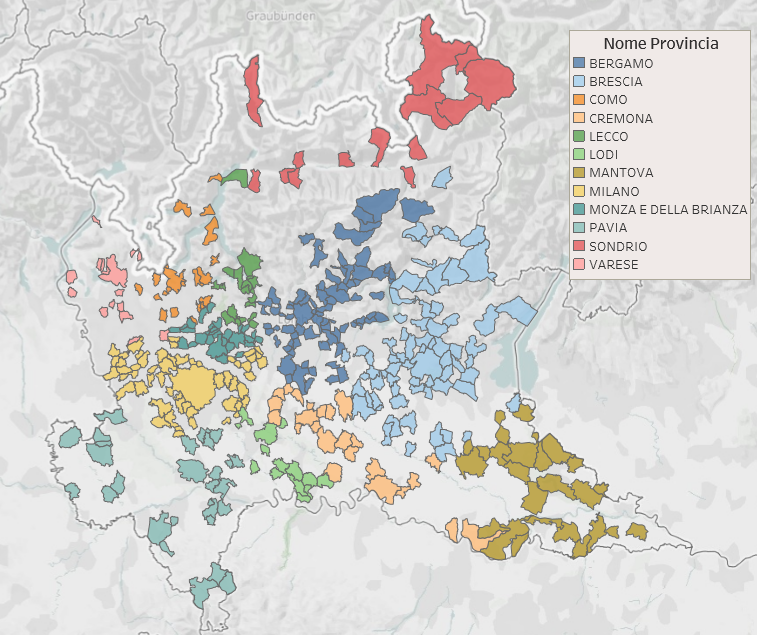
**2 Dataset Description**

Analyzed data were made available by ISTAT and contain informations about deaths distinguished by women, men and total, town and day, data are referred to the period from 1st January to 20th April from 2015 to 2020 [2]. The analyses were carried out on aggregated data at regional level, excluding municipalities whose data were not available for 2020 (‘DECESSI’ = 9999).

With regard to the attribute 'DATA\_INIZIO\_DIFF', it is not clear what it refers to. The interpretation we have given is that of beginning date of sharing data from the municipalities to ISTAT, despite this we have observed that municipalities groups of the Lombardia region present in every level of ‘DATA\_INIZIO\_DIFF’ were disjointed for this it was proceeded excluding only the municipalities whose value of ‘DATA\_INIZIO\_DIFF’ was equal to ‘Data 2020 n.a.’. After cleaning data phase it was able to conduct the analysis on 622 Lombard municipalities.



*Figure 1 Analysis available municipalities in Lombardia.*

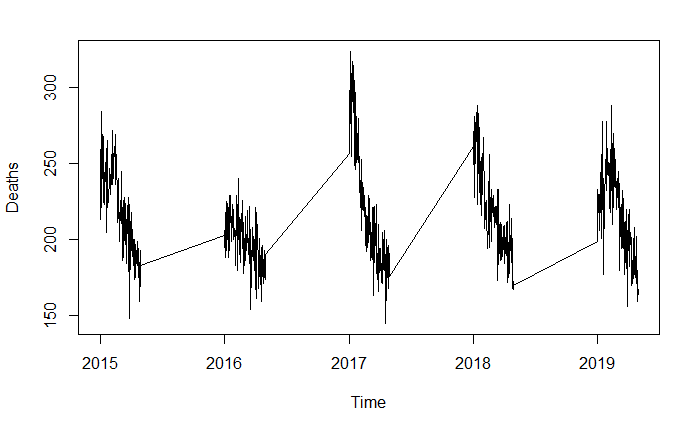
*Data source:* [*http://www.geoportale.regione.lombardia.it/*](http://www.geoportale.regione.lombardia.it/)

*Interactive version: https://public.tableau.com/profile/giulia.boschi#!/vizhome/comuni\_disponibili/Dashboard1*

Finally it was compared the expected deaths in the first quarter of 2020 with Covid verified deaths from Department of Civil Protection for Lombardia region [3].

**3 Modelling**

The available data design a particular daily time series with the following behavior.



*Figure 2 Lombardia deaths time series.*

*Source [2]*

As a rule, several trends can be identified in a time series [4, pp 14]:

* Trend component indicating the medium to long-term trend;
* Cyclical component that represents succession phases of prosperity and depression, usually included in the trend component, therefore named cycle-trend;
* Seasonal component representing the seasonal interim fluctuations;
* Random component that is the irregular and unobservable behavior present in each historical series.

In this particular case we don’t have any information about several months per year, therefore the identification of above mentioned component result incorrect considering a standard year. Anyway, we decided to proceed with a time series model imagining an unique time series due to forecast the three consecutive months.

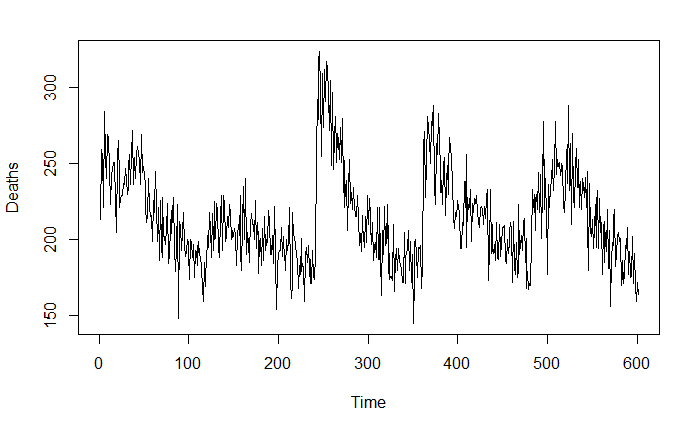


Figure 3 Time series considered behavior

The analysis has been based on the models of time series ARIMA for whose identification has been followed the Box and Jenkins procedure based on three steps:

1. Identification;
2. Estimation;
3. Diagnostics.

At the first step it was studied the general trend of the historical series: it was observed an important peak on 2017 that turned out to be a particularly black year for the influence [5] (it was tested a model with intervention on 2017, but did not result significance). It was identified the Box and Cox λ with value 0.14, so it was chose to not adopt a logarithmic transformation [6, pp 101].

It was conducted the Dickey Fuller test (: time series is nonstationary) which reported a p-value of 0.09 that would result in an acceptance of the null hypothesis for the usual thresholds and, consequently, a differencing operator should be introduced [6, pp 128].

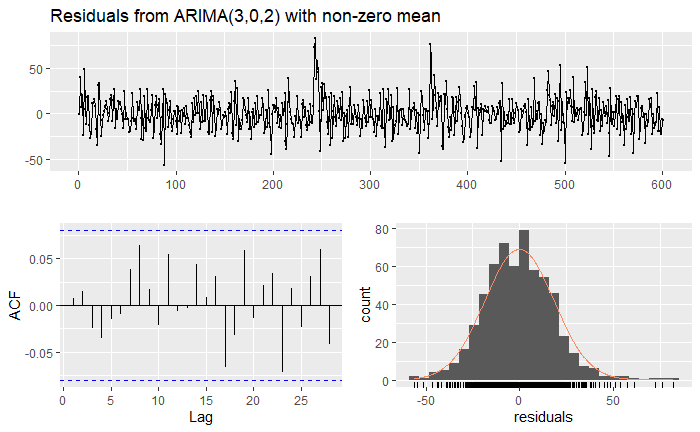
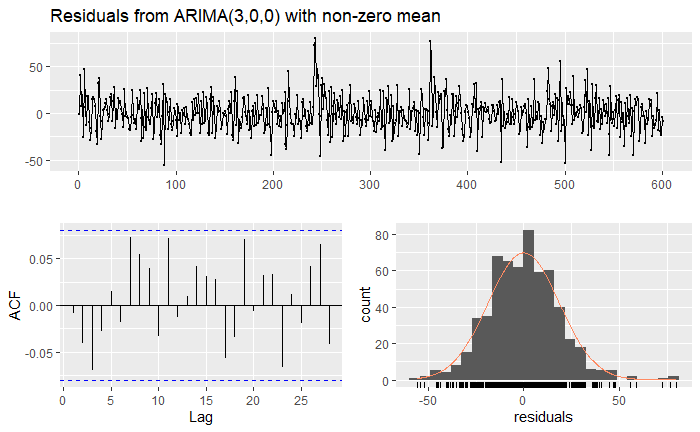
Because of the peculiarity of the series, it was preferred to not apply any difference in order to not introduce bias caused by differentiating not consecutive data.

Due to test the statistical significance of the ARIMA parameters estimated it was used the T test

At the diagnostic step it was applied Jarque-Bera test for residuals normality, which is based on the fact that a normal distribution has zero skewness and zero kurtosis. In addition to looking at residual correlations at individual lags, the Ljung Box test was used: with [6, pp 183]. The model comparison was based on the mean absolute percentage error (MAPE) and Akaike’s information criterion (AIC) indicators.

The choice fell on the models AR(3) as specified: and ARMA(3,2): .

In the second model the values of AIC and MAPE are slightly lower than the first; in both models the residuals are not normal for the Jarque Bera test, but show a good adaptation of the distribution to the Normal curve. The second model was chosen because the Q value for the Ljung Box test is halved, ensuring non correlated error at all meaningfulness thresholds.



*Figure 4 Residuals diagnostic plot for the two considered models obtained with the R function ‘checkresiduals()’*

**Criticità sul modello da aggiungere in Conclusions & Criticalities**

The biggest problem we have encountered in the analysis of time series is the lack of data for most of the year. This made us to imagine a time series consisting of the first months for each year. The analysis would have been more complete and correct if it had been possible to consider all the months of the year, so as to be able to identify the seasonality that distinguishes them and the general trend.

**Riferimenti**

[2] <https://www.istat.it/it/archivio/240401>

[3] [httppc/COVID-19/blob/master/dati-regioni/dpc-covid19-ita-regioni.csv](https://github.com/pcm-dpc/COVID-19/blob/master/dati-regioni/dpc-covid19-ita-regioni.csv)

[4] Vajani L. (1982), Analisi statistica delle serie temporali Volume primo, Cleup

[5] <https://www.repubblica.it/salute/2017/03/18/news/l_anno_nero_dell_influenza_morti_ventimila_anziani_in_piu_-160814115/>

[6] Cryer J., Chan K. (2008), Time Series Analysis, Springer