Tropospheric Pollutant Concentration Database for the Mediterranean Basin: Addressing Data Gaps and Enhancing Air Quality Assessment

Francisco Sánchez-Jiménez[1], Leandro Segado-Moreno[1], Eloisa Raluy-López[1], Ester García-Fernández[1], Pedro Jiménez-Guerrero[1,2] , Juan Pedro Montávez[1].

[1]Physics of the Earth, Regional *Campus of International Excellence (CEIR) "Campus Mare Nostrum",* University of Murcia, Spain.
[2]Biomedical Research Institute of Murcia (IMIB-Arrixaca), Spain.

Introduction

The Mediterranean basin is a region particularly vulnerable to atmospheric pollution. At the tropospheric level, pollutants such as O_3, PM_{10}, $PM_{2.5}$, NO or NO2 are particularly harmful. The need for sufficiently reliable air quality databases becomes essential to obtain valuable information on the mitigation of air pollution on the population, environment and economy.

What the scientific literature provides is that **existing air quality databases have notable weaknesses** such as lack of inclusion of data from areas close to emission sources, exclusion of cities whose population does not exceed a threshold (Schwela et al., 2020) or the challenge of comparing values due to the presentation of air quality information using city-specific air quality indices (Baldasano et al., 2003).

This study presents a **database of tropospheric pollutant concentrations in the Mediterranean basin for the last two decades**. The data were evaluated using a rigorous quality control process that included detecting manipulation errors, verifying consistency and coherence limits, and assessing spatio-temporal coherence.

Data Collection

The database was constructed from pollution records acquired from thousands of automated air quality stations acquired from:

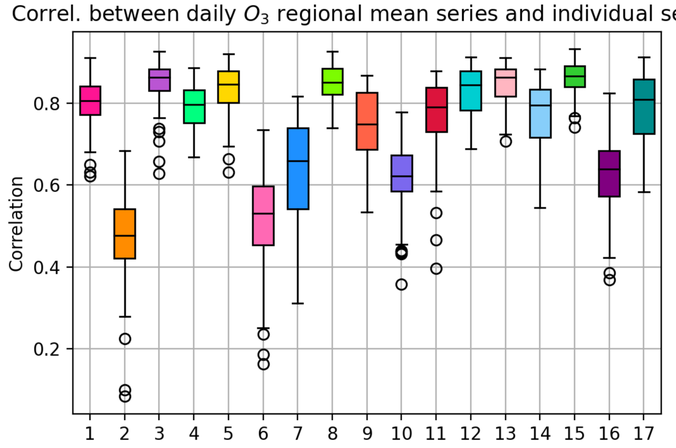
- **AirBase**, provided by the European Environmental Agency (EEA) through the **European Air Quality Portal**. Hourly time series were recorded although work has been carried out on a **maximum daily scale** (Coverage: 2000-2022).

Results

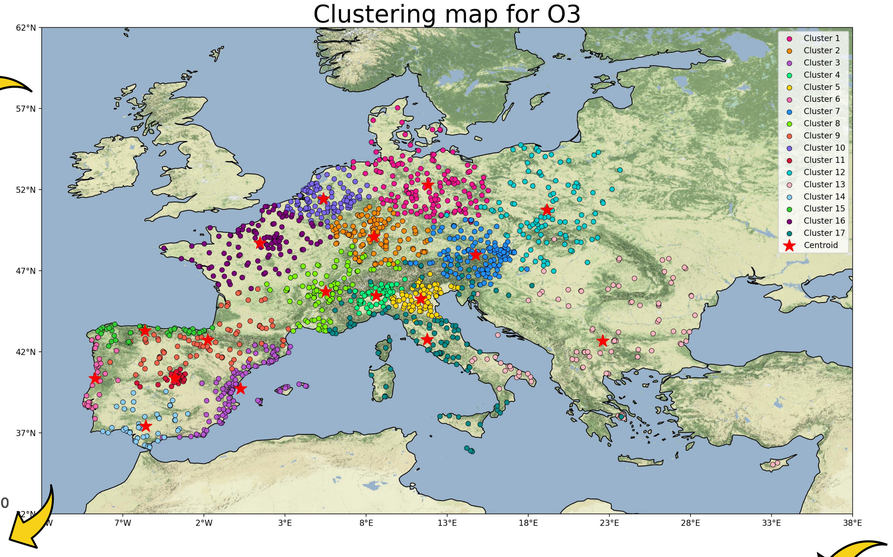
To identify those regions with similar temporal patterns of contamination behavior, a clustering algorithm based on the k-means method has been applied.

$$\min E(\mu_i) = \min \sum_{i=1}^{k} \sum_{x_j \epsilon S_i} \left|\left| x_j - \mu_i \right|\right|^2$$

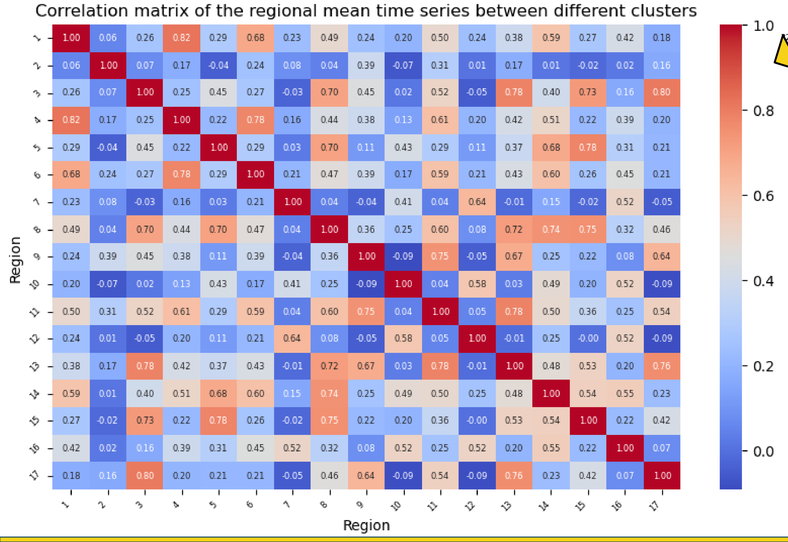
Correl. between daily O_3 regional mean series and individual series

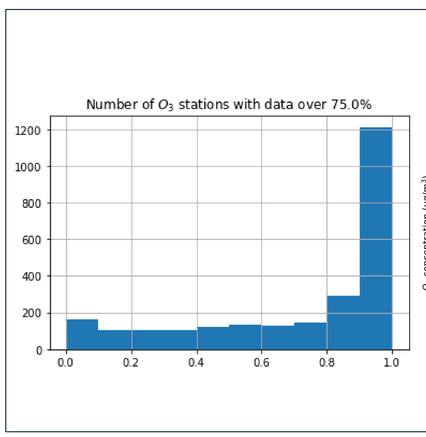


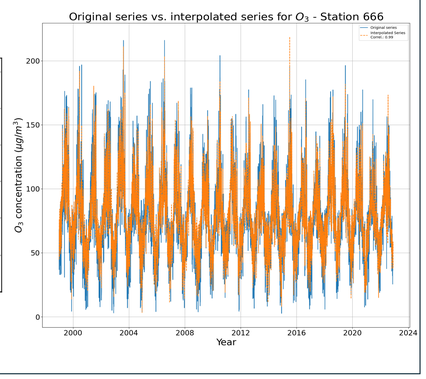
Clustering map for O3



Correlation matrix of the regional mean time series between different clusters







Data Quality Methodology

Deseasonalization process: The first step was to remove the annual cycle (AC). The AC is removed by subtracting the multiyear annual daily mean calculated from a 31-day runmean of the original daily time series.

Data consistency. Measuring stations were found that had been relocated and their time series were supplemented by others. Other measuring stations showed inactivity for some time, which affects the consistency of the data.

Colocalization. For nearby stations:
- If they do not have coincident time steps, they are considered the same and unified.
- If they have time steps, the RMSE is calculated. If this value is less than a certain threshold, these stations are unified.

Temporary coverage. For each pollutant, a minimum percentage of valid data for daily scale data is established for the entire period.

	O3	PM2.5	PM10	NO	NO2
Total stations	3323	2317	4727	3446	4933
No. stations > 75% valid data	1590	838	1946	1342	2205
Rejection threshold	500 $\mu g/m^3$	500 $\mu g/m^3$	800 $\mu g/m^3$	800 $\mu g/m^3$	600 $\mu g/m^3$
Threshold for concentration variation	200 $\mu g/m^3$	200 $\mu g/m^3$	300$\mu g/m^3$	350 $\mu g/m^3$	200 $\mu g/m^3$

Conclusions and future lines

The regionalization carried out through the clustering process and its subsequent characterization through the boxplots associated with each cluster, as well as the correlation graph of the mean series, reveals that the process of building the database for Ozone (the results for PM10, PM2.5, NO and NO2 are similar) has improved the observational data that were available at the beginning. The algorithm spontaneously detects patterns of similar temporal behavior.

In this respect, although the database improves the data gaps and makes it possible to distinguish different regions in Mediterranean science with similar behavioural patterns, there are still certain challenges that we need to face, i.e. there is the possibility of further improving the data density.

For this reason, a process of temporal reconstruction backwards in time is being worked on, to fill in missing data consistently. To this end, the **CAMS European air quality reanalyses** database. Interpolating in the locations where there are air quality measurement stations. The graph shows the high accuracy achieved for the specific case for a particular O3 station. The complete process for all contaminants is underway.

References

Baldasano J. M., Valera E., Jim énez P. (2003). Air quality data from large cities. Science of the Total Environment 307 (1-3), 141–165.

Schwela D. H., Haq G., et al. (2020). Strengths and weaknesses of the who global ambient air quality database. Aerosol and Air Quality Research 20(5), 1026–1037

Lorente-Plazas, R., Montavez, J. P., Jimenez, P. A., Jerez, S., Gomez-Navarro, J., Garcia-Valero, J. and Jimenez-Guerrero, P. (2015b), 'Characterization of surface winds over the iberian peninsula', INTERNATIONAL JOURNAL OF CLIMATOLOGY 35(6), 1007–1026. Article.