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 15, 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 whene 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 ;
- inderectly (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.

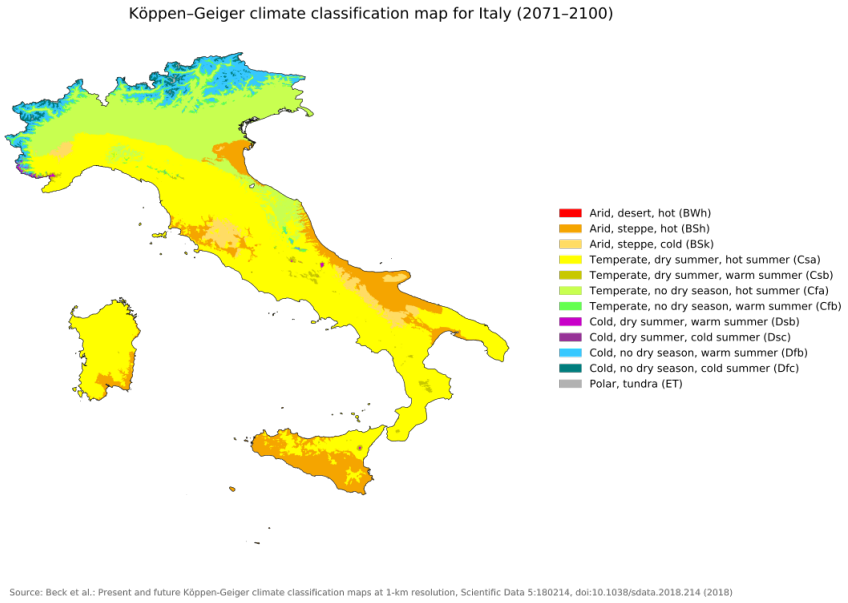


Figure 3.3: Köppen Climate Classification of Italy

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

$$C_r = \frac{\text{number of confirmed contagions to date}}{\text{number of people tested to date}} * \frac{\text{population of the region}}{\text{population of the state they belong to}} * \frac{1}{100} \tag{3.1}$$

Fatality rate

$$C_r = \frac{\text{number of confirmed deaths to date}}{\text{number of confirmed cases to date}} * \frac{\text{population of the region}}{\text{population of the state they belong to}} * \frac{1}{100} \tag{3.2}$$

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 11 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.

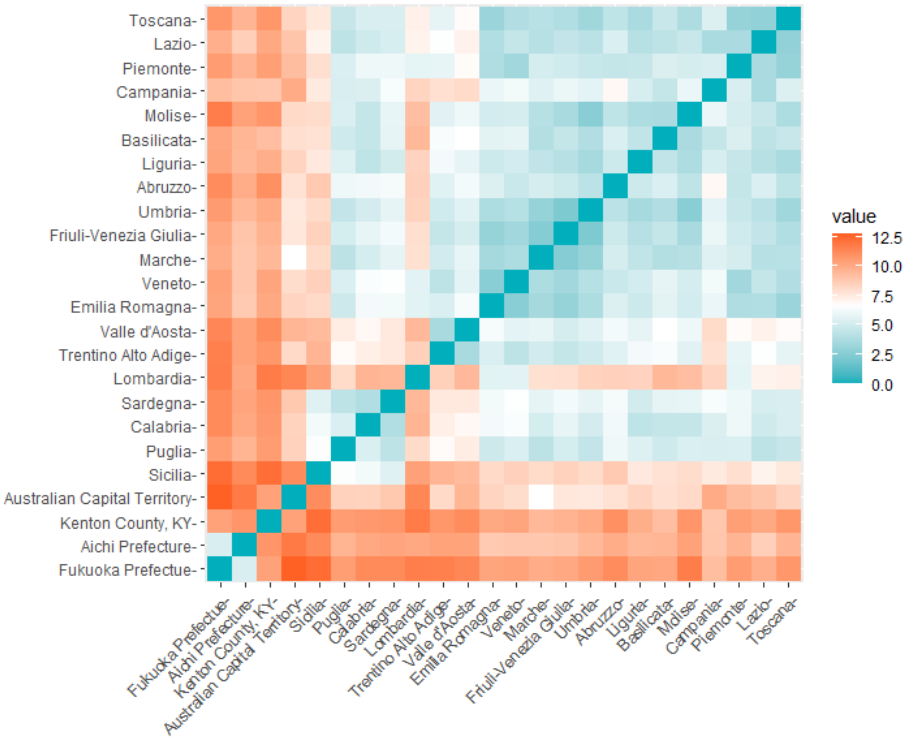


Figure 4.1: Distance matrix without Eastern Cape and Rio Grande do Sul

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.

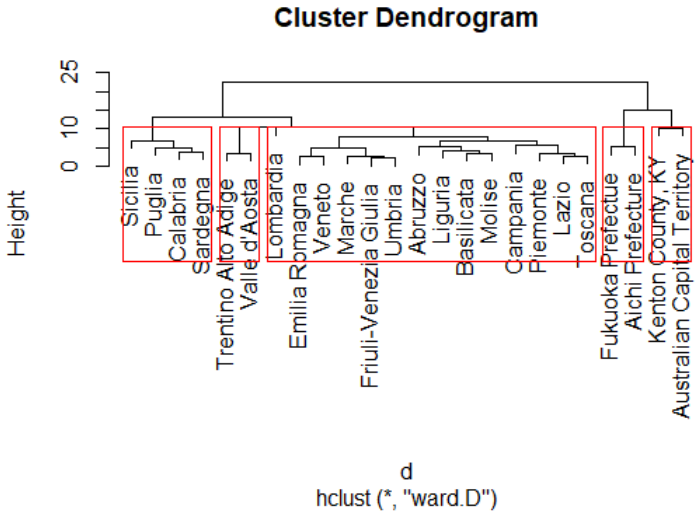


Figure 4.2: Dendrogram resulting from the hierarchical clustering

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.

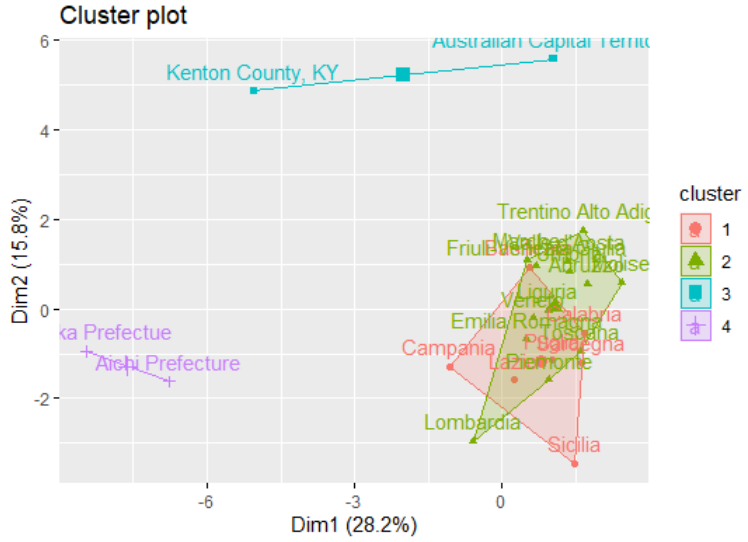


Figure 4.3: K-means algorithm output

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

Time Series Analysis

5.1 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.

5.2 Time series

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]

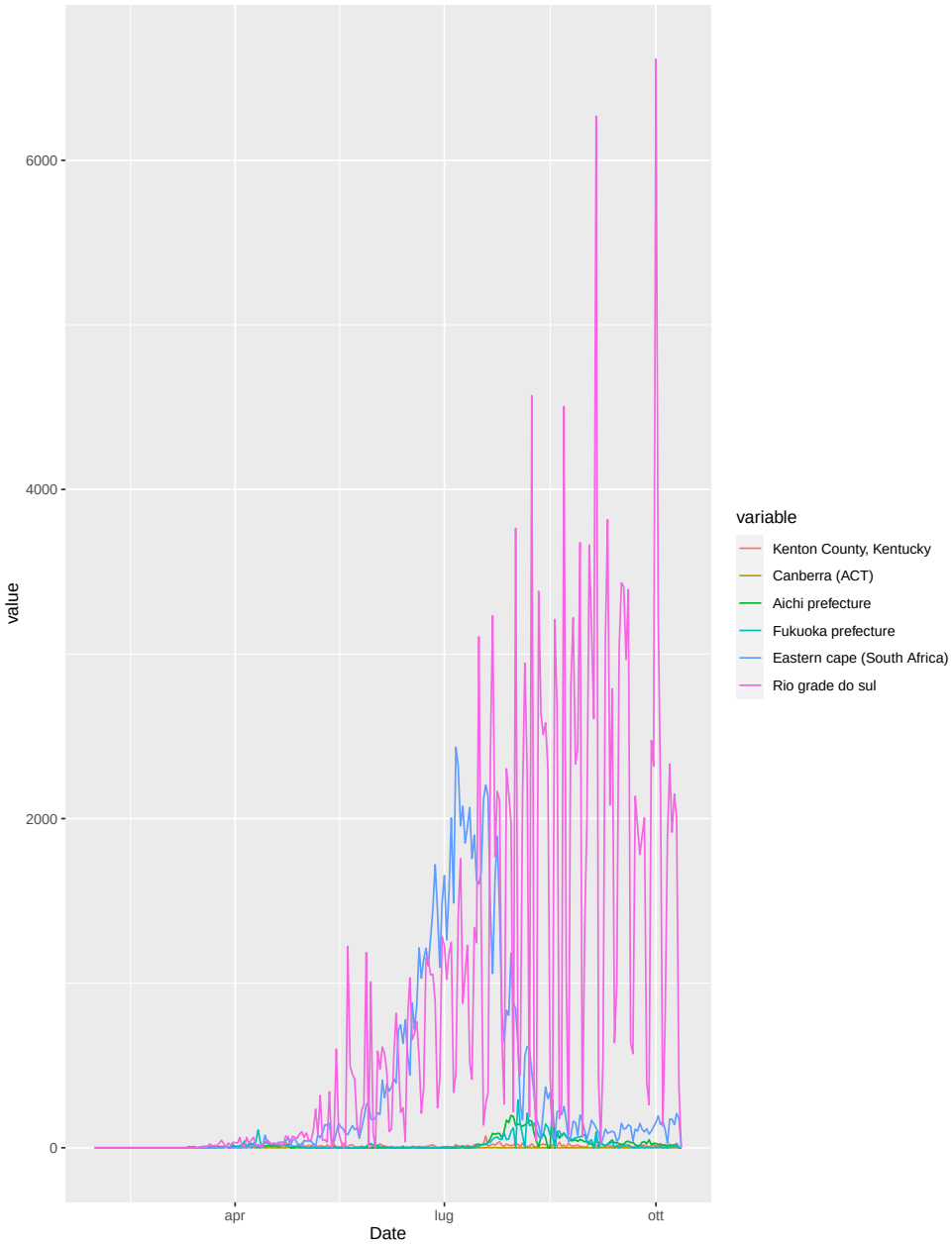
Based on how these three elements combine it is possible to distinguish between multiplicative and additive time series. In general having an additive model would be preferred but it is possible to convert a multiplicative time series into an additive by taking a log of the time series.

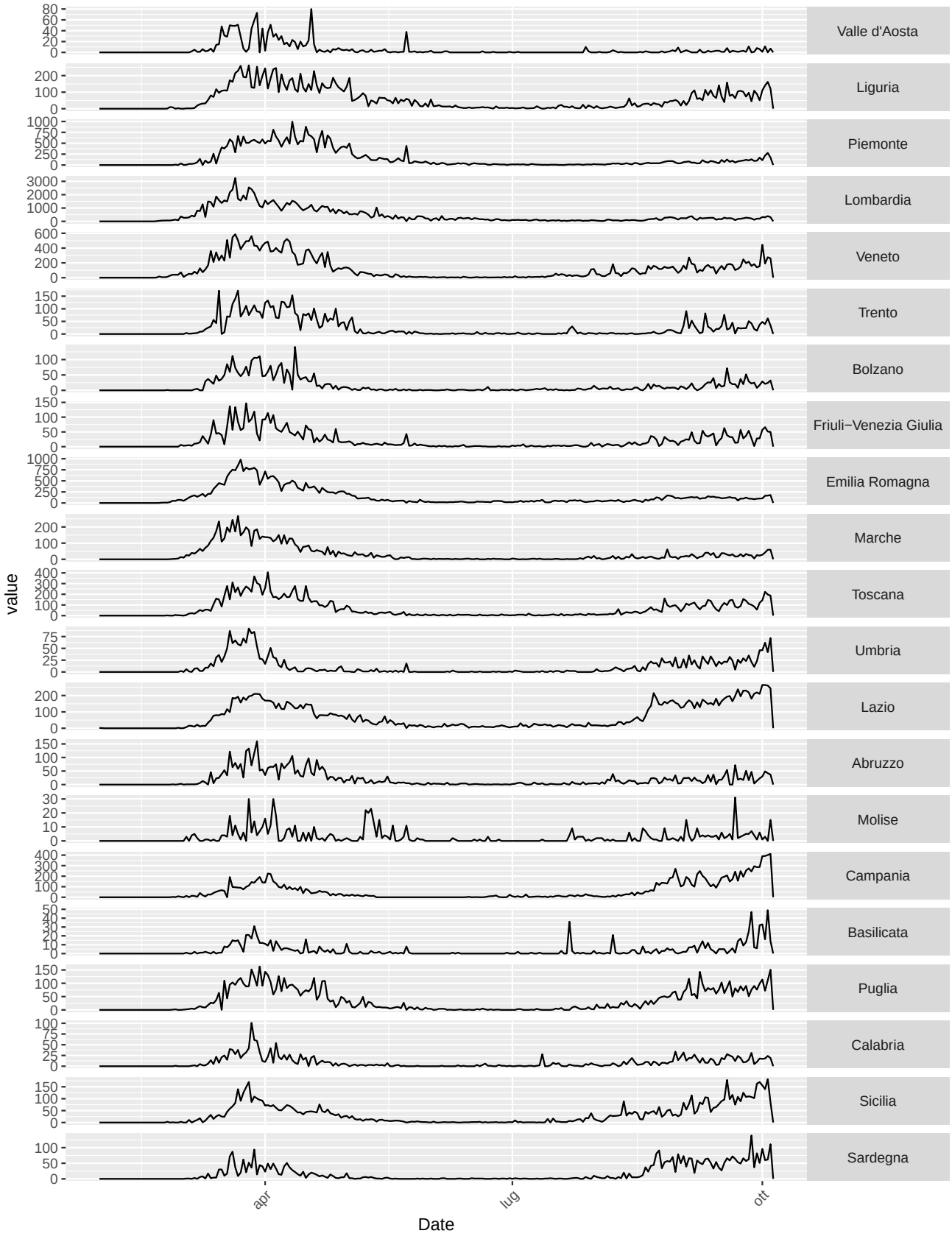
5.2.1 Stationarity

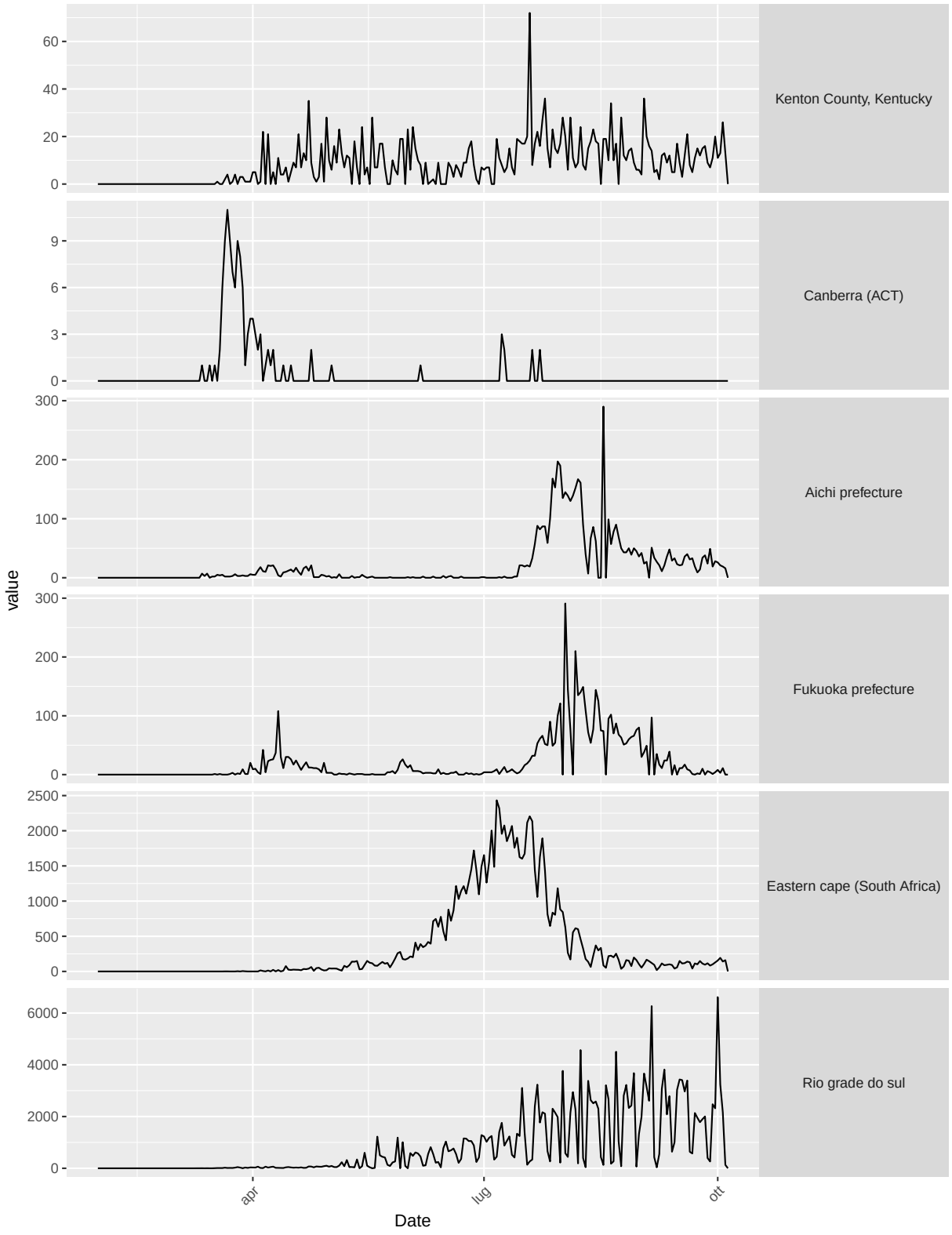
The first step to perform time series analysis is checking for stationarity, otherwise it is not possible to build a time series model. Stationarity means that the statistical properties of a process generating a time series do not change over time.[13]

The procedure to apply when performing a time series analysis comprises a few steps:

- checking for stationarity;
-


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


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


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

Chapter 6

Regression

The following linear model have been proposed to describe both the contagion and the fatality rate of the COVID-19 for each region:

$$\text{cr}_i = \alpha_0 + \alpha_1 bsh_i + \alpha_2 bsk_i + \alpha_3 csa_i + \alpha_4 csb_i + \alpha_5 cfa_i + \alpha_6 cfb_i + \alpha_7 dfb_i + \alpha_8 dfc_i + \alpha_9 et_i + \alpha_{10} CAQI_i + \alpha_{11} emp_i + \alpha_{12} area_i + \alpha_{13} density_i + \alpha_{14} hosp_i$$

$$\text{fr}_i = \alpha_0 + \alpha_1 bsh_i + \alpha_2 bsk_i + \alpha_3 csa_i + \alpha_4 csb_i + \alpha_5 cfa_i + \alpha_6 cfb_i + \alpha_7 dfb_i + \alpha_8 dfc_i + \alpha_9 et_i + \alpha_{10} CAQI_i + \alpha_{11} emp_i + \alpha_{12} area_i + \alpha_{13} density_i + \alpha_{14} hosp_i$$

Where:

- i is the identifier of each of the regions of the study;
- cr_i is the dependant variable that stands for the contagion rate;
- fr_i is the dependant variable that stands for the fatality rate;
- bsh_i is the dummy variable for BSh climate;
- bsk_i is the dummy variable for BSk climate;
- csa_i is the dummy variable for Csa climate;
- csb_i is the dummy variable for Csb climate;
- cfa_i is the dummy variable for Cfa climate;
- cfb_i is the dummy variable for Cfb climate;
- dfb_i is the dummy variable for Dfb climate;
- dfc_i is the dummy variable for Dfc climate;
- et_i is the dummy variable for ET climate;
- $CAQI_i$ is the level of pollution;
- emp_i is the employment rate;
- $area_i$ is the surface of the region;
- $density_i$ is the density fo the region;
- $hosp_i$ is the number of hospital beds.

Chapter 7

Conclusion

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