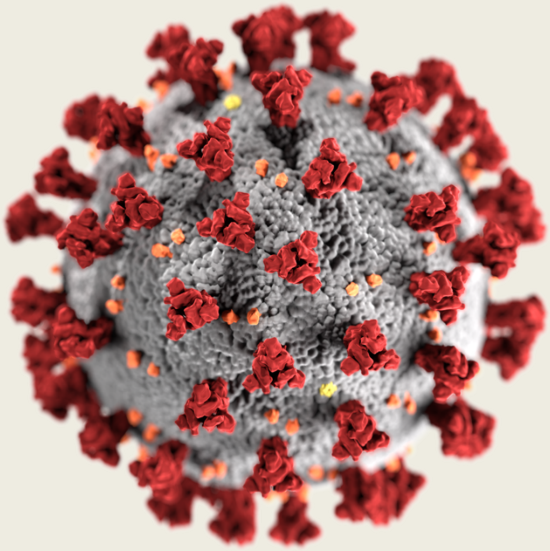
COVID-19 ANALYSIS

STUDYING THE VIRUS EXPANSION WITH REGRESSION AND ML

April 16th, 2020



Predictive Analytics Course
@ MIP – Politecnico di Milano

Group 4

Guida Giulia, Muttoni Davide, Ticozzi Paolo

Table of Contents

Coronavirus Analysis

1. Introduction

The Virus

The disease caused by the infamous “Coronavirus”, also known as COVID-19, is an infectious illness that is causing serious problems to humanity. Apparently born in China around November 2019, and quickly spread all over the world. Most of the countries in the previous months have declared emergency status with various levels of collective quarantine in order to reduce and prevent the spread of the virus. People affected by COVID-19 may experience symptoms such as cough, fever, tiredness, and in cruel respiratory illness that in worst cases may lead to death that luckily regards only a small percentage.

Our Goal

Our aim for this project is to apply Machine Learning algorithms in order to find the best way to predict the short-term advancement of the spreading of COVID-19 in the whole world. In particular, the authors of the project decided to focus on short-term forecasting, developing several analyses and implementing different models to extrapolate anything which can give to the reader a deeper vision of how Coronavirus tends to spread among and across countries.

It’s worth adding that our analysis is purely based on the provided numerical data (especially the binome date - number of cases), without considering the hundreds of variables that can influence the prediction in real life. For example, most of the countries have begun a process of quarantine, trying to stop the exponential growth of the cases that risks to put the healthcare system in overload. However, we did our best to come up with the most accurate schema for this kind of predictive analytics problem.

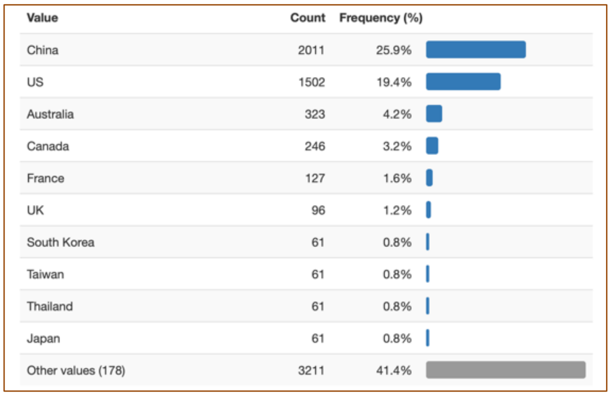
In short, our scope is to search for the tools that provide an approximate short-term prediction of the future COVID19 cases, deaths and recovery, based mostly on historical data. In order to do so, we established four key questions that will be answered in this report.

1. Can we apply log-transformation in our model?”
2. Is it possible to predict the trend of coronavirus cases for each country and region through autoregression components?
3. Is it possible to predict the trend of coronavirus cases for the whole world?
4. Is it possible to improve the analysis? How?

At first stages we decided to develop a simple linear regression on single countries (focusing on how coronavirus spreads in a single country), lately we implemented more sophisticated models aggregating all the countries to try to identify a tendency of Coronavirus in its spread across the globe.

The Dataset

The 'covid_19_data.csv' data contains 7 variables (3 numeric and 4 categorical), 7926 observations, and 3433 missing cells (mostly related to the "Province" subcolumn).



Value	Count	Frequency (%)
China	2011	25.9%
US	1502	19.4%
Australia	323	4.2%
Canada	246	3.2%
France	127	1.6%
UK	96	1.2%
South Korea	61	0.8%
Taiwan	61	0.8%
Thailand	61	0.8%
Japan	61	0.8%
Other values (178)	3211	41.4%

Figure 1. List of countries with most updates in the dataset

After a deep cleaning of the dataset, we identified 188 countries, where China and US (together with Australia, Canada, France and UK) appear more frequently than the others (Fig. 1), but only because they receive daily updates from all the different areas within the same country.

In fact, for these specific nations, it's possible to make deeper analysis to understand the trends for each region.

Down below, a brief explanation of the most important variables for us; some of them (*) have been added or updated in our data cleaning phase:

- Date*: it represents the exact date when the updates have been published;
- Country/Region: the country publishing the update; all the data in that row are related to its territory:
- Province/State: for some states (US, China and Australia) and some rare cases (French Guiana, for example), there was a second, more precise variable about the exact location inside the country
- Country_prov*: join between Country/Region and Province/State, created in order to simplify the regression procedure
- Confirmed: the amount of total confirmed cases in the country. It's a cumulative data, so it's a sum of the total cases.
- Deaths: sum of the total deaths in the country
- Recovered: total amount of people that recovered from COVID-19 in that zone
- Active*: the amount of people that are currently infected but not in emergency (result of the operation Confirmed - Recovered - Deaths)

We decided in the data ingestion phase to merge our data with world population indicators. We found a trusted dataset with several important KPI's such as the population density, total population in the country, the average age and the urbanization index for each country.

The idea was to train our models with these features too, to capture valuable correlations with the spreading growth. Unfortunately we didn't have the time to develop this stage of the analysis.

2. Data Cleaning & Preprocessing

Several adjustments to the original dataset have been made in order to meet our goal. Cleaning activities required the usage of Excel, OpenRefine tool and Python, to obtain a final base on which we could effectively work.

The date format

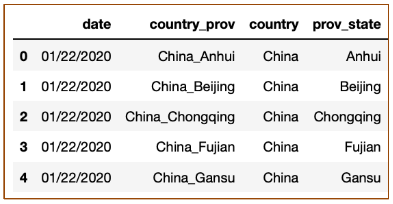
First of all, we transformed the date variable format (string) into datetime64 format. In this way, we had a standardized format that made the following steps easier for our calculations. The dataset contains updates from January 22nd to March 22nd of 2020, with 61 days of records.

Latitude & Longitude

Afterwards, we identified the latitude and longitude parameters for each state, and for each region available, in order to create a graphical representation of the coronavirus cases update, as will be shown later in the report. In order to do so, we used *geopy.geocoders* to abstract the service's API, and the geocoder *Nominatim* to extract the geolocation from the name of the countries.

Regions and Provinces

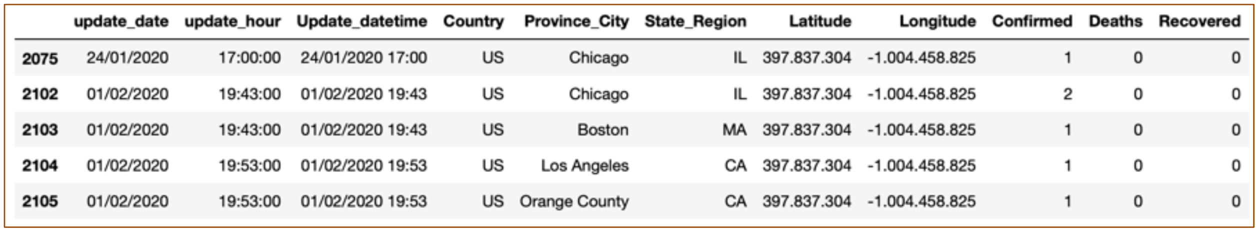
We distinguished for each country their own state/region, and their related province/city. For reasons of simplicity, we have joined together the two variables *Country/Region* and *Province/State* in a unique column called *country_prov* as shown in figure 2.



	date	country_prov	country	prov_state
0	01/22/2020	China_Anhui	China	Anhui
1	01/22/2020	China_Beijing	China	Beijing
2	01/22/2020	China_Chongqing	China	Chongqing
3	01/22/2020	China_Fujian	China	Fujian
4	01/22/2020	China_Gansu	China	Gansu

Figure 2. screenshot showing *country_prov*

Moreover, we did an additional cleaning with OpenRefine to separate country, state_region (states in the US and regions in the case of China), and Province_City (when available) to perform graphical representation of our data. By doing so, we obtained the dataset showed in figure 3.



	update_date	update_hour	Update_datetime	Country	Province_City	State_Region	Latitude	Longitude	Confirmed	Deaths	Recovered
2075	24/01/2020	17:00:00	24/01/2020 17:00	US	Chicago	IL	397.837.304	-1.004.458.825	1	0	0
2102	01/02/2020	19:43:00	01/02/2020 19:43	US	Chicago	IL	397.837.304	-1.004.458.825	2	0	0
2103	01/02/2020	19:43:00	01/02/2020 19:43	US	Boston	MA	397.837.304	-1.004.458.825	1	0	0
2104	01/02/2020	19:53:00	01/02/2020 19:53	US	Los Angeles	CA	397.837.304	-1.004.458.825	1	0	0
2105	01/02/2020	19:53:00	01/02/2020 19:53	US	Orange County	CA	397.837.304	-1.004.458.825	1	0	0

Figure 3. screenshot of *cleanedDF_by_regions.csv*

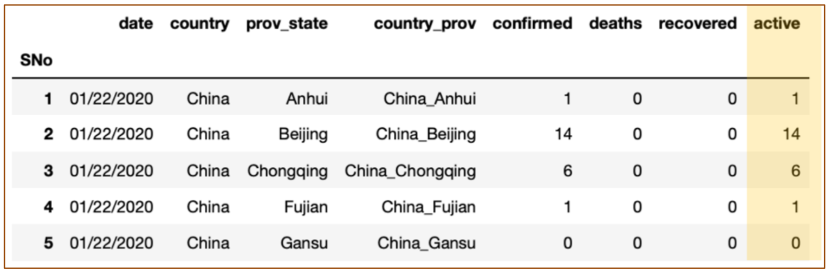
Missing/Wrong Values

There are null values just in the prov_state column, which means that not all countries have provinces or states within their territory.

Regarding 'wrong variables', we noticed some minor mistakes while analyzing the "Country" column. We manually edited to standardize the countries (for example we changed "Congo (Kinshasa)" into the correct name of the country, "DR Congo").

The "Active" Variable

Our target variables are the cumulative number of confirmed cases, deaths and recovered which are not fully capturing the spreading tendencies. Indeed, more important is the number of healthy carriers (we identified them as 'active') those people in quarantine who can still spread involuntarily the virus if not properly handled. The healthy carrier curve is probably the most important to be forecast properly, due to the high impact an active asymptomatic or not-in-emergency case can have on society.



	date	country	prov_state	country_prov	confirmed	deaths	recovered	active
SNo								
1	01/22/2020	China	Anhui	China_Anhui	1	0	0	1
2	01/22/2020	China	Beijing	China_Beijing	14	0	0	14
3	01/22/2020	China	Chongqing	China_Chongqing	6	0	0	6
4	01/22/2020	China	Fujian	China_Fujian	1	0	0	1
5	01/22/2020	China	Gansu	China_Gansu	0	0	0	0

Figure 4. screenshot of dataset showing new variable 'active'

Until the curve represented by healthy carrier shows an increasing trend, the lockdown is the suggested choice in most of the countries. Even in China which shows a bell-shaped active curve (see figure 5) the lockdown has not been a choice for quite long.

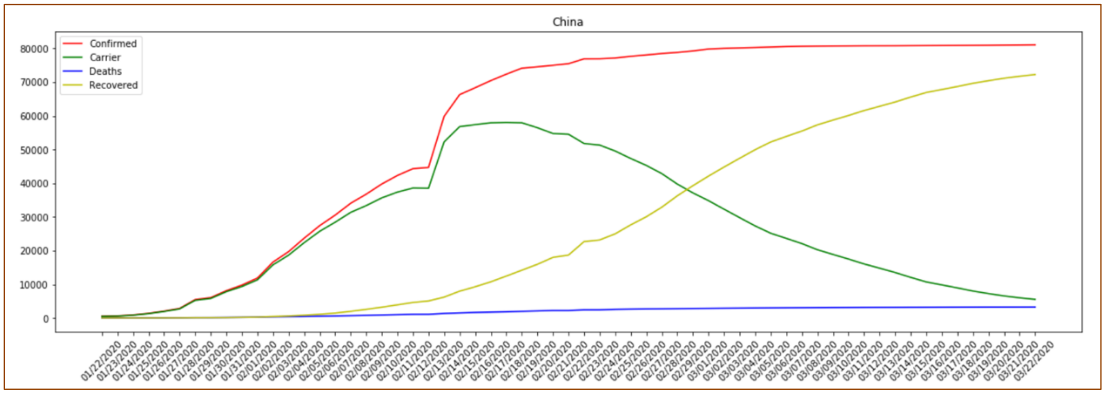
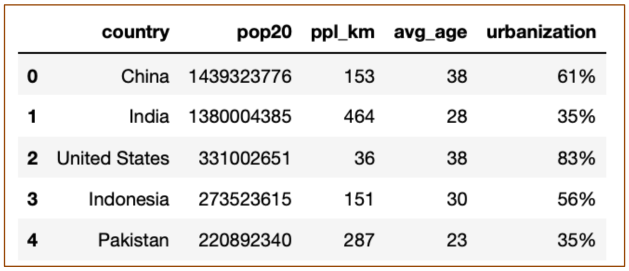


Figure 5. Plot of China's trends for Confirmed, Carrier, Deaths, and Recovered cases

Population Information

We added an additional dataset that contains data about world population. It is quite important to understand also how the population is structured in the country to better understand the spreadout of coronavirus. See figure 6 below for a screenshot of the mentioned dataset.



	country	pop20	ppl_km	avg_age	urbanization
0	China	1439323776	153	38	61%
1	India	1380004385	464	28	35%
2	United States	331002651	36	38	83%
3	Indonesia	273523615	151	30	56%
4	Pakistan	220892340	287	23	35%

Figure 6. Screenshot of *ppl.csv* dataset containing information regarding world population

Data Visualization with maps

Using the kepler.gl extension on jupyter, we were able to represent graphically the cumulative confirmed Coronavirus cases worldwide (figure 7) of the whole dataset. China is the country with the majority of confirmed cases (3,449,730), followed by Italy (440,823), Iran (231,132), South Korea (172,802), Spain (157,439), Germany (136,100), USA (120,075), and France (102,412). All other countries in the world have less than 40 thousands confirmed cases as of March 22nd, 2020.

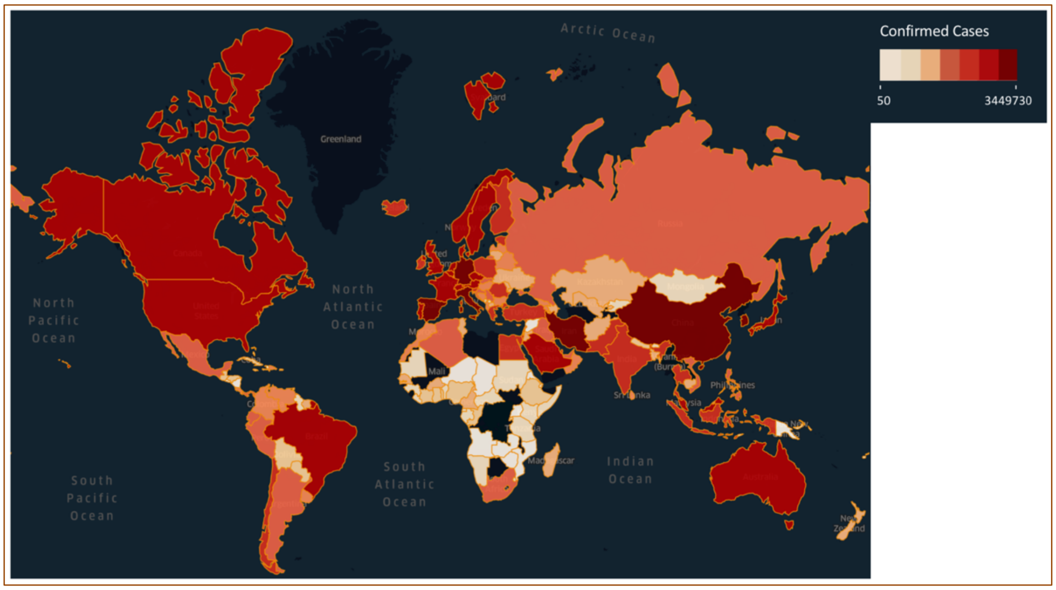


Figure 7. Coronavirus confirmed cases on March 22nd, with 61 days of records

Using Tableau, we were able to use the latitude and longitude that we found previously, to represent the development of COVID-19 within all Chinese regions (figure 8) and within all US states (figure 9) .

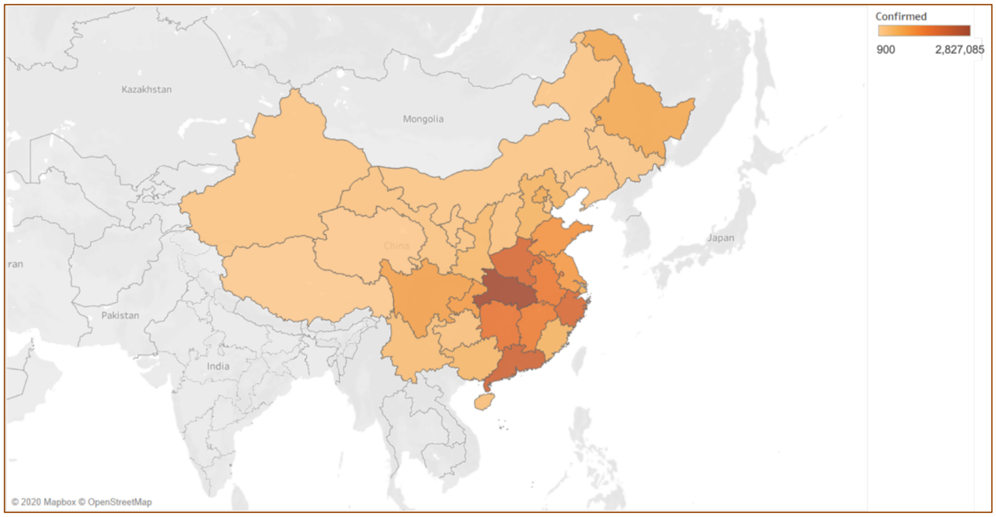


Figure 8. COVID-19 confirmed cases in Chinese regions

Hubei is the Chinese region with the majority of Coronavirus cases (2,827,085), not only in China, but all around the world. In fact, this is the region where the virus was found at first, and was locked down as soon as the seriousness of the situations was detected.

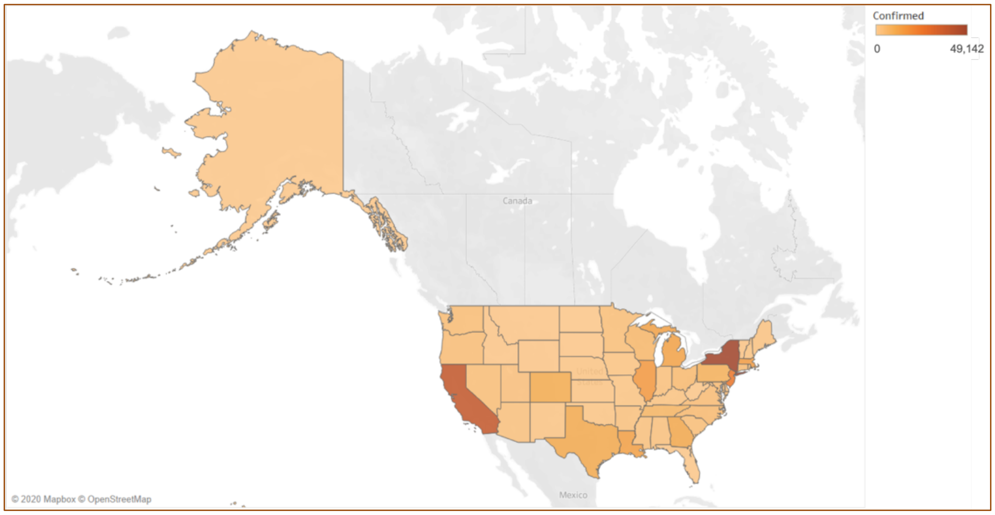


Figure 9. COVID-19 confirmed cases in US states

New York is the state having the majority of Coronavirus cases (49,142), followed by District of Columbia (12,721), and California (9,461). The numbers are not as big as they were in the Chinese regions as of March 22nd, in fact the US are the 7th country having confirmed cases.

Lastly, we quickly represented graphically on Tableau also the amount of recovered and death cases worldwide (figure 10).

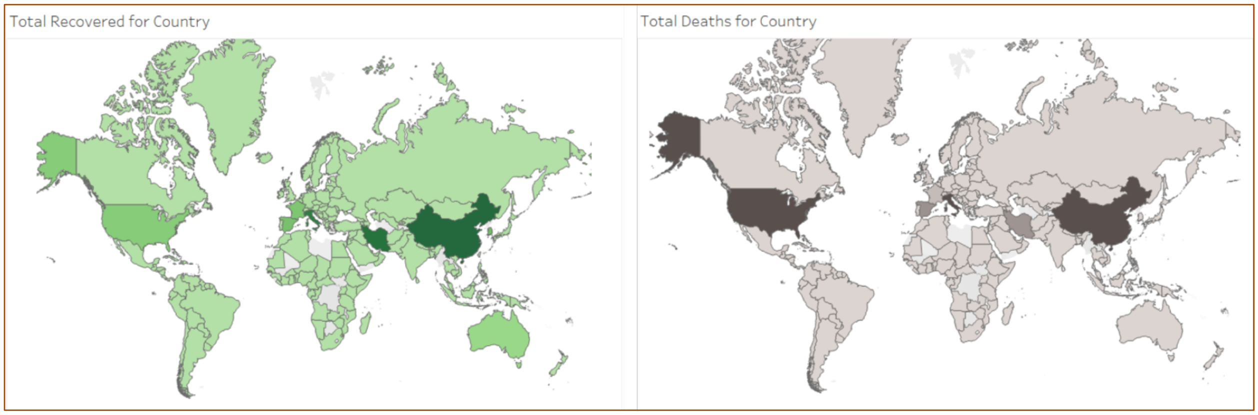


Figure 10. Recovered and death cases worldwide

It's clear from the two maps above that the countries having the majority of deaths, are also the countries with the biggest number of recoveries and, more in general, Coronavirus cases.

Encoding

This dataset presents as explanatory variables only categorical features, such as the date as the country. Preprocessing technique on categorical variables might require:

- Dummies Encoder, which converts each unique value of categorical feature in a column with Boolean values (1/0). This was not the case because we had too many different countries and too few observations; it would have required a further step with PCA, which might produce a less comprehensible result.
- Label Encoder, which assigns to each unique value of categorical feature in a column an integer (from 0 to n). This number must be treated as categorical, not as a continuous variable. We decided to use the second technique, which seemed to fit the best our data structure.

3. Models

In this section we will present the models performed and the relative questions of our goal. In this report we will show only the results regarding the confirmed cases, but the same algorithms can be ran for deaths and recovered variables.

3.1 Linear Regression

3.1.a Linear Regression with log-transformation

With the Linear Regression, we want to answer to the first question "Can we apply log-transformation in our model?".

The graphs performed during the EDA were showing that at first stage the virus tends to spread semi-exponentially. This induced us to think of log-transforming the target variables, especially the number of confirmed cases, for specific analysis at early stages. Why doing so? Because it would allow us to obtain more linear data and the linear regression analysis would be more precise. The graphs below (figure 11) shows the log-transformation of confirmed cases applied to Italy from February 25th, 2020. The lockdown for Italy started a week later but Milanese schools and major companies closed preventively that week.

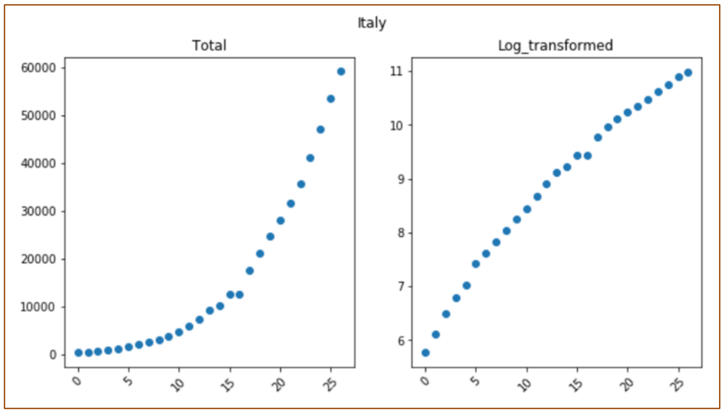


Figure 11. logarithm-transformation graphs for Italy

In this first stage of our analysis we tried to understand if log-transformation is relevant for our analysis. Accordingly, this first model detects the exponential growth in the so-called phase 1, which requires the government intervention and the lockdown for all the activities.

We developed a simple linear regression having:

- **Explanatory variable**: date;
- **Target** (one at a time): confirmed, deaths, active.

This model is quite unstable and tends to overestimate or underestimate the curve as the following graphs show (figure 12), for example Italy is overestimated, while Spain is underestimated. This result means that the exponentiality at first stage is true only for a very small interval.

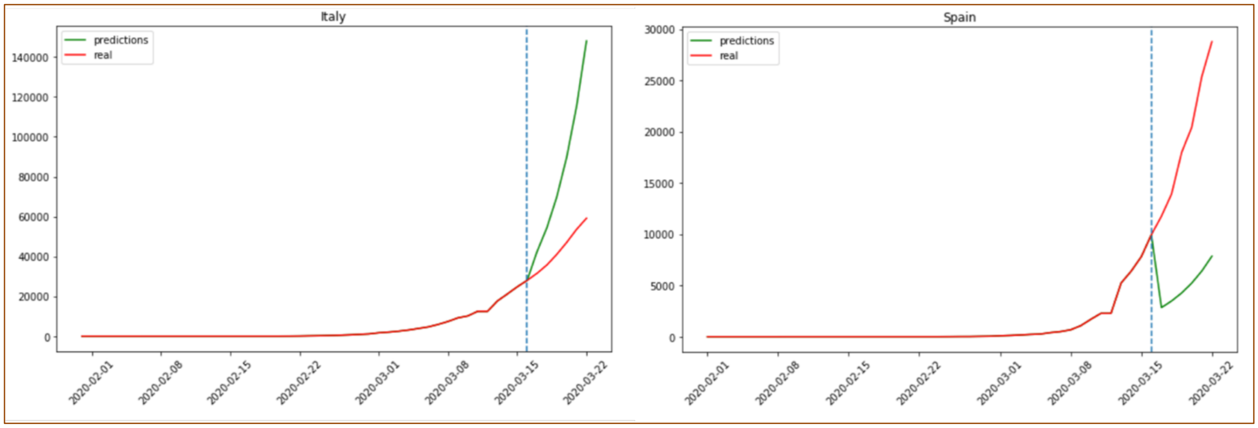


Figure 12. Linear Regression (with log-tranformation) for countries Italy and Spain - Prediction from March 16th

In conclusion, applying the log-transformation is not a good solution, because the exponentiality is not constant, as it takes place only at the beginning of the curve in each country (due to the high speed of spreading of the virus), and then the trend slows down (thanks to locking down countries).

3.1.b Linear Regression with lags

With the Linear Regression we also tried to answer to another question “Is it possible to predict the trend of coronavirus cases for each country and region through autoregression components?”.

With this linear regression we introduced temporal lags to capture the evolution of coronavirus within the countries and regions at any phase in the short-term, using autoregressive components.

This analysis in future steps might be enlarged with cross-analysis with world population data and to capture if the general structure of population (such as average age and density) are indeed strong factors for the diffusion of the virus.

The cumulative number of targets are highly correlated to the previous observations, which led us to try to implement autoregressive models.

In order to implement ML algorithms, we created 6 features containing temporal lags to capture the influence of the previous daily values up to a week before (lag 1 = t-1, lag 2=t-2… lag 6=t-6). In this way, the prediction was considering the previous days’ results, increasing the final accuracy. Figure 13 shows graphs that visually represent the autocorrelation of t in t-1.

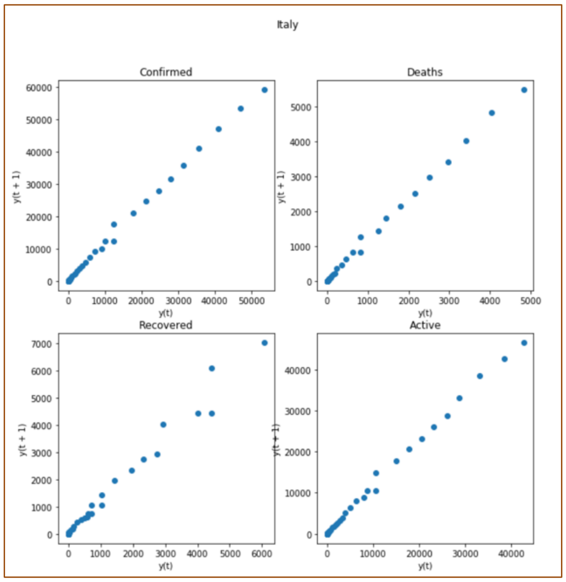


Figure 13. Lag plots for Italy

To perform the linear regression, we used:

- **Explanatory variable**:
  - date
  - target_variable_t-1, target_variable_t-2, target_variable_t-3
- **Target** (one at a time): confirmed, deaths, active.

We defined functions for splitting and training data, and with a Linear Regression algorithm we started to predict our values and comparing them with the real data of the dataset: our target period was about the last 7 days of the dataset, from 16th to 22nd March.
The coefficients of our regression, after being computed through the train set, need to be fitted in the testset, but actually lags are tricky. Indeed, at each single day predicted in the testset we had to recompute lags with the new predicted value.

This is key in the analysis to test the actual performances of our models. Using the temporal lags of real values (temporal lags of y_test) implicitly doesn’t predict anything, because of the strong autocorrelation demonstrated before. Introducing instead the temporal lags recomputed with the new predictions let us observe and clearly establish the goodness of our model. In this way, the final outcome was extremely precise. Figure 14 down below shows four examples that we analyzed about the confirmed cases of 3 important countries and the Hubei, the Chinese region.

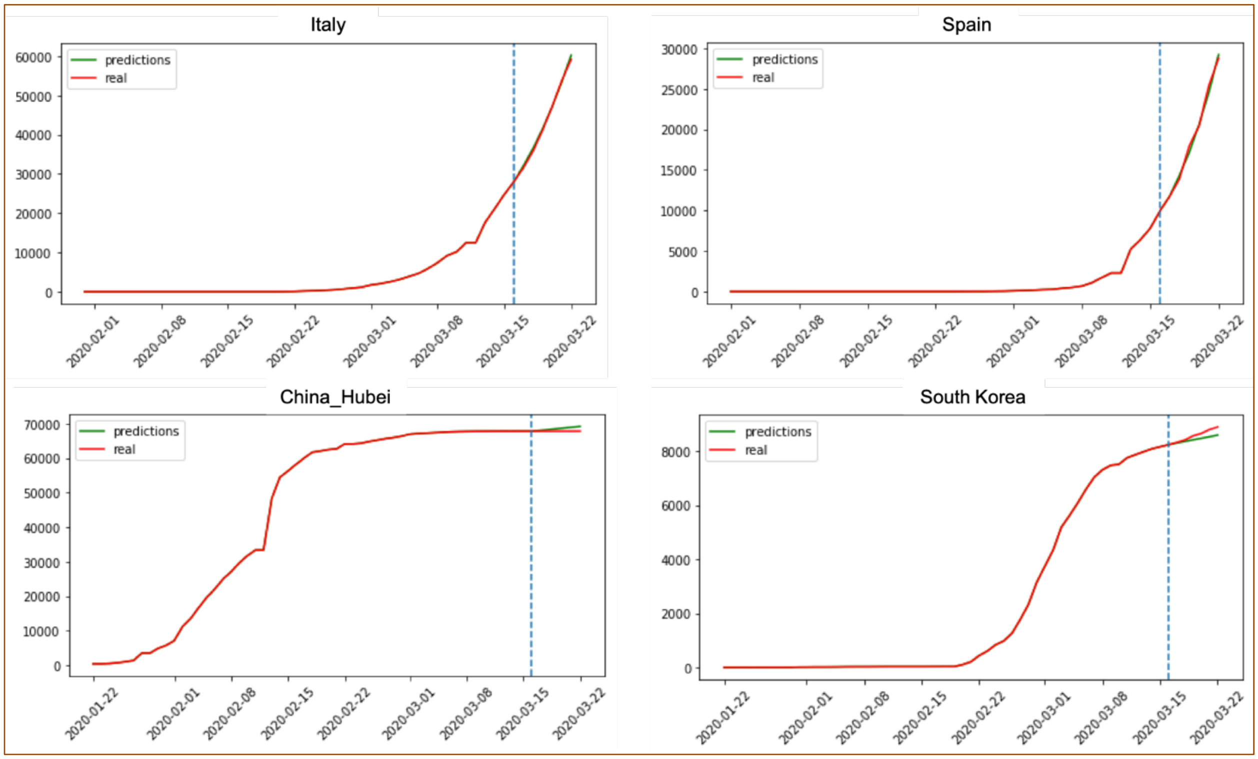


Figure 14. Linear Regression (with lags) for countries Italy, Spain, South Korea, and Chinese region Hubei - Prediction from March 16th

As shown in figure 14, Italy and Spain predictions are very precise. On the other hand, the regression for the two countries that were already outside phase 1, captures the flow only partially, but still sufficient to answer positevily to our question.

Linear Regression is a simple model, but considering the lags variable and updating it correctly, it was capable of predicting with a good amount of accuracy.

The aim of this model was to test the consistency of our hypothesis on temporal lags. The more precise the models would come out, the more we were convinced of using lags components. The good performances we obtained from single regions and single countries using regression with lags, led us to more generic models, which comprehend the whole globe at once, to capture general flows across countries, as will be described in paragraph 3.2.

3.2 Ridge & Lasso

Considering the third question of our goal "Is it possible to predict the trend of coronavirus cases for the whole world?", the last passage of our predictions is focused on building a model to generalize as much as possible the flow of coronavirus across all countries.

In this section we want to describe our models to predict the worldwide tendencies, using not only temporal lags but also countries as explanatory variables. We want to train our model to be as much general as possible, hoping that learning from all the countries altogether will allow to learn and predict each single country with a single, general function, instead of one model per each single country (which is too specific to understand the complex environment in which coronavirus spreads out).

In both cases, we had the same issue we faced with linear regression with lags. We had to recompute the test set after predicting each single day, to avoid to use always the real lags values, which would have make our analysis very poor.

To do so, we decided to develop the two multi-regression models: Ridge and Lasso. These two models have a penalty factor (L1 for Lasso and L2 for Ridge that will be explained later) that allows to better manage the multicollinearity.

3.2.a Ridge Regression

To perform the ridge regression, we used:

- **Explanatory variable**:
    - date
    - countries
    - target_variable_t-1, target_variable_t-2, target_variable_t-3
- **Target** (one at a time): confirmed, deaths, active.

Ridge model introduces a regularization term (L2 penalty) to the Cost function (highlighted in figure 15).

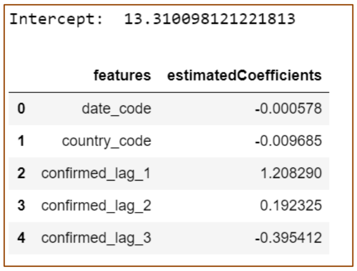
$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Figure 15. Cost function (L2 highlighted)

It is useful to reduce overfitting and multicollinearity, highly present in our data in between our temporal lags. Indeed a temporal lag in time t can be predicted through the temporal lags in time t-1,t-2,…,t-n.

In particular, the coefficient of predictor in t-1 and t-2 are quite impactful with respect to the others, as shown in figure 16. The models predict t through 3 levels of lags (confirmed cases in t-1, t-2, t-3); the objective is to evaluate the confirmed cases through the evolution of coronavirus spreading in the past week.

This model captures the trend quite well, especially for those countries where the coronavirus is in its phase 1 (semi-exponential growth) and low confirmed cases (Colombia, Malaysia,Taiwan), while for those countries with low observations and no phase 1 it tends to overestimate (Kenya, Macao, Nepal).



Intercept: 13.310098121221813

	features	estimatedCoefficients
0	date_code	-0.000578
1	country_code	-0.009685
2	confirmed_lag_1	1.208290
3	confirmed_lag_2	0.192325
4	confirmed_lag_3	-0.395412

Figure 16. Ridge Regression coefficients

On the other side, taking a look at the most hit countries such as the US, Spain and Italy (figure 17) it understands the actual increasing trend, but treats them as outliers and tends to underestimate the increase.

We performed a standard splitting, training and prediction phase using the Ridge Regression, obtaining rather precise results. Figure 17 below shows some results on 4 countries we selected, about the confirmed cases.

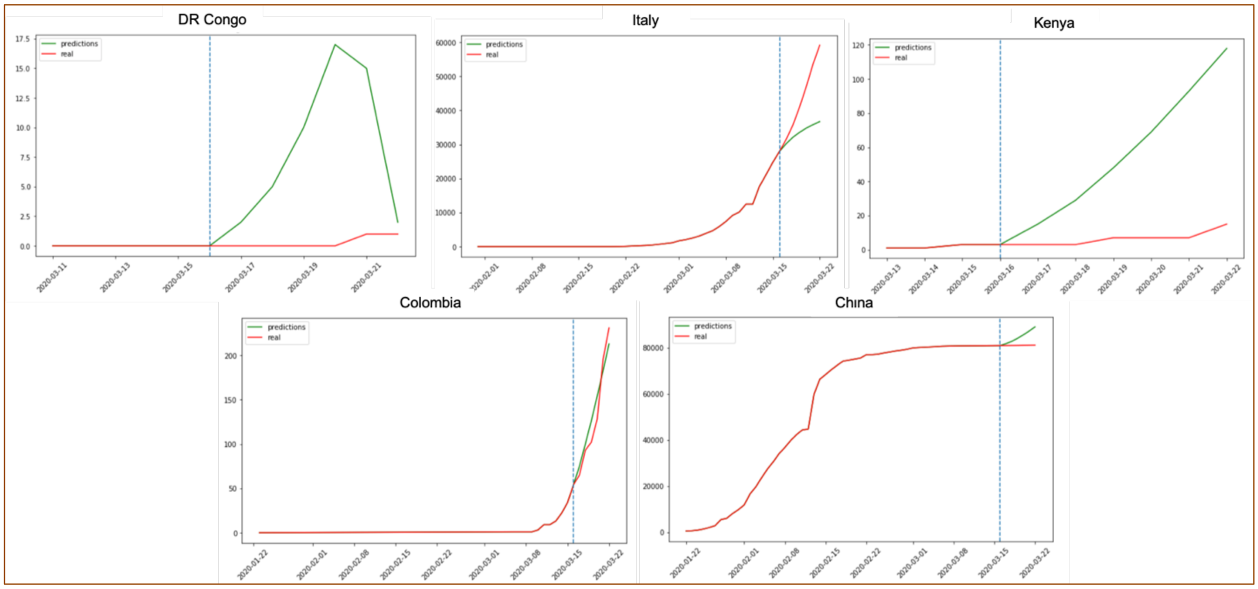


Figure 17. Ridge Regression for some countries - Prediction from March 16th

Interesting the case of DR Congo, which finds at the beginning an increasing trend (untrue) but through precedent knowledge the prediction drops dramatically to almost the real values. Italy, which is highly hit by the virus is treated as an outlier and is underestimated. Kenya which doesn’t have many observations is definitely overestimated. Colombia’s prediction is pretty close to actual data. Lastly China, is also overestimated.

In conclusion, we can say that the ridge regression model gave positive results as it found a common tendency, but in order to do so, it had to mediate within all countries, leading to having bad predictions for outliers (those having a strong increase or those close to zero records confirmed), and very good prediction for other countries.

3.2.b Lasso Regression

To perform the lasso regression, we used the same variables and target as for the ridge:

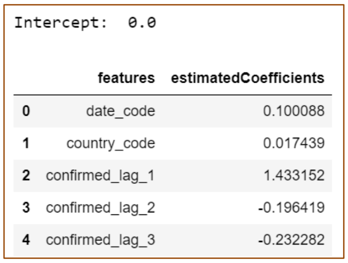
- Explanatory variable:
    - date
    - countries
    - target_variable_t-1, target_variable_t-2, target_variable_t-3
- Target (one at a time): confirmed, deaths, active.

Lasso Regression process and results were, approximately, similar to the Ridge ones (with slightly better performances on MAE indicator); thus we merged the two chapters together. Lasso multi-regression introduces an alternative penalty to Ridge (L1) which allows to reduce (or even eliminate) the weight those features that provoke multicollinearity.

Lasso model introduces a regularization term (L1 penalty) to the cost function (highlighted in figure 18). Figure 19 shows the lasso regression coefficients of our model.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Figure 18. Cost function (L1 highlighted)



```
Intercept:  0.0
```

	features	estimatedCoefficients
0	date_code	0.100088
1	country_code	0.017439
2	confirmed_lag_1	1.433152
3	confirmed_lag_2	-0.196419
4	confirmed_lag_3	-0.232282

Figure 19. Lasso Regression coefficients

It is still present the problem of underestimation of the countries with high infection and overestimation for those countries with very low numbers of confirmed cases.

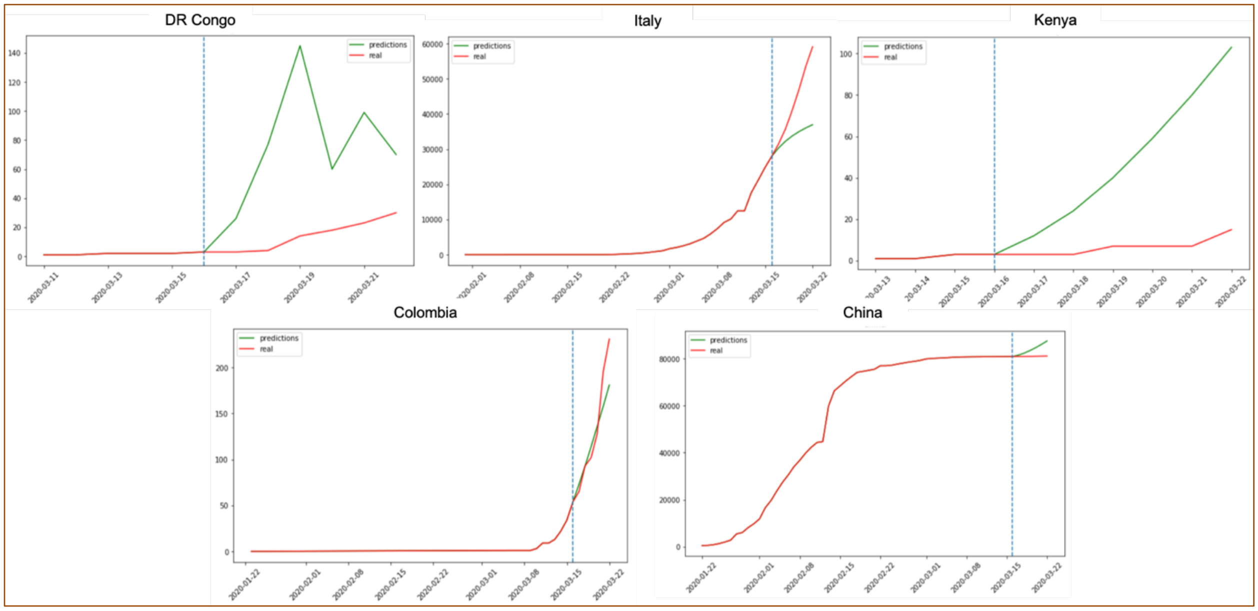


Figure 20. Lasso Regression for some countries - Prediction from March 16th

As mentioned before, we can see from figure 20 that predictions with Lasso Regression are very similar to those with Ridge Regression.

4. Conclusions

Providing a precise prediction is extremely hard, considering the small dimensions of the dataset and the unpredictable behavior of the disease we are analyzing. However, after a detailed process of data cleaning and data enrichment, we have managed to build a solid base where we could start to use our tools. We had cleaned wrong values, merged variables, formatted the data values and provided an initial analysis of the data.

Then, we have tried out various different strategies (for example, Polynomial Regression), but, in the end, we have mostly worked with Linear Regression, Ridge and Lasso. From this kind of data, we had a lot of difficulties with the training part of the project: the size was extremely small. However, we managed to define a rather precise model.

As already stated in chapter 1, you can notice, when observing the code, that we were working on a data enrichment process: we wanted to compare our results with some external data, like the average age of the nation, but we didn't have time, so that this process is left for an hypothetical future step. The model is exactly the same for predicting confirmed, deaths and active cases. In order to make predictions about them: it's possible to substitute the parameter in the prediction notebooks and run the prediction algorithms in a couple of clicks.

Finally, we can conclude the report by saying that this project was a perfect example of how the notions we learned can be applied in real life. In fact, we have worked on the most important topic in 2020, the COVID-19 epidemic!

5. References

All datasets and notebooks that we used to perform our analysis and elaborate this report are available at the following link: https://github.com/covidteam4/covid