Al-Qadisiyah Journal of Pure Science
ISSN(Printed): 1997-2490 ISSN(Online): 2411-3514
DOI: /10.29350/jops.
http://qu.edu.iq/journalsc/index.php/JOPS

Cubic Spline Interpolation for Data Infections of COVID-19 Pandemic in Iraq

Authors Name

Jehan Mohammed Al-Ameri

Article History

Received on:23/8/2021
Revised on:15/10/2021
Accepted on: 8/11/2021

Keywords:

COVID-19, cubic spline interpolation, data fitting.

DOI:
https://doi.org/10.29350/jops.2021.26. 5.1443

ABSTRACT

In this paper, we use an empirical equation and cubic spline interpolation to fit Covid-19 data available for accumulated infections and deaths in Iraq. For Scientific visualization of data interpretation, it is useful to use interpolation methods for purposes fitting by data interpolation. The data used is from 3 January 2020 to 21 January 2021 in order to obtain graphs to analyse the rate of increasing the pandemic and then obtain predicted values for the data infections and deaths in that period of time. Stochastic fit to the data of daily infections and deaths of Covid-19 is also discussed and showed in figures. The results of the cubic splines and the empirical equation used will be numerically compared. The principle of least square errors will be used for both these interpolations. The numerical results will be indicated that the cubic spline gives an accurate fitting to data.

1. Introduction

At the early beginning of 2020, coronavirus (Covid-19) started to spread widely and large number of people were infected in the city of Wuhan in China. It is spreading rapidly in the other cities of China and around the world, even it was effectively controlled by the domestic outbreak to avoid increasing the infections rate outside Wuhan. Meanwhile, the World Health Organization (WHO) declared "global pandemic outbreak" in the end of March when Iraq reported over 630 cases of infections with 46 deaths since the start of the Covid-19 outbreak in the country on 3rd of January 2020 when Iraq becomes the second country with the highest death rate of Covid-19 in the WHO's Eastern Mediterranean Region. On the other hand, 152 patients had been recovered as reported from three laboratories in Iraq which became operational for Covid-19 testing in many cities of Basrah, Najaf, Arbil, and Baghdad Medical City in Baghdad.

In 2019, Covid-19 pandemic has become the biggest debated issue. Many Mathematics researches work on modeling the spread of Covid-19 pandemic locally and globally. The most interesting researches are for modeling and forecasting the number of infected and recovered cases comparing with the curve estimation models such as Box-Jenkins (ARIMA) and Brown-Holt linear exponential smoothing to epidemic cases in Canada, Japan, France, Italy, etc. [7]. Mathematical modeling and simulation to Covid-19 transmission moderation procedures for spreading Covid-19 in

[a] Mathematics Department, Basrah University, College of Science, Basrah, Iraq, E-Mail: jehan.khudhir@uobasrah.edu.iq

Canada and Italy is discussed [3,6]. Moreover, for predicting Covid-19 cases by, for example, fitting of the Logistic models using the least squares error [4], C++, Matlab, Java and Python codes were implemented which reflect useful results about the rapid spreading the virus. Khan and Atangana in 2020[12] presented the mathematical modeling and dynamics of spreading the pandemic using fractional model to the real data infections in Wuhan in January 2020. In 2021, the authors Appadu et.al. [2] considered the STEIR model of spreading Covid-19 in Lebanon and accounting the effect of travel of the virus.

It is scientifically confirmed later that the total number of confirmed infection cases of Covid-19 is growing exponentially around the world as mentioned by Dong et al. in [5]. There has been unlimited works in the literature which are using numerical methods for future prediction of spreading the pandemic and analysis the data of confirmed infections of Covid-19 (see: [13],[1]). The forecasted growth rates of infections in the range 0% -120% of the calculated growth rate are claimed briefly by Perc et al. in [13]. Lubbe and Botha in [11] proposed a 3-dimensional iterative model to the spread of Covid-19 in the world using fitting parameters technique to determine the total infection cases and number of confirmed daily cases.

Data fitting or as called data interpolation is a practical form for visualizing the data that obtained from numerical experiments or collected data of spreading epidemics. In 2021, Appadu, et.al. [2] has been discussed the results of using cubic spline interpolation, Euler method and Hybrid Euler method for the pandemic data interpretation in five countries, which concluded that cubic splines interpolation gives impressive and accurate results in the sense of using linear regression and non-linear regression of least squares error.

In this paper, we study the data of infections and deaths in Iraq during the outbreak of Covid-19. The data is picked from a published daily and accumulated number of coronavirus infections and deaths worldwide which predict future development of the virus https://www.worldometers.info/coronavirus/coronavirus-cases/#total-cases. We consider the accumulated infections and deaths in Iraq cover the period starting from 3 January 2020 until 21 January 2021.

The paper is organized as follows. Section 2 includes figures of cumulative data infections and deaths in Iraq, and relative error of daily cases. Curve fitting for the real data of total number of infections and deaths by using an imperial function will be considered in details. In Section 3, we will explain how to use cubic spline interpolation method for large amount of data. We will describe briefly how cubic spline method is used for fitting data.

2. Data Infections and Deaths: Spreading of Covid19

In this section we will study data infections and deaths in Iraq from starting the virus to rapidly spreading in Iraq. We consider the conventional approximation task to study a function $F(t)$ which may coincide in some sense with given measurements (data) at the corresponding locations (data sites). We can define $F(t) = F(t, \theta)$ where $\theta \in \mathbb{R}^k$ is a vector of unknown parameters. Normally an approximant is a continuous mapping from $\mathbb{R} \times \mathbb{R}^k$ to $\mathbb{R}$ with proper values for θ. If the region on which the data sites from a uniform or a regular grid then the process is called grid or mesh data approximation, otherwise, it is called scattered data approximation.

Let us consider the fitting curve to data in the following function:

$$F(t, \theta) = \theta_1 * \tanh\left(\frac{t - \theta_2}{\theta_3}\right) + \theta_4 \qquad (1)$$

To estimate the values of the unknown parameters $\theta_1, \theta_2, \theta_3, \theta_4$ we define a cost function and perform an optimization problem, y is the collected data of infection and death cases in Iraq from 3rd of January 2020 until 21st of January 2021.

We consider here the data points $(t_1, y_1), (t_2, y_2), \cdots, (t_n, y_n)$ and the model in Equation (1) depends on the parameters $\theta = (a_1, a_2, a_3, a_4)$. We assume that there exists $\hat{\theta}$ such that:

$$y(t_i) = F(t_i, \hat{\theta}) + \varepsilon_i$$

$$= \hat{\theta}_1 * \tanh\left(\frac{t_i - \hat{\theta}_2}{\hat{\theta}_3}\right) + \hat{\theta}_4 + \varepsilon_i\,, \quad i = 1,2,\cdots,n$$

where $\hat{\theta}$ represent the vector of the estimated values of θ and $\varepsilon_i\,, i = 1,2,\cdots,n$ are (measurement) errors on the data. Hence, we have to solve the cost function:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^4} \sum_{i=1}^{n} \left(y(t_i) - F(t_i, \theta)\right)^2 \qquad (2)$$

We start using the initial guess of parameters $\theta_0 = (10, 10, 0.5, 1)$. Stochastic fit was performed and repeated 100 times to reach best convergence. For this particular problem $n = 385$ and t_i formed an aquascape grid in [1, 385] with $t_{i+1} - t_i = 1$ for all $i = 1,2,\cdots,n-1$.

For the least squares fit the parameters are determined as the minimizer $\hat{\theta}$ of the sum of squared residuals. This is seen to be a problem of the form in Equation (2) with $k = 4$. The graph of $F(t_i, \hat{\theta})$ is shown by dot red in Fig. 1 for infections and deaths. The least squares problem is a special variant of the more general problem to find an argument of F that gives the minimum value of the objective function.

a b

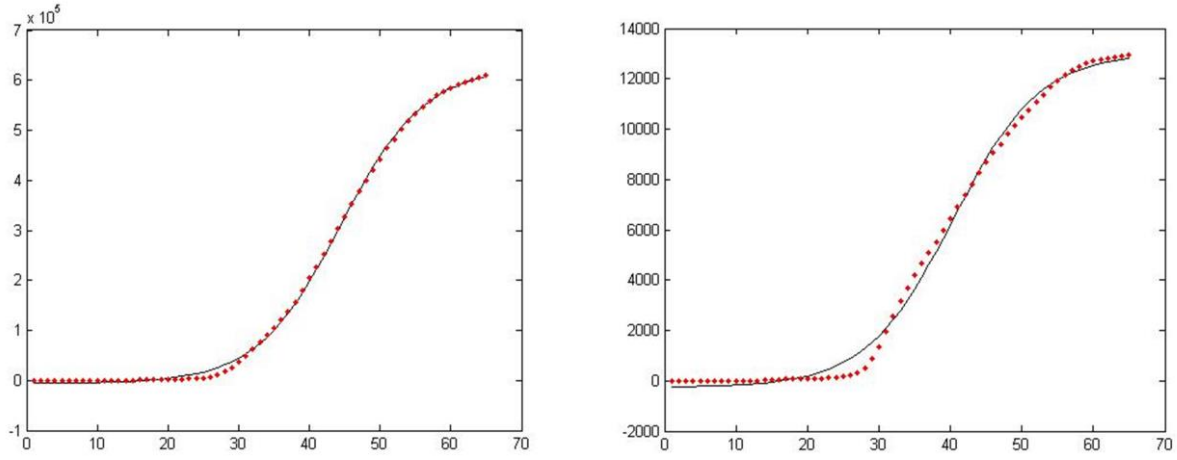

Fig.1- The data of discrete 65 infection and death cases of COVID-19 from January 3, 2020 to January 21, 2021. Stochastic fit is performed 1000 times. The figures of infected cases are shown in (a) and of death cases are shown in (b). The real data and fitting data are represented by deep black line and red dots, respectively.

As we notice in Fig. (1) that the relative error for both cases of infections and deaths is exponentially increasing in the middle of the period due to various prevention measures and there was no enough awareness about how the virus is spreading between people.

The least square error in Equation (2) was measured between $y(t)$ and $F(t, \hat{\theta})$ which are approximately 1.186472×10^{10} for infections and 2.759882×10^{10} for deaths. Moreover, we notice that trajectories of $y(t)$ and $F(t, \hat{\theta})$ are not nearly coincide with maximum relative error $e(t) = (y(t) - F(t, \hat{\theta}))/\|y\|_{\infty,[t_1, t_1+\infty]}$ for infection and death cases, approximately 0.0245 for infections and 1.002 for deaths as it is clearly shown in Fig. 2. The predicted error is shown in Table (1) and Table (2) for infections and deaths, respectively.

a b

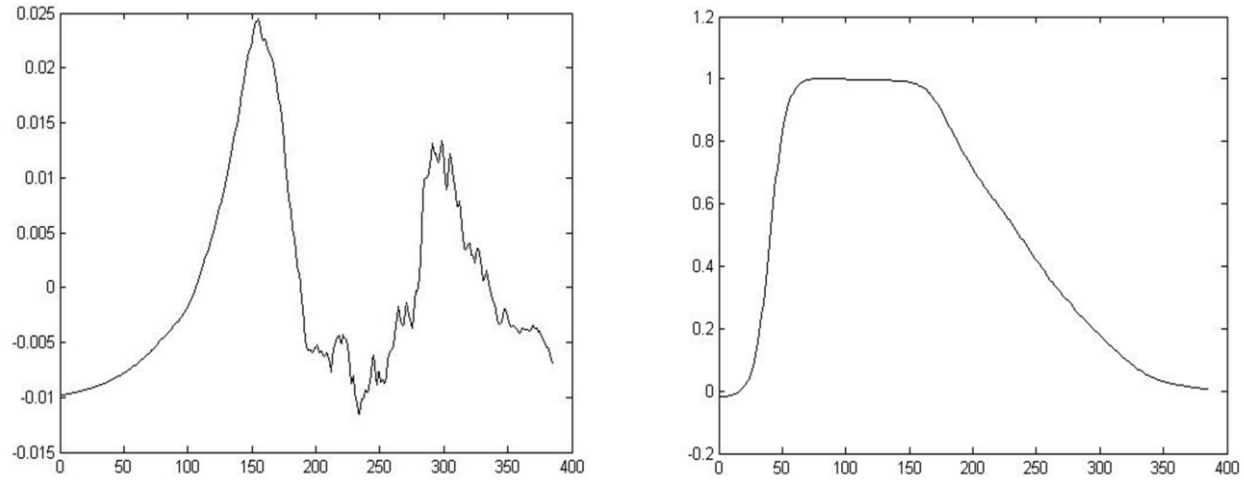

Fig. 2 - The values of relative error $e(t) = \frac{y(t)-F(t,\hat{\theta})}{\|y\|_{\infty,[t_1,t_1+\infty]}}$ as a function of t for the data of infections and deaths of Covid-19 from January 3, 2020 to January 21, 2021. The relative error curve of infected cases is shown in (a) and of death cases is shown in (b).

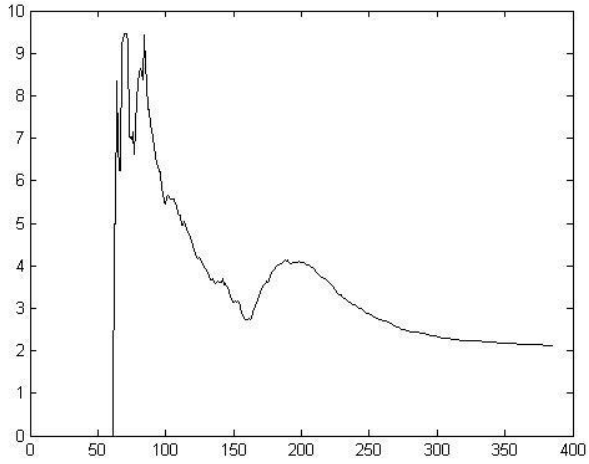

Fig. 3- Percentage ratio of cumulative number of deaths to cumulative number of infections.

In Fig.3, we calculate the percentage ratio of cumulative number of deaths to infections $(100 * \text{Deaths})/\text{Infections}$. It is noticed that infection cases are rapidly increased starting from the day 50 (21^{st} of February 2020) until the day 89 (31^{st} of May 2020) as shows the ratio in high level. This means the death cases are not high in that period of time, however, it is changed by increasing the death cases more than 10 cases every day. The daily infections can be found from the cumulated data and as it is shown in Fig. (4) the curve is not smooth but instead it takes many perturbations which is difficult in some sense to fit the data by using any empirical equation. This is the same for the data of daily death cases.

a b

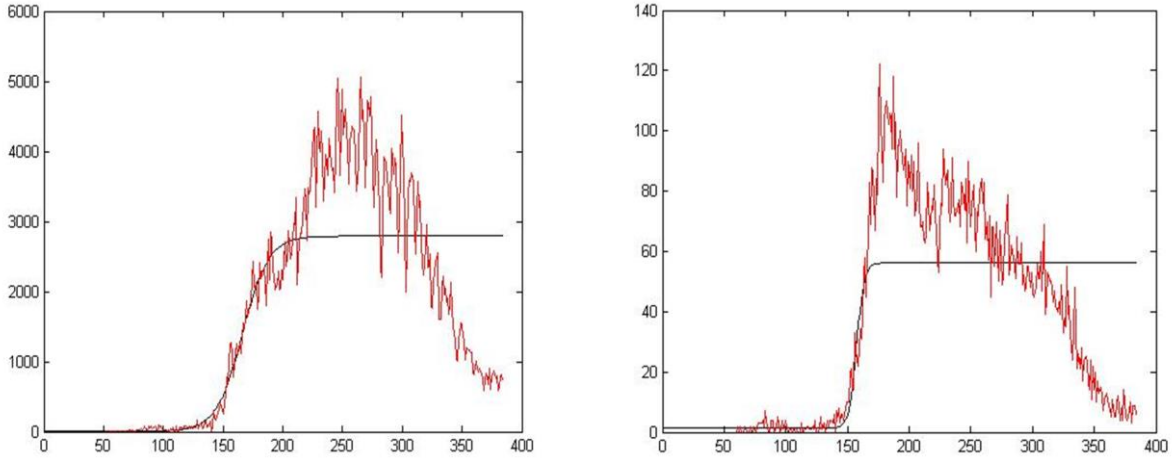

Fig. 4 - The data of daily infections and deaths of Covid-19 from January 3, 2020 to January 21, 2021. Stochastic fit is performed 1000 times. The figures of infected cases are shown in (a) and of death cases are shown in (b). The real data and fitting data are represented by red line and deep black line, respectively.

The data we use show an interesting behavior. The daily rate of infections continued to increase relatively slow until it began to reach high level of infections is 5055 in day 266 at 24/9/2020 and high-level deaths is about 122 death cases in day 176 at 27/6/2020 which may reflect the middle of high infections level in Iraq, However, it may be exceeded by the fast increasing at about the beginning of the period of spreading Covid-19 in Europe globally and China.

In order to fit and model such data empirically, one need to consider series of mathematical functions with different parameters, such as cubic spline interpolation. These functions must have some inherited properties.

The classical and most commonly used empirical functions offer convenient results to fit a broad range of data. They do, however, have shortcomings and limitations. These require a considerable number of iterations in order to estimate optimal unknown parameters which lead to approach better fit to data. This, however, may not always be optimal, it is not fitting properly with very small error, and, therefore, they do not provide approximated points at all.

A possible way to derive such approximations is to approximate the function using some "well-behaved" approximants. "Well-behaved" here is understood as a property that the derivatives of the approximating function could be controlled in some sense.

A conventional approximation task now is to construct a function $\mathrm{F}(\mathrm{t},\theta)$ which coincides in some sense with given measurements (data) at the corresponding locations (data sites). Normally an approximant is a continuous mapping from $\mathbb{R}\times\mathbb{R}^{\mathrm{k}}$ to $\mathbb{R}$. If the region on which the data sites from a uniform or a regular grid then the process is called grid or mesh data approximation, otherwise, it is called scattered data approximation. To better control behavior of the function between nodes one may relax the approximation requirement at all or some nodes, and allow for some errors there. We will discuss that by using cubic spline interpolation method.

3. Cubic spline interpolation method

We analyze the collected online available data of the cumulative number of infections and number of deaths in Iraq. The data was collected from 3rd of January 2020 until 21st of January 2021. Cubic spline interpolation is used to sort out the drawbacks of some methods such as Runge phenomenon [9]. This method provides an approximation polynomial, which is smoother and provides accurate results with smaller errors than some other interpolating polynomials such as Lagrange interpolation. Moreover, cubic spline interpolation is more suitable for solving smooth functions which do not have oscillating behavior [10], in addition to solve problems with high perturbations. It obtains a system of linear equations could be described by a system with a tridiagonal matrix.

The cubic spline interpolation is defined in the interval $[t_{i-1}, t_i]$; $i = 1,2,\dots,385$ and is shown in the following equation:

$$S_i(t) = a_i + b_i t + c_i t^2 + d_i t^3; \quad i = 1,2,\cdots,n \qquad (3)$$

where a_i, b_i, c_i and d_i ; $i = 1,2,\cdots,n$ are the $4n$ coefficients of the spline equations.

The function $S(t)$ in (3) satisfies the following conditions:

$$S_i(t_i) = y(t_i) \qquad (4)$$

$$S_i(t_{i+1}) = S_{i+1}(t_{i+1}) = y(t_{i+1}) \qquad (5)$$

$$S_i'(t_{i+1}) = S_{i+1}'(t_{i+1}) \qquad (6)$$

$$S_i''(t_{i+1}) = S_{i+1}''(t_{i+1}) \qquad (7)$$

Thus, we have $N = n - 1 = 384$ piecewise cubic spline equation that satisfies condition (4) until (7). We can analyse the cubic spline equations as follows:

- The first equations in (4) and (5) require that the spline function passing through the first and last point t_i of the interval for all $i = 1, 2, \cdots, 385$, yielding $2n - 2 = 768$ equations.
- Equation (6) requires that the first derivative is continuous at each interior point in the interval, yielding $n - 2 = 383$ equations.
- The last equation in (7) requires that the second derivative is continuous at each interior point yields $n - 2 = 383$ equations.

Then we get $4n - 6 = 1534$ equations which include $4(n - 1) = 1536$ unknown coefficients required to be determined and estimated. This, however, needs at least two more conditions to possibly have a unique solution or estimated values for the coefficients. Then, we can consider the boundary conditions:

$$S_0''(x_0) = S_n''(x_n) = 0 \qquad (8)$$

By this system of 1536 equations and 1536 unknown parameters, we get tri-diagonal matrix form of linear equations. Due to that big number of equations and space limitations, the spline equations are not displayed here. However, predicted values will be displayed in Table (1) and (2) for 21 days of infections and deaths data. The system of spline equations is in the following form:

$$A\,C = B \qquad (9)$$

where C is the vector which includes the unknown parameters $\{a_{ij}, b_{ij}, c_{ij}, d_{ij}; i = 1,2,\dots,354;\ j = 1,2,\dots,354\}$; A is the matrix of coefficients of the parameters and B is the right-hand side (constants) of the spline equations. The system of linear equations (9) with a square matrix can be solved with the backslash technique. The backslash can also be used as well when the matrix has more rows than

columns, however, it is, in general, not possible to satisfy all the equations. Therefore, the system is solved in the sense of minimizing the sum of the squares of errors (residuals). By using Matlab, we can find the coefficients of the spline that fits the data as well as possible with $C = A\backslash B$. Then the linear system (9) can be solved by the Gauss elimination method or the Lower-Upper (LU) decomposition method.

The approximation problem can be cast as a constrained optimization problem by using the principle of least square errors:

$$\min \sum_{i=2}^{n-1} \left(y(t_i) - S_i(t_i)\right)^2$$

Instead of dealing with continuous-time signals, y(t), one may re-formulate the above results in the setting in which model responses and data are compared at $n-2$ discrete points $\{t_i\}$ in $[t_2, t_{n-1}]$. In this case we need to find fitted values for the first and the last points of the data, i.e., the first and last functions $S_1(t)$ and $S_{385}(t)$ must pass through the first and last end points $(t_1, S_1(t_1))$ and $(t_{385}, S(t_{385}))$ in the form:

$$S_0(t_0) = y(t_0) \quad , \quad S_n(t_n) = y(t_n) \qquad (10)$$

where the coefficients of equation in (10) are considered in before in the boundary conditions in (8). This yields 2 more conditions for the system of spline equations, then the objective function of the constrained optimization problem will be as follows:

$$\min \sum_{i=1}^{n} \left(y(t_i) - S_i(t_i)\right)^2 \qquad (11)$$

and the constraints are 385 spline equations. For parameter estimation routine we used the Nelder-Mead algorithm [8]. The values of reflection, expansion, and contraction coefficients in the algorithm were set to 1, 2, and 0.5, respectively.

Day	Recorded value	Predicted value by Cubic Spline	Predicted value by Equation (1)	Prediction error (ε)
20/7/2020	92530	92530	89176.95	3353.049
21/7/2020	94693	94693	91484.71	3208.285
22/7/2020	97159	97159	93837.78	3321.217
23/7/2020	99865	99865	96236.54	3628.462
24/7/2020	102226	102226	98681.34	3544.655
25/7/2020	104711	104711	101172.5	3538.455
26/7/2020	107573	107573	103710.5	3862.547
27/7/2020	110032	110032	106295.4	3736.640
28/7/2020	112585	112585	108927.5	3657.471
29/7/2020	115332	115332	111607.2	3724.806
30/7/2020	118300	118300	114334.6	3965.439
31/7/2020	121263	121263	117109.8	4153.196
1/8/2020	124609	124609	119933.1	4675.938
2/8/2020	126704	126704	122804.4	3899.555
3/8/2020	129151	129151	125724	3426.977
4/8/2020	131886	131886	128691.8	3194.167
5/8/2020	134722	134722	131707.9	3014.128
6/8/2020	137556	137556	134772.1	2783.900
7/8/2020	140603	140603	137884.4	2718.565
8/8/2020	144064	144064	141044.8	3019.244
9/8/2020	147389	147389	144252.9	3136.104

Table 1 - The commutative number of infections in Iraq using the cubic spline interpolation method for only 21 days.

In Table 1 and Table 2, the real data and the results of cubic spline interpolation $(S(t))$ and the empirical function $F(t)$ explained in section 2. The error prediction (absolute error ε) by using $F(t)$ is computed to show the accuracy of both interpolations.

Day	Recorded value	Predicted value by Cubic Spline	Predicted value by Equation (1)	Prediction error (ε)
1/6/2020	205	205	878.769	673.769
2/6/2020	215	215	907.849	692.849
3/6/2020	235	235	937.595	702.595
4/6/2020	256	256	968.021	712.020
5/6/2020	271	271	999.136	728.135
6/6/2020	285	285	1030.952	745.952
7/6/2020	318	318	1063.482	745.482
8/6/2020	346	346	1096.736	750.736
9/6/2020	370	370	1130.726	760.726
10/6/2020	392	392	1165.464	773.464
11/6/2020	426	426	1200.960	774.960
12/6/2020	457	457	1237.226	780.226
13/6/2020	496	496	1274.274	778.273
14/6/2020	549	549	1312.113	763.113
15/6/2020	607	607	1350.756	743.756
16/6/2020	652	652	1390.212	738.212
17/6/2020	712	712	1430.493	718.492
18/6/2020	773	773	1471.608	698.608
19/6/2020	856	856	1513.568	657.568
20/6/2020	925	925	1556.382	631.383
21/6/2020	1013	1013	1600.062	587.061

Table 2 - The commutative number of deaths in Iraq using the cubic spline interpolation method for only 21 days.

We see in Fig. (1) that the infections rate is exponentially decreasing in the middle of disease spread due to various reasons. It shows that the number of total infection cases is still increasing and reaches the high level in about 266 days after January 3, 2020, that is, September 25, 2020. This rate of infections is slightly decrease and leads to decrease the deaths about the day 360 as it is clear no large changing in the curves of figures. This indicates that some susceptible people were well protected by medicines or taking vaccine, and there is a small chance of transmission by infected people. This may can make predictions about the future trend of the disease spread. Figure 4(a) and (b) shows the dynamics of the number of the daily infection cases and death cases in the 385 days after January 3, 2020. The number of daily new cases has begun to decrease slightly compared with the middle of the period. However, the data show that the maximum value of the new infections rate is changing and approximately the same over the days even there is perturbations in the deaths. The infections rate declines to 0.15 on 27/12/2020. It also indicates that the number of newly confirmed cases will continue to decline in the begging of 2021. The number of newly confirmed cases will decrease to reach 0 infections and 0 deaths in about after February 2021, which is not considered in this paper.

Conclusion

This paper discussed the use of Equation (1) as an interpolating equation for fitting curve to Covid-19 data for infections and deaths in Iraq from the period 3rd of January 2020 till 21st January 2021. Cubic spline interpolation is considered in details for that particular data of infections and deaths. The numerical results showed that there is a big difference in the accuracy of the interpolations. Cubic spline interpolation gives much more smooth interpolating curves compare to the interpolating curves by using Equation (1). The interpolating curves by using Equation (1) tend to simulate the data on certain given interval and takes the same outline curve for any estimated values of coefficients to fit the real data of both infections and deaths which does not preserve the shape of the given data. As a result, cubic spline works well for all data points. Error analyses by using Equation (1) and cubic spline for data interpolation also have been showed in details In Table (1) and (2). Other studies for fitting data of recovered cases can take the same scenario of using cubic spline interpolation.

References

[1] A.E. Botha, W. Dednam, A simple iterative map forecast of the covid-19 pandemic, ArXiv preprint arXiv:2003.10532. 1–6 (2020).

[2] Appadu, A.R., Kelil, A.S. and Tijani, Y.O., 2021. Comparison of some forecasting methods for COVID-19. Alexandria Engineering Journal, 60(1), pp.1565-1589.

[3] A.R. Tuite, D.N. Fisman, A.L. Greer, Mathematical modelingof covid-19 transmission and mitigation strategies in the population of Ontario, Canada, Canadian Med. Assoc. J. 192(19) (2020) E497–E505.

[4] A. Xavier, A C++-code for predicting COVID-19 cases by least-squares fitting of the Logistic model,
Preprint (10) (2020) 1–23.

[5] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. 20 (2020) 533–534.

[6] G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. Di-Fillipo, A. Di Matteo, M. Colaneri, Modeling
covid-19 epidemic and implementation of population-wide interventions in Italy, Nat. Med. 26 (2020)
855–869.

[7] H. Yonar, A. Yonar, M.A. Tekindal, Modeling and forecasting for the number of cases of the covid-19 pandemic with the curve estimation models, the Box-jenkins and exponential smoothing methods, Eurasian J. Med. Oncol. 4 (2) (2020) 160–165.

[8] J.A. Nelder, R. Mead. A simplex method for function minimization. Comp. J., 7 (1965), 308–313.

[9] J. De Villiers, Mathematics of Approximation, Volume 1, Springer, 2012.

[10] K.E. Atkinson, An Introduction to Numerical Analysis, John Wiley & Sons, 2008.

[11] Lubbe W, Botha E, Niela-Vilen H, Reimers P. Breastfeeding during the COVID-19 pandemic–a literature review for clinical practice. International breastfeeding journal. 2020 Dec;15(1):1-9.

[12] M.A. Khan, A. Atangana, Modeling the dynamics of novel coronavirus (2019-ncov) with fractional
derivative, Alexandria Eng. J. 1–11 (2020).
[13] M. Perc, N.G. Miksic, M. Slavinec, A. Stozer, Forecasting COVID-19, Front. Phys. (2020)
8–127.