SF2943 Time Series Analysis - Project

Study of the Suicides Rate Evolution in the Unites States

Thomas Bouquet, Rodrigo Castellano Ontiveros, Alejandro Garcia Castellanos, Paul Lieutier, Luca Marini

Members personal numbers

Alejandro Garcia Castellanos: 990722-5552
Thomas Bouquet: 20000301-T095
Rodrigo Castellano Ontiveros: 19970119-4632.
Paul Lieutier: 20000111-T178
Luca Marini: 19980422-1357

1 Introduction

Using the tools of time series analysis, we conducted a study on the evolution of the suicide rates in the United States between 1999 and 2020 in order to highlight certain structures inherent in these data and thus better understand and anticipate this sad phenomenon.

2 Methods

The Python packages that we have used for the time series analysis is `statsmodels` [3], used to decompose seasonality, trend, hypothesis testing and model fitting. Also, `pmdarima` [4] was used to look for the best model.

3 Visualisation and study of the dataset

3.1 Raw dataset

To conduct this study, we used the dataset [1]. At first glance of the dataset on Figure 1, the suicide rates has an increasing trend throughout time. Besides, in spite of this increase, there seems to be a recurring pattern in the form of four *bumps* per year which could correspond to seasonal peaks consistent with particular events of the year As a consequence, the time series is most likely not stationary.

Note that we are using the suicide rates rather than the absolute number of deaths because we already see a clear increase of population in USA [2] (see Figure 2). Therefore, we prefer that our model represents the amount of suicides, regardless of the amount of population.

3.2 Seasonality

As explained in the previous section, a recurring pattern seems to appear in the time series. In order to better visualize this apparent seasonality, it is easiest to display the annual figures all together so that they can be directly compared and it is easier to see if their changes over time coincide. Figure 3 gathers this annual data in one plot.

It is now evident that similar patterns appear every year at the same time. The easiest example to observe is the significant drop in the number of suicides in February. This temporal series thus has a notable seasonality which we will examine later.

3.3 Trend

To begin with, we computed the corresponding Rolling Average and Rolling Standard Deviation with a period of 12 months (1 year) since our study of the previous section revealed an evident yearly behaviour. As we can see on Figure 4, we obtain a non constant trend which indicate non-stationarity of the time series.

3.4 Dickey-Fuller Test

Nevertheless, in order to be fully convinced that we have a non-stationary time series we used the Dickey-Fuller Test. We obtained a p-value of 0.795277, which verify our hypothesis.

4 Removing trend and seasonality

4.1 Model

First we have to consider our possible models: additive model and multiplicative model. The additive model is based on the assumption that our time series $\{X_t\}$ can be written as:

$$\forall t,\ X_t = Y_t + m_t + s_t\,,$$

where Y_t is a stationary noise, m_t is the trend and s_t is the seasonality (with period 12 based on our observations). While the multiplicative assumes that that our time series $\{X_t\}$ can be written as:

$$\forall t,\ X_t = Y_t \cdot m_t \cdot s_t\,.$$

Using the *seasonal_ decompose* method from [3], which uses moving averages, we can obtain the trend and seasonality for each of the models. By applying the Dickey-Fuller test to the residuals of each model we see that the p-value of the additive model ($1.68e-13$) is slightly better than the multiplicative one ($8.36e-13$). Therefore we will use the additive model (see Figure 5)

4.2 Differentiation

However, forecasting using the moving average can be harder than using reverting differentiation. Therefore, we will first do first a lag 12 differentiation for removing the seasonality. We obtained a p-value of $1.26e-03$ in the DF test which would be enough to consider the time series stationary. However we can see in Figure 6 that the mean is not completely constant. Therefore, after applying one step differentiation we obtain a p-value of $8.77e-08$, and almost constant mean (see Figure 7).

5 Autocorrelation and Partial Autocorrelation functions

Figure 8 contains the autocorrelation and partial autocorrelation functions. We can see that in the acf most of the points are inside the confidence interval when $lag > 2$. This gives us certain confidence towards trying an MA(2) process, or a MA(14) one if we want a higher confidence. Furthermore, we do not see a clear geometrical decay in the acf, which has a somehow indication that it may not be an AR process, even though we see that pacf might indicate that it could be an AR(8) process (or a AR(12) process).

6 ARMA modeling

6.1 Finding the best ARMA model

In this section we use the corrected Akaike information criterion (AICC) to find which is the best ARMA(p,q) model. These models are based on the assumption that our time series $\{X_t\}$ is stationary and it can be written as:

$$\forall t, X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

where $\{Z_t\} \sim \text{WN}\left(0, \sigma^2\right)$ and the polynomials $(1 - \phi_1 z - \ldots - \phi_p z^p)$ and $(1+\ \theta_1 z + \ldots + \theta_q z^q)$ have no common factors.

For the simpler models AR and MA, the results are shown in table 1. We can see how higher order in both MA and AR give lower AICC values and thus better results, which could be due to over-fitting. The model that gave the lowest AICC was MA(14).

We also executed an extensive grid search to see which ARMA(p,q) is the best, with p and q ranging from 0 to 20. And we obtained that the model that minimizes the AICC is actually the ARMA$(0,14)$ (MA(14)). However, looking at the p-values of the coefficients we see that a great majority have a high value (see Figure 9). Therefore, we might be having some case of overfiting by using the MA(14) model. Regardless, the obtained model is

$$\begin{gathered}\theta(z) = 1 - 0.5010z - 0.1705z^2 - 0.0724z^3 + 0.0382z^4 - 0.0303z^5 - 0.0185z^6 - 0.0128z^7 - 0.0761z^8 \\ +0.1230z^9 - 0.0923z^{10} + 0.1264z^{11} - 0.8804z^{12} + 0.3430z^{13} + 0.2633z^{14}\,, \\ \phi(z) = 1\,, \text{ and } \sigma^2 = 0.0010\,.\end{gathered}$$

In Figure 11 we can also notice that the residuals follow a normal distribution, therefore the fitted parameters are reliable.

6.2 Finding the best SARIMA model

In order to have a wider range of models, and since we are already using differentiating plus ARMA models, then we also implemented the SARIMA model. The SARIMA model is defined as follows:
If d and D are nonnegative integers, then $\{X_t\}$ is a seasonal ARIMA $(p,d,q)\times\ (P,D,Q)_s$ process with period s if the differenced series $Y_t = (1-B)^d\,(1-B^s)^D\,X_t$ is a causal ARMA process defined by

$$\phi(B)\Phi\left(B^s\right)Y_t = \theta(B)\Theta\left(B^s\right)Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$, $\acute{}\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P$, $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$, and $\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$.

We performed a stepwise search using the *auto_arima* function from [4], and the best parameters found were $(p,d,q) = (1,1,1), (P,D,Q)_s = (1,0,1)_{12}$. The AICC score of this SARIMA model is equal to -1026.30, which is way lower than the previous ARMA$(0,14)$ process. In addition, all the p-values of the coefficients of this model are almost equal to zero (see Figure 10).The obtained model is

$$\phi(z) = 1 + 0.3314z\,, \Phi(z) = 1 + 0.7454z\,, \theta(z) = 1 - 0.8619z\,, \Theta(z) = 1 + 0.9880z\,, \sigma^2 = 0.0010\,.$$

In Figure 12 we can also notice that the residuals follow a normal distribution, therefore the fitted parameters are reliable.

6.3 Predictions of the ARMA model and SARIMA model

Figure 13 and 14 show the suicide rate predictions for the years 2021 and 2022 given respectively by the ARMA$(0, 14)$ and the SARIMA$(1, 1, 1) \times (1, 0, 1)_{12}$ processes.

The predictions are promising for 2021 and 2022 taking into account the similarity of the plotted curves of the real values and the predictions from previous years. The difference between the ARMA model and SARIMA model is very small, the predictions are very similar and a slight difference can be seen in the confidence intervals (grey area).

7 Discussion

In conclusion, if we want to achieve the best overall fit we could use the SARIMA$(1, 1, 1) \times (1, 0, 1)_{12}$ model. However, if we want to prioritize interpretability (at the cost a little worse fit) we could use the ARMA$(0, 14)$ model with the differentiating step.

8 Peer review of group 24

In their report an analysis of the daily temperatures in Dehli between the years 2013 and 2017 is presented.

First, seasonality and trend are subtracted. In order to test for stationarity they performed the Augmented Dickey-Fuller Test, with a p-value of 0.001. Next, the best model was found by looking to the acf and pacf and later comparing AICC of different AR, MA and ARMA models, resulting in the final choice of ARMA(2,2). Finally, a prediction was done for 2018 based on the seasonality of 2017 and the k-value of the trend.

Their work is relevant in the sense that predicting temperature values could be described as an ARMA(2,2) model, with the respective equation shown. Besides, it is interesting the part in which they mention that by not taking into account the first three years, the results are quite different. Nevertheless, it is not clear up to what point data should not be taken into account and the reasons why. The model assumptions regarding acf and pacf were right, but the model selection by AICC showed a different selection of the most suitable model (they correctly comment this). Maybe it could have been written with more detail some parts such as how the trend and seasonality was subtracted, but the report in general is good and the presentation seems correct.

References

[1] CDC. Centers for disease control and prevention. URL: `https://wonder.cdc.gov/controller/saved/D76/D285F319`.

[2] FRED. Population of the usa from 1999 to 2022. URL: `https://fred.stlouisfed.org/series/POPTHM`.

[3] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[4] Taylor G. Smith. pmdarima, 2 2022. URL: `https://github.com/alkaline-ml/pmdarima`.

A Figures and tables

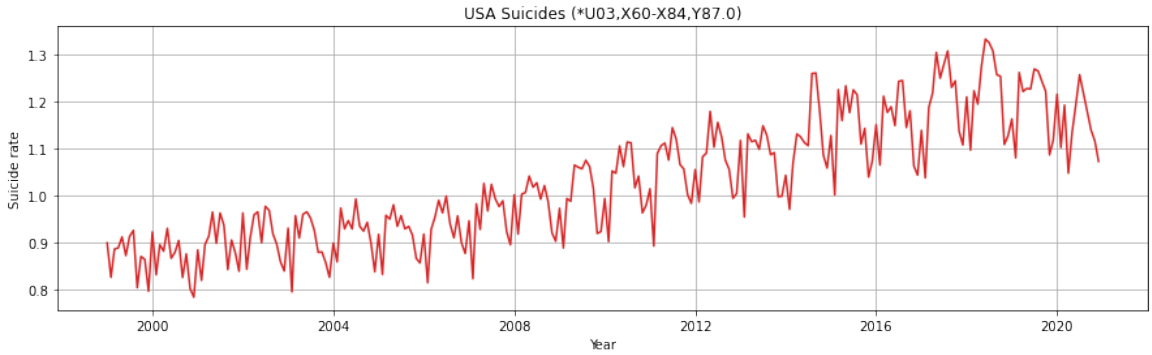

Figure 1: Evolution of the number of suicides in the United States

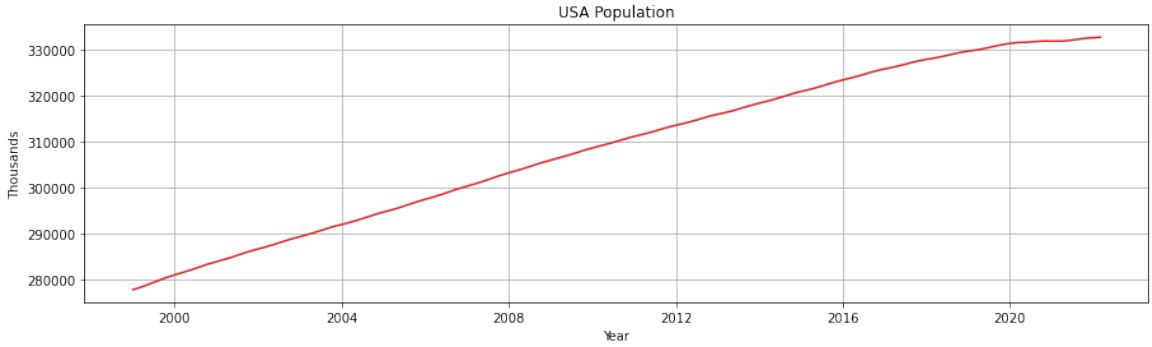

Figure 2: USA population from 1999 to 2022

MODEL	AICC
MA(2)	-887.50
MA(14)	-974.06
AR(8)	-880.43
AR(12)	-929.24

Table 1: AICC for AR and MA models

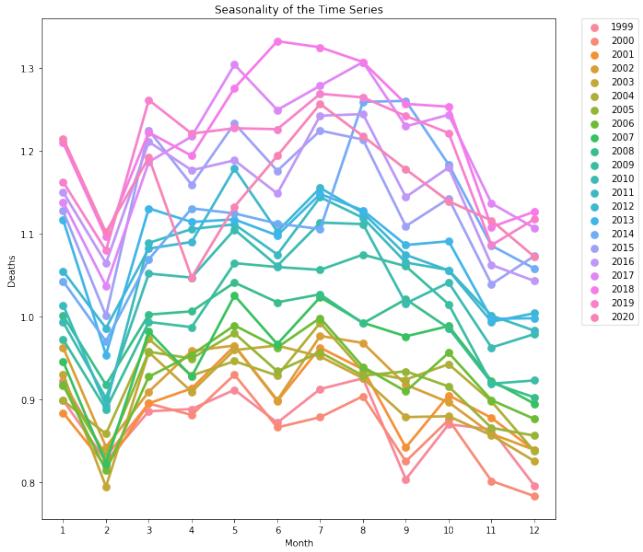

Figure 3: Evolution of the number of suicides per year

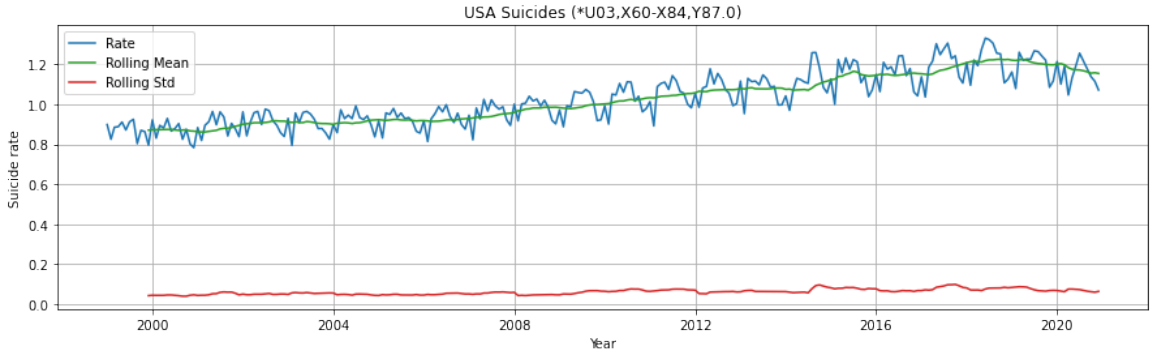

Figure 4: Moving average / standard deviation

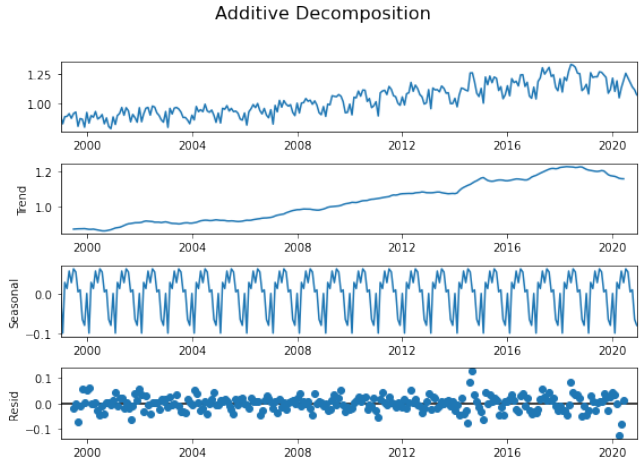

Figure 5: Additive decomposition of the time series

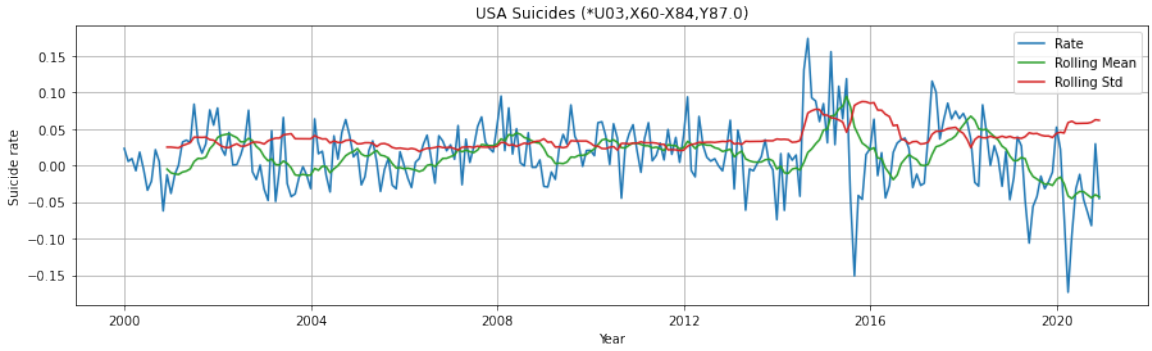

Figure 6: Differentiating with a periodicity of 12

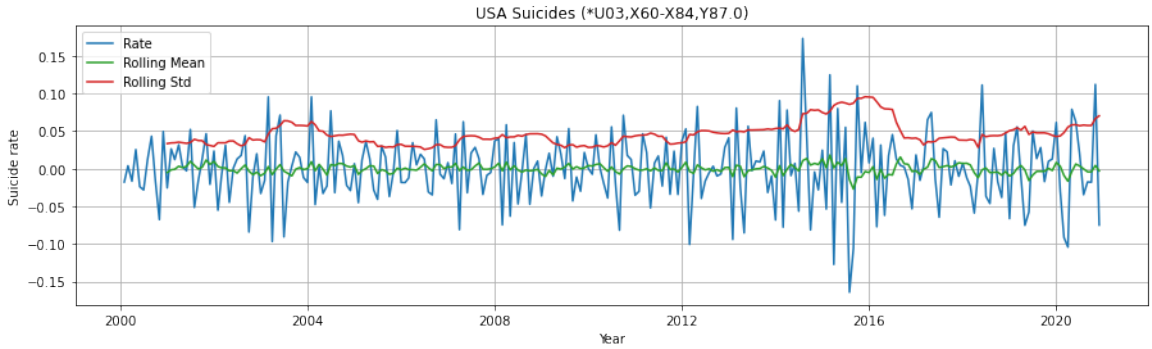

Figure 7: Differentiating with a periodicity of 1

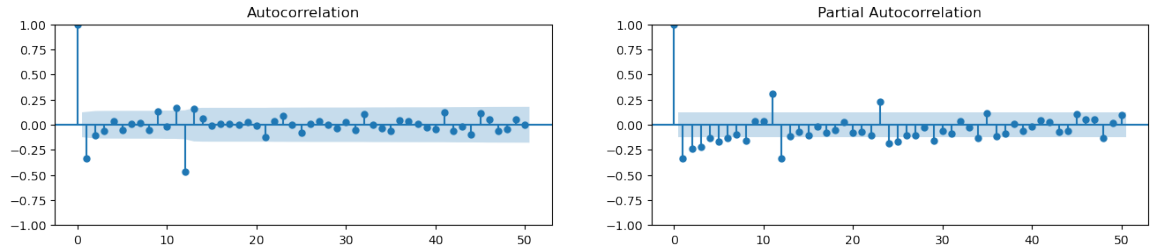

Figure 8: Autocorrelation and partial autocorrelation functions

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.5010	0.056	-8.899	0.000	-0.611	-0.391
ma.L2	-0.1705	0.076	-2.248	0.025	-0.319	-0.022
ma.L3	-0.0724	0.058	-1.257	0.209	-0.185	0.041
ma.L4	0.0382	0.055	0.698	0.485	-0.069	0.146
ma.L5	-0.0303	0.049	-0.624	0.533	-0.126	0.065
ma.L6	-0.0185	0.066	-0.280	0.779	-0.148	0.111
ma.L7	-0.0128	0.066	-0.196	0.845	-0.141	0.116
ma.L8	-0.0761	0.062	-1.238	0.216	-0.197	0.044
ma.L9	0.1230	0.058	2.110	0.035	0.009	0.237
ma.L10	-0.0923	0.066	-1.401	0.161	-0.221	0.037
ma.L11	0.1264	0.062	2.032	0.042	0.004	0.248
ma.L12	-0.8804	0.057	-15.484	0.000	-0.992	-0.769
ma.L13	0.3430	0.064	5.363	0.000	0.218	0.468
ma.L14	0.2633	0.066	3.974	0.000	0.133	0.393
sigma2	0.0010	8.23e-05	12.135	0.000	0.001	0.001

Figure 9: Coefficients and their confidence of the ARMA$(0, 14)$ process

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3314	0.064	5.148	0.000	0.205	0.458
ma.L1	-0.8619	0.045	-19.274	0.000	-0.950	-0.774
ar.S.L12	0.9880	0.009	108.858	0.000	0.970	1.006
ma.S.L12	-0.7454	0.061	-12.278	0.000	-0.864	-0.626
sigma2	0.0010	5.83e-05	17.852	0.000	0.001	0.001

Figure 10: Coefficients and their confidence of the SARIMA$(1, 1, 1) \times (1, 0, 1)_{12}$ process

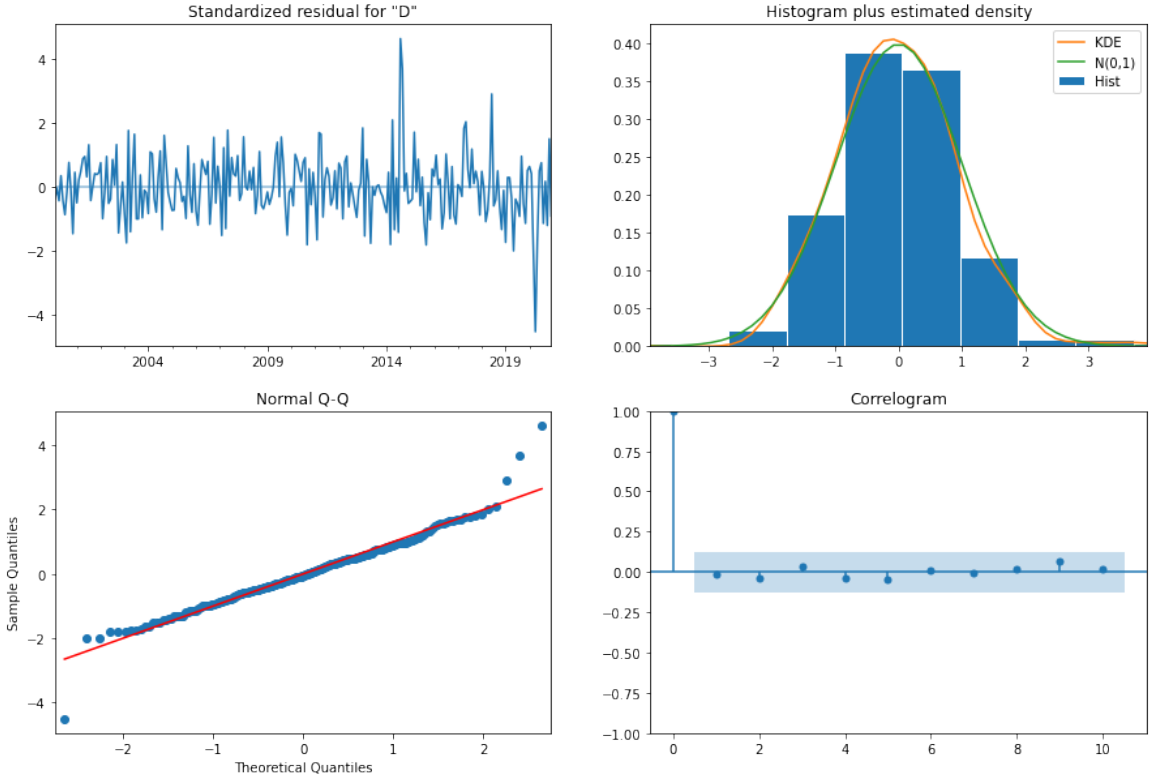

Figure 11: Diagnostic plots of ARMA$(0, 14)$ process

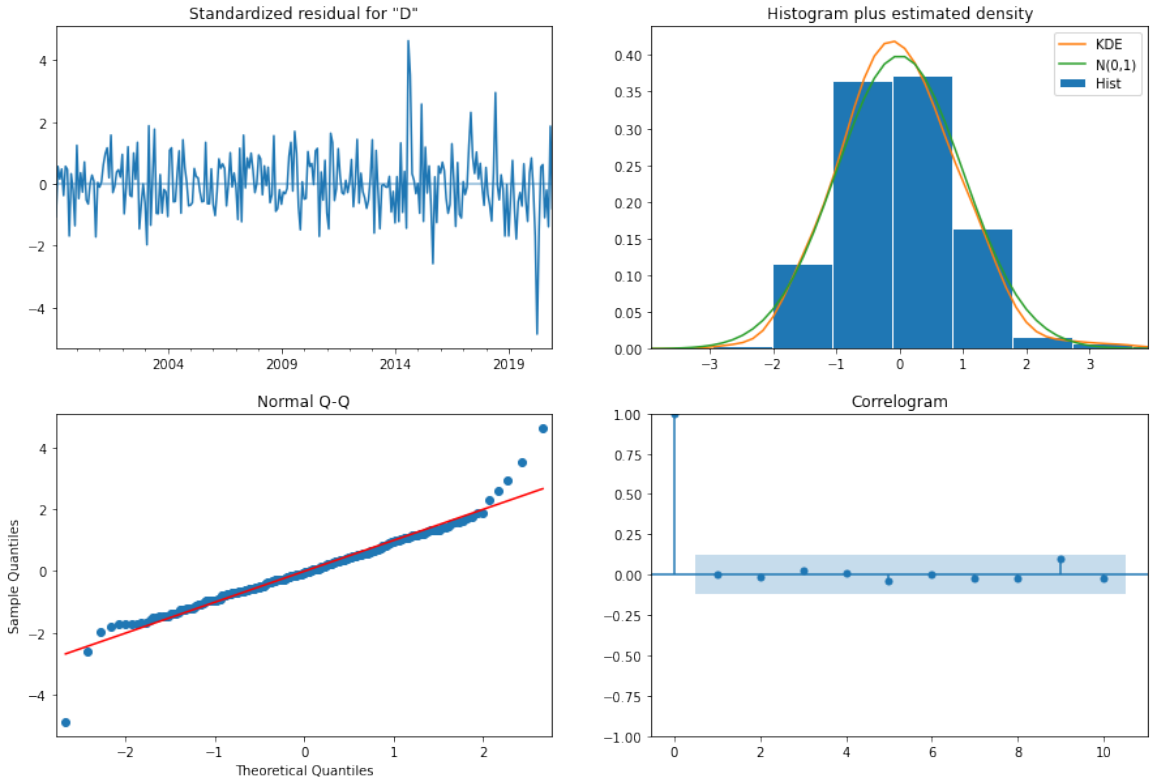

Figure 12: Diagnostic plots of SARIMA$(1,1,1) \times (1,0,1)_{12}$ process

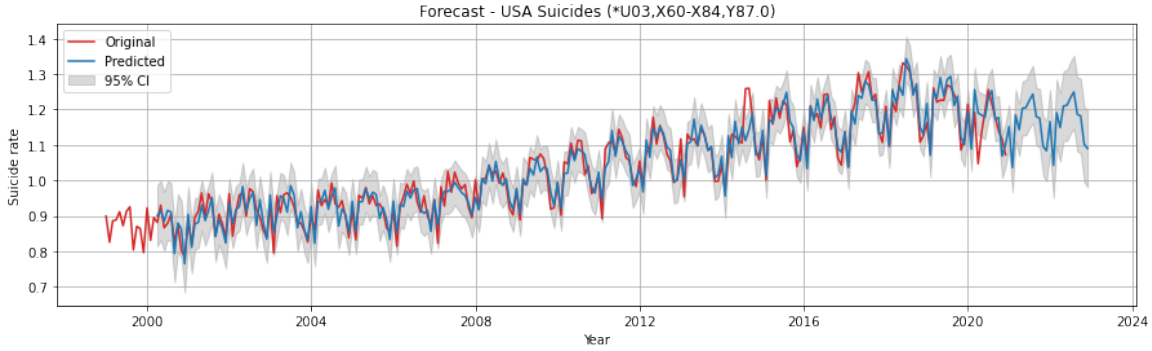

Figure 13: Predictions using ARMA$(0,14)$ process

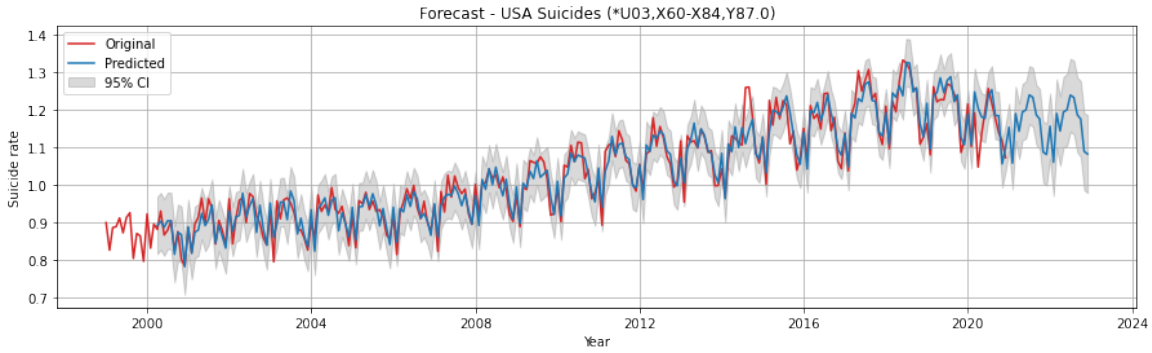

Figure 14: Predictions using SARIMA$(1,1,1) \times (1,0,1)_{12}$ process