UNIVERSITÉ LE HAVRE

U.F.R OF SCIENCE AND TECHNOLOGY

Modeling COVID-19 data

RATA Lotfi

Proposed by Mr.AMBROSIO & Mr.ALAOUI

College year 2021-2022

1 Abstract

In this article, we use a SIRD model to analyze the evolution of the COVID-19 pandemic in New-York State, This model is composed by a nonlinear system of differential equations that allows us to detect trends in the pandemic and make reliable predictions of the course of the infection in the short term. The disadvantage of this model lies in these parameters. That is, a slight change in the transmission rate (k), recovery rate (r) or mortality rate (d) of the latter can change the prediction results. Moreover, it is difficult to make a long-term forecast. However, the “SIRD” model is very effective for short-term forecasting and for adjusting its parameters to the data.

In this study, the "SIRD", "ARIMA" and "POLY" models are used to predict the evolution of infected, cured and deceased cases from real New-York state population data modeled by statisticians. . We can then compare the prediction results to validate the most suitable model for predicting future infected, cured and deceased cases.

2 Introduction

The coronavirus (COVID-19) is an infectious disease that can be dangerous. Indeed, it can cause respiratory problems, muscle pain and loss of taste...). Transmission of the disease is possible with every contact between infected people and people likely to catch up with the virus, which can increase the cases of death for some people, especially those with chronic illnesses and who are frail. During the COVID 19 outbreak, more than 200 countries were severely affected and faced significant challenges, including New-York State , which went through a great economic crisis [**7**]. More than 79,000 deaths have been reported by the state governor [**13**]. new cases are still on the rise and the development of a vaccine is overdue for this new epidemic, which has forced the state to impose confinement on its population in order to limit the spread of the disease and have time to develop an effective treatment [6] . In order to find a solution to the problem of the pandemic, research and predictions have been made by scientists and state statisticians [3]. However, several strategic public health questions have been asked, How many infected cases will there be in the coming months ? How to predict the new infected cases of the population after one or two months from the real data ? These questions could have answers thanks to epidemiological and statistical forecasting models. The "SIRD" model was proposed by the two scientists William O. Kermack and Anderson G. McKendrick. It divides the population into four compartments : Susceptible, Infected, Recovered, and Deceased. This will be detailed below. [10] [12] [4] [2], The auto-regressive integrated moving average model "ARIMA" is based on stationary time series analysis [9] [11] [12] and the "POLY" polynomial regression model uses least squares fitting to find the coefficients of a polynomial of degree n that best fits a data curve [12]. However, the results of the forecasts are not necessarily reliable. It is therefore necessary to evaluate the performance of these models in order to verify the reliability of the prediction thanks to the statistical indicators of the error such as the mean absolute error "MAE" and the root mean square error "RMSE".

3 Curve fitting and forecasting models

3.1 The SIRD Model

The SIRD model is one of the simplest models used in epidemiology. It separates the population into four compartments :

↪ S (Susceptible) : Represents the people who could become infected.

↪ I (Infected) : Represents the people who are infected by the epidemic .

↪R (Recovered) : Represents the people who have had the disease and are now healthy.

↪ D (Deceased) : Represents the people who have died of the disease.

If we assume that the population affected by the epidemic is $N = S + I + R + D$ The model SIRD is governed by the following system of differential equations :

$$\begin{cases} S^{'}(t) = -k\ S(t)\ I(t) \\ I^{'}(t) = k\ S(t)\ I(t) - r\ I(t) - d\ I(t) \\ R^{'}(t) = r\ I(t) \\ D^{'}(t) = d\ I(t) \\ (S(0), I(0), R(0), D(0)) = (S_0, I_0, R_0, D_0) \end{cases}$$

With

↪ k : The rate of transmission of the epidemic from an infected person to a person who has not yet been infected (is a probability that has a value in the interval $[0, 1]$ [**1**]) .

↪ r : The cure rate (is a probability that has a value in the interval $[0, 1]$ [**1**]).

↪ d : The death rate,(is a probability that has a value in the interval $[0, 1]$ [**1**]).

Figure 1 shows the usual way in which the variables S, I, R, and D evolve :

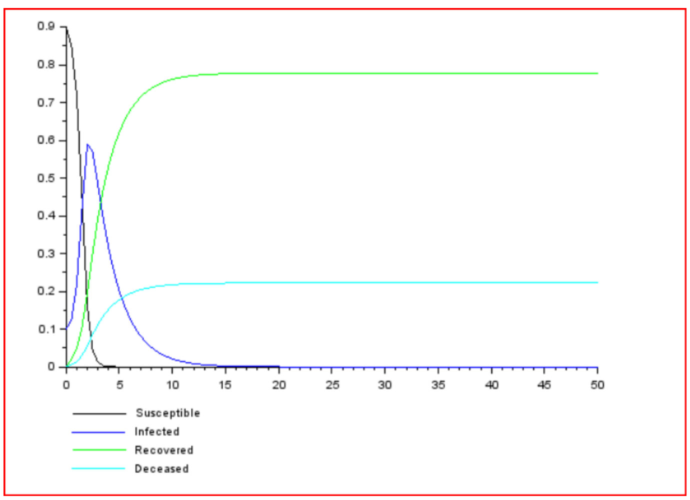

FIGURE 1 – Typical behavior of an epidemic's evolution over time

3.2 Time Series Forecasting with ARIMA

The ARIMA method is a forecasting method which makes it possible to analyze the time series proposed by Box-Jenkins (1970) [**8**][**9**] [**11**], The use of this method requires a study of the stationarity of the series temporal. In the case of non-stationarity of the series, this one will have to be differentiated in order to obtain the stationarity. The purpose of ARIMA models is to make savings on the future evolution of a phenomenon, it is a model that is widely used in the field of econometrics, statistics and epidemiological analysis. The ARIMA model can be broken down into three parts :

3.2.1 AR(p) auto-regressive model

Each observation of the auto-regressive model consists of a random component (white noise) and a linear combination of previous observations, i.e. :

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t$$

With :

ϵ_t : is a random variable called "white noise" which is independently distributed.

$\phi_1, ..., \phi_p$: they are real coefficients of the auto-regressive model.

y_t : Observation at the moment t

p : The order of autoregressive models

3.2.2 The integrated part (I(d))

This part represents the number of differentiation observations needed to make the time series stationary with :

d : The order of differentiation of non-stationary series

3.2.3 MA(q) moving average pattern

For each moving average pattern, the observation is composed of a random error component (white noise) and a linear combination of past random errors, in other words :

$$y_t = \epsilon_t - \theta_1 \epsilon_{t-1} - ..... - \theta_q \epsilon_{t-q}$$

With :

ϵ_t : is a random variable called "white noise" which is independently distributed.

$\theta_1, ..., \theta_q$: they are actual moving average model coefficients.

q : The order of the moving average pattern.

The ARIMA model which is denoted ARIMA(p,d,q) is a combination of an auto-regressive model of order p and a moving average of order q on the variable y differentiated d times.

Example of ARIMA model :

$$ARIMA(1,1,2): \quad y'_t = \phi_1 y'_{t-1} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2}$$

3.3 Prediction with the polynomial regression model

Polynomial regression (POLY) [**12**] uses least squares fitting to find the coefficients of a polynomial of degree n that best fits a curve. A certain curve Y can be approximated by the polynomial function with coefficients C_0 à C_n. In other words by the following polynomial :

$$Y_a = C_0 \ X^n + C_1 \ X^{n-1} + ... + C_{n-1} \ X + C_n$$

In our case, X is the time unit and Y_a is the forecast value for the time unit X (per day).

4 Experimental Work

We consider Worldometer data provided by New-York State statisticians from March 12, 2020 to April 06, 2022 expressed in millions [**3**]. Our objective is to seek to approach the three curves of the data. We will therefore first build the three curves corresponding to the Worldometer data : "Number of people infected (I)", "Total number of cases (I+R+D)" and "Number of deaths (D)"

Our work consists in this section in finding the coefficients k, r, d such that the model $SIRD$ approximates the curves of the worldometer data.

↪ In this work, we have chosen to use the Scilab software to present the real worldometer data and the data estimated by the SIRD model during this period.

↪ The initial conditions corresponding to March 12, 2020 are given by :

$$S_0 = 20000000, \quad I_0 = 328, \quad R_0 = 0, \quad D_0 = 0$$

↪The variation of the parameters k, r, d estimated by the model SIRD during each period are given in table 1 and the results obtained are presented in figures 2, 3, 4 and 5.

Période	k	r	d
12 March - 21 March 20	0.3	0.001	0.07
22 March - 27 March 2020	0.18	0.03	0.12
28 March - 09 April 2020	0.1	0.21	0.13
10 April - 24 April 2020	0.0232	0.05	0.075
25 April - 12 May 2020	0.0135	0.09	0.015
13 May - 29 May 2020	0.008	0.09	0.01
30 May - 01 July 2020	0.0061	0.09	0.003
02 July - 09 August 2020	0.001	0.55	0.0019
10 August - 17 August 2020	0.001	0.55	0.0019
18 August - 26 September 2020	0.0082	0.17	0.018
27 September - 16 November 2020	0.02	0.17	0.0018
17 November - 25 November 2020	0.02	0.17	0.0018
26 November - 31 December 2020	0.025	0.0008	0.0057
01 January 2020 - 19 January 2021	0.0205	0.03	0.0046
20 January - 29 January 2021	0.015	0.076	0.0042
30 January - 18 February 2021	0.00854	0.1	0.0032
19 February - 19 April 2021	0.0101	0.29	0.0019
20 April - 24 May 2021	0.0025	0.22	0.00145
25 May - 04 July 2021	0.0029	0.46	0.00087
05 July - 23 July 2021	0.011	0.7	0.0009
24 July - 25 July 2021	0.02775	0.198	0.002
26 July - 22 August 2021	0.02775	0.198	0.002
23 August - 21 September 2021	0.024	0.045	0.00255
22 September - 14 October 2021	0.0128	0.045	0.0019
15 October - 01 November 2021	0.0138	0.18	0.00145
02 November - 07 Décember 2021	0.0138	0.18	0.00145
08 Décember 2021 - 09 January 2022	0.045	0.095	0.0015
10 January - 06 February 2022	0.0155	0.0388	0.00158
07 February - 07 March 2022	0.0024	0.08	0.00052
08 March - 06 April 2022	0.002	0.65	0.0002

TABLE 1 – **Variation in transmission rate (k), cure rate (r) and death rate (d) during the period from March 12, 2020 to April 06, 2022**

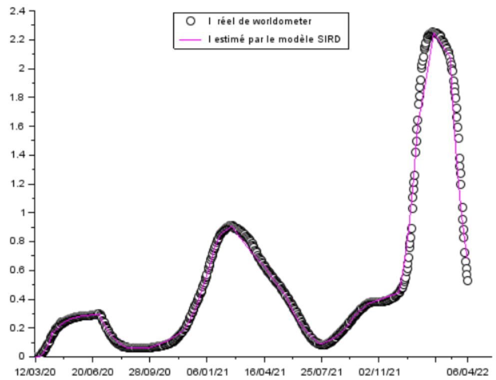

FIGURE 2 – *The adjustment of the number of infected people estimated by the SIRD model to the number of infected people corresponds to the data of the worldometer during the period March 12, 2020 - April 06, 2022*

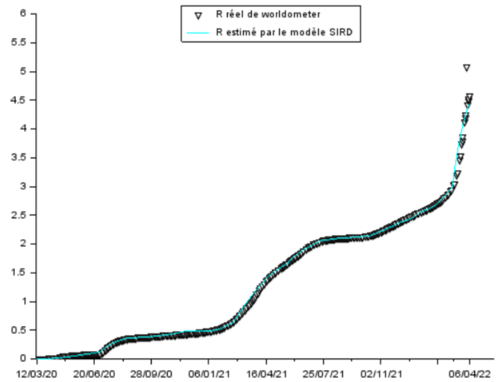

FIGURE 3 – The adjustment of the number of cured people estimated by the SIRD model to the number of cured people corresponds to the data of the worldometer during the period March 12, 2020 - April 06, 2022

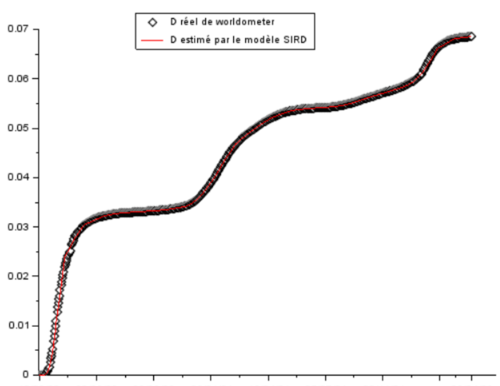

FIGURE 4 – The adjustment of the number of deceased people estimated by the SIRD model to the number of deceased people corresponds to the data of the worldometer during the period March 12, 2020 - April 06, 2022

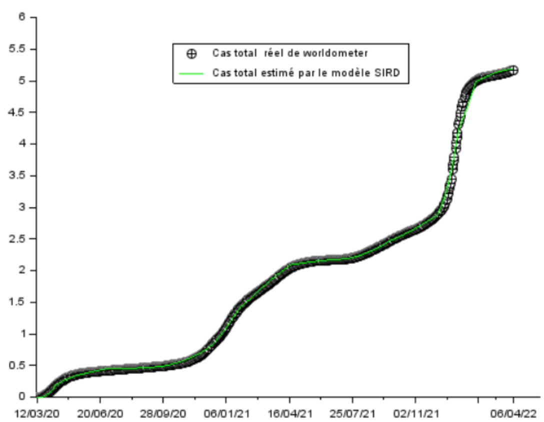

FIGURE 5 – The adjustment of the total number of cases estimated by the SIRD model to total the number of deceased people corresponds to the data of the worldometer during the period March 12, 2020 - April 06, 2022

4.1 Prediction of the evolution of COVID-19 in the state of New York after 1 month (30 days)

4.1.1 prediction with the ARIMA model

In this part, we applied the following ARIMA models (with differentiation) :

↪ For infected cases : The model $ARIMA(6,1,9)$ was chosen

↪ For cured cases : The model $ARIMA(1,2,3)$ was chosen

↪ For deceased cases : The model $ARIMA(1,2,2)$ was chosen

We will now make a 30-day forecast (April 06 - May 06, 2022) with the ARIMA model with a differentiation of the actual data that corresponds to the numbers of people infected, cured and died. The forecast results are shown in Figures 6, 7 and 8 :

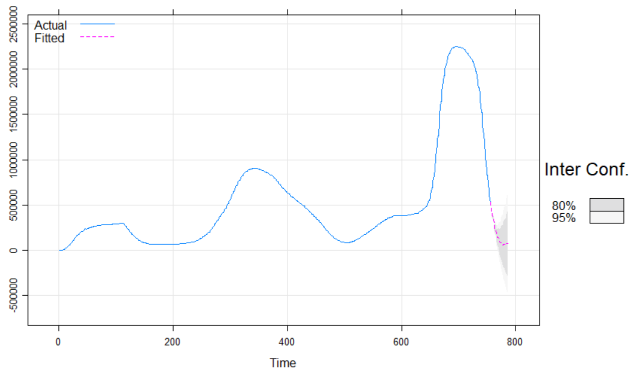

FIGURE 6 – *The prediction of the evolution of the number of infected people after 30 days (06 April - 06 May 2022) with the ARIMA model (6,1,9)*

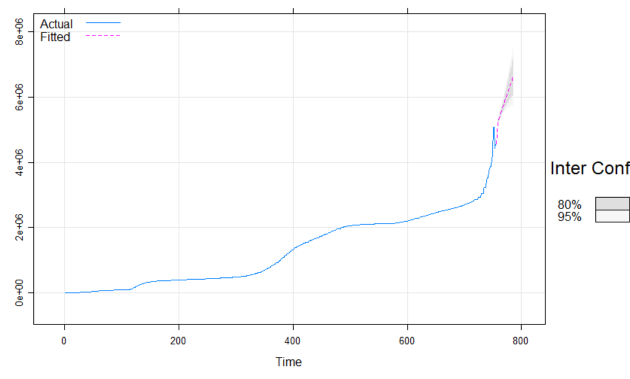

FIGURE 7 – *The prediction of the evolution of the number of people cured after 30 days (06 April-06 May 2022) with the ARIMA model (1,2,3)*

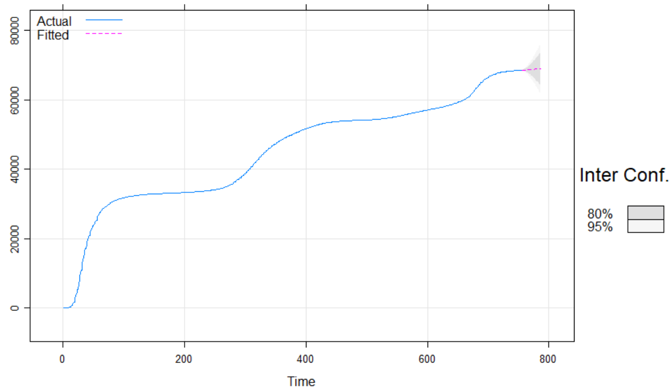

FIGURE 8 – The prediction of the evolution of the number of deceased people after 30 days (06 April-06 May 2022) with the ARIMA model (1,2,2)

4.1.2 Prediction with the SIRD model

SIRD model is one of the models to predict an outbreak using real data, In this part we will predict the outbreak in New York State after 1 month (April 06-May 06, 2022) , for this we consider the following initials the conditions :

$$S_0 = 14.812034 \times 10^6, I_0 = 0.527661 \times 10^6, R_0 = 4.57251 \times 10^6, D_0 = 0.068537 \times 10^6$$

We will estimate the parameters of this model using ARIMA in order to perform the prediction, the results of the SIRD model parameter estimates are shown in the following table :

Days	Transmission rate (k)	Cured rate (r)	Death rate (d)
1	0.002	0.641	0.0002
2	0.002	0.633	0.0002
3	0.002	0.624	0.0002
4	0.002	0.616	0.0002
5	0.002	0.609	0.0002
6	0.002	0.601	0.0002
7	0.002	0.593	0.0002
8	0.002	0.586	0.0002
9	0.002	0.579	0.0002
10	0.002	0.572	0.0002
11	0.002	0.565	0.0002
12	0.002	0.558	0.0002
13	0.002	0.551	0.0002
14	0.002	0.545	0.0002
15	0.002	0.538	0.0002
16	0.002	0.532	0.0002
17	0.002	0.526	0.0002
18	0.002	0.52	0.0002
19	0.002	0.514	0.0002
20	0.002	0.508	0.0002
21	0.002	0.503	0.0002
22	0.002	0.497	0.0002
23	0.002	0.492	0.0002
24	0.002	0.486	0.0002
25	0.002	0.481	0.0002
26	0.002	0.476	0.0002
27	0.002	0.471	0.0002
28	0.002	0.466	0.0002
29	0.002	0.461	0.0002
30	0.002	0.456	0.0002

TABLE 2 – **The variation in transmission rate (k), cure rate (r) and death rate (d) during the period April 06, 2022 - May 06, 2022**

The prediction results are shown in Figures 9, 10, and 11 :

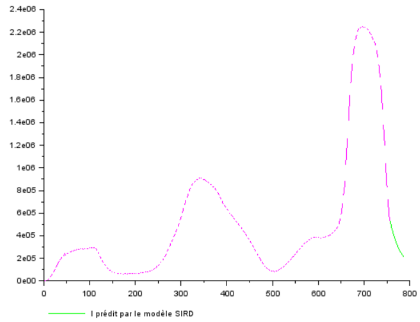

FIGURE 9 – *The prediction of the evolution of the number of infected people after 30 days (06 April - 06 May 2022)*

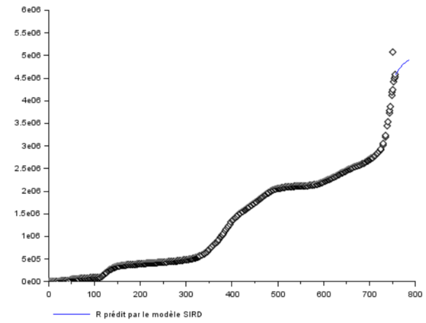

FIGURE 10 – The prediction of the evolution of the number of cured people after 30 days (06 April - 06 May 2022)

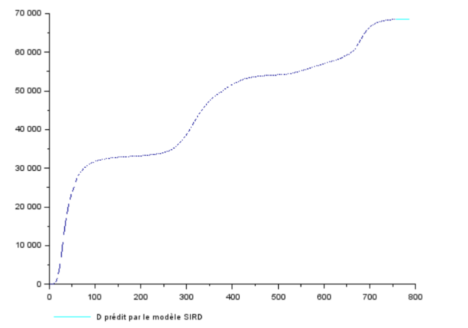

FIGURE 11 – The prediction of the evolution of the number of deceased people after 30 days (06 April - 06 May 2022)

4.1.3 Prediction with the polynomial regression model

Several polynomial degrees have been tested in our experiment on real data corresponding to the infected, recovered and deceased compartments of the state of New York. We applied the polynomial regression of degree 9 and 10 to the number of infected cases, of degree 5 and 7 to the number of cured cases and of degree 5 and 6 to the number of deceased cases. The prediction results after 30 days are shown in Figures 12, 13, 14 , 15, 16 and 17 :

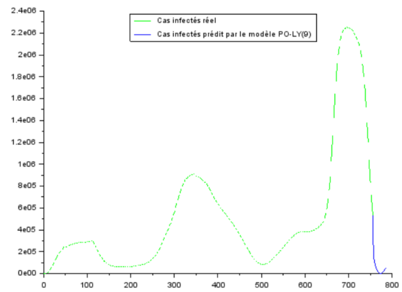

FIGURE 12 – *Prediction of the evolution of the number of infected people after 30 days (06 April - 06 May 2022) with the POLY(9) model*

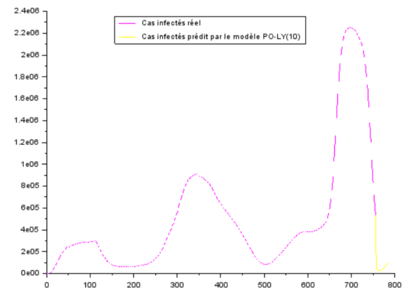

FIGURE 13 – *The prediction of the evolution of the number of infected people after 30 days (06 April - 06 May 2022) with the POLY(10) model*

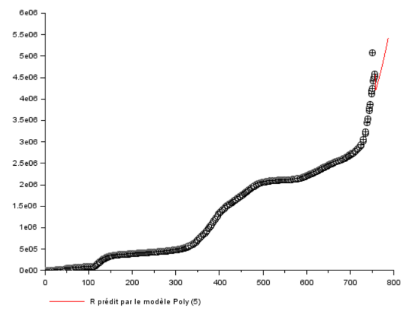

FIGURE 14 – The prediction of the evolution of the number of people cured after 30 days (06 April-06 May 2022) with the POLY(5) model

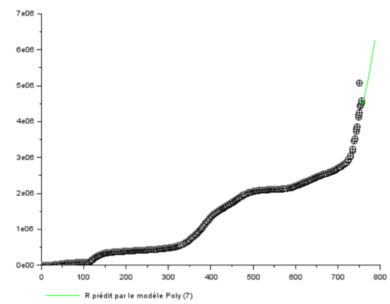

FIGURE 15 – The prediction of the evolution of the number of people cured after 30 days (06 April-06 May 2022) with the POLY(7) model

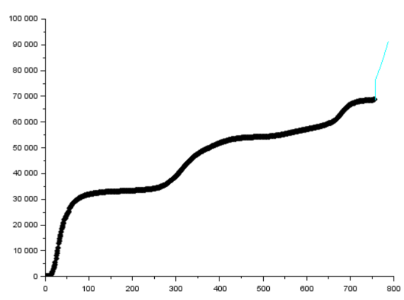

FIGURE 16 – The prediction of the evolution of the number of deceased people after 30 days (06 April-06 May 2022) with the POLY(5) model

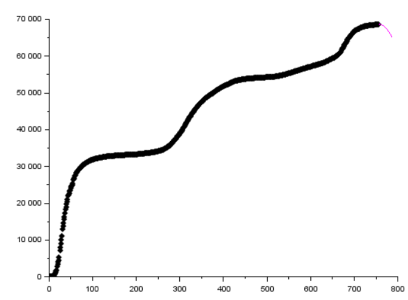

FIGURE 17 – The prediction of the evolution of the number of deceased people after 30 days (06 April-06 May 2022) with the POLY(6) model

5 Evaluate the performance of forecasting models

The results of the prediction of the evolution of the epidemic in the state of New York after 30 days with different models are presented in figures 18, 19 and 20 :

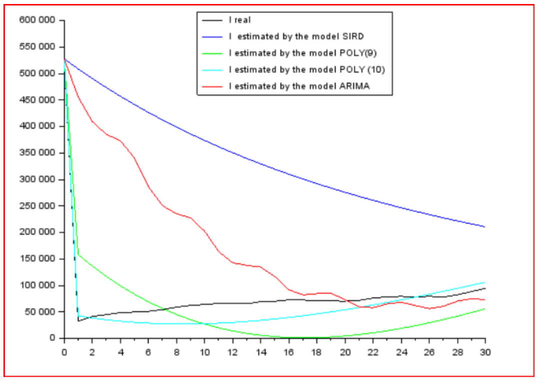

FIGURE 18 – The prediction of the evolution of the number of infected people after 30 days (06 April - 06 May 2022) with the ARIMA, SIRD and POLY models

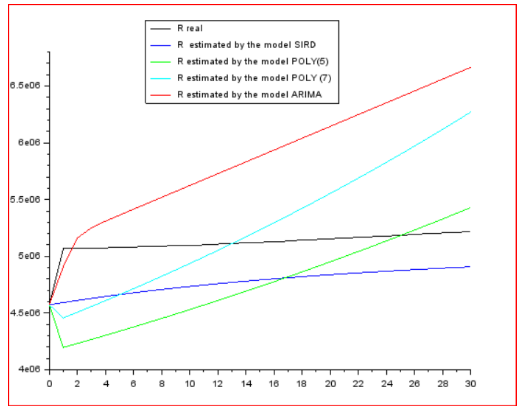

FIGURE 19 – The prediction of the evolution of the number of recovered people after 30 days (06 April-06 May 2022) the ARIMA, SIRD and POLY models

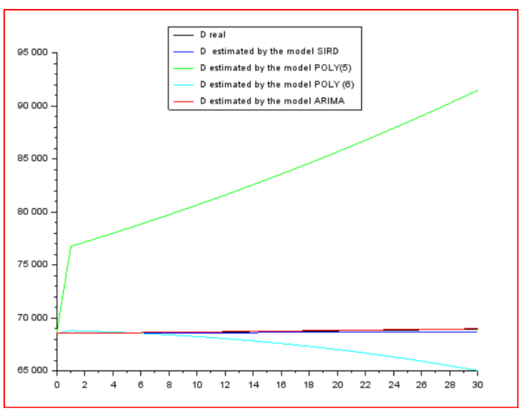

FIGURE 20 – *The prediction of the evolution of the number of deceased people after 30 days (06 April-06 May 2022) with the ARIMA, SIRD and POLY models*

Discussion

Visualization is not enough, we must therefore evaluate the performance of our models in order to verify the reliability of our prediction and to ensure our estimates. For this we will apply two error indicators in order to validate our prediction model for each compartment. The two commonly used error types are :

5.0.1 The mean absolute error MAE

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

With y_i is the value of the ith observation in our dataset, $\hat{y}_i$ is the predicted value of the ith observation and n corresponds to the observation number.

5.0.2 The root mean square error RMSE

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$

With y_i is the value of the ith observation in our dataset, $\hat{y}_i$ is the predicted value of the ith observation and n corresponds to the observation number.

These two indicators can judge the reliability of our model and give the best idea of the quality of the prediction, the principle is : "the closer the indicators are to 0, the better the prediction".

Tables 1, 2, and 3 show the model error flags applied to each compartment :

↪ For infected cases :

	ARIMA(6,1,9)	POLY (9)	POLY(10)	SIRD
MAE	107786.479	54152.0952	20108.2357	263321.634
RMSE	165508.816	59120.8745	23138.3271	282121.571

Table 3 – **Model errors for infected cases**

↪ For cured cases :

	ARIMA(1,2,3)	POLY (7)	POLY(5)	SIRD
MAE	777808.5333	437924.633	398055.867	352343.11
RMSE	881.93454	661.758743	630.916688	593.584964

Table 4 – **Model errors for cured cases**

↪ For deceased cases :

	ARIMA(1,2,2)	POLY (6)	POLY(5)	SIRD
MAE	29.03533333	1391.533	14868.6963	115.358833
RMSE	32.4251934	1840.38467	15471.0198	145.004407

Table 5 – **Model errors for deceased cases**

After analyzing the indicators of each model applied to each compartment, it is well confirmed that the model $POLY(10)$ is the most suitable as a model to predict future infected cases because it reaches the smallest average and the square errors (MAE and RMSE) compared to other models. Additionally, Table 4 shows that the model $SIRD$ had the best 30-day forecast. We therefore confirm that the model $SIRD$ is the most appropriate as a prediction model for future cured cases. on the other hand and according to table 5, the model $ARIMA(1,2,2)$ makes the fewest errors compared to the other models. therefore the latter proved to be the most suitable as a predictive model for future death cases.

Références

[1] *OFFICE PARLEMENTAIRE D'ÉVALUATION DES CHOIX SCIENTIFIQUES ET TECHNOLOGIQUES*, (page consultée le 12 mai 2022) [en ligne], adresse URL : https://www.senat.fr/fileadmin/Fichiers/Images/opecst/quatre_pages/OPECST_modelisation_covid_19.pdf

[2] *Images des mathématiques*, (page consultée le 12 mai 2022) [en ligne], adresse URL :http://images.math.cnrs.fr/Modelisation-d-une-epidemie-partie-1.html#nb2

[3] *Worldometer*, (page consultée 02 février 2022) [en ligne], adresse URL :https://www.worldometers.info/coronavirus/usa/new-york/

[4] *SIR et Simulation*, (page consultée 04 février 2022) [en ligne], adresse URL :http://mpechaud.fr/scripts/SIR/SIR.html?fbclid=IwAR1hWRZQEMuUUsS8-xa1jDpap4getI_C9iOoPyZul3otyoFavGqImbodDhY#simul

[5] *Pandémies de COVID-19 aux États-Unis*, (page consultée 18 février 2022) [en ligne], adresse URL :https://fr.wikipedia.org/wiki/Pandémie_de_Covid-19_aux_États-Unis?fbclid=IwAR1hWRZQEMuUUsS8-xa1jDpa

[6] *New-York prolonge le confinement d'un mois*, (page consultée 18 février 2022) [en ligne], adresse URL :https://www.huffingtonpost.fr/entry/new-york-coronavirus-confinement_fr_5e98aea2c5b6a92100e44f41?fbclid=IwAR0Mh6rj0Il58pbofgLaittv39wIGTu5P-HT0c--VrX7QMJcpLsMkecq2OY

[7] *France24 COVID19 aux États unis*, (page consultée 18 février 2022) [en-ligne], adresse URL :https://www.france24.com/fr/20200426-covid-19-aux-états-unis-possible-reprise-des-activités-dans-l-état-de-new-y

[8] *Didier Delignières, Analyse des séries temporelles, Septembre 2007*

[9] *Box, G.E.P., Jenkins, G.M. (1970). Time Series Analysis, Forecasting and Control., San Francisco, CA : Holden-Day.*

[10] *Cours de Monsieur Benjamin Ambrosio, the course Dynamical Systems Networks and Applications, given to students of M2 MASC, Le Havre Normandie University.This lecture is about SIR models and their networks extensions. The lecture addresses the modeling problem around an example of SIR coupled equations fitting COVID19 Data in New York and New Jersey states (page consultée 28 Mai 2022) [en-ligne], adresse URL :*`https://youtu.be/Z0k-6dmrzQA`

[11] *Exposé de Jacques Demongeot (25 février 2021) (page consultée 28 Mai 2022) [en-ligne], adresse URL :*`https://youtu.be/j_gpBLSzsrg`

[12] *Epidemiology SIR with Regression, Arima, and Prophet in Forecasting Covid-19 International Conference on Time Series and Forecasting, Gran Canaria, Spain, (19–21 July 2021)*

[13] *La ville de New York en confinement au moins jusqu'en juin (page consultée 24 Février 2022) [en-ligne], adresse URL :*`https://www.lapresse.ca/international/etats-unis/2020-05-11/la-ville-de-new-york-en-confinement-au-moins-jusqu-en-juin?fbclid=IwAR1hWRZQEMuUUsS8-xa1jDpap4getI_C9iOoPyZul3otyoFavGqImbodDhY`

[14] *Covid-19 : les dernières restrictions levées en Californie et dans l'Etat de New York (Le monde) (page consultée 24 Février 2022) [en-ligne], adresse URL :*`https://www.lemonde.fr/planete/article/2021/06/15/les-dernieres-restrictions-levees-dans-l-etat-de-new-york-ou-plus-de-70-des-adulte`
`6084267_3244.html`

[15] *États-Unis : la campagne de vaccination contre le Covid-19 commence lundi (France 24) (page consultée 25 Février 2022) [en-*

ligne], adresse URL :*https://www.france24.com/fr/amériques/20201212-états-unis-la-campagne-de-vaccination-contre-le-covid-19-commencera-lundi*

[16] *France 24 (page consultée 25 Février 2022) [en-ligne], adresse URL* :*https://www.france24.com/fr/amériques/20211207-covid-19-à-new-york-les-salariés-du-privé-devront-être-vaccinés-pour-alle*

[17] *Mesures sanitaires et épidémiologie de la COVID-19 de certains territoires (Public Health Ontario) (à jour au 2 février 2022) , (Publié le 22 février 2022)*