Estimation of the determinants of the spread and fatality rate of COVID-19

University of Milan

Rapso, Ricardo Murillo
ricardo.murillorapso@studenti.unimi.it

Papallazi, Lirida
lirida.papallazi@studenti.unimi.it

October 29, 2020

Abstract

The present research aims focuses of the diffusion of the Corona Virus Disease 19 that starting from China has hit the entire world. The goal is finding out if there is a correlation between the contagion rate of COVID-19 and the socio-demographic characteristics, health infrastructure indicators, and pollution levels of the considered regions. In order to do this and also to create a predictive model for the remaining seasons of autumn and winter (since Covid-19 hit Italy in the Spring season), similar regions for humidity and temperatures were searched together with the same data about pollution and ”macroscopic” characteristics.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

On January 30th 2020, the World Health Organization declared the outbreak COVID-19 as a Public Health Emergency of International Concern.[1]

This might be the beginning of an apocalyptic movie but it is our current reality, a reality that will probably affect everyone's life and also future.

As of 12 October 2020, more than 37.5 million cases have been confirmed, with more than 1.07 million deaths. The virus has hit 188 countries with different intensities: at the top of the list of the countries that mostly suffered the effects of Covid-19 there is the USA at the top of the list both for the contagion rate and especially a fatality rate, followed by India, Brazil and Russia. Since the disease spreads when people are physically close through air via small droplets and aerosols the main preventive measures, in absence of a vaccine, have been hand washing, wearing face masks and social distancing. Many countries in order to flatten the contagion curve and to not make their healthcare systems collapse also adopted lockdowns and travel restrictions.

Chapter 2

Epidemiology of Covid-19, SARS and Influenza

2.1 Initial purpose

The paper aim changed during its development in order to adapt to the actual data available. The starting goal of the paper was, with a focus on the sole Italy, of comparing the contagion and lethality curves of Covid-19 with those of the more common Influenza.

At this point it seems useful to briefly introduce both Covid-19 and Influenza from an epidemiological point of view.

2.2 Influenza

Influenza is an infectious disease caused by an influenza virus an RNA virus. Symptoms can be fever, sore throat, pain in muscles and headache and they generally last less than a week. Complications of influenza include pneuomonia and worsening of pre-existing health problems, such as asthma and heart failure. This virus is typically spread via cough, sneezes but also touching contaminated surfaces and then touching the face and eyes. The positive thing about influenza is that there exists a vaccine for it.

2.3 SARS

The Severe Acute Respiratory Syndrome is the disease caused by SARS-CoV-1 that outbroke in 2003 in Asia and the first disease for with the WHO declared an emergency status. It causes a severe illness that often begins with fever, headache, respiratory symptoms and pneumonia. In the outbreak of 2003 about 9% of patients with SARS infection died.

2.4 Covid-19

Covid-19 is a new coronavirus strain; coronaviruses are viruses that circulate among animals and some of them can also infect humans and it is believed to also be the origin of Covid-19 which also belongs to the same family of the SARS virus.[3]

2.4.1 Origins

The sudden and unexpected outburst of Covid-19 resulted in misinformation and conspiracy theories about the origins, the prevention, the diagnosis and the treatment of the disease. What can be said at a distance of 10 months from its appearance is that on the 31st of December the WHO received reports of a cluster of viral pneumonia cases with unknown origins. After some investigations the it was confirmed that the infected people had visited the Huanan Seafood Wholesale Market and therefore the virus is thought to have zoonotic origins.[2]

2.4.2 Transmission

The main modality of transmission are:

- direct ;
- indirectly (via contaminated objects or surfaces);
- by close contact with infected people through secretions from the mouth and nose.

The last one in particular is the main reason for which the preventive techniques of frequent sanitization, social distancing and surgical masks are required.

Considering that people remain infectious for 7-14 days another important issue of Covid-19 are the asymptomatic individuals.

2.4.3 Symptoms

The symptoms of Covid-19 can be very various and some of them are quite similar to the common cold, making it even more difficult to diagnose especially during the mid-seasons, when it is more probable to catch the cold. The most frequent however are:

- fever;
- dry cough;
- fatigue;
- loss of the sense of smell and/or taste.

In extreme cases it might cause pneumonia, acute respiratory distress syndrome, sepsis and kidney failure. When combined with previous debilitating illnesses it might lead to death and this happens especially with elderly people and those people with underlying conditions such as hypertension, cardiac problems, those being treated with immunosuppressive drugs, and so on and so forth.

2.4.4 Diagnosis

The diagnosis of Covid-19 can be done via:

- RNA testing of secretions collected via nasopharyngeal swab (RT-PCR);
- CT imaging of the chest in order to check for pleural effusions[4] or
- serological test which detect antibodies produced by the body in response to the infection.

.

Chapter 3

Dataset

The building of the data-set for the analysis of Covid-19 was extremely difficult and tricky and although the issues of such a project will be deeper explained later in the paper, this section will introduce the data and features that were collected.

Figure 3.1: Coffins of deceased in Bergamo loaded on military vehicles and transported to nearby provinces

3.1 Italian provinces and regions

As it has been already said countries have been hit with different intensity by Covid-19, but this difference can be seen also within the single states. In Italy there is an enormous difference in effects between North and South: in particular during the highest peak, the case of the city of Bergamo (region of Lombardy) received enormous notoriety for its elevated numbers of both fatality and contagion rates. Now it is assumed that they are a consequence of hospital contamination.

The issue with finding the daily contagion and lethality rates at province level is that such data (like most of the data on Covid-19) are not available and the ones that the authors were able to find about all the provinces arrive until June 24th.

The solution adopted was therefore that of conducting the analysis at a regional level. This decision was also strengthened by the fact that during the peak of the pandemic hospitals became overwhelmed in a short amount of time. As a result, many people who had to be hospitalized because seriously ill were transported by ambulances to hospitals located in other provinces.

3.2 Features collected

As it has been already stated, the original purpose of the analysis was that of finding a possible relation between the contagion rate and the fatality rate of the italian provinces and the air quality, controlling for some macroscopic feature.

The logic would be that since Covid-19 is an infectious respiratory disease and the effects of air pollution on health, now well documented, concern acute and chronic problems, mainly of the respiratory and cardiovascular systems[6], finding a correlation between these variables could explain why its impact has been so different from region to region.

Even though the focus moved from the provinces to the regions, the original purpose remained the same.

Table 3.1: Macroscopic features

Variable	Type
Unemployment	Numerical
Population	Numerical
Region area (km2)	Numerical
Density (ab/km2)	Numerical
PM10 (μg/m3)	Numerical
PM2.5(μg/m3)	Numerical
CAQI index	Numerical
Hospital Beds	Numerical
Heart disease	Fatality rate (per 10.000 abs)
Stroke	Fatality rate (per 10.000 abs)
Malignant Tumor	Fatality rate (per 10.000 abs)
Tracheal, Bronchus and Lung Cancer	Fatality rate (per 10.000 abs)
Influenza and Pneumonia	Fatality rate (per 10.000 abs)
Chronic Lower Respiratory Disease	Fatality rate (per 10.000 abs)
Tuberculosis	Fatality rate (per 10.000 abs)

3.2.1 CAQI

The CAQI index is used as an air quality index since 2006 and its definition was necessary to accommodate the introduction of limit values for PM2.5 in the index.[5]

The earth's atmosphere is an aerosol of dispersions of liquid and solid particles in a gaseous envelope made up of a gas mixture composed of Nitrogen (N2), Oxygen (O2), water vapor, Argon (Ar), Carbon Dioxide (CO2) and rare gases.

PM10

Particulate matter is the atmospheric pollutant that causes the greatest damage to human health since its particles since its particles are made up of different chemical components that penetrate deep into the lungs. The fraction of PM10 due to human activities would be 40-50% emitted directly into the atmosphere, while the remaining 50-60% results from chemical reactions.[7]

PM2.5

As per the PM10 case, fine particulate pollution (PM2.5, namely particulate matter with a diameter of less than 2.5 microns) is made up of solid and liquid particles so small that they not only penetrate deep into our lungs, but also enter our bloodstream, exactly like oxygen.

Ozone

Ozone (O3) is a highly reactive form of oxygen and is made up of three oxygen atoms. In the stratosphere, one of the highest layers of the atmosphere, ozone protects us from dangerous ultraviolet radiation from the sun but it also reduces the ability of plants to perform photosynthesis and hinders their absorption of carbon dioxide, weakening the growth and reproduction of plants. In the human body instead, high levels of O3 cause inflammation of the lungs and bronchi.

Nitrogen dioxide (NO2)

Nitrogen dioxide (NO2) is a reactive gas which in the short term can cause decreased lung function while in the long term it can cause susceptibility to respiratory infections. The excess of nutrient nitrogen can also cause changes in aquatic and marine ecosystems and loss of biodiversity.

Benzene (C6H6)

Benzene is a liquid and colorless chemical substance widely used as a solvent in many industrial and craft activities. The best known effect of chronic exposure concerns the carcinogenic potential of benzene on the hematopoietic system.

There are numerous other agents that affect air quality and for further information we refer the reader to other sources. In the previous section, perhaps the best known were listed and for the project the focus were PM10 and PM2.5 especially due to lack of data.

3.3 Comparison group

Italy was affected by Covid-19 mainly in the months from February to April, which can be considered the spring season. The second step of the analysis therefore involved the search for areas and regions in the world with a climate and humidity level comparable to the Lombardy region, which were affected by Covid in months and seasons different from those in Italy in order to define a forecasting model of infections and deaths for the Lombardy region in other periods.

This search lead to the areas of:

- Covington, Kentucky (USA);
- Fukuoka, Japan;
- Nogoya, Japan;
- Bhisho (South Africa);

- Canberra (Australia);

- Santa Maria, Rio Grande do Sul (Brazil)

As per Italy an analysis that focused on provinces was not possible due to lack of data, therefore the previous list was modified into considering the "regions" or states they belong to and taking into consideration in the computations the population, respectively:

- Kenton County, Kentuky (USA);

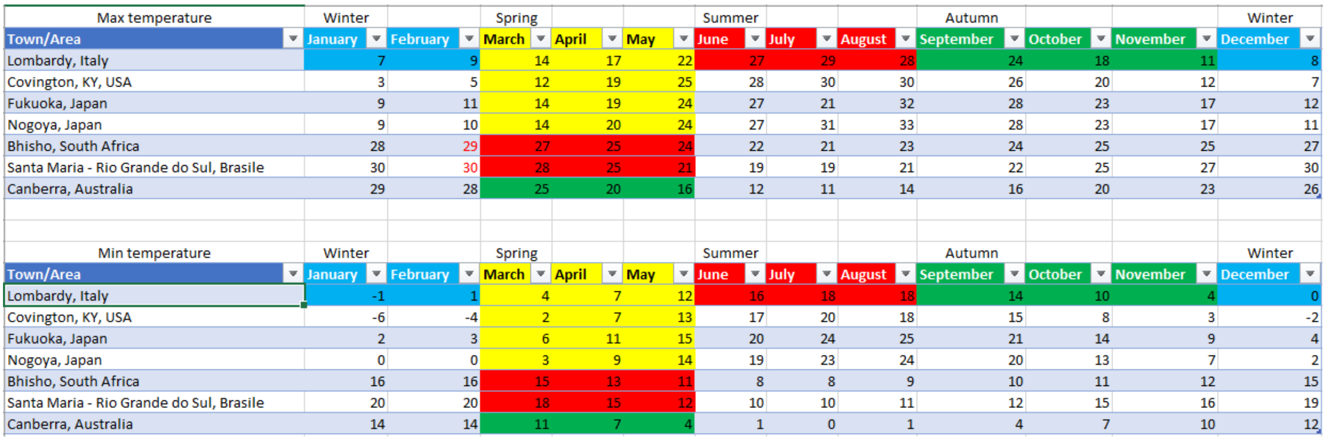
- Fukuoka Prefecture (Japan);

- Aichi Prefecture (Japan);

- Eastern Cape (South Africa);

- Australian Capital Territory (Australia);

- Rio Grande do Sul (Brazil).



Max temperature	Winter		Spring			Summer			Autumn			Winter
Town/Area	January	February	March	April	May	June	July	August	September	October	November	December
Lombardy, Italy	7	9	14	17	22	27	29	28	24	18	11	8
Covington, KY, USA	3	5	12	19	25	28	30	30	26	20	12	7
Fukuoka, Japan	9	11	14	19	24	27	21	32	28	23	17	12
Nogoya, Japan	9	10	14	20	24	27	31	33	28	23	17	11
Bhisho, South Africa	28	29	27	25	24	22	21	23	24	25	25	27
Santa Maria - Rio Grande do Sul, Brasile	30	30	28	25	21	19	19	21	22	25	27	30
Canberra, Australia	29	28	25	20	16	12	11	14	16	20	23	26

Min temperature	Winter		Spring			Summer			Autumn			Winter
Town/Area	January	February	March	April	May	June	July	August	September	October	November	December
Lombardy, Italy	-1	1	4	7	12	16	18	18	14	10	4	0
Covington, KY, USA	-6	-4	2	7	13	17	20	18	15	8	3	-2
Fukuoka, Japan	2	3	6	11	15	20	24	25	21	14	9	4
Nogoya, Japan	0	0	3	9	14	19	23	24	20	13	7	2
Bhisho, South Africa	16	16	15	13	11	8	8	9	10	11	12	15
Santa Maria - Rio Grande do Sul, Brasile	20	20	18	15	12	10	10	11	12	15	16	19
Canberra, Australia	14	14	11	7	4	1	0	1	4	7	10	12

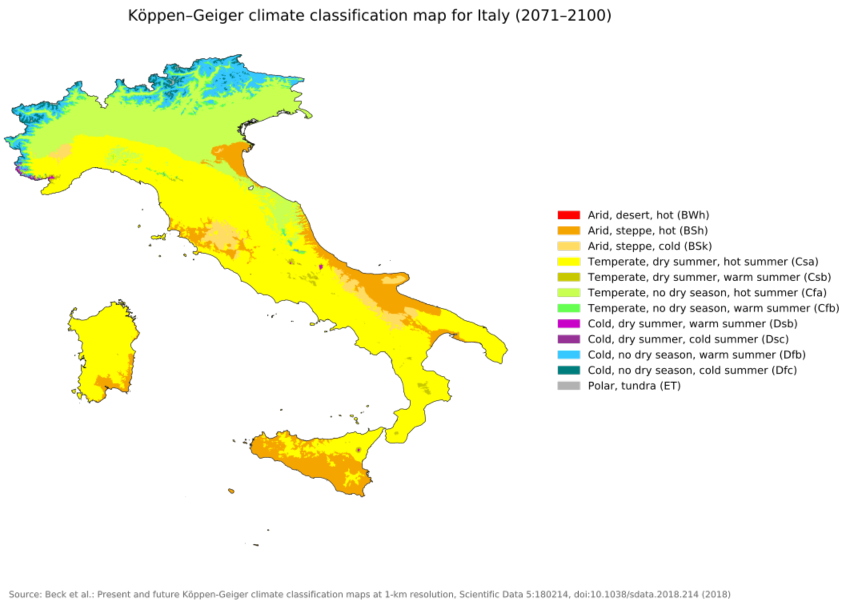
Figure 3.2: Comparison Regions temperatures

And the same data that were collected for the italian regions were collected for these areas too, with the addiction of:

Table 3.2: Macroscopic features with comparison regions

Variable	Type
Koppen Climate classification	Dummy
Total population of the State	Numerical
Contagion rate	Numerical as index

Let's focus on the first and the third new variables added.



Köppen–Geiger climate classification map for Italy (2071–2100)

Arid, desert, hot (BWh)
Arid, steppe, hot (BSh)
Arid, steppe, cold (BSk)
Temperate, dry summer, hot summer (Csa)
Temperate, dry summer, warm summer (Csb)
Temperate, no dry season, hot summer (Cfa)
Temperate, no dry season, warm summer (Cfb)
Cold, dry summer, warm summer (Dsb)
Cold, dry summer, cold summer (Dsc)
Cold, no dry season, warm summer (Dfb)
Cold, no dry season, cold summer (Dfc)
Polar, tundra (ET)

Source: Beck et al.: Present and future Köppen-Geiger climate classification maps at 1-km resolution, Scientific Data 5:180214, doi:10.1038/sdata.2018.214 (2018)

Figure 3.3: Köppen Climate Classification of Italy

3.3.1 Köppen Climate Classification

It was previously stated that the search of the regions to use as comparison for the Lombardy region would have been based on the similarities in the climate. In order to do so the Köppen climate classification, one of the most widely used climate classification systems.

This classification divides climates into five main climate groups based on seasonal precipitation and temperature patterns.[8] Italy has a variety of climate systems, therefore for each region were collected the climate classifications, which could be more than one; they can be considered categorical variables and therefore one-hot encoded, and this was done also for the comparison regions.

3.3.2 Epidemics parameters for contagion and fatality

Modelling mathematically an infectious disease has always been a difficult taks. In order to measure how transferable a disease is

Basic reproduction number R_0

An important parameter in an epidemic of an infectious disease is the so-called R0 or the "basic reproduction number" which represents the average number of secondary infections produced by each infected individual in a population where all individuals are susceptible to infection.[9]

When when $R_0 > 1$ the infection will be able to start spreading in a population and in general the larger this value, the harder it is to control the epidemic. For simple models, the proportion of the population not susceptible to infection, necessary to prevent sustained spread of the infection has to be larger than $1 - \frac{1}{R_0}$.[10]

Effective reproduction number R_t

In reality, varying proportions of the population are immune to any disease at any given time. To account for this, the effective reproduction number R_e is used, usually written as R_t, or the average number of new infections caused by a single infected individual at time t in the partially susceptible population.

3.3.3 Contagion and Fatality rate

Unfortunately neither the R_0 nor the R_t are available for all the regions on which this paper focuses and estimating them was not the aim of the analysis. The authors therefore have decided to "create" two indices, one for contagion and one for fatality that also take into account the population of the region r.

Contagion rate

$$C_r = \frac{\text{number of confirmed contagions to date}}{\text{number of people tested to date}} * \frac{\text{population of the region}}{\text{population of the state they belong to}} * 100 \tag{3.1}$$

Fatality rate

$$C_r = \frac{\text{number of confirmed deaths to date}}{\text{number of confirmed cases to date}} * \frac{\text{population of the region}}{\text{population of the state they belong to}} * 100 \tag{3.2}$$

3.3.4 Issues

As previously mentioned, data collection was not easy: the absence of the daily contagions and death at provincial level and the lack of R_0 and/or the R_t parameters at regional level table were only the first examples.

Another problem with reference to table 3.1 on page 12 was the impossibility of finding the data concerning the health indicators for both Eastern Cape (South Africa) and Rio Grande do Sul. Moreover the data on the number of people tested are not available neither Kentucky nor South Africa. The solution adopted therefore, was quite brutal: derive the number of test performed in Kenton County and Eastern Cape dividing the number of tests performed to date in respectively, Kentucky and South Africa, by the number of states. For example in South Africa the number of tests performed is 90515 and there are 9 states, therefore to obtain the number of tests per each state we simply divided the first value by the second.

Due to the lack of data on the "health" of the region (namely fatality rate for stroke, tumors, etc.) the regression was not based on these aspects, as well as the clustering.

Chapter 4

Clustering

Clustering is an unsupervised machine learning technique that given a set of data points performs a grouping on them. There are various clustering algorithms based on the cluster method they adopt. The ones used in this project are the hierarchical and the k-means algorithms.

4.1 Hierarchical Clustering

Hierarchical clustering needs a distance matrix to be performed and it starts by treating each observation in a separate cluster. Then it identifies the two clusters that are closest to each other and merges the two most similar clusters together; the distance between two clusters is computed.

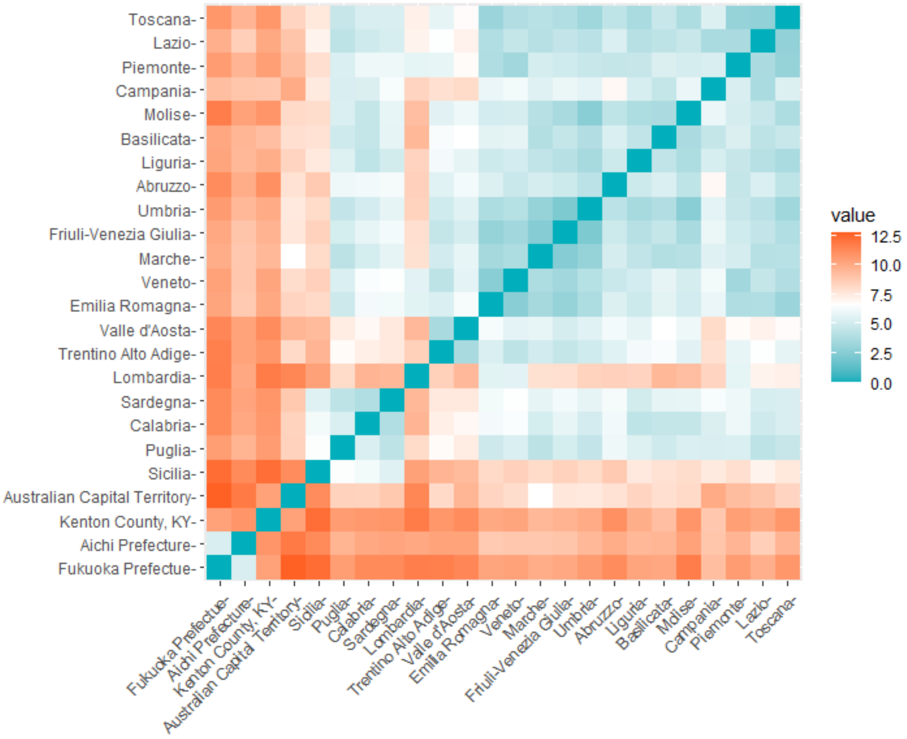


Figure 4.1: Distance matrix without Eastern Cape and Rio Grande do Sul

The clusters formed in this method form a tree-type structure based on the hierarchy. Through this technique we obtained visible clusters of trusts.[11] The criterion applied to the hierarchical clustering in this project is the Ward's minimum variance criterion that minimizes the total within-cluster variance.

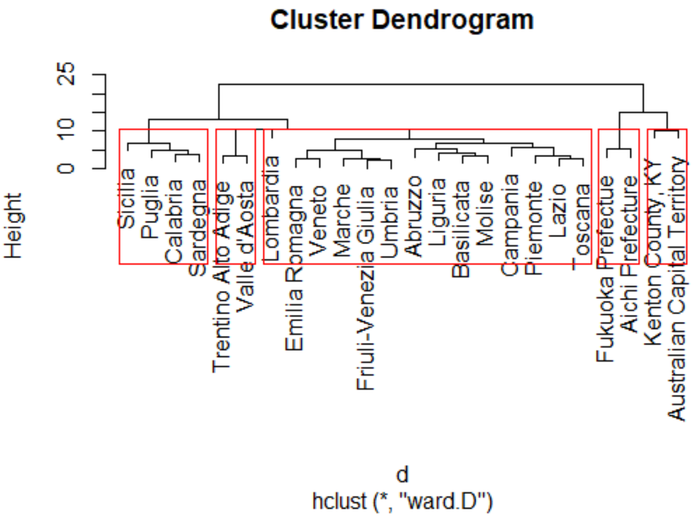


Figure 4.2: Dendrogram resulting from the hierarchical clustering

4.2 K-means algorithm

K-means clustering iteratively partitions a given data set into k groups or clusters. Objects belonging to the same cluster are as similar as possible, while the different groups are as dissimilar as possible. Each cluster has a center called centroid. Although we run the k-means algorithm with 4 clusters, we used and considered the results of the hierarchical clustering.

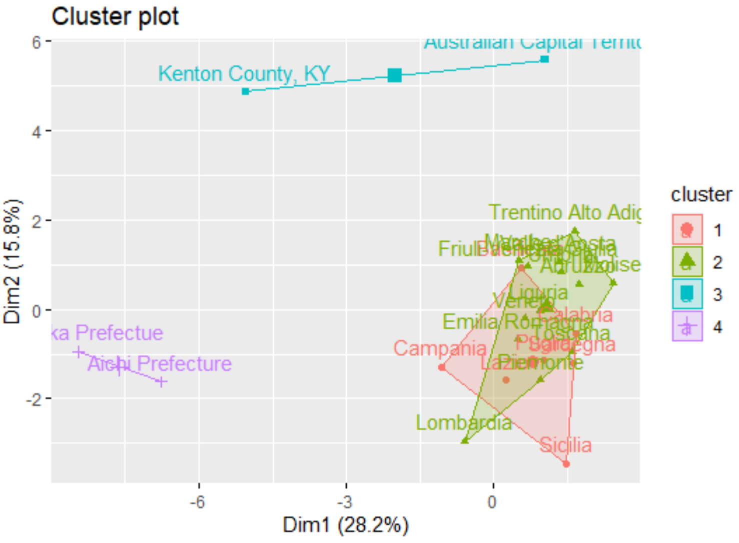


Figure 4.3: K-means algorithm output

4.3 Clustering results

It seems useful to recall that the clustering analysis was performing without taking into consideration Eastern Cape and Rio Grande do sul because of their lack of data on the "health" of the states.

It is also important to note that the number k of clusters has to be decided before the implementation of the k-means algorithm, in this case k was set to 4 but as it is possible to observe that there is a bit of overlap between two clusters in particular.

From the analysis the two prefectures of Aichi and Fukuoka form the first cluster, Kenton County and ACT the second one while Italy could be seen as a sole group (if we take into account the overlap of the two clustes) or as two clusters: one containing Basilicata, Lazio, Campania, Sicilia and Sardegna and the other one formed by the remaining regions.

Chapter 5

Regression

5.1 The Data-set

The regression analysis aims at finding a correlation between the contagion rate first, and the fatality rate second, and the pollution level controlling for some macroscopic characteristics of the regions of interest.

In particular the variables composing the data-set built to perform this analysis and the logic behind their choice is the following:

- Area/Region;
- State;
- Contagion rate;
- Fatality rate;
- BSh Koppen classification;
- BSk Koppen classification;
- Csa Koppen classification;
- Csb Koppen classification;
- Cfa Koppen classification;
- Cfb Koppen classification;
- Dfb Koppen classification;
- Dfc Koppen classification;
- ET Koppen classification;
- Unemployment rate, used as a proxy for the mobility within the region/area;
- Population (inhabitants) of the region;
- Area (km2);
- Density of population;

- PM10 micrograms/cubic meter;
- PM25 micrograms/cubic meter;
- CAQI index;
- Hospital beds, used as a proxy for the capacity of the local health system;
- Heart disease fatality rate, used as a proxy for the general health of the population;
- Stroke fatality rate, used as a proxy for the general health of the population;
- Malignant tumors fatality rate, used as a proxy for the general health of the population;
- Tracheal, Bronchus and Lung Cancer fatality rate, used as a proxy for the percentage of smokers;
- Influenza and Pneumonia fatality rate;
- Chronic Lower Respiratory Disease fatality rate, used as a proxy for the general health of the population;
- Tuberculosis fatality rate.

5.2 The analysis

As it was previously stated, the data-set created for the regression analysis does not contain the observations for the variables of interest for all the regions, in particular for Eastern Cape and Rio Grande do Sul.

To deal with this issue the authors decided to split the original data-set and therefore the analysis in two parts: the first data-set will contain all the variables but it won't consider either Eastern Cape or Rio Grande do Sul.

The second part will contain all the regions of interest but it won't consider the variables for which Eastern Cape and Rio Grande do Sul do not have observations, namely:

- Heart disease fatality rate;
- Stroke fatality rate;
- Malignant tumors fatality rate;
- Tracheal, Bronchus and Lung Cancer fatality rate;
- Influenza and Pneumonia fatality rate;
- Chronic Lower Respiratory Disease fatality rate;
- Tuberculosis fatality rate.

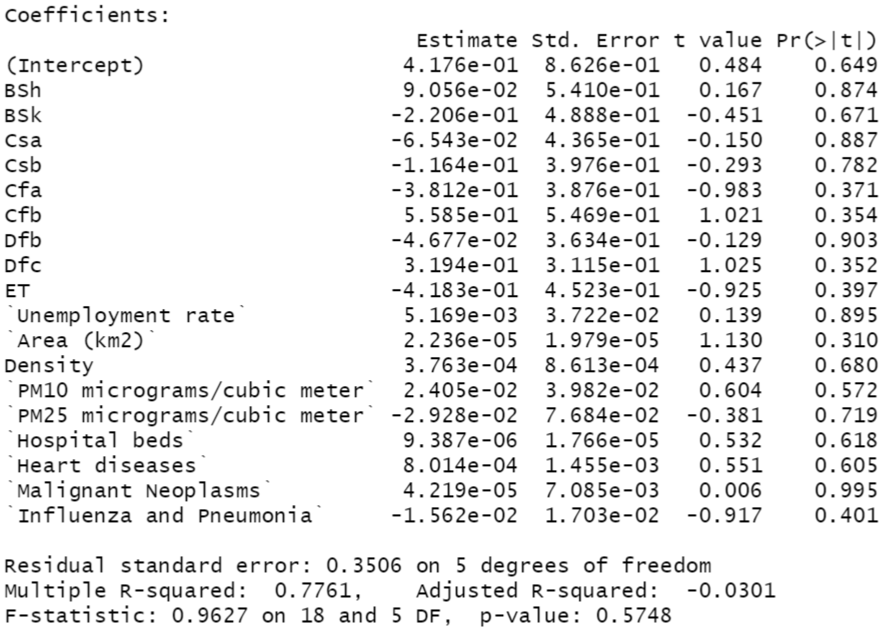
5.3 Analysis on the data-set without Eastern Cape and Rio Grande

5.3.1 Regression analysis of the Contagion rate

As for what concerns the analysis of the contagion rate three simple regressions were run with respect to the density of population, the CAQI index (for air quality) and the unemployment rate (for the mobility within the region/area). However no significant correlation was found.

Multiple Regression

The following steps involved Multiple regression, namely regression on multiple variables and in particular both the Population variable and the Fatality rate (of Covid-19) were excluded from the analysis. It is also important to say that also the correlations between the variables contained in the data-set were taken into account, for example: the Stroke fatality rate is a subclass of the Heart Diseases fatality rate, as the Tracheal, Bronchus and Lung Cancer fatality rate is a subclass of the Malignant tumors fatality rate. This means that when running the regression with one of them the other one was not included in the model.



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.176e-01  8.626e-01   0.484    0.649
BSh                          9.056e-02  5.410e-01   0.167    0.874
BSk                         -2.206e-01  4.888e-01  -0.451    0.671
Csa                         -6.543e-02  4.365e-01  -0.150    0.887
Csb                         -1.164e-01  3.976e-01  -0.293    0.782
Cfa                         -3.812e-01  3.876e-01  -0.983    0.371
Cfb                          5.585e-01  5.469e-01   1.021    0.354
Dfb                         -4.677e-02  3.634e-01  -0.129    0.903
Dfc                          3.194e-01  3.115e-01   1.025    0.352
ET                          -4.183e-01  4.523e-01  -0.925    0.397
`Unemployment rate`          5.169e-03  3.722e-02   0.139    0.895
`Area (km2)`                 2.236e-05  1.979e-05   1.130    0.310
Density                      3.763e-04  8.613e-04   0.437    0.680
`PM10 micrograms/cubic meter`  2.405e-02  3.982e-02   0.604    0.572
`PM25 micrograms/cubic meter` -2.928e-02  7.684e-02  -0.381    0.719
`Hospital beds`              9.387e-06  1.766e-05   0.532    0.618
`Heart diseases`             8.014e-04  1.455e-03   0.551    0.605
`Malignant Neoplasms`        4.219e-05  7.085e-03   0.006    0.995
`Influenza and Pneumonia`   -1.562e-02  1.703e-02  -0.917    0.401

Residual standard error: 0.3506 on 5 degrees of freedom
Multiple R-squared:  0.7761,    Adjusted R-squared:  -0.0301
F-statistic: 0.9627 on 18 and 5 DF,  p-value: 0.5748
```

Figure 5.1: Contagion rate regression without CAQI index, Stroke fatality rate, Tracheal cancer fatality rate and Chronic fatality rate

Results

No significant correlation was found between any of the variables showing that the contagion rate, as it was defined by the authors of this project, is not caused by or dependent of any of the characteristics investigated here.

5.3.2 Regression analysis of the Fatality rate

The regression analysis of Fatality rate did not take into consideration the Population of the region but it seems logical that fatality rate might be caused or related with the contagion rate and as a matter of fact, running a regression model with respect to the latter it is evident a positive and significant relation.

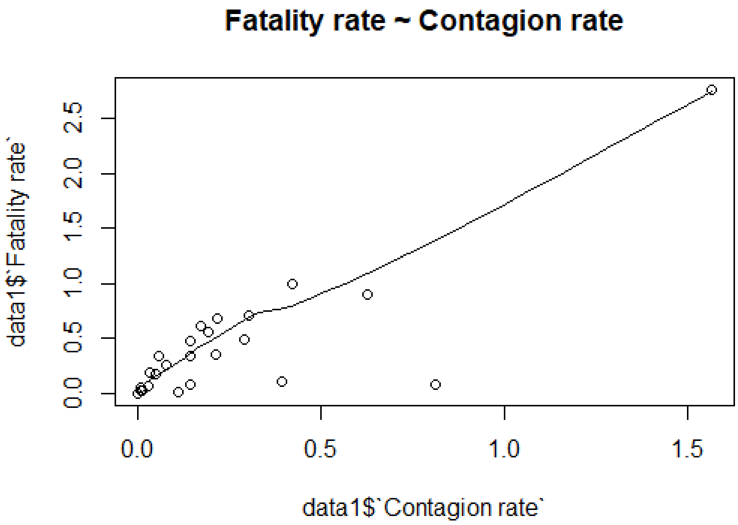
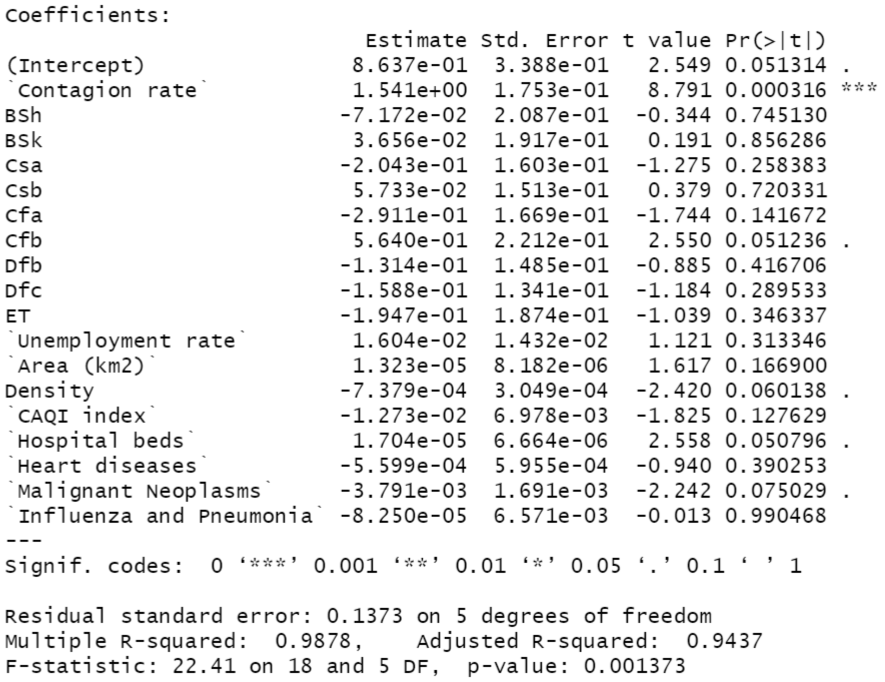


Figure 5.2: Plot of the regression model of the fatality rate on the contagion rate

Multiple Regression

Unlike the contagion analysis, the one on fatality leads to more interesting results: as a matter of fact it is possible to observe that the fatality rate significantly depends on many factors.

Replacing the CAQI Index with the PM10 and the PM25, the Heart disease fatality rate with the Stroke fatality rate and sub-setting the previous data-set to consider only the variables on which the fatality rate is dependent, it is possible to observe the following:



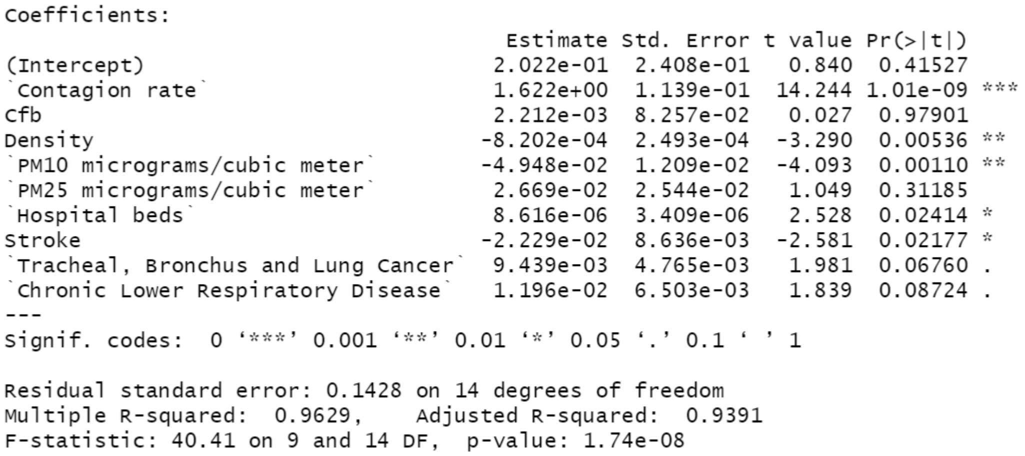
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  8.637e-01  3.388e-01   2.549 0.051314 .  
`Contagion rate`             1.541e+00  1.753e-01   8.791 0.000316 ***
BSh                         -7.172e-02  2.087e-01  -0.344 0.745130    
BSk                          3.656e-02  1.917e-01   0.191 0.856286    
Csa                         -2.043e-01  1.603e-01  -1.275 0.258383    
Csb                          5.733e-02  1.513e-01   0.379 0.720331    
Cfa                         -2.911e-01  1.669e-01  -1.744 0.141672    
Cfb                          5.640e-01  2.212e-01   2.550 0.051236 .  
Dfb                         -1.314e-01  1.485e-01  -0.885 0.416706    
Dfc                         -1.588e-01  1.341e-01  -1.184 0.289533    
ET                          -1.947e-01  1.874e-01  -1.039 0.346337    
`Unemployment rate`          1.604e-02  1.432e-02   1.121 0.313346    
`Area (km2)`                 1.323e-05  8.182e-06   1.617 0.166900    
Density                     -7.379e-04  3.049e-04  -2.420 0.060138 .  
`CAQI index`                -1.273e-02  6.978e-03  -1.825 0.127629    
`Hospital beds`              1.704e-05  6.664e-06   2.558 0.050796 .  
`Heart diseases`            -5.599e-04  5.955e-04  -0.940 0.390253    
`Malignant Neoplasms`       -3.791e-03  1.691e-03  -2.242 0.075029 .  
`Influenza and Pneumonia`   -8.250e-05  6.571e-03  -0.013 0.990468    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1373 on 5 degrees of freedom
Multiple R-squared:  0.9878,	Adjusted R-squared:  0.9437 
F-statistic: 22.41 on 18 and 5 DF,  p-value: 0.001373
```

Figure 5.3: Regression model of the fatality rate on the data-set without: PM10, PM25, stroke, Tracheal cancer, Chronic disease of the lower respiratory system and tuberculosis

Results

The analysis shows a clear dependence of the fatality rate by Covid-19 on the Contagion rate (as it was defined by the authors), the population density, the PM10 levels, the hospital beds capacity (and therefore by the effectiveness of the healthcare system), the fatality rate by stroke and by chronic lower respiratory disease (and by "induction" on the health of the population) and finally on the fatality rate of tracheal, bronchus and lung cancer (namely on the smoking percentage).



```
Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            2.022e-01  2.408e-01   0.840  0.41527    
`Contagion rate`                       1.622e+00  1.139e-01  14.244 1.01e-09 ***
Cfb                                    2.212e-03  8.257e-02   0.027  0.97901    
Density                               -8.202e-04  2.493e-04  -3.290  0.00536 ** 
`PM10 micrograms/cubic meter`         -4.948e-02  1.209e-02  -4.093  0.00110 ** 
`PM25 micrograms/cubic meter`          2.669e-02  2.544e-02   1.049  0.31185    
`Hospital beds`                        8.616e-06  3.409e-06   2.528  0.02414 *  
Stroke                                -2.229e-02  8.636e-03  -2.581  0.02177 *  
`Tracheal, Bronchus and Lung Cancer`   9.439e-03  4.765e-03   1.981  0.06760 .  
`Chronic Lower Respiratory Disease`    1.196e-02  6.503e-03   1.839  0.08724 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1428 on 14 degrees of freedom
Multiple R-squared:  0.9629,	Adjusted R-squared:  0.9391 
F-statistic: 40.41 on 9 and 14 DF,  p-value: 1.74e-08
```

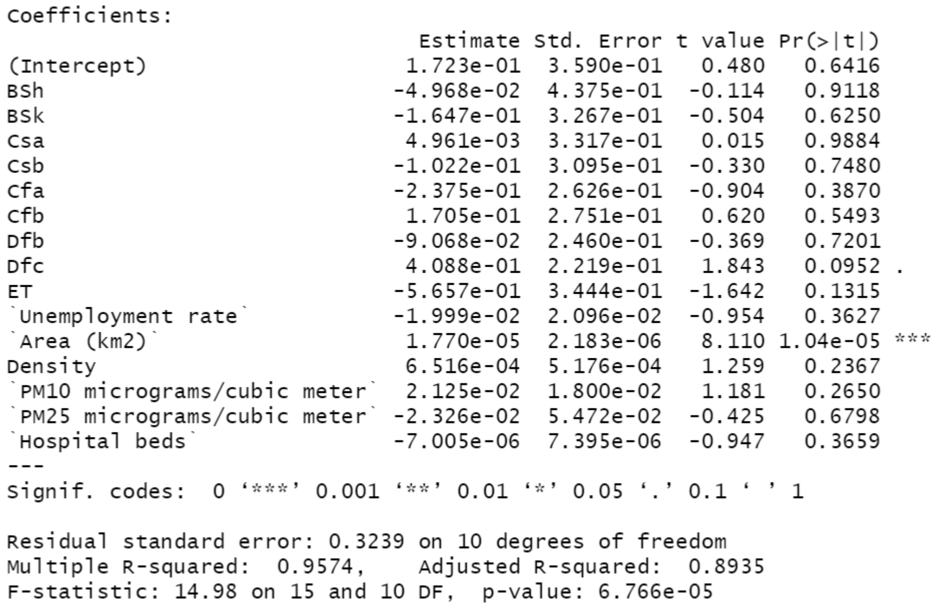
Figure 5.4: Regression model of the fatality rate on the data-set composed by: Contagion rate, Cfb, Density, PM10, PM25, Stroke, Tracheal cancer, Chronic disease of the lower respiratory system

5.4 Analysis on the incomplete data-set containing Eastern Cape and Rio Grande

The analysis on the alternative data-set that takes into consideration also the observations on Eastern Cape and Rio Grande do sul but lacking all the variables on the "health" of the population, followed and the steps are quite similar to those for the previous data-set.

5.4.1 Regression analysis of the Contagion rate

As for what concerns the analysis performed on Contagion rate, the results seem to confirm the previous results with the exception of a mild dependence on the Dfc Koppen climate classification and on the area of the region.



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  1.723e-01  3.590e-01   0.480   0.6416    
BSh                         -4.968e-02  4.375e-01  -0.114   0.9118    
BSk                         -1.647e-01  3.267e-01  -0.504   0.6250    
Csa                          4.961e-03  3.317e-01   0.015   0.9884    
Csb                         -1.022e-01  3.095e-01  -0.330   0.7480    
Cfa                         -2.375e-01  2.626e-01  -0.904   0.3870    
Cfb                          1.705e-01  2.751e-01   0.620   0.5493    
Dfb                         -9.068e-02  2.460e-01  -0.369   0.7201    
Dfc                          4.088e-01  2.219e-01   1.843   0.0952 .  
ET                          -5.657e-01  3.444e-01  -1.642   0.1315    
`Unemployment rate`         -1.999e-02  2.096e-02  -0.954   0.3627    
`Area (km2)`                 1.770e-05  2.183e-06   8.110 1.04e-05 ***
Density                      6.516e-04  5.176e-04   1.259   0.2367    
`PM10 micrograms/cubic meter`  2.125e-02  1.800e-02   1.181   0.2650    
`PM25 micrograms/cubic meter` -2.326e-02  5.472e-02  -0.425   0.6798    
`Hospital beds`             -7.005e-06  7.395e-06  -0.947   0.3659    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3239 on 10 degrees of freedom
Multiple R-squared:  0.9574,	Adjusted R-squared:  0.8935 
F-statistic: 14.98 on 15 and 10 DF,  p-value: 6.766e-05
```

Figure 5.5: Regression model of the contagion rate on the entire data-set without CAQI Index and Population (inhabitants)

5.4.2 Regression analysis of the Fatality rate

The Fatality rate analysis started again with a regression of the fatality rate by Covid-19 on the contagion rate. In this case there does not seem to be a significant dependence between the two variables: as Figure 6.5 shows the regression curve seems to be more normal-like although left-skewed and with many outliers. Not even the regressions run with respect to the CAQI Index, on the density and on the Hospital beds seem to result in a significant dependence.

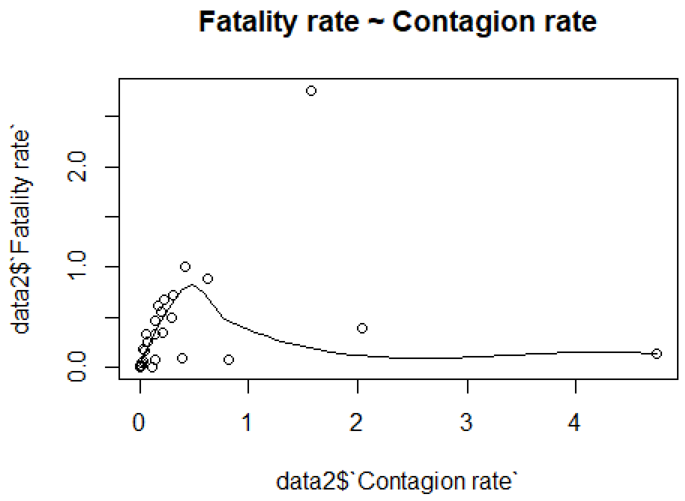
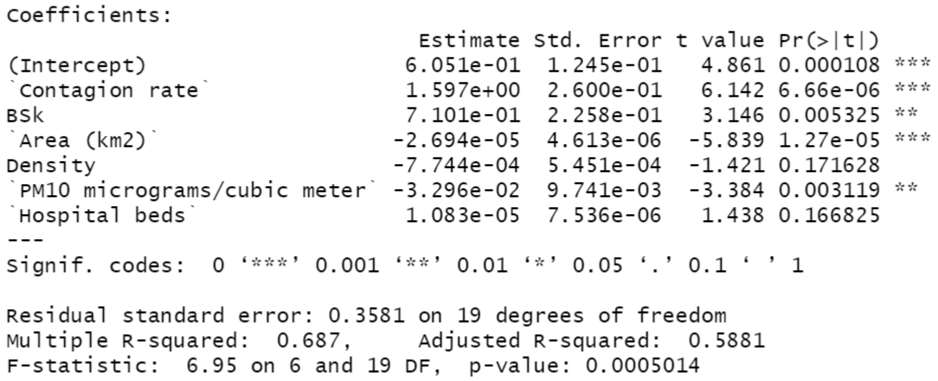


Figure 5.6: Regression model of the fatality rate on the contagion rate, considering Eastern Cape and Rio Grande

Multiple Regression

Running the Multiple Regression on the full data-set (without the CAQI Index and of course, the Population) seems to lead to much more significant results.



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  6.051e-01  1.245e-01   4.861 0.000108 ***
`Contagion rate`             1.597e+00  2.600e-01   6.142 6.66e-06 ***
BSk                          7.101e-01  2.258e-01   3.146 0.005325 ** 
`Area (km2)`                -2.694e-05  4.613e-06  -5.839 1.27e-05 ***
Density                     -7.744e-04  5.451e-04  -1.421 0.171628    
`PM10 micrograms/cubic meter` -3.296e-02  9.741e-03  -3.384 0.003119 ** 
`Hospital beds`              1.083e-05  7.536e-06   1.438 0.166825    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3581 on 19 degrees of freedom
Multiple R-squared:  0.687,	Adjusted R-squared:  0.5881 
F-statistic:  6.95 on 6 and 19 DF,  p-value: 0.0005014
```

Figure 5.7: Regression model of the fatality rate on the contagion rate, BSk Koppen classification, PM10, hospital beds considering Eastern Cape and Rio Grande

Results

It is possible to observe from Figure 6.7 a quite strong and significant dependence between the Fatality rate of Covid-19, the Contagion rate, the BSk Koppen climate classification, on the Area of the region and on the levels of PM10.

Chapter 6

Time Series Analysis with ARIMA models on the daily Contagions

6.1 Considerations

For the regions under analysis were collected the data on the daily contagions and deaths starting from the 31st of January 2020, the day when the first cases of Covid-19 in Italy were confirmed, also because none of the other regions reports previous data. Therefore since these data are reported over time, they will be time series.

Again, we refer the reader to other sources to deepen the topic but a brief introduction on the steps followed in this analysis will follow.

6.2 Time series and ARIMA models

Time series data are data points collected over a period of time as a sequence of time gap and among the main time series analysis techniques there is that of ARIMA modelling.

ARIMA is the abbreviation for AutoRegressive Integrated Moving Average where, Auto Regressive (AR) refer to the lags of the differenced series while Moving Average (MA) refers to the lags of errors and ”I” is the number of difference used to make the time series stationary.

There are three components to a time series:

- trend, how things are overall changing;
- seasonality, how things change within a given period, for example a year, month or week;
- error or residuals, not explained by the trend or the seasonal value.[12]

Based on how these three elements combine it is possible to distinguish between multiplicative and additive time series. In general having an additive model would be preferred but it is possible to convert a multiplicative time series into an additive one by taking a log of the time series.

6.2.1 Stationarity

The first step to perform time series analysis is checking for stationarity, otherwise it is not possible to build a time series model.

Stationarity means that the statistical properties of a process generating a time series do not change over time.[13] Trends or seasonality will affect the value of the time series at different times.

Differencing

Computing the differences between consecutive observations is one possibility and it is known as differencing, which stabilises the mean of the time series by removing changes in the level of a time series, reducing both trend and seasonality. Logarithmic transformation may instead help to stabilise the variance of a time series.[14]

Auto-Correlation Function (ACF) plots

Another way to detect a non-stationary time series is looking at the Auto-correlation plot (ACF). For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly.

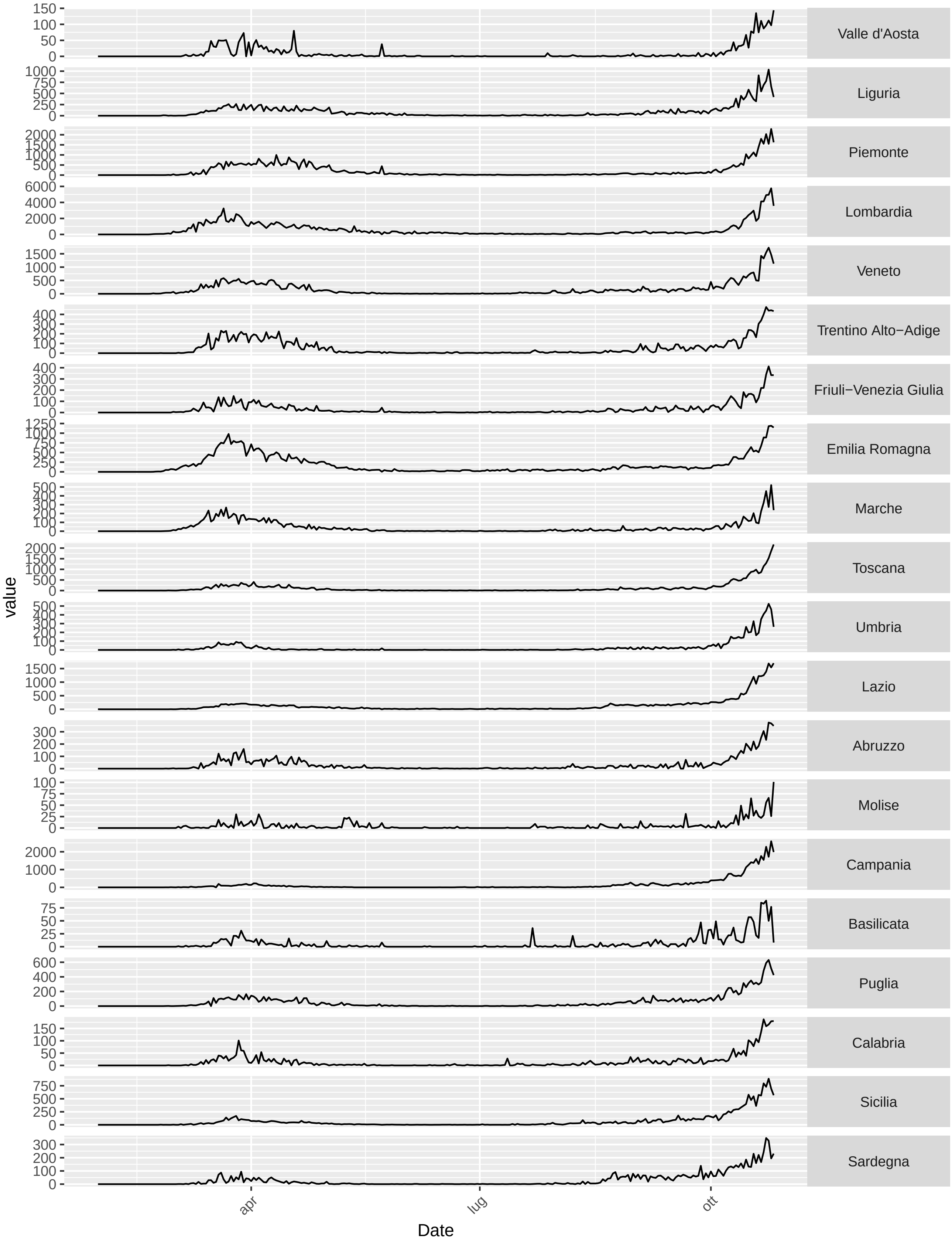
Unit Root Test

Unit root tests are used to be more objectively sure whether differencing is required is to use a unit root test. One of the most used test is the we use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test: in this test, the null hypothesis is that the data are stationary therefore small p-values (for example less than 0.05) suggest that differencing is required.

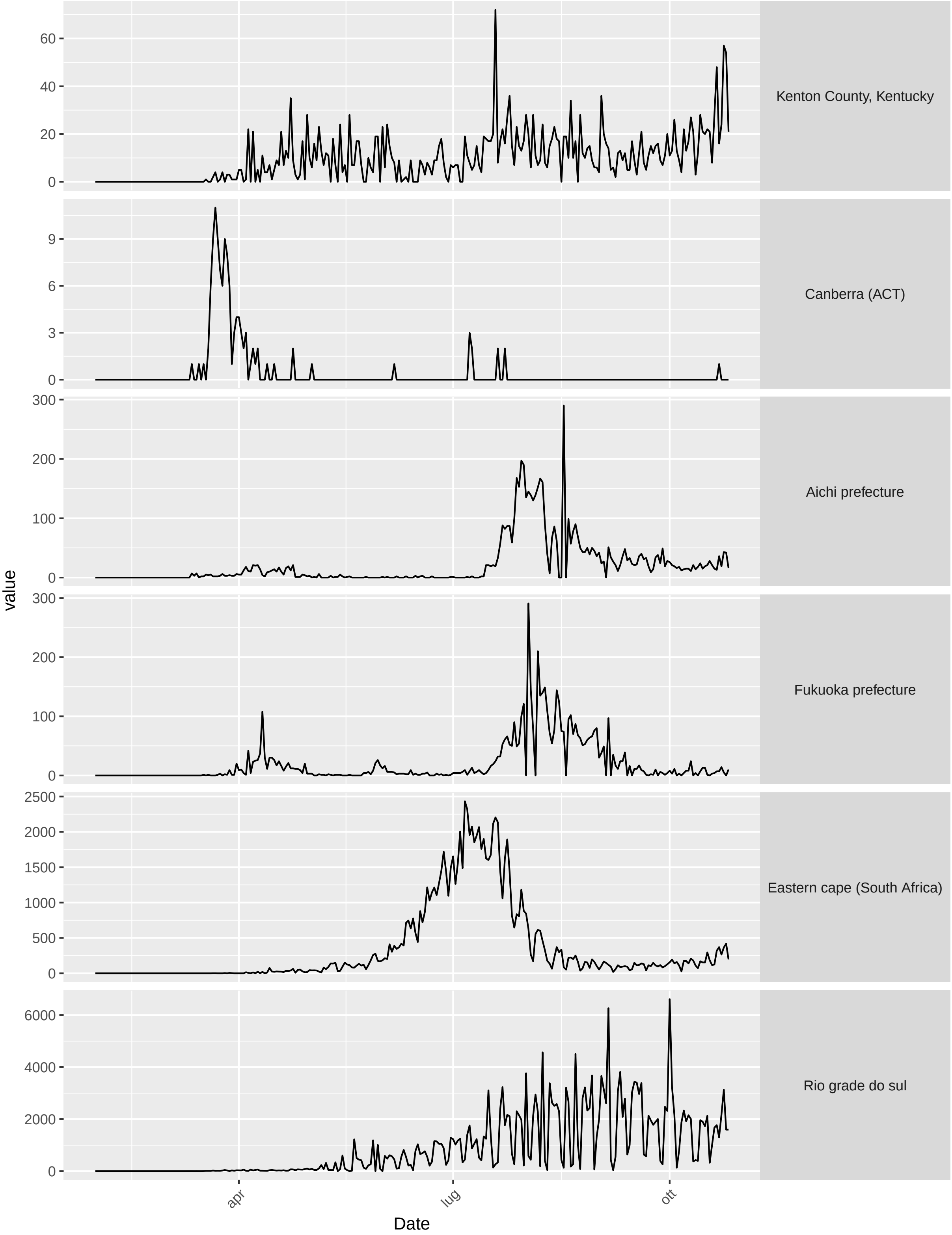
6.2.2 ARIMA modelling

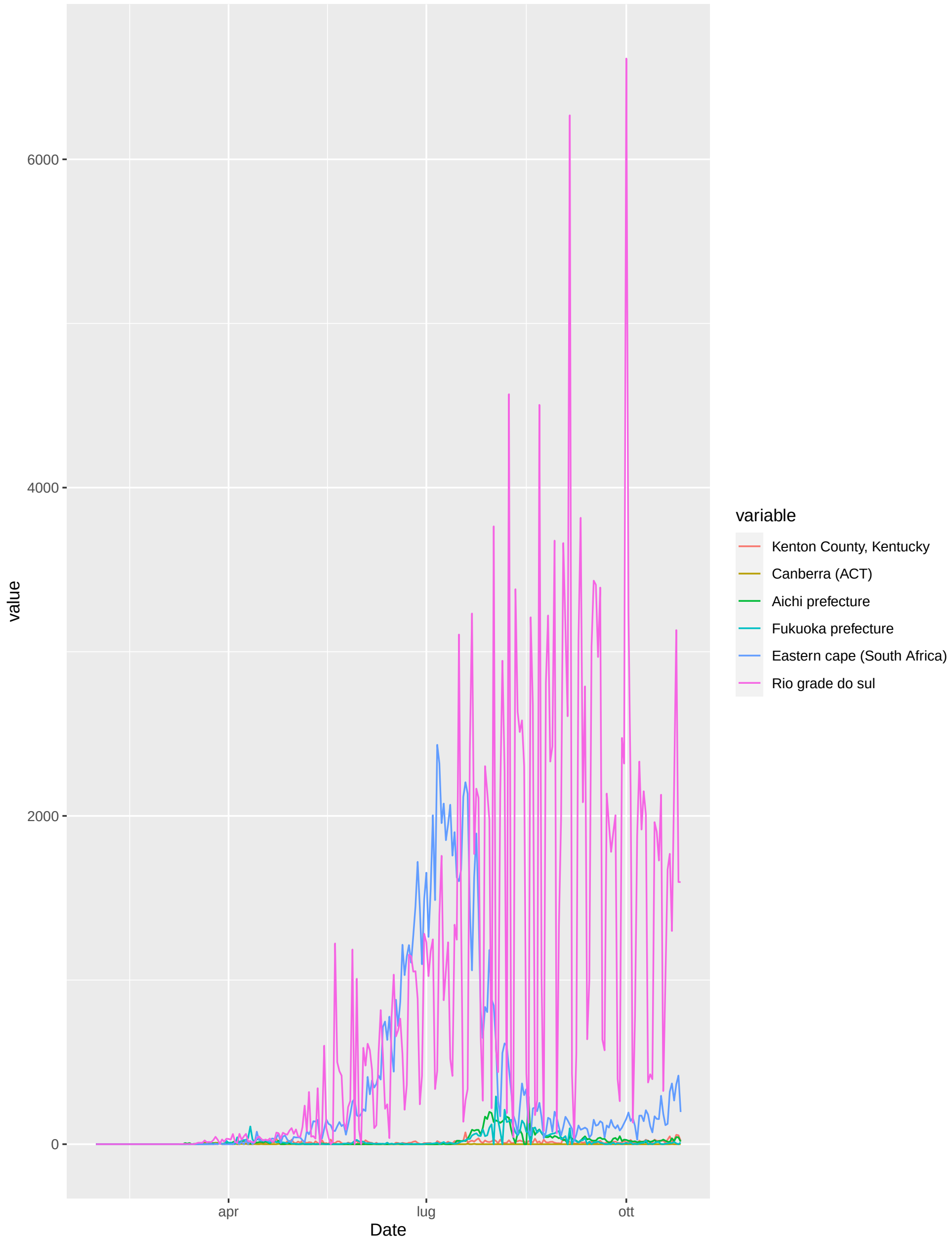
The previous parenthesis on stationarity was much needed since one of the assumptions at the basis of the implementation of ARIMA models is that data should be stationary.

It is quite important to say that the data at disposal for a time series analysis do not even reach one year, which makes it difficult to investigate both trends and seasonality, and the time series analysis itself.



Valle d'Aosta
Liguria
Piemonte
Lombardia
Veneto
Trentino Alto−Adige
Friuli−Venezia Giulia
Emilia Romagna
Marche
Toscana
Umbria
Lazio
Abruzzo
Molise
Campania
Basilicata
Puglia
Calabria
Sicilia
Sardegna
value
Date
apr
lug
ott


Kenton County, Kentucky
Canberra (ACT)
Aichi prefecture
Fukuoka prefecture
Eastern cape (South Africa)
Rio grade do sul
value
Date
apr
lug
ott
0
20
40
60
0
3
6
9
0
100
200
300
0
100
200
300
0
500
1000
1500
2000
2500
0
2000
4000
6000


value
6000
4000
2000
0
apr
lug
ott
Date
variable
Kenton County, Kentucky
Canberra (ACT)
Aichi prefecture
Fukuoka prefecture
Eastern cape (South Africa)
Rio grade do sul

Chapter 7

Conclusion

Due to the novelty of the Covid-19 and beyond this project has undergone many changes in progress.

The initial aim was to investigate the presence of a correlation between the number of infections and deaths and the quality of the air with a focus on Italy alone, at the provincial level. This analysis was motivated by the evident different impact that the Covid-19 had not only between the regions, but between the provinces of the same region. However the lack of data for all the complete Italian provinces has diverted this paper towards a regional perspective.

The case of Lombardy (and Bergamo and Milan above all), has then determined a strong interest not only in the authors of this project, in wanting to investigate the reasons for this impact as unexpected as it is terrible.

This reasoning has further modified the ultimate purpose of the analysis which led to the search for areas and regions in other states that, having climatic conditions similar to Lombardy (for humidity and temperature), were affected by the Covid-19 in periods of the year that may be comparable to other seasons other than spring, during which the pandemic hit the Italian region. And so it led us to analyze the case of Kenton County, Kentucky, the prefectures of Aichi and Fukuoka in Japan, the Australian capital territory (in Australia), the Eastern Cape province in South Africa and Rio grande do sul in. Brazil, to encounter new problems in finding the data necessary for the analysis, the homogeneity and also, unfortunately, the reliability of such data.

This collection was intended not only to investigate the presence of a correlation between Covid and air quality, which as seen in chapter 5 exists only as regards the fatality rate and must be combined with the one that has been defined as the state of health of the population of the regions under analysis. But also to try to predict the future trend of the Lombardy contagion curve, in case some similarity with those of the other regions under analysis were found. As it is possible to observe from the plots in chapter 6 however, such a comparison is not possible since the contagion curves of the regions, although the similar temperature, humidity and in general climate conditions might have told us differently, are completely different.

Furthermore, a complete time series analysis would require at least a year of data which, however, no one currently has. The authors' intentions would therefore be to continue this analysis, of which the foundations have already been laid, when we will have more and above all more certain information on the current pandemic so as to provide a more complete and equally reliable investigation.

Bibliography

[1] Statement on the second meeting of the International Health Regulations, Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)(30/01/2020), https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov)

[2] Background on Covid-19, Wikipedia, https://en.wikipedia.org/wiki/COVID-19$_{p}andemicEpidemiology$

[3] Nuovo coronavirus, Ministero della Salute, http://www.salute.gov.it/portale/nuovocoronavirus/dettaglioFaqNuovoCoronavirus.jsp?lingua=italianoid=228

[4] Imaging, Wikipedia, https://en.wikipedia.org/wiki/COVID-19$_{p}andemicTransmission$

[5] CAQI Air quality index, https://www.airqualitynow.eu/download/CITEAIR-Comparing$_{U}rban_{A}ir_{Q}uality_{a}cross_{B}orders.pdf$

[6] Qualità dell'aria ed effetti sulla salute: un problema sempre attuale, ARS Toscana, https://www.ars.toscana.it/aree-dintervento/determinanti-di-salute/ambiente/news/3297-inquinamento-qualita-dell-aria-e-effetti-sulla-salute-un-problema-sempre-attuale.html

[7] Inquinanti e Indice della Qualità dell´Aria (IQA), https://www.arpae.it/dettaglio$_{g}enerale.asp?id = 3883idlivello = 2074$

[8] Koppen Climate Classification, Wikipedia, https://en.wikipedia.org/wiki/K%C3%B6ppen$_{c}limate_{c}lassificationCfa :_{H} umid_{s}ubtropical_{c}limates$

[9] Che cos'è R0 e perché è così importante, Istituto Superiore di Sanità, https://www.iss.it/primo-piano/-/asset$_{p}ublisher/o4oGR9qmvUz9/content/id/$5268851

[10] Basic reproduction number, Wikipedia, https://en.wikipedia.org/wiki/Basic _reproduction_number

[11] George Seif, "The 5 Clustering Algorithms Data Scientists Need to Know". In: towards data science, (2018), https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-knowa36d136ef68

[12] Selva Prabhakaran, Time Series Analysis, http://r-statistics.co/Time-Series-Analysis-With-R.html

[13] Shay Palachy, Stationarity in time series analysis (2017),https://towardsdatascience.com/stationarity-in-time-series-analysis-90c94f27322

[14] Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of Econometrics, 54(1-3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y