Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Informatica per il Management

Analisi di similarità di serie temporali per la validazione di osservazioni meteorologiche: confronto tra i tool MeteoNetWork vs NOAA

Relatore:
Chiar.mo Prof.
Marco Di Felice

Presentata da:
Nicolò Sguerso

Co-relatrici:
Chiar.ma Prof.ssa
Bianca Federici
Dott. Ing.
Ilaria Ferrando

Sessione
Anno Accademico 2020/2021

Indice

Ringraziamenti

Ringrazio il Prof. Di Felice per avermi permesso di sviluppare questo lavoro di tesi che mi ha appassionato, e che mi ha permesso di applicare le conoscenze informatiche, sviluppate nel percorso di studi, alle esigenze del trattamento dei dati in ambito meteorologico.

Ringrazio il Laboratorio di Geomatica dell'Università di Genova, nello specifico la Prof.ssa Bianca Federici e la Dott.ssa Ilaria Ferrando, per la l'ospitalità ricevuta nonostante le modalità telematiche dettate dal periodo pandemico.

Voglio ringraziare la mia famiglia per il sostegno ricevuto, in particolare i miei genitori che mi hanno supportato e sopportato nei momenti di incertezza, e per avermi permesso di intraprendere il percorso universitario che ha contribuito a formare la persona che oggi sono diventato.

Inoltre, voglio ringraziare tutte quelle persone che in questi tre anni sono state al mio fianco e hanno reso Bologna, la città per cui senza di loro, Bologna non è più la stessa.

Infine voglio ringraziare Sofia, per avermi spronato ad andare avanti anche nei momenti di maggiore sconforto e per essere stata al mio fianco per tutto questo tempo.

Introduzione

Il presente lavoro di tesi affronta l'analisi di osservazioni di carattere ambientale, di particolare utilità per fornire informazioni preziose per la meteorologia. Varie sono le organizzazioni che realizzano reti di stazioni per osservare diversi parametri, tra i quali la temperatura, il punto di rugiada, la pressione, le precipitazioni, velocità e direzione del vento ed altre ancora. Tali reti si differenziano nella distribuzione spaziale, nell'intervallo di campionamento, nei parametri osservati ed anche nella loro gestione. Lo studio effettuato si origina nell'esperienza di tirocinio svolto presso il Laboratorio di Geomatica dell'Università di Genova. L'obiettivo del lavoro è verificare la similarità tra osservazioni di differenti reti, in particolare quella gestita da parte di volontari denominata MeteoNetWork, diffusa per lo più in Italia e Francia, e quella del National Oceanic and Atmospheric Administration (NOAA), diffusa internazionalmente e da considerarsi quale riferimento. Per tale confronto si è optato di analizzare diverse serie temporali, con frequenza di osservazione oraria, a partire dal primo gennaio 2010 all'estate 2021. Non avendo ovviamente la co-locazione tra queste stazioni, se ne sono individuate sette coppie tra le due organizzazioni, distribuite nel centro-nord Italia, con distanze massime di 10 km. Per la validazione delle osservazioni di MeteoNetWork nei confronti di quelle del NOAA è necessario un'approccio critico tramite analisi di similarità delle serie temporali.

Dopo un inquadramento delle tematiche di riferimento del progetto, descritte nel capitolo 1, nel successivo vengono presentati gli obiettivi del lavoro di tesi, seguiti dalle scelte progettuali e dalla descrizione dei dati a disposizione. Nel terzo capitolo si affronta l'implementazione di codici autonomamente sviluppati per affrontare l'intero processo di trattamento dei dati; dalla loro raccolta, al Data-Cleaning e alla loro analisi critica. Particolare attenzione viene dedicata alla valutazione della coerenza di comportamento tra le serie temporali di stazioni limitrofe, tenendo successivamente conto delle differenze di quote presenti tra le stesse per l'analisi di similarità. Nel capitolo quarto, infine, sono tratte le conclusioni che riepilogano i principali passaggi affrontati nel lavoro di tesi, per effettuare il confronto atto alla validazione delle stazioni di MeteoNetWork nei confronti del NOAA.

1 Stato dell'Arte

1.1 Serie Temporali

1.1.1 Cenni alle serie temporali

Le serie temporali non sono nient'altro che un insieme di osservazioni ottenute attraverso ripetitive misurazioni all'interno di una determinata finestra temporale.[1] Sicuramente l'informazione maggiormente rilevante, per un dato di una seri temporale, è per l'appunto il suo time-stamp, ovvero il momento in cui è stata effettuata la rilevazione. Inoltre i time series data possono essere classificati in due tipologie di data set, semplicemente in base al fatto che le misurazioni vengano effettuate ad intervalli regolari di tempo, o che vengano campionate ad intervalli non regolari, venendo cosi chiamati eventi.[1]

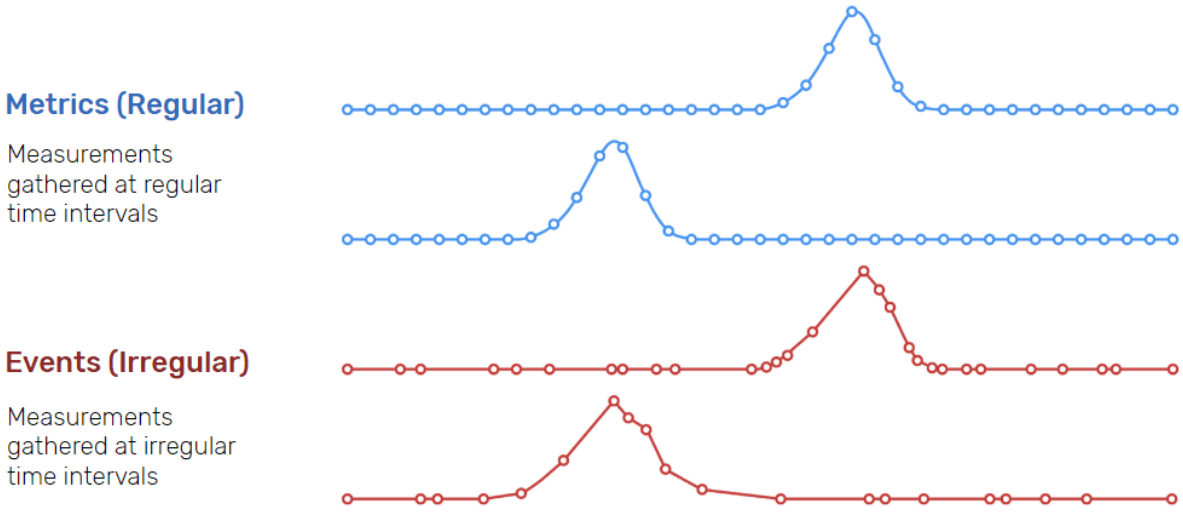


Figura 1: Differenza tra una serie temporale regolare (blu) ed irregolare (rossa)

Nelle serie temporali basati su eventi, gli stessi sono imprevedibili in quanto vengono rilevati ad intervalli irregolare, di conseguenza risultano complessi da modellizzare o ad effettuare delle previsioni, poiché la previsione presuppone che ciò che sia accaduto nel passato sia un buon indicatore di ciò che accadrà nel futuro. Mentre per serie temporali dove i dati vengono rilevati ad intervalli di tempo regolari, nel momento in cui si analizza la serie nella sua interezza, è possibile ottenere una corretta rappresentazione del comportamento e delle variazioni che caratterizzano un determinato sistema. Attraverso lo studio e l'analisi di queste

serie di dati è possibile estrarre statistiche ed informazioni proprie della serie in questione. La disciplina prende il nome di *Time series data mining*,ed include le seguenti tecniche:

- Query by content
- Clustering
- Classificazione
- Segmentazione
- Predizione
- Rilevamento di anomalie
- Motif Discovery

Query by content L'interrogazione per contenuto è l'area di ricerca più attiva nell'analisi delle serie temporali. Si basa sul recupero di un insieme di soluzioni che sono più simili ad una query fornita dall'utente.

Clustering Il processo di Clustering si pone l'obiettivo di trovare dei gruppi naturali, chiamati *cluster*, all'interno di un set di dati. L'obbiettivo è di trovare cluster che siano il più omogenei possibili e ben distinti da altri cluster. Più formalmente, il raggruppamento dovrebbe massimizzare la varianza inter-cluster mentre minimizza la varianza intra-cluster. L'algoritmo dovrebbe quindi individuare automaticamente quali gruppi sono intrinsecamente presenti nei dati. La difficoltà principale dei problemi di clustering (anche al di fuori dell'ambito del time series data mining) risulta la definizione del numero corretto di cluster.[2]

Classificazione Un *task* di classificazione cerca di assegnare etichette ad ogni serie di un insieme. La differenza principale rispetto al compito di clustering è che le classi sono note in anticipo e l'algoritmo è addestrato su un set di dati di esempio. L'obiettivo è innanzitutto imparare quali sono le caratteristiche distintive che distinguono le classi l'una dall'altra. Successivamente, quando un set di dati non etichettati viene inserito nel sistema, esso può automaticamente determinare a quale classe appartiene ogni serie.

Segmentazione Il compito di segmentazione (o sintesi) mira a creare un'approssimazione accurata delle serie temporali, riducendo la sua dimensionalità pur mantenendo le sue caratteristiche essenziali. La figura 2 mostra l'output di un sistema di segmentazione. L'obiettivo è quindi minimizzare l'errore di ricostruzione tra una rappresentazione ridotta e la serie temporale originale.

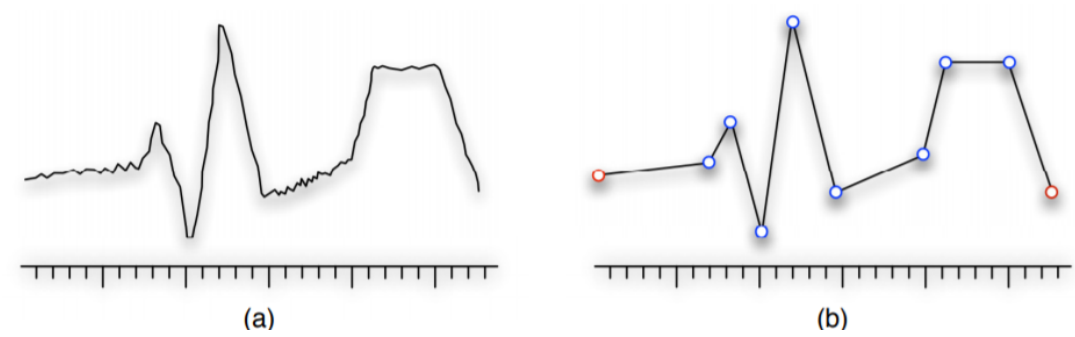


Figura 2: Esempio di applicazione di un sistema di segmentazione

Predizioni Le serie temporali sono di solito molto lunghe e vengono considerate *morbide*(*smooth*), nel momento in cui i valori variano all'interno di intervalli prevedibili [3] Il compito della predizione è volto a modellare esplicitamente tali dipendenze variabili per prevedere i futuri valori di una serie.

Rilevamento di anomalie Il rilevamento di anomalie cerca di trovare sotto-sequenze anomale in una serie. Questa branca della *Time Series Data Mining* è caratterizzata da ampi ambiti applicativi e di ricerca dovuta al rapido evolversi delle tecnologie wireless e alla sempre più semplice connessione alla rete internet. Grazie alla veloce crescita dell'Internet of Things la rilevazione di anomalie è uno strumento che viene sempre più adottato dalle industrie 4.0, per poter monitorare ed identificare possibili comportamenti che potrebbero portare ad anomalie nella filiera produttiva.

Motif dicovery La scoperta di *motif* consiste nel trovare ogni sotto-sequenza (chiamata motivo) che appare in modo ricorrente in una serie temporale più lunga. Questa idea è stata trasferita dall'analisi dei geni in bioinformatica. I motivi sono stati definiti originariamente come tipiche sotto-sequenze non sovrapposte. [4]

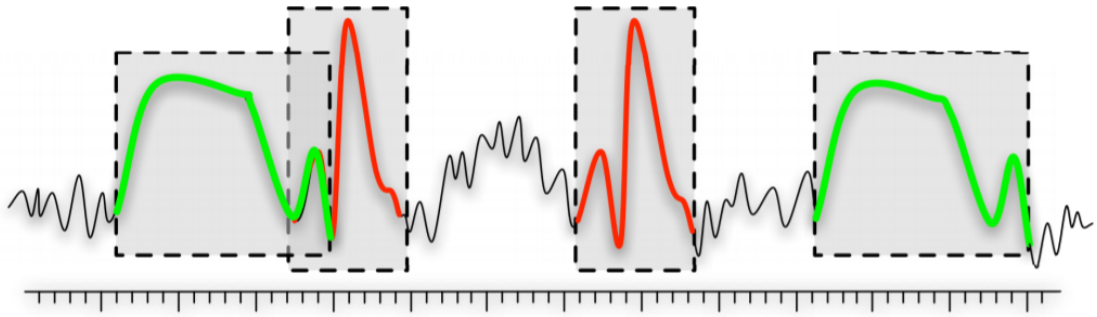


Figura 3: Esempio di applicazione motif discovery

1.1.2 Esempi di ambiti applicativi

L'analisi delle serie temporali non è uno ambito scientifico nuovo, nonostante la tecnologia ne faciliti l'accesso. Molti dei testi raccomandati che insegnano le teorie e le pratiche fondamentali dell'argomento esistono da diversi decenni. Inoltre il metodo stesso è ancor più datato, tanto da aver permesso lo studio del movimento planetario e la navigazione di civiltà antiche. L'analisi delle serie temporali viene utilizzata per i dati non stazionari - ossia che fluttuano costantemente nel tempo o che sono influenzate dal tempo. Industrie come la finanza, la vendita al dettaglio e l'economia usano spesso l'analisi delle serie temporali in quanto la valuta e le vendite cambiano continuamente.[6] L'analisi del mercato azionario è un eccellente esempio di analisi di serie temporali in azione, specialmente con algoritmi di trading automatizzati. Allo stesso modo, l'analisi delle serie temporali è ideale per prevedere i cambiamenti del tempo, aiutando i meteorologi ad effettuare previsioni. Esempi di analisi delle serie temporali in azione includono:

- Dati meteorologici
- Misure di pioggia
- Letture di temperatura
- Monitoraggio della frequenza cardiaca (EKG)
- Monitoraggio del cervello (EEG)
- Vendite trimestrali
- Prezzi delle azioni

- Trading azionario automatizzato
- Previsioni dell'industria
- Tassi di interesse

1.2 Big-Data

1.2.1 Cenni ai big-data

La definizione di Big Data si riferisce a dati che contengono una maggiore varietà, che arrivano in volumi crescenti e con maggiore velocità [7]. Questo concetto è anche noto come le tre V.

- Volume
- Velocità
- Varietà

Volume Indica elevati volumi di dati non strutturati e a bassa densità. Può trattarsi di dati di valore sconosciuto, come feed di dati di Twitter, clickstream su una pagina Web o un'app mobile o apparecchiature abilitate per sensori. Per alcune organizzazioni, potrebbero essere decine di terabyte di dati. Per altre, potrebbero essere centinaia di petabyte.

Velocità Indica la velocità con cui i dati vengono ricevuti e su cui si agisce. Normalmente, i dati vengono gestiti in memoria invece di essere scritti sul disco. Ne sono esempio tecniche di sicurezza informatica dove analizzano in real-time le proprie reti e sistemi informatici affinché possano prevenire ed identificare possibili cyber-attack.

Varietà Si riferisce ai molti tipi di dati disponibili. I tipi di dati tradizionali erano strutturati e si adattavano perfettamente a un database relazionale. Con l'avvento dei Big Data, i dati arrivano come nuovi tipi di dati non strutturati. I tipi di dati non strutturati e semistrutturati, come testo, audio e video, richiedono un'ulteriore elaborazione preliminare per ricavare significato e supportare i metadati.

Nell'analisi di questi dati è sempre bene assicurarsi della loro veridicità in modo da evitare il così detto fenomeno *Garbage in, garbage out*. Fenomeno per cui elaborando dati insensati, viziati o non veritieri, queste stesse caratteristiche caratterizzeranno il risultato delle analisi.[9]

1.3 Internet of Things

1.3.1 Cenni al Iot

L'Internet of things (IoT), internet delle cose, descrive la rete di oggetti fisici - "cose" - che sono incorporati con sensori, software e altre tecnologie allo scopo di connettersi e scambiare dati con altri dispositivi e sistemi su internet. Questi dispositivi vanno da oggetti domestici ordinari a sofisticati strumenti per le rilevazioni meteorologiche. Di fatti una delle motivazioni che spiega l'evoluzione dell'IoT è data dalla'aumento delle nuove strumentazioni dedicate al monitoraggio ed al rilevamento di sempre nuove tipologie di dato.

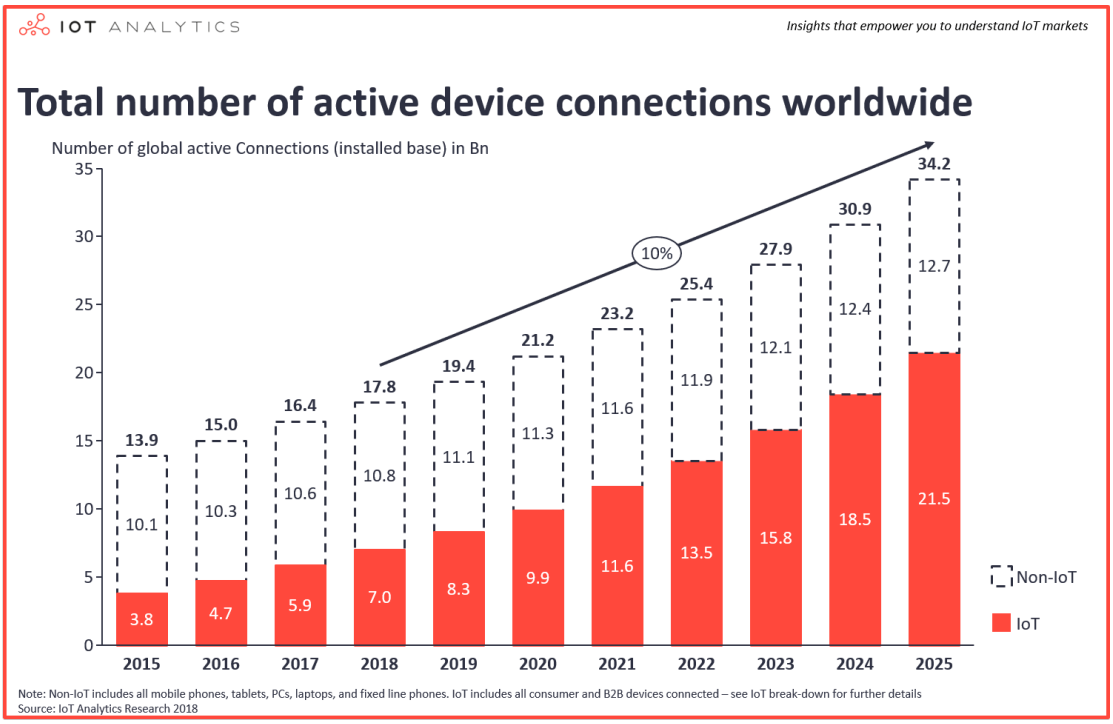


Figura 4: Telefoni, smartphone, tablet e computer sono contati nel Non-Iot. Nle Iot sono contati solamenti la sstrumentazione di clienti e B2B [10]

1.3.2 Esempi di ambiti applicativi

IoT è un componente tecnologico fondamentale nei progetti d'Industria 4.0. Esso rende intelligenti macchine e linee di produzione attraverso l'integrazione di sensori IoT, attuatori e componenti di Edge computing.[11] Tutto questo consente elaborazione in tempo reale e quindi avvio di processi automatici e allarmi.

I dati operativi, selezionati e sintetizzati dai sistemi Edge sono resi disponibili per ulteriori elaborazioni. Ad esempio per alimentare il pannello di controllo di una fabbrica, alimentare sistemi esterni di machine learning ed AI, utili per fare manutenzione predittiva e ottimizzare la produzione.

La maggiore disponibilità d'informazioni e la capacità di controllo consentono previsioni più attendibili, benefici a livello della flessibilità e del bilancio energetico, riduzioni degli scarti. Le capacità IoT integrate in elettrodomestici, uffici e infrastrutture critiche consentono di ridurre i consumi energetici, introdurre la manutenzione predittiva in nuovi ambiti.

1.4 Sistemi di monitoraggio meteorologico

Il campo della meteorologia, come tutti gli altri campi, con l'avvento di Internet, il minor costo delle telecomunicazioni e la velocità dello sviluppo dell'elettronica, è stato anch'esso travolto dal mondo dell'Internet of Things. Conseguentamente ha reso più semplice l'acquisizione di dati meteorologici, permettendo l'espansione delle zone monitorate così da avere previsioni e visioni d'insieme sempre più precise, grazie a nuove tecniche di studio di questi dati.

1.4.1 Acquisizione dati per la meteorologia

Nella meteorologia moderna ora mai vi sono svariati strumenti per il rilevamento delle caratteristiche meteorologiche in un dato momento. Le strumentazioni maggiormente diffuse e utilizzate sono:

- Termometro : Misuratore della temperatura dell'aria
- Pluviometro: Misuratore delle precipitazioni
- Barometro: Misuratore della pressione atmosferica
- Igrometro: Misuratore dell'umidità relativa dell'aria

Inoltre negli ultime 50 anni sono stati effettuati numerosi passi avanti nelle rilevazioni di questi parametri. Principalmente 2 sono stati due i passi che hanno trasformato il mondo della meteorologia:

- Satelliti meteorologici
- Computer e telecomunicazioni

Satelliti meteorologici Prima del 1959 tutte le osservazioni venivano effettuate a *terra*, per l'appunto tramite la strumentazione che abbiamo sopra elencato. Dopo di che nel 17 febbraio 1959 venne lanciato nello spazio il primo satellite meteorologico dalla NASA per il progetto *Vanguard* [24]. Da quella data fino ad'oggi i satelliti si sono sviluppati in due categorie:

- Satelliti geostazionari
- Satelliti polari

La differenza tra queste due tipologie di satelliti risiede nel modo in cui si muovono nei confronti della terra. Il satellite geostazionario è un satellite che si trova in orbita geostazionaria, ovvero la sua posizione è costante nei confronti della terra, portandolo a seguire l'orbita terrestre, da est ad ovest, e sopra l'equatore. Il satellite polare, invece è uno speciale tipo di satellite che segue l'orbita polare, ovvero passando da nord a sud periodicamente sopra i poli. Grazie a questo suo movimento è possibile monitorare longitudinalmente tutte le aree del pianeta, compresi i poli.

I satelliti raccolgono informazioni sull'atmosfera e sulla superficie terrestre grazie a strumenti chiamati radiometri. Tali strumenti misurano una grandezza fisica chiamata radianza, ossia la densità di flusso di radiazione elettromagnetica per angolo solido: si tratta di una misura dell'intensità della radiazione elettromagnetica misurata in una banda di frequenze. [12]

Computer e telecomunicazioni Come sopra accennato, grazie all'avvento dei moderni computer e delle telecomunicazioni, anche quest'ambito è stato oggetto di nuovi sviluppi nel mondo dell'Iot. A tale proposito sono state rivoluzionate le modalità con le quali vengono rilevati i parametri meteorologici, infatti tutte le strumentazioni sopra citate sono oggi compattate in una unica stazione meteorologica digitale, che essendo collegata ad un computer remoto, fornisce la

possibilità di avere rilevazioni meteorologiche aggiornate in tempo reale, ponendo così le basi al *now-casting*, ovvero alle previsioni meteorologiche a brevissimo termine (entro poche ore) su un particolare territorio d'interesse. L'utilità del *now-casting* è data dalla capacità di analizzare in una piccola porzione di territorio, i vari parametri dell'atmosfera e lo stato attuale del tempo, cosa che un modello internazionale che lavora su grandi porzioni di territorio, non riesce a fare.[13]

2 Progetto di Tesi

2.1 Introduzione e obiettivo del progetto

Questo progetto nasce dal lavoro effettuato durante il tirocinio, svolto presso il Laboratorio di ricerca di geomatica presso l'Università di Genova e focalizzato sull'analisi di dati meteorologici di interesse del laboratorio stesso. Il lavoro svolto nel progetto di tesi si è posto l'obiettivo di poter dare validità alle osservazioni meteorologiche che vengono effettuate ogni giorno dalle stazioni digitali di MeteoNetWork, d'ora in poi ci si riferirà come *MNW*[8]. Il *MNW* è un'organizzazione senza scopo di lucro che gestisce la più vasta rete di osservazioni meteorologiche in Italia, grazie al supporto di una ancora più vasta rete di appassionati e volontari che mettono a disposizione le proprie stazioni. Ovviamente per poter concludere una qualche affermazione per analizzare al qualità dei dati osservati dalle stazioni meteorologiche, è stato deciso di effettuare delle analisi a campione e, cosa ancor più importante, di effettuare delle analisi di similarità con delle stazioni limitrofe appartenenti ad un altro ente. Il secondo ente di cui abbiamo utilizzato i dati è la *National Oceanic and Atmospheric Administration*, d'ora in poi ci si riferirà con la sigla di *NOAA*. Il NOAA è l'agenzia federale statunitense che si interessa di oceanografia, meteorologia e climatologia, riconosciuta a livello mondiale. I dati da loro rilevati vengono gestiti dal *NCEI*, *National Centers for Environmental Information*, ente che ospita i dati raccolti dal *NOAA*. Tale ente ospita e fornisce accesso pubblico a uno dei più significativi archivi di dati ambientali sulla Terra, con oltre 37 petabyte di dati atmosferici, costieri, oceanici e geofisici completi messi a disposizione.[14]

Prima di effettuare il calcolo di similarità tra le stazioni meteorologiche *NOAA* e *MNW*, che rilevano e memorizzano serie temporali, è necessario effettuare un *Data-Cleaning*, ossia un processo di correzione o rimozione di dati errati, corrotti, formattati in modo errato, duplicati o incompleti all'interno di un set di dati affinché si possa utilizzare solamente dati validi che permettano risultati affidabili.[15]

Infine, una volta ottenuti i dati puliti, saranno effettuate le analisi di similarità, di queste differenti serie temporali, per verificare la coerenza e veridicità delle osservazioni compiute dalle stazioni *MNW*.

2.2 Scelte progettuali

Questa sezione verterà sulla motivazione delle principali scelte che hanno guidato il progetto, al fine di rendere più trasparente possibile la logica che ha indirizzato questo lavoro. Le principali scelte compiute prima di iniziare con le vere e proprie analisi sono le seguenti:

- Scelta degli Enti
- Scelta dei Data-Set
- Scelta delle stazioni meteorologiche da analizzare
- Scelta dei parametri meteo da analizzare
- Scelta del Calcolo di similarità

2.2.1 Scelta degli enti

Come descritto nella parte introduttiva di questa sezione, il progetto si fonda sulla comparazione di 2 set di dati, uno del *MNW* e l'altro del *NOAA*. La scelta del secondo data set si è basata sulla ricerca di diversi enti.

Qui di seguito si analizzano quali sono stati gli enti presi in considerazione e come mai non sono stati selezionati.

- Copernicus, ex Global Monitoring for Environment and Security
- European Climate Assessment and Dataset, ECAD
- Arpa Regionali

Copernicus Il progetto Copernicus, è il programma di osservazione della Terra dell'Unione Europea, che osserva il pianeta e il suo ambiente a beneficio di tutti i cittadini europei. Offre servizi di informazione che attingono all'Osservazione della Terra via satellite e ai dati in situ (non spaziali).[16] Nonostante Copernicus si basi su numerosissimi satelliti che realizzano svariate osservazioni quotidiane su una rete globale di migliaia di sensori terrestri, aerei e marini che riproducono immagini della Terra il più dettagliate possibili e con una produzione attuale di 12 terabyte al giorno [17], queste informazioni non sono di nostro interesse. Questo perché le osservazioni con cui dovrebbero essere paragonate, ovvero quelle

effettuate da *MNW*, sono rilevazioni effettuate a *terra* tramite strumentazioe digitale. Di conseguenza effettuare paragoni con rilevazioni satellitari non sarebbe opportuno. Inoltre le osservazioni a *terra* che vengono effettuate dal progetto Copernicus non registrano gli stessi parametri analizzati da *MNW* oppure si trovano in località differenti.

ECAD ECAD,*European Climate Assessment Dataset*, rappresenta la spina dorsale del nodo dei dati climatici nel Centro Climatico Regionale (RCC) per la Regione VI del WMO (Europa e Medio Oriente) dal 2010. I dati e i prodotti informativi contribuiscono al Global Framework for Climate Services (GFCS). Ad oggi, *ECAD* riceve dati da 79 partecipanti per 65 paesi e il dataset ECAD contiene 74317 serie di osservazioni per 13 elementi in 20131 stazioni meteorologiche in tutta Europa e nel Mediterraneo. La partecipazione a *ECAD* è aperta a chiunque mantenga dati di stazioni giornaliere.[18] Proprio per questo motivo *ECAD* fornisce dati che di fatto non sono registrati da stazioni di loro proprietà o da loro gestite, ma fornisce dati che vengono registrati da privati o pubbliche amministrazioni che poi inviano a loro. Infatti al'interno della lista dei loro partecipanti si possono trovare enti come *Arpa Lombardia*, *Arpa Calabria* e altri enti di simile natura.[19] Un ulteriore motivo per cui *ECAD* non è stato selezionato deriva dal fatto che le sue stazioni non sono prossime a quelle di MeteoNetWork oggetto di studio.

Arpa Regionali Un'agenzia regionale per la protezione dell'ambiente (in acronimo ARPA) è un ente della pubblica amministrazione italiana, costituito e operante in ogni Regione d'Italia. Ciascuna Regione ha costituito la propria Agenzia. Le 19 ARPA regioni più le due ARPA delle province autonome di Trento e Bolzano e ISPRA compongono il Sistema nazionale per la protezione dell'ambiente (*SNPA*).[20] Nonostante le svariate stazioni meteorologiche di cui *SNPA* dispone, i dati provenienti da questi enti non sono stati presi in considerazione in quanto il reperimento dei dati risultava maggiormente ostico in confronto alla facilità di reperimento dei dati con cui si ha avuto esperienza con il *NOAA*.

2.2.2 Scelta dei DataSet

Successivamente è stato necessario decidere quale data-set utilizzare per effettuare il successivo paragone con i dati di *MNW*. Queste scelte sono state necessarie solamente per il *NOAA*, in quanto dispone di un'enormità di dati e differenti set

di dati, utili per ogni tipo di studio dall'oceanografia alla rilevazione meteo. Il *NOAA* per le rilevazioni meteorologiche offre i seguenti set di dati, tutti rilevati con metodologie e strumentazioni differenti.

Land-Based Station Le osservazioni terrestri, o di superficie, includono temperatura, punto di rugiada, umidità relativa, precipitazioni, velocità e direzione del vento, visibilità, pressione atmosferica e tipi di eventi meteorologici come grandine, nebbia e tuoni, raccolti per località di ogni continente.

Satellite I satelliti geostazionari e in orbita polare forniscono dati grezzi di radianza raccolti da stazioni a terra per aiutare a monitorare e prevedere eventi meteorologici e ambientali.

Radar Acronimo di *Radio Detection and Ranging*, è un sistema di rilevamento di oggetti che utilizza le onde radio per determinare la portata, l'altitudine, la direzione del movimento e la velocità degli oggetti, producendo dati grezzi e generando prodotti di analisi.

Weather Balloon Rileva dati meteorologici dall'atmosfera, a partire da tre metri sopra la superficie terrestre. Questi dati sono ottenuti da radiosonde, che sono pacchetti di strumenti legati a palloni aerostatici che trasmettono dati alla stazione ricevente.

Ovviamente il set di dati che meglio rispondeva alle nostre necessità è *Land-Based Station*, in modo tale che i dati provenienti da questo set di dati fossero totalmente paragonabile con quelli di *MNW*. Conseguentemente in questo data-set sono presenti numerosi altri sotto-data-set tutti divisi per funzionalità, per zona geografica coperta, per finestra temporale coperta o per la frequenza di osservazione, giornaliera o oraria. In seguito alle dovute ricerche è stato deciso di utilizzare i dati provenienti dal' *ISD*, *Integrated Surface Database*, un database globale che consiste in osservazioni di superficie orarie e sinottiche compilate da numerose fonti in un unico formato ASCII e modello di dati comune. *ISD* integra i dati provenienti da più di 100 fonti di dati originali, compresi numerosi formati di dati che sono stati inseriti da moduli cartacei durante il periodo 1950-1970.

2.2.3 Scelta delle stazioni

Il processo decisionale che ha portato alla scelta delle stazioni meteorologiche può essere diviso in tre parti. La prima in cui sono state effettuate le dovute ricerche relativamente a quali stazioni fossero presenti sul territorio italiano per entrambi gli enti. La seconda per identificare le stazioni *MNW* che avessero una continuità dell'ordine di una decade in modo da poter dare rilevanza e struttura agli studi successivi. In questo passaggio per le stazioni *NOAA* non c'è stato il bisogno di effettuare questa operazione in quanto tutte le loro serie temporali riescono a tornano indietro nel tempo di oltre 50 anni. Mentre nella terza fase sono state ricercate delle possibili coppie di stazioni meteo, *NOAA* e *MNW*, che non fossero planimetricamente distanti l'una dall'altra più di 10 km, in modo tale da non rischiare di viziare le future analisi di similitudine a causa di una eccessiva distanza geografica che impedisca la rilevazione dei medesimi fenomeni meteorologici.

A posteriori di queste fasi sono state identificate 13 stazioni, 7 MeteoNetWork e 6 NOAA, su cui è stato basato il progetto. Per maggiore informazione sulle stazioni si veda la sezione 2.3.

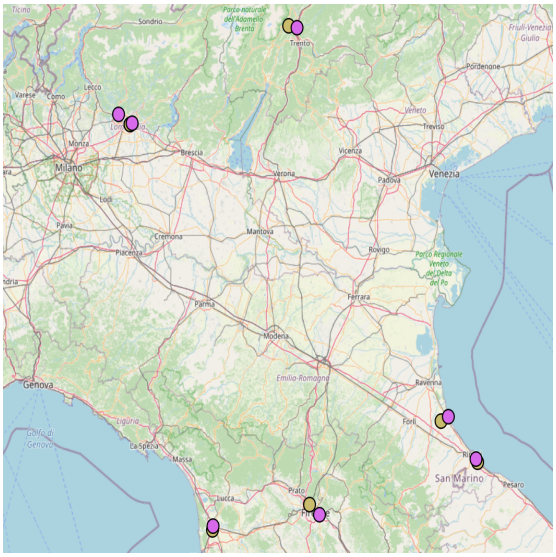


Figura 5: Colore fucsia: Stazioni MeteNetWork, Colore beige: Stazioni NOAA

Per questo passaggio è stato fondamentale l'utilizzo di QGIS, un'applicazione desktop GIS open-source che permette di visualizzare, organizzare, analizzare e rappresentare dati spaziali.[21] Infatti è bastato l'inserimento delle coordinate delle stazioni sia di *NOAA* sia di *MNW* per ottenere una mappa con tutte le sta-

zioni presenti in Italia. Infine è stata effettuata una *query* nella quale si chiedeva di mostrare solamente le stazioni *NOAA* e *MNW* che fossero ad una distanza inferiore di 10km l'uno dall'altra.

Noaa	Mnw	Località	Provincia
160220	trn013	Paganella	Provincia autonoma di Trento
160760	lmb079	Aereoporto Orio Al Serio	Provincia di Bergamo
160760	lmb102	Aereoporto Orio Al Serio	Provincia di Bergamo
161480	ero099	Cervia	Provincia di Ravenna
161490	ero121	Rimini	Provincia di Rimini
161580	tsc001	Pisa	Provincia di Pisa
161700	tsc007	Firenze	Provincia di Firenze

2.2.4 Scelta dei parametri

Normalmente, nel campo delle misurazioni meteorologiche, spesso capita che i parametri con maggiori rilevazioni, speso rilevati da più stazioni, siano i seguenti:

- Temperatura
- Pressione
- Precipitazione
- Punto di Rugiada
- Umidità

La motivazione principale è il fattore economico, infatti questi parametri sono spesso presenti nei modelli di stazioni meteo a minor costo. Per questo motivo, nelle osservazioni su cui si é lavorato, i parametri che presentavano il minor numero di dati mancanti erano proprio temperatura, pressione e punto di rugiada.

2.2.5 Scelta del Calcolo di similarità

Per perseguire l'obbiettivo del progetto sono state valutate diverse analisi per poter asserire che le serie temporali, del NOAA e di MeteoNetWork, rilevassero le medesime condizioni meteo. All'inizio si era optato per l'ausilio di cluster. Come primo passo si sarebbe verificata l'esistenza di cluster nei dati di una singola stazione, successivamente si sarebbe verificata la presenza di un altro cluster nella

stazione coppia, ossia la stazione corrispondente dell'altro ente. Una volta verificate queste condizioni sarebbe bastato calcolare il medesimo cluster su l'unione dei dati di entrambe le stazioni, NOAA e MeteoNetWork. Se in quest'ultimo passaggio le singole osservazioni di entrambe le stazioni venissero identificate tutte sotto lo stesso cluster, questo significherebbe che le osservazioni effettuate da entrambi gli enti sono coerenti, che possiedono le stesse caratteristiche discriminative e che rappresentano i medesimi fenomeni meteorologici, come se fossero misurati dalla stessa stazioni. In questo modo si sarebbe potuto affermare la similarità delle serie temporali raggiungendo così lo scopo del progetto.

Dopo qualche test si è osservato che non si riusciva a delineare correttamente dei cluster, tramite algoritmo *K-means*, in quanto l'insieme di dati non formavano un cluster di forma circolare, forma per il quale funziona particolarmente bene questo algoritmo.

Successivamente è stato scelto di portare avanti lo studio effettuando una semplice differenza tra le osservazioni del *NOAA* con quelle rilevate dal *MeteoNetWork*, solamente su tutte quelle osservazioni per cui vi è una osservazione gemella dell'altro ente. Per osservazioni gemelle intendiamo le osservazioni con medesimo *time-stamp*. In questo modo, partendo dal presupposto che le stazioni siano ad una distanza planimetrica massima di 10 km, dovrebbero misurare i medesimi fenomeni meteorologici. Tramite questo approccio possiamo facilmente asserire se le stazioni del *MeteoNetWork* effettuino delle rilevazioni valide, coerenti con quanto osservato dal *NOAA*. Più i valori ottenuti a posteriori della sottrazione sono prossimi allo zero, maggiore è la vicinanza numerica con le osservazioni che vengono effettuate dal *NOAA*, pertanto maggiore è la similitudine con delle serie temporali. Infine una volta ottenute le differenze per ogni parametro analizzato, è stata calcolata la media e deviazione standard di questi valori così da poter analizzare in maniera critica i risultati globali ottenuti.

2.3 Dati di partenza

In questa sezione vengono presentate le caratteristiche dei dati allo stato originale, ovvero nel momento in cui questi sono stati raccolti. Le caratteristiche elencate sono relative alle stazioni che sono state identificate nella sotto sezione 2.2.3. Verranno descritti i data-set sia di *NOAA* che di *MNW*

2.3.1 Dati del NOAA

I dati del *NOAA*, come anticipato nella sotto sezione 2.2.2, fanno parte del data set ISD lite, Integrated Surface Database, in cui vengono memorizzati 12 differenti campi, ognuno per una rilevazione differente. Per l'*ISD-Lite*, nel momento in cui i dati vengono rilevati, viene definita una "finestra di cattura" per estrarre gli elementi osservativi più vicini all'inizio dell'ora. Questa finestra di cattura permette la memorizzazione di tutte le osservazioni con minuto di osservazione maggiore o uguale a dieci minuti prima dell'ora fino allo scadere dell'ora incluso. Gli elementi *ISD* con un minuto di osservazione al di fuori della finestra di cattura vengono ignorati.[22]

Tipo di osservazione	Unità	Note
Anno		
Mese		
Giorno		
Ora		arrotondata all'ora più vicina
Temperatura dell'aria	C°	fattore di scala 10
Punto di rugiada	C°	fattore di scala 10
Pressione sul livello del mare	hPa	fattore di scala 10
Direzione del vento	Gradi angolari	fattore di scala 1
Velocità del vento	m/s	fattore di scala 10
Condizioni del cielo		
Precipitazioni per 1 ora	mm	fattore di scala 10
Precipitazioni per 6 ora	mm	fattore di scala 10

Nel momento in cui una delle osservazioni dovesse mancare, tale osservazione viene marcata con il valore *-9999*. Inoltre per fattore di scala si intende che i valori sono stati moltiplicati per 10. Questa è una tipica tecnica affinché le virgole dei valori non creino problemi con il formato con cui vengono memorizzati, in questo caso sono memorizzati in formato *'tsv'* e salvati con l'estensione *.txt*

I dati ottenuti sono relativi a 6 stazioni, come si possono vedere nella figura [5], sparsi tra la Toscana, Emilia-Romagna, Lombardia e Trentino-Alto Adige. Qui di seguito sono riportate le meta-informazioni delle singole stazioni.

Codice	Località	Provincia	Longitudine	Latitudine	Altitudine(m)
160220	Paganella	Provincia autonoma di Trento	11.033	45.150	2129
160760	Aereoporto Orio Al Serio	Provincia di Bergamo	9.704	45.674	238
161480	Cervia	Provincia di Ravenna	12.307	44.224	5
161490	Rimini	Provincia di Rimini	12.612	44.020,	12.5
161580	Pisa	Provincia di Pisa	10.393	43.684,	1.8
161700	Firenze	Provincia di Firenze	11.205	43.810	44

Infine, i dati reperiti fanno riferimento al periodo che intercorre dal *1 gennaio 2010* al *18 agosto 2021* per un totale di osservazioni pari a 488'316.

2.3.2 Dati del MNW

I dati del *MNW*, come precedentemente anticipato non fanno parte di alcun tipo di data set in quanto gli unici tipi di dati reperibili sono le osservazioni delle stazioni meteorologiche che hanno a disposizione. Di seguito riportiamo i parametri che sono disponibili per ogni stazione.

Tipo di osservazione	Unità	Note
Data		time-stamp dell'osservazione
Temperatura dell'aria	C°	
Punto di rugiada	C°	
Umidità	%	fattore di scala 10
Pressione sul livello del mare	hPa	
Direzione del vento	Gradi angolari	
Velocità del vento	m/s	
Raffica del vento	m/s	
Precipitazioni	mm	oraria
Precipitazione	mm	giornaliera
Radiazione solare	rad	fattore di scala 10
Radiazione ultravioletta	valore tra 0 e 11+	

I dati del *MNW*, nel momento in cui ci siano delle osservazioni mancanti, a differenza del *NOAA*, non vengono marcati con il numero *-9999*, ma il campo viene lasciato vuoto, facendo si che a livello di analisi sia riconosciuto come *null*

A differenza del *NOAA* le stazioni prese in considerazione non sono 6 ma 7, questo perché due stazioni meteorologiche del *MNW* sono limitrofe ad una del *NOAA*. In questo caso è stato deciso di mantenere entrambe le stazioni. Come si può vedere nella figura [5], queste stazioni sono *lmb102* e *lmb079* per *MNW* e *160760* per il *NOAA*. Di seguito sono riportate le stazioni scelte per*MNW*.

Codice	Località	Provincia	Longitudine	Latitudine	Altitudine(m)
trn013	Lavis	Provincia autonoma di Trento	11.101	46.139	219
lmb079	Paladina	Provincia di Bergamo	9.609	45.724	268
lmb102	Seriate	Provincia di Bergamo	9.720	45.680	246
ero099	Cervia	Provincia di Ravenna	13.369	44.246	0
ero121	Rimini	Provincia di Rimini	12.600	44.038	7
tsc001	Pisa	Provincia di Pisa	10.398	43.704	5
tsc007	Firenze	Provincia di Firenze	11.289	43.761	51

Infine i dati reperiti fanno riferimento al periodo che intercorre dal *1 gennaio 2010* al *15 maggio 2021* per un totale di osservazioni pari a 6'144'388. Questo perché le osservazioni vengono effettuate ogni 5 minuti e non ogni ora come per il *NOAA*.

3 Implementazione di codici per il trattamento e l'analisi dei dati

Questa sezione illustrerà le operazioni che sono state effettuate per la prosecuzione del progetto. Conseguentemente non verranno esposte solamente le operazioni di analisi vere e proprie, ma saranno illustrate anche tutte quelle *"operazioni nascoste"*, ossia passi necessari che non sono strettamente legati alle analisi, andando così ad indicare tutte le operazioni di *Data-Cleaning* e di raccolta dati. Ne consegue che in questa sezione verrà spiegato nello specifico i passi necessari per l' ottenimento dei risultati ottenuti, con anche l'inserimento di *"snippet"*, porzioni di codice. Per maggiore chiarezza è bene indicare che il codice sviluppato per le operazioni che verranno descritte di seguito, è stato sviluppato in linguaggio Python 3.9.6. Nello specifico si analizzeranno i seguenti argomenti:

- Raccolta dei dati
- Data-Cleaning
- Calcolo di similarità

3.1 Raccolta dei dati

La fase di raccolta dei dati si è sviluppata in maniera differente tra *MNW* e *NOAA*, nonostante entrambi abbiano delle *API* che permettano lo scaricamento dei dati delle stazioni a cui si è interessati. Le differenze tra i due enti per questa operazione verranno discusse di seguito.

3.1.1 NOAA

I dati del *NOAA* come sopra anticipato sono stati ottenuti tramite l'*API* che messa a disposizione, il quale si basa su protocollo *FTP*. Tale *API* richiede in input i seguenti dati:

- Codice identificativo della stazione
- Anno di cui si vuole l'osservazione

Ovviamente per effettuare la richiesta serve inoltre l'*end-point* di riferimento del data-set di cui siamo interessati, che in questo caso è il seguente:

ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite/

Per specificare qual è la stazione di cui siamo interessati e sopratutto di quale anno vogliamo analizzare le osservazioni dobbiamo effettuare un'operazione di concatenamento dei dati di input con l'indirizzo URI del'*end-point*, affinché l'indirizzo finale sia della seguente forma:

ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite/ YEAR / CODE -9999- YEAR.gz

Dove :

- YEAR : Indica l'anno a cui ci si riferisce
- CODE : Indica il codice identificativo della stazione
- -9999-: Porzione di indirizzo fissa che è presente per tutte le stazioni di tutti gli anni disponibili

Come si nota, l'estensione del file indica che otteniamo un file compresso, per cui sarà necessario decomprimerlo per poter leggere i dati richiesti. Inoltre come si può intuire dalla struttura del'*URI* viene scaricato un anno alla volta di osservazioni per singola stazione. Per ottenere i dati di tutte le stazioni è stato necessario effettuare un doppio ciclo *for* aventi come indici gli anni da scaricare e i codici delle stazioni. Qui di seguito si riporta il codice sviluppato.

```
import urllib.request
import gzip
import os

#variabili
anno = [2010 , 2011, 2012, 2013, 2014, 2015, 2016, 2017,
    2018, 2019, 2020, 2021]
codice_stazione = [161480, 161490, 160760, 160220, 161580,
    161700]

#links
percorso_FTP='ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite/'

for code in codice_stazione:
```

```
file_name = str(code)+'.csv'
mode = 'xb'
f = open(file_name,mode)
f.close()
for year in anno:
link ="ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite/"+
str(year)+"/"+str(code)+"-99999-"+str(year)+".gz"

ftp_r = urllib.request.urlopen(link)
data = gzip.decompress(ftp_r.read())
f2 = open(file_name, 'ab')
f2.write(data)
f2.close()
print( str(code)+" "+str(year)+" file appended")
print("dati scaricati correttamente")
```

Come è possibile notare, una volta effettuata la chiamata *FTP*, come primo passo viene decompresso il file ottenuto e successivamente viene eseguita l'operazione di *append* ad un file che conterrà tutte le osservazioni di una specifica stazione. In questo modo si riesce ad avere una migliore gestione dei file; in quanto se ne otterranno solamente 6, uno per stazione, e non 66 file, uno per ogni anno e per ogni stazione. Infine i dati all'interno dei singoli file che si scaricano seguono la formattazione *.tsv*. Questo vuole dire che i valori sono separati da tabulazioni. Nella fase di *Data-Cleaning* i file verranno modificati in formato *.csv* affinché si possa garantire una migliore compatibilità con gli strumenti di analisi quali Pyhton e MySQL.

3.1.2 MNW

Per *MNW*, nonostante anch'essa metta a disposizione una *API*, non è stato possibile effettuare in maniera autonoma la richiesta dei dati necessari, in quanto l'*API* fornita pone dei limiti al numero delle richieste che possono essere effettuate. Inoltre pone un'altro vincolo sul numero di osservazioni che possono essere richieste.[23] Infatti è possibile effettuare un massimo di 20 richieste ogni 10 minuti, ed ogni richiesta può chiedere solamente le osservazioni giornaliere di una singola stazione. Di conseguenza dati delle stazioni su cui verranno effettuate le analisi sono state ottenute tramite una richiesta, scritta per email, al gestore del Data-Base di *MNW*, il quale ha fornito tutte le osservazioni dal 2010 al 2021 delle

stazioni richieste.

3.2 Data-Cleaning

Questa sezione verte sull'esposizione dei passaggi che sono stati svolti affinché i dati ottenuti, tramite i passaggi spiegati precedentemente, possano considerarsi sufficientemente *"puliti"* per effettuare l'analisi di similarità.

3.2.1 NOAA

Per il set di dati ottenuto dal *NOAA*, le operazioni svolte sono state le seguenti:

- Conversione dei file *.txt* in file *.csv*
- Sostituzione dei valori -9999
- Rimozione delle colonne non necessarie
- Eliminazione del fattore di scala

Per convertire i file dal formato *.tsv* al *.csv* è stato sufficiente effettuare una lettura del file di partenza utilizzando il metodo di lettura *.read_csv* della libreria *Pandas*, impostando come separatore gli spazi bianchi. Inoltre nella porzione di codice che segue viene effettuato anche l'inserimento del'*header* in quanto i dati scaricati tramite *API* non la forniscono.

```
codice_stazione = [161480, 161490, 160760, 160220,
    161580, 161700]

#creo una cartella in cui ci metto tutti i nuovi file csv
os.mkdir("file-csv-1")
col_name = ['year', 'month', 'day', 'hour',
    'temperature','dewPointTemperature',
    'slp_msl','wind_direction', 'wind_speed_rate',
    'sky_condition', 'prec_1h', 'prec_6h' ]

for code in codice_stazione:
file_name = str(code)+'.csv'
file_r = pd.read_csv(file_name, delim_whitespace=True,
    header= None)
file_r.to_csv("file-csv-1/"+file_name, index=None)
```

Una volta formattati correttamente è stato necessario eliminare tutte le colonne di cui non abbiamo dati. Qui di seguito viene mostrata la tabella che illustra il numero di dati mancanti di ogni stazione.

Codice	Temperatura	P. Rugiada	Pressione	Precipitazioni
160220	0	0	101253	x
160760	0	0	69015	x
161480	0	0	67311	x
161490	0	0	17795	x
161580	0	0	8	x
161700	0	0	68157	x

Tutte le colonne che non sono state elencate non presentano alcuna rilevazione. Conseguentemente sono tutte colonne che verranno eliminate.

Successivamente, affinché i dati ottenuti possano essere salvati su un Data-Base, e per ogni osservazione si possa identificare quale stazione abbia effettuato tale rilevazione, è necessario aggiungere una colonna che indichi il codice identificativo della stazione operante.

```
path = os.getcwd()
pathOldDir = path+"\\noaa-old-dir\\"
pathNewDir = path+"\\noaa-new-dir"
list_dir = os.listdir(pathOldDir)

os.mkdir(pathNewDir)

for file_name in list_dir:
df = pd.read_csv(pathOldDir+file_name)
df.insert(0,"id_stazione", file_name[0:6])
print(df)
df.to_csv(pathNewDir+"\\"+file_name, index=None)

print("programma finito con successo")
```

Questi file possono essere trasferiti su un Data-Base MySQL creato *ad-hoc* per poter effettuare ulteriori operazioni di manipolazione delle colonne, in modo da poter mantenere solamente i dati che siano strettamente necessari ai fini del progetto. Tutte le colonne che non indicano la temperatura, punto di rugiada

e pressione atmosferica sono state tutte eliminate in quanto non contenevano osservazioni valide.

Una volta mantenuti tutti i dati necessari si è passati alla sostituzione di tutte le osservazioni che sono state marcate con il valore *-9999* con il valore *NULL* in modo tale che nelle future analisi questi valori non vengano conteggiati e non sia necessario dover eliminare l'intera tupla rischiando di perdere dati utili.

```
# sostitzione dei valori -9999 con NULL con MySQL

update Dati_NOAA
set
temperatura = null
where temperatura = -9999;

update Dati_NOAA
set
pressione = null
where pressione = -9999;

update Dati_NOAA
set
punto_di_rugiada = null
where punto_di_rugiada = -9999;
```

Per completare la parte di Data-Cleaning, è stato necessario eliminare il fattore di scala dalle osservazioni rimanenti. Come descritto nella sezione 2.3.1 il fattore di scala è pari a 10, per tanto sarà necessario dividere per 10 ogni osservazione. Per svolgere questa operazione è stato deciso di utilizzare Python e la libreria Pandas in modo da sfruttare al massimo le capacità di elaborazione dai metodi offerti per le manipolazioni di *DataFrame*. Qui di seguito viene riportato il codice che svolge tale operazione.

```
pathDati = path+"\\Dati_NOAA"
listDir = os.listdir(pathDati)

for cartella in listDir:
print("Inizio ", cartella)

listFile = os.listdir(pathDati + "\\" + cartella)
```

```
# apro il file noaa
pathFile = pathDati + "\\" + cartella + "\\" + listFile[0]
df = pd.read_csv(pathFile)

# divido tutte le osservazioni di pressione per 10
df["pressione"] = df["pressione"]/10

# divido tutte le osservazioni di temperaura per 10
df["temperatura"] = df["temperatura"]/10

# divido tutte le osservazioni di punto_di_rugiada per 10
df["punto_di_rugiada"] = df["punto_di_rugiada"]/10

#salvo i cambiamenti
df.to_csv(pathFile, index= None)
```

3.2.2 MNW

I dati del *MNW* necessitano trattamenti simili a quelli eseguiti sui dati del *NOAA* . Le operazioni necessarie sono state le seguenti:

- Rimozione delle colonne non necessarie
- Inserimento delle colonne Anno, mese, giorno, ora
- Aggregazione dei valori

A differenza del *NOAA*, verranno eliminate sia le colonne che non sono d'interesse, come per esempio i dati delle Radiazioni Solari, sia le colonne che a causa dell'eccessiva mancanza di dati sono state eliminate nel data set del *NOAA*, rendendo in questo modo inutile il loro mantenimento in quanto lo scopo del presente lavoro è di confrontare i due data set. Qui di seguito viene mostrata la tabella che illustra i dati mancanti di ogni stazione.

Codice	Temperatura	P. Rugiada	Pressione	Precipitazioni
trn013	1	1	11	10693
lmb079	2	2	1	1
lmb102	2	36487	0	0
ero099	728	733	348	57545
ero121	305	299	6	0
tsc001	0	0	0	0
tsc007	18430	18430	1	1

Conseguentemente ad un confronto con la tabella 3.2.1 possiamo notare come la colonna delle precipitazioni, nonostante abbia molti dati, non possa essere utile per questo progetto in quanto non ci sono dati di precipitazioni per il *NOAA*. Di conseguenza, oltre a tutte le colonne che non sono state reputate utili ai fini di questo progetto, verrà eliminata anche la colonna delle precipitazioni. Di seguito mostriamo il codice relativo all'operazione di *"drop"* di queste colonne.

```
# path dei file di mnw ancora da pulire
pathDir = "C:\\user\\....\\dati-mnw-grezzi\\"
pathNewFile = path+ "\\mnw_10_anni"
#prendo solamente i file
list_file =listdir(pathDir)

drop_column = [ 'rh', 'wind_speed', 'wind_direction',
'wind_gust','daily_rain', 'rad',
'uv'
]

for file_name in list_file:
print("Inizio - File : "+ file_name )
# leggo il file
df = pd.read_csv(pathDir+"\\"+file_name, sep=";")

#elimino le colonne a me inutili
df.drop(drop_column, inplace=True, axis=1)

#salvo il file
df.to_csv(pathNewFile+"\\"+file_name, index=None)
```

Ora che le osservazioni sono relative solo ai campi d'interesse, bisogna aggiungere delle ulteriori colonne, che non sono utili per le successive analisi ma permettono di sfruttare al meglio la funzione di *GROUP BY* di *MySQL* che, come illustrano le prossime porzioni di codice, servirà per poter trasformare questo data set con osservazioni di frequenza 5 minuti ad un data set con osservazioni orarie. Ovviamente prima di procedere è stato effettuato l'upload del data set fin qua ottenuto sul db precedentemente creato.

```
# per creare una tabella del mnw con colonna anno mese
    settimana giorno ora

create table mnw_definitivo as
select id_stazione , `date` ,
year(`date`) as anno, month(`date`) as mese, day(`date`) as
    giorno, hour(`date`)as ora ,
temperatura , punto_di_rugiada ,pressione
from mnw_10_anni ;
```

Completata la precedente *query* è possibile eseguire l'ultima operazione necessaria per poter effettuare le dovute analisi di somiglianza tra questa serie temporale e quella del *NOAA*. Come anticipato sopra, viene effettuata la trasformazione del data-set passando da una frequenza di osservazione di 5 minuti ad una frequenza oraria.

```
# creazione tabella di MNW di tutti i 10 anni, con
    osservazioni che sono ogni ora e non ogni 5 min
create table mnw_10_anni_definitivo as
select id_stazione ,
year(dataOra) as y, month(dataOra) as m, day(dataOra) as d,
    hour(dataOra)as h ,
avg(temperature) as temperature , avg(dew_point) as
    dew_point , avg(smlp) as smlp
from mnw_10_anni ma
group by id_stazione , y, m , d, h
;
```

3.2.3 Estrazione delle osservazioni in comune

In questa sezione viene illustrata quella che è stata l'ultima operazione della fase di pulizia dei dati. Il codice mostrato di seguito permette di estrarre, dalle serie temporali sia da *MNW* sia dal *NOAA*, le sole osservazioni corrispondenti in entrambe le serie temporali. Questo indica che se non esiste una *osservazione gemella* nel rispettivo data set, allora questa osservazione non verrà presa in considerazione e conseguentemente non verrà utilizzata per effettuare le analisi di similarità. Inoltre in questo modo, non è più rilevante quante osservazioni ci siano per ogni tipologia di osservazione, ma conteranno solamente quante coppie di osservazioni ci siano, per poi effettuare i calcoli di similarità come descritti nella sezione 2.2.5

```
# leggo i file noaa_definitivo e mnw_definitivo
noaa = pd.read_csv(pathNoaa , sep=";", index_col=None)
mnw  = pd.read_csv(pathMnw , sep=";", index_col=None)

# relazione tra i codici noaa e quelli del mnw
stazioni = {   "trn013": 160220 ,
   "lmb079": 160760 ,
   "lmb102": 160760 ,
   "ero099": 161480 ,
   "ero121": 161490 ,
   "tsc001": 161580 ,
   "tsc007": 161700
}

# loop in cui prendo mano a mano tutte le stazioni mnw e
    noaa accoppiati
# poi popolo un df con tutte le osservazioni che hanno
    stessa ora e data , uno per noaa e uno per mnw
# salvo il data frame

for cod_stazione in stazioni:
dfMnw  = mnw.loc[ mnw["id_stazione"] == cod_stazione ]
dfNoaa = noaa.loc[ noaa["id_stazione"] ==
    stazioni[cod_stazione] ]

#variabili temporanee in cui vengono salvate le
```

```
    osservazioni con una corrispondenza
tempMnw = pd.DataFrame(columns=
    dfMnw.columns.values.tolist())
tempNoaa = pd.DataFrame(columns=
    dfNoaa.columns.values.tolist())

# ciclo sulle date di mnw e verifico la corrispondeza con
    il noaa, in caso positivo memorizzo tale istanza che poi
    verra' salvata a fine iterazione
for dataOra in dfMnw["date"]:
print(dataOra)
b = dfNoaa.loc[ ( dfNoaa["date"] == dataOra) & (
    dfNoaa["id_stazione"] == stazioni[cod_stazione] ) ]
m = dfMnw.loc[ ( dfMnw ["date"] == dataOra) & ( dfMnw
    ["id_stazione"] == cod_stazione  ) ]


if( len(b) == 1 ): # controllo se esiste questa data in NOAA
# aggiungo l'istanza di NOAA corrispondente all'istanza MNW
    nella variabile temporanea
tempNoaa.loc[tempNoaa.shape[0]] = b.values[0]
tempMnw.loc[tempMnw.shape[0]] = m.values[0]

# creo una cartella per ogni coppia di stazioni NOAA-MNW
cartella = pathDir +
    "\\mnw_"+str(cod_stazione)+"_noaa_"+str(stazioni[cod_stazione])+"_"
os.mkdir(cartella)

fileNoaa = cartella + "\\"+str(stazioni[cod_stazione])
fileMnw = cartella + "\\"+str(cod_stazione)
tempNoaa.to_csv( fileNoaa , index=None)
tempMnw.to_csv( fileMnw , index=None)
```

3.2.4 Detezione degli outliers

Questo paragrafo mostra la metodologia adottata affinché i data-set del *NOAA* e del *MNW* vengano puliti da tutti i valori anomali, ossia *outliers*. Per l'individuazione degli *outliers* è stato deciso di utilizzare la tecnica del *Range Inter-*

Quantile, per cui viene reputato che un valore sia anomalo nel momento in cui non sia compreso tra : $[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$. Dove :

- Q1 : Indica il primo interquartile
- Q3 : Indica il terzo interquartile
- IQR: Indica differenza tra il terzo e il primo quartile, ovvero l'ampiezza della fascia di valori che contiene la metà "centrale" dei valori osservati.

In questo modo è stato possibile non prendere in considerazione valori anomali, che rischiano di inficiare la precisione delle successive analisi, assicurando migliori risultati. Di seguito è riportato il codice sviluppato.

```
listDir = os.listdir(pathDati)
for d in listDir:
dirTemp = pathDati + "\\"+ d
listFile = os.listdir(dirTemp)
df =[1,2]

df[0] = pd.read_csv(dirTemp + "\\" + listFile[0])
df[1]  = pd.read_csv(dirTemp + "\\" + listFile[1])
c = 0
# cartela in cui salvo i risultati
savePath = pathRisultati+"\\"+d
os.mkdir(savePath)
for dfTemporaneo in df:

#estratto tutti solamente temp, presisnoe e rugiada con indici
    la data
dfEstratto = pd.DataFrame(list(zip(
    dfTemporaneo["temperature"],dfTemporaneo["dew_point"],dfTemporaneo["smlp"])
    columns=[ "temperature", "dew_point", "smlp"], index=
    dfTemporaneo["date"])
dfBase = pd.DataFrame(list(zip(dfTemporaneo["id_stazione"])),
    index= dfTemporaneo["date"], columns=["id_stazione"])
temp = [0,1,2]
count = 0
#tengo solamente i dati Inliener di ogni variabile
for i in dfEstratto:
#calcolo il range interquantile per temeratura, rugiada e
    pressione
```

```
Q1 = dfEstratto[i].quantile(0.25)
Q3 = dfEstratto[i].quantile(0.75)
IQR = Q3 - Q1
#creo il singolo df che contiene le osservazioni inliers per
    temp , pressione , rugiada
temp[count] = pd.DataFrame(dfEstratto[i][ ((dfEstratto[i] >
    (Q1 - 1.5 * IQR)) & (dfEstratto[i] < (Q3 + 1.5 * IQR))) ] )

count +=1

# ora devo fare il merge dei dati basandosi sull'indice ,
    ovvero la data
mergeBase = pd.merge(dfBase, temp[0], how="left", on="date")

dfSx = pd.merge( mergeBase, temp[1], how="left", on="date")
noaaInliers = pd.merge( dfSx, temp[2], how="left", on="date")

#salvo il file

filePath = savePath + "\\"+ listFile[c]

noaaInliers.to_csv(filePath)
c += 1
```

Come viene mostrato, questo codice svolge quanto descritto precedentemente; è bene sottolineare come la detezione degli *outliers* sia stata effettuata singolarmente per ogni parametro. I valori dei parametri vengono estratti dal data set di partenza per poi selezionare solamente i dati validi, ed infine tutti i valori che sono stati riconosciuti come non anomali, che vengono nuovamente messi insieme. In questo modo è possibile scartare un singolo valore di una tupla, senza essere obbligati ad eliminare anche valori potenzialmente validi.

Nella seguente tabella è riportata la numerosità degli *outliers* identificati tramite il processo del *Range Interquartile* sopra descritto.

Come si nota, la numerosità degli outliers presenti nel data set del *NOAA* è maggiore rispetto al data set del *MNW* per il parametro della pressione; mentre per temperatura e punto di rugiada il *NOAA* presenta meno outliers. I risultati di

Codice	Temp.	P.Rug.	Press.	Codice	Temp.	P.Rug.	Press
trn013	1	42	1810	160220	155	451	83555
lmb079	2	77	1564	160760	6	291	54810
lmb102	2	36013	3282	160760	6	291	54810
ero099	728	741	2623	161480	14	21	52687
ero121	105	242	1004	161490	4	33	16574
tsc001	1	212	1268	161580	3	158	1283
tsc007	18290	18348	3333	161700	84	724	59862

Tabella 1: Sono riportati la numerosità di outliers di MNW, a sinistra, e di NOAA, a destra

tale identificazione, mostrati in tabella, relativamente a MNW sono incoraggianti nel considerare tale realtà un utile contributo al monitoraggio dei parametri ambientali, anche tenuto conto della estrema distribuzione delle stazioni di osservazione. Nella figura seguente sono riportate le serie temporali della temperatura ottenute a valle della fase di Data-Cleaning appena terminata.

Serie temporali delle temperature osservate da NOAA e MeteoNetWork, rispettivamente per le stazioni 161490 ed ero121, nel periodo 2010-2021

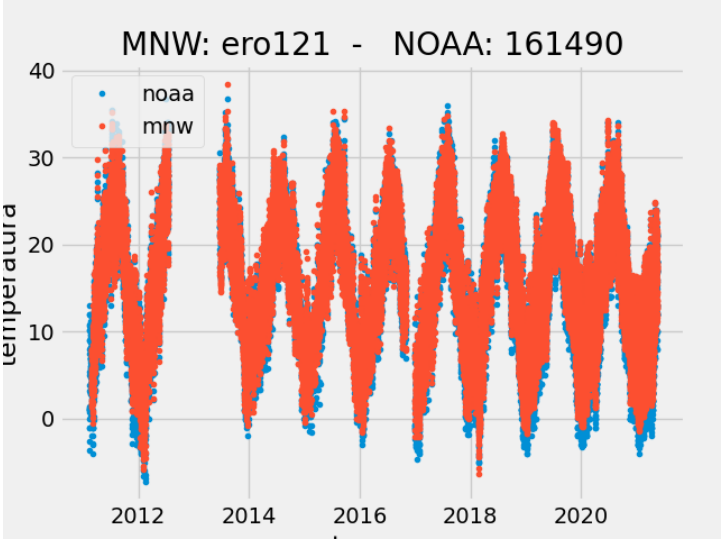


Figura 6: Serie temporali delle temperature osservate da NOAA e MeteoNetWork, rispettivamente per le stazioni 161490 ed ero121, nel periodo 2010-2021

La figura sopra esposta è rappresentativa delle serie temporali delle sette coppie di stazioni NOAA e MeteoNetwork che sono riportati in appendice.

3.3 Analisi di similarità

Questa sezione illustrerà quali sono stati i passaggi necessari per analizzare attivamente i dati ottenuti nella sezione precedente. Nello specifico verrà illustrato il codice che permette di analizzare la similarità tra le serie temporali di *NOAA* con *MNW*.

Come anticipato nella sezione 2.2.5, l'analisi di similarità adottata si basa sull'idea di verificare la differenza puntuale tra le osservazioni effettuate dal *NOAA* e quelle eseguite da *MNW*. In questo modo, grazie alla pulizia dei dati, è possibile effettuare una sottrazione semplice tra l'osservazione del *NOAA* con quella del *MNW*; questo perché per ogni osservazione, di entrambe le serie temporali, è presente un'osservazione con il medesimo *time-stamp* nella serie opposta. In questo modo è possibile verificare che vi sia coerenza tra le osservazioni e che rilevino le medesime condizioni meteo. Qui di seguito viene mostrato il codice.

```
listDirectory = os.listdir(pathDati)

for cartella in listDirectory:
print("Inizio Cartella : ", cartella)
pathTemp = pathDati+"\\"+cartella
listFile = os.listdir(pathTemp)

noaa= pd.read_csv( pathTemp + "\\"+ listFile[0])
mnw=  pd.read_csv( pathTemp + "\\"+ listFile[1])

differenzaTemp = noaa["temperature"] - mnw["temperature"]
differenzaDew = noaa["dew_point"] - mnw["dew_point"]
differenzaSmlp = noaa["smlp"] - mnw["smlp"]

#creao la cartella in cui salvo le differenze per Temperatura ,
    DewPoint e forse pressione
differenza = pd.DataFrame( list(zip(noaa["date"] ,
    differenzaTemp, differenzaDew, differenzaSmlp)) ,
    columns=["Data", "DiffTemp", "DiffDew", "DiffSmlp" ] )
name = "_" + listFile[0][0:6] + "__" + listFile[1][0:6] + "_"
nameFile = pathRisultati + "\\"+ name + ".csv"

differenza.to_csv( nameFile, index=False)
```

Una volta eseguito il codice sopra riportato si ottengono 7 file contenenti, per ogni stazione, le singole differenze tra le osservazioni *NOAA* e *MNW*. Per le osservazioni in assenza di corrispondenza con dato valido (valore *NULL*) il valore risultante sarà imposto come *NULL*.

Una volta ottenuti i file con le differenze si può procedere al calcolo della media e della deviazione standard dei risultati ottenuti, in modo da poter comprendere meglio il comportamento delle serie temporali.

```
listFile = os.listdir(pathDati)

# creao il df in cui memorizzo tutte le medie delle differenze
df = pd.DataFrame(columns=["noaaID", "mnwID", "temperatura",
    "sdtTemp" ,"difTempMax","difTempMin", "DewPoint",
    "stdDew","difDewMax","difDewMin","SMLP",
    "stdSmlp","difSmlpMax","difSmlpMin"])

for f in listFile:
print("Inizio ", f)
pathFile = pathDati + "\\" + f
dfTemp = pd.read_csv(pathFile)

#calcolo la media
mediaTemp = dfTemp.DiffTemp.mean()
mediaSMLP = dfTemp.DiffSmlp.mean()
mediaDew = dfTemp.DiffDew.mean()

maxTemp = dfTemp.DiffTemp.max()
maxSMLP = dfTemp.DiffSmlp.max()
maxDew  = dfTemp.DiffDew.max()

minTemp = dfTemp.DiffTemp.min()
minSMLP = dfTemp.DiffSmlp.min()
minDew  = dfTemp.DiffDew.min()

# calcolo della deviazione standard
stdTemp = dfTemp.DiffTemp.std()
stdSmlp = dfTemp.DiffSmlp.std()
stdDew  = dfTemp.DiffDew.std()
#ottengo codice noaa e mnw
```

```
noaaID = f[1:7]
mnwID = f[9:15]

#aggiungo quanto calcolato al df con i dati delle altre sstazioni
df.loc[df.shape[0]] = [noaaID, mnwID, mediaTemp, stdTemp
    ,maxTemp, minTemp , mediaDew, stdDew, maxDew, minDew,
    mediaSMLP, stdSmlp, maxSMLP, minSMLP]
pathFile = path
    +"\\..\\Risultati\\definitivi\\medie_differenze_NOAA-MNW.csv"
df.to_csv( pathFile , index=None)
```

Da notare che tutti i valori *NULL* non vengono presi in considerazione per il calcolo della media e della deviazione standard.

3.4 Risultati e analisi critica

Questa sezione verte sulla visualizzazione e sull'analisi dei risultati finali ottenuti attraverso l'implementazione dei codici che sono stati mostrati nella sezione precedente.

Una volta terminato con successo l'ultimo script mostrato nella sezione 3.3, si otterrà la seguente tabella:

Noaa	Mnw	Temp. media [C°]	Temp. stdev [C°]	Dew_Point media [C°]	Dew_Point stdev [C°]	SMLP media [hPa]	SMLP stdev [hPa]
160220	trn013	-10.2	4.8	-9.0	3.1	-	-
160760	lmb079	-0.2	1.1	-0.6	1.3	0.4	0.9
160760	lmb102	-0.5	0.9	-1.2	1.0	1.8	1.0
161480	ero099	-0.6	1.5	-0.2	2.1	0.6	0.8
161490	ero121	-0.6	1.3	0.3	1.9	0.3	0.6
161580	tsc001	-1.2	1.4	-0.6	1.3	0.9	0.7
161700	tsc007	-0.3	1.1	-0.6	1.3	-0.3	0.7

Tabella 2: In tabella sono riportate le stazioni NOAA e MNW corrispondenti con medie e deviazioni standard di temperatura, dew point (punto di rugiada) e pressione Sea Mean Level Pressure (SMLP)

Come si nota la stazione trn013 differisce rispetto alla stazione 160220 per valori significativi dei parametri considerati, che saranno oggetto di successive analisi.

Per quanto riguarda le altre stazioni i valori delle medie della temperatura, del Dew point e della pressione hanno valori compresi tra:

- [-1.2 ; -0.2]C per la temperatura
- [-1.2 ; 0.3]C per il punto di rugiada
- [-0.3 ; 1.8]hPa per la pressione sul livello del mare

Tali valori sono ovviamente da integrare con i valori di deviazione standard che risultano essere :

- [0.9 ; 1.5]C per la temperatura
- [1.0 ; 2.1]C per il punto di rugiada
- [0.6 ; 1.0]hPa per la pressione sul livello del mare

Nella figura di seguito riportata si evidenzia come il picco mostrato presenti valori coerenti ma con uno scarto di 1,5 C in linea con i valori di deviazione standard riportati in tabella.

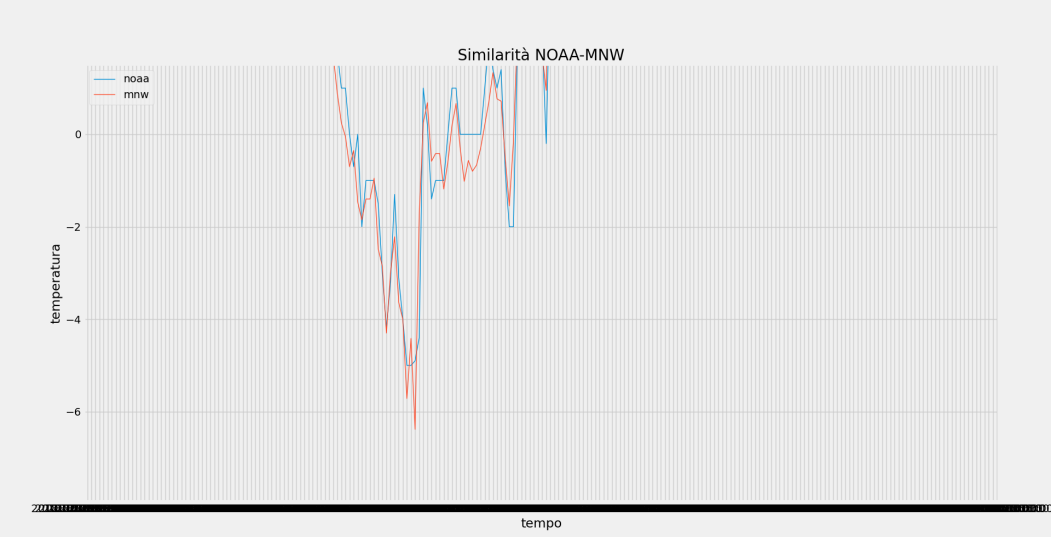


Figura 7: Particolare della serie temporale delle temperature osservate da NOAA e MeteoNetWork, rispettivamente per le stazioni 161490 ed ero121

Da tenere in conto che i valori derivano dalle osservazioni relative ad una finestra temporale di quasi 11 anni, con frequenza di osservazione oraria laddove vi sia contemporaneità delle osservazioni. Come precedentemente evidenziato, le

stazioni oggetto di tale confronto non sono posizionate nelle stesso punto ma possono risultare ad una distanza massima di 10 km. Come riportato nella tabella seguente:

Noaa	Mnw	ΔEst [km]	ΔNord [km]	ΔAltitudine [m]
160220	trn013	-7.3	1.2	1910
160760	lmb079	10.3	-5.4	-30
160760	lmb102	-1.7	-0.6	-8
161480	ero099	-6.7	-2.4	6
161490	ero121	1.3	-1.9	6
161580	tsc001	-0.5	-2.2	-3
161700	tsc007	-9.1	5.3	-7

Per quanto riguarda la stazione trn013, si nota una differenza planimetrica dell'ordine di 8km ed una differenza altimetrica di quasi 2000m. La stazione,(160220), si trova in trentino sulla cima del monte Paganella mentre la stazione prossima di MNW, trn013, è presso l'abitato di Lavis, ai piedi della montagna stessa. Questo aspetto è sicuramente non trascurabile in quanto le osservazioni risentono di differenza di quota così elevata. Si vuole allora tenere conto di questi dislivelli ipotizzando una atmosfera standard per la quale i gradienti verticali possono essere cosi valutati:

- gradiente verticale temperatura : -6.5 gradi/1000m
- gradiente verticale del dew point: non preso in analisi
- gradiente verticale barimetrico:
  $((1 + z * 2.2557696 * 10^{(-5)})^{(-5.2559324)-1)} * 1013.25$ [hPa]

Per quanto riguarda la temperatura l'espressione del gradiente è lineare con la differenza di quota; per la linearità dell'operatore media è possibile quindi applicare l'espressione del gradiente verticale alla media precedentemente ottenuta. Tale applicazione porta un valore della media delle differenze di temperatura riportate alla stessa quota pari a : 2,2 C. Tale risultato, tenendo conto che risente di una approssimazione di atmosfera standard, non differisce dai valori medi delle altre stazioni.

Per quanto riguarda la pressione, essendo un parametro non osservato dalla stazione *NOAA* nulla si può dire, mentre per quanto riguarda il dew point questo risente

della umidità presente in atmosfera e della temperatura ambiente; per questo motivo la temperatura di rugiada della stazione di Lavis (219 m) è stata corretta tenendo conto dell'umidità relativa valutata nella stazione della Paganella e della temperatura di Lavis portata alla medesima quota (2129 m).

A partire dalla temperatura media sul decennio di 2.6 C per la stazione della Paganella, e dal valore del punto di rugiada per lo stesso periodo di -2.8 C, mediante il diagramma psicrometrico [25] si ottiene una valore medio di umidità dell'ordine del 66%. La temperature della stazione di Lavis per il decennio risulta essere di 13 C che riportati alla quota della stazione della Paganella diventano 0.6 C ai quali corrisponde una temperatura di rugiada intorno a -4.4 C per valori di umidità relativa del 66%. Il confronto con la media dei *dew_point*, punto di rugiada, della stazione della Paganella (-2.8 C) risulta pertanto dell'ordine di 1.6 C contro i -9 C presenti in tabella. Questo risultato avvalora ulteriormente la corrispondenza delle osservazioni della stazione di MeteoNetWork nei confronti della stazione *NOAA*.

4 Conclusioni

Nel presente lavoro di tesi è stato affrontato il processo che ha portato alla validazione delle osservazioni che vengono effettuate dalle stazioni di *MeteoNetWork*. La presentazione inizia attraverso la descrizione del contesto di riferimento per quanto riguarda le conoscenze teoriche relative alle serie temporali di dati ambientali. Si forniscono inoltre alcuni riferimenti anche per Internet Of Things, Big data e per le moderne metodologie per la raccolta di osservazioni meteorologiche, in modo da presentare il più ampio e chiaro quadro di riferimento possibile. Successivamente inizia la fase di progetto in cui vengono riportati gli obiettivi e le scelte progettuali, per le quali si forniscono motivazioni adeguate alla scelta effettuata. Sono stati successivamente mostrati i dati di partenza relativi alle 13 stazioni prese in analisi. Sono quindi descritti i passaggi ritenuti necessari per poter completare la fase di *Data-Cleaning* per tutte le stazioni, affinché si possa successivamente passare alla fase di analisi in cui vengono paragonate le serie temporali del MeteoNetWork con quelle del NOAA. Viene quindi effettuata una analisi critica dei risultati da cui è possibile affermare che la presenza di una organizzazione come MeteoNetWork sia di grande vantaggio per future analisi ambientali. I risultati ottenuti infatti mostrano come entrambe le serie temporali, per ogni coppia di stazioni, siano assolutamente coerenti e come rilevino le medesime condizioni meteorologiche, a fronte di lievi differenze tra i valori osservati. Tali differenze sono del tutto giustificabili, in quanto le stazioni non sono co-locate, ma possono risultare ad una distanza massima di 10 km l'una dall'altra. Inoltre la differenza altimetrica gioca un ruolo importante nell'analisi dei risultati, in quanto le stazioni con maggiori differenze di rilevazioni sono le stesse che soffrono di una differenza altimetrica importante; ne sono un esempio le stazioni TRN013 e 160220, rispettivamente site in Lavis (219 m) e Paganella (2129 m). I risultati di questo progetto, relativamente a MeteoNetWork, sono incoraggianti nel considerare tale realtà un utile contributo al monitoraggio dei parametri ambientali, anche tenuto conto della estrema distribuzione delle stazioni di osservazione che risulterà essere un aspetto chiave per futuri studi. L'importanza di una estrema diffusione di stazioni di osservazione è di primaria importanza per lo studio dei fenomeni meteorologici, fornendo informazioni preziose per l'innesco dei modelli meteorologici a scala regionale; per i medesimi motivi ad una distribuzione spaziale capillare, va ad affiancarsi l'importanza di una risoluzione temporale. Le tecniche adottate nel presente lavoro potranno essere applicate ad altri campi, nei quali l'integrazione di differenti fonti di dati permetta una

maggiore conoscenza del fenomeno in studio; all'eterogeneità dei dati si deve affiancare una validazione in termini di similarità degli stessi. Può essere questo il caso di sensori distribuiti per la misura del particolato atmosferico, così come l'integrazione di differenti sensori per il monitoraggio di fertilizzanti, concimi e agrofarmaci dei quali il rapido evolversi della così detta agricoltura smart potrà beneficiare.

Riferimenti bibliografici

[1] What is time series data? - influxdata. [Online]. Available: https://www.influxdata.com/what-is-time-series-data/

[2] Philippe Esling, Carlos Agon. Time-series data mining. ACM Computing Surveys, Association for Computing Machinery, 2012, 45 (1), pp.12. ff10.1145/2379776.2379788ff. ffhal-01577883f

[3] High Performance Discovery In Time Series: Techniques and Case Studies, Shasha e Zhu 2004

[4] Finding Motifs in Time Series, Jessica Lin Eamonn Keogh Stefano Lonardi Pranav Patel, 2002

[5] What is big dat data?- oracle [online]. Available : https://www.oracle.com/it/big-data/what-is-big-data/

[6] Serie Storiche Finanziarie:Studio e Previsione - Michele Caselle. [online] Available: https://terna.to.it/tesi/gomitoni.pdf

[7] Cosa sono i Big Data ? - Oracle [Online] Available : https://www.oracle.com/it/big-data/what-is-big-data/

[8] Home Page - MeteoNetwork [Online] Available : https://www.meteonetwork.it/

[9] Garbage in, garbage out - wikipedia. [Online] Available : https://it.wikipedia.org/wiki/Garbage-in-garbage-out/

[10] State of the IoT 2018: Number of IoT devices now at 7B – Market accelerating- oracle [online]. Available : https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/

[11] L'Internet delle cose (IoT): cos'è e come rivoluzionerà prodotti e servizi [online]. Available : https://www.zerounoweb.it/analytics/big-data/internet-of-things-iot-come-funziona/

[12] I satelliti meteorologici- MeteoTrentino [online]. Available : http://content.meteotrentino.it/dati-meteo/meteosat/satelliti.pdf

[13] NOWCASTING: significato e strumenti- IlMeteo [online]. Available : https://www.ilmeteo.it/notizie/nowcasting-significato-e-strumenti-meteo-previsioni

[14] National Centers for Environmental Information- NCEI [online]. Available : https://www.ncei.noaa.gov/

[15] Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data - Tableau [online]: https://www.tableau.com/learn/articles/what-is-data-cleaning

[16] About Copernicus - Copernicus [online]: https://www.copernicus.eu/en/about-copernicus

[17] Accesso ai dati - Copernicus [online]: https://www.copernicus.eu/it/accesso-ai-dati

[18] Home page - ECAD [online]: https://www.ecad.eu/

[19] Partecipants - ECAD [online]: https://www.ecad.eu/dailydata/datadictionaryparticipants.ph

[20] Agenzia regionale per la protezione ambientale - Wikipedia [online]: https://it.wikipedia.org/wiki/Agenzia-regionale-per-la-protezione-ambientale

[21] Home page - Qgis [online]: https://www.qgis.org/it/site/index.html

[22] isd lite format - NOAA [online]: https://www.ncei.noaa.gov/pub/data/noaa/isd-lite/isd-lite-format.pdf

[23] MeteoNetwork API v3 (REV2) - MNW [online]: https://api.meteonetwork.it/documentation.html#tag/Archive-Data

[24] Progetto VANGUAR - Wikipedia [online]: https://it.wikipedia.org/wiki/Vanguard_2

[25] Termodinamica e trasmissione del calore - Yunus A.Cengel, McGraw-Hill, terza edizione

Appendice

Di seguito vengono riportati i grafici delle serie temporali per temperatura, pressione e punto di rugiada, dal 2010 al 2021, relativamente delle stazioni NOAA e MeteoNetWork limitrofe, per le quali:

- il colore arancione identifica le stazioni di MeteoNetWork mentre il coloro azzurro quelle del NOAA;
- i grafici delle stazioni trn013 (Lavis) e 160220 (Paganella) risentono della differenza di altitudine di 1910 m;
- si rilevano assenze di dati in taluni periodi temporali in particolare per le osservazioni di pressione del NOAA.

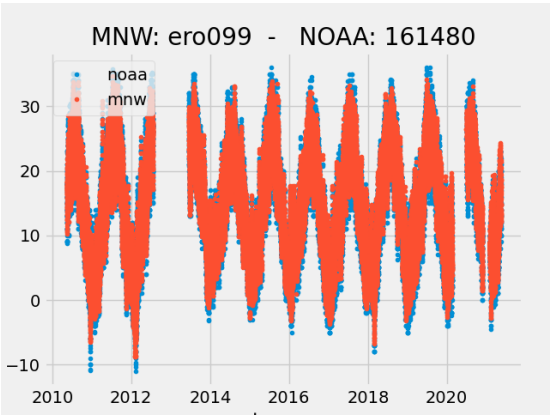
NOAA

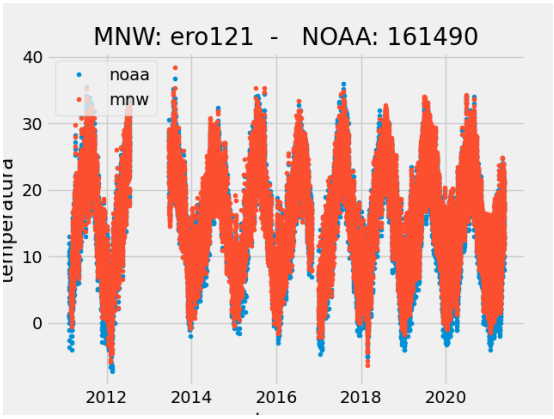
Codice	Località	Provincia	Longitudine	Latitudine	Altitudine(m)
160220	Paganella	Provincia autonoma di Trento	11.033	45.150	2129
160760	Aereoporto Orio Al Serio	Provincia di Bergamo	9.704	45.674	238
161480	Cervia	Provincia di Ravenna	12.307	44.224	5
161490	Rimini	Provincia di Rimini	12.612	44.020,	12.5
161580	Pisa	Provincia di Pisa	10.393	43.684,	1.8
161700	Firenze	Provincia di Firenze	11.205	43.810	44

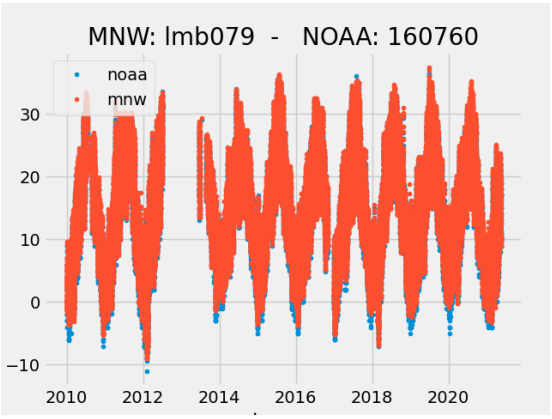
MeteoNetWork

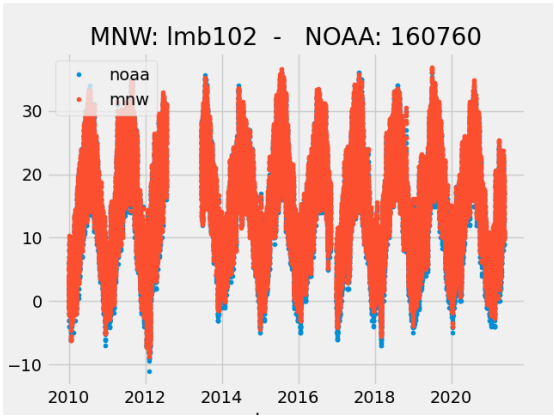
Codice	Località	Provincia	Longitudine	Latitudine	Altitudine(m)
trn013	Lavis	Provincia autonoma di Trento	11.101	46.139	219
lmb079	Paladina	Provincia di Bergamo	9.609	45.724	268
lmb102	Seriate	Provincia di Bergamo	9.720	45.680	246
ero099	Cervia	Provincia di Ravenna	13.369	44.246	0
ero121	Rimini	Provincia di Rimini	12.600	44.038	7
tsc001	Pisa	Provincia di Pisa	10.398	43.704	5
tsc007	Firenze	Provincia di Firenze	11.289	43.761	51

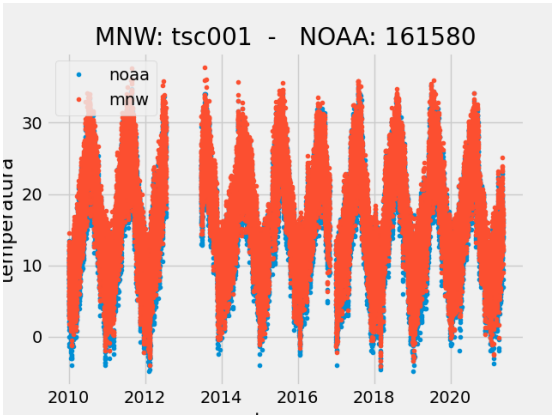
Temperatura

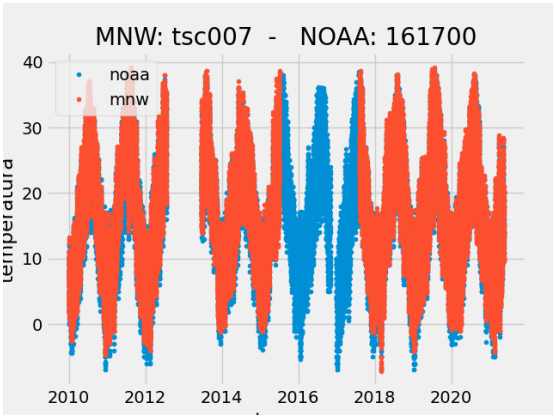


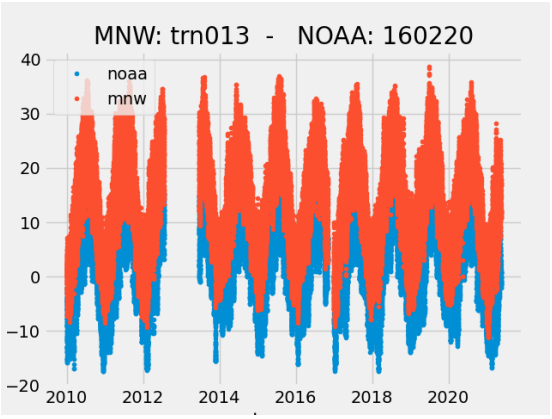




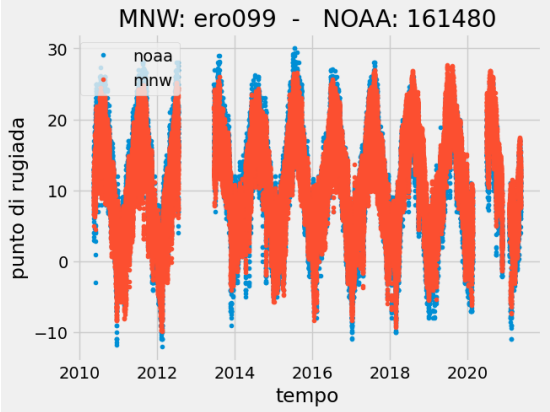


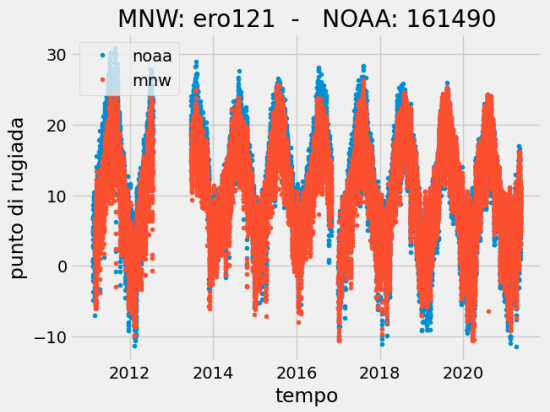


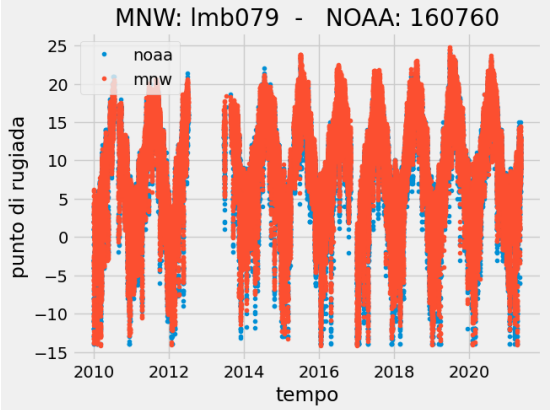


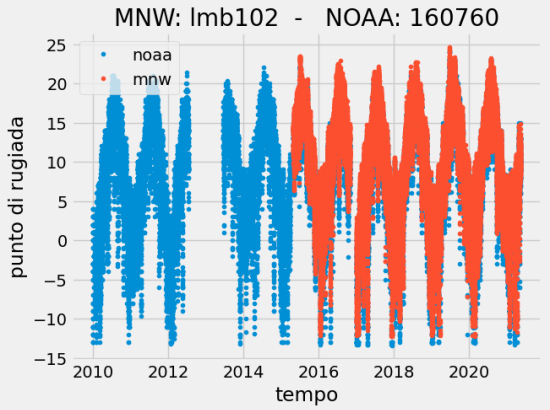


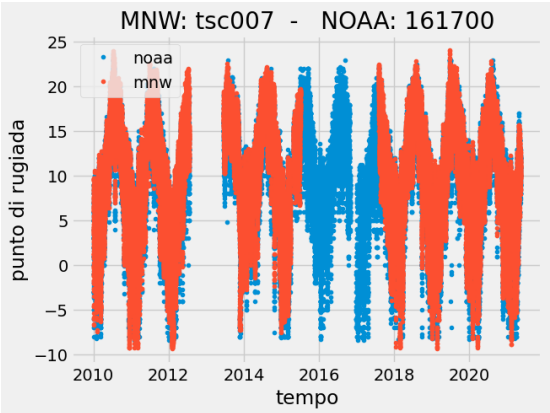
Punto di Rugiada

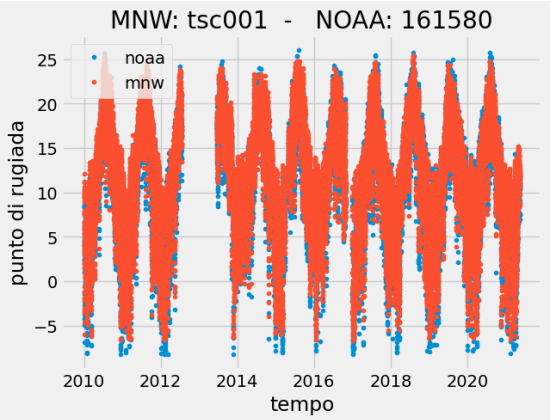


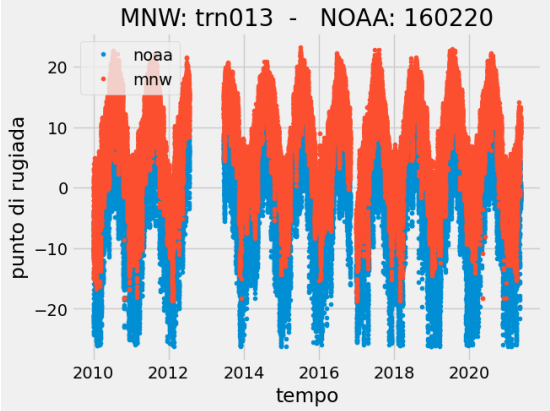




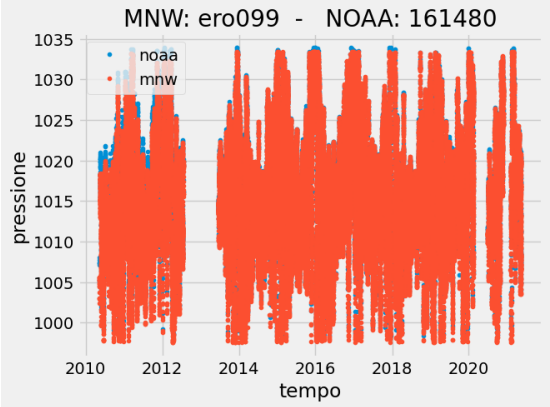


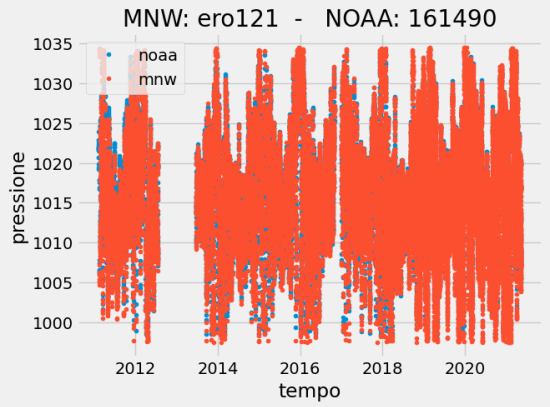


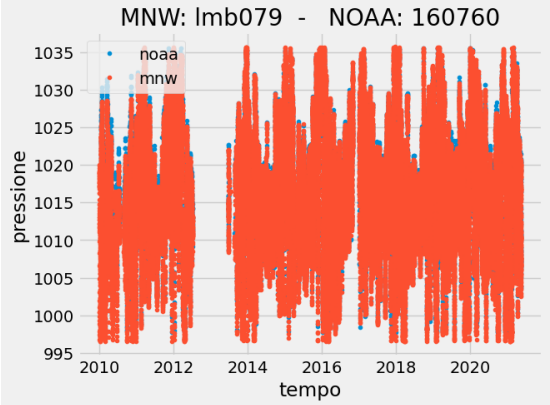


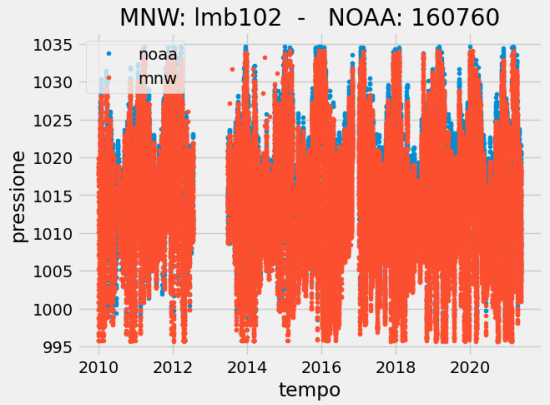


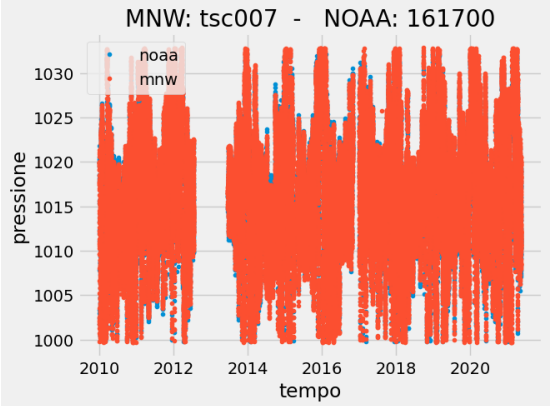
Pressione

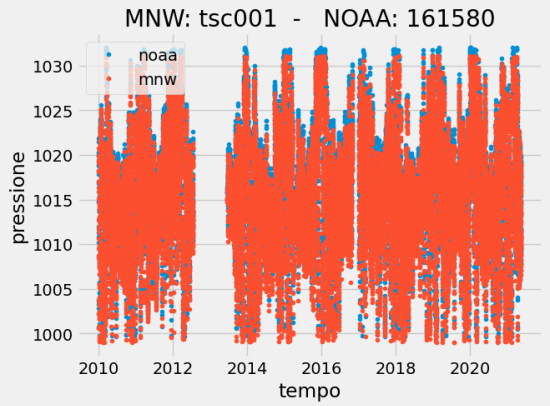

MNW: ero099 - NOAA: 161480
noaa
mnw
pressione
tempo

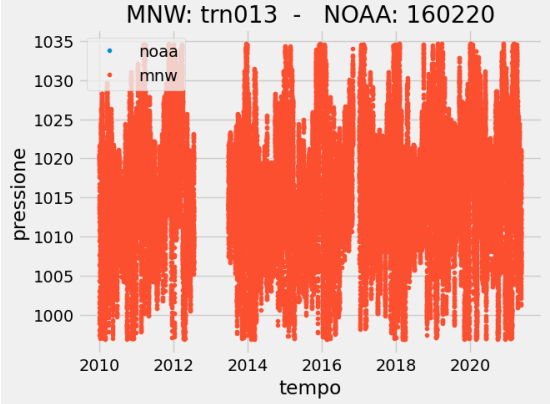

MNW: ero121 - NOAA: 161490
noaa
mnw
pressione
tempo


MNW: lmb079 - NOAA: 160760
noaa
mnw
pressione
tempo


MNW: lmb102 - NOAA: 160760
noaa
mnw
pressione
tempo


MNW: tsc007 - NOAA: 161700
noaa
mnw
pressione
tempo


MNW: tsc001 - NOAA: 161580
noaa
mnw
pressione
tempo


MNW: trn013 - NOAA: 160220
noaa
mnw
pressione
tempo