

**UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO**

**DIPARTIMENTO DI INFORMATICA**

**CORSO DI LAUREA MAGISTRALE**

**IN**

**DATA SCIENCE**

**Caso di studio di**

**Gestione di dati strutturati e non strutturati**

**#TITOLO**

**Docenti: Studente:**

**Mario Alessandro Bochicchio Marco Casamassima 730346**

**Corrado Loglisci**

**Anno Accademico 2022-2023**

Sommario

[Obiettivi 2](#_Toc134891004)

[Fonti dati e diagramma architettura 2](#_Toc134891005)

[ETL 2](#_Toc134891006)

[Modellazione 2](#_Toc134891007)

[Diagramma ER 2](#_Toc134891008)

[Diagramma Relazionale 2](#_Toc134891009)

[Stima della dimensione database 3](#_Toc134891010)

[Queries 3](#_Toc134891011)

[Presentazione dei risultati 3](#_Toc134891012)

[Matrice delle responsabilità 3](#_Toc134891013)

[Bibliografia-Sitografia 3](#_Toc134891014)

# Obiettivi

L'inquinamento atmosferico, causato dalle attività umane, rappresenta un fattore di rischio ampiamente riconosciuto per la salute umana e per gli ecosistemi. Negli Stati occidentali, la lotta contro l'inquinamento atmosferico ha una storia di oltre sessant'anni di studi e ricerche volti a comprendere i meccanismi degli effetti dannosi degli inquinanti, nonché a individuare strategie e tecniche di mitigazione. Le reti di monitoraggio costituiscono il principale strumento per la valutazione della qualità dell'aria, intesa come il complesso di attività mirate a verificare il rispetto dei valori limite e l'ottenimento degli obiettivi stabiliti al fine di prevenire, eliminare o ridurre gli effetti negativi dell'inquinamento atmosferico sulla salute umana e sull'ecosistema in un determinato territorio statale [1].

Secondo uno studio dell’Organizzazione Mondiale della Sanità (OMS) ogni anno muoiono circa 7 milioni di persone a causa dell’inquinamento ambientale [2] e circa 9 persone su 10 al mondo respirano aria con elevati livelli di inquinanti [3] che causano malattie respiratorie e polmonari, leucemia, problemi cardiovascolari, malattie cardiache, ictus, cancro ai polmoni, ecc.

Un indicatore che può essere utilizzato per valutare la qualità dell’aria è la quantità del particolato atmosferico, ossia una miscela complessa di particelle solide e liquide di sostanze organiche ed inorganiche sospese in aria come, ad esempio, alcuni metalli tossici (piombo, cadmio e nichel) che hanno un’elevata tossicità [4]. In generale quanto più piccola è la dimensione delle particelle tanto maggiore è la loro capacità di penetrare nei polmoni e dunque di produrre effetti dannosi sulla salute umana. Il particolato è suddiviso in base al diametro aerodinamico:

* Particelle PM10: con diametro aerodinamico inferiore a 10 µm, penetrano nel tratto superiore delle vie aeree (cavità nasali, faringe e laringe);
* Particelle PM2.5: con diametro aerodinamico inferiore a 2.5 µm, possono giungere fino alle parti inferiori dell’apparato respiratorio.

Le sorgenti di queste particelle possono essere di tipo naturale (erosione del suolo, vulcani, incendi boschivi, dispersione di pollini, etc.) o antropogenico (industrie, riscaldamento, traffico veicolare e processi di combustione in generale) [5].

Secondo l’OMS, per il particolato non è possibile definire un valore limite al di sotto del quale non si verificano nella popolazione effetti sulla salute: per questo motivo le concentrazioni di PM10 e PM2.5 nell'aria dovrebbero essere mantenute a livelli più bassi possibile [6]. Tuttavia, le linee guida dell'OMS sulla qualità dell'aria, aggiornate al 2019, stabiliscono delle soglie di concentrazioni di particolato atmosferico consigliate da raggiungere. In particolare:

* Per il PM2.5: le concentrazioni medie annuali non devono superare i 10 µg/m3, mentre l’esposizione media giornaliera non deve superare i 25 µg/m3 per più di 3 - 4 giorni all'anno.
* Per il PM10: le concentrazioni medie annuali non devono superare i 20 µg/m3, mentre l’esposizione media giornaliera non deve superare i 50 µg/m3 per più di 3 - 4 giorni all'anno.

Come precedentemente menzionato, il traffico veicolare rappresenta una delle fonti di particolato PM2.5 e PM10, il quale assume un carattere sempre più urgente e grave, soprattutto nelle grandi città dove l'uso predominante dell'automobile privata influisce sulla mobilità delle persone in modo eccessivo. Pertanto, risulta essenziale individuare soluzioni atte a ridurre il numero di veicoli circolanti nelle aree urbane, promuovendo parallelamente modalità di trasporto più sostenibili, al fine di mitigare le emissioni di gas serra e di sostanze dannose per l'ambiente. Un'opzione di particolare interesse potrebbe essere rappresentata dalla condivisione di veicoli (automobili elettriche, bici o monopattini), il quale si configura come un complemento al trasporto, contribuendo così alla diminuzione del numero di veicoli privati in circolazione e dei chilometri percorsi. Uno studio condotto dall'Università di Palermo, condotto da Marco Migliore et al. [7], sull'impatto ambientale del car sharing a Palermo, mette in luce come questa forma di condivisione possa costituire una soluzione per ridurre l'inquinamento ambientale nelle aree urbane.

Lo scopo del presente elaborato consiste nel creare un database contenente dati raccolti dalle stazioni di monitoraggio riguardanti le concentrazioni di particolato PM2.5 e PM10 nel territorio della regione Puglia, nonché il numero di decessi potenzialmente attribuibili all'inquinamento ambientale e un’analisi mirata a identificare le città urbane maggiormente colpite da tale problematica prendendo in considerazione lo studio effettuato da Alessia Longo e Eleonora Miccoli [8] dal titolo *Studio e analisi della correlazione tra inquinamento ambientale e mortalità nella popolazione italiana*.

I principali stakeholder individuati per questo progetto risultano essere i dirigenti delle imprese di condivisione di veicoli, i quali potranno avvalersi di tale database e dell'analisi corrispondente per determinare le aree pugliesi in cui operare.

# Fonti dati e diagramma architettura

Il database utilizzato in questo studio è stato creato a partire da dataset provenienti da diverse fonti di dati, tra cui:

* il numero di morti nelle diverse province pugliesi, estratto dal sito dell'Istituto Nazionale di Statistica (ISTAT);
* la popolazione delle varie città pugliesi, estratta dal sito ISTAT;
* i valori di PM10 e PM2.5 registrati dalle diverse stazioni di rilevamento dell'inquinamento sul territorio pugliese, provenienti dal sito dell’Agenzia Regionale per la Prevenzione e la Protezione dell’Ambiente (ARPA).

Immagine che contiene testo, diagramma, schermata, linea

Descrizione generata automaticamente

# ETL

In questo paragrafo si descrive la pipeline ETL (estrazione, trasformazione e caricamento) effettuata[[1]](#footnote-1):

## Estrazione

Come menzionato nel paragrafo precedente, i dati utilizzati sono stati estratti dai siti Arpa Puglia e Istat. In particolare:

* I dati relativi al numero di morti nelle diverse province pugliesi sono stati estratti dal seguente link <http://dati.istat.it/Index.aspx?DataSetCode=DCIS_CMORTE1_RES> in formato csv. Le query utilizzate per l’estrazione sono:
  + Territorio: sono state selezionate le caselle relative alle province pugliesi;
  + Tipo dato: è stata selezionata la casella “morti”;
  + Sesso: sono state selezionate le caselle “maschi” e “femmine”;
  + Causa Iniziale di morte – European Short List[[2]](#footnote-2): sono state selezionate le caselle “tumori”, “malattie endocrine, nutrizionali e metaboliche”, “malattie del sistema nervoso e organi di senso”, “malattie del sistema circolatorio”, “malattie del sistema respiratorio” e “malattie dell’apparato digerente”;
  + Seleziona periodo: sono stati selezionati gli anni 2019 e 2020;
  + Per le dimensioni Pivot:
    - Filtra dimensioni: è stato inserito “Tipo dato”;
    - Dimensione verticale: è stato inserito “Territorio”;
    - Dimensione orizzontale: sono stati inseriti “Causa iniziale di morte - European Short List”, “Seleziona periodo” e “Sesso”.

La query è stata salvata nel seguente link [http://dati.istat.it//Index.aspx?QueryId=60835](http://dati.istat.it/Index.aspx?QueryId=60835). Nella Figura 1 è riportato il risultato della query.

Immagine che contiene testo, schermata, numero, software

Descrizione generata automaticamente

Figura Risultato della query sul dataset delle morti

* I dati relativi alla popolazione risedente nei diversi comuni pugliesi sono stati estratti dal seguente link [http://dati.istat.it/Index.aspx?DataSetCode=DCIS\_POPRES1#](http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1) in formato csv. Le query utilizzate per l’estrazione sono:
  + Territorio: sono state selezionate le caselle relative ai comuni pugliesi;
  + Tipo di indicatore demografico: è stata selezionata la casella “popolazione inizio periodo”;
  + Sesso: sono state selezionate le caselle “maschi” e “femmine”;
  + Seleziona periodo: sono stati selezionati gli anni 2019 e 2020;
  + Per le dimensioni Pivot:
    - Filtra dimensioni: è stato inserito “Tipo di indicatore demografico”;
    - Dimensione verticale: sono stati inseriti “Territorio” e;
    - Dimensione orizzontale: sono stati inseriti “Seleziona periodo” e “Sesso”.

La query è stata salvata nel seguente link <http://dati.istat.it//Index.aspx?QueryId=60836>. Nella Figura 2 è riportato il risultato della query.

Immagine che contiene testo, schermata, numero, software

Descrizione generata automaticamente

Figura Risultato della query sul dataset della popolazione

* I dati relativi alle registrazioni effettuate dalle stazioni di rilevazione di PM2.5 e PM10 in Puglia sono stati estratti dal seguente link <http://old.arpa.puglia.it/web/guest/meta-aria> in formato csv. Dopo aver cliccato su “Dati per centralina”, è stata utilizzata la seguente query per l’estrazione:
  + Seleziona la provincia: è stato selezionato “Tutto”;
  + Seleziona l’inquinante: è stato selezionato “Tutto”;
  + Report Format: è stato selezionato “csv”;
  + Seleziona la data o un intervallo di date: sono stati selezionati i giorni dal 01-01-2019 al 31-12-2020.

## Trasformazione

La fase di trasformazione è stata eseguita con i seguenti file:

* Comuni.ipynb per i comuni;
* Decessi.ipynb per i decessi;
* Indicatori.ipynb per gli indicatori;
* Popolazioni.ipynb per le popolazioni;
* Province.ipynb per le province;
* Stazioni, Rilevazioni, Rileva.ipynb per le stazioni, rilevazioni e rileva.

Di seguito sono riportate le descrizioni dei file.

### Comuni

Importazione delle librerie pandas e geopandas per la gestione dei dataframe

import pandas as pd  
import geopandas as gpd

Configurazione dei path dei dataset regioni e comuni

path\_regioni = f'../dataset/dati\_spaziali/regioni/regioni.shp'  
path\_comuni = f'../dataset/dati\_spaziali/comuni/comuni.shp'

Lettura dei dataset

comuni = gpd.read\_file(path\_comuni)  
regioni = gpd.read\_file(path\_regioni)

Estrazione dei comuni pugliesi

codice\_puglia = regioni[regioni['DEN\_REG'] == 'Puglia']['COD\_REG'].values[0]  
comuni\_puglia = comuni[comuni['COD\_REG'] == codice\_puglia]  
comuni\_puglia.reset\_index(inplace=True, drop=True)

Rinominazione delle colonne del dataframe

comuni\_puglia = comuni\_puglia[['COD\_PROV', 'PRO\_COM', 'COMUNE', 'geometry']]  
comuni\_puglia.rename(columns={'COMUNE': 'nome',  
 'COD\_PROV': 'id\_prov',  
 'PRO\_COM': 'id\_comune'},   
 inplace=True)

Esportazione del dataset trasformato

comuni\_puglia.to\_file('../dataset/dati\_spaziali/comuni/comuni\_finale.shp', index=None)

### Decessi

Importazione delle librerie pandas e geopandas per la gestione dei dataframe

import pandas as pd  
import geopandas as gpd

Configurazione dei path dei dataset decessi e province

path\_decessi = '../dataset/decessi/decessi\_malattie\_2019\_2020.csv'  
path\_province = f'../dataset/dati\_spaziali/province/province.shp'

Lettura dei dataset

province = gpd.read\_file(path\_province)  
decessi = pd.read\_csv(path\_decessi)

Proiezione del dataframe decessi

columns = [‘Territorio’, ‘Sesso’, ‘Causa iniziale di morte – European Short List’, ‘TIME’, ‘Value’]  
decessi = decessi[columns]

Creazione del dataframe finale

decessi\_finale = pd.DataFrame(columns = ['nome\_malattia', 'numero\_decessi\_maschi', 'numero\_decessi\_femmine', 'anno', 'id\_provincia'])

Il dataset iniziale contiene gli attributi Sesso e Value. Di seguito si fondono queste due colonne in numero\_decessi\_maschi e numero\_decessi\_femmine.

for p in set(decessi['Territorio']):  
 decessi\_territorio = decessi[decessi['Territorio'] == p]  
 for c in set(decessi\_territorio['Causa iniziale di morte - European Short List']):  
 decessi\_territorio\_malattia = decessi\_territorio[decessi\_territorio['Causa iniziale di morte - European Short List'] == c]  
 anni = set(decessi['TIME'])  
 for a in anni:  
 decessi\_territorio\_malattia\_maschi = decessi\_territorio\_malattia[decessi\_territorio\_malattia['Sesso'] == 'maschi']  
 decessi\_territorio\_malattia\_maschi\_anno = decessi\_territorio\_malattia\_maschi[decessi\_territorio\_malattia\_maschi['TIME'] == a]['Value'].values[0]  
   
 decessi\_territorio\_malattia\_femmine = decessi\_territorio\_malattia[decessi\_territorio\_malattia['Sesso'] == 'femmine']  
 decessi\_territorio\_malattia\_femmine\_anno = decessi\_territorio\_malattia\_femmine[decessi\_territorio\_malattia\_femmine['TIME'] == a]['Value'].values[0]  
 id\_provincia = province[province['DEN\_UTS'] == p]['COD\_PROV'].values[0]  
   
 row = [c, decessi\_territorio\_malattia\_maschi\_anno, decessi\_territorio\_malattia\_femmine\_anno, a, id\_provincia]  
 decessi\_finale.loc[len(decessi\_finale)] = row

Esportazione del dataset trasformato

decessi\_finale.to\_csv('..\dataset\decessi\decessi\_finale.csv', index=False)

### Indicatori

Importazione delle librerie pandas per la gestione dei dataframe

import pandas as pd

Creazione del dataframe finale contenente i dati

indicatori = {  
 'sigla': ['PM2.5', 'PM10'],  
 'nome': ['Particolato PM2.5', 'Particolato PM10'],  
 'soglia\_media\_giornaliera': [10, 20],  
 'giorni\_superamento': [4, 4],  
 'soglia\_media\_annuale': [25, 50]  
}  
indicatori\_df = pd.DataFrame(indicatori)

Estrazione del dataset

indicatori\_df.to\_csv('..\dataset\indicatori\indicatori\_finale.csv', index=False)

### Popolazioni

Importazione delle librerie pandas per la gestione dei dataframe

import pandas as pd

Configurazione dei path dei dataset decessi e province

path\_popolazioni = f'../dataset/popolazioni/popolazioni.csv'

Lettura del dataset

popolazioni = pd.read\_csv(path\_popolazioni)

Proiezione del dataframe popolazioni

popolazioni = popolazioni[[‘ITTER107’, ‘Sesso’, ‘TIME’, ‘Value’]]

Creazione del dataframe con colonne 'id\_comune', 'anno', 'numero\_maschi', 'numero\_femmine'

popolazioni\_finale = pd.DataFrame(columns = ['id\_comune', 'anno', 'numero\_maschi', 'numero\_femmine'])

Il dataset iniziale contiene gli attributi Sesso e Value. Di seguito si fondono queste due colonne in numero\_maschi e numero\_femmine.

### Province

Importazione delle librerie pandas e geopandas per la gestione dei dataframe

import pandas as pd  
import geopandas as gpd

Configurazione dei path dei dataset regioni, province e comuni

path\_path\_regioni = f'../dataset/dati\_spaziali/regioni/regioni.shp'  
path\_province = f'../dataset/dati\_spaziali/province/province.shp'

Lettura dei dataset

regioni = gpd.read\_file(path\_regioni)  
province = gpd.read\_file(path\_province)

Estrazione delle province pugliesi

codice\_puglia = regioni[regioni['DEN\_REG'] == 'Puglia']['COD\_REG'].values[0]  
province\_puglia = province[province['COD\_REG'] == codice\_puglia]  
province\_puglia.reset\_index(inplace=True, drop=True)

Rinominazione delle colonne del dataframe

province\_puglia = province\_puglia[['DEN\_UTS', 'COD\_PROV', 'geometry']]  
province\_puglia.rename(columns={'DEN\_UTS': 'nome',   
 'COD\_PROV': 'id\_prov'},   
 inplace=True)

Esportazione del dataset trasformato

province\_puglia.to\_file('../dataset/dati\_spaziali/province/province\_finale.shp', index=None)

### Stazioni, rilevazioni, rileva

## Caricamento

# Modellazione

## Diagramma ER

## Diagramma Relazionale

## Stima della dimensione database

# Queries

# Presentazione dei risultati

# Matrice delle responsabilità

# Bibliografia

|  |  |
| --- | --- |
| [1] | S. n. p. l. p. dell'ambiente, «PARTE-INIZIALE-STATO-E-TREND.pdf,» 02 12 2020. [Online]. Available: https://www.snpambiente.it/wp-content/uploads/2020/12/PARTE-INIZIALE-STATO-E-TREND.pdf. |
| [2] | «500 actions to take, including 82 measures to reduce both outdoor and indoor air pollution that can help prevent 7 million premature deaths worldwide,» 2 09 2021. [Online]. Available: https://www.who.int/multi-media/details/500-actions-to-take-including-82-measures-to-reduce-both-outdoor-and-indoor-air-pollution-that-can-help-prevent-7-million-prematuer-deaths-worldwide. |
| [3] | «9 out of 10 people worldwide breathe polluted air, but more countries are taking action,» 2 05 2018. [Online]. Available: https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action#:~:text=Air%20pollution%20levels%20remain%20dangerously,outdoor)%20and%20household%20air%20pollution.. |
| [4] | A. R. P. A. V. d'Aosta, «ll particolato atmosferico - PM10 e PM2.5,» [Online]. Available: https://www.arpa.vda.it/it/aria/l-inquinamento-atmosferico/2536-il-particolato-atmosferico. |
| [5] | A. Lombardia, «PM10 E PM2,5,» [Online]. Available: https://www.arpalombardia.it/Pages/Aria/Inquinanti/PM10-PM2,5.aspx. |
| [6] | «PM10 - Particolato atmosferico o polveri sottili,» [Online]. Available: https://www.issalute.it/index.php/la-salute-dalla-a-alla-z-menu/p/pm10-particolato-atmosferico-o-polveri-sottili#effetti-sulla-salute. [Consultato il giorno 28 02 2018]. |
| [7] | M. Migliore, G. D'Orso e D. Caminiti, «The environmental benefits of carsharing: the case study of Palermo,» *Transportation Research Procedia,* vol. XXXXVIII, pp. 2127-2139, 2020. |
| [8] | A. Longo e E. Miccoli, «Studio e analisi della correlazione tra inquinamento ambientale e mortalità nella popolazione italiana». |
| [9] | R. News, «A Trento a 30 all'ora, ecco il piano del Comune per la sicurezza,» [Online]. Available: https://www.rainews.it/tgr/trento/articoli/2023/04/a-trento-a-30-allora-ecco-il-piano-del-comune-per-la-sicurezza-1a2c52be-39f6-4c89-97b7-7d4866fc3b07.html. |
| [10] | I. Macchi, «In Italia crescono le città con limite a 30 km/h,» [Online]. Available: https://www.alvolante.it/news/italia-crescono-citta-limite-30-kmh-381060. |
| [11] | N. I. o. E. H. Sciences, «Air Pollution and Your Health,» [Online]. Available: https://www.niehs.nih.gov/health/topics/agents/air-pollution/index.cfm. [Consultato il giorno 06 05 2023]. |

1. I dati relativi al numero di morti sono disponibili fino all’anno 2020 e quelli relativi alla popolazione a partire dall’anno 2019. Sono stati quindi considerati solo i dati relativi a questi due anni. [↑](#footnote-ref-1)
2. Secondo il National Institute of Environmental Health Sciences, alcune delle malattie causate dall'inquinamento ambientale includono il cancro, le malattie cardiovascolari, le malattie respiratorie, il diabete mellito, l'obesità e i disturbi riproduttivi, neurologici e del sistema immunitario [9]. [↑](#footnote-ref-2)