Exposure to air pollution and COVID-19 expansion in Western-Europe

Darja Nonaca & Tim Tuuva

May 2020

1 Abstract

Many of the pre-existing conditions that increase the risk of death in those with COVID-19 are the same diseases that are affected by long-term exposure to air pollution. Hence the doubt: Recent studies of Harvard University [1] showed that an increase of 1 $\mu g/m^3$ in $\text{PM}_{2.5}$ is associated with an 8% increase in the COVID-19 death rate in the United States. In this work we investigated whether a similar correlation could be found in Western-Europe.

2 Introduction

The question that we try to answer is whether an average long-term exposure to the air pollution influences the mortality rate of COVID-19 in Western-Europe. We hypothesise that an average long-term exposure to PM2.5 increases the mortality rate due to COVID-19. We choose to analyse the regions where the COVID-19 caused the highest fatality in order to avoid noise on the data. We select the following regions for our study: The United Kingdom and the French region Ile-de-France, Grand-Est, Auvergne-Rhone-Alpes, Provence-Alpes-Cote d'Azur. We use PCA to investigate if there is a correlation between the deaths due to COVID-19 and $\text{PM}_{2.5}$. In section 3 we present the tool and its limitations, we demonstrate PCA using the synthetic data. The last part of section 3 describes the real data found online. We give the results for the real data in section 4, we do the discussion in the section 5 and the conclusion in the section 6.

3 Materials and Methods

3.1 Test of the PCA tool

For any dataset we could arrange an one-to-one mapping to a linear vector space. Indeed, we can represent features as a vector, within each component carries a different feature. To explain how PCA works, we generate a synthetic dataset with the following features:

average age	presence particle 1	presence particle 2	disease 1	disease 2

Table 1: Features of the synthetic data.

We define $c_m = (c_{m_1}, \ldots, c_{m_5})$ to be the *features vector*, where m represent the number of samples within the dataset.
We distribute the particles 1 and 2 within the dataset according to Gaussian distribution as follows:

- c_{m1} (age) is i.i.d. distributed in the interval [0-99]
- c_{m2} (particle 1) $\sim \mathcal{N}(\mu_1, \sigma^2)$
- c_{m3} (particle 2) $\sim \mathcal{N}(\mu_1, \mu_2, \sigma^2)$

Then we correlate the diseases 1 and 2 to the first three features (c_{m1}, c_{m2}, c_{m3}) as follows:

- c_{m4} (desease 1), computed as $c_{m_4} = \alpha\ c_{m_2} + \beta\ c_{m_1}$
- c_{m5} (desease 2), computed as $c_{m_5} = \gamma\ c_{m_3}$

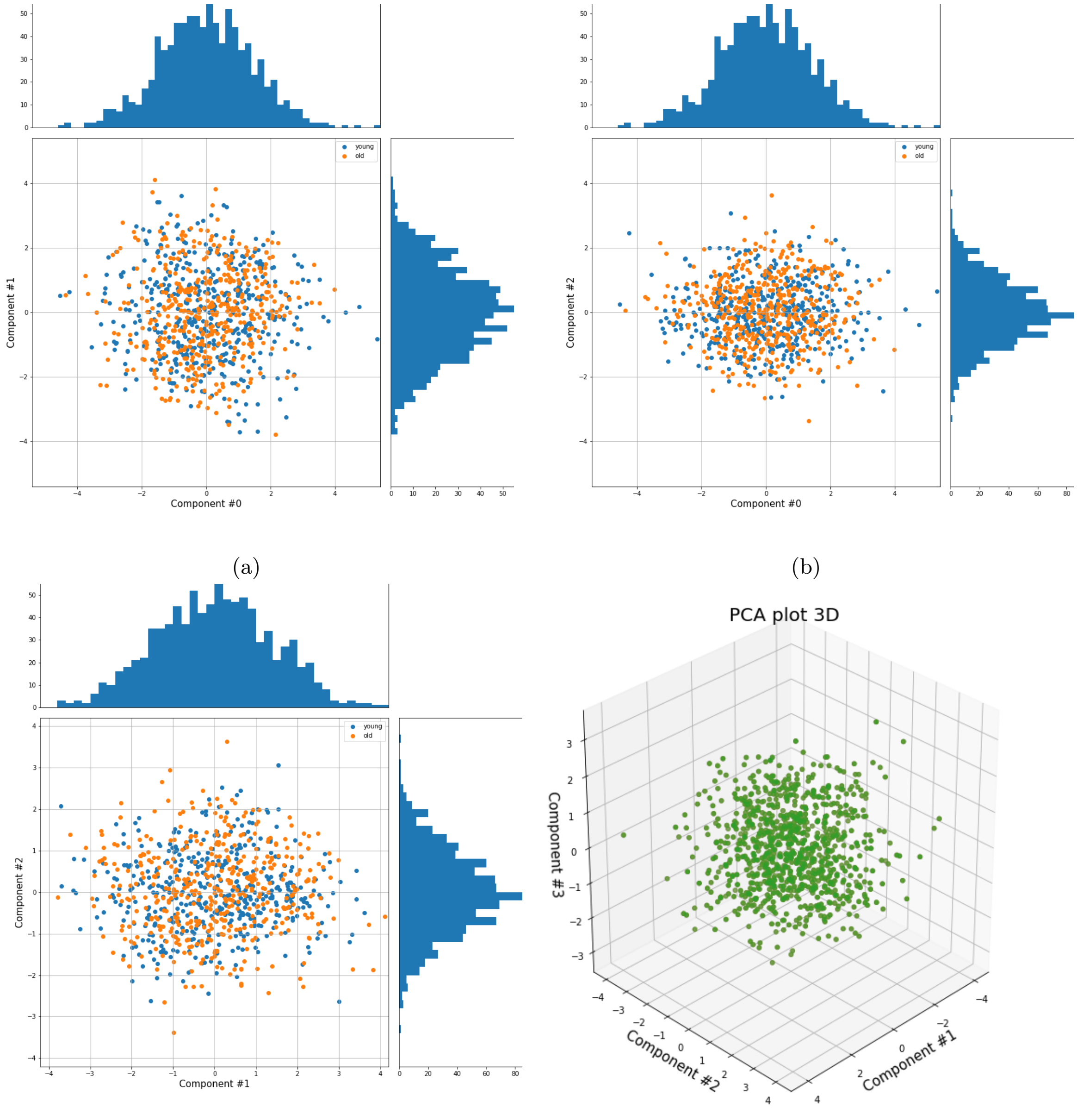
The parameters $\mu_i, \sigma_i (i = 1, 2), \alpha, \beta$ and γ are chosen randomly and do not represent any real behaviour. After applying the PCA on the whole dataset we find the following singular values:

λ_1	λ_2	λ_3	λ_4	λ_5
4.26086984e+01	4.21239011e+01	3.00844774e+01	2.44462381e−14	9.54265059e−15

Table 2: Features of the synthetic data.

The singular values represent the importance of a certain feature within the dataset. In our example, we find singular values λ_4 and λ_5 to be almost 0. This is not a surprise as the features c_4 and c_5 carry an information that is dependent on (correlated with) features c_1, c_2, c_3. Hence, c_4 and c_5 do not add any supplementary (independent) information. The most significant singular values are λ_1, λ_2, λ_3, which means that the features c_1, c_2, c_3 are the most relevant features. We call them *principal components*.

We plot the 3 principal components one vs. the other in the Figure 1. The presence of a cluster indicates that the within points are correlated, i.e. they are a linear combination of principal components (main features). The mean of the cluster indicates the mean of the main features in the principal components space. From the plots in Figure 1 we observe an unique cluster which would indicate that all the data are correlated.



(a)

(b)

(c)

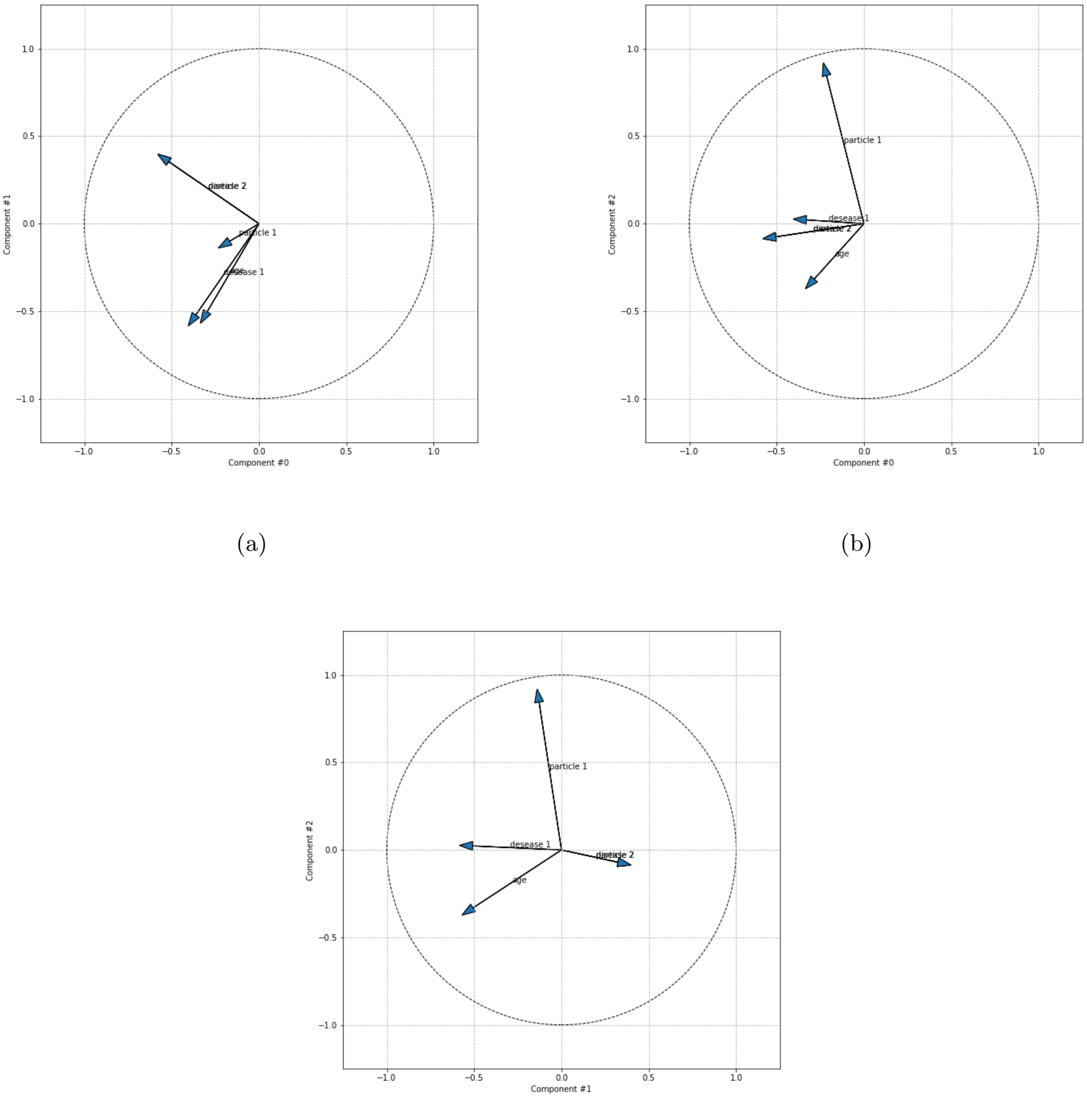
(d)

Figure 1: PCA plots of the synthetically generated data.

Nevertheless, we have all the information about the data construction and we know that the points are not all correlated. Hence, probably what we observe in Figure 1 are several clusters which means are so close to make it look as there was an unique cluster. To support the principle component analysis, we use a complementary tool called *correlation circle* that is discussed in the next paragraph.

3.2 Advanced tool: Correlation Circle

The correlation circle is a representation of the features vectors from the original dataset with respect to the principle component vectors basis. It allows to see the association between features and principle components and it is a very helpful support for the PCA. Figure 2 shows these plots.



(a)

(b)

(c)

Figure 2: Correlation circles to support the PCA on synthetic data.

3.3 Where does PCA fail

Although PCA is a useful tool, it also has some limitations:

- When we select the number of principal components by approximating some singular values to 0, we have a loss of information.
- PCA operates by finding the direction of maximum variance within the dataset and it could happen that a projection into that direction is not useful, for example when the noise is too high. We could discard like this some important information.
- In order to make relevant conclusions on the feature's correlations using PCA, dataset needs to contain as many samples as possible.

3.4 PCA on real data

Ideally, we would like to have death rates due to COVID-19 per city, but this is very difficult to find at the moment. Hence we settle upon a more bigger spatial resolution. We want to consider the regions in Western-Europe where the virus has hit the most. Table 3 summarises the regions that were analysed in this work.

Information on the dataset			
Country	Most affected Regions	data COVID-19	data $PM_{2.5}$
France	Ile-de-France Grand Est Provence-Alpes-Côte d'Azur Auvergne-Rhône-Alpes	[2]	[3]
UK	All territory	[5]	[4]

Table 3

The data are composed of 4 features which are:

NO_2 $[\mu g/m^3]$	PM10 $[\mu g/m^3]$	PM2.5 $[\mu g/m^3]$	*covid_death_rate*/population

Table 4: Features of the real data.

We find some of the points to be outliers which prevents the PCA to process correctly (as the PC vectors point towards the highest variance in the data). Hence, we decide to remove such outlier points by computing the Interquartile range:
Let Q_1, Q_3 the first and third quartiles:

$$IRQ = Q_3 - Q_1$$

$$a = Q_1 - 1.5 \times IRQ$$

$$b = Q_3 + 1.5 \times IRQ$$

Then we keep only the points that are in the interval $[a : b]$. In the end of this process, we have 55 valid data points.

4 Results

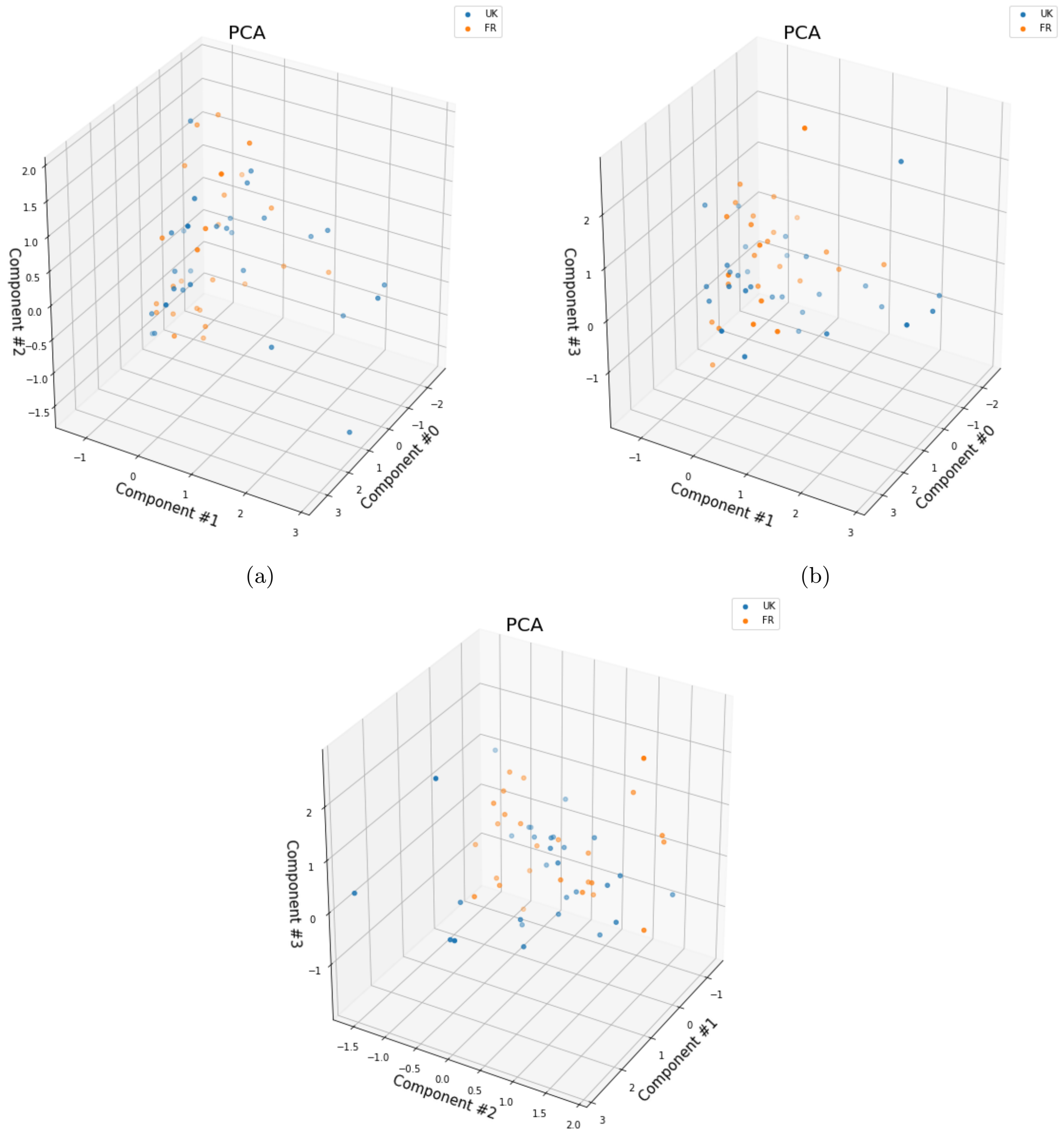
4.1 Singular values

λ_1	λ_2	λ_3	λ_4
9.82809116	7.09382197	6.24462426	5.48552476

Table 5: Features of the synthetic data.

We can see that all the singular values are relevant thus we perform the PCA by using 4 principles components.

4.2 PCA plots



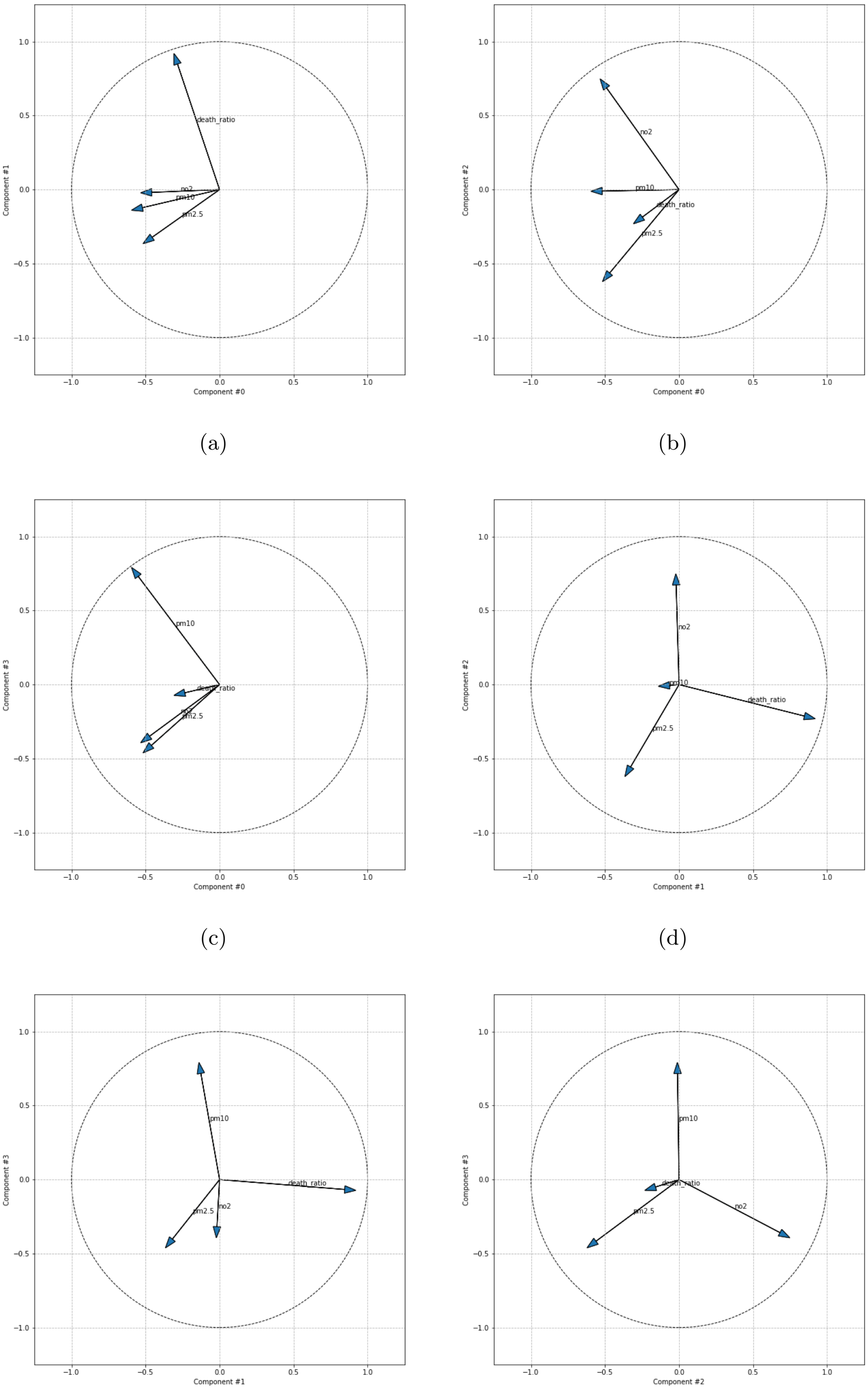
(a)

(b)

(c)

Figure 3: Results after applying PCA on the UK dataset (in blue) and on the France dataset (in orange).

From the Figure 3 we observe that the data points are poorly correlated to the principle component 1. To understand which features is represented by the latter we plot the correlation circles in Figure 4.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4: Correlation circles.

From the Figure 4 we observe that the component 1 has a similar direction of the death rate. We also see that the death rate is almost perpendicular to all the pollutants.

5 PCA on the data used in [1]

The study of Harvard University [1] found a correlation between the pollutant PM2.5 and death rate due to COVID-19. We apply PCA tool on the data used in the Harvard's research to see if we can confirm their conclusion by using a different tool. The Figure 5 shows the correlation circle applied to such data.

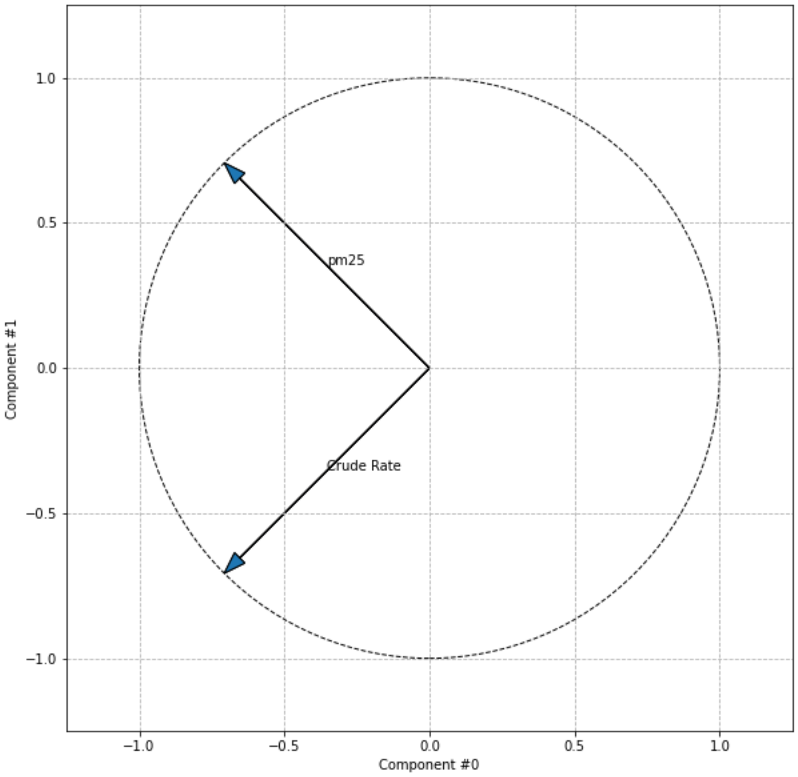


Figure 5: Correlation circle analysis using the data assembled by Harvard University in [1]. Crude rate is the death rate due to COVID-19 every 1000 people.

From the Figure 5 we see that the features PM2.5 and the death rate are perpendicular to each other thus they are uncorrelated.

6 Discussion and conclusions

We had a lot of troubles to find the European information on the COVID-19 mortality per "county". In fact, many countries publish a global information for their territory. For this reason we worked only with 55 data points which are not enough to perform a strong analysis using PCA. Furthermore, by performing our analysis on the data assembled by Harvard University 5, we found that PM2.5 and death rate are uncorrelated which is contradictory to their article. This make us conclude that the way we do the correlation analysis should be rethought.

7 GitHub

The GitHub repository is public and it is accessible at: https://github.com/donaca/COVID19PM2.5

References

[1] X. Wu, R. Nethery, M. Sabath, D. Braun, F. Dominici. *Exposure to air pollution and COVID-19 mortality in the United States: A nationwide cross-sectional study.* 2020.

[2] Données de certification électronique des décès associés au COVID-19 (CEPIDC)
https://www.data.gouv.fr/fr/datasets/donnees-de-certification-electronique-des-deces-associes-au-cov

[3] Féderation des associations de la survelliance de la qualité d'air,
https://atmo-france.org/

[4] ENV02 - Air quality statistics in UK,
https://www.gov.uk/government/statistical-data-sets/env02-air-quality-statistics

[5] Office for National Statistics,
https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/datasets/death