Available online at www.HEFJournal.org

Journal of
Human, Earth, and Future

Vol. 2, No. 1, March, 2021

Analysis on the COVID-19 Impact on the Deaths Tendency in Italy and Europe

Sofia Montagna [1], Angelica Lo Duca [2*], Andrea Marchetti [2]

[1] *Department of Information Engineering, University of Pisa, Pisa, Italy*

[2] *Institute for Informatics and Telematics, National Research Council, Pisa, Italy*

Received 19 December 2020; Revised 14 February 2021; Accepted 25 February 2021; Published 01 March 2021

Abstract

Due to the arrival of COVID-19 in Italy and Europe, there has been a significant increase in deaths recorded in the year 2020. This increase is not justified by the number of deaths recorded for COVID-19. The hypothesis is that the deaths recorded for COVID-19 are underestimated. This study aims to estimate the possible number of unrecorded COVID-19 deaths using a predictive model built based on historical deaths recorded from 2015 to 2019. The estimate was calculated by comparing the number of deaths expected according to the prediction for the year 2020 under normal conditions with the deaths recorded during the pandemic in the same period, which runs from March to September 2020. Through the comparison, it was possible to obtain an estimate of the number of excess deaths, which represented how much the arrival of the coronavirus had affected the increase in deaths recorded. From the number of extra deaths, the number of COVID-19 deaths that were reported and recorded by official national sources was subtracted to get an idea of how many COVID-19 deaths might not have been recorded.

Keywords: Data Analysis; Time Series Analysis; Predicted Models; SARIMA Models; COVID-19 Deaths Unrecorded.

1. Introduction

Between the end of 2019 and the beginning of 2020 in the city of Wuhan in China, numerous cases of contagion related to the diffusion of a new virus called COVID-19 were recorded. It is a highly contagious virus belonging to the coronavirus family that mainly affects the respiratory tract but can cause symptoms that affect all organs and systems. In China, in the following months, the virus began to spread all over the world, so much so that the World Health Organization (WHO) was forced to declare a pandemic on March 11, 2020. Italy was one of the first nations to register numerous cases related to the pandemic. The first infections of people not coming from China were recorded in Northern Italy on February 21, 2020, while the first death related to COVID-19 was registered on February 22, 2020 in the hospital of Padua. The situation has progressively worsened. Italy suddenly found itself having to cope with a health emergency with numerous hospitalizations and deaths, so much so that the government, to counter the diffusion of the virus and limit the number of deaths, on March 9, 2020, has decided to take containment measures by declaring a national lockdown. Italy was the first European state to adopt such severe and restrictive measures. Despite the measures adopted, the number of deaths continued to grow, reaching a maximum peak on March 27, 2020, the date on which they recorded 969 deaths. The assumption on which the analysis is based is that the number of deaths recorded for COVID-19 has been underestimated, as many patients, precisely because of the sudden health

* Corresponding author: angelica.loduca@iit.cnr.it

http://dx.doi.org/10.28991/HEF-2021-02-01-01

emergency, could not receive the correct diagnosis that resulted in death. As proof of this, the increase in deaths recorded in the most affected areas was analyzed compared to the mean of the past years, which in the province of Bergamo has increased by about 365%. Therefore, the study proposes, using predictive models, to make a prediction of deaths from all causes for the period March–September 2020 based on the historical deaths recorded from 2015 to 2019. The result obtained from the prediction represents the number of deaths expected in normal conditions. Using some metrics (SDE and SDS), the value of expected deaths was compared with the number of deaths from all causes recorded in Italy since the beginning of the pandemic to determine the estimate of excess deaths presumably caused by the virus. From the estimate of excess deaths obtained (SDE), the number of deaths from COVID-19 released by official national sources was detracted to be able to estimate the possible number of COVID-19 unrecorded deaths (SDS). The analysis was carried out both at the Italian (national and regional) level and for some European countries, including the United Kingdom. The results obtained demonstrated a greater number of COVID-19 unrecorded deaths in the areas most affected by the virus. Specifically, at the Italian level, the northern regions such as Lombardy, with a mean estimate of 8,481 unrecorded deaths, up to a maximum of 18,550 deaths. Following Piedmont, Emilia Romagna, and some southern regions such as Sicily and Puglia, subject to a greater extent by the return of residents. In Europe, however, the United Kingdom, with a mean estimated death submerged of 29,005 for up to 111,486 value, was followed by Germany, Italy, Spain, and France.

2. State of the Art

During the first phase of the pandemic, several studies have been carried out and published around the world that detect and demonstrate the presence of an excess of mortality for the year 2020 that is higher than the deaths officially documented for COVID-19. It is likely that the excess value includes COVID-19 unrecorded deaths.

2.1. Excess Mortality in the United States

The University of Cambridge, in October 2020, published a study [1] that analyzed the excess mortality recorded in the United States during the COVID-19 pandemic. The study is based on weekly data on mortality recorded from all causes, pneumonia, and influenza in the states where a high concentration of COVID-19 deaths was reported, especially in the states of California, Connecticut, Florida, Illinois, Indiana, Louisiana, Massachusetts, Michigan, New Jersey, New York, Pennsylvania, and Washington. The recorded mortality weekly data from September 27, 2015 to May 9, 2020, was obtained from the September 11, 2020 release of the National Center for Health Statistics Mortality Surveillance System. The object of the study was therefore to determine the real impact that the pandemic had on the increase in deaths in the states listed during the year 2020 through the creation of models that could predict the excess mortality recorded for all causes, including influenza and pneumonia. In particular, the study made use of two models: a conventional model, which estimates the excess of deaths compared to previous years as the difference between observed and predicted deaths by the model under normal conditions, and a model semiparametric, which, on the contrary, estimates the excess deaths as the difference of two expected values: the expected mortality taking into account an indicator of the pandemic period, which represents the turning point in the increase in mortality; and the expected mortality without taking into account the pandemic period indicator. The semiparametric model returned tighter confidence intervals, so the study was concentrated on the results produced.

Table 1. Results for all causes, for influenza and for pneumonia of the 95% confidence intervals obtained from the semi parametric model compared with the COVID-19 deaths recorded

State	COVID-19 deaths	All causes excess mortality	Influenza excess mortality	Pneumonia excess mortality
California	2.849	(3.338, 6.344)	(-75, 52)	(1.729, 2.370)
Colorado	1.130	(1.175, 1.730)	(-26, 15)	(620, 803)
Connecticut	2.932	(3.095, 3.952)	(-20, 5)	(651, 844)
Florida	1.840	(1.271, 2.856)	(-39, 21)	(1.100, 1.439)
Illinois	3.525	(4.646, 6.111)	(-20, 44)	(1.974, 2.422)
Indiana	1.490	(1.400, 2.078)	(-28, 14)	(679, 882)
Louisiana	2.267	(2.341, 3.183)	(-3, 27)	(1.042, 1.263)
Massachusetts	5.050	(5.562, 7.201)	(-100, 39)	(2.044, 2.456)
Michigan	5.036	(5.581, 7.171)	(-52, 28)	(2.386, 2.926)
New Jersey	10.465	(13.170, 16.058)	(2, 46)	(5.550, 6.539)
New York	26.584	(32.538, 39.960)	(694, 911)	(12.016, 14.310)
Pennsylvania	3.793	(5.125, 6.560)	(-65, -8)	(1.757, 2.135)
Washington	925	(559, 1633)	(-3, 72)	(358, 623)
United States	73.834	(100.013, 127.501)	(15, 1.385)	(40.066, 47.391)

Through the analysis, specifically, of deaths from pneumonia and influenza, it emerged a significant increase compared to the expected number. This increase was recorded in the city of New York where despite the number of influenza cases recorded steadily decreasing throughout the month of March 2020, influenza deaths increased until April 2020.

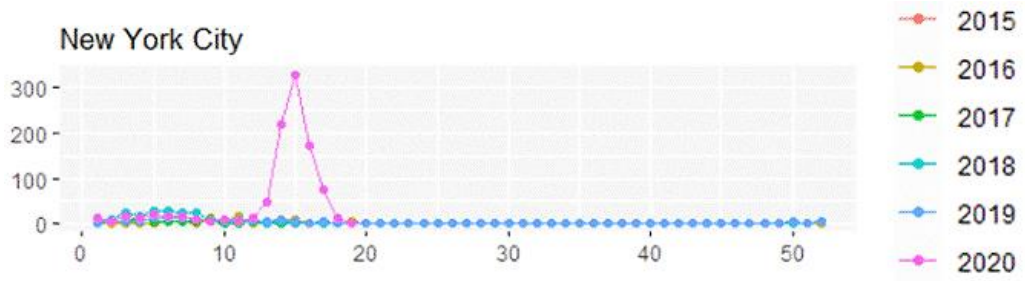

Figure 2. Weekly influenza mortality recorded in New York City broken down by years

The study succeeded in demonstrating the possibility that, in the absence of molecular swabs, many COVID-19 deaths have been misclassified as influenza or pneumonia deaths as the symptoms between diseases are nearly identical. Thus, a percentage of the estimated total excess mortality was captured by the excessive pneumonia and influenza mortality recorded in the same period. A further percentage was instead captured by deaths indirectly generated by the virus, or those deaths caused by the delay in health care due to the overloading of the structures or by the fear of the population to go to hospitals and contract the infection. In fact, the study found that many patients suffering from heart attacks or strokes have delayed the search for treatments and this has led to an increase in deaths. Ultimately, only increased availability of molecular testing, including post mortem, can lead to more accurate pandemic period mortality counts and reports.

2.2. Excess mortality in England and Wales

Based on the footprint of the previously discussed study, a survey was carried out on the excess deaths recorded in England and Wales during the early stages of the pandemic. The article [2] demonstrated whether, and to what extent, the number of unrecorded COVID-19 deaths has increased compared to what was expected in the absence of the virus. The study was based on provisional weekly data, released by the Office for National Statistics (ONS) updated to May 12, 2020. The data document mortality recorded in Wales and England broken down by gender, age, and region, for years from 2015 to 2020. To calculate the value of deaths not officially related to the pandemic, the study used the total number of deaths, regardless of cause, and the number of deaths that in the certificate was mentioned COVID-19. The estimate obtained, calculated as the difference of the two values, was compared with the average number of deaths recorded in the first eighteen weeks of the previous five years. The results, also in this case, show an excess of mortality that did not officially involve the virus, it would be about 968 additional weekly deaths compared to the weekly average of the years 2015-2019.

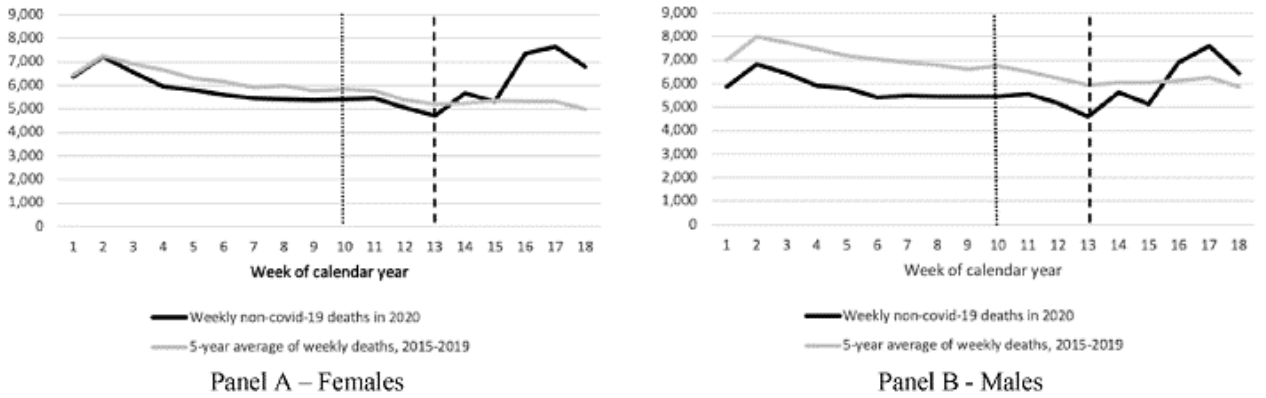

Figure 3. Weekly deaths, by gender, recorded in England and Wales as unrelated to COVID-19 compared to the 2015-2019 weekly average. The dashed vertical line indicates the first recorded death from COVID-19

The reasons for the excess mortality are the same: lack of molecular tests that lead to a lack of diagnosis of the virus and delay in providing or receiving help due to the overload of health facilities and the fear of citizens in going there.

3. Research Methodology

The objective of this study is to estimate the differences between the total death recorded in Italy during the pandemic period from March to September 2020 and the possible value of death expected for the same period, in the scenario in which pandemic does not have occurred. The prediction of the expected death was determined based on the analysis of the number of deaths from all causes recorded in Italy during the five-year period 2015-2019. The official

number of registered COVID-19 deaths was subtracted from the difference in deaths thus obtained, to obtain an estimate of the possible number of unrecorded COVID-19 deaths.

The metrics applied were as follows:

- *EED*: measures the estimated number of excess deaths due to the pandemic and is calculated as the differences between all-causes deaths recorded in the year 2020 and all-causes deaths expected according to the model predictive based on historical deaths recorded from 2015 to 2019.
- *UED*: measuring the estimation of unrecorded COVID-19 deaths according to the model predictive based on historical deaths recorded from 2015 to 2019.

The analysis was carried out in the first place at the Italian level (national and regional), and then moved to analyze the situation of some European nations as well, such as: Belgium, France, Germany, Greece, Portugal, United Kingdom, Romania, and Spain. These are countries for which, considering the period 2015-2020, grater completeness is available in the data collected.

4. Data Collection and Cleaning

4.1. Total Number of Deaths in Italy (ISTAT)

The data regarding the number of deaths registered in Italy for all causes have been collected from the official ISTAT website. ISTAT periodically provides the dataset containing the daily number of deaths divided by municipalities, provinces, and regions for years from 2015 to 2020. Thanks to the update of December 3, 2020, it was possible to acquire complete data to September 30, 2020 for all 7.903 Italian municipalities. From the downloaded dataset, the time series was created which reports the total deaths registered throughout the county from January 2015 to September 2020. Figure 1 shows the data.

Table 2. Information on collected data.

Title	Source	Format	Cadence	Level	Period
Total deaths recorded in Italy	ISTAT	csv	Daily	National and regional	January 2010 – September 2020
COVID-19 deaths national	Civil Protection	csv	Daily	National	March 2020 – September 2020
COVID-19 deaths regional	Civil Protection	csv	Daily	Regional	March 2020 – September 2020
Total deaths recorded in Europe	EUROSTAT	csv	Weekly	National	January 2010 – September 2020
COVID-19 deaths registered worldwide	Johns Hopkins University Center	csv	Daily	National	March 2020 – September 2020

By time series we mean a set of values ordered with respect to time that express the dynamics of a certain phenomenon: in this case the phenomenon of mortality in Italy. It was chosen to work on monthly data because they are more representative for the general trend of deaths. To do this, it was necessary to reprocess the dataset by combining the total value of deaths based on the month of registration. The time series of total deaths relating to each region of the peninsula were also obtained from the same dataset. To carry out this operation, it was sufficient to divide the initial ISTAT dataset from each region and subsequently combine again the total value of deaths based on the month of registration for each time series.

	REG	NOME_REGIONE	PROV	NOME_PROVINCIA	COD_PROVCOM	NOME_COMUNE	DATA	TOTALE
0	12	Lazio	58	Roma	58091	Roma	2015-10-20	76
1	12	Lazio	58	Roma	58091	Roma	2016-10-20	70
2	12	Lazio	58	Roma	58091	Roma	2017-10-20	74
3	12	Lazio	58	Roma	58091	Roma	2018-10-20	66
4	12	Lazio	58	Roma	58091	Roma	2019-10-20	61
5	12	Lazio	58	Roma	58091	Roma	2015-12-28	69
6	12	Lazio	58	Roma	58091	Roma	2016-12-28	103
7	12	Lazio	58	Roma	58091	Roma	2017-12-28	98
8	12	Lazio	58	Roma	58091	Roma	2018-12-28	105
9	12	Lazio	58	Roma	58091	Roma	2019-12-28	81

Figure 4. Extract from the ISTAT dataset

4.2. Number of Deaths from COVID-19 in Italy (Civil Protections)

The data regarding the number of deaths registered for COVID-19 were instead retrieved from the official profile GitHub of the Italian Civil Protection which is updated daily (Figure 5). To the analysis, it was necessary to obtain the deaths from March 2020 to September 2020 to be compared with the total death of the same period. The Civil Protection reports for each date the total value of deaths due to COVID-19, for which, to derive the increase compared to the previous day and therefore the actual number of deaths recorded daily, it was necessary to subtract from the current value of death, the previous value. To construct the time series, the values of daily deaths were then combined based on the month of registration. The Civil Protection also provides the dataset that reports COVID-19 deaths divided by region (Figure 6).

	Date	Deaths
0	2020-02-24	7
1	2020-02-25	10
2	2020-02-26	12
3	2020-02-27	17
4	2020-02-28	21
...	...	...
215	2020-09-26	35818
216	2020-09-27	35835
217	2020-09-28	35851
218	2020-09-29	35875
219	2020-09-30	35894

220 rows × 2 columns

Figure 5. Civil Protection's dataset of COVID-19 deaths recorded in Italy

	Date	Region's name	Deaths
126	2020-03-01	Abruzzo	0
127	2020-03-01	Basilicata	0
128	2020-03-01	Calabria	0
129	2020-03-01	Campania	0
130	2020-03-01	Emilia-Romagna	8
...	...	...	...
4615	2020-09-30	Sicilia	311
4616	2020-09-30	Toscana	1164
4617	2020-09-30	Umbria	85
4618	2020-09-30	Valle d'Aosta	146
4619	2020-09-30	Veneto	2178

4494 rows × 3 columns

Figure 6. Civil Protection's dataset of COVID-19 deaths recorded in Italian's regions

4.3. Total Number of Deaths in Europe (EUROSTAT)

The dataset containing the deaths of European nations was obtained from the EUROSTAT website which periodically provides the total number of deaths for all causes and for all ages recorded weekly by each nation in the union (Figure 7). From the latter, the total value of monthly deaths recorded from January 2015 to September 2020 was obtained with the aim of creating the monthly time series to analyze the following countries: Belgium, France, Germany, Greece, Portugal, United Kingdom, Romania, and Spain.

	TIME	GEO	Value
0	2010W01	Belgium	2280
1	2010W01	Bulgaria	2262
2	2010W01	Czechia	2155
3	2010W01	Denmark	1159
4	2010W01	Germany (until 1990 former territory of the FRG)	NaN
...	...	...	...
21011	2020W47	Albania	NaN
21012	2020W47	Serbia	NaN
21013	2020W47	Andorra	NaN
21014	2020W47	Armenia	NaN
21015	2020W47	Georgia	NaN

21016 rows × 3 columns

Figure 7. Extract from the EUROSTAT's dataset

4.4. Number of COVID-19 Deaths in Europe (Johns Hopkins University Center)

The data relating to COVID-19 deaths registered at national European level were collected from GitHub profile managed by the Johns Hopkins University Center which provides the number of deaths due the pandemic for nations around the world (Figure 8). From were obtained the total of COVID-19 deaths recorded in the month from March 2020 to September 2020 to build the time series of previously listed nations.

	Date	Geo	Province/State	Value
0	2020-01-22	Afghanistan	NaN	0
1	2020-01-23	Afghanistan	NaN	0
2	2020-01-24	Afghanistan	NaN	0
3	2020-01-25	Afghanistan	NaN	0
4	2020-01-26	Afghanistan	NaN	0
...	...	...	...	...
87728	2020-09-26	Zimbabwe	NaN	227
87729	2020-09-27	Zimbabwe	NaN	227
87730	2020-09-28	Zimbabwe	NaN	228
87731	2020-09-29	Zimbabwe	NaN	228
87732	2020-09-30	Zimbabwe	NaN	228

68563 rows × 4 columns

Figure 8. Johns Hopkins University Centre's dataset of COVID-19 deaths

5. Data Analysis

To analyze the data collected with the aim of obtaining an estimate of the unrecorded COVID-19 deaths, a software was created using the Jupyter application and the Python program language. The libraries used were: Statsmodels [3], Pandas [4], Numpy [5] and Matplotlib [6]. The collected data were interpreted as time series. For convenience, the analysis on the time series relating to the total death recorded in Italy, from January 2015 to September 2020, used to carry out the analysis at national level was taken as an example. The procedure explained below can be generalized to all the time series analysis.

The first operation performed was to study the time series, to create a predictive model for each of them, from which to obtain the prediction of the number of deaths expected for the year 2020 under normal conditions, based on time series data recorded from January 2015 to December 2019. Figure 9 shows a certain seasonality in the time series, which is repeated annually, and which determines the fluctuations in Italian's deaths. In fact, seasonality refers precisely to periodic fluctuations in observations.

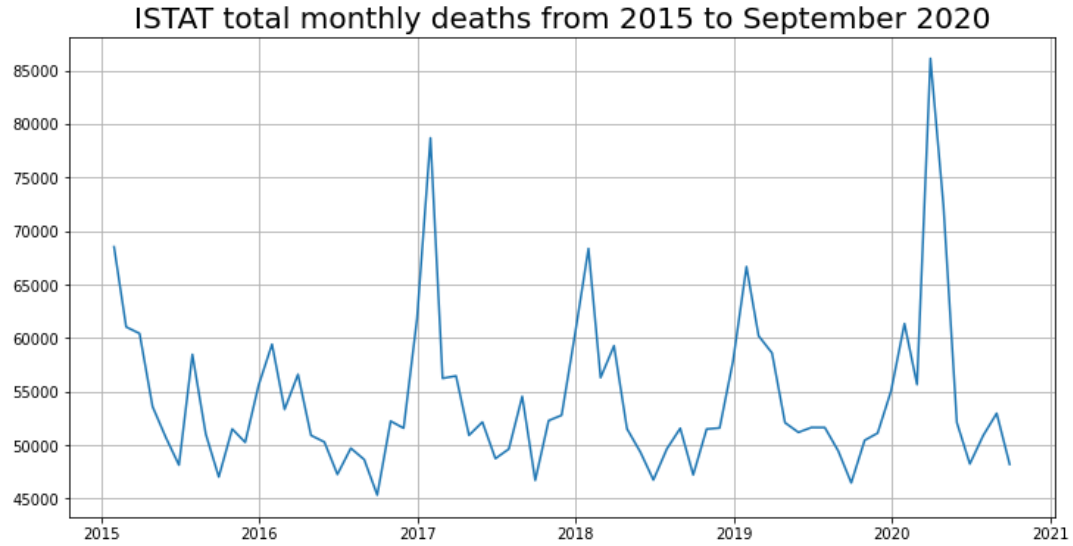

Figure 9. Monthly trend of total Italian's deaths

Every year there is an increase in deaths roughly coinciding with the month of November, probably due to the flu epidemic. The trend tends to decrease until May, to then increase again, in more slight way, in conjunction with the summer months and the increase in temperatures. Analyzing the year 2020 in more detail, it is possible to note an anomaly in the seasonal trend due to the increase in deaths caused by the arrival in Italy of the Coronavirus.

The study of the time series [7] was divided into two phases. The first phase consists in analyzing the series before a given breaking point, in this case it corresponds with the onset of the Coronavirus pandemic. The object of this phase is to detect the characteristics of the series, defining a predictive model for the representation of the time series before the breaking point. The model used in this phase is the model SARIMA (Seasonal Autoregressive Integrated Moving Average). The second phase consists instead in the use of the predictive model during the period following the breaking point, to establish a comparison with the real data. The use of the model allows us to calculate ac accurately ad possible what could be the future values of the series if the breaking point had not occurred.

5.1. First Phase: Analysis before the Pandemic

Before proceeding with the decomposition of the series, it was necessary to verify that the time series was stationary; this is because the modeling of a stationary series yields more reliable results in the medium to long term. By stationary we mean when the statistical properties of the time series do not vary over time, so it is possible to use its history to be able to predict its possible future behavior under normal conditions. To verify that the series in stationary, you can take advantage of statistical tests. The method considered to be the fastest and most effective is the use of the test Dickey-Fuller (ADF) based on unit root test. In the case of the time series analyzed, the test gave a negative result: the series was not stationary (Figure 10). To make it stationary, it must be transformed. In the specific case it has been differentiated, as better descripted below.

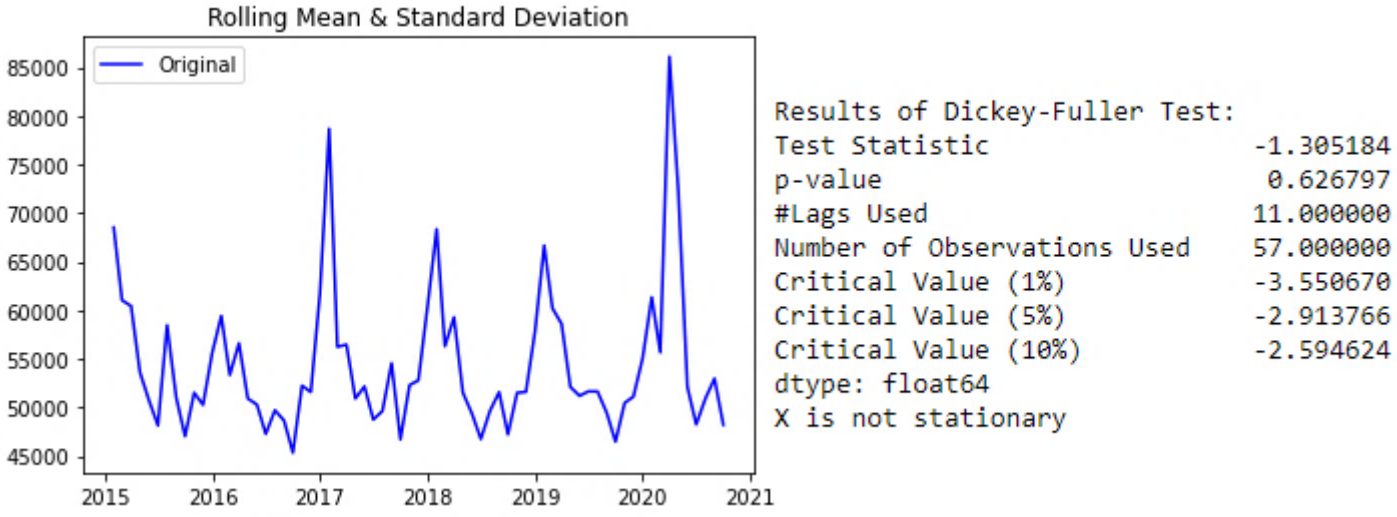

Figure 10. Result obtained from the Dickey-Fuller test applied to the time series

The model that was chosen to apply to the time series analyzed was the SARIMA model (Seasonal Autoregressive Integrated Moving Average). This model derives from the ARIMA model (Autoregressive Integrated Moving Average) which trivially tries to predict the future values of the time series based on its past values. The ARIMA model, however, does not perform well in time series in which seasonality is present. For this reason, it was chosen to use the SARIMA version which instead considers the seasonal variable present in the analyzed series.

The SARIMA (p, d, q) model receives three variables as input. The variable p indicates the order of autoregressive that is the degree of dependence between the current value and the previous values and corresponds to the number of observations included in the model. It can be obtained through the analysis of the graph of the Autocorrelation function (ACF) applied to the time series. The function measures the correlation between observation in a time series based on time. The possible value of the variable p corresponds in the graph by the light blue area (Figure 11). Based on what has been explained, the variable p has been assigned the value of 12, as it appears to be the outermost value of the area that represents the confidence interval.

The variable q indicates the moving average order, and its value can be obtained by analyzing the graph of the Partial Autocorrelation function (PACF) for the analyzed series. Also in this case, the possible value to be attributed to the variable q corresponds to the maximum value outside the confidence interval (Figure 12). The best result obtained by the model was that achieved by assigning to the variable q the value of 1.

Finally, the variable d which indicates the order of integration. It can be obtained from the number of differentiations that are necessary for the time series to make stationary. As explained above, the time series analyzed needs to be differentiated as it was not stationary, therefore was assigned the variable d the value of 1. In summary, to obtain the prediction of the number of deaths expected in Italy for the year 2020 it was chosen to apply the SARIMA (12,1,1) model.

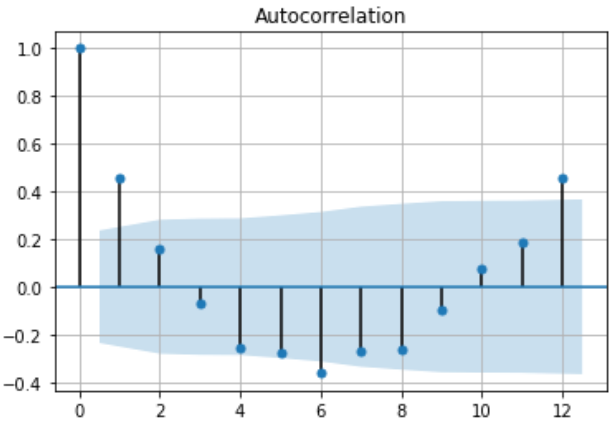

Figure 11. Autocorrelation function graph for the time series of total Italian's deaths

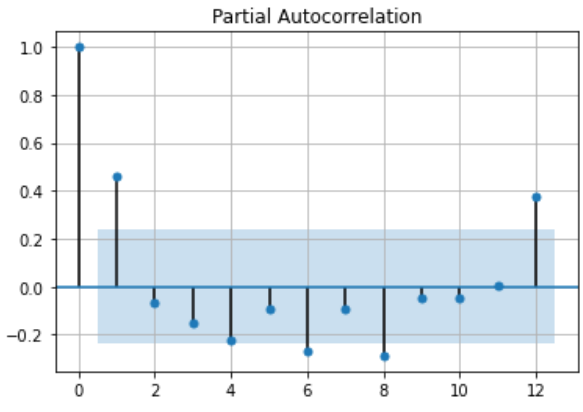

Figure 12. Partial Autocorrelation function graph for the time series of total Italian's deaths

Once the values to be assigned to the variables were obtained, the time series was divided into *train* and *test*. The *train* is formed by the values of the series ranging from January 2015 to October 2019 and corresponds to the part of the series that was used by the software to train the model. Figure 13 compares the observed values of the *train* with the predicted values obtained from the SARIMA (12,1,1) model.

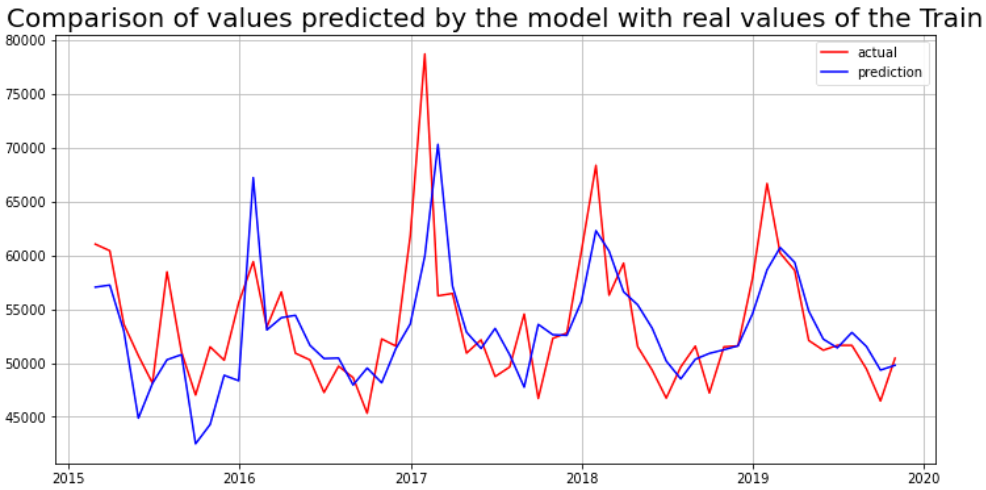

Figure 13. Comparison of the *train* with model prediction

From the graph it is visible how initially the model tends to make more mistakes and then be able to adapt to the current values with greater precision. Once the model was created, it was necessary to verify that the residuals, which measure the size of the errors emitted by the model, were stationary. The verification was performed again using the Dickey-Fuller test. Subsequently, the model created was tested by making a prediction on the *test* part into which the time series had previously been divided. The part of the *test* is made up of the values of the time series for the month of November and December 2019. By superimposing in a graph, the observed values of the *test* with the values obtained from the prediction for the same months, it was possible to give an initial evaluation to the model by displaying how far the predicted values are from the observed value.

The relation between the prediction obtained and the observed values was analyzed more specifically using the following metrics:

- MAPE: it measures the mean absolute percentage error.
- ME: it measures the mean error.
- MPE: it measures the mean percentage error.
- NMRSE: it measures the differences between the predicted values from the model and the observed values, according to which a value of 0 indicates a perfect model for the data.

This metrics were also used to compare different models with each other, with the aim of finding the most optimal one. For the historical series analyzed it was found to be the SARIMA (12,1,1) model. Table 3 shows the results obtained for that model.

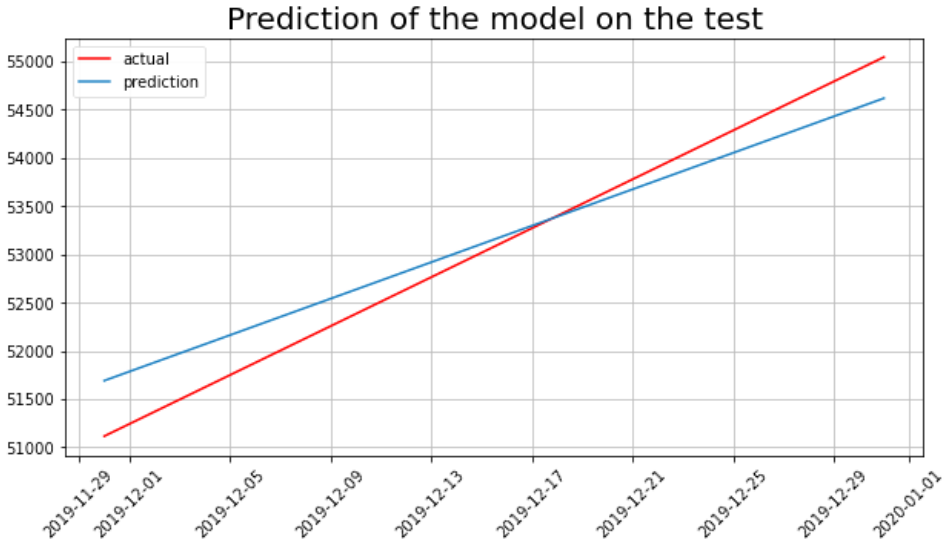

Figure 14. Comparison of the *test* with the prediction

Table 3. Accuracy metrics of the SARIMA (12,1,1) model

Nation	Model	Accuracy metrics			
		MAPE	ME	MPE	NRMSE
Italia	SARIMA(12,1,1)	0,001	74	0,001	0,13

5.2. Second Phase: Analysis from the Beginning of the Pandemic

Once the model was created and validated, it was possible to extend the prediction until September 2020, to derive the value of the expected deaths in Italy for the year 2020 based on historical mortality recorded from 2015 to 2019. In Figure 15, the prediction data were compared with the real data.

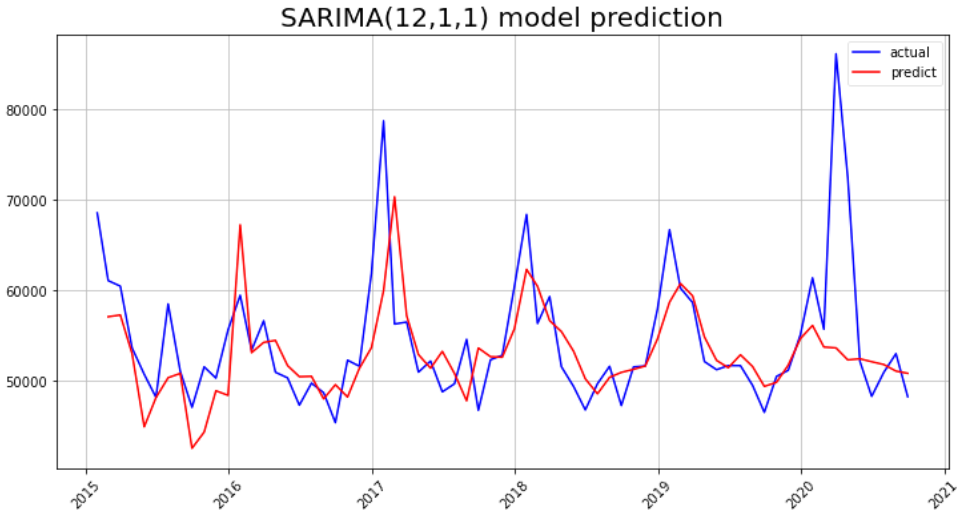

Figure 15. Prediction of the model respect the deaths observed from 2015 to September 2020 in Italy

To calculate the estimate of the possible number of unrecorded COVID-19 deaths, two metrics were applied. The first is the estimate of excess deaths (EED), which indicates the value of additional deaths in the period from March to September 2020. The value of EED was calculated by deducting from the deaths for all causes recorded by ISTAT from March to September 2020, the average estimate, and the minimum estimate of the number of deaths expected for the same period according to the prediction obtained from SARIMA model. The SARIMA model predicts according to an average value and a confidence interval for each time interval, so it is possible to calculate three EED values: minimum, average, and maximum. In some cases, the minimum value of EED obtained is negative, so only the average and maximum value of EED have been considered. Through the calculation of EED it was possible to obtain the average estimate and the maximum estimate of the number of excess deaths recorded in Italy from March to September 2020.

The second metric is the UED, which indicates the average number (or maximum) of unrecorded COVID-19 deaths. The UED value was calculated by deducting from the average (and maximum) EED estimate obtained, the number of COVID-19 official deaths registered in the months from March to September 2020, to obtain the estimate of the possible average (and maximum) value of COVID-19 deaths that differ from the officially registered in the same months.

6. Results

6.1. Italian Analysis

Figure 16 shows the prediction of deaths from all causes expected in Italy in normal condition compared whit the deaths recorded for all causes and the official deaths due to the Coronavirus pandemic. The gap between predictions and deaths from all causes is greatest in the first month in which the spread of the virus began (March and April 2020), and then diminishes from May onwards.

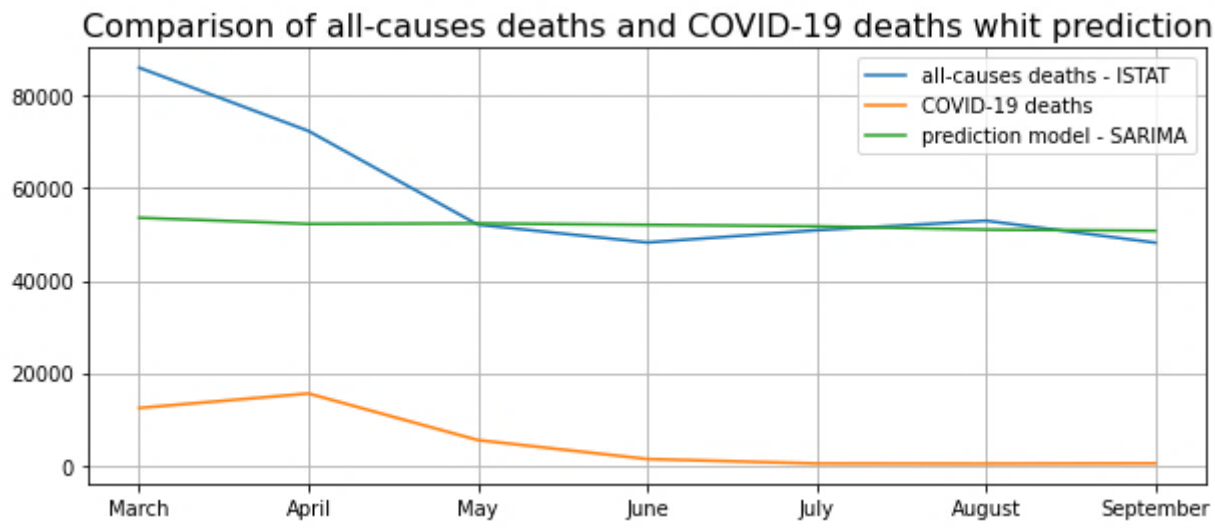

Figure 16. Comparison of expected deaths according to the prediction with deaths recorded in Italy for all causes and for COVID-19 from March to September 2020

From the differences between the curve of deaths from all causes and the curve of expected deaths according to the prediction of the SARIMA model, was obtained the curve of excess deaths recorded in Italy from March to September 2020. Figure 17 shows the comparison between the curve of estimated excess deaths (average and maximum value) and the official COVID-19 deaths' curve.

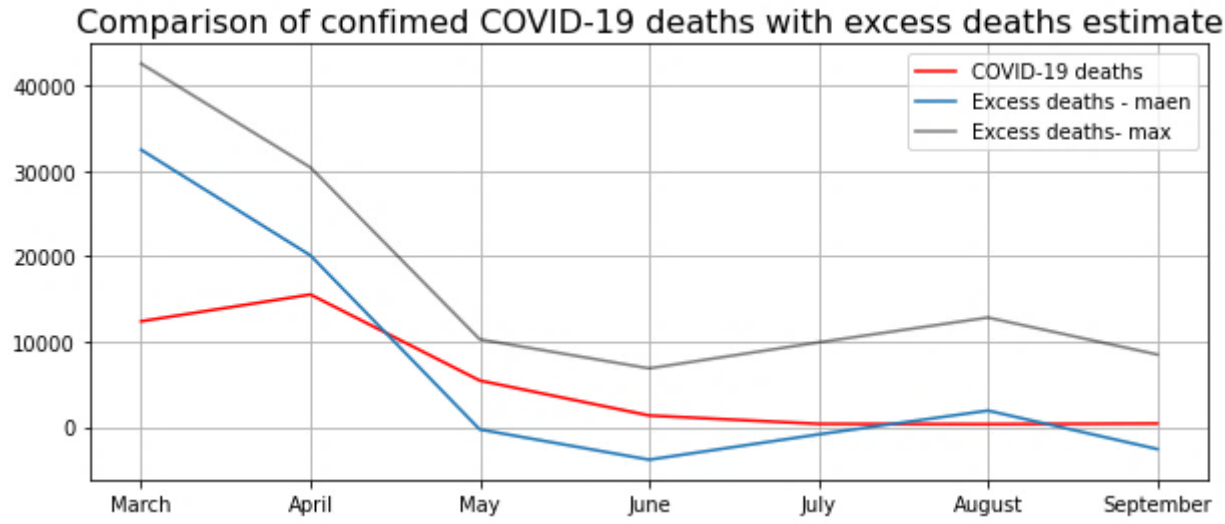

Figure 17. Estimate of excess deaths compared with officially registered COVID-19 deaths from March 2020 to September 2020

Again, a greater difference was found in March and April. This shows that probably in those months, due to the sudden arrival of the Coronavirus which caught the facilities and health personnel unprepared, many COVID-19 deaths were not recorded. In fact, in those months, as Figure 18 shows, was recorded a greater concentration of unrecorded COVID-19 deaths. Over the months, following the implementation of the government plan to combat the

pandemic and the arrival of summer temperatures that limited the spread of the virus, COVID-19 deaths have decreased since May a decrease also in the evolution of estimates of the unrecorded COVID-19 deaths.

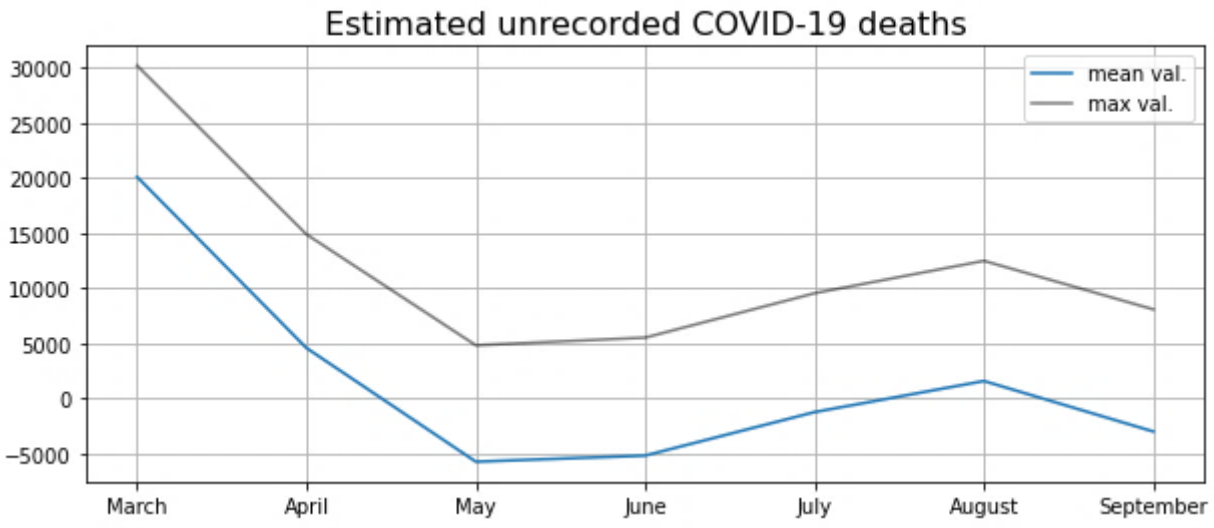

Figure 18. Trend of estimated unrecorded COVID-19 deaths in Italy

Ultimately, the analysis revealed that in Italy the mean value of deaths expected for the months from March to September 2020, under normal condition and based on the predictive model applied id 363.891 deaths up to a minimum of 289.426 deaths. In the same period, 410.899 deaths were observed and officially registered by ISTAT. Through the EED metric, the mean and maximum estimated value of the excess deaths for the period March-September 2020 was calculated, which resulted in a mean value of 47.008 deaths, up to a maximum value of 121.473 deaths. This means that compared to what was expected under normal conditions, there were on mean 47.008 excess deaths due to the pandemic, up to a maximum of 121.473 deaths. Officially in Italy, have been attributed 35.865 COVI-19 deaths in the months from March to September. Therefore, through the UED metric, the unrecorded COVID-19 death amounted to a mean value of 11.143 up to a maximum of 85.608 deaths.

Table 4. Results obtained from the analysis of Italian deaths (reference period March-September 2020)

Nation	Total deaths recorded	Total deaths expected from the model		EED: excess estimated deaths		COVID-19 deaths	UED: unrecorded COVID-19 estimated deaths	
		mean	min.	mean	max		mean	max
Italy	410.899	363.891	289.426	47.008	121.473	35.865	11.143	85.608

6.2. Analysis of the Italian's Regions

In addition to the national level, the method discussed was also applied to the regional time series. Table 5 details the regional models applied, and the related accuracy metrics calculated.

The first column shows for each region the number of total deaths recorded as reported by ISTAT in the months from March to September 2020, subsequently the mean and minimum value of the expected deaths in the regions for the same period in normal conditions according to the predictive model used. From the difference of the first two columns, the EED metric was determined which measures the estimate of the number of excess deaths (mean and maximum value) referring to the period March-September 2020. Subtracting from the latter, the value reported in the column of COVID-19 confirmed deaths in each region for the same period, the UED metric (mean and maximum value), was calculated which measures, for each region, the estimated number of unrecorded COVID-19 deaths from March 2020 to September 2020. A negative value of the metrics, as recorded in some regions such as: Campania, Lazio, Umbria, and Basilicata, means that fewer deaths were recorded in the months analyzed than those predicted. In fact, there are regions of central and southern Italy that in the phase of pandemic analysis were less affected by the COVID-19 virus.

In first place, the region in which a greater number of unrecorded COVID-19 deaths have been estimated, is Lombardy, with a mean estimate of 8.481 unrecorded deaths, up to a maximum of 18.550 deaths. In fact, the region that has been drastically affected by the spread of the pandemic since the beginning has recorded the highest number of COVID-19 deaths. Following are neighboring regions Piedmont and Emilia Romagna and some regions of Southern Italy such as Puglia and Sicily, regions that in March 2020 were most affected by the return of workers and students, giving way to a real "exodus" from Northern Italy. On the other hand, the regions with a lower population density such as Valle d'Aosta and Molise are in the last places, which consequently recorded lower numbers of deaths.

Table 5. Regional models and accuracy metrics

Regions	Models	Accuracy metrics			
		MAPE	ME	MPE	NRMSE
Abruzzo	SARIMA(12,1,1)	0,02	18,82	0,02	0,21
Basilicata	SARIMA(12,1,6)	0,02	7,07	0,01	0,17
Calabria	SARIMA(12,1,1)	0,06	13,14	0,01	1,83
Campania	SARIMA(12,1,8)	0,05	226,43	0,05	0,47
Emilia Romagna	SARIMA(12,1,3)	0,01	-4,65	0	0,17
Friuli Venezia Giulia	SARIMA(12,0,2)	0,02	-0,15	0	0,14
Lazio	SARIMA(12,1,8)	0,04	179,8	0,04	0,41
Liguria	SARIMA(12,0,5)	0	0,32	0	0,02
Lombardy	SARIMA(12,0,4)	0,01	99,59	0,01	0,17
Marche	SARIMA(12,1,1)	0,04	65,65	0,04	0,65
Molise	SARIMA(12,1,3)	0,03	8,26	0,03	inf
Piedmont	SARIMA(12,0,6)	0,02	69,18	0,02	0,53
Puglia	SARIMA(12,1,3)	0,01	7,8	0	0,09
Sardinia	SARIMA(6.1,8)	0	-1,87	0	0,03
Sicily	SARIMA(6,1,5)	0,02	85,85	0,02	0,31
Tuscany	SARIMA(6,1,2)	0,01	-16,3	0	0,24
Trentino Alto Adige	SARIMA(12,0,3)	0,03	26,39	0,03	0,41
Umbria	SARIMA(12,1,5)	0,02	19,74	0,02	0,33
Valle d'Aosta	SARIMA(6,1,1)	0,02	2,93	0,02	0,55
Veneto	SARIMA(6,1,4)	0,02	-9,16	0	0,3

Through the prediction obtained from the models it was possible to carry out a complete analysis for each region by calculating the EED and the UED metrics for each of them.

Table 6. Results obtained from the analysis of the deaths recorded in each Italian's regions (reference period March-September 2020)

Regions	Total deaths reorded	Total deaths expected from the model		EED: excess deaths estimated		COVID-19 deaths	UED: unrecorded COVID-19 estimated deaths	
		mean	min.	mean	max		mean	max
Lombardy	82.533	57.097	47.028	25.436	35.505	16.955	8.481	**18.550**
Piedmont	35.694	31.258	22.979	4.436	12.715	4.164	272	**8.551**
Sicily	30.216	31.065	21.858	-849	8.358	311	-1.160	**8.047**
Emilia Romagna	34.184	29.021	24.436	5.163	9.748	4.484	679	**5.264**
Puglia	24.052	23.142	18.629	910	5.423	595	315	**4.828**
Tuscany	26.053	25.777	20.279	276	5.774	1.164	-888	**4.610**
Veneto	30.322	28.399	24.049	1.923	6.273	2.178	-255	**4.095**
Liguria	14.625	12.226	9.703	2.399	4.922	1.604	795	**3.318**
Campania	30.414	31.593	26.820	-1.179	3.594	463	-1.642	**3.131**
Marche	11.461	10.169	8.106	1.292	3.355	990	302	**2.365**
Calabria	11.794	11.640	9.373	154	2.421	100	54	**2.321**
Sardinia	10.000	95.41	7.641	459	2.359	154	305	**2.205**
Lazio	33.119	34.076	30.084	-957	3.035	918	-1.875	**2.117**
Friuli Venezia Giulia	8.394	8.181	6.512	213	1.881	351	-138	**1.531**
Abruzzo	8.734	8.472	6.771	262	1.963	481	-219	**1.482**
Trentino Alto Adige	6.527	5.456	4.475	1.071	2.052	698	373	**1.354**
Umbria	5.947	6.192	4.882	-245	1.065	85	-330	**980**
Molise	2.229	2.159	1.549	70	680	24	46	**656**
Basilicata	3.622	3.786	3.110	-164	512	29	-193	**483**
Valle d'Aosta	979	861	626	118	353	146	-28	**207**

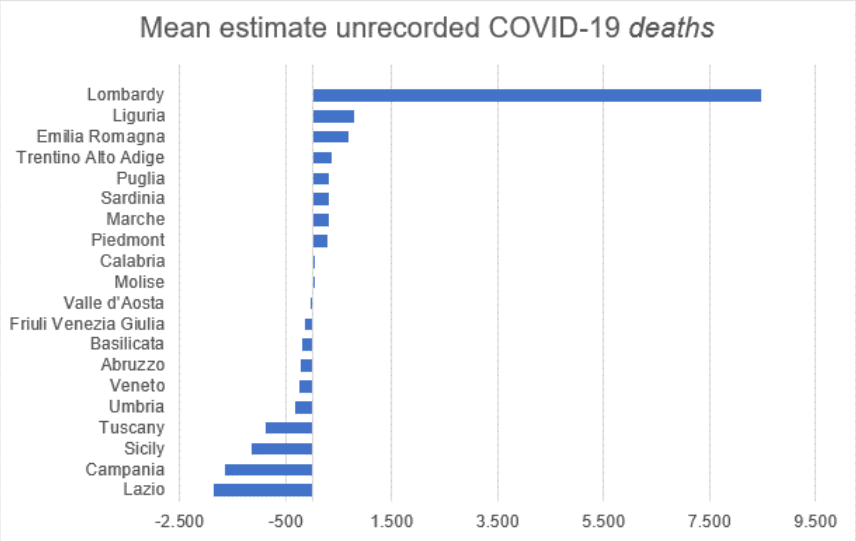

Figure 19. Histogram of the calculated mean estimates of unrecorded COVID-19 deaths in the Italian's regions. Negative values are present in cases where the values of officially registered COVID-19 deaths exceeds the mean value of excess deaths estimated for the region.

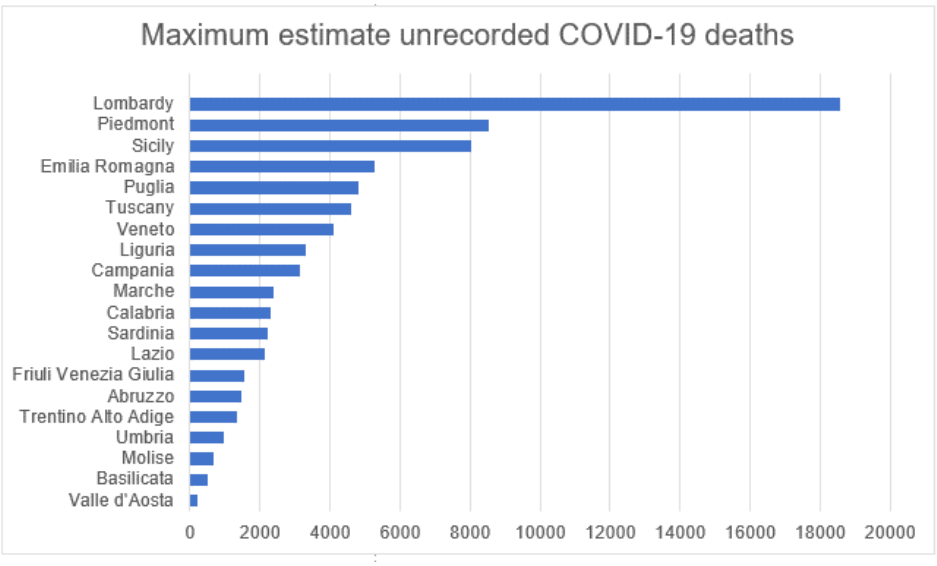

Figure 20. Histogram of the calculated maximum estimates of unrecorded COVID-19 deaths for the Italian's regions

6.3. Analysis of European Nations

The study was also applied to some European nations, including the United Kingdom, to determine a comparison between the different areas of the Union. Table 7 shows the predictive models applied and calculated accuracy metrics.

Table 7. Models and accuracy metrics of the analysed nations

Nations	Models	Accuracy metrics			
		MAPE	ME	MPE	NRMSE
Belgium	SARIMA(12,1,5)	0,06	-218,06	-0,01	0,23
France	SARIMA(12,1,4)	0,11	-6658,26	-0,12	0,5
Germany	SARIMA(12,0,3)	0,08	-6707,21	-0,08	0,31
Greece	SARIMA(12,1,8)	0,08	-414,93	-0,03	0,31
Italy	SARIMA(12,1,1)	0,001	74	0,001	0,13
Portugal	SARIMA(12,1,4)	0,08	-912	-0,08	0,47
United kingdom	SARIMA(12,0,8)	0,06	-979,52	-0,01	0,31
Romania	SARIMA(12,0,8)	0,05	243,12	0,02	0,16
Spain	SARIMA(12,1,7)	0,1	-29,51	0,01	0,39

Table 8 shows the results obtained for each country in references to the period March-September 2020.

Table 8. Results obtained from the analysis of the deaths recorded for each country (reference period March-September 2020)

Nations	Total deaths recorded	Total deaths expected from the model		EED: excess deaths estimated		COVID-19 deaths	UED: unrecorded COVID-19 estimated deaths	
		mean	min.	mean	max.		mean	max.
United Kingdom	405.191	333.953	251.472	71.238	153.719	42.233	29.005	**111.486**
Germany	553.752	551.126	438.174	2.626	115.578	9.495	-6.869	**106.083**
Italy	410.899	363.891	289.426	47.008	121.473	35.865	11.143	**85.608**
Spain	293.491	233.793	180.846	59.698	112.645	31.791	27.907	**80.854**
France	376.133	342.789	265.523	33.344	110.610	31.984	1.360	**78.626**
Romania	154.591	149.691	127.067	4.900	27.524	4.825	75	**22.699**
Portugal	68.227	61.287	45.504	6.940	22.723	1.971	4.969	**20.752**
Greece	63.700	59.479	50.700	4.221	13.000	391	3.830	**12.609**
Belgium	71.323	62.691	49.594	8.632	21.729	10.016	-1.384	**11.713**

At the top it is located the United Kingdom with a mean estimate of unrecorded COVID-19 deaths of 29.005 for up to 111.486 value. Following are Germany, Italy, Spain, and France. At the bottom of the list are Greece and Belgium with maximum estimates of unrecorded deaths of 12.609 deaths and 11.713 deaths. Some nations repost a negative mean estimate of COVID-19 deaths, this means that the value of officially registered deaths for COVID-19 exceeds the mean value of excess deaths (Figures 21 and 22).

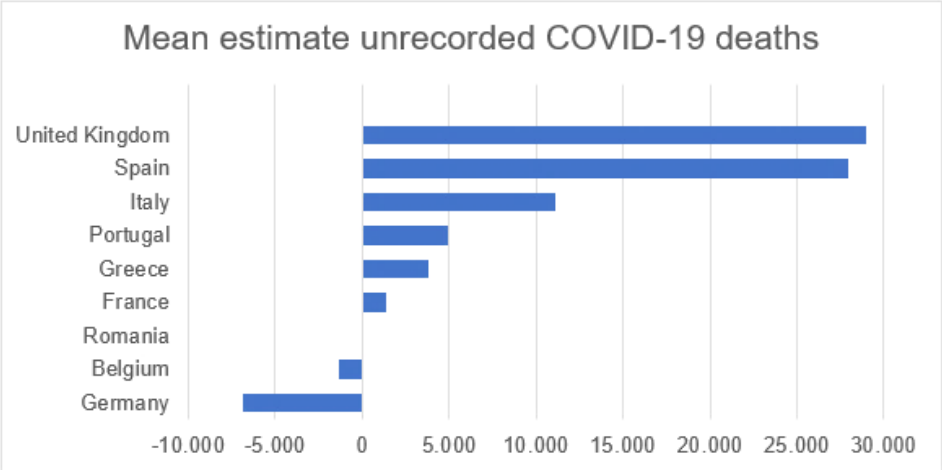

Figure 21. Histogram of calculated mean estimates of unrecorded COVID-19 deaths of the nations surveyed. Negative values are present in cases where the value of officially registered COVID-19 deaths exceeds the mean value of excess deaths estimated for the nation.

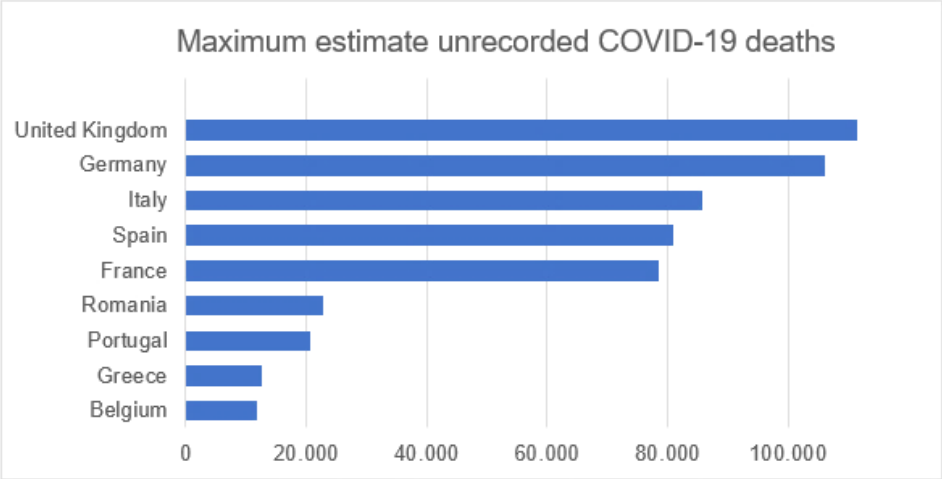

Figure 22. Histogram of calculated maximum estimates of unrecorded COVID-19 deaths in the nations surveyed

7. Conclusions

Even today, more than a year after the pandemic began, Italy, like the rest of Europe and the world, is still fighting an invisible enemy that has killed millions worldwide. The whole society, to be able to reduce the circulation of the virus, is still forced to make numerous sacrifices that have gradually led the nation to force, in addition to the health emergency, also an economic and social crisis. The analysis carried out was able to demonstrate the existence of a discrepancy between the excessive number of deaths recorded for all causes and those declared for COVID-19 during the first two months (March and April) of the pandemic, highlighting a fair number of unrecorded COVID-19 deaths. This makes us reflect on what could be the real figure to be attributed to the deaths caused by the virus.

The limitations produced by the study, such as the confidence in the data provided and the predictive models applied, should also be considered. A limitation is also given by the fact that all excess deaths are considered unrecorded COVID-19 deaths. There is also the possibility that a percentage may be made up of deaths caused indirectly by the virus but potentially resulting from the decrease in routine diagnosis and treatment of other conditions due to the health emergency or other causes not necessarily related to the spread of the virus.

An interesting in-depth analysis could be given by extending the entire analysis up to the current year, to be able to compare the results obtained from this analysis, relating to the initial phases of the pandemic that found an unprepared system, with those produced in the subsequent phases, in which methods have been devised to counter it. The objective could be to evaluate if and to what extent there are differences in the level of unrecorded COVID-19 cases and how much the applied methods have produced beneficial effects on the spread of the virus. It would also be interesting to analyze the trend of deaths recorded from flu, pneumonia, and other causes (not necessarily related to the symptoms of COVID-19) related to the pandemic period to be able to ascertain in a more precise and detailed way how the arrival of the coronavirus influenced the increase in deaths in 2020 in relation to the number of unrecorded deaths.

8. Declarations

8.1. Author Contributions

S.M. collected data, contributed to design the analysis, performed the analysis, wrote the code and wrote the paper. A.L.D. contributed to design the analysis and wrote the paper. A.M. contributed to design the analysis and revised the paper. All authors have read and agreed to the published version of the manuscript

8.2. Data Availability Statement

The data presented in this study are available in article.

8.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

8.4. Declaration of Competing Interest

The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

9. References

[1] Rivera, R., Rosenbaum, J. E., & Quispe, W. (2020). Excess mortality in the united states during the first three months of the COVID-19 pandemic. Epidemiology and Infection, n. 148, 264. doi:10.1017/S0950268820002617.

[2] Vandoros, S. (2020). Excess mortality during the Covid-19 pandemic: Early evidence from England and Wales. Social Science and Medicine, 258, 0277–9536. doi:10.1016/j.socscimed.2020.113101.

[3] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference, 92–96. doi:10.25080/majora-92bf1922-011.

[4] Jbrockmendel, J.R., McKinney, W., Van den Bossche, J.,…, Li, T. (2022). Pandas-dev/pandas: Pandas Software (v1.4.3). Zenodo. doi:10.5281/zenodo.6702671

[5] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. doi:10.1038/s41586-020-2649-2.

[6] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science and Engineering, 9(3), 90–95. doi:10.1109/MCSE.2007.55.

[7] Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). Introduction to time series analysis and forecasting. John Wiley & Sons, New Jersey, United States.