Analysis of non-linear air pollution time series

Kelyan Hangard
Semester project hosted by the Swiss Data Science Center

Spring semester 2023

Abstract

This study applies Information Theory methodologies to examine Nitrogen Dioxide (NO2) and fine particulate matter (PM2.5) air pollution across Europe. The comprehensive spatio-temporal time series analysis provides a deeper understanding of global and local properties of these pollutants, particularly noting key patterns and structures that inform our comprehension of air quality distribution. Through an exploratory data analysis, we determined certain features such as higher NO2 concentrations in Central Europe, seasonal and diurnal patterns, and notable variances between rural and urban areas. Semivariogram analysis was applied to determine spatial and temporal correlation. These semivariograms unveil significant distances and times beyond which the correlation between pollutant levels drops sharply. This insight offers a deeper understanding of pollutant propagation and their atmospheric lifetime. The Fisher Shannon analysis shed light on the predictability and structure of data, highlighting distinct differences between urban and rural areas. Furthermore, it identified an unusually high complexity in the distribution of pollutants in the Gran Canary Islands. Empirical orthogonal function analysis confirmed greater variability of distribution in Mediterranean countries and Atlantic Islands than in Central Europe. This research's findings aid in the development of more precise prediction models and form a significant contribution to environmental policy planning and public health strategies. Insights into the spatial and temporal dynamics of air pollution have the potential to inform more effective mitigation strategies, improving air quality management across Europe and beyond.

1 Introduction

Air pollution continues to pose a significant global challenge, affecting the quality of life, public health, and climate change. Among the many pollutants in the atmosphere, Nitrogen Dioxide (NO2) and fine particulate matter (PM2.5) are particularly hazardous to human health [1]. These pollutants are capable of penetrating deep into the respiratory system, causing a range of severe health problems[2], from respiratory irritation and asthma to cardiovascular and lung diseases[3]. The World Health Organization (WHO) has highlighted the urgency of addressing these pollutants due to their adverse impact on the environment and human health.

NO2 is primarily generated from combustion processes, particularly from vehicle engines and power plants that burn fossil fuels. It is a significant component of urban air pollution, contributing to the formation of photochemical smog. In rural areas, agricultural practices such as the use of fertilizers can also contribute to NO2 emissions. Exposure to NO2 can lead to respiratory problems, including reduced lung function and increased susceptibility to allergens [4].

PM2.5, or fine particulate matter, has a diameter less than 2.5 micrometers, making it small enough to be inhaled deep into the lungs and even enter the bloodstream. Major sources of PM2.5 include combustion processes such as those in motor vehicles, power plants, residential wood burning, and some industrial processes. Additionally, PM2.5 can be formed in the atmosphere from gases such as SO2 and NO2. Also, certain agricultural practices and the burning of biomass contribute to PM2.5 levels. Exposure to PM2.5 is linked to a variety of health effects, including respiratory and cardiovascular diseases, and premature death [5].

Understanding and monitoring air pollution is crucial in order to mitigate its harmful effects. Monitoring allows us to evaluate the effectiveness of policies aimed at reducing pollution, to predict future pollution levels, and to provide timely warnings to the public when pollution reaches dangerous levels. This process of monitoring and analysis requires extensive data collection, often taking the form of time series data, collected at various spatial and temporal scales.

For this study, we leveraged data from 6095 stations across Europe, which hourly measured pollutant concentrations at different periods between 2013 and 2019. The analysis of these time series is not straightforward due to their inherent complexity. They often exhibit non-linearities and non-stationarities, making traditional linear methods inadequate for understanding and predicting pollutant levels [6].

Recognizing these challenges, the aim of this work is to investigate advanced tools for spatio-temporal data exploration, to identify these non-linearities and non-stationarities in spatio-temporal pollution data. To this end, we utilized sophisticated measures such as variography, Fisher Information Measure, and Shannon Entropy Power.

This investigation into the changing distribution patterns is not merely an academic exercise; it also has practical applications. As we uncover these patterns, they can be used for feature engineering, enhancing the capacity of our spatio-temporal predictive models to account for these changes in data distribution. This added layer of complexity in the models would allows for more accurate temporal forecast and spatial interpolation, as they can adjust to shifts in pollution patterns.

The knowledge obtained from this analysis is expected to support policy makers in making informed decisions on how to manage pollutant sources more effectively, and consequently, reduce the health and environmental impacts of air pollution.

The remainder of this paper is structured as follows. Section 2 Materials and Methods introduces the data and methods used for this study. Section 3 Results and Discussion presents our main findings and the implications of these results. Finally, Section 4 concludes the paper by summarizing the key points and outlining potential future work. Appendices permits to understand further other details of the analysis.

2 Materials and Methods

2.1 Data collection and exploratory data analysis

The data for this study was provided by the European Environmental Agency, which maintains a vast network of 6095 monitoring stations throughout Europe. These stations offer excellent spatial resolution in tracking air pollution, with an inevitable higher density in urban areas compared to rural ones [7].

For each of these stations, time series data of two key pollutants were obtained: Nitrogen Dioxide (NO2) and particulate matter (PM2.5). These time series span six years, from 2013 to 2019, with measurements taken hourly, providing robust temporal precision. Each station's area type, classified as Urban, Suburban, Rural, Rural-Regional, or Rural-NearCity, was also recorded. This classification can provide valuable insights into the potential sources of pollutants, as it can be linked to different human activities characteristic of each area type. The location of the station and the mean value of the considered pollutant are shown in figure A1 of the Appendices section.

In total, the data for this study amounted to over 10 gigabytes, presenting significant challenges in data analysis. The data is not without its imperfections, with between 30% and 50% of missing data depending on the pollutant under consideration. For some plots of this study that couldn't support missing values, we chose

to focus on stations with less than 5% of missing values, linearly interpolating any missing points to maintain continuity in the data.

While this study focuses on NO2 and PM2.5, both recognized as among the most hazardous pollutants to human health, the analytical pipeline developed can be readily adapted to any pollutant of interest [8].

A preliminary exploratory data analysis (EDA) was conducted to visualize spatial and temporal trends in the data. Then, Hovmoller plots [9], have been used to jointly visualize spatial and temporal patterns of the data.

In the Hovmoller plot of NO2 concentration with respect to longitude and time (Figure 1 (b)), we notice a higher NO2 concentration in central Europe. Indeed, central Europe is notably dense with urban areas and industrial facilities, which contribute significantly to NO2 emissions [10]. The plot also shows the seasonality of NO2 concentration, which tend to be higher during winter months compared to summer. This could be attributed to several factors. During winter, the increased use of heating systems, especially those using fossil fuels, can elevate emission levels. Furthermore, the meteorological conditions during colder months, characterized by lower temperatures and less mixing in the atmosphere, can result in higher concentrations of NO2 [11]. A subtle decrease in NO2 concentrations over recent years is also noticeable in the Hovmoller plot. This might be the effect of some policies that have positive effect on pollutant concentrations [11].

The Hovmoller polts for NO2 and PM2.5 with respect to latitude and time (Figures 1 (a) and 1 (c)) show higher pollutant concentrations between degree 40 and 60 of latitude than in the extreme north part of Europe and in the extreme south parts of Europe. In the case of the northern regions of Europe, the lower pollutant concentration might be attributed to a lesser degree of industrialization and a lower population density. Much of Northern Europe, particularly regions within the Arctic circle, are sparsely populated with limited industrial activity, resulting in lower emission of pollutants like NO2 and PM2.5[12]. Furthermore, the strong winds common in these regions could aid in the dispersion of pollutants, preventing accumulation and reducing local concentrations. Regarding Southern Europe, the lower pollutant concentration could be influenced by prevailing meteorological conditions, including wind patterns and atmospheric dispersion properties which may contribute to a faster dilution and dispersion of pollutants[13]. Furthermore, there isn't any big city located below the latitude 40, which also strongly encourages the decrease of PM2.5 and NO2 concentration.

Considering the Hovmoller plot of PM2.5 concentration with respect to longitude and time (Figure 1 (d)), it seems that PM2.5 concentrations are higher in eastern Europe than in western Europe, providing an interesting contrast to the spatial distribution of NO2. The higher PM2.5 concentration observed in eastern Europe might be attributable to the region's reliance on solid fuels for heating and cooking, as well as industrial activities that generate particulate matter [14].

Further examination of our data can be found in the appendices, where we present Hovmoller plots that study the average percentage of missing values in NO2 and PM2.5 concentration data with respect to time and geographical coordinates, both latitude and longitude (see Appendix, Figure A5). This analysis provides additional insights into the completeness and reliability of our dataset across space and time.

These initial findings not only provide a glimpse into the spatial and temporal patterns of NO2 and PM2.5 concentrations but also underscore the influence of human activity and seasonal variations on air pollution.

2.2 Advanced exploration of spatio-temporal data

Traditional exploratory data analysis (EDA) tools often fall short in their capacity to fully capture the intricate interplay of spatial and temporal trends [15]. Thus, to delve deeper into the spatio-temporal correlations and evolving distribution patterns within our data, we employed a more advanced analytical approach.

A key aspect of this advanced exploration is the study of changes in data distribution over time and space [16]. This allows us to understand the level of predictability of our data, the presence of non-linearity and non-stationarity, and how these factors evolve over different spatial and temporal scales. Non-linearity refers to the complex

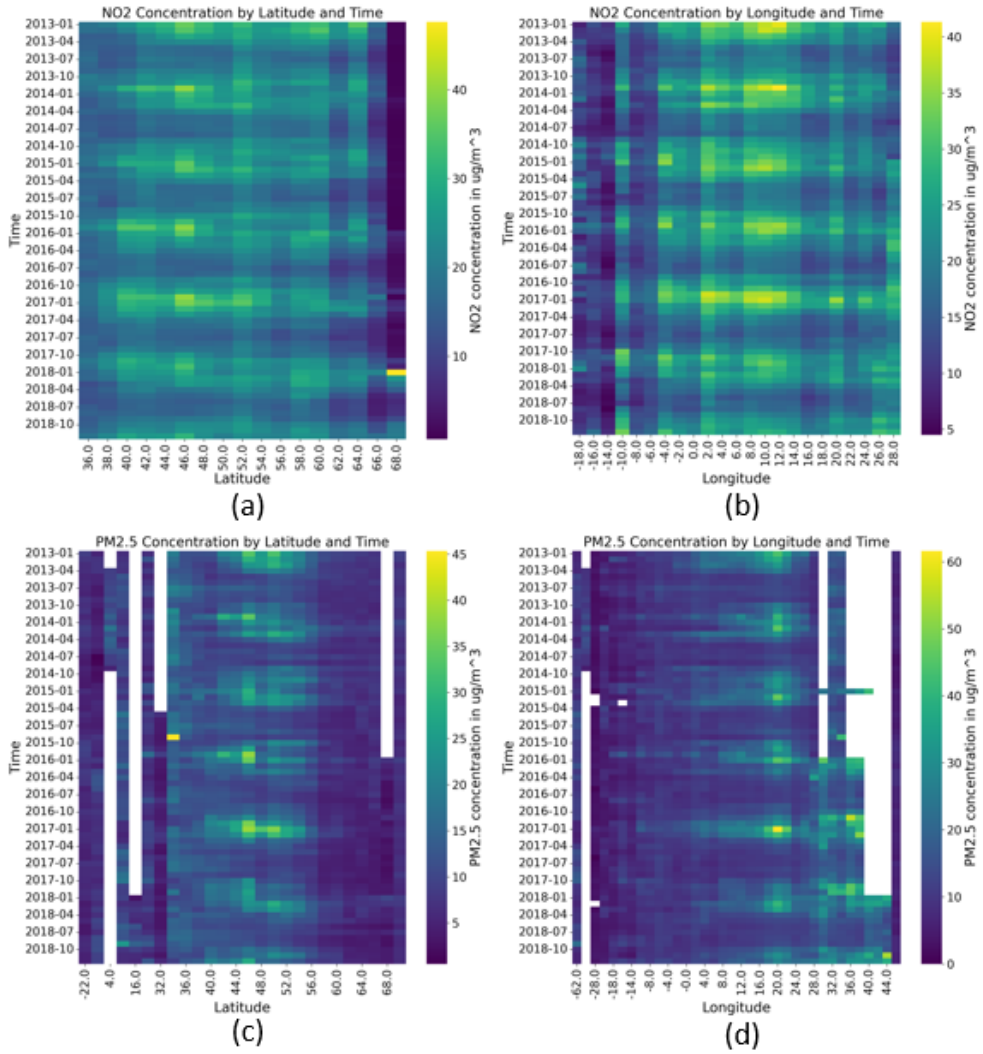


Fig. 1: **Hovmoller plots of pollutant concentration data:** Hovmoller plots for NO2 (top row) and PM2.5 (bottom row) average concentration data with respect to latitude (left column) and longitude (right column)

dependencies between variables that cannot be represented by a straight line, while non-stationarity suggests that the statistical properties of the data, such as mean and variance, change over time.

Investigating the advanced spatio-temporal structures in our data provides a more nuanced understanding of pollution dynamics, unveiling the intricate mechanisms that drive changes in air quality. This insight is fundamental in informing more effective strategies for air pollution management and control.

2.2.1 Spatio-temporal variography

Understanding the joint spatio-temporal dependence structure of a spatio-temporal process is crucial for optimal prediction, such as in the case of kriging and gaussian processes. To gain a measure of this joint spatio-temporal dependence, we employ the tool of empirical spatio-temporal semivariograms[9].

The semivariogram provides a mathematical function that quantifies the degree of spatial and temporal dependence between data points. Let Z be a univariate continuous random variable, in our case, this random variable will be pollutant concentration It's formally defined as:

$$\gamma_z(s_i, s_k; t_j, t_l) = \frac{1}{2}\text{var}(Z(s_i; t_j) - Z(s_k; t_l)) \tag{1}$$

Where s_i or s_k and t_j or t_l denote spatial and temporal locations, respectively. In situations where the covariance depends solely on differences in space lag ($h = s_k - s_i$) and time lag ($\tau = t_l - t_j$), the semivariogram can be expressed as:

$$\gamma_z(h;\tau) = \frac{1}{2}\text{var}(Z(s+h;t+\tau) - Z(s;t)) = C_z(0;0) - C_z(h;\tau) \tag{2}$$

where $C_z(0;0)$ is the variance of the process, and $C_z(h;\tau)$ is the covariance function, representing the expected degree of similarity between two data points separated by a spatial lag h and a temporal lag τ.

Nonetheless, there may be cases where the semivariogram γ_z is a function of h and τ, but a stationary covariance function $\gamma_z(h;\tau)$ does not exist. To circumvent these models of covariability, we typically fit trend terms that are linear or quadratic in the spatio-temporal coordinates.

If the covariance function of the process is well-defined, the semivariogram typically displays three key characteristics: the nugget effect, the sill, and the partial sill. The nugget effect represents $\gamma_z(h;\tau)$ as both $\mathbf{h}$ and τ approach 0, while the sill corresponds to $\gamma_z(h;\tau)$ when both $\mathbf{h}$ and τ tend towards infinity. The partial sill is defined as the difference between the sill and the nugget effect.

The employment of spatio-temporal semivariograms in our study is interesting because they provide a statistical measure of the variability in our data across both space and time. It enables us to understand and quantify the complex spatio-temporal dependence structure inherent in our dataset, thereby enhancing our ability to create accurate predictive models.

2.2.2 Fisher Information Measure and Shannon Entropy Power

Information theory is a good way to study structure and predictability levels of the data. In order to study these features, we delve into the estimations of Fisher Information Measure (FIM) and Shannon Entropy Power (SEP). In brief, FIM serves to assess the information contained in a signal X, while SEP gauges its degree of disorder and predictability [17].

FIM and SEP act as summary quantities that partially describe a Probability Density Function (PDF). Let X be a univariate continuous random variable following a PDF f(x), in our case, this random variable will be pollutant concentration. The SEP of X is defined as SEP via the relationship:

$$SEP = \frac{1}{2\pi e}e^{2H_X}, \quad H_X = E[-\log f(x)], \tag{3}$$

where H_X is the differential entropy of X. The FIM of X, denoted FIM, is defined as:

$$FIM = E\left[\left(\frac{\partial}{\partial x}\log f(X)\right)^2\right]. \tag{4}$$

FIM and SEP can be visualized together in the Fisher-Shannon Information Plane (FSIP) to study the PDF of X. It can be shown that $SEP \cdot FIM \geq 1$, with equality satisfied if and only if X is a Gaussian random variable. Therefore, the only reachable points in the FSIP belong to the set:

$$D = (SEP, FIM) \in \mathbb{R}^2 | SEP > 0, FIM > 0, SEP \cdot FIM \geq 1 \tag{5}$$

The quantity $FSC = SEP \cdot FIM$, named the Fisher-Shannon Complexity (FSC), is occasionally used as a statistical complexity measure. The FSC can be interpreted as a measure of non-Gaussianity of X. Thus, the boundary of D is reached if and only if X has a unitary FSC, in which case it is a Gaussian random variable.

To estimate SEP and FIM from data, $f(x)$ and its derivative $f'(x)$ are replaced by their kernel density estimators (KDE) in the integral forms. This is accomplished by approximating the PDF f(x) as:

$$\hat{f}_{h(x)} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{6}$$

Where $h(x)$ is a bandwidth parameter and $K()$ is a kernel assumed to be a unimodal probability density function symmetric around zero with integral over $\mathbb{R}$ equal to 1. By using a Gaussian kernel with zero mean and unit variance, the estimator takes the form:

$$\hat{f}_{h(x)} = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right). \tag{7}$$

The estimates are sensitive to the choice of a proper bandwidth. In this work, this parameter is selected using the Sheather-Jones direct plug-in method, which approximates the optimal bandwidth with respect to the Asymptotic Mean Integrated Squared Error of $\hat{f}$. Operationally, the non-parametric estimation of the SEP, FIM and FSC are obtained with the FiShPy package.

2.2.3 Empirical orthogonal function

To further elucidate the spatial patterns of the spatio-temporal FIM and SEP values calculated, an Empirical Orthogonal Function (EOF) decomposition will be performed on the data of the two estimations. EOF analysis, as the spatio-temporal equivalent of principal component analysis, allows data to be deconstructed into orthogonal basis functions - the Empirical Orthogonal Functions (EOFs) - and temporal coefficients, which are detailed by the principal components (PC) time series [9].

Let $X_{tj} = [X(s_1; t_j), ..., X(s_m; t_j)]' \in \mathbb{R}^m$ be observations for the spatial locations $s_i : i = 1, ..., m$ and times $t_j : j = 1, ..., T$. The empirical spatial covariance matrix can then be computed as:

$$\hat{C} = \frac{1}{T} \sum_{j=1}^{T} (X_{tj} - \hat{\mu})(X_{tj} - \hat{\mu})'. \tag{8}$$

As this real matrix is symmetric and non-negative definite, it can be spectrally decomposed as:

$$\hat{C} = \Phi \Lambda \Phi' \tag{9}$$

where $\Lambda = diag(\lambda_1, ..., \lambda_m)$ is the diagonal matrix of the non-negative eigenvalues decreasing down the diagonal, and $\Phi = (\Phi_1, ..., \Phi_m)$ is the matrix of the corresponding spatially indexed eigenvectors $\Phi_k = (\Phi_k(s_1), ..., \Phi_k(s_m))'$, for k=1,...,m also called EOFs. The EOFs form a discrete orthonormal basis. The k-th PC time series which is the time series of coefficient of the corresponding EOF, or equivalently the contribution of the k-th spatial basis at time t_j is then given by $a_k(t_j) = \Phi_k' X_{tj}$.

3 Results and discussion

3.1 Variography

The empirical spatio-temporal semivariograms of NO2 in figure 2a and PM2.5 in figure 2b, each computed with a lag of 12 hours, offer valuable insights into the dependence structure of the pollutants' concentrations over both space and time.

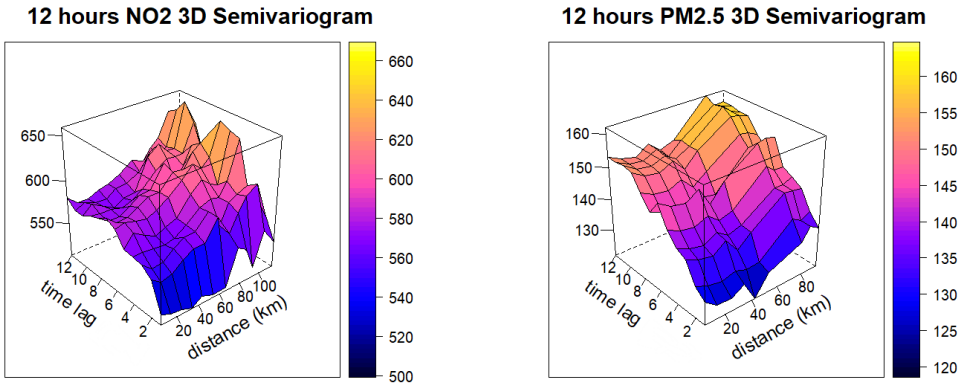
For both NO2 and PM2.5, we observe an increase in semivariogram values with respect to space and time, which aligns with our expectations. This suggests that the spatial and temporal correlations between pollutant measurements decrease as the separation in space (distance h between stations) or time (time τ between measurements) increases. This knowledge is essential for constructing accurate prediction models and making informed decisions about pollutant control. We surprisingly notice that the semivariogram value is not clearly increasing with time when the time lag is 0 hours. This suggests that there is no spatial structure in the data (for NO2 and PM2.5) and that it will then be difficult to fit spatial interpolation models.

However, if we look at the semivariogram values with respect to the space lag for higher time lags, we can see some trends. In the case of PM2.5 (figure 2b), we observe a sharp increase in the semivariogram value around the 70-kilometer mark on the spatial axis. This infers that the spatial correlation of PM2.5 concentrations between stations

rapidly decreases beyond this distance. A potential explanation for this behavior might be that PM2.5 particles, being small and lightweight, are susceptible to atmospheric dispersion and dilution processes over longer distances [18]. We can draw exactly the same conclusion for the NO2 semivariogram (figure 2a) but the space lag where we see this difference in the semivariogram value is more around 60 kilometers.

In Figure 2b, an abrupt rise in the PM2.5 semivariogram value after a time lag of 5 hours can be seen, suggesting that PM2.5 particulates might typically stay for around 5 hours in the atmosphere. This can obviously vary with respect to meteorological conditions, such as wind speed, precipitation, and temperature that can affect the atmospheric lifetime of PM2.5. Moreover, during periods of low wind speed and stagnant air, these particles can remain suspended in the atmosphere for several hours [19]. However, more investigations are needed to validate this hypothesis.On the other hand, while we observe an increase in the semivariogram value of NO2 over time in Figure 2a, no distinctive temporal patterns emerge from the data.

The examination of these semivariograms reinforces our understanding of the spatio-temporal dependence structure of NO2 and PM2.5 concentrations. The obtained insights can inform more effective pollution control strategies and allow for more precise predictive modelling.



(a) (b)

Fig. 2: **12 hours pollutant spatio-temporal semivariograms:** Semivariogram with respect to time and distance based on NO2 concentration data (a) and PM2.5 concentration data (b)

3.2 Information theory

The Fisher Information Measure (FIM) and Shannon Entropy Power (SEP) were utilized to gain a more nuanced understanding of the distribution patterns of NO2 and PM2.5 concentrations. These measures were computed for each monitoring station using a one year moving window on pollutant concentration data. Measures are displayed in the presented Fisher-Shannon information planes in figure 3, delineating the distribution complexities for both pollutants based on data from the first year of monitoring (2013) and the most recent year (2018).

One striking observation from these plots is the proximity of most stations FIM and SEP product to the value of 1, a result that signifies that the distribution of pollutant concentrations at these stations are closely approximating Gaussian distributions. This outcome suggests that these distributions demonstrate balanced complexity in terms of structure and predictability.

Interestingly, the overall distribution pattern of points in the Fisher-Shannon information plane remains largely consistent between 2013 and 2018 for both NO2 and PM2.5, implying that the temporal change in pollutant concentrations does not considerably alter their distribution complexities. This could be indicative of consistent emission sources and regulatory influences within the monitored areas over the five-year period.

Examining the NO2 plots, a discernible difference between urban and rural monitoring stations emerges. Urban stations generally exhibit higher SEP and lower FIM, indicating that the NO2 concentrations in these areas possess lower predictability and structure, which could be attributed to a complex mix of variable emission sources and traffic patterns prevalent in urban environments.

On the contrary, rural stations display a more varied range. Some rural stations align with urban ones, exhibiting low predictability and structure. However, a significant portion of rural stations register notably low SEP and high FIM values, suggesting that these locations demonstrate predictable and structured NO2 distributions. This intriguing divide in the rural dataset could potentially be linked to the varying agricultural practices and related policies across different regions, contributing to differential emission profiles.

For PM2.5, discerning clear trends within the Fisher-Shannon information planes proves challenging due to the overlapping distributions from stations located in diverse areas. This may suggest that PM2.5 emission sources and dispersion patterns are more homogeneous across urban and rural areas, leading to similar distribution complexities.

These insights pave the way for more targeted air quality monitoring strategies, highlighting the value of information theoretic tools in the context of environmental data analysis. Nonetheless, the observed trends and hypotheses warrant further investigation for comprehensive understanding and validation.

In the figure 4 showing FSC time-series of station located in the extreme corners of their Fisher-Shannon information planes, we notice again that some rural stations have much higher FSC values than urban stations. This difference is important here because the y axis is log scale. We can notice some interesting peaks and drops in these FSC time-series. We especially have a big drop in the rural PM2.5 FSC time-serie and in the urban PM2.5 FSC time-serie around the begining of 2017. This complexity drop might have a lot of different causes that we don't have time to investigate in this report, but this would be an intresting subject to study in future studies. The four stations locations are presented in the map in figure 5. We ca notice that the two urban stations located at the bottom right corner of the NO2 and PM2.5 Fisher-Shannon information plane are both located in the United Kingdom, one near London and one in Leicester. It is interesting that the two stations with the lowest level of predictability and structure in the NO2 or PM2.5 data are both located in the same country, in urban areas, not far from each other.

We can see the two stations location with highest mean NO2 and PM2.5 FSC in the map represented in figure 5 (green points). It is very interesting to see that the station with highest mean FSC based on NO2 concentration data and the station with highest mean FSC based on PM2.5 concentration data are both located in very rural areas in Las Palma de Gran Canaria islands while stations located outside continental

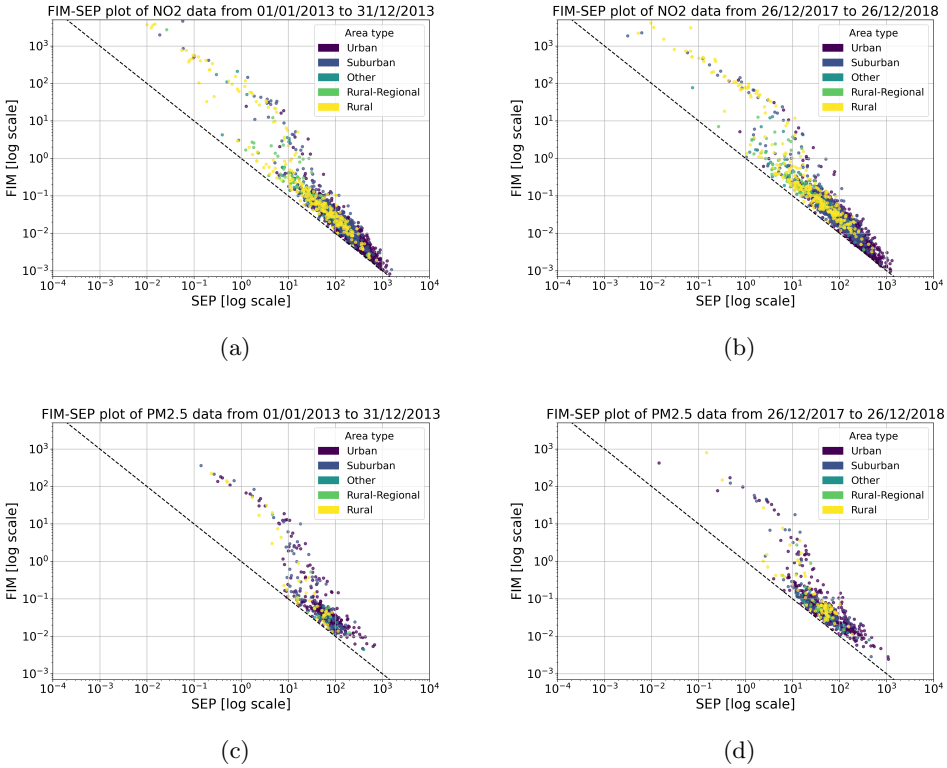


Fig. 3: **Fisher-Shannon information planes of different areas and time windows:** Fisher-Shannon information planes representing FIM and SEP values of each station. The FIM and SEP values are computed using the first year of data (left column) or the most recent year of data (right column) of NO2 concentration data (top row) or PM2.5 concentration data (bottom row). Each color associated to a point corresponds to the type of area in which the station is located. The dashed line correspond to the boundaries where the FIM and SEP product is equal to 1. Points located on this boundary have Gaussian distribution.

Europe represent a tiny percentage of all the stations. The high FSC indicates that the pollutant distributions at these stations are significantly complex, challenging to predict, and deviate noticeably from Gaussian distributions.

There are a few possible reasons that could explain this finding:

- **Geographical Factors:** Gran Canaria, as an island, is significantly influenced by maritime air masses, which can introduce a high degree of variability in the concentrations of pollutants like NO2 and PM2.5. The effect of wind and ocean currents, which vary depending on seasons and weather patterns, can lead to irregular and unpredictable changes in air pollution levels[20].
- **Population and Tourism:** The Gran Canaria Islands are a popular tourist destination, experiencing significant seasonal population flux. This variability in population density, coupled with activities related to tourism (like increased vehicular traffic, energy use, etc.), may cause unpredictable shifts in pollutant concentrations[21].
- **Industrial Activities:** While Gran Canaria isn't highly industrialized, there are certain activities such as shipping, waste processing, and some localized industries that can contribute to air pollution. These activities can lead to sporadic increases in pollutant levels[22].
- **Topography:**The island's diverse topography, with a combination of mountains, valleys, and coastal areas, could also contribute to variability in air pollutant distributions. The complex terrain can influence wind patterns and atmospheric dispersion, leading to a non-uniform and complex spread of pollutants[23].

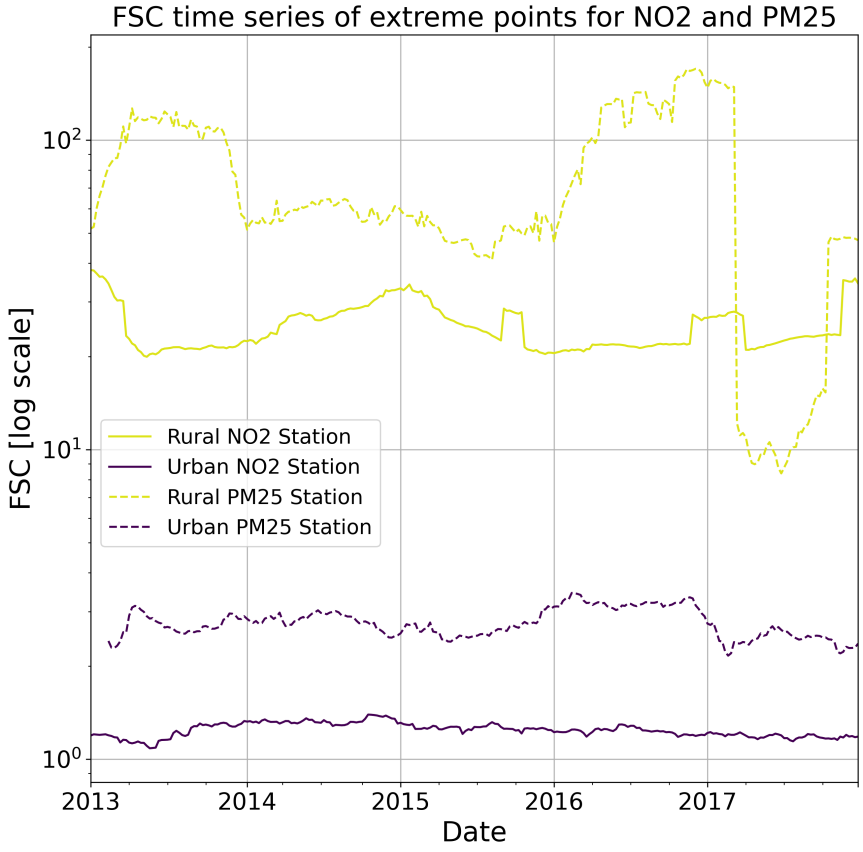


Fig. 4: **Fisher Shannon Complexity time-series:** Fisher Shannon Complexity of four station. The rural NO2 station (yellow line) and the rural PM2.5 station (yellow dashed line) correspond to two stations located on the top left extreme of their respective Fisher-Shannon information plane. The urban NO2 station (purple line) and the urban PM2.5 station corresponds to two stations located on the bottom right extreme of their respective Fisher-Shannon information plane. Each FSC value is computed using pollutant concentration value of the next year.

- **Microclimatic Conditions:** Islands like Gran Canaria often have unique microclimates which can vary significantly over short distances. This can cause diverse and unpredictable local weather patterns that significantly affect pollutant dispersion and concentration[24].

These reasons might contribute to the high complexity and lower predictability in the air pollutant distributions observed at the stations on the Gran Canaria Islands. Nonetheless, these hypotheses should be investigated further in the context of local emission sources, meteorological conditions, and other relevant factors to confirm their validity and importance.

In figure 6, we present an intricate exploration of the Fisher Shannon Complexity (FSC) of the stations that exhibit the highest mean FSC based on the NO2 and PM2.5 concentration data, respectively.

Our analysis of Fisher Information Measure (FIM) and Shannon Entropy Power (SEP) time trajectory for the station with the highest mean Fisher Shannon Complexity (FSC) based on NO2 data reveals intriguing patterns (figure 6a and figure 6b). The FIM and SEP values predominantly align along a curve where their product appears to be constant. This suggests a stationary behavior of the NO2 data. The corresponding time-series graph (figure 6b) further corroborates this, showing that the

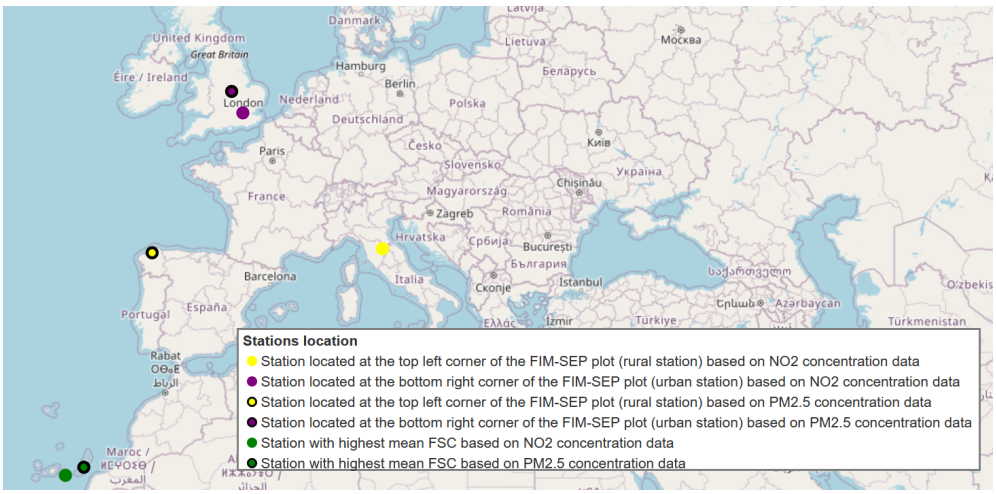


Fig. 5: **Interesting stations location:** Europe map with markers representing stations that are located in extreme corners of their NO2 (circles without border) or PM2.5 (circles with black border) Fisher-Shannon information planes (purple and yello circles), or stations with highest mean FSC based on NO2 (green circle) or PM2.5 (green circle with black border).

FSC primarily ranges between 90 and 150. Interestingly, we observe that both the significant drop and peak in the FSC time series coincide with the maximum occurrence of missing NO2 values for this station (figure B7a). These fluctuations could be influenced by these missing data points.

In contrast, when we consider the time trajectory of the FIM and SEP values for PM2.5 for the station with the highest mean FSC (figure 6c and figure 6d), we notice a distinct pattern. Instead of a constant product, the FIM and SEP values seem to move towards the origin over time, signaling a non-stationary behavior in the PM2.5 data. A more detailed look at the FSC time series (figure 6d) reveals a notable decrease between mid-2016 and mid-2017. This drop aligns with the period of the highest percentage of missing PM2.5 data for this station (figure B7b in Appendices). Therefore, this non-stationarity and decrease in FSC could be attributed to the significant amount of missing data during this period. This drop in the Fisher Shannon complexity is very interesting, this means that a machine learning model that trained on this station data from 2013 to 2015 will probably fail to forecast the PM2.5 concentration from 2015 to 2018 since the pollutant distribution is drastically changing between these two periods.

In figure 7, it's apparent that the spatial and temporal average FSC values vary based on geographical location. These hovmoller plots were computed on data from locations with at least five measurement stations to avoid outliers.

The plot 7a specifically shows a concentration of higher average FSC values for NO2 in central Europe (latitude degrees between 25 and 45). This suggests more complex NO2 emission distributions in this region, possibly due to increased industrial activities or vehicular emissions.

On a similar note, the plot 7b indicates that NO2 measurement stations located in the western part of Europe (longitude between -20 and -15) record higher mean FSC values than those in other regions of Europe.

Interestingly, the same spatial pattern is observed for PM2.5 measurements. As per figure 7d, the western part of Europe also registers the highest FSC values for PM2.5.

Overall, these Hovmoller plots enable a comprehensive understanding of the temporal and spatial variability of FSC associated with NO2 and PM2.5 emissions in Europe. This, in turn, could potentially help in developing targeted strategies for air pollution control and mitigation.

You can access to the missing values data corresponding to these Hovmoller plots in the Appendices section, if you look at the figure B8.

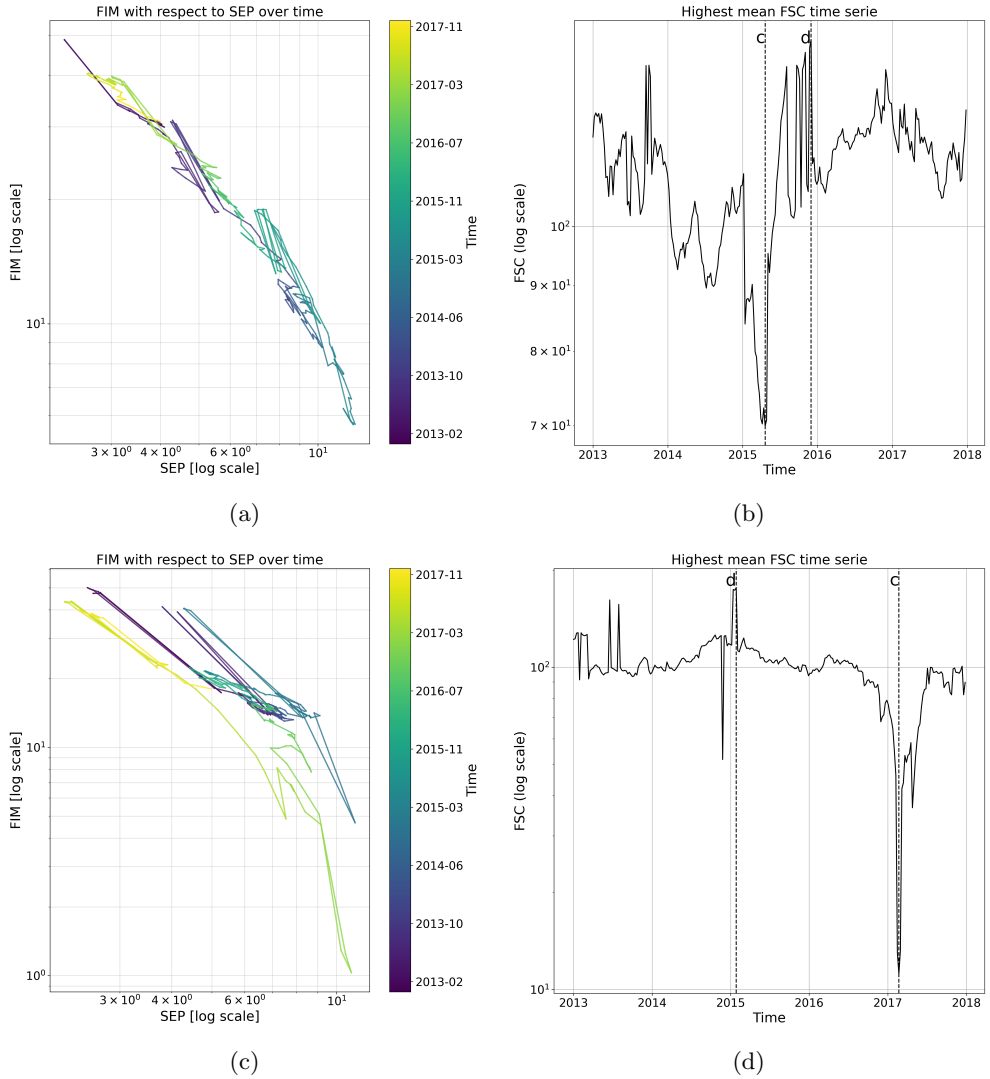


Fig. 6: **Fisher Shannon analysis of the station with highest mean FSC :** (a) represents the FIM-SEP graphic of the highest mean NO2 FSC time-serie. This highest mean NO2 FSC time-serie is represented in (b), the minimum and maximum values of this time-serie are annotated with the points c and d. (c) represents the FIM-SEP graphic of the highest mean PM2.5 FSC time-serie. This highest mean PM2.5 FSC time-serie is represented in (c), the minimum and maximum values of this time-serie are also annotated with the points c and d.

3.3 Empirical orthogonal function

Since we are applying principal component analysis on spatio-temporal data (FIM and SEP), we obtain spatial coefficients and temporal bases. The figure 8 shows the first PC temporal bases and spatial coefficients for the SEP and FIM of the NO2 concentration data. We can notice a positive trend between the first PC bases of SEP and the time (figure 8a). We observe the same positive trend for FIM first PC bases (figure 8c). Furthermore, we observe higher first PC coefficients for FIM and SEP in the south of Europe (Spain, Italy, Greece) than in central Europe (Germany, France, UK) as you can see in figure 8b and 8d. It means that the unpredictability levels, structure levels and then the complexity of NO2 distribution in Mediterranean countries are more increasing than in central Europe countries. It can be due to several factors:

- **Meteorological conditions:** Factors such as high temperatures and increased sunlight, typical of Mediterranean climates, can heighten photochemical reactions and suddenly increase NO2 levels[25]. Similarly, diverse coastal wind patterns, seasonal

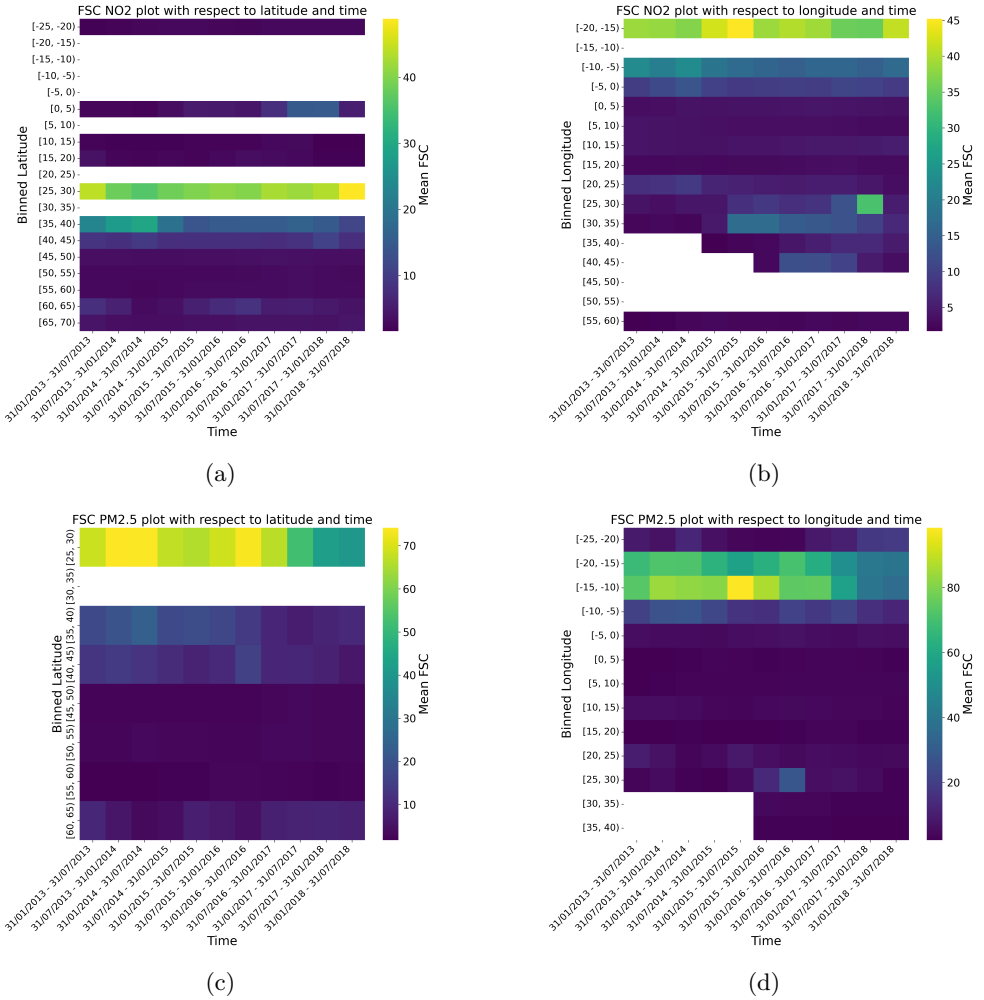


Fig. 7: **Hovmoller FSC plots:** Hovmoller plots representing the average FSC computed over NO2 (first row) or PM2.5 (bottom row) time series with respect to latitude (left column) or longitude (right column) and time

precipitation, and atmospheric stability variations in the Mediterranean can contribute to the unpredictability of NO2 distribution. On the other hand, Central European countries, characterized by more stable weather conditions, less extreme temperatures, and consistent wind patterns, might display more predictable and less variable NO2 distributions.

- **Geographical Factors:**The topographical diversity of the Mediterranean region, with its mix of coastal areas, inland regions, mountains, and valleys, can influence the dispersion and concentration of air pollutants. Complex terrain can lead to varying wind patterns that influence where pollutants are carried[26], and mountains can trap pollutants in certain areas, leading to higher concentrations.
- **Economic Activities:** The Mediterranean region is known for its tourism, agriculture, and industry, all of which can contribute to fluctuations in air pollution[27]. For instance, tourism-related activities tend to peak in the summer, possibly leading to increased pollution from transportation and energy use.

The same analysis as been conducted on PM2.5 data in the figure 9. We needed to reduce the number of station where we performed the PCA because the main part of PM2.5 stations have too much missing values. Trends are then less clear for the PM2.5 PCA. However we can still notice that highest PC coefficients are located in the Canary islands (figure 9b) and that there exist a positive trend between SEP bases and time (figure 9a). This means that Canary islands PM2.5 distributions unpredictability is strongly increasing with time (it is also the case for Canary islands NO2 distributions). The reasons of this high SEP values in the Canary islands have been explained before.

More globally, in figures 8a, 8c and 9a, we observe a gradual increase over time of the Fisher Shannon complexity of NO2 and PM2.5 concentration distributions in Europe. This positive trend suggests that NO2 and PM2.5 concentration distributions are progressively deviating from Gaussian behavior, becoming more complex over time.

Several factors could be contributing to this phenomenon, where pollutant distributions are progressively becoming less Gaussian:

- **Urbanization and Industrial Growth:** As urban areas expand and industries grow, they can lead to increased emissions from vehicles, factories, and power plants. This increasing pollution output may contribute to the complexity of pollutant distribution over time[28].
- **Changing Traffic Patterns:** Over time, changes in the volume, flow, and types of vehicles on the road, as well as advancements in engine technologies[11], can alter emission patterns, thus affecting the structure of pollutant distributions.
- **Climate Change:** Rising global temperatures[29] may amplify photochemical reactions, leading to an increase in NO2 and PM2.5 levels. This shift, in turn, could alter the distribution patterns of pollutants, causing them to deviate from typical Gaussian distributions.

Regarding the PM2.5 FIM bases in figure 9c, we notice that the first principal component bases of FIM is varying interestingly, with a step between the end of 2014 and the end of 2015, and a big drop in the beginning of the year 2017 until the end. These particular behaviors would need to be investigated in another project.

You can find the second component PCA of NO2 and SEP in Appendices (figure C9 and C10)

4 Conclusion

This study explored the utility of methods derived from Information Theory for examining surface air pollution, the largest environmental health risk in Europe and in the World. We discovered that applying the three measures, FIM, SEP, and FSC, could generate significant understanding of the global and local characteristics of the referenced spatio-temporal time series. In particular, we successfully identified spatio-temporal structures within the data, revealing patterns that would otherwise be challenging to discern.

In conclusion, this study has provided valuable insights into the spatial and temporal characteristics of NO2 and PM2.5 air pollutants across Europe, with an emphasis on understanding their distributions, predictability, and structure using a suite of statistical tools rooted in Information Theory.

Through our Exploratory Data Analysis, we identified several key features, such as higher NO2 concentrations in Central Europe, seasonal fluctuations in both NO2 and PM2.5 levels with higher concentrations observed during winter months, and diurnal patterns showing higher concentrations during mornings, evenings, and weekdays. We also highlighted high PM2.5 concentrations in Eastern Europe, as well as a noticeable decline in NO2 levels over recent years.

Further, our semivariogram analysis suggested that the spatial and temporal correlation between pollutant levels decrease significantly beyond certain distances and times: roughly 60 kilometers and 70 kilometers for NO2 and PM2.5 respectively, and approximately 5 hours for PM2.5. This indicates the potential extent of pollutant propagation and the duration pollutants remain in the atmosphere.

Our Fisher Shannon analysis revealed interesting differences in pollutant concentration distributions between rural and urban areas. Rural NO2 stations tended to have high predictability and structure in their data, contrasting sharply with the low predictability and structure observed for urban stations. This analysis also pinpointed higher average FSC values for NO2 and PM2.5 in Central Europe, with notable outliers in Gran Canaria Islands having the highest mean FSC values. Conversely, the two stations with the highest SEP and lowest FIM were found near London, indicating significantly low predictability and structure in these regions.

Finally, our empirical orthogonal function analysis demonstrated that predictability and structure levels of NO2 and PM2.5 distribution in Mediterranean countries

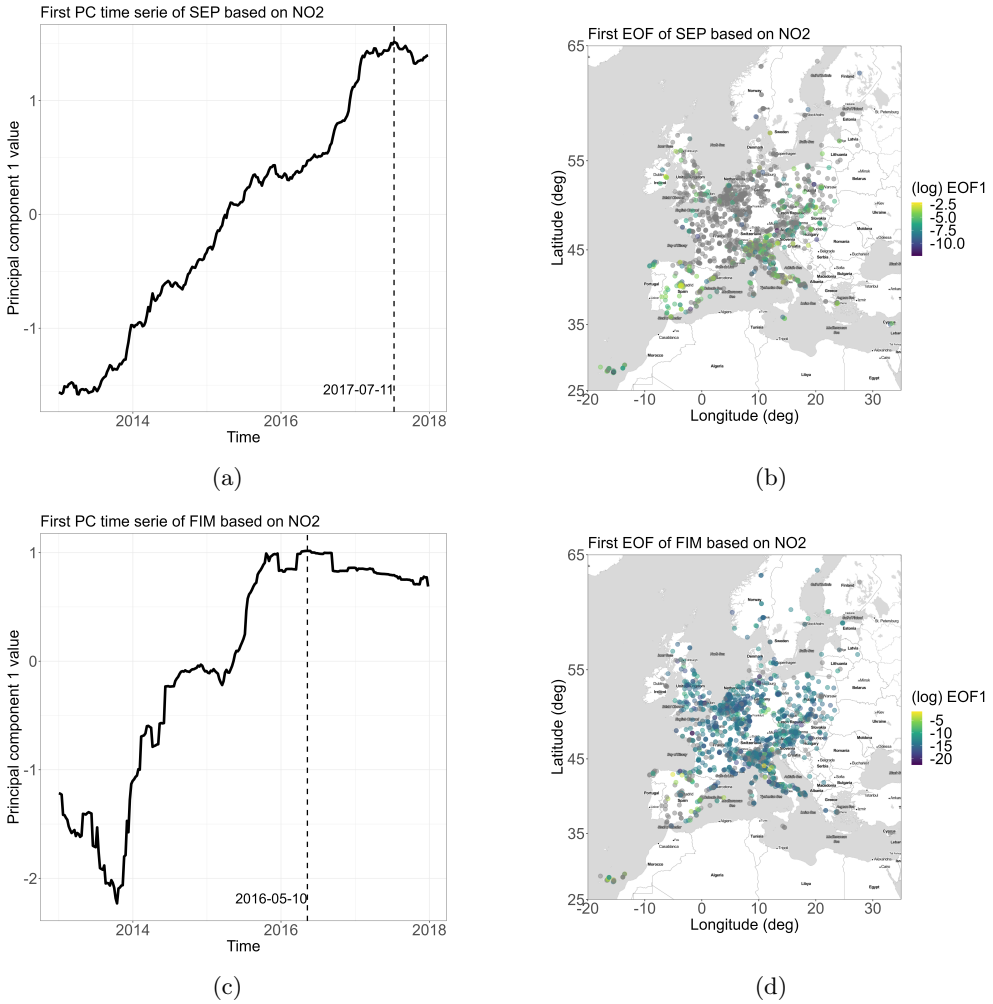


Fig. 8: **First principal component analysis of SEP and FIM based on NO2 concentration data:** In (a) and (c), we see the first principal component temporal basis computed over the SEP (a) or FIM (c) based on NO2 concentration data with respect to time. A dashed line referenced the date corresponding to the maximum value of this time serie. In (b) and (d), we can analyze the first spatial principal component coefficients map of SEP (b) and FIM(d) based on NO2 concentration data. The logarithmic of this first principal component coefficients is represented through colors of the points in the Europe map

and Atlantic Islands show more variability and are more increasing than in Central Europe.

Additionally, this exploration into the evolving patterns of distribution is not simply an intellectual pursuit; it bears practical significance too. By figuring out these patterns, we can use them to improve our prediction models, helping them to adapt to changes in how pollution is distributed over space and time. This extra detail in the models means they can give more accurate predictions by adjusting to shifts in pollution patterns.

The findings of this study are highly relevant as they provide a deeper understanding of the complex spatial and temporal dynamics of air pollution, which is crucial for environmental policy planning and public health strategies. By harnessing the power of Information Theory, we can better characterize and predict air pollution trends, contributing to more effective mitigation strategies and improved air quality management in Europe and beyond.

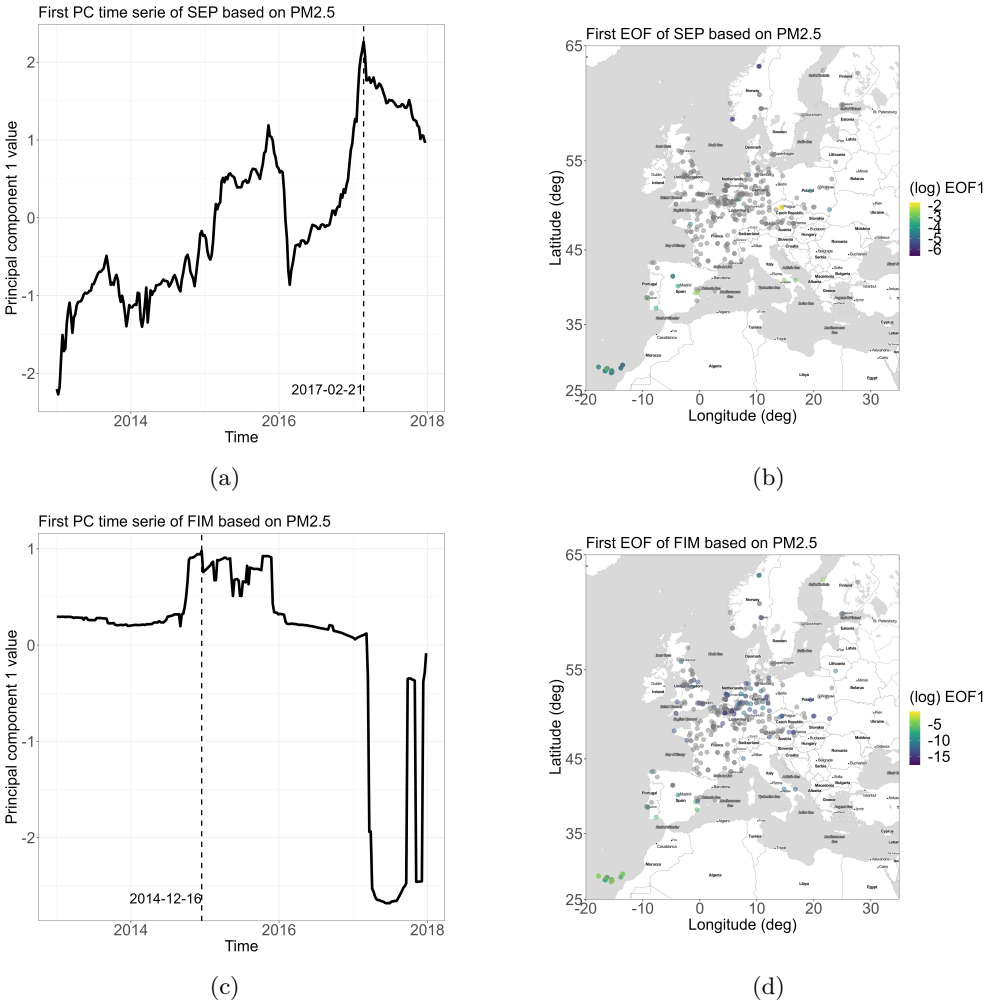


Fig. 9: **First principal component analysis of SEP and FIM based on PM2.5 concentration data:** In (a) and (c), we see the first principal component temporal basis computed over the SEP (a) or FIM (c) based on PM2.5 concentration data with respect to time. A dashed line referenced the date corresponding to the maximum value of this time serie. In (b) and (d), we can analyze the first spatial principal component coefficients map of SEP (b) and FIM(d) based on PM2.5 concentration data. The logarithmic of this first principal component coefficients is represented through colors of the points in the Europe map

References

[1] Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., *et al.*: Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. The lancet **389**(10082), 1907–1918 (2017)

[2] Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z.J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., *et al.*: Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 european cohorts within the multicentre escape project. The lancet **383**(9919), 785–795 (2014)

[3] Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope III, C.A., Apte, J.S., Brauer, M., Cohen, A., Weichenthal, S., *et al.*: Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter.

Proceedings of the National Academy of Sciences **115**(38), 9592–9597 (2018)

[4] Beelen, R., Hoek, G., Den Brandt, P.A., Goldbohm, R.A., Fischer, P., Schouten, L.J., Jerrett, M., Hughes, E., Armstrong, B., Brunekreef, B.: Long-term effects of traffic-related air pollution on mortality in a dutch cohort (nlcs-air study). Environmental health perspectives **116**(2), 196–202 (2008)

[5] Pope Iii, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D.: Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama **287**(9), 1132–1141 (2002)

[6] Jhun, I., Coull, B.A., Schwartz, J., Hubbell, B., Koutrakis, P.: The impact of weather changes on air quality and health in the united states in 1994–2012. Environmental research letters **10**(8), 084009 (2015)

[7] Tamburlini, G., Ehrenstein, O.S., Bertollini, R.: Children's Health and Environment: a Review of Evidence: a Joint Report from the European Environment Agency and the WHO Regional Office for Europe vol. 1. Office for Official publications of the European communities, ??? (2002)

[8] Thurston, G.D., Kipen, H., Annesi-Maesano, I., Balmes, J., Brook, R.D., Cromar, K., De Matteis, S., Forastiere, F., Forsberg, B., Frampton, M.W., et al.: A joint ers/ats policy statement: what constitutes an adverse health effect of air pollution? an analytical framework. European Respiratory Journal **49**(1) (2017)

[9] Wikle, C.K., Zammit-Mangion, A., Cressie, N.: Spatio-temporal Statistics with R vol. 1. CRC Press, ??? (2019)

[10] Beekmann, M., Vautard, R.: A modelling study of photochemical regimes over europe: robustness and variability. Atmospheric Chemistry and Physics **10**(20), 10067–10084 (2010)

[11] Grange, S.K., Lewis, A.C., Moller, S.J., Carslaw, D.C.: Lower vehicular primary emissions of no2 in europe than assumed in policy projections. Nature Geoscience **10**(12), 914–918 (2017)

[12] Stohl, A., Berg, T., Burkhart, J., Fjæraa, A., Forster, C., Herber, A., Hov, Ø., Lunder, C., McMillan, W., Oltmans, S., *et al.*: Arctic smoke? record high air pollution levels in the european arctic due to agricultural fires in eastern europe. Atmospheric Chemistry and Physics Discussions **6**(5), 9655–9722 (2006)

[13] Querol, X., Tobías, A., Pérez, N., Karanasiou, A., Amato, F., Stafoggia, M., García-Pando, C.P., Ginoux, P., Forastiere, F., Gumy, S., *et al.*: Monitoring the impact of desert dust outbreaks for air quality for health studies. Environment international **130**, 104867 (2019)

[14] Karagulian, F., Belis, C.A., Dora, C.F.C., Prüss-Ustün, A.M., Bonjour, S., Adair-Rohani, H., Amann, M.: Contributions to cities' ambient particulate matter (pm): A systematic review of local source contributions at global level. Atmospheric environment **120**, 475–483 (2015)

[15] Cressie, N., Wikle, C.K.: Statistics for Spatio-temporal Data. John Wiley & Sons, ??? (2015)

[16] Atkinson, P., Foody, G.: Uncertainty in remote sensing and gis: fundamentals. Uncertainty in remote sensing and GIS, 1–18 (2002)

[17] Guignard, F., Laib, M., Amato, F., Kanevski, M.: Advanced analysis of temporal data using fisher-shannon information: theoretical development and application in geosciences. Frontiers in Earth Science **8**, 255 (2020)

[18] Van Donkelaar, A., Martin, R.V., Brauer, M., Hsu, N.C., Kahn, R.A., Levy, R.C., Lyapustin, A., Sayer, A.M., Winker, D.M.: Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. Environmental science & technology **50**(7), 3762–3772 (2016)

[19] Tai, A.P., Mickley, L.J., Jacob, D.J.: Correlations between fine particulate matter (pm2. 5) and meteorological variables in the united states: Implications for the sensitivity of pm2. 5 to climate change. Atmospheric environment **44**(32), 3976–3984 (2010)

[20] Calvo, A., Pont, V., Olmo, F., Castro, A., Alados-Arboledas, L., Vicente, A., Fernández-Raga, M., Fraile, R., *et al.*: Air masses and weather types: a useful tool for characterizing precipitation chemistry and wet deposition. Aerosol and Air Quality Research **12**(5), 856–878 (2012)

[21] Saenz-de-Miera, O., Rosselló, J.: Modeling tourism impacts on air pollution: The case study of pm10 in mallorca. Tourism Management **40**, 273–281 (2014)

[22] López-Villarrubia, E., Ballester, F., Iñiguez, C., Peral, N.: Air pollution and mortality in the canary islands: a time-series analysis. Environmental Health **9**(1), 1–11 (2010)

[23] Basart, S., Pérez, C., Cuevas, E., Baldasano, J.M., Gobbi, G.P.: Aerosol characterization in northern africa, northeastern atlantic, mediterranean basin and middle east from direct-sun aeronet observations. Atmospheric Chemistry and Physics **9**(21), 8265–8282 (2009)

[24] Köhler, L., Gieger, T., Leuschner, C.: Altitudinal change in soil and foliar nutrient concentrations and in microclimate across the tree line on the subtropical island mountain mt. teide (canary islands). Flora-Morphology, Distribution, Functional Ecology of Plants **201**(3), 202–214 (2006)

[25] Wallace, J., Kanaroglou, P.: The effect of temperature inversions on ground-level nitrogen dioxide (no2) and fine particulate matter (pm2. 5) using temperature profiles from the atmospheric infrared sounder (airs). Science of the Total Environment **407**(18), 5085–5095 (2009)

[26] Elminir, H.K.: Dependence of urban air pollutants on meteorology. Science of the total environment **350**(1-3), 225–237 (2005)

[27] Zhang, N., Ren, R., Zhang, Q., Zhang, T.: Air pollution and tourism development: An interplay. Annals of Tourism Research **85**, 103032 (2020)

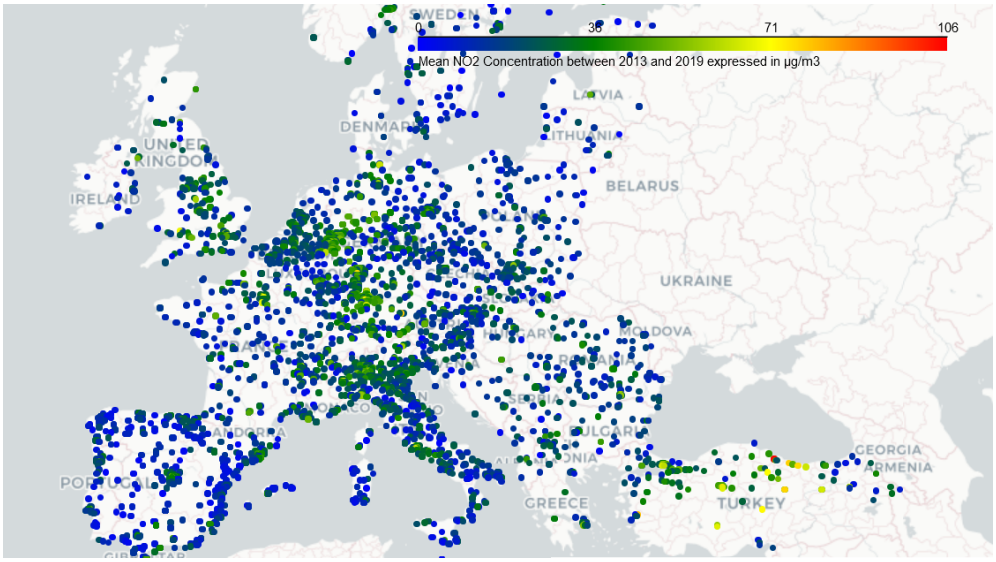
[28] Wang, Q., Kwan, M.-P., Zhou, K., Fan, J., Wang, Y., Zhan, D.: The impacts of urbanization on fine particulate matter (pm2. 5) concentrations: Empirical evidence from 135 countries worldwide. Environmental pollution **247**, 989–998 (2019)

[29] Jacob, D.J., Winner, D.A.: Effect of climate change on air quality. Atmospheric environment **43**(1), 51–63 (2009)

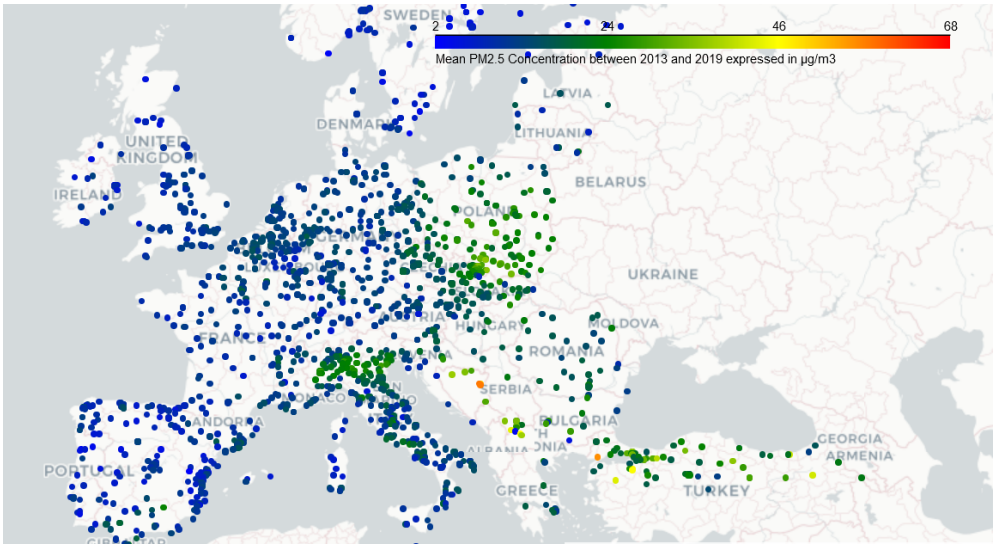
Appendix A First spatio-temporal EDA

The two plots A1a and A1b display maps of mean NO2 and PM2.5 concentrations between 2013 and 2019, expressed in micrograms per cubic meter ($\mu g/m^3$). From this visual representation that you can see, it is evident that central Europe registered higher NO2 concentrations than the rest of Europe. Regarding PM2.5 it is especially eastern Europe that have higher mean PM2.5 concentration than the rest of Europe.

The plots A2 and A3 delve into the temporal dynamics of NO2 ans PM2.5 concentrations.The first plots A2a and A3a depict the mean NO2 and PM2.5 concentrations



(a)



(b)

Fig. A1: **Spatial mean of NO2 (a) and PM2.5 (b)**: Europe map showing concentration of NO2 (a) and PM2.5 (b) between 2013 and 2019 expressed in $\mu g/m^3$. Each color corresponds to an average NO2/PM2.5 value.

by hour of the day. Clear patterns emerge, with higher concentrations observed during the morning hours (6h-11h) and the evening (18h-23h) compared to the night (23h-6h) and afternoon (11h-18h). This can be attributed to daily human routines, which often involve commuting during morning and evening hours, leading to increased traffic-related emissions during these periods. In contrast, the night and afternoon periods often see less vehicular activity, resulting in lower pollutant levels.

The second plots A2b and A3b demonstrate the mean NO2 and PM2.5 concentration by day of the week. An interesting observation from these plots is that the average pollutant concentration is significantly higher during weekdays than on weekends for NO2. This could potentially be linked to the pattern of human activity, with typically higher traffic and industrial activity during the work week compared to the weekend, leading to increased emissions.

The third plots A2c and A3c illustrate the mean NO2 and PM2.5 concentration grouped by month. A discernible pattern emerged from the A2c plot, suggesting that NO2 concentrations are higher in winter than in summer. The seasonal trend is exactly the same for the graphic A3c.

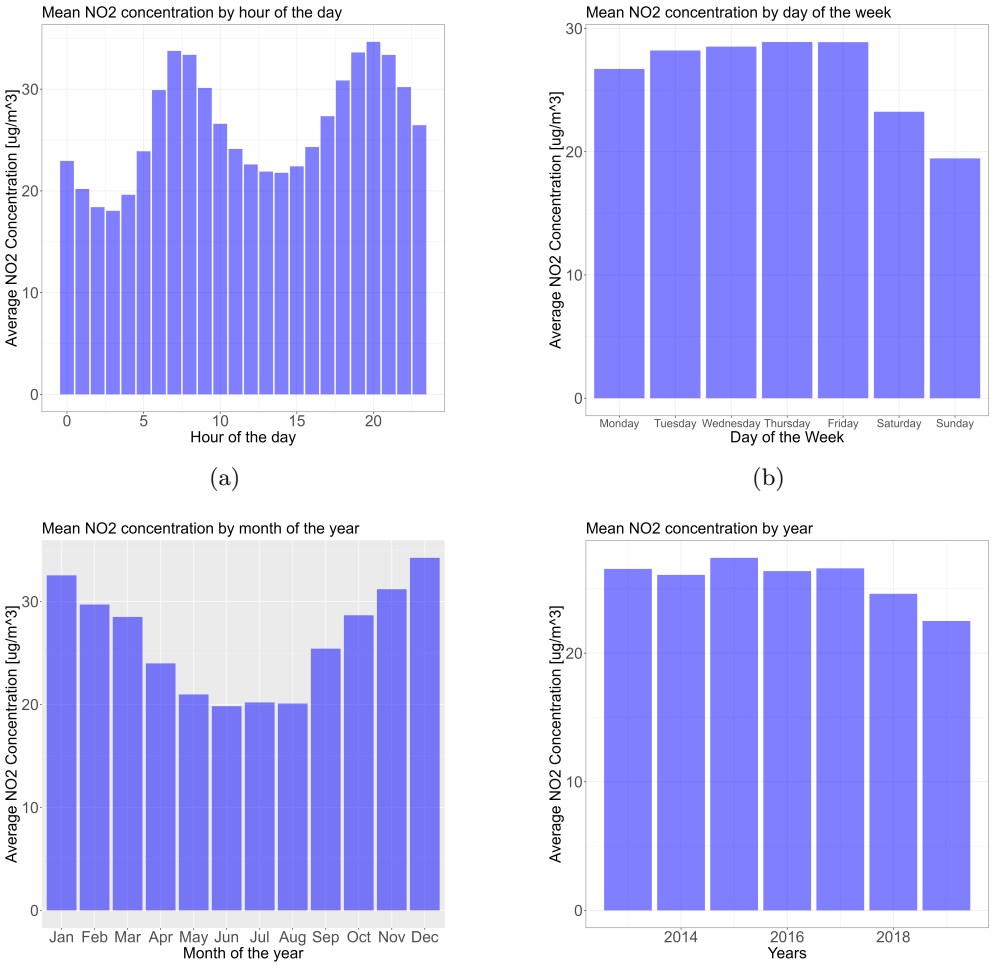


Fig. A2: **Temporal NO2 Analysis**: Mean NO2 concentration by hour of the day (a), day of the week (b), month of the year (c), and years (d) expressed in $\mu g/m^3$

In the two figures A4a and A4b, we clearly see again some seasonalities in the NO2 and PM2.5 concentration data. Showing that these values are usually higher in winter than in summer.

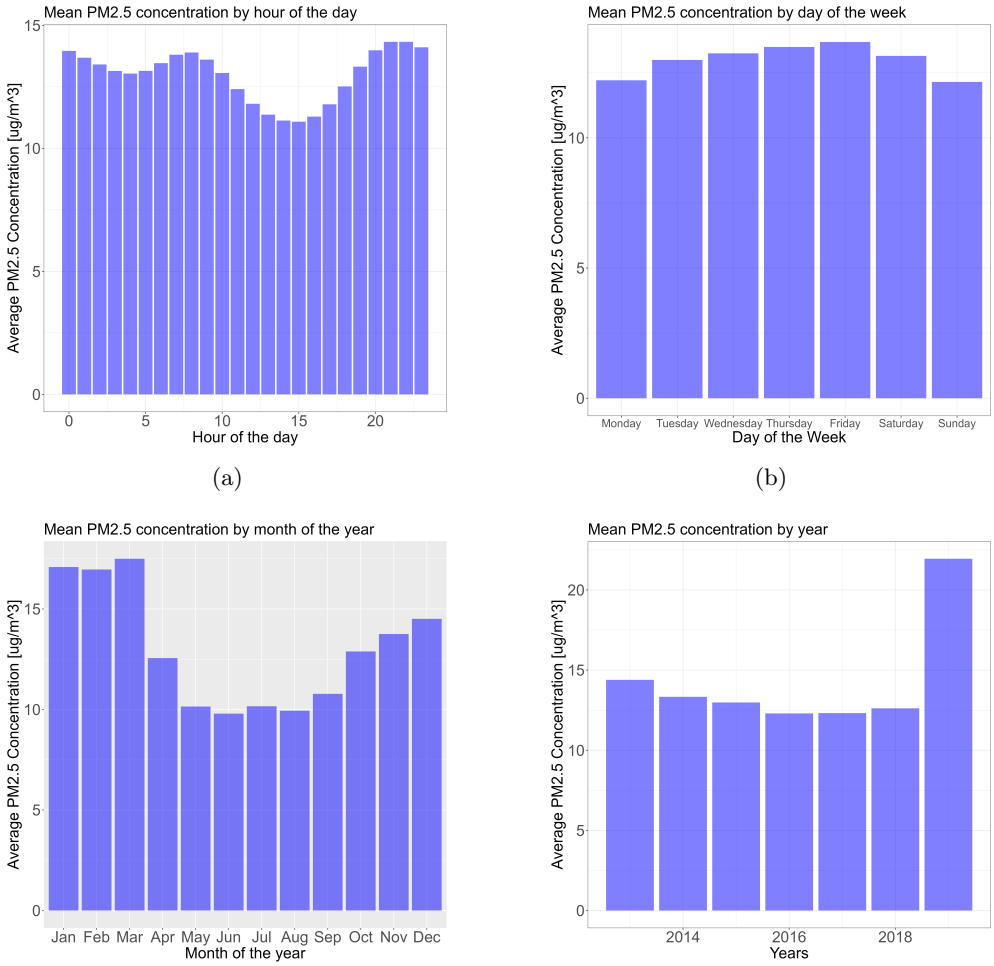


Fig. A3: **Temporal PM2.5 Analysis**: Mean PM2.5 concentration by hour of the day (a), day of the week (b), month of the year (c), and years (d) expressed in $\mu g/m^3$

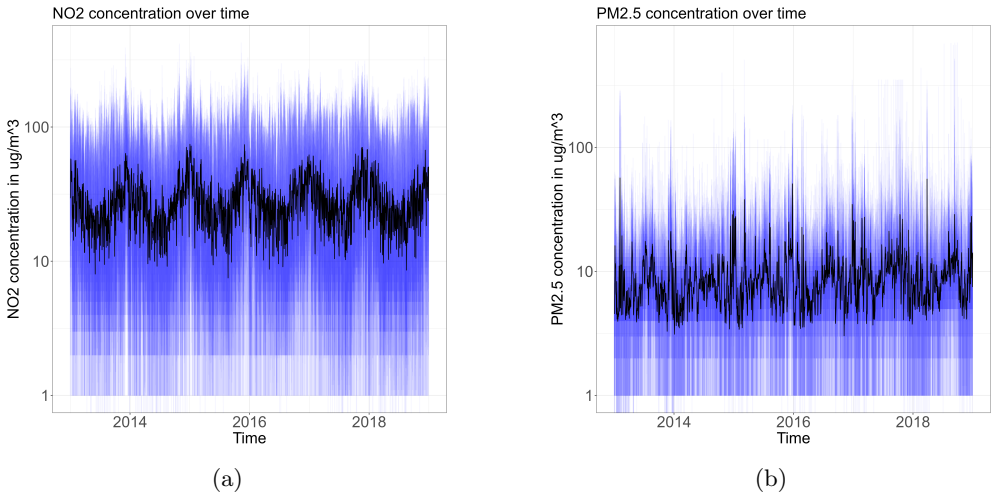


Fig. A4: **Time-series superposition:** NO2 concentration (a) and PM2.5 concentration (b) time series superposition between 2013 and 2019 expressed in $\mu g/m^3$ with average NO2/PM2.5 concentration time-serie in black.

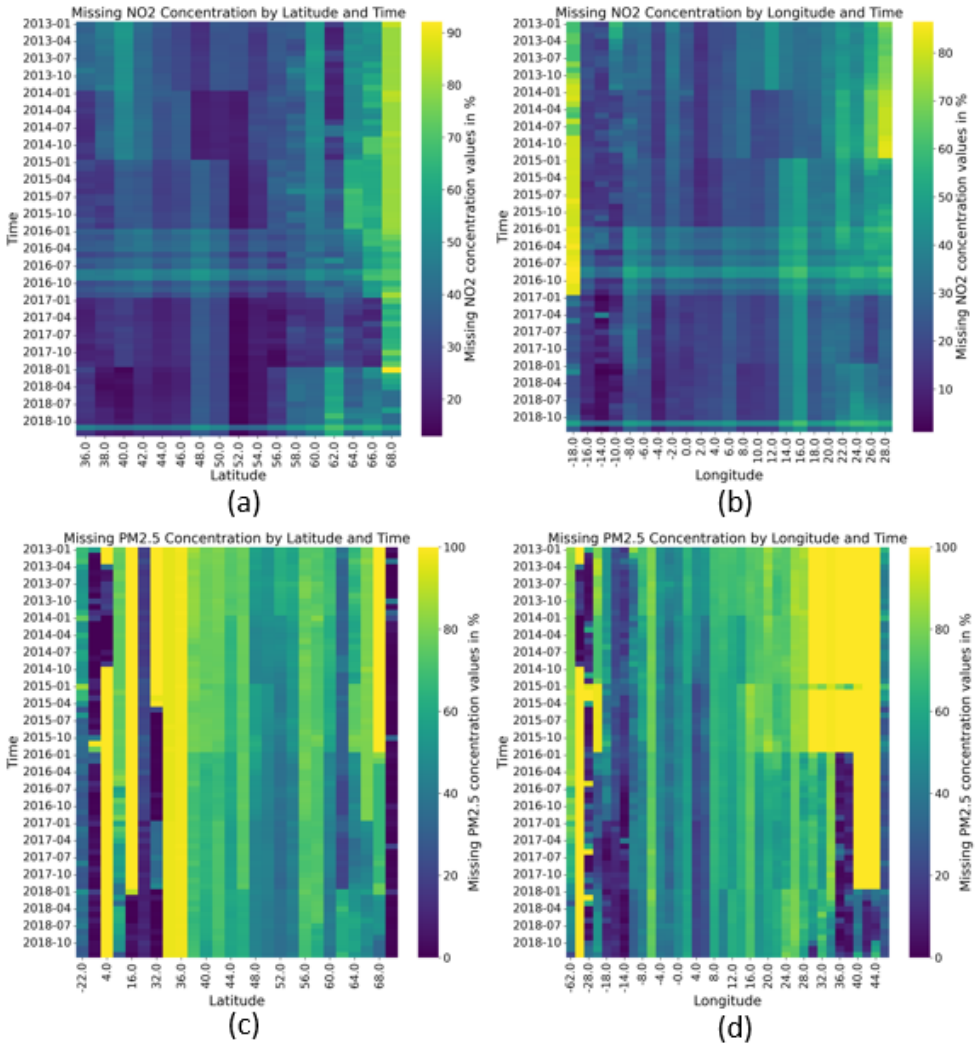


Fig. A5: **Hovmoller plots of missing concentration data:** Hovmoller plots representing the percentage of missing values for NO2 (top row) and PM2.5 (bottom row) with respect to latitude (left column) and longitude (right column)

Appendix B Information Theory

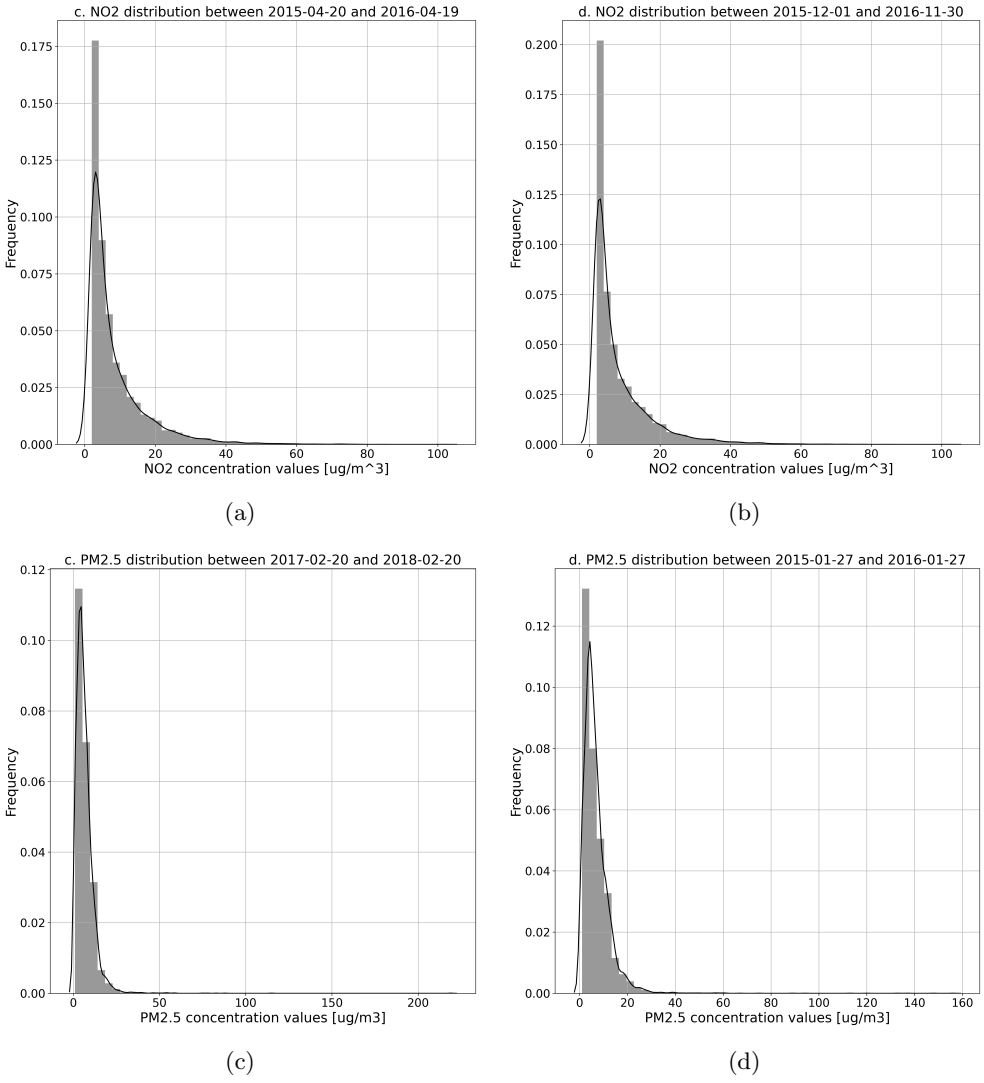
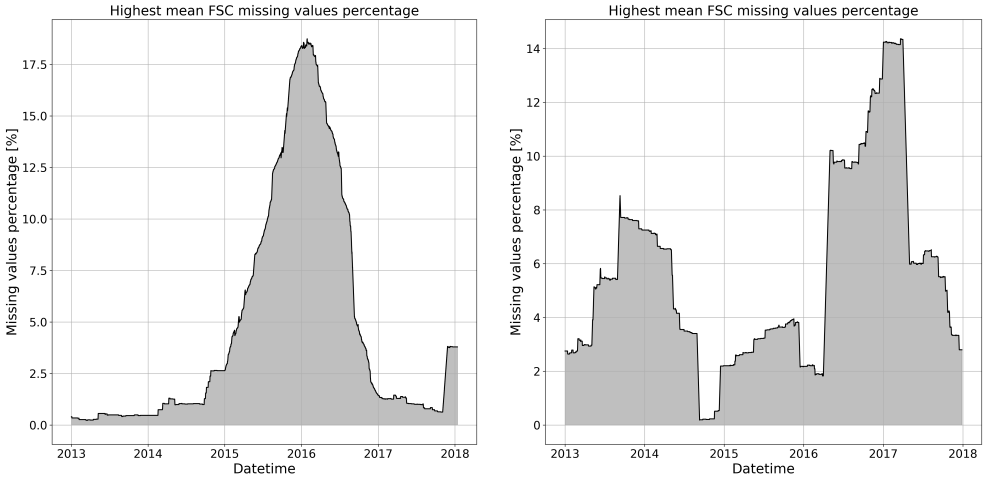


Fig. B6: **NO2 and PM2.5 distributions associated to the minimum and maximum peaks of the station with highest mean FSC based on NO2 or PM2.5 concentration data:** In the graphics (a) and (b), we can find the NO2 concentration distributions corresponding to the year after the dates associated to the minimum (a) and maximum (b) peaks of the station with highest mean FSC based on NO2 concentration data. In the graphics (c) and (d), we can find the PM2.5 concentration distributions corresponding to the year after the dates associated to the minimum (c) and maximum (d) peaks of the station with highest mean FSC based on PM2.5 concentration data.



(a) (b)

Fig. B7: **Highest mean FSC missing value percentages:** Missing value percentage with respect to time of the concentration data corresponding to the station with highest mean FSC based on NO2 concenctration data (a) or PM2.5 concentration data (b)

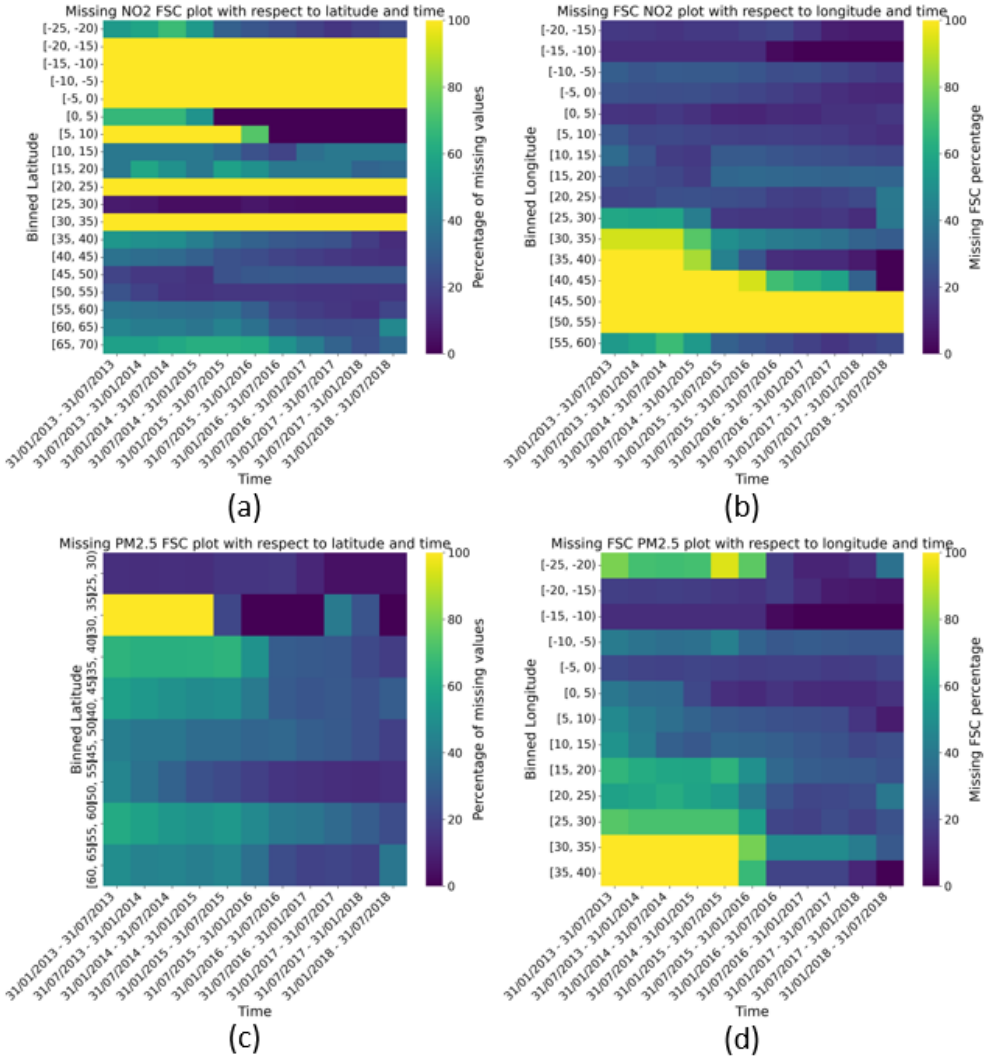


Fig. B8: **Hovmoller FSC missing plots:** Hovmoller plots representing the average missing FSC values computed over NO2 (first row) or missing PM2.5 values (bottom row) time series with respect to latitude (left column) or longitude (right column) and time

Appendix C Empirical orthogonal function

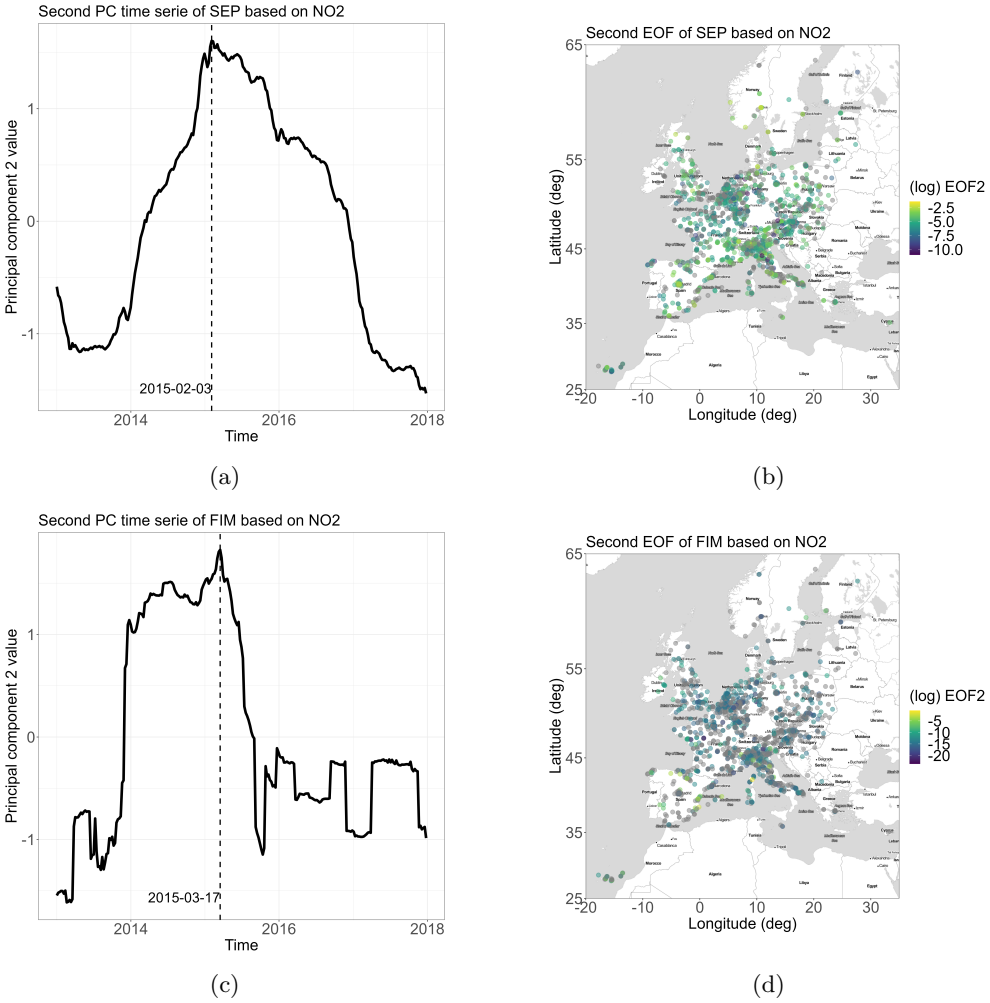


Fig. C9: **Second principal component analysis of SEP and FIM based on NO2 concentration data:** In (a) and (c), we see the second principal component temporal basis computed over the SEP (a) or FIM (c) based on NO2 concentration data with respect to time. A dashed line referenced the date corresponding to the maximum value of this time serie. In (b) and (d), we can analyze the second spatial principal component coefficients map of SEP (b) and FIM(d) based on NO2 concentration data. The logarithmic of this second principal component coefficients is represented through colors of the points in the Europe map

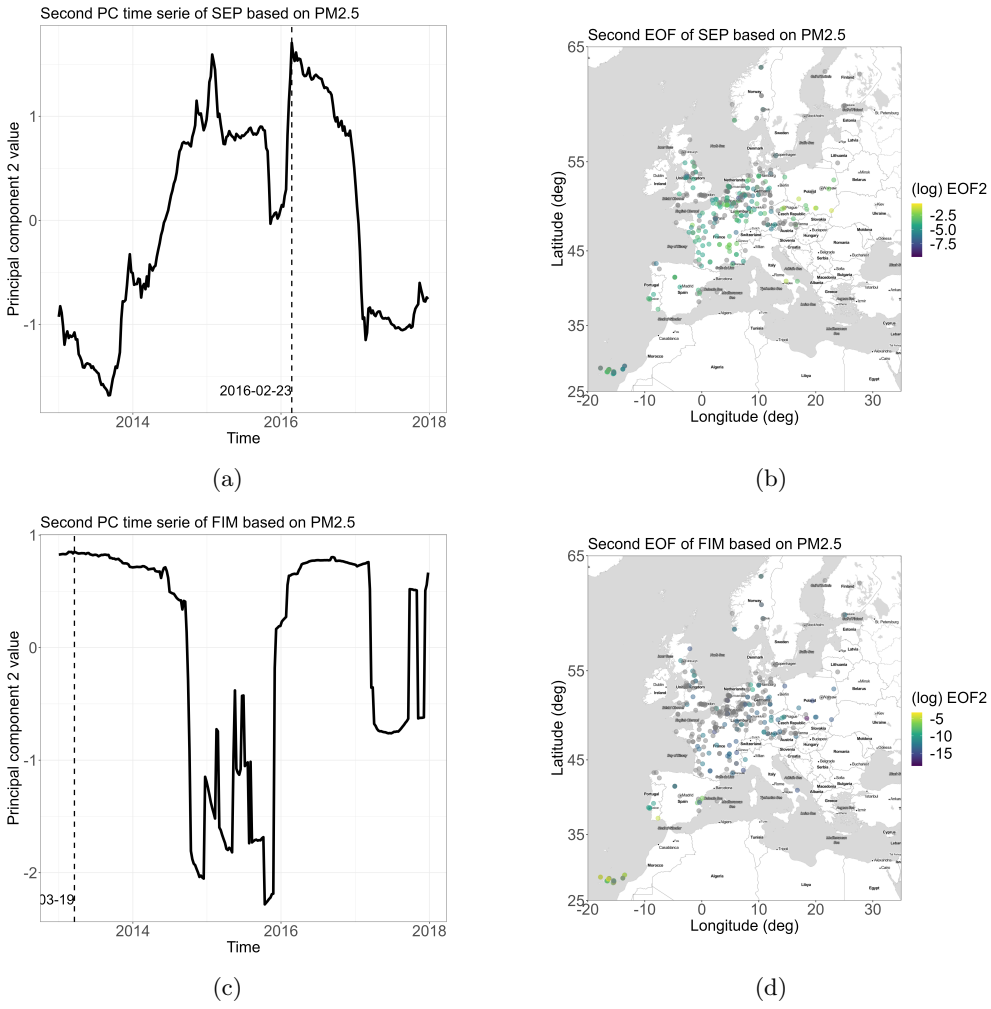


Fig. C10: **Second principal component analysis of SEP and FIM based on PM2.5 concentration data:** In (a) and (c), we see the second principal component temporal basis computed over the SEP (a) or FIM (c) based on PM2.5 concentration data with respect to time. A dashed line referenced the date corresponding to the maximum value of this time serie. In (b) and (d), we can analyze the second spatial principal component coefficients map of SEP (b) and FIM(d) based on PM2.5 concentration data. The logarithmic of this second principal component coefficients is represented through colors of the points in the Europe map