Università di Pisa

Department of Computer Science
MSc in Data Science and
Business Informatics

Distributed Data Analysis & Mining

U.S. Air Pollution Analysis in Apache Spark

Autori

Gaetano Antonicchio - 616685
Salvatore Lusito - 609000
Martina Sustrico - 533252
Luca Palla - 533605

2021/2022

Indice

1 Introduzione al Dataset

Il seguente DataSet contiene dati riguardanti i livelli di **inquinamento atmosferico** registrati negli U.S. dal 2000 al 2016. I dati sono stati forniti e documentati dall'**EPA U.S**, ente che fornisce informazioni riguardanti le rilevazioni giornaliere effettuate su tutto il territorio americano in un finestra temporale lunga 16 anni. Gli inquinanti per i quali sono state eseguite le misurazioni sono: **SO2, O3, NO2, CO**.
Complessivamente il Dataset ha una dimensione pari a 405.95MB ed è composto da 1,746,661 osservazioni. Di seguito riportiamo la descrizione di alcuni attributi presenti del dataset, i quali necessitano una descrizione più dettagliata.

- **State Code** : Codice identificativo assegnato dall'US EPA ad ogni stato
- **County code** : Codice della Contea di uno specifico Stato assegnato dall' US EPA.
- **Site Num** : Numero unico del Sito all'interno della specifica Contea assegnato dall' US EPA.
- **Address**: Indirizzo del sito di monitoraggio.
- **State** : Stato del sito di monitoraggio.
- **County** : Contea del sito di monitoraggio.
- **City** : Città del sito di monitoraggio.
- **Date Local** : Data del monitoraggio.

2 Data Pre-Processing

Dopo aver caricato il DataSet, un subset di features è stato rimosso in quanto poco utile ai fini delle analisi successivamente condotte. Tali features riguardano le **Units** (unità di misura) dei vari inquinanti, composte da un singolo valore ripetuto in ogni record. Inoltre è stato rimosso un **ID** rappresentante la singola misurazione ed il campo **Address** che, pur facendo riferimento all'indirizzo nel quale è collocata la stazione di rilevazione, non rappresenta un'informazione essenziale ai fini delle analisi geografiche condotte successivamente, già supportate da altre features quali Stato, Contea e Città.

I dati sono stati inseriti in un **Data Frame**, permettendo così una più semplice manipolazione.

```
root
 |-- State Code: integer (nullable = true)
 |-- County Code: integer (nullable = true)
 |-- Site Num: integer (nullable = true)
 |-- State: string (nullable = true)
 |-- County: string (nullable = true)
 |-- City: string (nullable = true)
 |-- Date Local: string (nullable = true)
 |-- NO2 Mean: double (nullable = true)
 |-- NO2 1st Max Value: double (nullable = true)
 |-- NO2 1st Max Hour: integer (nullable = true)
 |-- NO2 AQI: integer (nullable = true)
 |-- O3 Mean: double (nullable = true)
 |-- O3 1st Max Value: double (nullable = true)
 |-- O3 1st Max Hour: integer (nullable = true)
 |-- O3 AQI: integer (nullable = true)
 |-- SO2 Mean: double (nullable = true)
 |-- SO2 1st Max Value: double (nullable = true)
 |-- SO2 1st Max Hour: integer (nullable = true)
 |-- SO2 AQI: double (nullable = true)
 |-- CO Mean: double (nullable = true)
 |-- CO 1st Max Value: double (nullable = true)
 |-- CO 1st Max Hour: integer (nullable = true)
 |-- CO AQI: double (nullable = true)
```

Figura 1: Schema dei Dati iniziali

2.1 Gestione dei Missing Values

Successivamente, è stata condotta un'analisi sui Missing Values. Si è notato come le uniche due features contenenti valori nulli siano quelle facenti riferimento all'indice AQI per gli inquinanti **SO2 e CO**, per un totale di **873,323** corrispondenti al **50% dei record totali**. Tuttavia, è stato notato un pattern concernente tali valori nulli: per ogni giorno considerato nei dati, infatti, vengono riportate varie misurazioni per ogni *Site Num*, alcune delle quali nulle.
Questo perchè per ogni giorno vengono eseguite un numero di misurazioni differenti in base al tipo di inquinante. Si è dunque deciso di sostituire tali valori nulli andando a calcolare i valori medi ottenuti per le rilevazioni non nulle della medesima giornata relative a quell'inquinante.
Inoltre, analizzando i dati si è notato che il numero identificativo di un *Site Num* (e quindi di un sensore) può essere lo stesso di un sensore diverso sito in una Contea diversa. E' stata dunque creata una **Window** per partizionare i dati, considerando la tripla **Site Num, Date Local e City**.
In questo modo i valori nulli sono stati sostituiti tenendo conto di tali partizioni senza perdere la correlazione naturale con le rilevazioni temporalmente e geograficamente vicine.

3 Studio della Correlazione

Per eseguire uno studio sulla correlazione tra le varie features, in primo luogo queste sono state filtrate considerando solo valori numerici, escludendo quindi tutti i dati geografici (di tipo stringa) e temporali. Nella figura sottostante **(Figura 2)** si ha la visualizzazione della matrice di correlazione ottenuta utilizzando il coefficiente di **Pearson** :

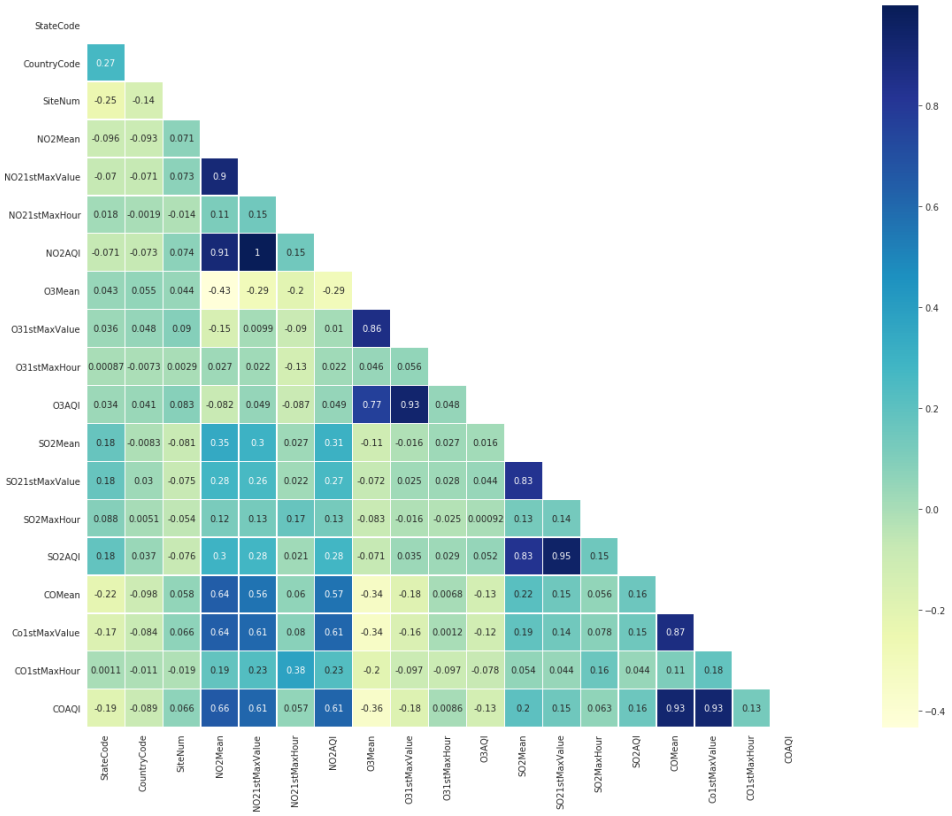


Figura 2: Matrice di Correlazione

Dalla matrice di correlazione non si evincono dipendenze statisticamente significative tra le varie features (ignorando quelle concernenti lo stesso inquinante), ad eccezione di **CO ed NO2**, i quali sembrano positivamente correlati. Tale correlazione viene di fatto confermata dalla comunità scientifica, che suggerisce come tali inquinanti siano in relazione tra loro in quanto scaturiti spesso da fenomeni medesimi (quali incendi o fenomeni di combustione).

4 Data Visualization

4.1 Geo-visualizzazioni e analisi dei trend stagionali

Dal momento che i livelli di inquinanti possano variare su base stagionale, sono stati analizzati i trend di questi ultimi con tale granularità. Abbiamo inizialmente preparato lo Spark Dataframe per eseguire un analisi stagionale in modo da ottenere rilevazioni medie in ogni stato per inquinante e stagione. I valori sono stati ottenuti eseguendo in primo luogo una trasformazione degli attributi in input, infatti sono state ricavate le colonne **month** e **year** dallo split della colonna **Date local**. Dopodichè è stata creata una funzione **"seasonalize"** la quale esegue una mappatura per stagione in base al mese e all'anno (crea una nuova colonna **season-year**).
Il **mapping** è stato così definito: ai mesi appartenti alla stagione invernale (da dicembre a febbraio) è stato assegnato il valore 1; a quelli della stagione primaverile (da marzo a maggio) è stato assegnato il valore 2; a quelli della stagione estiva (da giugno ad agosto) è stato assegnato il valore 3; infine, a quelli della stagione autunnale (da settembre a novembre) è stato assegnato il valore 4. I valori medi sono stati calcolati tramite una query sql applicata sullo spark dataframe, mediante la quale abbiamo eseguito una group by sugli attributi **State** e **season-year** e calcolato la media come funzione di aggregazione.

4.1.1 Analisi dell'O3 (Ozono)

Abbiamo deciso di osservare le variazioni stagionali per anno dei livelli di O3 utilizzando una mappa interattiva (**choropleth**) offerta dalla libreria **Plotly**. Eseguendo una sliding completo partendo dal 2000 fino al 2016, abbiamo notato subito un aumento dei livelli di O3 nei periodi estivi di ogni anno. E' stato interessante osservare come, durante la stagione invernale, i livelli di questo inquinante siano minimi, per poi aumentare guadualmente fino a raggiungere un picco durante il periodo estivo. A fine dimostrativo, si mostra la variazione annuale relativa all'anno 2012 (come si può notare in **Figura3**).

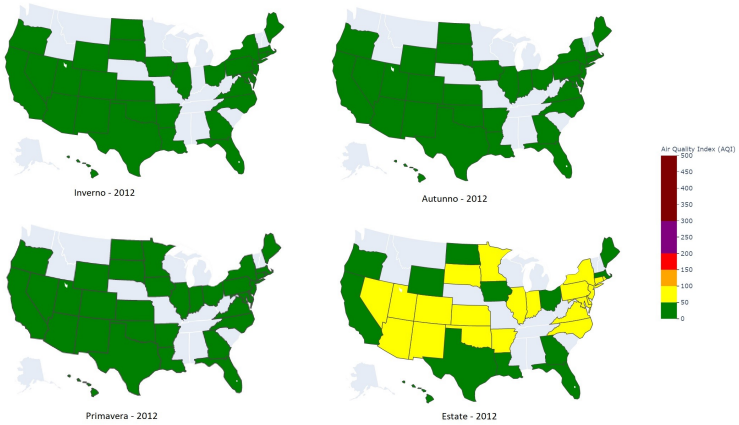


Figura 3: Variazioni Ozono(O3) nel 2012 per stagione

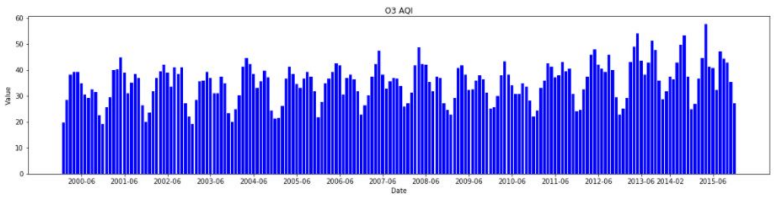


Figura 4: Trend Ozono(03) dal 2000 al 2016

La stagionalità dell'O3 è grafiamente osservabile dal barplot rappresentato in **Figura 4**, che mostra nel dettaglio come variano i livelli dell'inquinante tra i vari anni. Notiamo dei picchi nei mesi di Giugno di ogni anno, con trend in leggero aumento a ridosso degli anni 2015 - 2016.

L'Ozono (O3) è l'inquinante più pericoloso, il quale può causare seri danni al sistema respiratorio. Il motivo per quale l'O3 raggiunge picchi durante i mesi estivi, è dovuto alle alte temperature ed al caldo torrido. Infatti, l'intensa radiazione solare dei mesi estivi, combinata all'inquinamento umano e naturale, da il via a una serie di reazioni chimiche che favoriscono la maggiore formazione dell'agente inquinante.
E' importante notare che il grafico in **Figura 3** mostra i valori medi calcolati per ogni stagione e anno per ogni stato, quindi alcune rilevazioni potenzialmente dannose potrebbero essere nascoste da tale manipolazione. Per eseguire un analisi più accurata, è stato deciso di analizzare un livello di granularità più dettagliato, osservando il contributo di ogni **città** all'AQI O3 generale dello stato, su base annua. In **Figura 5** vengono mostrati i livelli di O3 estivi misurati nelle città del territorio americano. E' interessante notare che dal grafico in **Figura 4**, sembrerebbe che in California i livelli di O3 durante il periodo estivo siano nella norma (confermato dall'indicatore verde AQI). Tuttavia, si osserva come le rilevazioni con l'indice più pericoloso (alto) di O3 siano proprio in Califorina. Difatti, le città **Rubidoux** (O3 pari a 87.53 medi) e **Fontana** (O3 pari a 95.25 medi) sono quelle con gli indici più alti. Al contrario di quanto ci si possa aspettare invece, le città più popolose, come Los Angeles, mostrano un livello di O3 medio che non supera i 35.23. Questo è dovuto dal fatto che, sebbene lo smog contribuisca fortemente nei livelli di O3 presenti nell'aria, le sostanze gassose (tossiche) emesse dalla maggior parte delle industrie della Silicon Valley (come quelle presenti in Fontana e Rubidaux) incrementano maggiormente i livelli di inquinamento. Concentrandoci invece sul versante Est, si nota che città come New York, Oklahoma e Boston, non superano i 55 di AQI medi durante i periodi estivi.

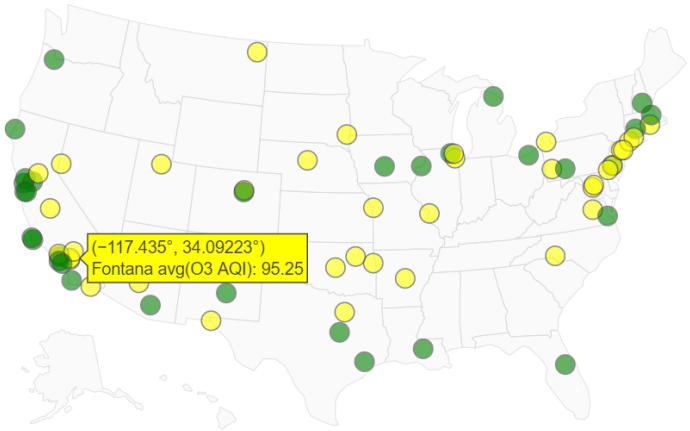


Figura 5: Trend 03 dal 2000 al 2016

Il bar plot in **Figura 6**, mostra le 30 città più inquinate (per O3) negli USA. I valori sono stati ottenuti eseguendo una group by sull'attributo *City* e applicando una media calcolata su un periodo che va dal 2000 al 2016.

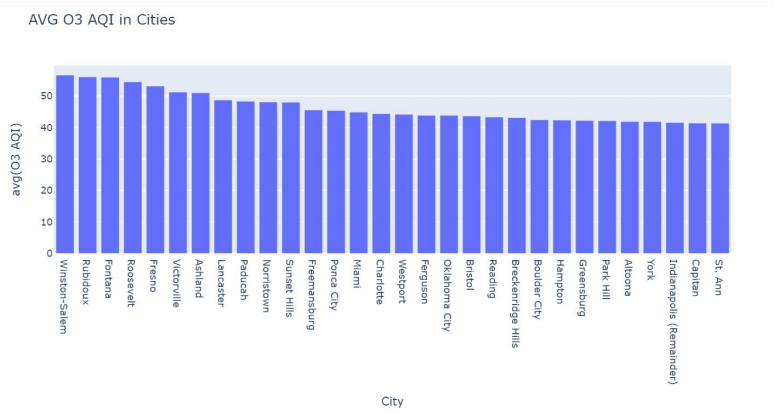


Figura 6: Top 30 Città per emissioni di O3 dal 2000 al 2016.

4.1.2 Analisi dell'SO2 (Anidride Solforosa)

L'anidride solforosa è un gas tossico che viene rilasciato naturalmente dall'attività vulcanica ed è scaturito da fattori come l'estrazione del rame e della combustione di combustibili fossili contenenti zolfo.
L'analisi dei dati disponibili non mostra una situazione generalmente preoccupante, in quanto i valori medi registrati nel corso degli anni non superano soglie critiche.

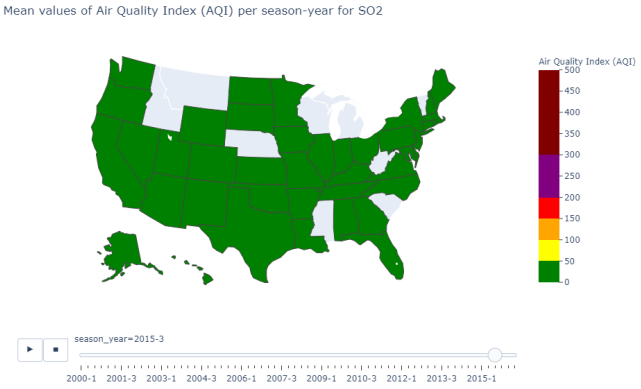


Figura 7: Valori medi di emissioni di SO2 negli Stati Uniti.

Inoltre, è possibile notare come visualizzando il trend dei valori medi di emissione negli ultimi anni, quest'ultimo sia in netta tendenza decrescente, mostrando un calo pressoché costate nel corso degli ultimi 15 anni.
Scendendo di granularità ed analizzando i dati inerenti le singole città, in **Figura 9** è possibile notare come le città con valori di emissioni più alte per questo inquinante siano **Reading e Beaver Falls**, con valori medi pari a 30. Tale AQI rientra nella "safe zone", non rappresentando un rischio per la salute.

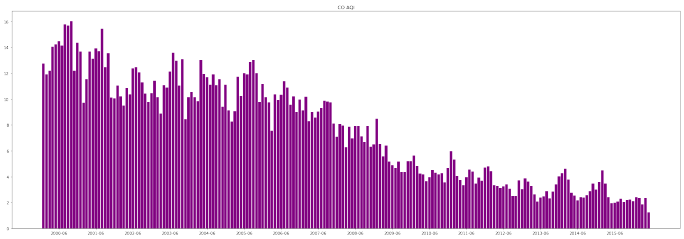


Figura 8: Trend delle emissioni di SO2 negli Stati Uniti.

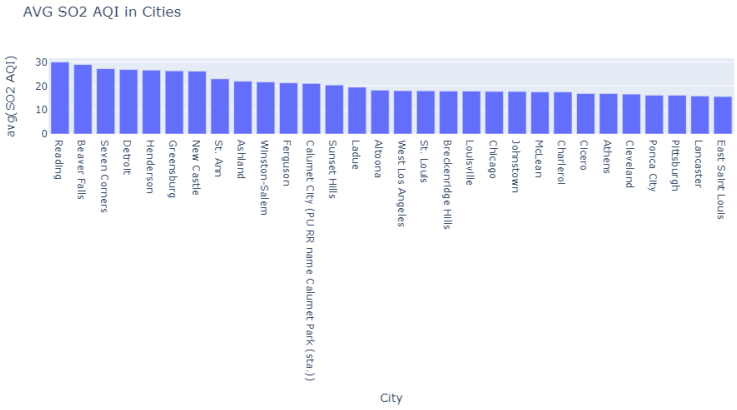


Figura 9: Top 30 Città per emissioni di SO2 dal 2000 al 2016.

4.1.3 Analisi dell'NO2 (Biossido di Azoto)

Il biossido di azoto è un gas tossico che si forma nei processi di combustione (centrali termoelettriche, riscaldamento, traffico) e processi produttivi senza combustione (produzione di acido nitrico, fertilizzanti azotati, ecc.) È un gas irritante per l'apparato respiratorio e per gli occhi che può causare bronchiti fino anche a edemi polmonari e decesso. Contribuisce alla formazione dello smog fotochimico e al fenomeno delle "piogge acide". Anche in questo caso, l'analisi delle emissioni medie non ha rilevato situazioni critiche sul territorio degli USA.

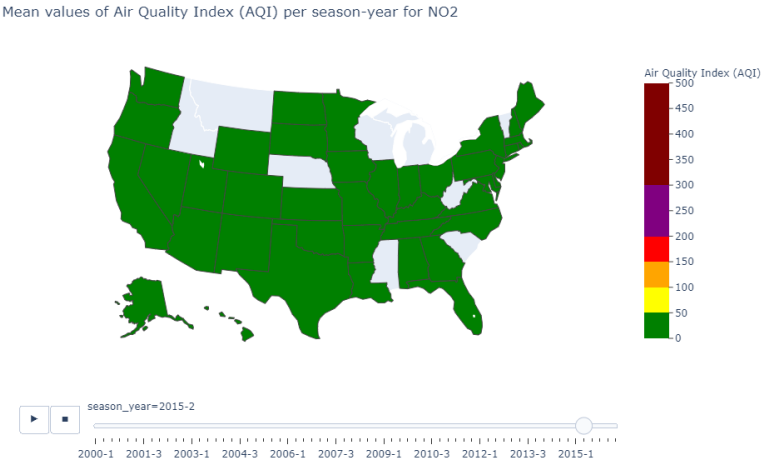


Figura 10: Valori medi di emissioni di NO2 negli Stati Uniti.

L'analisi dei trend rileva come vi sia stato un trend decrescente anche di questo inquinante, seppur in modo meno significativo rispetto alla SO2. L'analisi sulle singole città mostra come quelle con presenza di NO2 più massiccia siano **Bakersfield e Chicago**, seppur con valori medi di poco superiori alle 40 unità di AQI.

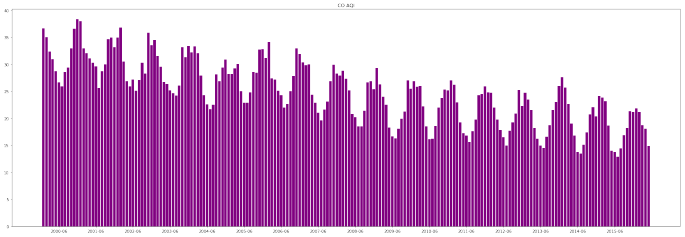


Figura 11: Trend di emissioni di NO2 negli Stati Uniti.

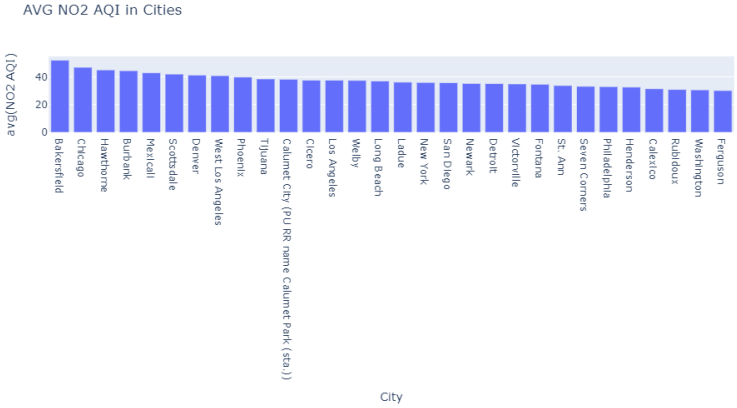


Figura 12: Top 30 Città per emissioni di NO2 dal 2000 al 2016.

4.1.4 Analisi del CO (Monossido di Carbonio)

Il monossido di carbonio assume particolare rilevanza tra gli inquinanti prodotti dalla combustione. E' un gas tossico prodotto per combustione incompleta di qualsiasi materiale organico. I valori medi registrati sul territorio statunitense non mostrano andamenti al di sopra della soglia minima di rischio.

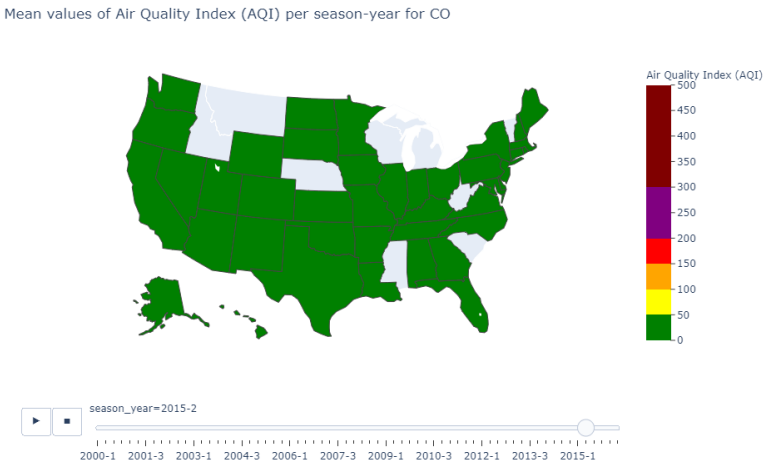


Figura 13: Valori medi di emissioni di CO negli Stati Uniti.

L'analisi dei trend, inoltre, dimostra una costante diminuzione della quantità di CO nell'aria avvenuta negli ultimi anni, con picchi nei mesi invernali sicuramente dovuti al maggiore impiego di fonti di riscaldamento a combustione. Le città con emissioni di CO più alte sono in questo caso **MexiCali e Hawthorne**, con valori medi tuttavia ben al di sotto di una soglia preoccupante (circa 20).

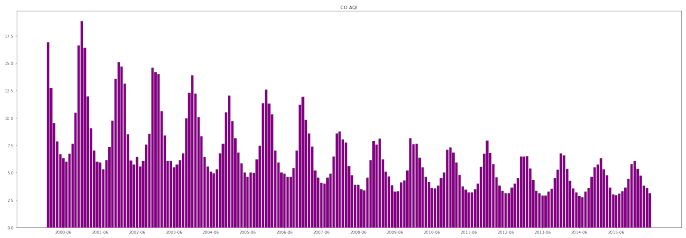


Figura 14: Trend di emissioni di CO negli Stati Uniti.

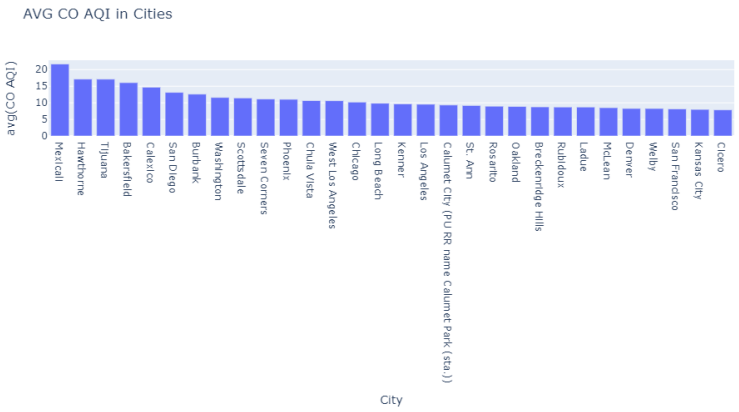
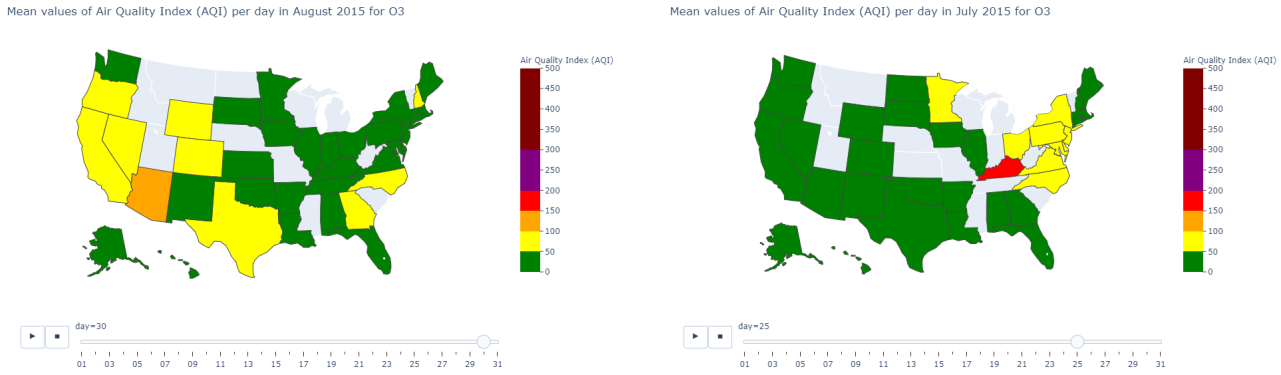


Figura 15: Top 30 Città per emissioni di CO dal 2000 al 2016.

4.1.5 Considerazioni Generali e accenni su analisi aggiuntive.

Dopo aver condotto un'analisi a più livelli di granularità sugli andamenti degli inquinanti sul territorio statunitense, ci si è soffermati sui trend dell'**03**: tale inquinante, come citato in precedenza, è il più pericoloso e tende a manifestarsi in quantità massicce durante i mesi estivi, a causa di incendi, caldo torrido e siccità. Si è dunque deciso di osservare con più attenzione questo pericoloso fenomeno considerando i dati relativi all'**estate 2015**, in quanto di più recente valutazione (è importante sottolineare che le rilevazioni del 2016 erano incomplete e quindi non adatte da prendere come riferimento). Sono stati dunque plottati i valori singoli dei giorni estivi più caldi (di Luglio e Agosto), rivelando uno scenario tutt'altro che roseo:



(a) Rilevazioni di O3 30 Agosto 2015

(b) Rilevazioni di O3 25 Luglio 2015

Figura 16: Analisi dei valori di O3 nei mesi estivi del 2015

Si noti come in molti stati i valori registrati superano la soglia minima di benessere, fissata ad un AQI di massimo 50. In particolare, lo stato dell'Arizona (in arancio nel primo plot) e del Kentucky (in rosso nel secondo) hanno valori particolarmente preoccupanti, superiori al tetto di 100 e rappresentanti un reale rishio per la salute dei cittadini. Il picco massimo è stato raggiunto proprio dal **Kentucky** nella giornata del 25 Luglio, con un valore pari a 174.

4.2 Analisi dei Trend negli USA

Si è osservato come negli USA l'emissione di inquinanti quali la CO, SO2 e NO2 sia in calo rispetto al 2000. In particolare, è interessante notare come i livelli di Anidride Solforosa (SO2) diminuiscano drasticamente a partire dal 2009, fino a raggiungere un minimo di AQI 2 nel 2015. Inquinanti come CO, NO2 e O3 presentano una stagionalità costante: infatti, il trend dell'O3 riporta un aumento delle emissioni durante i periodi estivi, mentre i trend degli altri due inquinanti riportano aumenti durante i mesi invernali. Livelli preoccupanti a livello nazionale, riguardano l'O3 (Ozono) il quale è l'unico agente per cui si è registrato un aumento negli ultimi anni. Notiamo che nel territorio americano gli inquinanti emessi in maggiore quantità siano O3 e NO2, i cui valori variano in un range da 25 a 40 AQI.

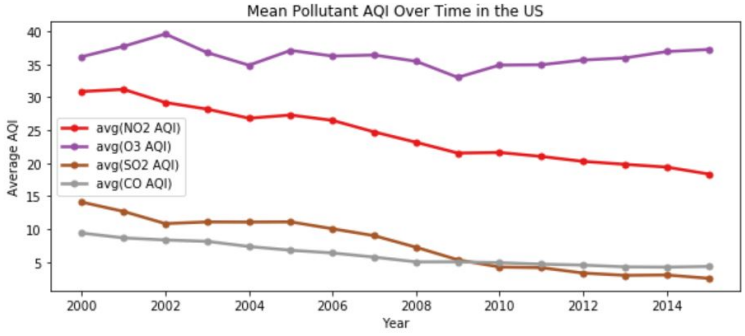


Figura 17: Valori Medi degli inquinanti negli USA dal 2000 al 2015.

4.3 Analisi dell'Inquinamento nello Stato della California

Dopo aver contato il numero di osservazioni per Stato, si ha che la maggior parte dei records appartengono allo stato della **California**, come conferma il bar plot in **Figura 18**. In California sono stati registrate circa 600k misurazioni, mentre negli altri Stati il numero di records si aggira intorno alle 50k.

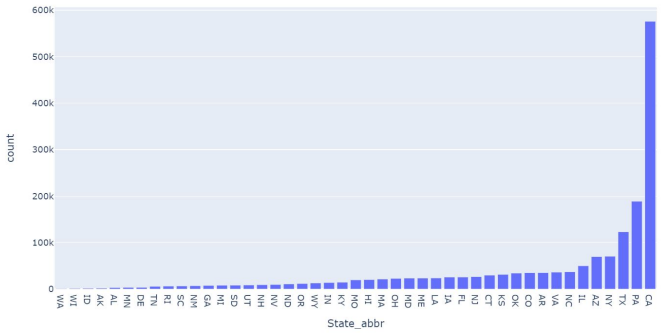


Figura 18: Numero di osservazioni per Stato dal 2000 al 2016.

Per questo motivo, la successiva analisi si focalizzata sulle caratteristiche dello stato della **California**, in modo tale da estrarre maggiori informazioni circa le regioni per le quali l'EPA effettua più misurazioni. Osservando i trend si nota una notevole diminuzione delle emissioni di SO2 ed un calo moderato di CO e NO2. Tuttavia, la quantità di O3 rilasciata nell'atmosfera da questo stato presenta una leggera crescita negli anni che vanno dal 2014 al 2016. Tali trend risultato essere completamente in linea con i trend generali a livello nazionale.

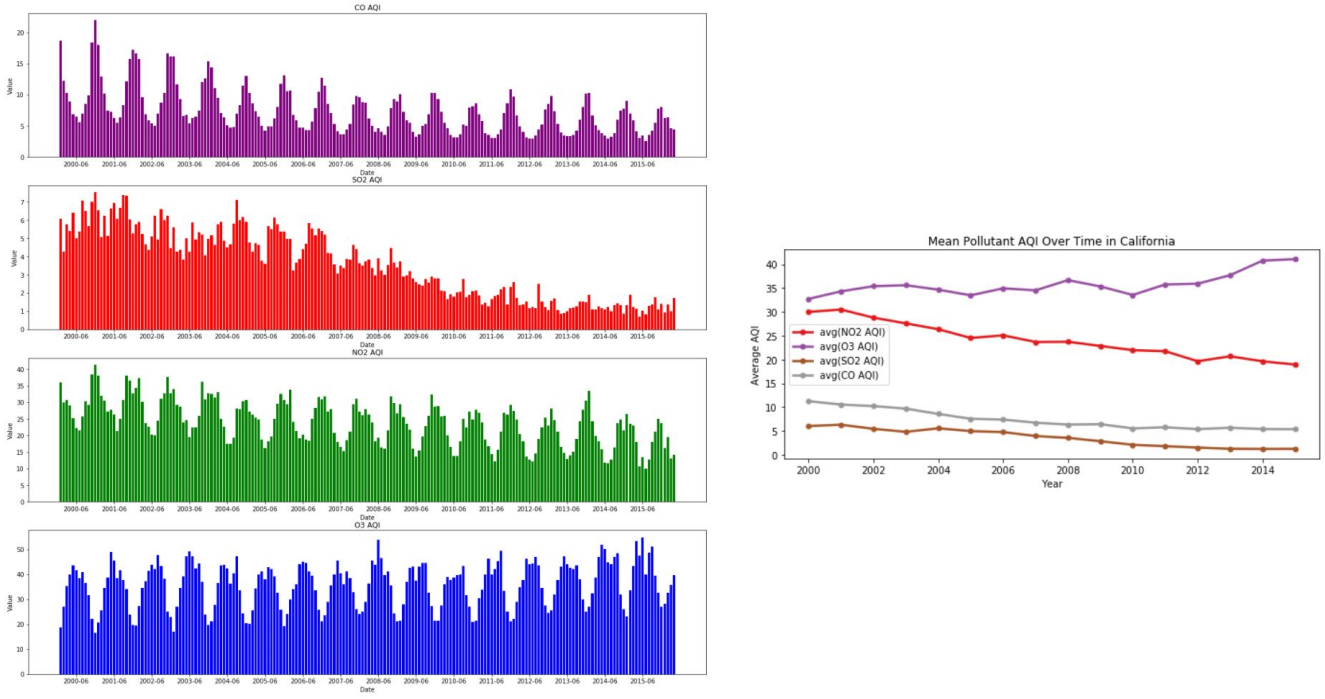


Figura 19: Trend degli inquinanti dal 2000 al 2016 nello Stato della California.

4.3.1 Analisi delle due città Californiane più inquinanti.

Guardando la **Figura 20**, si evidenzia come le città di Fontana e Rubidoux mostrino livelli di O3 molto alti (linea gialla), i quali raggiungono un picco di circa 120 AQI durante i periodi estivi. Invece, i livelli di SO2 e CO (linea viola e nera) risultano ottimali in quanto non superano in media i 10 AQI. Nella fase successiva,

i dati relativi alle due città menzionate sono stati confrontati con le due città più popolose della California: San Diego e Los Angeles. Si evince che suddette città contribuiscono in parte (sebbene in misura ridotta rispetto alle città descritte precedentemente) all'inquinamento di O3 e NO2. Los Angeles mostra un calo delle emissioni di NO2 a partire dal 2004, mentre i livelli di O3 sembrano in leggera crescita negli ultimi anni.

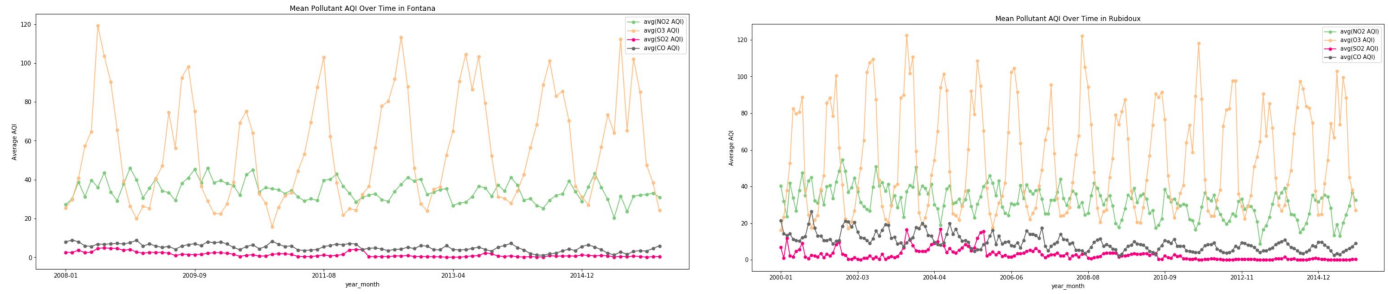
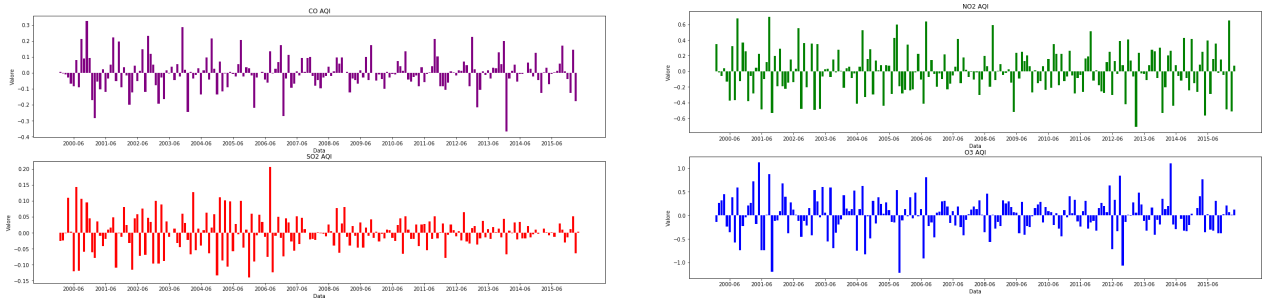


Figura 20: A sinistra i livelli di inquinamento di Fontana, mentre a destra quelli di Rubidoux.

4.3.2 Analisi delle Contee Californiane con più rilevazioni

Per comprendere meglio eventuali similitudini nell'evoluzione del tasso di inquinamento in America tra varie zone geografiche, è stata intrapresa un'analisi che si concentra sulle variazioni giornaliere degli inquinanti intercorse tra una rilevazione e l'altra. Il livello di aggregazione, scelto a livello geografico, riguarda le contee Californiane, nello specifico **San Diego, Los Angeles, Santa Barbara e Contra Costa**, le quali presentano il numero più elevato di osservazioni. Inoltre, si è così potuto evitare eventuali problematiche nelle comparazioni dovute a possibili shoutdown dei singoli siti di rilevazione.
La costruzione del Dataset dedicato al task ha visto come primo step l'eliminazione dei valori duplicati al suo interno, con riferimento alle colonne **Site Num, Year, County, Month, Day, CO AQI** [1]. Sapendo che ad ogni contea corrisponde uno o più sensori, si è costruita una media fra le osservazioni derivanti da più siti, ottenendo un valore unico giornaliero per ognuna e per ogni inquinante. Successivamente sono state selezionate solo le osservazioni in riferimento alla California. Avendo i dati organizzati correttamente, si è poi andati a costruire quattro nuove colonne, con al loro interno le variazioni giornaliere (una per ogni AQI). È stato deciso, per convenzione, che la variazione al 01/01/2000 fosse 0, in quanto non è stato possibile ricavare i dati del giorno 31/12/1999. A tal punto sono state plottate le variazioni dei vari inquinanti (raggruppate in una media mensile) al fine di darne una rappresentazione di facile interpretazione. Nella **Figura 21** è riportato come esempio il caso della contea di Santa Barbara.



(a) CO/SO2 (b) NO2/O3

Figura 21: Variazioni dei vari inquinanti nella contea di Santa Barbara

[1] Year, month e day sono state ottenute dalla disgregazione di Date Local. La scelta di CO AQI, inoltre, è totalmente casuale e sarebbe stato indifferente scegliere uno degli altri inquinanti. Difatti, per come è composto il Dataset, non sussistono variazioni nelle rilevazioni che appartengono ad un medesimo giorno.

Oltre a tale vista, è stata anche effettuata una comparazione tra le varie contee, alla ricerca di qualche similarità nell'evoluzione delle variazioni tra di esse. In aggiunta, è stata incorporata la rappresentazione grafica delle oscillazioni dell'intera California. La **Figura 22** mostra la comparazione delle fluttuazioni medie mensili di O3.

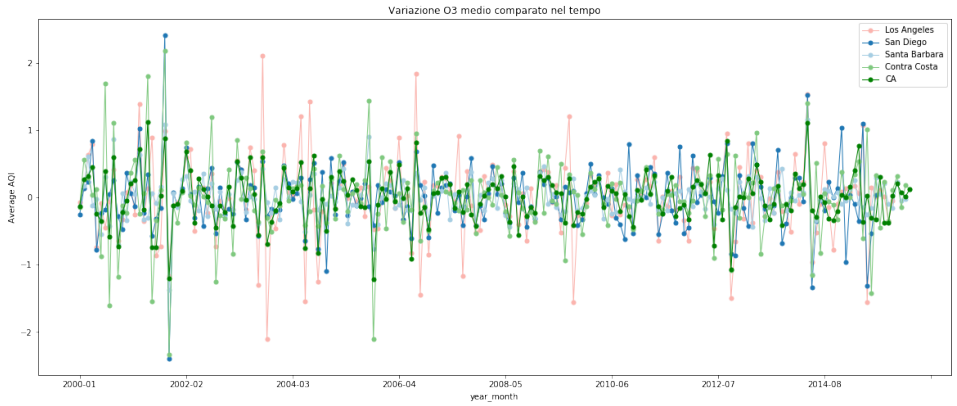


Figura 22: Fluttuazioni mensili di O3 in diverse contee Californiane, confrontate con quelle globali dello stato

Trovare una somiglianza osservando solo le variazioni mensili risultava però essere complicato. A tal proposito, per rendere l'analisi più esaustiva, sono state create altre quattro variabili. Esse racchiudono al loro interno la somma delle oscillazioni intercorse nei vari anni, con l'intento di andare a ricreare il pattern percorso dai vari inquinanti. Viene mostrato, in maniera sequenziale, quelli che sono i risultati delle comparazioni (raggruppamento mensile). Nel primo confronto (O3), pur avendo tendenze molto simili fino al 2012, San Diego e Santa Barbara presentano maggiori somiglianze, con incrementi dei valori più contenuti rispetto agli anni 2000.

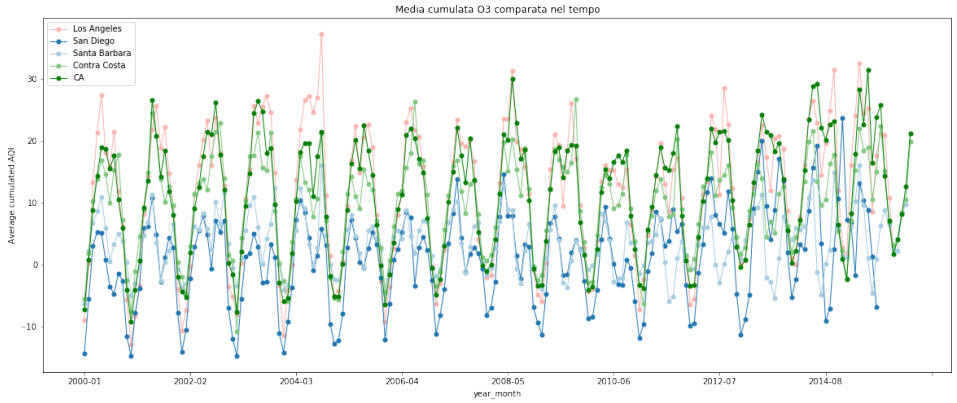


Figura 23: Pattern dell'inquinante O3 in varie contee della California

Per l'SO2 si ha invece una situazione molto diversa fra quasi tutte le contee, con una tendenza di riduzione dell'inquinamento ovunque, eccetto per Santa Barbara.

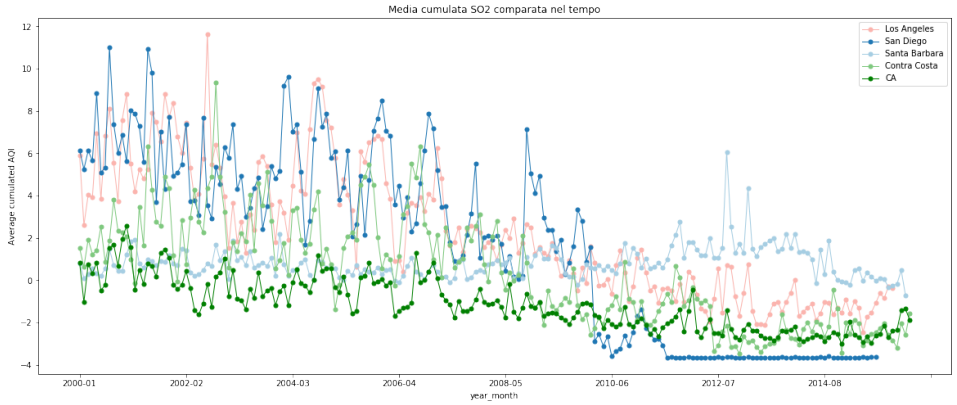


Figura 24: Pattern dell'inquinante SO2 in varie contee della California

L'evoluzione nelle rilevazioni dell'NO2 evidenzia nuovamente una tendenza anomala per Santa Barbara, la quale non sembra risentire della stagionalità caratteristica degli inquinanti presi in esame. San Diego, a partire dal 2014, inverte tale stagionalità rispetto alle altre contee ed alla California in generale.

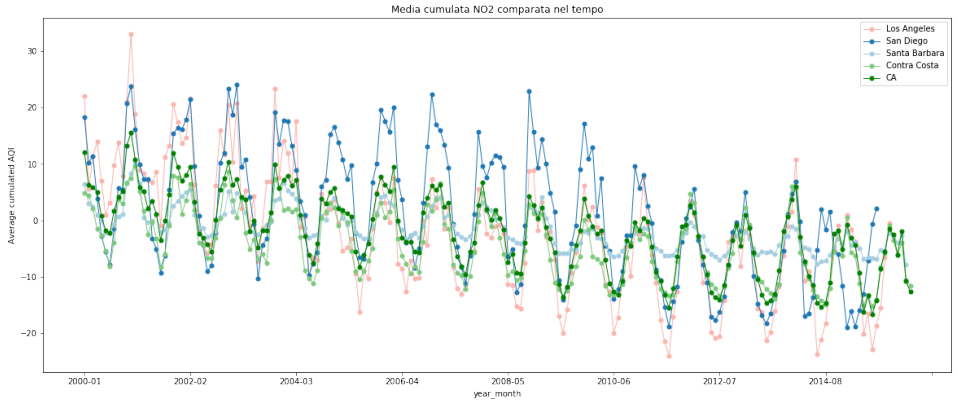


Figura 25: Pattern dell'inquinante NO2 in varie contee della California

Infine, per la CO, si evidenzia una tendenza marcatamente positiva a Los Angeles, che si distacca nettamente dalle altre contee e riduce drasticamente la sua presenza nell'aria. Santa Barbara presenta il solito andamento atipico, mentre San Diego è caratterizzata da un'inversione della stagionalità già riscontrata nel confronto precedente. Tali riscontri evidenziano ancora una volta la tendenza al ribasso dell'inquinamento in California, con Contra Costa che, tra le contee analizzate, sembra meglio rappresentarne l'evoluzione.

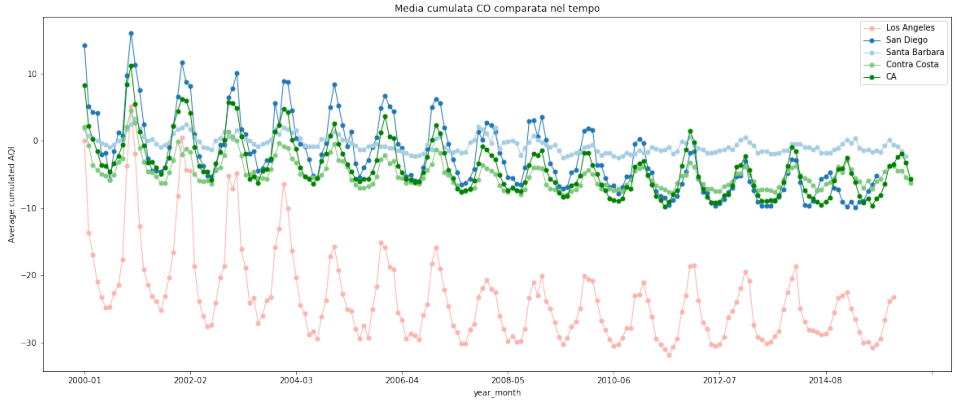


Figura 26: Pattern dell'inquinante CO in varie contee della California

5 Clustering dei Sensori delle Contee Californiane.

Uno dei task eseguiti sul DataSet riguarda il **Clustering** dei sensori, al fine di identificare gruppi di sensori caratterizzati da comportamenti simili durante tutte le loro misurazioni. Poichè i valori della variabile *Site Num* corrispondono al numero identificativo di sensori all'interno di una singola Contea, nel nostro DataSet potremmo avere Contee diverse con lo stesso numero identificativo dei sensori installati. Per risolvere questo problema, abbiamo effetuato un pre-processing andando a generare una nuova chiave identificativa dei sensori (chiamata *C-Sensor*) la quale corrisponde alla concatenazione degli attributi *County* e *Site Num*. Successivamente, considerando il vasto numero di sensori presente su tutto il territorio americano e considerando che il maggior numero di rilevazioni sono state eseguite nello Stato della California, abbiamo deciso di eseguire un clustering sui sensori delle Contee dello stato Californiano.

Poichè le singole rilevazioni sono organizzate in **Time Series**, è stata necessaria un'ulteriore fase di pre-processing, al fine di estrarre da tali Time Series delle features significative e rappresentative dei comportamenti generali dei vari sensori (le quali saranno descritte più nel dettaglio nel capitolo seguente).

5.1 Pre-processing e preparazione del Dataset

La prima trasformazione eseguita è stata l'estrazione delle varie misurazioni per ogni anno.
Avendo a disposizione tale informazione, si è deciso di raggrupparle per **Quadriennio** (2000-2003, 2004-2007, 2008-2011, 2012-2015), sia per ridurre il numero di Clustering da eseguire (computazionalmente molto costoso anche in ambiente distribuito) che per avere un'idea più generica delle variazioni dei fenomeni registrati (essendo l'inquinamento un fenomeno in trasformazione su lunghi periodi di tempo).
Dopo aver quindi generato una feature indicante il quadriennio di appartenenza di ogni Time Series, queste sono state raggruppate per C-sensor e quadriennio, calcolando valori massimi, minimi e medi di ogni inquinante.

5.2 Estrazione delle features tramite Regressione Lineare

Le informazioni estratte dalla prima fase di elaborazione dati, tuttavia, non sono sufficienti per condurre un clustering accurato dei sensori, in quanto avere solo un riferimento annuale sui valori dei vari inquinanti non garantisce una precisa informazione sul possibile trend che questi hanno nel corso del tempo. Per fornire all'algoritmo di Clustering delle features aggiuntive da considerare, abbiamo estratto i **coefficienti di regressione** delle singole Time Series. Per semplificare le Time Series senza però perdere informazioni riguardanti il loro andamento, abbiamo applicato una media mensile in modo da ridurre i time stamps da passare al regressore. I coefficienti di regressione per ogni inquinante per ogni quadriennio, che sono stati aggiunti ai relativi Dataset, sono rispettivamente: *'min(NO2 1st Max Value)', 'min(O3 1st Max Value)', 'min(SO2 1st Max Value)', 'min(CO 1st Max Value)', 'max(NO2 1st Max Value)', 'max(O3 1st Max Value)', 'max(SO2 1st Max Value)', 'max(CO 1st Max Value)', 'avg(NO2 Mean)', 'avg(O3 Mean)', 'avg(SO2 Mean)', 'avg(CO Mean)', 'x1_NO2' , 'x1_O3' , 'x1_SO2' , 'x1_CO'.*

C-Sensor	x1_NO2_0003	x1_SO2_0003	x1_CO_0003	x1_O3_0003
2007-San Diego	1.5646759892268332	2.17599066268434	17.596747750163537	1008.7720242679804
3001-Contra Costa	1.4800099593172749	4.7339853316308655	19.65749046023339	1091.5002600581045
306-San Bernardino	1.520069508396114	2.2220310303230204	16.610283589551138	1881.3177037440066
1-San Diego	1.3307549645303938	1.1102040904118304	16.692736951106763	922.8422448219388
5001-Los Angeles	0.9956185514120106	2.348068515334792	20.013350730719445	1038.7778720214028

only showing top 5 rows

Figura 27: Coefficienti di Regressione calcolati per ogni sensore

5.3 Esecuzione Clustering ed Analisi dei Risultati

Le misurazioni sono state dunque divise in 4 DataSet distinti, uno per quadriennio. Per ognuno di essi, i dati sono stati accorpati in 2 featues principali, una indicante il **C-Sensor (usato come chiave)** e l'altra contenente un **vettore di features** estratte nelle fasi precedenti, necessario per l'esecuzione dell'algoritmo di Clustering della libreria MLLib. Per ogni DataSet, è stato innanzitutto calcolato e plottato il valore della

Silhouette per un numero variabile di Cluster possibile, da 2 a 15. Il valore ottimale del parametro K è stato individuato selezionando il numero di cluster che massimizza la Silhouette. Si è notato come per tutti i quadrienni, il valore ottimale sia 2 (come mostrato nella **Figura 28**).

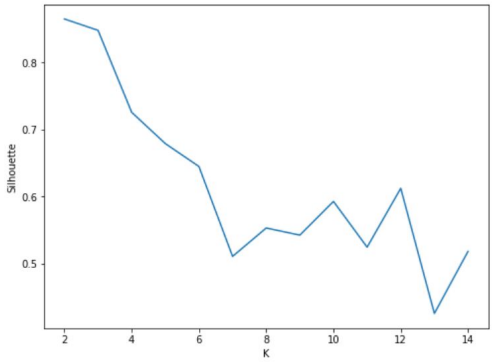
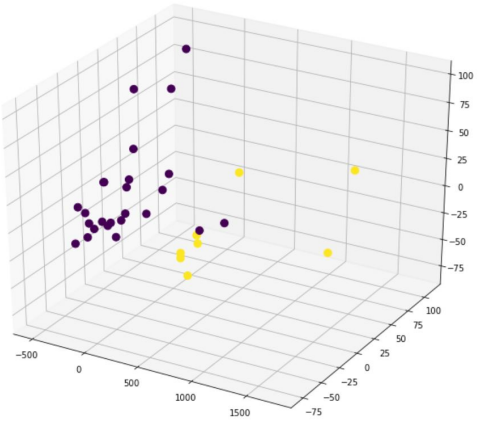
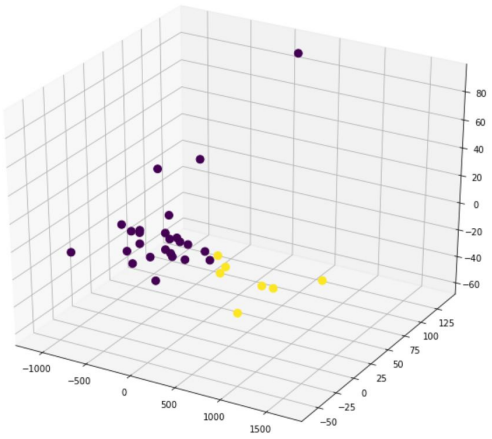


Figura 28: Plot Silhouette al variare del parametro k per il quadriennio 2000-2003. Valore ottimale k = 2 con Silhouette = 0.8653.

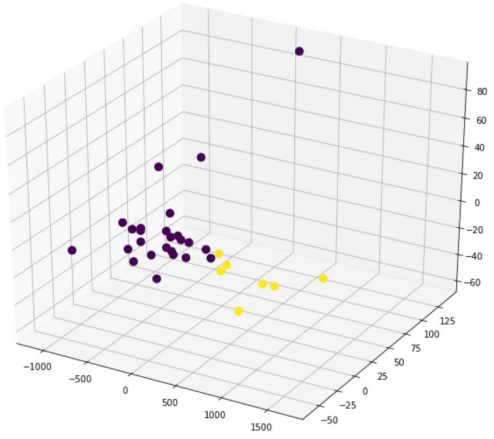
Per ogni quadriennio è stato dunque performato l'algoritmo di clustering **KMeans**. I risultati ottenuti, sono stati plottati in un grafico tridimensionale sfruttando la **PCA** dei dataset. Successivamente, è stata condotta un'analisi più approfondita dei risultati analizzando i *C-Sensor* appartenenti ai vari clusters e identificandone pattern all'interno delle suddivisioni ottenute.



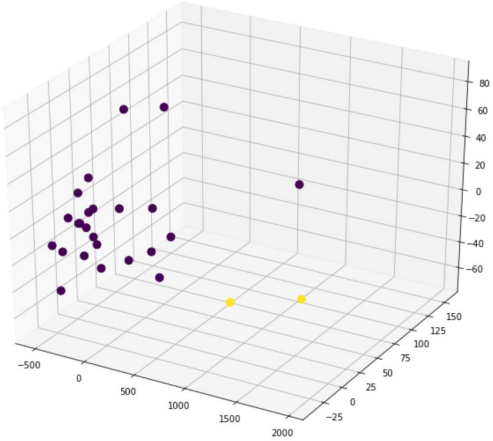
(a) Clusters 2000-2003



(b) Clusters 2004-2007



(c) Clusters 2008-2011



(d) Clusters 2012-2015

È importante sottolineare la composizione iniziale dei cluster dei sensori per i quali si hanno informazioni per tutti i quadrienni:

Cluster 0: [*'1002-Contra Costa', '1003-Orange', '1004-Contra Costa', '1025-Santa Barbara', '2-Contra Costa', '2004-Santa Barbara', '4-Solano', '4002-Los Angeles', '4003-Santa Barbara', '5-Imperial'*].

Cluster 1: [*'1002-Los Angeles', '1103-Los Angeles', '306-San Bernardino', '6-Sacramento', '8001-Riverside'*].

Studiando le caratteristiche di ogni cluster ottenuto, abbiamo notato peculiarità che li accomunano nei vari quadrienni:

- **Cluster 0**: Alti i valori di *avg(SO2)* e *min(NO2)*, Bassi valori di *max(NO2 1st max value)*
- **Cluster 1**: Alti valori di *max(NO2 1st max value)*, *avg(O3)*, *avg(NO2)* e *min(CO)*

Dopo l'esecuzione degli algoritmi di clustering, è stato creato uno spark dataframe contenente, per ogni *C-Sensor*, il numero del cluster di appartenenza. In questo modo, eseguendo delle query sql è stato possibile individuare i *C-Sensor* che non cambiano mai cluster, e quelli che cambiano cluster a partire da certi quadrienni. Di seguito (figura 30) si riportano le liste dei *c-sensor* per cui si hanno rilevazioni in tutti i quadrienni.

C-Sensor
5-Imperial
1025-Santa Barbara
1003-Orange
4-Solano
2004-Santa Barbara
1002-Contra Costa
4003-Santa Barbara
1004-Contra Costa
2-Contra Costa
8001-Riverside
4002-Los Angeles

(a) Sensori che non cambiano mai cluster.

C-Sensor
6-Sacramento
1103-Los Angeles
306-San Bernardino
1002-Los Angeles

(b) Sensori che cambiano cluster a partire dal quadriennio 2008-2011.

Figura 30: Risultati query sql su spark dataframe dei clusters.

Possiamo notare in **Figura 31(b)** che i sensori quali *"6-Sacramento", "1103-Los Angeles", "306-San Bernardino"* e *"1002-Los Angeles"* appartengono inizialmente al cluster 1, mentre dal 2008 in poi risultano comportarsi più similmente ai sensori appartenenti al cluster 0. Il cambiamento di cluster di alcuni C-sensor potrebbe essere in linea con i risultati osservati dalla precedente analisi dei trend, proprio per le caratteristiche tipiche dei due insiemi. Infatti, i sensori inizialmente appartenenti al cluster 1 (caratterizzato da valori medi degli inquinanti più alti), tendono a spostarsi nel cluster 0.

Poichè la Contea di Los Angeles ha il maggior numero di sensori installati, si è ritenuto interessante osservare il comportamento degli stessi durante le varie iterazioni (i risultati sono visibili nella **Figura 32**). Abbiamo notato che inizialmente tutti i sensori, ad eccezione di *1002-Los Angeles* e *1103-Los Angeles* appartengono al cluster 0.
E' importante sottolineare che il valore **null** indica che il sensore non era attivo durante quel quadriennio.

C-Sensor	cluster 2000-2003	cluster 2004-2007	cluster 2008-2011	cluster 2012-2015
5001-Los Angeles	0	0	null	null
4002-Los Angeles	0	0	0	0
30-Los Angeles	0	null	null	null
1002-Los Angeles	1	1	0	0
1103-Los Angeles	1	1	0	0
31-Los Angeles	0	null	null	null
4006-Los Angeles	null	null	0	0

Figura 31: Cluster labels attribuite ai sensori della Contea di Los Angeles.

6 Classificatori per predire la qualità dell'aria

6.1 Random Forest

Prima di eseguire il **task di classificazione** è stato effettuato un pre-processing del DataSet, andando a selezionare i seguenti attributi: : *'NO2 Mean','NO2 1st Max Value','NO2 1st Max Hour','NO2 AQI','O3 Mean','O3 1st Max Value', 'O3 1st Max Hour','O3 AQI','SO2 Mean','SO2 1st Max Value','SO2 1st Max Hour','SO2 AQI','CO Mean', 'CO 1st Max Value','CO 1st Max Hour','CO AQI'*.
Dopodichè, è stato creato un nuovo attributo, *AQI* in modo da avere un livello generale di inquinamento per ogni giorno. L'*AQI generale* (come riportato dall'EPA) corrisponde al valore massimo tra i singoli AQI dei vari inquinanti. Per calcolare questa label, abbiamo utilizzato la funzione spark ".greatest" al set di attributi *"NO2 AQI", "O3 AQI", "SO2 AQI", "CO AQI"*.
Il risultante AQI (avente un range da 0 a 500) è stato convertito in una label avente 5 possibili valori numerici (0, 1, 2, 3, 4) in base alla qualità dell'aria (dalla più ottimale alla più pericolosa). Utilizzando un VectorAssembler, abbiamo generato un vettore comprendente le features che verranno utilizzate dal modello in fase di training e test. La composizione per classe del DataSet è la seguente:

Classe	Numero di Records
0	1472380
1	229348
2	35534
3	4234
4	4766

Tabella 1: Numero di records per classe.

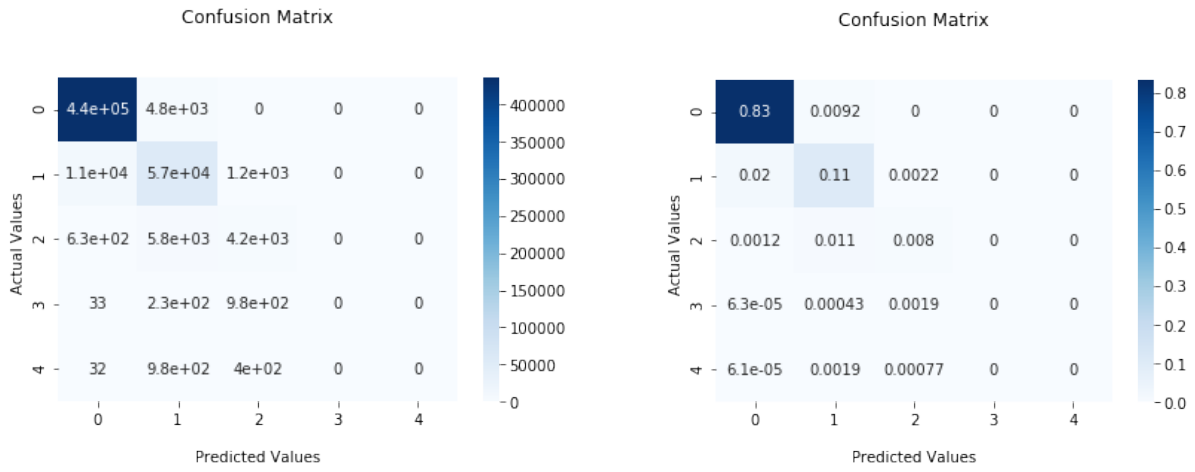
Il dataset è stato segmentato in training set (70%) e test set (30%) utilizzando un sistema di partizione hold-out. Il training set è composto da 1,222,348 records, mentre il test set da 523,914. Eseguendo il fit sul training set e un predict sul test set, abbiamo osservato ed analizzato le performance in termini di accuracy, precision e recall. Poichè il dataset è fortemente sbilanciato, si notano delle performance per nulla soddisfacenti per le classi 3 e 4.

Classe	Precision (%)	Recall (%)
0	97.45	98.91
1	82.74	82.69
2	62.15	39.24
3	0.0	0.0
4	0.0	0.0

Tabella 2: Report di Classificazione Random-Forest. Dataset Sbilanciato

Dal report di classificazione (**Figura 33**) si nota come, sebbene l'**accuracy** complessiva sia pari al 94.64% con un errore pari al 5.36%, la recall e la precision delle classi 3 e 4, sono pari a 0. Questo vuol dire

che il classificatore non ha mai predetto correttamente osservazioni appartenenti alla classe 3 e 4. La recall e precision più alta (rispettivamente 97.45% e 98.91%) si ottengono sulla classe 0 (la quale rappresenta la classe maggioritaria). Di seguito riportiamo i confusion matrix espressi in termini assoluti e percentuali.



(a) In valori assoluti

(b) In percentuale

Figura 32: A sinistra il confusion matrix della Random Forest in valori assoluti mentre a destra in valori percentuali.

6.1.1 Imbalanced Learing: Tuning del modello

Poichè i risultati ottenuti con la Random Forest non sono stati ritenuti soddifacenti per alcune classi, in questo capitolo si discuterà dei risultati ottenuti allenando il modello di classificazione sul training set bilanciato. A questo fine sono state eseguite tecniche di **imbalanced learning**. In primo luogo, è utilizzato un algoritmo di **oversampling**, il quale però, considerando il nostro framework di lavoro, non si è rivelato adatto. Si è deciso quindi di applicare un **undersampling** sulle classi maggioritarie in modo tale da ri-equilibrare lo sbilanciamento presente, consentendo al modello di riuscire a catturare patterns caratterizzanti le classi 3 e 4 in modo migliore rispetto al modello proposto nella Sezione precedente. Per eseguire l'undersampling, abbiamo utilizzato la funzione ".sample" fornita da pyspark sul Training Set. La distribuzione delle classi si può leggere nella **Tabella 3**.

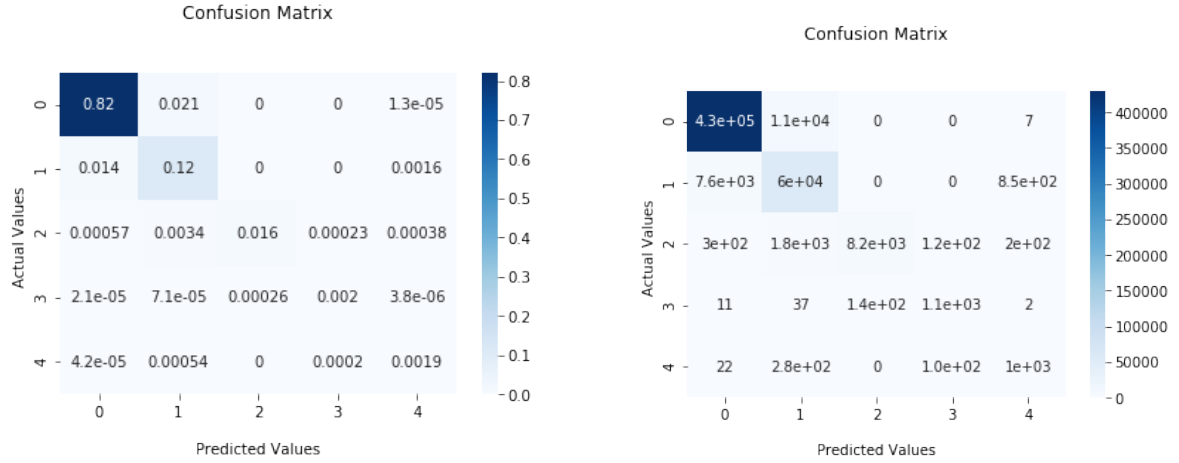
Classe	Numero di Records
0	12093
1	6784
2	4100
3	2989
4	3352

Tabella 3: Numero di records per classe nel training set con Undersampling.

Il training sul dataset bilanciato, ha mostrato significativi miglioramenti sulla capacità predittiva del modello. Infatti, osservando il report di classificazione in **Tabella 4**, si osserva un miglioramento delle prestazioni relative alla classe 2. La precision passa da 62.15% all'98.35%, mentre la recall subisce un incremento del circa 38.39%. La predizione delle classi 3 e 4, inoltre, subisce un netto miglioramento per la precision, la quale passa rispettivamente a 82.51% e 48.57%. Anche per la recall è stato ottenuto un miglioramento: sono stati ottenuti valori pari al 84.89% per la classe 3 e all' 70.93% per la classe 4. Infine, l'**accuracy** complessiva aumenta del 1.1% (passando da 94.64% a 95.74%). In conclusione, il modello ottenuto con questa tecnica di imbalanced learning, risulta nettamente migliore rispetto a quello proposto nella Sezione 6.1.

Classe	Precision (%)	Recall (%)
0	98.19	97.45
1	81.90	87.73
2	98.35	77.63
3	82.51	84.89
4	48.57	70.93

Tabella 4: Report di Classificatione Random-Forest - Dataset con Undersampling.



(a) In valori assoluti

(b) In percentuale

Figura 33: A sinistra il confusion matrix della Random Forest con Undersampling in valori assoluti mentre a destra in valori percentuali.

6.2 Logistic Regression

Per utilizzare una regressione logistica, è stata eseguita una classificazione binaria delle osservazioni in: **hazardous(label 1) e non-hazardous(label 0)**. Le osservazioni che non rappresentano un rischio per la salute sono quelle aventi Classe 0 (in riferimento alle label descritte nella sezione del Random Forest), mentre quelle che potrebbero rappresentare un serio rischio sono quelle appartenenti alle Classi 3 e 4, i cui record sono stati raggruppati in un unica Classe. Per quanto riguarda le osservazioni appartenenti alle classi 1 e 2 nel Random forest, esse non state incluse nella seguente analisi in quanto border-line tra "hazardous" e "non-hazardous".
Il Dataset si ripresenta sbilanciato con un rapporto percentuale pari a 99.40% (1472380) per la classe non-hazardous e 0.60% (9000) per la classe hazardous, ed è stato successivamente partizionato in training set (70%) e test set(30%). Considerando il forte sbilanciamento verso la classe 0 (non-hazardous) il modello non risulta essere idoneo all'individuazione e corretta classificazione dei records appartenenti alla classe 1 (hazardous). Infatti, sebbene l'accuracy complessiva è pari al 99.38%, la recall e la precision risultano essere pari a 0.0% (quindi questa classe non viene mai predetta dal modello).

6.2.1 Tuning del modello

Per migliorare le prestazioni sulla Classe hazardous (label 1), anche in questo caso, è stata applicata una tecnica di undersampling, in modo da riequilibrare il numero dei record presenti nel training set.

Classe	Numero di Records
0 (non-hazardous)	10204
1 (hazardous)	6249

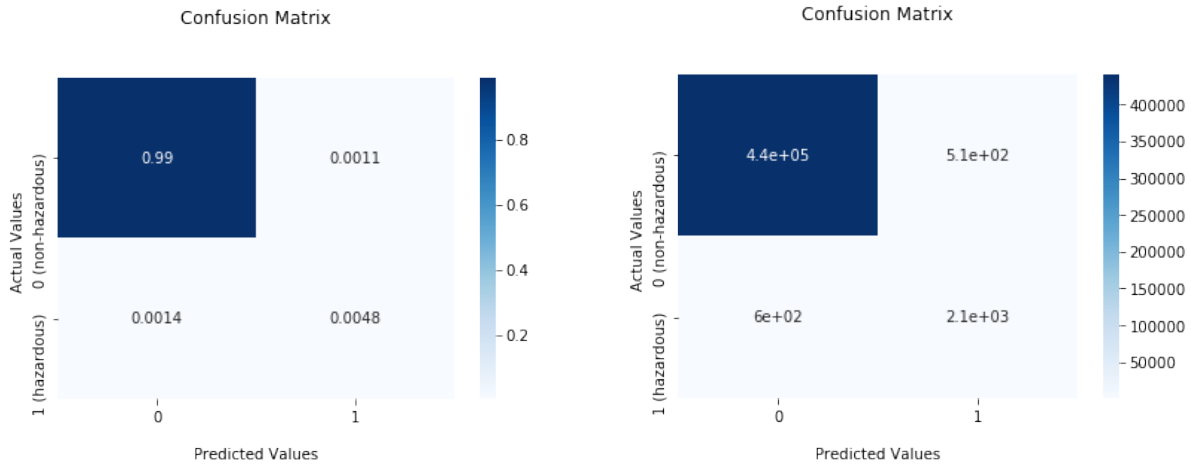
Tabella 5: Numero di records per classe dopo l'Undersampling nel training set.

Per migliorare ulteriormente le performance è stata anche istanziata una grid search con cross validation a 5 fold, applicata al training set bilanciato, in modo tale da poter ricercare la configurazione di attributi che massimizzasse le prestazioni del modello. Sono stati testati i seguenti valori: **regParam** = {0.0, 0.01, 0.5, 1, 2}, **elasticNetParam** = {0.0, 0.5, 1}, **maxIter** = {1, 5, 10}. La configurazione ottimale è risultata essere: **regParam** = 0.5, **elasticNetParam** = 0.0 e **MaxIter** = 10. Con questo settaggio di parametri e con l'undersampling applicato dal DataSet, l'accuracy complessiva aumenta dello 0.37% (accuracy pari al 99.75%). L'AUC score invece migliora raggiungendo un picco del 99.44%. Di seguito si riportano i risultati in termini di precision e recall delle due classi.

Classe	Precision (%)	Recall (%)
0 (non-hazardous)	99.86	99.88
1 (hazardous)	80.94	78.11

Tabella 6: Report di Classificatione Logistic Regression - Dataset con Undersampling.

Notiamo un netto miglioramento nelle prestazioni sulla classe 1. Precision e recall passano dallo 0.0% all' 80.94% e 78.11%. La recall indica che quasi l'78.11% dei record appartenenti alla classe 1 sono stati correttamente classificati. Nella **Figura 34** , mostriamo i confusion matrix espressi in termini percentuali e valori assoluti.



(a) In valori assoluti

(b) In percentuale

Figura 34: Confusion Matrix Logistic Regression. A sinistra in termini di composizione percentuale mentre sulla destra in valori assoluti.

6.3 SVM Lineare

Il task di predizione proposto, risulta essere lo stesso descritto nella Sezione 6.2, dunque la classificazione dei record in "hazardous" (classe 1) e "non-hazardous" (classe 0). A causa del dataset sbilanciato, anche applicando il modello SVM Lineare, i risultati ottenuti si sono rivelati simili a quelli restituiti dalla Logistic Regression. Per cui, si è deciso di applicare nuovamente l'Undersampling, procedendo dunque all'allenamento di un nuovo modello di classificazione. Il training set ribilanciato è così suddiviso:

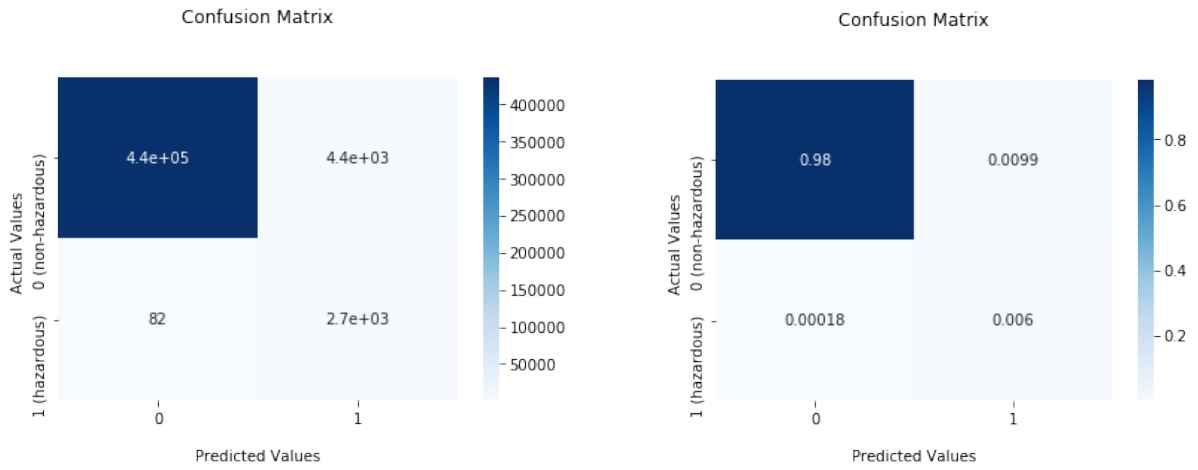
Classe	Numero di Records
0 (non-hazardous)	10204
1 (hazardous)	6249

La configurazione ottimale dei parametri del modello, è stata ricavata eseguendo un tuning mediante una Grid Search con cross validation a 5 fold. I parametri testati sono: **regParam** = {0.0, 0.1, 0.2, 0.5, 1}, **MaxIter**= {10, 20}. La configurazione ottimale risulta essere: **regParam** = 0.1, **MaxIter**= 10. Dopo aver allenato il modello eseguento un fit sul training set, abbiamo predetto l'output sul test set. Nella **Tabella**

8, mostriamo i risultato ottenuti al modello. L'accuracy risulta essere pari al 91.68% (con un AUC score del 96.96%). Notiamo come la recall di entrambe le classi sia molto alta (rispettivamente 99.0 % e 97.01 %) il che fornisce informazioni circa l'abilità del modello di predire correttamente quasi tutti i record. Si noti come la precision della classe 1 è pari al 37.65%, a causa di un alto numero di Falsi Positivi (come si può osservare dalla **Tabella 7**).

Classe	**Precision (%)**	**Recall (%)**
0 (non-hazardous)	99.98	99.0
1 (hazardous)	37.65	97.01

Tabella 7: Report di Classificazione SVM - DataSet con Undersampling.



(a) In valori assoluti

(b) In percentuale

Figura 35: Confusion Matrix - SVM.

A parità di task, confrontando i risultati ottenuti con quelli riportati dal modello di Logistic Regression, osserviamo alcune differenze. Il modello SVM presenta una recall per la classe 1 (hazardous) più alta rispetto alla Logistic Regression (rispettivamente 97.01% e 78.11%). D'altra parte, la Logistic Regression dimostra una precisione nettamente più alta in confronto all'SVM (80.94 % nella Logistic Regression e 37.65% nell'SVM). Pertanto, dato che il focus è di identificare correttamente quante più osservazioni caratterizzate da una scarsa qualità dell'aria, se si considera l'F1-Score il modello della Logistic Regression sembra più performante rispetto a quello della SVM.

6.4 Conclusioni

L'intero processo di analisi ha condotto a risultati più che soddisfacenti: è stato possibile analizzare i dati in vari livelli di granularità, portando a considerazioni sull'andamento degli inquinanti a livello nazionale, regionale e locale con relativi confronti a livello temporale e geografico degli indici di inquinamento.
Il task di clustering ha permesso di effettuare un confronto tra i comportamenti delle rilevazioni dei vari sensori, permettendoci di individuare zone geografiche con andamenti simili o in cambiamento nel tempo. Tale approccio di analisi è stato ulteriormente potenziato dall'impiego delle regressioni sulle varie Time Series, evitando problemi di generalizzazione statistica.
L'impiego dei classificatori ha permesso la realizzazione di un sistema automatico di analisi di rischio correlato ai livelli di inquinanti registrati, con risultati più che promettenti.
Le analisi sono state inoltre condotte in un ambiente distribuito, permettendo l'utilizzo di grandi quantità di dati ed ottenendo i risultati in tempi ragionevoli.