



Corso di laurea in Management and Computer Science

Cattedra di Databases & Big Data

**Integrating Exposome: the relevance of
databases in the implementation of a
comprehensive diagnosis approach**

Prof. Blerina Sinimeri

Relatore

270541

Candidato

Anno Accademico 2023/2024

INDEX

Introduction

Literature

- What is exposomics
- History and future perspective
- Dataset relevance in diagnosis administration
- Nosql datasets and mongodb

Method

- Risk factor data selection
- Risk factors datasets` specifics
- Pollutants` impact data selection
- `Deaths by territory of residence` dataset specifics

Analysis and interpretation

- The importance of data preprocessing
- Analysis

Conclusions

Bibliography

INTRODUCTION

The aim of this thesis is to investigate the Exposome, its relationship to the branches of research that converge in it, and the characteristics it will need to assume for future development. Additionally, it's important to consider the relationship with database technology as well, delving into the problems and the proposed solutions.

Exposomics' goal is to identify the elements, inside and outside the individual, that have an influence on the individual's overall health. The exposome takes into account the natural predispositions, the habits and lifestyle, and the general environment with which the individual comes into contact during the whole course of their life, distinguishing the moments in which certain expositions might have greater relevance. The goal of the study of the Exposome is to use knowledge obtained from these analyses to formulate guidelines for improving the overall health of members of society.

The approach is inevitably tied to extensive data analysis, and because of this it is tied to the issues related to the use of big data and their management. One of the most considerable obstacles in this regard is the evolution of database technology and in recent years the shift from relational databases to NOSQL models. Additionally, MongoDB, one of the most prominent NOSQL implementations, is used as example to explain this technology's architectural aspect further.

Finally, the work conducted on datasets regarding risk factors on Italian territory and deaths by respiratory causes aims at exemplifying the valuable contributions of exposomics.

To understand the topic of Exposomics, the structure of this research will be demonstrated firstly by the details of the data acquired, as well as the sources for them. Next, an explanation of the process to refine the datasets is shown, such as the aspect of getting rid of anomalies that would have been a risk to the results' reliability. Lastly, the outputs derived from the analysis are shown and explained.

LITERATURE

PART 1 - What is exposomics

The exposome is composed of every exposure to which an individual is subjected from conception to death (Christopher. P. Wild).

The definition given by Wild [1], father of the term “Exposome” and former Director of the International Agency for Research on Cancer, can be dissected in two points of focus, the first being the factors that compose the exposome: the general external environment, the specific external environment, and the internal environment. To delve deeper into the factors, it is essential to define them as follows:

- an internal environment to include internal biological factors such as metabolic factors, gut micro-flora, inflammation, oxidative stress.”
- a specific external environment with specific contaminants, diet, physical activity, tobacco, infections, etc
- a general external environment to include factors such as the urban environment, climate factors, social capital, stress;

The internal exposome basis can be derived from the term “omics” which can be defined as the probing and analyzing of large amounts of data representing the structure and function of an entire makeup of a given biological system at a particular level [2]. These omics can analyze both large and small molecules, gene expression profiles, and reactive electrophiles. Recent animal and invitro studies have demonstrated connections between risk factors and epigenetic (methylation and miRNA) modifications, and similar correlations have been found in the more limited human studies as well. The hope for improving these studies would be to find a unique matrix that could play an equivalent role to the DNA sequence in ‘genome-wide association studies’. However, this is where biomarkers face challenges such as cost and sample quality

In regard to external exposome, a distinction is made between specific external environments at the individual level and those mainly configured with data from questionnaires. In contrast, geographical mapping assesses the general external exposome at the community level. In the paper `The Exposome: a new paradigm to study the impact of environment on health` (2014) [3] the author provides the opinion that the individual level surveys could be improved by combining questionnaires with biomarkers alongside smartphone-linked diaries and imaging and that the community level surveys could be improved by studying how humans move through their environment, estimated inhalation rates that could be collected through a smartphone that can measure physical activity, and people's movement through the microenvironment that can also be collected through a smartphone's GPS.

Back to the initial definition, the second point of focus concerns the exposome's dynamic approach, as the field aims at taking into account the nature of exposition over time. Experimental studies have found that a given exposure or dose can have different effects because one specific instance cannot characterize all impacts of the environment. One possible solution to this problem may be developing key points throughout the life cycle, such as prenatal life or infancy, which are extremely vulnerable stages to environmental risk factors.

Lastly, another consideration from Wild, who strongly emphasized how, despite the possibility for a spillover into personal medicine, the development of exposomics should be mainly considered as a tool for the study of epidemiology.

PART 2 - History and future perspective

In 2003, the Human genome project had just reached its successful end after more than a decade of efforts from a diverse team of international scientists. The achievement laid the foundation for personalized medicine, providing a tool to trace early indicators of diseases such as cancer and cardiovascular diseases. In reflection of this extraordinary result, the field of exposomics was born.

Because of this strong relationship with genomics, in a 2020 essay for the International Journal of Environmental Research and Public Health [4], scientist and author Stefano

Canali argued that exposome is not necessarily an entirely new paradigm, but an advancement in the interaction between different fields of study, which in turn leads to the formation of a shared pool of technologies and different expertise. Three main contributors are recognized in the birth of exposomics: genome, exposure science and biomarkers science.

Canali also focused on the challenges faced by exposomics innovation, discussing how even some of the latest projects exhibited a lack of much needed long-sightedness, prioritizing short term projects that often struggle to even get to completion because of the excessively limited time at disposal. These concerns relate to the point made by Wild, who recognized that the lack of a moon-shot goal, like the one genome had, makes it difficult to focus funding, suggesting eventually to concentrate efforts in the recognizable divisions of emphasis [...] by assigning specific international teams with defined goals and shared expertise [1].

While the authors' concerns are justified, it is important to recognize that there has been movement in the right direction, such as the infrastructure developed in the context of EHEN (European Health Data and Evidence Network): the project, begun in 2018, aims at building a pan-european system which will allow simplified sharing of Health Data among researchers while complying with the GDPR framework.

Both Wild and Canali recognize the Human Genome Project as a guideline for what the evolution of exposomics should be, identifying the international commitment to the development of the research infrastructure and substantial financial backing as key factors not only in the realization of the objective, but also in the consolidation of progress.

Even though the figures available online suggest that we are still far from substantial funding, especially when comparing to the expenses that led to the completion of the genome project (estimated 2.7B from the US alone), it is still possible to assume that there is interest in developing the approach further. A testimony of this is the renovation of funding throughout the Horizon projects, the EU programs targeting scientific development (Horizon 2020, evolved in Horizon Europe since 2021), and the investments made by US institutions, another key player in exposomics research.

Finally, I think it is important remember the emphasis from Wild on the significant impact advancements in exposome understanding will have not only for the countries who will likely contribute the most to it, but also and perhaps mainly to the mid-to-low income countries, for it is in these regions that the burden of cancer and other non-communicable diseases is set to rise the most dramatically [1].

PART 3 - Databases relevance in healthcare

In addition to the logistical problems exposome research faces, as previously mentioned, there are technical problems regarding data processing and retrieval. By highlighting these issues, databases' relevance emerges as a defining factor in the development of this approach: it is fundamentally important that efficient management of different information becomes a point of focus in the research to select the right tools and consequently foster innovation on the right path.

The healthcare industry holds massive amounts of data, referred to as big data. This data is so overwhelmingly large as it stores clinical data, patient data, machine-generated data, web pages, and emergency care data. Within this massive amount of stored data, there is an opportunity, because by understanding and analyzing it there is the possibility to improve health care. A few possible improvements that these data analytics can help include selecting at-risk patients, identifying treatments and managing population health. Research and progress, however, do not have to be the only driving factors in the push for solutions to these technical obstacles, as it has been demonstrated that big data analytics in healthcare allow, while maintaining the increase in quality of treatments, for significantly lower costs: McKinsey estimated in 2021 that about \$500 billion would be saved in the US healthcare system alone [5].

This discussion of big health data is characterized by volume, velocity, variety, and veracity. Volume refers to the massive amount of data accumulated, it is reflected at the architectural level in the necessity to break down processing and execute it across different nodes.

Velocity refers to the constant flow of data accumulated in real-time, and variety refers to the type of data collected, including structured, unstructured, and semi-structured.

Veracity refers to the scaling up of architectures and platforms, algorithms, methodologies, and tools to match the demands of big data.

These properties have determined a significant divide between tools traditionally used for healthcare analytics and the ones suitable for big data use: While the algorithms and models are similar, the user interfaces of traditional analytics tools and those used for big data are entirely different; traditional health analytics tools have become very user friendly and transparent. Big data analytics tools, on the other hand, are extremely complex, programming intensive, and require the application of a variety of skills [6].

Another problem incurred by big data in the context of healthcare concerns the majority of database structures fit for storing and handling of these large datasets being developed as open-source projects. This, despite meaning that the development costs will be lower, implies drawbacks in security and technical support availability.

In conclusion, the overlook reveals a relatively embryonic state. The potential is immense, but it might be too early to be fully taken advantage of. However, strong incentives and initial signs of efforts from institutions are a promising sign towards the future.

PART 4 - NOSQL Databases and MongoDB

While relational DBMS has long proved to be an essential resource in contexts which required deep reliability of information, which concerned directly the possibility to update data whenever needed. NOSQL databases were born because the new trends brought by the characteristics of Big data required the development of new solutions.

After starting their evolution in the first decade of the 2000s, NOSQL databases grew to become the answer to the questions raised by the new approach, as they allow for increased flexibility and are inherently more oriented towards horizontal scaling. Opposedly to relation-based databases, these structures come in different forms: key-based, graph based and document based.

Key based is the simpler structure, graph based is instead used for all the instances that, intuitively, present a graph structure. Document based NOSQL is used to store data,

instead of in columns and rows, in document format to which is associated a key. The retrieval process can happen in two distinct ways, either leveraging the key association or by specification of formal conditions. The querying offers a third retrieval opportunity, as it allows further specification of the conditions for the retrieval. Other possibilities offered by the queries are creation, update and removal of documents.

Usually, when referring to document based NOSQL applications, the immediate connection is to MongoDB, the most representative and diffused tool. MongoDB supports different data formats, mostly JSON, which works well for interchange and maintains high human readability, and BSON, essentially a binary version of JSON allowing for increased efficiency at the expense of interpretability . In the model, fields correspond to the tables of a relational database, they aggregate into collections. A cluster of collections forms a database.

The database`s architecture is split in components [7] :

- MongoD manages data access
- Mongos routes the user application to the actual database
- A configuration server stores metadata in order to be able to locate the operations required by the user
- Replica set: divided between primary and secondary sets, keep copies of the data
- shard : similar to the replica sets, but its role is to store only a portion of the data in order to facilitate horizontal scaling between different processors

METHOD

As exposomics represents a promising field to investigate the relationship between environmental exposures and human health outcomes, the objective is to implement an experimental study to inquire about the relationship between air quality and the health of individuals subject to airborne pollutants.

PART 1 - Risk factors data selection

The first step of the project consisted in identifying which factors should be used to define pollution levels and how to address the health situation. Guidance to answer the first question came from the latest Legambiente report on air pollution in Italian cities, *Mal'aria di città 2024* [8], which takes into consideration three primary risk factors to determine air quality: PM10, PM25 and NO2

- Pm10: small particles measuring less than 10 micrometers, which can remain suspended in the atmosphere for a long time, reducing visibility and infiltrating the respiratory system, ultimately causing adverse health effects. The current Italian directive, D.Lgs 155/2010, fixes the daily threshold for human health protection to 50µg/mc. The annual limit for human protection, determined by the same law, corresponds to an average of 40µg/mc.
- Pm25: a mixture of various chemicals, all airborne and all measuring less than 2.5 micrometers. It is particularly dangerous for humans, as its dimension allows it to penetrate deep into the respiratory system and infiltrate the bloodstream. As of today, Italian legislation has fixed the threshold at 25µg/mc for daily and annual average concentration.
- No2: Nitrogen dioxide can cause in individuals subject to exceeding exposition both short- and long-term effects, on top of this it is also a contributor of photochemical smog formation. Italian directives take as reference both an hourly level and a yearly level. The threshold of 200µg/mc should not be exceeded for

more than an hour no more than 10 times per year, while the general aim for the yearly average is $40\mu\text{g}/\text{mc}$.

The report also mentioned obtaining data from air quality control units. Given that these units fall under regional jurisdiction, the initial idea was to make use of Lombardy's air quality dataset to focus the study solely on my home region. However, the restricted localization of the dataset proved to be a hurdle, as accessing and navigating it represented a significant challenge during consultation. Consequently, the focus shifted towards the website of the national institute for environmental research and protection (ISPRI) [9]. This alternative granted access to wide datasets encompassing not only the three risk factors mentioned above, but also ozone (O₃), another pollutant posing threats to human health

- O₃: Ground-level ozone affects human health by impairing respiratory and cardiovascular function, which leads to more hospital admissions, school and work absences, medication use, and even premature mortality. Short-term exposure to ozone is associated with respiratory symptoms, reduced lung function and airway inflammation; long-term exposure with aggravated asthma and an increased incidence of strokes (European climate and health observatory) [10]. Currently, Italian legislation requires hourly ozone levels to not exceed $180\mu\text{g}/\text{mc}$, while the long-term goal for human health protection requires a threshold of $120\mu\text{g}/\text{mc}$ as daily maximum.

PART 2 - Risk factors datasets` specifics

This research led to four separate datasets, one for each risk factor. These datasets were downloaded from the ISPRA website mentioned above, and all were already available in form of .xlsx Excel spreadsheets. The document dimensions were:

- 1.8 MB for the PM₁₀ dataset
- 694 KB for the PM₂₅ dataset
- 2.1 MB for the NO₂ dataset
- 1.5 MB for the O₃ dataset

Each dataset contains the following columns:

- Station_eu_code: a European code to identify each measurement unit
- Regione: The name of the region in which the control unit is
- Tipo_zona: For every unit, the category of area in which it is located is specified. Three are the main categories: URBAN, SUBURBAN and RURAL
- yy: Each row of the dataset reports the measurements from one unit over one specific year, this value indicates the year
- Media_yy: The value showcased corresponds to the yearly average of the measurement recorded by the specific station

As it will be explained in the forthcoming section, the datasets encompass other information regarding data quantity, statistical metrics including percentiles, and more detailed information about the position of the units. However, these were determined to have limited relevance to the specific analysis being conducted, and as a result will not be further elaborated upon.

Another point taken into account in the following chapter is the different durations of recording activity for each risk factor, as

- PM10 dataset spans from 2002 to 2022
- PM25 has been recorded in the dataset from 2004 to 2022
- NO2's records in the dataset go from 2001 to 2022
- O3 goes from 2002 to 2022

PART 3 - Pollutants' impact data selection

Regarding the way in which to address the impact of air quality on human health, the choice was made to utilize the ISTAT (Italian National Institute of Statistics) website [11] to access datasets referring to deaths associated with respiratory causes. ISTAT categorizes deaths based on various attributes, including the location of the event and the residence of the deceased. A decision was made to prioritize the residence, as it seemed

pertinent to attribute the 'cause' to the place where individuals likely spend the majority of their time.

PART 4 - 'Deaths by territory of residence' dataset specifics

The chosen option, 'death by territory of residence,' includes a wide and very general array of data, hence necessitating subsequent refinement. As the institute's web page allows filtering operations before even exporting the dataset, the process focused on the geographical area and the specific causes of death. The retained data are divided by region, with no concern for specific provinces. Moreover, causes of death unrelated to air quality, such as suicides and brain cancers, were removed from the dataset. It is also worth mentioning that the years taken into consideration in the dataset are 2004 to 2021.

The final document was exported in .xlsx format, maintaining continuity with the other datasets, and has a dimension of 438 KB.

ANALYSIS AND INTERPRETATION

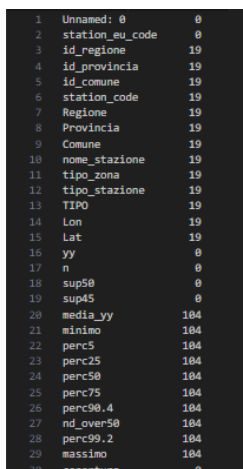
The following study was conducted using Python programming language. Additionally, during most stages of the analysis were used commands from various Python-supported libraries: Pandas, a software library offering data structures and operations for the handling of numerical tables; Numpy, which simplifies mathematical operation in Python; Seaborn and Matplotlib, libraries focused on data visualization.

Part 1 - The importance of data preprocessing

To perform an analysis on data, they must first undergo a cleaning process. The aim is to avoid the flawed results that would derive from any inaccuracies and inconsistencies contained in the dataset. In the specific case at hand, all the datasets were derived from structured data, as they were all imported from the websites in form of excel spreadsheets

i. PM10 dataset cleaning

The first dataset cleaned was PM10, containing the detected values for the polluting agent from 2002 to 2022. The preprocessing started by filtering out useless columns, as they did not correspond to any column listed in the dataset legend provided on the source website. The following step was the removal of rows containing duplicate values with the 'df.drop_duplicates' command, and the result showed there were none. Next, to handle missing data, the functions '.isna' and '.sum' were combined to get a clearer overview on the situation.



```
1 Unnamed: 0      0
2 station_en_code  0
3 id_regione      19
4 id_provincia    19
5 id_comune       19
6 station_code    19
7 Regione         19
8 Provincia       19
9 Comune          19
10 nome_stazione  19
11 tipo_zona      19
12 tipo_stazione  19
13 TIPO           19
14 Lon            19
15 Lat            19
16 yy             0
17 n              0
18 sup50          0
19 sup45          0
20 media_yy       104
21 minimo         104
22 perc5          104
23 perc25         104
24 perc50         104
25 perc75         104
26 perc90.4       104
27 nd_over50      104
28 perc99.2       104
29 massimo        104
30 costruttore    0
```

Figure 1

Since two main groups of columns were shown to have the same number of missing values (figure 1), a decision was made to handle them by creating a subgroup restricting the action of '.dropna' to one column for each set. This solution proved able to solve the problem at once.

Lastly, to get the data to coincide with the dataset regarding the deaths over the years, all the rows dated before 2004 and after 2021 were eliminated.

ii. PM25 dataset cleaning

For the PM25 dataset the same actions were performed, and again no duplicates were found. The few missing value rows were taken care of by deleting them altogether as it had been done for PM10. Finally, we filtered the rows by year again and got them to conform to the ‘deaths over the years’ dataset.

iii. NO2 dataset cleaning

NO2 dataset showed to be similar to the previous two, as the only operations necessary were column drop and the handling of missing values. In this instance as well no duplicates resulted.

iv. O3 dataset cleaning

Once again, columns with no significant connection to the scope of the analysis and the ones that did not match the description of the dataset available on the source webpage were dropped, and no duplicates were found. However, the missing value situation proved to be slightly trickier than it had been for the other risk factor, requiring a different solution.

The inspection, visible in figure 2, revealed that some columns presented only a few empty cells, one notable group of columns showed to have the same number of missing values, and a higher number of missing values concentrated in the columns ‘aot40v_s’ and ‘val_ob’.

```

1 station_eu_code 0
2 id_regione 1
3 id_provincia 1
4 id_comune 1
5 station_code 5
6 Regione 1
7 Provincia 1
8 Comune 1
9 nome_stazione 5
10 tipo_zona 5
11 tipo_stazione 1
12 TIPO 1
13 Lon 1
14 Lat 1
15 yy 0
16 n 0
17 n_estate 0
18 n_inverno 0
19 n_aot40v 0
20 n_aot40f 0
21 sup180h 0
22 sup240h 0
23 media_yy 17
24 minimo 17
25 perc5 17
26 perc25 17
27 perc50 17
28 perc75 17
29 perc95 17
30 perc99.8 17
31 massimo 17
32 aot40v_m 0
33 aot40f_m 0
34 aot40v_s 150
35 aot40f_s 77
36 copertura 0
37 copertura_aot40v 0
38 copertura_aot40f 0
39 n_giorni_validi 0
40 giorni_over_180 0
41 giorni_over_240 0
42 giorni_mm8_over_120 0
43 val_ob 542
44 giorni_mm8_over_100 0
45 giorni_mm8_over_160 0
46 percentile99_mm8 13
47 peak_season 75
48 dtype: int64

```

Figure 2

Apart from the two specific cases represented by the columns just mentioned, it was decided to implement 'Pandas.dropna' and eliminate the rows containing missing cells entirely, since the limited reduction would have not represented a significant change on the overall dataset. On the other hand, to better understand the distribution of the two particular columns, two histograms were hatched. The results showed the two columns to have a left-skewed distribution. 'Pandas.fillna' function was decided to be applied in both instances, and because of the skewness in the distribution of the existing values the median was chosen to substitute the missing ones.

In the end, all the rows pertaining to years before 2004 and after 2021 were filtered out.

v. ‘Deaths by territory of residence’ dataset cleaning

The dataset regarding the deaths caused by respiratory causes constituted, to some extent, a different instance from the ones observed above, as we had the opportunity to filter out most of the unwanted information through the functions offered by the website before even exporting the data, as was mentioned in the ‘method’ chapter. Nonetheless, data underwent the same basic steps, removing duplicates and missing values, which prompted removing the column ‘Flags’ altogether, as it was automatically added during the export process to warn users in case of data corruption, which luckily did not occur.

Part 2 – Analysis

i. PM10 analysis

Analysis on the PM10 dataset began with a barplot representing the yearly average. The data in the ‘media_yy’ column was filtered by year using the ‘.groupby’ function of the Pandas library, and the yearly subsets were averaged through the ‘.mean’ command. The obtained output, in Figure 3, allows visualization of the PM10 situation of the Italian territory over the last 20 years.

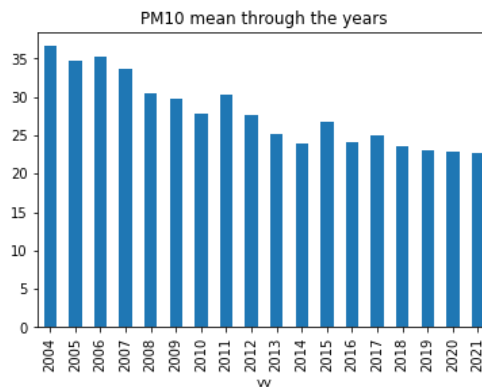


Figure 3

The trend has seen a significant decline in values, with sporadic setbacks, such as in 2011, the year in which the current regulation entered its period of validity, and overall has managed to reach values lower than 25 $\mu\text{g}/\text{mc}$ in 2021. While this represents a considerable accomplishment, given the current annual limit being 40 $\mu\text{g}/\text{mc}$, it is important to recognize the relative weight of the attainment, as the World Health Organization (WHO) guidelines recommend to not exceed 15 $\mu\text{g}/\text{mc}$. This perspective explains that even though current strategies lay on the right track, the recorded values are still far from providing a safe environment for human life, hence highlighting the need for further work.

The analysis continues with a representation of PM10 levels in the different kinds of areas for two years of reference. It should be noted that the dataset for the risk factor took originally into consideration 6 classes: urban, suburban, rural, rural_remote, rural_nearcity and rural_regional. As for the finality of this analysis there would be no notable benefit in considering subclasses of the ‘rural’ area, the rows containing the subsets were simply renamed as one collective area. Another issue concerned some classes being recorded in all caps while others only had the first letter capitalized, in this case the ‘Pandas.str.upper’ command was sufficient to solve the issue.

The first year referenced in the graphic is 2004 (figure 4), which recorded values slightly higher than 25 $\mu\text{g}/\text{mc}$ in rural areas, slightly lower than 35 $\mu\text{g}/\text{mc}$ in suburban areas and slightly higher in urban ones. The second year taken into consideration is 2021 (figure 5), in which the recorded values are noticeably lower, as values in rural areas have decreased on average to be less than 20 $\mu\text{g}/\text{mc}$, with urban areas recording less than 25 $\mu\text{g}/\text{mc}$.

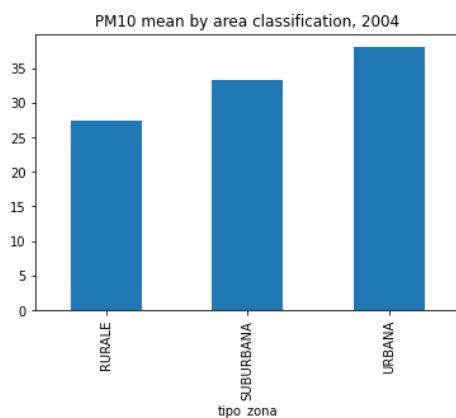


Figure 4

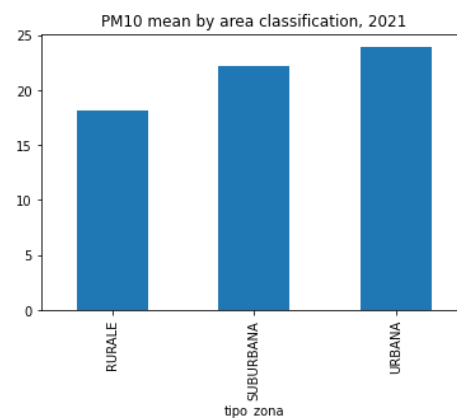


Figure 5

Lastly, heatmaps were created to visually represent regional pollution in different years. The process began by performing an operation of grouping and averaging similar to the one used before, but this time the subsets were created from the 'Regione' column, thus obtaining values for each Italian region in the given year. These values were represented with a single variable, 'A'. On this variable was implemented the 'Pandas.reset_index' function, deriving a second variable 'B', reassigning the default index. Later 'Pandas.set_index' was used on the 'B' variable, specifying as index the 'Region' column, this way we were finally able to compute the heatmap using the Seaborn library.

The first heatmap, shown in figure 6a, regards the year 2004. The results show a range of values between 20 and 50 micrograms of cubic meters concentration, with regions such as 'Marche' and 'Lombardia' showing higher levels of particulate dispersed, while regions like 'Valle d'Aosta' and 'Friuli-Venezia Giulia' have significantly lower average values.

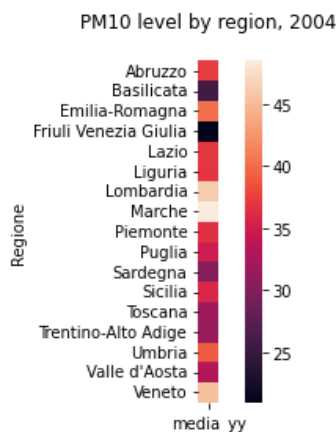


Figure 6a

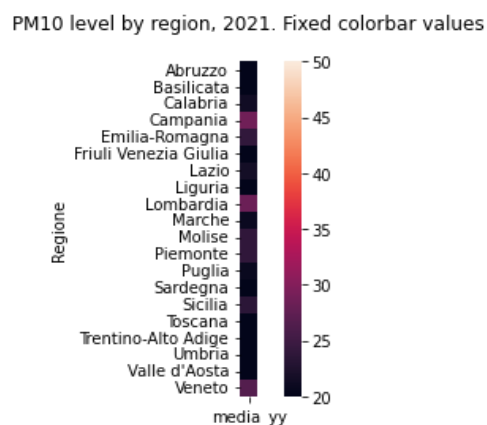


Figure 6b

The second heatmap (figure 6b) shows the regional situation in 2021, but in order to get a fair comparison between years, the range of shown values in the colorbar of the heatmap was set to be the same as in 2004. The represented circumstance shows a clear improvement, as all the regions are now characterized by a darker color range.

The analysis was concluded by drawing a heatmap representing the situation in 2021 (figure 7). This time, however, the colorbar boundaries were left to be automatically assigned by the function. The heatmap shows clearly how the densely industrialized regions are still showing the highest levels of pollution, but the fact that the majority of regions now range between 26 and 18 micrograms of airborne particulate represents a promising signal.

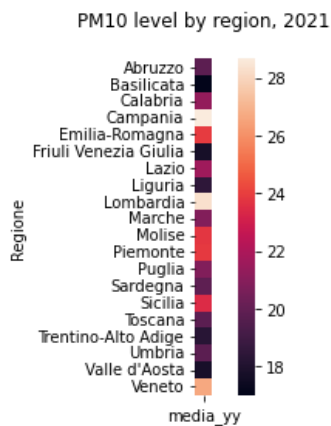


Figure 7

ii. PM25 analysis

The analysis of the PM25 dataset reveals a contrasting landscape. If the reference for average annual values were to be the 25 $\mu\text{g}/\text{mc}$ suggested by Italian regulation, then the barplot (figure 8) would represent a consolidated safety in 2021, and the last year of concern would be all the way back in 2007, as the last two decades have shown an encouraging decreasing trend, despite some marked fluctuations. This interpretation must, however, be reconsidered in light of the WHO most recent suggestions for a much lower limit, fixed at 5 $\mu\text{g}/\text{mc}$ for the yearly average. While it would be wrong to discard the achieved results, the perspective should be a call for reconsideration of directories on practical applications in Italy, especially given the elevated threat to human health posed by this pollutant.

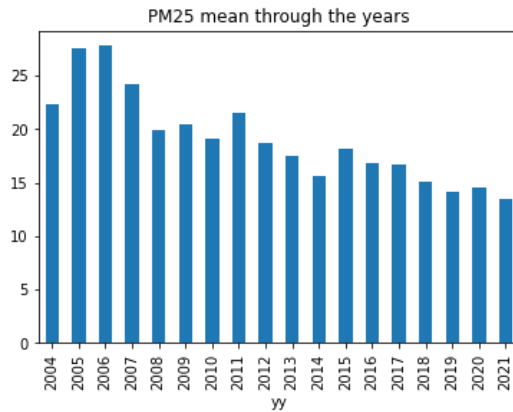


Figure 8

The same situation results when 2021 is investigated by area kind (figure 9), as even rural areas show levels of particulate exceeding $10\mu\text{g}/\text{mc}$

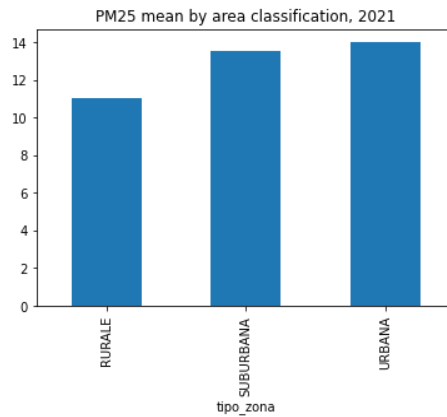


Figure 9

The next step of the analysis, heatmap representation, incurred the issue of some early years not accounting for all the regions. This prompted the choice to use 2013 as first year of reference, as it was the first usable year. In this first year (figure 10a) the values ranged between 26 and 10 micrograms of particulate matter per cubic meter, with the region of Lombardy showing the most concerning situation, having an average concentration level above $25\mu\text{g}/\text{mc}$.

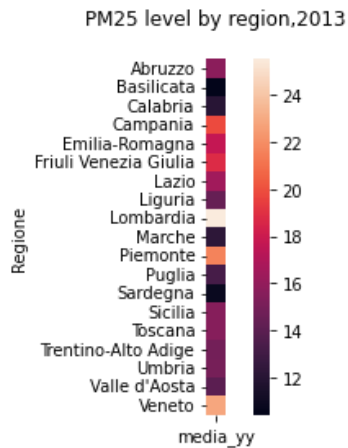


Figure 10a

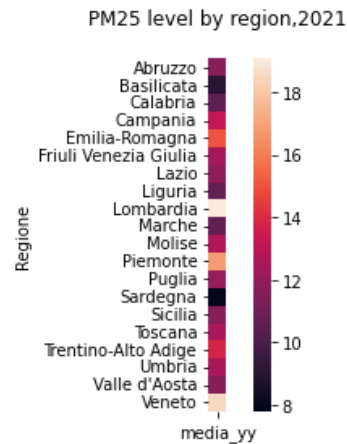


Figure 10b

Figure 10b shows the development in 2021. Even though it is important to consider the relevant reduction in the scale of values, the situation has remained mostly unchanged: Lombardy and Veneto are still by far the most polluted regions, while Emilia-Romagna and Trentino-Alto Adige seem to have not been able to maintain a linear decrease. Even though this is ultimately a testimony to the successful efforts of domestic administrations to reduce air contamination, the future might require strategies to aim for a more even distribution of the sources of particulate.

iii. NO2 analysis

From the barplot represented in figure 11, it is possible to observe how the management of NO2 has halved the risk factor's level over the years. In 2004, nitrogen dioxide had a concentration close to 40µg/mc, while in 2020, it reached its recent minimum of just under 20µg/mc, followed by a negligible rise in 2021.

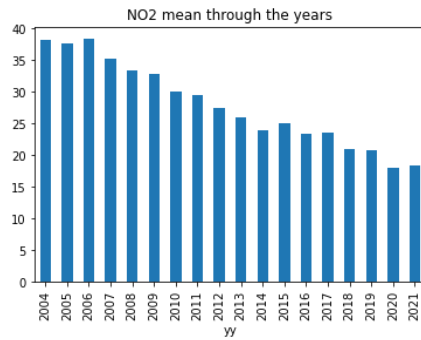


Figure 11

This critical decrease is further highlighted by the graphs depicting the pollutant's situation in different area kinds in 2006 and 2021, respectively Figures 12 and 13. For every area class, the value is half of what it was in the first year of reference.

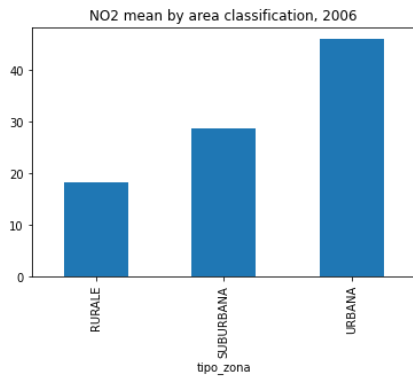


Figure 12

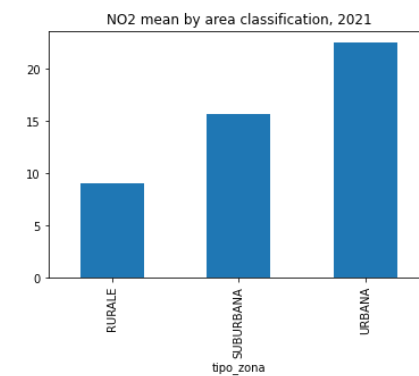


Figure 13

The heatmaps, created using the same years of reference as the area type barplots above, depict the progress achieved by the regions over time. Indeed, in 2006 (figure 14a) the situation seemed to be rather concerning, as most regions showed colors from the uppermost part of the colorbar, thus ranging between 35 and 50 micrograms per cubic meter. When the same range is applied to the 2021 regional averages, in figure 14b, it is easy to see the progress, as every region is characterized by the darker shades of the colorbar.

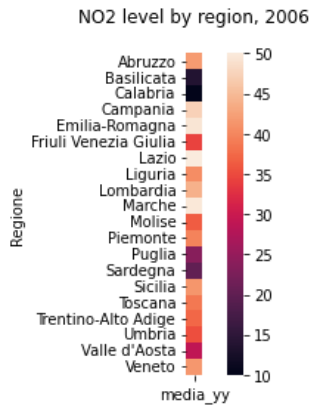


Figure 14a

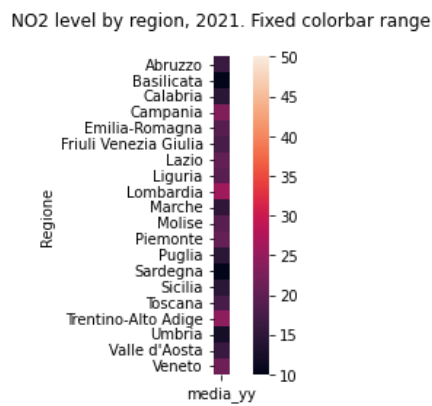


Figure 13b

iv. O3 analysis

Contrastingly from the other risk factors analyzed, ozone showed a remarkably different trend. As results from the barplot in figure 15, the levels of O3 in Italy since 2013 have fluctuated lightly, maintaining a concentration in the air between 50 and 60 micrograms.

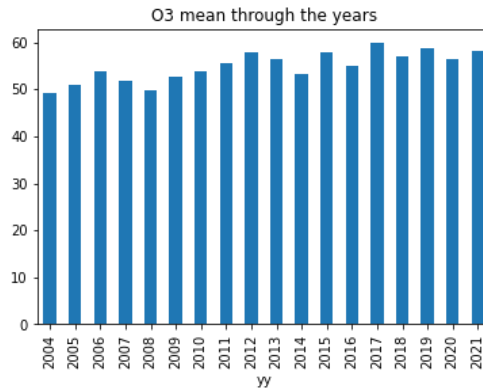


Figure 15

The graphs (figure 16, 17) comparing the situation in urban, suburban, and rural areas in different years confirm this result and reveal another uncommon characteristic: The observed amount is higher further out from the centers, where we found that other risk factors were more concentrated. This phenomenon was explained by R. Bertollini, director of WHO Health and Environment Europe, by saying: ‘Paradoxically, ozone is less concentrated in more trafficked urban areas. Reacting with NO₂, mainly emitted by cars, it is destroyed by the same emissions that produce it in the first place. It accumulates instead in green areas, such as parks, or in rural areas because here the same autodestructive mechanism that happens in city traffic is not in place’[12]

Another factor influencing this result is described by the European Climate and Health Observatory: ‘Since photochemical formation of ozone takes hours, winds can transport the peak of pollution before O₃ is actually created.’[10]

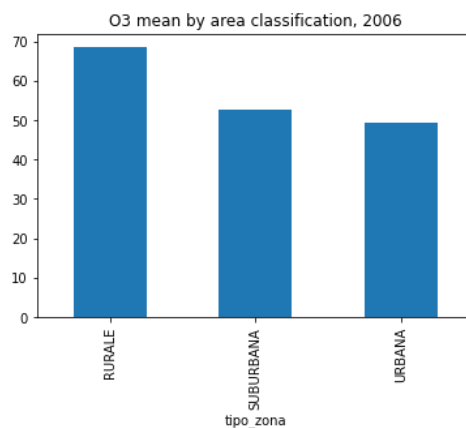


Figure 16

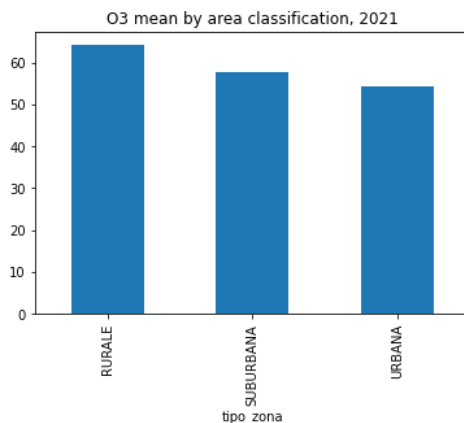


Figure 17

The heatmaps reveal that, at a regional level, a complex situation. 2006, in figure 18a, was characterized by a relatively varied distribution, with regions such as Abruzzo and Marche recording less than 45 micrograms of ozone per cubic meter and Calabria representing the pollution peak at over 70 $\mu\text{g}/\text{m}^3$. The heatmap for 2021 (figure 18b), to which is applied the same range as 2006 for the colorbar values, is instead more uniform, with the majority of regions observing high risk factor levels and more than one region in the colorzone of 70 $\mu\text{g}/\text{m}^3$. At the head of this undesirable standing is Sicily, with Basilicata and Calabria closing the podium.

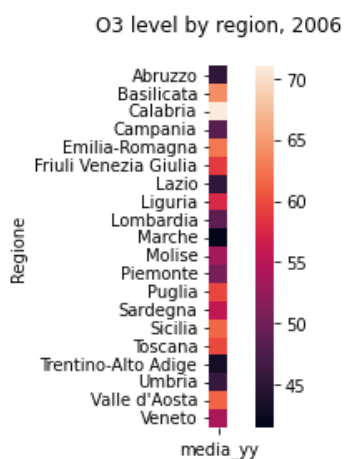


Figure 18a

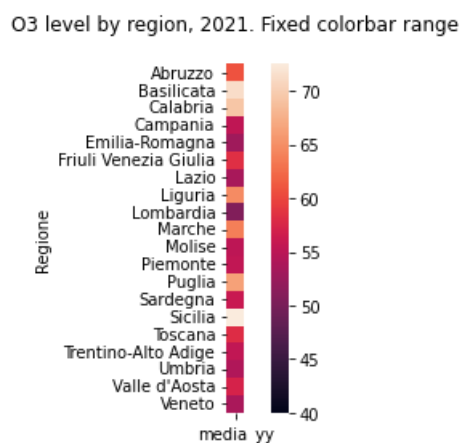


Figure 18b

v. 'Deaths by territory of residence' analysis

The analysis of the dataset regarding deaths caused by respiratory causes began with the creation of the barplot summing all the deaths by year, without distinction on the specific respiratory symptoms at their root. The initial dataset was chosen to include COVID-19 in the symptoms leading to demise, explaining the noticeable spike in 2020 and 2021. Nonetheless, figure 19 shows an increasing trend in the number of deaths by year, even before the pandemic.

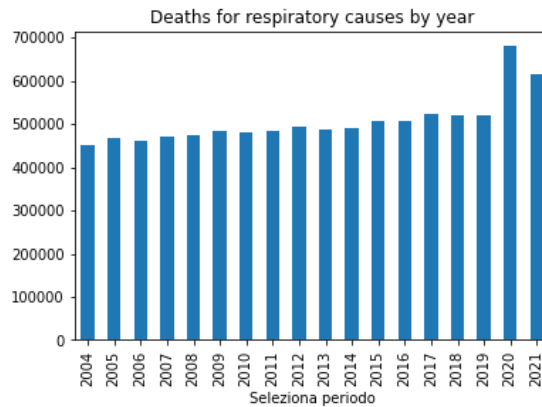


Figure 19

The next step was inspecting the total mortality by region. In order to obtain a significant result, it was important to make sure that the difference between the populations did not impact the final output. Normalization was thus implemented on the regions' totals. Figure 20 demonstrates an unsettling overview, as the deceased in Lombardy are more than double the number of any other region. Furthermore, this is not the only concern: five regions, excluding Lombardy, registered relatively high death rates.

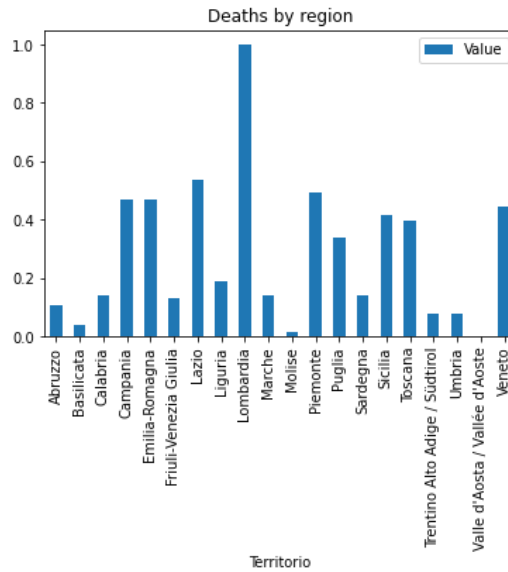


Figure 20

To ensure that COVID-19 wasn't a determining factor in the regional count, a filtered copy of the dataset was created, excluding all pandemic-related causes. Instead of showing differences, the resulting graph (figure 21) confirms the general data: Lombardy maintains the undesirable leading position, with a wide array of runner ups.

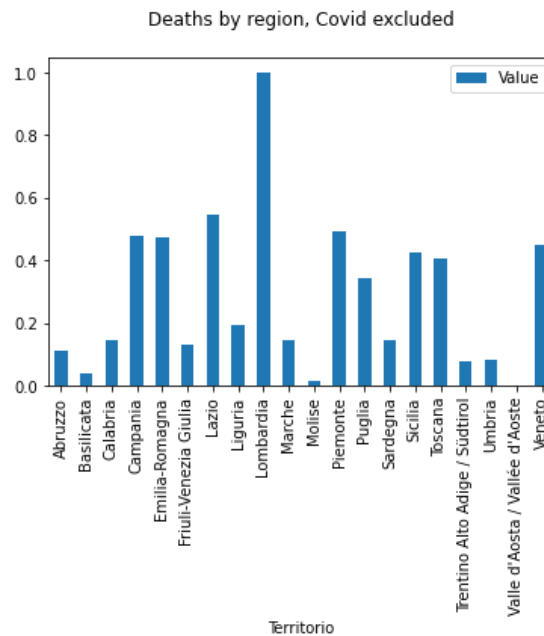


Figure 21

To increase interpretability, the same graph was sorted in descending order in figure 22. This way, the subdivisions appear clearer: Lombardy is the sole element of the first group, a second group

is composed of the five regions between Lazio and Puglia, and a last group in which seemingly respiratory causes do not affect the population as severely.

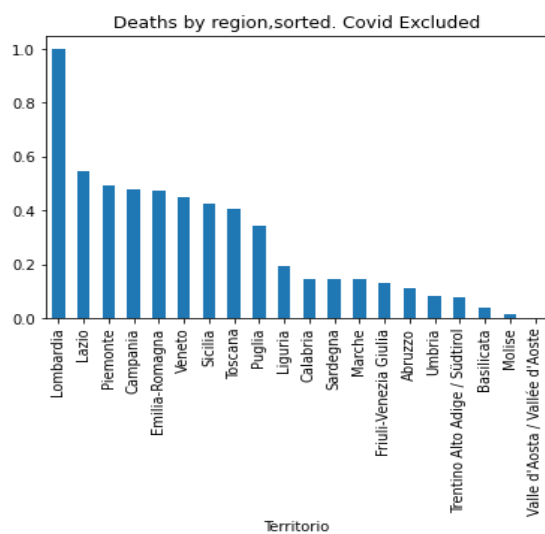


Figure 22

CONCLUSION

Through the analysis of the risk factors datasets, a recurring result was that there are high values registered by a limited set of regions: Lombardy and Veneto steadily appeared as the most polluted territories, followed by Piemonte, Campania and Emilia Romagna. Although they demonstrated to have the ability to drive down the concentration of Particulates and Nitrogen Dioxide, still after almost two decades of effort their air quality is still nearing regulation limits. It is thus unsurprising to find a match in the graph for the number of deaths originated by respiratory symptoms.

Even in the case of Sicily, seemingly unrelated to the set mentioned above, the high mortality rate for ventilatory problems is traceable to the high levels of Ozone. It is, however, difficult to imagine that the O₃ problem alone could have contributed that much to a spike in the number of mortalities, as some other regions showing similarly high concentrations, such as Basilicata and Calabria, ranked significantly lower in figure 22.

Other instances of tough interpretation are Lazio, Tuscany and Puglia. All three regions had average pollution levels at every step of the analysis but recorded a relatively high number of deaths.

While it is reasonable to assume that some other factors might be playing a role in the 'exceptional' results, and we can think of the difficulties of the Sicilian healthcare system, big industrial centers causing a spike in pollution for limited territories inside the region, such as Piombino in Tuscany and the Taranto area in Puglia, whose influence would be harder to trace at regional level due to the averaging operations performed, these results end up being equally as relevant.

While the study points toward a relevant correlation between the amount of risk factors concentrated in the air and the health of a region's inhabitants, the difficulties in explaining part of the findings reveal the same approach criticalities discussed in the

literature chapter, as all of these factors are extremely complex to be accounted for in a single study.

BIBLIOGRAPHY

1. Wild, C. 2012. The exposome: from concept to utility. *International Journal of Epidemiology*, 41: 24-32
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9289742/>, consulted on 15/05/2024
3. Vrijheid, M. 2014. The exposome: a new paradigm to study the impact of environment on health. *Thorax* ,69 :876–878.
4. Canali, S. 2020. What Is New about the Exposome? Exploring Scientific change in Contemporary Epidemiology. *International Journal of Environmental Research and Public Health*, 17, 2879
5. <https://www.mckinsey.com/industries/healthcare/our-insights/administrative-simplification-how-to-save-a-quarter-trillion-dollars-in-us-healthcare> , consulted on 20/05/2024
6. Raghupathi, W. and V. 2014. Big data analytics in healthcare: promise and potential. *Health, Information, Science and Systems*, 2: 3
7. G. Ovando-Leon, L. Veas-Castillo, M. Marin and V. Gil-Costa. 2019. "A Simulation Tool for a Large-Scale Nosql Database," *2019 Spring Simulation Conference (SpringSim)*, Tucson, AZ, USA, pp. 1-12
8. https://www.legambiente.it/wp-content/uploads/2021/11/Report_Malaria-2024.pdf?_gl=1*fr19uv*_up*MQ..*_ga*MTk0NzQ2OTQzNC4xNzE2NTA1OTE3*_ga_LX7CNT6SDN*MTcxNjUwNTkxNS4xLjAuMTcxNjUwNTkxNS4wLjAuMA.. , consulted on 20/03/2024

9. [homepage — English \(isprambiente.gov.it\)](http://isprambiente.gov.it) , consulted on 26/03/2024
10. <https://climate-adapt.eea.europa.eu/en/observatory> , consulted on 21/03/2024
11. http://dati.istat.it/Index.aspx?DataSetCode=DCIS_CMORTE1_EV , consulted on 24/03/2024
12. Faiella, M. G. 2006. 'L'ozono colpisce anche in campagna'. Corriere della Sera, July 03.