete pluviometrica, è riportata in Tab. 7.4:

	Modello SEVIRI				
Misure	secco	leggera	moderata	intensa	estrema
secco	23934	224776	161959	67139	0
leggera	486	22912	21056	16376	0
moderata	129	8758	8607	12653	0
intensa	60	1415	1800	3462	0
estrema	1	48	83	106	0

Tabella 7.4: Matrice di confusione complessiva per i giorni 22 e 23 settembre 2019

La percentuale di conteggi corretti rispetto ai conteggi totali per ciascuna classe è la seguente: secco 23934 su 477808 (5%), leggera 22912 su 60830 (37.7%), moderata 8607 su 30147 (28.6%), intensa 3462 su 6737 (51.4%) e estrema 0 su 238 (0%). Anche nel contesto complessivo il modello non ha stimato nessun valore di precipitazione estrema, confermando la sua sottostima a maggiori intensità di precipitazione.

Le metriche di valutazione calcolate sono riportate nella Tabella 7.5.

	secco	leggera	moderata	intensa
Precision	0.97	0.088	0.045	0.035
Recall	0.050	0.38	0.29	0.52
F1-Score	0.095	0.14	0.077	0.065
CSI	0.05	0.078	0.040	0.034
POD	0.050	0.38	0.29	0.52
FAR	0.028	0.91	0.95	0.96

Tabella 7.5: Sommario degli indici statistici, 22 e 23 settembre 2019

L'Accuracy complessiva del modello è del 10.23%, indicando una precisione globale limitata nel classificare correttamente le precipitazioni.

La Precision per la categoria secco è molto alta (97.25%), ma significativamente bassa per le altre classi indicando che il modello è quasi sempre corretto nella predizione di secco presentando tuttavia un alto numero di falsi positivi per tutte le altre categorie.

Il Recall è molto basso per la classe secco (5.01%), mentre per la classe di pioggia intensa raggiunge il 51.39%, suggerendo che il modello cattura correttamente solo una piccola frazione dei casi reali complessivi. Questa è un'ulteriore conferma della sovrastima dell'area di precipitazione, a causa dell'assegnazione di valori di pioggia a condizioni secche.

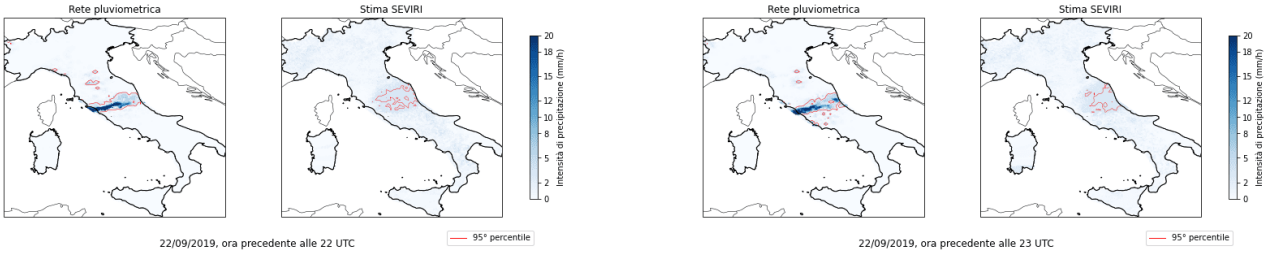
Gli F1-Score, che rappresentano un bilancio tra precisione e richiamo, e la probabilità di rilevamento (POD) sono generalmente bassi, soprattutto in assenza di precipitazione, indicando una scarsa performance complessiva.

Il False Alarm Ratio (FAR), che indica la proporzione di falsi allarmi, è molto basso per "secco" (2.75%), ma estremamente alto per le altre categorie, indicando che la maggior parte delle predizioni di precipitazione sono falsi allarmi.

L'analisi statistica ha evidenziato quindi l'alto numero di falsi positivi per le classi di precipitazione e un numero di stime di condizioni secche, seppur corretta, decisamente sottostimato. Inoltre il modello non riesce a prevedere correttamente l'intensità di precipitazione estrema, stimando valori di precipitazione inferiore. É confermata quindi la sovrastima dell'area di precipitazione e la sottostima dell'intensità per precipitazioni più intense.

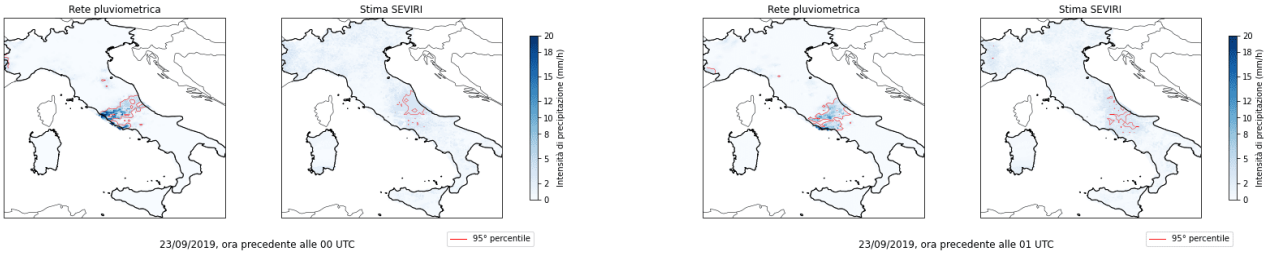
7.3.1 Esempio di sequenza su un intervallo

Si riporta una breve sequenza evolutiva di 4 ore, dalle 22 UTC del 22 settembre alle 02 UTC del 23 settembre 2019, del transito di una perturbazione dalla Toscana alla Campania, in Fig.7.9. Si rappresenta anche una linea di contorno per i valori di precipitazione sopra al 95° percentile, sia per la misura dei pluviometri che per la stima del modello, in modo da individuare le zone di maggior precipitazione:



(a) 22/09/19 ore 22 UTC

(b) 22/09/19 ore 23 UTC



(c) 23/09/19 ore 00 UTC

(d) 23/09/19 ore 01 UTC

Figura 7.9: Confronto tra la stima di precipitazione del modello ML su SEVIRI e le misure della rete di pluviometri, dalle 22 UTC del 22/09/2019 alle 02 UTC del 23/09/2019

Nonostante i problemi evidenziati nelle capacità predittive del modello, si osserva comunque una discreta capacità nel seguire il nucleo di maggior intensità di precipitazione, in corrispondenza di nubi fredde rilevate da SEVIRI, suggerendo l'importanza di questo paramento per un eventuale miglioramento del modello.

In riferimento alla validazione del modello sui valori dei pluviometri, i **risultati erano attesi** in quanto le stime di precipitazione del DPR, utilizzato per addestrare il modello, evidenziano differenze importanti in confronto con i valori della rete di pluviometri per il suolo italiano, come descritto dal lavoro di Petacca et al (2018) [53], in accordo con i seguenti indici statistici [54]: ME = -0.70 mm/h, RMSE = 3.36 mm/h, errore standard frazionario FSE = 151% e coefficiente di correlazione di Pearson CC = 0.55.

Conclusioni

L'obiettivo del lavoro di tesi è la stima quantitativa della precipitazione (QPE) tramite misure satellitari. Nello specifico, si utilizzano per questo scopo gli undici canali, 3 nel visibile in riflettanza (%) e 8 nell'infrarosso in temperatura di brillanza (K), di **SEVIRI**, strumento ottico equipaggiato sul satellite MSG dell'EUMETSAT. Il prodotto di precipitazione del **DPR** in coincidenza con SEVIRI, fornito dall'ISAC CNR di Roma, è utilizzato per l'addestramento dei valori di SEVIRI tramite un **modello machine learning a doppio livello**: classificazione Random Forest e regressione Gradient Boosting.

É stata svolta un'analisi statistica volta alla **descrizione del dataset** di SEVIRI relativo all'anno 2017, evidenziandone la frequenza di occorrenza, tramite grafici di distribuzione, dei valori suddivisi nelle diverse classi di intensità di precipitazione. Le classi di precipitazione scelte sono: secco ([0, 0.1) mm/h), leggera ([0.1, 1) mm/h), moderata ([1, 5) mm/h), intensa ([5, 30) mm/h) e estrema ($\geq 30 mm/h$). É importante osservare come i grafici di distribuzione nei diversi canali mostrino istogrammi con valore massimo distinguibile tra le differenti classi: all'aumentare della precipitazione la distribuzione tende ad avere un massimo a riflettanza maggiore, nel visibile, oppure a temperatura di brillanza minore, nell'infrarosso. Tuttavia gli istogrammi hanno aree di intersezione molto grande, soprattutto per i canali a minor lunghezza d'onda, mostrando quanto siano deboli le relazioni tra queste osservazioni e la precipitazione. Si valuta quindi l'utilizzo di algoritmi in machine learning basati su alberi decisionali, di tipo Random Forest e Gradient Boosting, sia per problemi in classificazione che in regressione, applicandoli al dataset di SEVIRI e addestrandoli sulla precipitazione fornita dal DPR.

Il dataset è stato preliminarmente preparato con un filtraggio sui valori errati del DPR, del CH-4 di SEVIRI e della temperatura del CH-9, sul posizionamento geografico per evitare deformazioni dell'immagine e sulla distinzione tra condizione diurna e notturna, basata sul valore di entropia delle immagini nel visibile. In seguito alla scelta delle classi viene effettuato un bilanciamento per preparare il **dataframe delle features**, comprendendo anche i valori di media e deviazione mobile su finestre di 5x5 su tutti i canali e una combinazione delle differenze tra i canali nell'infrarosso.

Applicando una Random Forest in **classificazione**, si ha una buona performance del modello avendo una Accuracy complessiva di 0.60. Il riconoscimento della situazione di assenza di precipitazione e ben ottenuta, con l'85.10 % di valori correttamente classificati, mentre per classi di maggior intensità di precipitazione peggiora la prestazione del modello, con i parametri di Precision, Recall e F1-Score minori e un numero elevato di falsi positivi. C'è un leggero vantaggio prestazionale per la condizione diurna a discapito della prestazione notturna e la classifica delle features più importanti mostra come svolgono un ruolo significativo le differenze dei canali 5 e 6, di giorno, e 10 e 11, di notte.

Nel problema di **regressione**, si valuta la prestazione di entrambi gli algoritmi Random Forest e Gradient Boosting. Ne risulta che il Gradient Boosting ha indici statistici leggermente migliori rispetto al Random Forest ottenendo, sullo spettro di intensità di precipitazione tra 0 e 30 mm/h, un R^2 di 0.41 per il Gradient Boosting e 0.37 per Random Forest. Tuttavia l'accuratezza complessiva di entrambi gli algoritmi non è ottima, e si osservano errori MSE (15.35 $(mm/h)^2$ per il GB) e MAE (2.56 mm/h per il GB) elevati. Dallo scatterplot è evidente come ci sia troppa dispersione nei dati e una tendenza a concentrarsi sul valore medio della classe, seppur meno accentuata con l'algoritmo Gradient Boosting. Per queste motivazioni si sceglie di utilizzare il Gradient Boosting come algoritmo in regressione nel modello completo.

Si realizza e valuta il **modello completo a doppio livello**, classificazione RF e regressione GB sulle 4 classi di precipitazione. Ne risulta che la classificazione risulta ancora buona, soprattutto per le classi secco e pioggia leggera, peggiorando la capacità di classificare per precipitazioni moderate, intense e estreme. I risultati in classificazione sono quindi in linea con le attese.
In regressione, invece, il modello non è buono e fornisce una bassa o nulla accuratezza. Per le classi con pioggia leggera i valori predetti dal modello tendono a concentrarsi attorno al valore medio della classe, mentre per le classi di intensità superiore tendono ad avvicinarsi sempre più al limite inferiore della classe, ottenendo come risultato una sottostima sempre maggiore dell'intensità delle precipitazioni da moderate a estreme.
Alla luce di queste osservazioni ci aspettiamo una debole **sovrastima dell'area di precipitazione**, a causa della predizione errata di circa un 20% dei valori della classe secca e una **sottostima dell'intensità di precipitazione** dalla moderata alla estrema per via del comportamento descritto nella regressione Gradient Boosting.

Il modello viene **validato** attraverso il confronto con i valori misurati dalla rete pluviometrica italiana gestita dalla Protezione Civile. Di conseguenza si adatta il dataset di SEVIRI per un confronto con la griglia dei pluviometri sul suolo italiano, per un evento di precipitazione avvenuto durante i giorni del 22 e 23 settembre 2019. La scelta di validare il modello sui pluviometri, nonostante l'addestramento sia stato effettuato tramite un radar satellitare, è dovuta al fatto che questi, in alcune nazioni come l'Italia, sono ancora gli **strumenti di riferimento** risultando un punto di riferimento per le stime e per la bontà delle prestazioni del modello.

I risultati del modello rispetto al dataset di validazione sono **scadenti**: i punti deboli del modello sono stati evidenziati dal confronto con la rete di pluviometri, ottenendo un alto numero di falsi positivi per tutte le classi di precipitazione. Si conferma quindi una sovrastima dell'area di precipitazione e una sottostima evidente dell'intensità per classi di pioggia moderata e intensa, mancando completamente la rivelazione di eventi di tipo estremo. Questo **risultato era atteso** in quanto l'addestramento è stato realizzato sui valori di precipitazione del DPR, che presentano differenze evidenti con le misure di una rete di pluviometri, come analizzato nello studio di Petracca et al. (2018) [53].

Tuttavia, individuando le aree con precipitazione sopra il 95° percentile per l'intensità, si evidenzia una **discreta concordanza dei nuclei più intensi di precipitazione**.

In conclusione del lavoro svolto, ed evidenziati i problemi riscontrati dal modello, è possibile individuare le seguenti **possibilità di sviluppo**:

- É possibile effettuare un bilanciamento delle classi meno riduttivo sul campione, permettendo di perdere meno informazione del dataset, a discapito di una maggior richiesta di potenza computazionale.
- Si potrebbe indagare un algortimo che gestisca meglio la regressione, soprattutto quella su classi. Si potrebbe provare usando reti neurali convoluzionali [55] invece che algortimi ad alberi decisionali.
- Individuare nuove features, come la variazione temporale su uno stesso punto o l'aggiunta dei dati sui fulmini.
- Adattare il modello al prossimo strumento ottico dell'EUMETSAT, equipaggiato sul satellite MTG [56]. Una miglior risoluzione spaziale, temporale e di qualità del segnale sicuramente permetteranno di indagare meglio le deboli relazioni tra le misure satellitari e la precipitazione.

Ringraziamenti

Questa attività è svolta all'interno del progetto H-SAF, progetto dell'EUMETSAT che si occupa di sviluppare, implementare e gestire prodotti satellitari per il monitoraggio idrologico e meteorologico.

Per la realizzazione di questo lavoro di tesi, è risultato fondamentale il supporto dei ricercatori del CNR-ISAC di Roma, che mi hanno fornito il dataset da loro elaborato con la coincidenza SEVIRI-DPR, per l'intero anno 2017: si ringraziano i dott.ri Daniele Casella, Paolo Sanò e Leo Pio D'Adderio.

Si ringrazia l'EUMETSAT, poichè attraverso il loro servizio a libero accesso ho potuto avere accesso ai dati di SEVIRI ad alta risoluzione.

Si ringrazia la NASA, poichè attraverso il loro servizio a libero accesso ho potuto avere accesso ai dati del DPR.

Si ringrazia il Dipartimento della Protezione Civile, per aver fornito i dati sul mese di settembre 2019 della rete di pluviometri gestita da loro in collaborazione con gli enti di tutela ambientale.

Ringrazio il mio relatore, il prof. Federico Porcù, per il tempo che mi ha dedicato e gli insegnamenti che mi ha donato, permettendomi di migliorare le mie conoscenze e capacità sotto tutti i punti di vista.

Riferimenti bibliografici

[1] John M. Wallace e Peter V. Hobbs. *Atmospheric Science: An Introductory Survey*. 2nd. Academic Press, 2006.

[2] Ulises Zavala Moràn. "Fabrication and characterization of III-V-Sb based optoelectronic devices for MWIR photodetection". Advisor: Philippe Christol, Francisco de Anda, Jean-Philippe Perez. Ph.D. Thesis. 2021. DOI: 10.13140/RG.2.2.29585.35685.

[3] Pao K. Wang. *Physics and Dynamics of Clouds and Precipitation*. Cambridge University Press, 2013.

[4] Grische Pierre_cb. *LFC-NCL*. https://commons.wikimedia.org/w/index.php?curid=32054437. File:LFC-NCL.png, CC BY-SA 3.0. 2023.

[5] Robert A. Jr. Houze. *Cloud Dynamics*. 2nd. Vol. 104. International Geophysics. Academic Press, 2014.

[6] Joo Wan Cha et al. "Analysis of Rain Drop Size Distribution to Elucidate the Precipitation Process using a Cloud Microphysics Conceptual Model and In Situ Measurement". In: *Asia-Pacific Journal of Atmospheric Sciences* 59 (dic. 2023), pp. 257–269.

[7] Aytaç Kubilay et al. "CFD simulation and validation of wind-driven rain on a building facade with an Eulerian multiphase model". In: *Building and Environment* 61 (mar. 2013). ETH Zurich, Université de Sherbrooke, pp. 69–81. DOI: 10.1016/j.buildenv.2012.12.005. URL: https://doi.org/10.1016/j.buildenv.2012.12.005.

[8] World Meteorological Organization. *Guide to Meteorological Instruments and Methods of Observation*. WMO-No. 8. 2014.

[9] Agenzia Regionale per la Protezione Ambientale del Veneto. *Radar Meteorologico*. https://www.arpa.veneto.it/dati-ambientali/dati-in-diretta/radar/radar.

[10] Metlink. *COMET program*.

[11] Ralph E. Taggard. *Weather Satellite Handbook*. 5th. American Radio Relay League, 1994. ISBN: 978-0-87259-448-7.

[12] Kuo-Nan Liou. *Radiation and Cloud Processes in the Atmosphere: Theory, Observation, and Modeling*. Oxford University Press, 1992.

[13] Encyclopædia Britannica, Inc. *Encyclopædia Britannica*. https://www.britannica.com/. 2022.

[14] Johannes Schmetz et al. "An introduction to Meteosat Second Generation (MSG)". In: *Bulletin of the American Meteorological Society* 83.7 (lug. 2002), pp. 977–992.

[15] D. M. A. Aminou, B. Jacquet e F. Pasternak. "Characteristics of the Meteosat Second Generation Radiometer/Imager: SEVIRI". In: *Proceedings of SPIE, Europto series* 3221 (1997), pp. 19–31.

[16] D. M. A. Aminou et al. "Meteosat Second Generation: On-ground Calibration, Characterisation and Sensitivity Analysis of the SEVIRI Imaging Radiometer". In: *Proceedings of SPIE "Earth Observing Systems IV"* 3750 (1999), pp. 419–430.

[17] J. Schmid. *The SEVIRI Instrument*. ESA/ESTEC, Keplerlaan 1, 2200 AG Noordwijk, The Netherlands.

[18] *NASA Global Precipitation Measurement (GPM) Mission*. https://gpm.nasa.gov/missions/GPM.

[19] *NASA Global Precipitation Measurement (GPM) Dual-frequency Precipitation Radar (DPR)*. https://gpm.nasa.gov/missions/GPM/DPR.

[20] Toshio Iguchi et al. *GPM/DPR Level-2 Algorithm Theoretical Basis Document*. Rapp. tecn. NASA Global Precipitation Measurement.

[21] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, 2010.

[22] Google Developers. *Decision Forests*. URL: https://developers.google.com/machine-learning/decision-forests.

[23] Leo Breiman et al. *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[24] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2ª ed. Springer, 2009.

[25] Google Developers. *Random Forests*. URL: https://developers.google.com/machine-learning/decision-forests/random-forest.

[26] Harris Drucker. "Improving Regressors using Boosting Techniques". In: *Proceedings of the Fourteenth International Conference on Machine Learning* (1997).

[27] Google Developers. *Gradient Boosting*. URL: https://developers.google.com/machine-learning/decision-forests/intro-to-gbdt.

[28] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python*. 2011. URL: https://scikit-learn.org/stable/index.html.

[29] C. Kidd e V. Levizzani. "Status of satellite precipitation retrievals". In: *Hydrology and Earth System Sciences* 15.4 (2011), pp. 1109–1116. DOI: 10.5194/hess-15-1109-2011. URL: https://hess.copernicus.org/articles/15/1109/2011/.

[30] Catherine Prigent. "Precipitation retrieval from space: An overview". In: *Comptes Rendus Geoscience* 342.4 (2010). Atmosphère vue de l'espace, pp. 380–389. ISSN: 1631-0713. DOI: https://doi.org/10.1016/j.crte.2010.01.004. URL: https://www.sciencedirect.com/science/article/pii/S1631071310000155.

[31] Mario Montopoli et al. "Investigation of Weather Radar Quantitative Precipitation Estimation Methodologies in Complex Orography". In: *Atmosphere* 8.2 (2017), p. 34. DOI: 10.3390/atmos8020034.

[32] Kyuhee Shin et al. "Quantitative Precipitation Estimates Using Machine Learning Approaches with Operational Dual-Polarization Radar Data". In: *Remote Sensing* 13.4 (2021). Author to whom correspondence should be addressed: GyuWon Lee, p. 694. DOI: 10.3390/rs13040694.

[33] Hanna Meyer et al. "Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals". In: *Atmospheric Research* 169.Part B (mar. 2016), pp. 424–433. DOI: 10.1016/j.atmosres.2015.04.015.

[34] A. Liaw e M. Wiener. "Classification and regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22.

[35] Tianqi Chen e Tong He. *xgboost: eXtreme Gradient Boosting*. https://github.com/dmlc/xgboost. Package Version: 1.7.7.1. Gen. 2024.

[36] Toshio Iguchi et al. *GPM/DPR Level-2 Algorithm Theoretical Basis Document*. Rapp. tecn. Revised October 2014, January 2015, April 2015, August 2015, February 2016, March 2016, April 2017 for V05, October 2018 for V06. Global Precipitation Measurement (GPM) Mission, dic. 2010.

[37] PROJ contributors. *PROJ coordinate transformation software library*. Open Source Geospatial Foundation. 2024. DOI: 10.5281/zenodo.5884394. URL: https://proj.org/.

[38] Satpy Developers. *Satpy Documentation*. https://satpy.readthedocs.io/en/stable/index.html. 2024.

[39] J. Schmid. *The SEVIRI Instrument*. Keplerlaan 1, 2200 AG Noordwijk, The Netherlands.

[40] M. Kühnlein et al. "Precipitation estimates from MSG SEVIRI daytime, nighttime, and twilight data with random forests". In: *Journal of Applied Meteorology and Climatology* 53.11 (2014), pp. 2457–2480. DOI: 10.1175/JAMC-D-14-0007.1.

[41] "A Brief Survey on Random Forest Ensembles in Classification Model". In: *Proceedings of the Conference on Random Forests and Classification Models*. First Online: 20 November 2018. Nov. 2018.

[42] Mohammed Bader-El-Den, Eleman Teitei e Todd Perry. "Biased Random Forest For Dealing With the Class Imbalance Problem". In: *IEEE Transactions on Neural Networks and Learning Systems* 30.7 (2019), pp. 2163–2172. DOI: 10.1109/TNNLS.2018.2878400.

[43] "Classifying convective and stratiform rain using multispectral infrared Meteosat Second Generation satellite data". In: *Original Paper* (dic. 2011). Published: 07 December 2011.

[44] Pertti Nurmi. *Recommendations on the Verification of Local Weather Forecasts*. Rapp. tecn. Finland: Finnish Meteorological Institute, Operations Department, dic. 2003.

[45] Scikit-learn Contributors. *sklearn.metrics.r2_score*. Scikit-learn Documentation. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html#r2-score. 2024.

[46] G. James et al. *An Introduction to Statistical Learning: With Applications in R*. First. New York: Springer, 2013.

[47] Daniele Corradini. "Enhancing Quantitative Precipitation Estimation over Vietnam through Stacked Random Forest Models using Satellite Multispectral Data". Relatore: Prof. Federico Porcù, Correlatore: Dott. Giacomo Roversi. Master's Thesis. Alma Mater Studiorum - Università degli studi di Bologna, 2023.

[48] SciPy Contributors. *scipy.stats.gaussian_kde*. SciPy Documentation. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html.

[49] Huajin Lei, Hongyu Zhao e Tianqi Ao. "A two-step merging strategy for incorporating multi-source precipitation products and gauge observations using machine learning classification and regression over China". In: *Hydrology and Earth System Sciences* 26 (2022), pp. 2969–2987. DOI: 10.5194/hess-26-2969-2022.

[50] EUMETSAT. *HRSEVIRI - High Resolution SEVIRI*. https://data.eumetsat.int/data/map/EO:EUM:DAT:MSG:HRSEVIRI. 2024.

[51] NOAA PSL. *NCEP/NCAR Reanalysis*. https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html. 2024.

[52] NASA Goddard Earth Sciences Data and Information Services Center. *Giovanni - NASA's Online Visualization and Analysis Tool*. https://giovanni.gsfc.nasa.gov/giovanni/. 2024.

[53] M. Petracca et al. "Validation of GPM Dual-Frequency Precipitation Radar (DPR) Rainfall Products over Italy". In: *Journal of Hydrometeorology* 19.5 (mag. 2018), pp. 907–925. DOI: 10.1175/JHM-D-17-0144.1.

[54] P. Nurmi. *Recommendations on the verification of local weather forecasts*. ECMWF Tech. Memo. 430. Reading, UK: European Centre for Medium-Range Weather Forecasts, 2003, p. 19.

[55] Cunguang Wang et al. "Infrared Precipitation Estimation Using Convolutional Neural Network". In: *IEEE Transactions on Geoscience and Remote Sensing* 58.12 (2020), pp. 8612–8625. DOI: 10.1109/TGRS.2020.2989183.

[56] EUMETSAT. *Meteosat Third Generation (MTG)*. 2024. URL: https://www.eumetsat.int/meteosat-third-generation.