Polynomial Interpolation and Cubic Spline to Determine Approximate Function of COVID-19 Tweet Dataset

Devi Munandar, Wahyu Suryaningrat, and Sri Purwani

***Abstract*—As we know theoretically if we are going to construct a polynomial interpolation function through a mapped base, we create an approximation function. In this study, we try to build an approximation function using all sample data available. The approximation function obtained represents the data whose graph goes through a given set of data points. We determine the value of a function at different points and specific intervals using the interpolation model. The first derivative of the function is obtained to find the growth rate of tweet data. The experimental data is a crawling tweet with the keyword COVID-19. Then we get the amount of data per time duration representing a value of the function at a node. The interpolation includes such as Lagrange, Newton's divided difference, and cubic spline. In this study, we compared polynomial interpolation with cubic splines to obtain optimal results. With the functional approach obtained, a pattern of tweets related to COVID-19 can be seen from its graph that passes through the given data points. The graph and the estimated values obtained show that the cubic spline is the optimal interpolation as an approximation function.**

***Index Terms*—Tweet COVID-19, Interpolation, Lagrange, Newton's divided difference, Spline, Approximation**

I. Introduction

SINCE the declaration of the Corona Virus Disease in 2019, known as COVID-19, by the world health organization (WHO) to become a pandemic in March 2020, social media has become an inseparable part of disseminating information, especially on Twitter [1]. Twitter is a social media used by many people as a medium for exchanging opinions, discussions, and research. The number of post tweets and retweets shows many information dissemination discussions [2][3]. A study to analyze the relationship between social media users vulnerability and the policy of cyber attacks between countries [4]. With the COVID-19 pandemic, this vulnerability can be used as a unifier for exchanging information in each country for Online Social Media users who can discuss between countries through an analysis considering data from two different countries.

Various application implementations can find procedures quickly and simply organize tweet data in millions of units to find important items [5]. To facilitate the data collection process, Twitter has also provided Application Programming Interfaces (APIs). With APIs, the user is granted access to Twitter data, such as tweets and profile information. This comprehensive data source can be used for several studies and research questions in different applications, such as health, culture, marketing strategy, and politics [6][7][8][9].

In mathematics, approximation theory is concerned with how functions or data can best come close to simpler functions and quantitatively describe the errors introduced thereby [10]. Data fluctuation in the number of tweets obtained at a particular time duration can vary greatly. The amount of information with text data format as an example tweet can also be approximated to represent the specific time duration pattern.

Polynomial interpolation is a commonly used approach in determining approximation functions based on mapping the data sample points [11]. Using polynomials interpolation is a simple evaluation than a non-polynomial approximation. Because of their rigidity (due to smoothness), the polynomial interpolation tends to over-fit the data for some cases. Therefore, other spline Interpolation methods have simple terms and many distinct polynomials; the resulting function is continuous, and its derivative is also continuous [12]. In other applications, linear interpolation is used to obtain missing data and detect anomalies in the SEIRS Epidemic Model with a point-to-point homoclinic linking the orbit to a saddle equilibrium. The endemic is located numerically as a bifurcation concerning the operational parameter. The use of polynomials involves a significant role in this modeling, especially the Lagrange polynomial [13] combined with the orthogonal collocation algorithm, which shows the role of the Lagrange polynomial in providing solutions to solving problems in epidemic modeling [14][15].

This study is the most important part of seeing the growth of tweets that discuss the COVID-19 phenomenon

Manuscript received December 29, 2020; revised September 17, 2021.

Devi Munandar is a Master candidate at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Padjadjaran University, Jl. Raya Bandung-Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia, and a researcher in the field of Data Science at Research Center for Informatics Department, National Research and Innovation Agency, Jl. Cisitu No.21/154D Komplek LIPI Gedung 20 Lantai 3, Bandung 40135, Indonesia. (e-mail:devi19010@mail.unpad.ac.id).

Wahyu Suryaningrat is a Master candidate at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Padjadjaran University, Jl. Raya Bandung-Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia. (e-mail:wahyu19005@mail.unpad.ac.id).

Sri Purwani is a Lecturer at the Department of Mathematics, Faculty of Mathematics and Natural Sciences, Padjadjaran University, Jl. Raya Bandung-Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia. (e-mail: sri.purwani@unpad.ac.id).

based on social media posts and obtaining an approximation of the function for the time interval the tweet is posted. It also predicts the number of tweets discussed at certain intervals, which is helpful to see how the pandemic fluctuates.

A. Literature Statistical Analysis

This session presents a bibliographical analysis of published papers on polynomial interpolation using the Lagrange, cubic spline, and Newton's divided difference methods. Data obtained from the Scopus database from 2011 to 2021 correlates with the keywords are polynomial interpolation, Lagrange, cubic spline, Newton's divided difference in the title, keywords, and abstracts of the paper are 440 documents. The results were only taken from articles and conference papers into 374 documents. Since we only analyzed 2011-2021, the data obtained was 191 documents (article 134, conference paper 57).

TABLE 1
TOP PAPERS PER CITATIONS ON POLYNOMIAL INTERPOLATION

Year	Reference	Citations
2014	[16]	78
2011	[17]	53
2011	[18]	49
2018	[19]	46
2018	[20]	43
2012	[21]	37
2011	[22]	37
2013	[23]	33
2013	[24]	32
2017	[25]	31

The results of Table 1 represent the top 10 citations generated. The total citations obtained are numbered 1314 cited to published papers in the polynomial interpolation using Lagrange, Newton's divided difference, and cubic spline research. Foundations of Computational Mathematics are sources cited in 78 citations from Springer New York LLC publisher. It can be seen that polynomial interpolation research mostly refers to computational and mathematics or numerical integration research papers, while the IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics as a source, ranks next at 53 citations. This source communicates and controls topics across humans, machines, and organizations at the structural or neural level. The two sources contributed 5.94% and 4.03% citations of total citations to the paper in the source, respectively. The number of citations obtained in this paper, so this research is still open to the topic of polynomial interpolation.

Conceptual Structure

A mapping study for the scientific discipline or research area topics can use a factorial approach as the main research area conceptual structure to determine a research question. This concept reduces dimensions of data and representing in low-dimensionally space. There are several alternative methodologies used to solve this case. In this discussion, we use Multiple Correspondence Analysis (MCA) [26] is a data analysis technique for nominal categorical data, to detect and represent underlying structures in a data set. It is summarizing and visualizing a data table containing more than two categorical variables.

In Fig. 1, each color represents a word cluster that refers to relevant keywords in polynomial interpolation. Each cluster is identified by hierarchical clustering. In the first cluster (red), keywords such as interpolation method, Lagrange interpolation, cubic spline, and numerical results are the most used. This cluster can be categorized as an analytical interpolation group. Whereas in the second cluster (blue) of keywords such as finite difference method, computational efficiency, numerical method can be categorized as a numerical computation group. In the third cluster (green), keywords such as fundamental theorems, fractional calculus, and fractional derivatives can be categorized as fractional theorems group. In the fourth

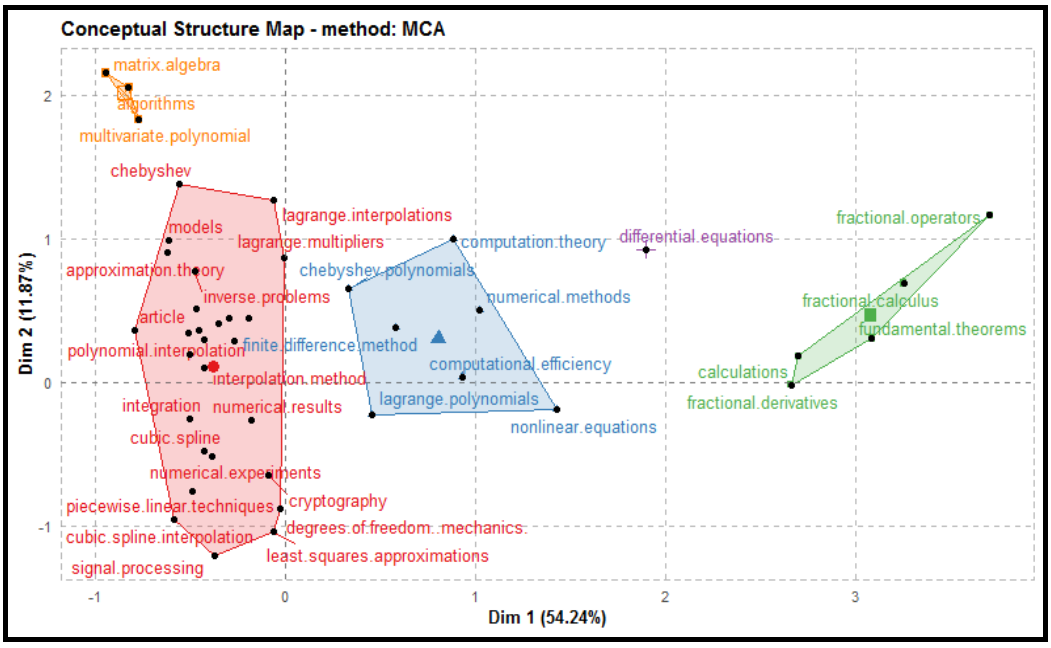

Fig. 1. Conceptual structure map with MCA method

cluster (orange) with the keyword algorithms, mix algebra, multivariate polynomial was classified as an integrated polynomial group. Based on the categories obtained, the study we conducted is an analytic interpolation group. It is based on modeling using Lagrange, Newton's divided difference, and cubic spline.

II. Materials and Methods

A. Materials

This study used a dataset from crawling twitter data using netizens keywords and then filtered the keywords. The crawling process was performed using the Application Programming Interface (API) obtained from the Twitter developer. These tweet data were used to construct an interpolation function that can be used to obtain data at even hours. The process was carried out for three consecutive days, from November 4-6, 2020 (see Appendix).

B. Methods

Lagrange Interpolation

Having discrete data points (x_i, y_i) where $i = 1, 2, ..., n$, we can construct a Lagrange polynomial interpolation defined as follows,

$$P_n(x) = \sum_{i=0}^{n} L_i(x)(y_i) \tag{1}$$

where is degree of polynomial approximation that interpolates (x_i, y_i) at distinct points, $L_i(x)$ is a Lagrange function defined as follows [27].

$$L_i(x) = \prod_{\substack{m=0 \\ m \neq i}}^{n} \frac{x - x_m}{x_i - x_m} \tag{2}$$

$$L_i(x) = \frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \tag{3}$$

where $i = 0, 1, 2, 3, \ldots n$. $L_i(x)$ has a degree equal to the degree of $P_n(x)$ and satisfies the following equations

$$L_i(x_j) = \delta_{ij}, 0 \le j \le n$$
$$\delta_{ij} = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases} \tag{4}$$

Equation (1) can be rewrite as follows,

$$P_n(x) = y_0 L_0(x) + y_1 L_1(x) + \ldots + y_n L_n(x) \tag{5}$$

Then the Lagrange polynomial interpolation is used to estimate data points (x_i, y_i), where $i = 1, 2, ..., n$. Having simple evaluation, polynomials are widely used on both numerical differentiation and integration [28].

Newton's divided difference Interpolation

Newton's divided difference is formulated as the same form as Lagrange, but the advantage is that it is more practical and efficient in computation and can be obtained P_n and P_{n+1}

With 2 data points, such as (x_0, y_0) and (x_1, y_1), we can construct a polynomial of degree one, $P_1(x)$, that interpolates the two points given as follows

$P_1(x) = \xi_0 + \xi_1(x - x_0)$, where

$\xi_0 = y_0$, $\xi_1 = \frac{y_1 - y_0}{x_1 - x_2}$ and at $x = x_1$, then

$f_1(x_1) = f(x_1) = \xi_0 + \xi_1(x_1 - x_0) = f(x_0) + \xi_1(x_1 - x_0)$

This gives a linear interpolation following

$$P_1(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) \tag{6}$$

Similarly, for degree two polynomial, we have

$P_2(x_2) = f(x_2) = \xi_0 + \xi_1(x_2 - x_0) + \xi_2(x_2 - x_0)(x_2 - x_1)$

$f(x_2) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x_2 - x_0) + \xi_2(x_2 - x_0)(x_2 - x_1)$

giving,

$$\xi_2 = \frac{\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_0} \tag{7}$$

Then the quadratic interpolation of Newton's divided difference is as follows [29]

$$\begin{aligned} P_2(x) &= \xi_0 + \xi_1(x - x_0) + \xi_2(x - x_0)(x - x_1) \\ &= y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + \\ &\quad \frac{\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_0}(x - x_0)(x - x_1) \end{aligned} \tag{8}$$

Hence, in general, we have $\xi_0 = f[x_0]$, $\xi_1 = f[x_1, x_0]$, $\xi_2 = f[x_2, x_1, x_0]$, ..., $\xi_{n-1} = f[x_{n-1}, x_{n-2}, ..., x_0]$, $\xi_n = f[x_n, x_{n-1}, ..., x_0]$

For the p^{th} divided differences, we have

$$\begin{aligned} \xi_p &= f[x_p, ..., x_0] \\ &= \frac{f[x_p, ..., x_1] - f[x_{p-1}, ..., x_0]}{x_p - x_0} \end{aligned}$$

Hence, Newton's divided difference interpolation that interpolates at the point (x_0, y_0), (x_1, y_1), (x_2, y_2), and (x_3, y_3) is as follows [30].

$$\begin{aligned} P_3(x) = f[x_0] &+ f[x_1, x_0](x - x_0) + f[x_2, x_1, x_0](x - x_0)(x - x_1) \\ &+ f[x_3, x_2, x_1, x_0](x - x_0)(x - x_1)(x - x_2) \end{aligned} \tag{9}$$

Spline Interpolation

Spline is another class of interpolating function widely used for interpolation. We use cubic spline interpolation function [31][32][33]. Apart from being the default boundary function, the cubic spline interpolation can also be used as a rational cubic / quadratic spline in its development. The process of boolean addition of cubic interpolation operators to combine a cubic / quadratic rational interpolation split into some boundary function [34].

Spline interpolation is preferable to use compared to high-order polynomial interpolation, where the

n^{th} derivative is continuous in a data point. Having data points such as

$$(x_0, y_0), (x_1, y_1), ..., (x_{n-1}, y_{n-1}), (x_n, y_n)$$

where $x_0 < x_1 < x_2 < ... < x_n$, we can construct cubic spline $\zeta(x)$ that satisfies the following conditions [35][36]:

- $\zeta(x)$ interpolates all data points, giving
$$\zeta_i(x_i) = \zeta(x_i) = y_i \quad (10)$$
where $\zeta_i(x)$ is defined on $[x_i, x_{i+1}]$
- $\zeta_i(x)$ continuous on $[x_0, x_n]$ where
$$\zeta_i(x_{i+1}) = \zeta_{i+1}(x_{i+1}) \quad (11)$$
- $\frac{d}{dx}\zeta(x)$ continuous on $[x_0, x_n]$ where
$$\frac{d}{dx}\zeta_i(x_{i+1}) = \frac{d}{dx}\zeta_{i+1}(x_{i+1}) \quad (12)$$
- $\frac{d^2}{dx^2}\zeta(x)$ continuous on $[x_0, x_n]$ where
$$\frac{d^2}{dx^2}\zeta_i(x_{i+1}) = \frac{d^2}{dx^2}\zeta_{i+1}(x_{i+1}) \quad (13)$$
- *Natural cubic spline interpolation* (*boundary condition*)
$$\frac{d^2}{dx^2}\zeta_0(x_0) = \frac{d^2}{dx^2}\zeta_{n-1}(x_n) = 0 \quad (14)$$

$$\zeta(x) = \begin{cases} a_0 + b_0(x-x_0) + c_0(x-x_0)^2 + d_0(x-x_0)^3, x_0 \le x < x_1 \\ a_1 + b_1(x-x_1) + c_1(x-x_1)^2 + d_1(x-x_i)^3, x_1 \le x < x_2 \\ ..... \\ a_i + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3, x_{n-1} \le x < x_n \\ a_{n-1} + b_{n-1}(x-x_{n-1}) + c_{n-1}(x-x_{n-1})^2 + d_{n-1}(x-x_{n-1})^3, x_{n-1} \le x \le x_n \end{cases} \quad (15)$$

To determine the coefficients of the equation system, we use the previous conditions. Using equations (10) and (11), we have the following equations,

$$\begin{aligned} &\zeta_0(x_0) = \alpha_0, \zeta_0(x_1) = \alpha_1 \\ &\zeta_1(x_1) = \alpha_1, \zeta_1(x_2) = \alpha_2 \\ &.... \\ &\zeta_{n-1}(x_{n-1}) = \alpha_{n-1}, \zeta_{n-1}(x_n) = \alpha_n \end{aligned} \quad (16)$$

Substituting results from equations (16) into equations (15) gives the following system of equations.

$$\begin{aligned} &a_0 = \alpha_0 \\ &a_0 + b_0(x_1 - x_0) + c_0(x_1 - x_0)^2 + d_0(x_1 - x_0)^3 = \alpha_1 \\ &a_1 = \alpha_1 \\ &a_1 + b_1(x_2 - x_1) + c_1(x_2 - x_1)^2 + d_1(x_2 - x_1)^3 = \alpha_2 \\ &a_2 = \alpha_2 \\ &a_2 + b_2(x_3 - x_2) + c_2(x_3 - x_2)^2 + d_2(x_3 - x_2)^3 = \alpha_3 \\ &\vdots \qquad \vdots \qquad \vdots \end{aligned} \quad (17)$$

or

$$\begin{aligned} &a_{n-1} = \alpha_{n-1} \\ &a_{n-1} + b_{n-1}(x_n - x_{n-1}) + c_{n-1}(x_n - x_{n-1})^2 + d_{n-1}(x_n - x_{n-1})^3 = \alpha_n \end{aligned}$$

The continuity of the polynomial first derivatives (12) and the second derivatives (13) can then determine the coefficients. The first and second derivatives of the cubic spline are respectively given as follows,

$$\frac{d}{dx}\zeta_i(x) = b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2 \quad (18)$$

and their second derivatives

$$\frac{d^2}{dx^2}\zeta_i(x) = 2c_i + 6d_i(x - x_i) \quad (19)$$

where $i = 0, 1, 2, \ldots, n-1$

Applying equations (12) into equation (18) results in the following equations.

$$\begin{aligned} &b_0 + 2c_0(x_1 - x_0) + 3d_0(x_1 - x_0)^2 = b_1 \\ &b_1 + 2c_1(x_2 - x_1) + 3d_1(x_2 - x_1)^2 = b_2 \\ &..... \\ &b_i + 2c_i(x_{i+1} - x_i) + 3d_i(x_{i+1} - x_i)^2 = b_{i+1} \\ &b_{n-2} + 2c_{n-2}(x_{n-2} - x_{n-1}) + 3d_{n-2}(x_{n-2} - x_{n-1})^2 = b_{n-1} \end{aligned} \quad (20)$$

Whereas applying continuity properties of equation (13) into equation (19) gives the following

$$2c_i + 6d_i(x_{i+1} - x_i) = 2c_{i+1} \quad i = 0, 1, ..., n-2$$

C. Data Preparation

Twitter Dataset

The dataset obtained through social media Twitter is crawled at odd hours for 5 minutes. We assume that time data collection is carried out to the retrieval process in order regulated through a computer that acts as a server. Data is stored in text format and extracted to get keywords from the generated tweets.

Data Cleaning

This means it is performed after crawled raw data is stored in the storage repository. On the Twitter structure, the variables that will be processed are *id*, *create_at*, *text*. Particularly for text variables, they must be cleaned to simplify the COVID-19 keyword calculation process. The cleanup starts with an update and removing ASCII characters representing unnecessary characters or symbols for each tweet posted. Then the whole tweet is converted to lowercase for uniformity to establish it more comfortable to count. The tweet variable is the text placed in one column as the initial counting.

Counting Keywords

The computations of keywords are obtained on a file containing cleaned tweets. That will calculate how many COVID-19 words emerge, which netizens talk about for the specified time duration. This process only uses the text attribute to calculate the number of tweet keywords in each tweet posted by the user. This calculation does not go through the process of converting each word data into a lowercase.

III. Result and Discussion

A. Cubic Spline Interpolation

Having the dataset (see Appendix), we can construct the

interpolating cubic spline passing through given points. With three data points, such as (35, 503), (37, 519), (39, 747), we can construct spline functions as follows.

$$\zeta_0(x) = a_0 + b_0(x-35) + c_0(x-35)^2 + d_0(x-35)^3, 35 \le x \le 37$$
$$\zeta_1(x) = a_1 + b_1(x-37) + c_1(x-37)^2 + d_1(x-37)^3, 37 \le x \le 39 \quad (21)$$

Based on equations (10), (11) and (21), we have

$$\zeta_0(35) = 503, \; a_0 = 503 \quad (22)$$
$$\zeta_0(37) = 519, \; a_0 + 2b_0 + 4c_0 + 8d_0 = 519$$
$$2b_0 + 4c_0 + 8d_0 = 16 \quad (23)$$
$$\zeta_1(37) = 519, \; a_1 = 519 \quad (24)$$
$$\zeta_1(39) = 747, \; a_1 + 2b_1 + 4c_1 + 8d_1 = 747$$
$$2b_1 + 4c_1 + 8d_1 = 228 \quad (25)$$

Using equation (20) we have

$$\frac{d}{dx}\zeta_0(x) = b_0 + 2c_0(x-35) + 3d_0(x-35)^2$$
$$\frac{d}{dx}\zeta_1(x) = b_1 + 2c_1(x-37) + 3d_1(x-37)^2$$

Using continuous condition (12) gives

$$\frac{d}{dx}\zeta_0(37) = \frac{d}{dx}\zeta_1(37),$$
$$b_0 + 4c_0 + 12d_0 = b_1 \quad (26)$$
$$\frac{d^2}{dx^2}\zeta_0(x) = 2c_0 + 6d_0(x-35)$$
$$\frac{d^2}{dx^2}\zeta_1(x) = 2c_1 + 6d_1(x-37)$$

Using continuous condition (13) gives

$$\frac{d^2}{dx^2}\zeta_0(37) = \frac{d^2}{dx^2}\zeta_1(37)$$
$$2c_0 + 12d_0 = 2c_1 \quad (27)$$

Applying the boundary conditions of natural cubic spline (14) gives the following equations for solving the coefficients,

$$\frac{d^2}{dx^2}\zeta_0(35) = \frac{d^2}{dx^2}\zeta_1(39) = 0 \quad (28)$$
$$\frac{d^2}{dx^2}\zeta_0(35) = 2c_0 = 0$$
$$c_0 = 0 \quad (29)$$
$$\frac{d^2}{dx^2}\zeta_1(39) = 0$$
$$\frac{d^2}{dx^2}\zeta_1(39) = 2c_1 + 12d_1 = 0 \quad (30)$$

We come to eight equations in eight unknown (22-30) to solve for the coefficients. This can be written as follow,

$$\begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} a_0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \begin{array}{c} b_0 \\ 0 \\ 2 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{array} \begin{array}{c} c_0 \\ 0 \\ 4 \\ 0 \\ 0 \\ 4 \\ 2 \\ 0 \\ 0 \end{array} \begin{array}{c} d_0 \\ 0 \\ 8 \\ 0 \\ 0 \\ 12 \\ 12 \\ 0 \\ 0 \end{array} \begin{array}{c} a_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \begin{array}{c} b_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \end{array} \begin{array}{c} c_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -2 \\ 0 \\ 0 \end{array} \begin{array}{c} d_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \left(\begin{array}{c} \\ 503 \\ 16 \\ 519 \\ 228 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right)$$

Solving the system of the equation gives the coefficients as follows,

$$a_0 = 503, b_0 = -18.5, c_0 = , d_0 = .625, a_1 = 19,$$
$$b_1 = 61, c_1 = 9.75, d_1 = -6.625$$

With the three data points, we then come to the interpolating cubic spline given as follows,

$$\zeta(x) = \begin{cases} 503 - 18.5(x-35) + 6.625(x-35)^3, 35 \le x \le 37 \\ 519 + 61(x-37) + 39.75(x-37)^2 - 6.625(x-37)^3, 37 \le x \le 39 \end{cases} \quad (31)$$

The graph of the cubic spline (31) is shown in Fig. 2.

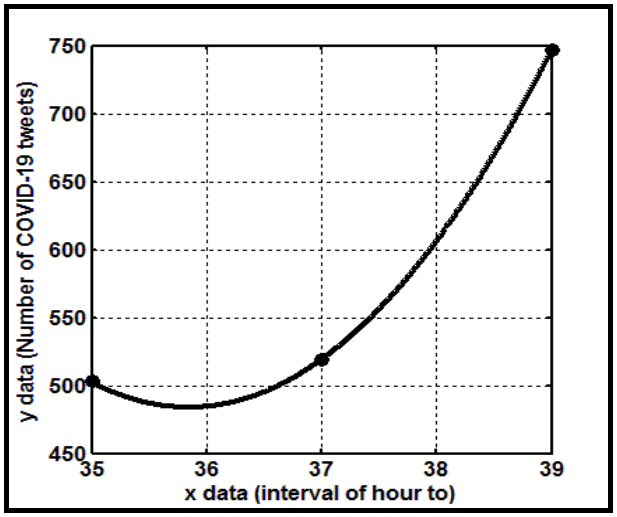

Fig. 2. Cubic spline interpolation with three points

Cubic spline (31) can be used to approximate the number of tweets discussing COVID-19 on the interval [36,38], which is calculated as follows,

$$\int_{36}^{38} \zeta(x)dx$$

The integral above is integral that is calculated using the piecewise integral of the intervals [36,37] and [37,38]

$$\int_{36}^{38} \zeta(x)dx = \int_{36}^{37} \zeta(x_0)dx + \int_{37}^{38} \zeta(x_1)dx$$
$$= \int_{36}^{37} (503 - 18.5(x-35) + 6.625(x-35)^3)dx$$

$$+\int_{37}^{38}(519+61(x-37)+39.75(x-37)^2-6.625(x-37)^3)dx$$

$$=500.09+561.10$$

$$=1061.19 \text{ Tweets approximation}$$

Approximately 1061.19 tweets emerge from hours 36-38.

The growth rate of the number of tweets can also be calculated using the first derivative of the approximation function (31), given as follows

$$v(x)=\frac{d}{dx}\zeta(x) \text{ at } x=38$$

$$\frac{d}{dx}\left(519+61(x-37)+39.75(x-37)^2-6.625(x-37)^3\right)$$

at $x=38$, $v(38)=120.62$

While the approximate value at 38 hours with interval [37,39] is 613.13 tweets.

B. Newton's divided difference interpolation

We then construct an interpolation using Newton's divided difference formula on the same data points. Starting with the calculation of divided difference functions, we have,

$$\xi_0=y_0$$
$$=503$$
$$\xi_1=\frac{y_1-y_0}{x_1-x_0}$$
$$=\frac{519-503}{37-35}$$
$$=8$$
$$\xi_2=\frac{\frac{y_2-y_1}{x_2-x_1}-\frac{y_1-y_0}{x_1-x_0}}{x_2-x_0}$$
$$=\frac{\frac{747-519}{39-37}-\frac{519-503}{37-35}}{39-35}$$
$$=\frac{30.914-27.148}{10}$$
$$=25.6$$

Following equation (8) we have

$$P_2(x)=y_0+(x-x_0)f[x_0,x_1]+(x-x_0)(x-x_1)f[x_0,x_1,x_2]$$
$$P_2(x)=503+(x-35)\times 8+(x-35)(x-37)\times 26.5$$
$$P_2(x)=503+(x-35)\times 8+(x^2-72x+1295)\times 26.5$$
$$P_2(x)=503+(8x-280)+(26.5x^2-1908x+34317.5)$$
$$P_2(x)=26.5x^2-1900x+34540.5 \quad (32)$$

The Newton quadratic divided difference formula for interpolating polynomial (32) is shown in Fig. 3.

For degree 2, this graph does not show a significant difference from the cubic spline (Fig. 2). However, they do for a higher degree of polynomial shown in the following.

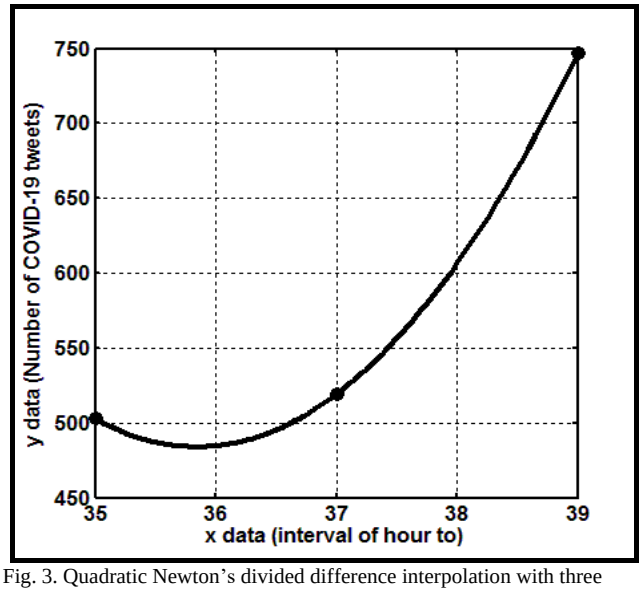

Fig. 3. Quadratic Newton's divided difference interpolation with three points

Visual Comparison of Polynomial and Cubic Spline Interpolation

We represent piecewise linear polynomial, piecewise quadratic polynomial, cubic spline, and Lagrange with degree 4 polynomial passing through the same 5 points whose graphs are shown in Fig.4 through Fig. 7.

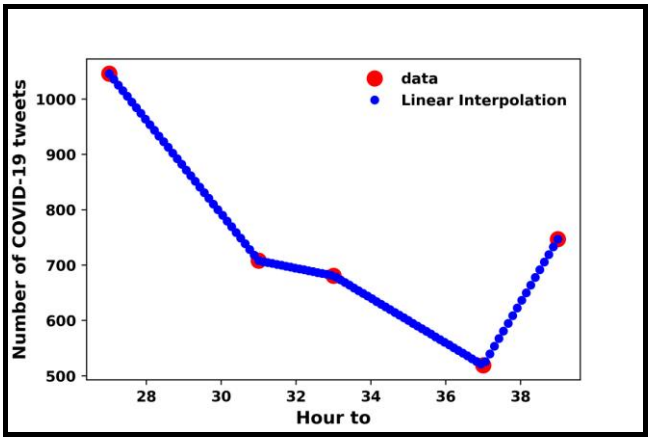

Fig. 4. Piecewise linear interpolation passing through 5 points

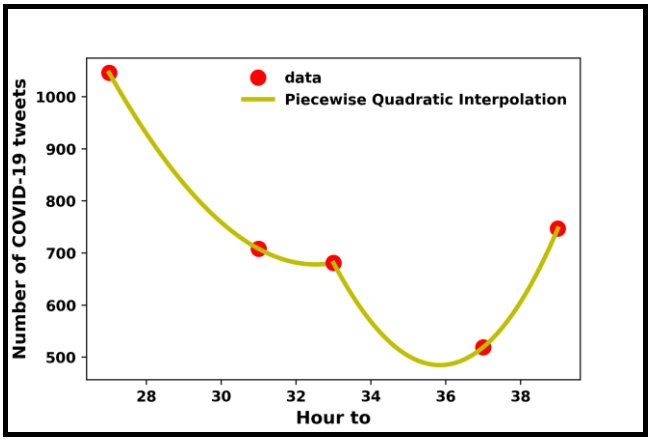

Fig. 5. Piecewise quadratic interpolation passing through 5 points.

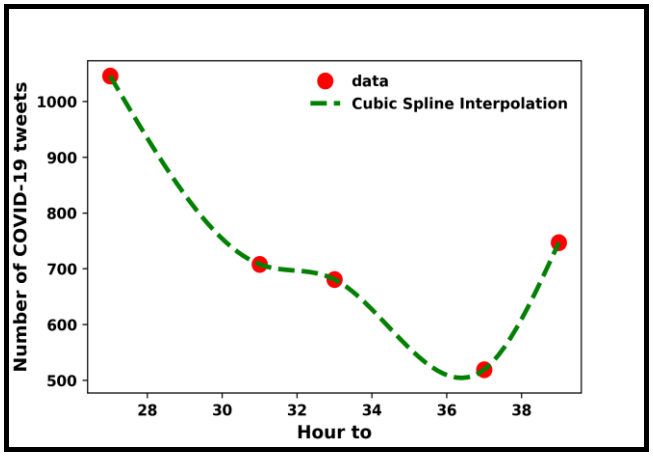

Fig. 6. Cubic spline interpolation passing through 5 points.

This shows that the piecewise linear is not smooth, even though it is good in approaching data points. The piecewise quadratic does not have a continuous derivative at point 33, shown by a corner. The higher degree of polynomials tends to change much between data points, whereas the cubic spline tends to be smooth and does not change much between data points. It is shown in Fig. 8 when all those figures are plotted together.

We then put together all those graphs in Fig. 8 to have a clear visual comparison.

As a whole cubic spline is the best in interpolating the data points as shown in the plot of Fig. 8. It combines the properties of piecewise linear and polynomial in terms of exact fitting data points and not too much changing between data points and smoothness.

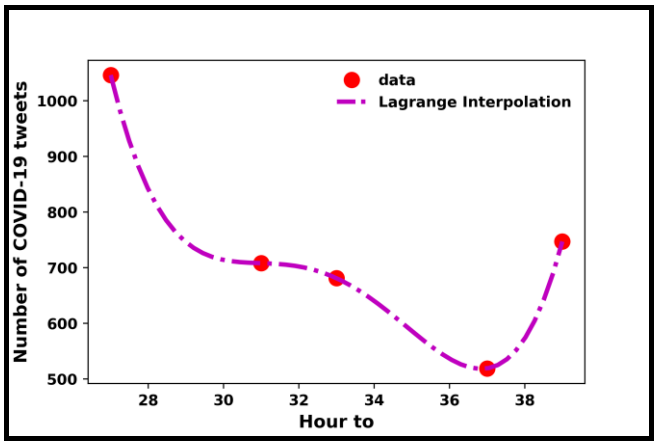

Fig. 7. Lagrange interpolation passing through 5 points.

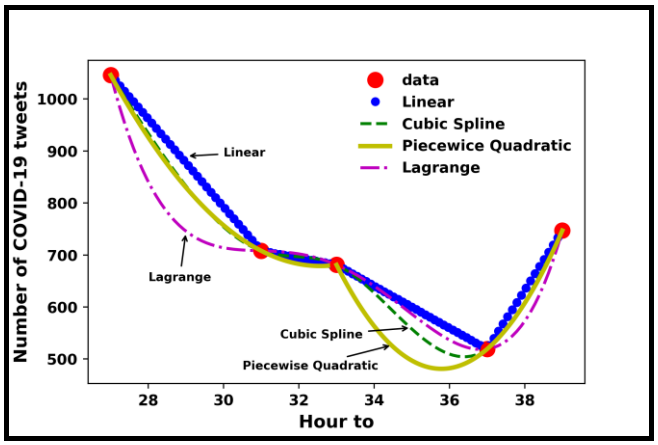

Fig. 8. Graph of piecewise linear, piecewise quadratic, Lagrange, and cubic spline interpolation passing through 5 points

Furthermore, using 11 data points gives a plot shown in Fig. 9. Piecewise linear shows unsmooth function. Whereas high degree polynomial shows significant changes between data points, especially between data points 3 and 4, then 10 and 11. This ensures that the cubic spline is the best amongst the other interpolations shown.

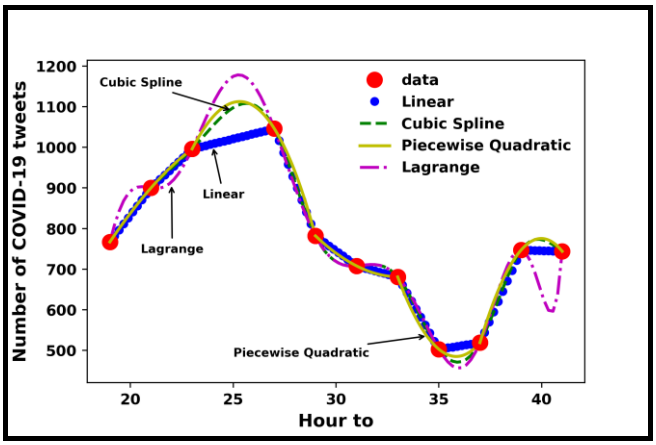

Fig. 9. Graph of piecewise linear, piecewise quadratic, Lagrange, and cubic spline interpolation passing through 11 points

These tweet data are taken by crawling at odd hours for three days with 5 minutes per retrieval. After the data cleaning process is carried out, the following calculation process counts the number of keywords in each tweet during these hours. The interpolation process is carried out to construct an approximation of the function that passes through the given data points. The models used are Lagrange interpolation, Newton's divided difference, and cubic spline. This Experiment using piecewise linear, piecewise quadratic, Lagrange, and cubic spline interpolation. The result shows the cubic spline produces the three approximation functions smoothest curve. With 3 points, continuous approximation functions at two subintervals with smooth curves without angles generated by a cubic spline, which is different from the other two interpolation models. Besides, the data plot represents patterns at certain hours. By utilizing the first derivative and integration of the approximation function, we approximate the increasing rate of tweets and the accumulated number of tweets in a subinterval, respectively. The denser the defined subinterval, the smoother the curve path will be. To solve another problem, we estimate the number of tweets at intermediate times at odd hours in the dataset for a specific interval. It allows analysis and comparison with actual data regarding the number of COVID-19 tweets posted by netizens. With the dataset information, calculating the increasing rate of tweets is also carried out, which helps calculate the accumulated number of tweets between 2 intervals.

IV. Conclusion

This study built an approximation function using interpolation on the Twitter dataset searched with the keyword COVID-19. According to the literature statistical analysis, our study belongs to the analytical interpolation group. The interpolations used are Lagrange polynomial, Newton's divided difference polynomial, and cubic spline

applied on the COVID-19 tweet data taken at a particular time duration.

In calculating the approximate number of twitters in the interval data, cubic splines are used by constructing and using comparisons to high degree interpolated polynomials. It starts with deriving n^{th} data points continuously. Then, the polynomial equation is derived from the first to the second derivative to obtain the coefficients. This construction is the main activity of this paper to approximate the number of twitters regarding COVID-19 in the j^{th} hour. It can be seen the contour lines of the cubic spline graph that it is not so extreme compared to Lagrange and Newton's divided difference when the line connects from one point to another. The cubic spline is considered the most optimal approximation function to the data points (see Fig. 8 and 9) in terms of smoothness, exact fit, and small changes between data points. The graph is significantly smooth through the whole interval.

The resulting function can be used to calculate the number of tweet data between data points and the growth rate of the number of tweet data. These results will become a potential part of the development of social media analysis.

APPENDIX

SAMPLE OF DATA TWEETS WITH THE KEYWORD "COVID-19"

Number of data	Date	Time	Hour to	Number of Tweets
1	11/04/2020	03:00-03:05	3	1369
2	11/04/2020	05:00-05:05	5	1002
3	11/04/2020	07:00-07:05	7	763
4	11/04/2020	09:00-09:05	9	546
5	11/04/2020	11:00-11:05	11	488
6	11/04/2020	13:00-13:05	13	503
7	11/04/2020	15:00-15:05	15	673
8	11/04/2020	17:00-17:05	17	619
9	11/04/2020	19:00-19:05	19	767
10	11/04/2020	21:00-21:05	21	900
11	11/04/2020	23:00-23:05	23	996
12	11/05/2020	03:00-03:05	27	1046
13	11/05/2020	05:00-05:05	29	782
14	11/05/2020	07:00-07:05	31	708
15	11/05/2020	09:00-09:05	33	681
16	11/05/2020	11:00-11:05	35	503
17	11/05/2020	13:00-13:05	37	519
18	11/05/2020	15:00-15:05	39	747
19	11/05/2020	17:00-17:05	41	744
20	11/05/2020	19:00-19:05	43	997
21	11/05/2020	21:00-21:05	45	1148
22	11/05/2020	23:00-23:05	47	1226
23	11/06/2020	03:00-03:05	51	1208
24	11/06/2020	05:00-05:05	53	1429
25	11/06/2020	07:00-07:05	55	836
26	11/06/2020	09:00-09:05	57	730

REFERENCES

[1] J. Kwon, C. Grady, J. T. Feliciano, and S. J. Fodeh, "Defining facets of social distancing during the COVID-19 pandemic: Twitter analysis," *J. Biomed. Inform.*, vol. 111, p. 103601, 2020, doi: https://doi.org/10.1016/j.jbi.2020.103601.

[2] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media," *Appl. Soft Comput.*, vol. 97, p. 106754, 2020, doi: https://doi.org/10.1016/j.asoc.2020.106754.

[3] F. Schultz, S. Utz, and A. Göritz, "Is the medium the message? Perceptions of and reactions to crisis communication via twitter, blogs and traditional media," *Public Relat. Rev.*, vol. 37, no. 1, pp. 20–27, 2011, doi: https://doi.org/10.1016/j.pubrev.2010.12.001.

[4] S. M. Abdullah, "A Two-phase Analyzer for Vulnerabilities of Online Social Media Users," *IAENG Int. J. Comput. Sci.*, vol. 47, no. 2, pp. 144–153, 2020.

[5] J. Groshek, V. de Mees, and R. Eschmann, "Modeling influence and community in social media data using the digital methods initiative-twitter capture and analysis toolkit (DMI-TCAT) and Gephi," *MethodsX*, vol. 7, p. 101164, 2020, doi: https://doi.org/10.1016/j.mex.2020.101164.

[6] P. Grover, A. K. Kar, and G. Davies, "'Technology enabled Health' – Insights from twitter analytics with a socio-technical perspective," *Int. J. Inf. Manage.*, vol. 43, pp. 85–97, 2018, doi: https://doi.org/10.1016/j.ijinfomgt.2018.07.003.

[7] B. K. Chae and E. O. Park, "Corporate social responsibility (CSR): A survey of topics and trends using Twitter data and topic modeling," *Sustainability*, vol. 10, no. 7, p. 2231, 2018.

[8] Y. Mejova, I. Weber, and M. W. Macy, Eds., *Twitter: A Digital Socioscope*. Cambridge: Cambridge University Press, 2015.

[9] R. C. Dunn, "'You sir are a hypocrite': responses to Pence's MLK Day tweets as attention intervention," *Atl. J. Commun.*, vol. 27, no. 5, pp. 354–365, 2019, doi: 10.1080/15456870.2019.1647206.

[10] L. N. Trefethen, *Approximation Theory and Approximation Practice, Extended Edition*. Philadelphia, PA, USA: SIAM-Society for Industrial and Applied Mathematics, 2019.

[11] B. Adcock and R. Platte, "A mapped polynomial method for high-accuracy approximations on arbitrary grids," *SIAM J. Numer. Anal.*, vol. 54, no. 4, pp. 2256–2281, 2016, doi: 10.1137/15M1023853.

[12] C. J. Sánchez and J. I. Yuz, "On the relationship between spline interpolation, sampling zeros and numerical integration in sampled-data models," *Syst. Control Lett.*, vol. 128, pp. 1–8, 2019, doi: https://doi.org/10.1016/j.sysconle.2019.04.006.

[13] G. Martinez-Guzman, M. Bustillo-Diaz, A. Rangel-Huerta, G. Juarez-Diaz, A. Ata-Perez, and N. Quiroz-Hernandez, "Approximation with Interpolation Conditions of an Asymptotic Function Using K-algebraic Lagrange Interpolation," *Eng. Lett.*, vol. 23, no. 2, pp. 77–81, 2015.

[14] P. . Douris and M. . Markakis, "Global Connecting Orbits of a SEIRS Epidemic Model with Nonlinear Incidence Rate and Nonpermanent Immunity," *Eng. Lett.*, vol. 27, no. 4, pp. 866–875, 2019.

[15] L. Li, Z. Wei, and Q. Huang, "A Numerical Method for Solving Fractional Variational Problems by the Operational Matrix Based on Chelyshkov Polynomials," *Eng. Lett.*, vol. 28, no. 2, pp. 486–491, 2020.

[16] A. Chkifa, A. Cohen, and C. Schwab, "High-Dimensional Adaptive Sparse Polynomial Interpolation and Applications to Parametric PDEs," *Found. Comput. Math.*, vol. 14, no. 4, pp. 601–633, 2014, doi: 10.1007/s10208-013-9154-z.

[17] T.-H. S. Li, Y.-T. Su, S.-W. Lai, and J.-J. Hu, "Walking motion generation, synthesis, and control for biped robot by using PGRL, LPI, and fuzzy logic," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 41, no. 3, pp. 736–748, 2011, doi: 10.1109/TSMCB.2010.2089978.

[18] L. Bos, J.-P. Calvi, N. Levenberg, A. Sommariva, and M. Vianello, "Geometric weakly admissible meshes, discrete least squares approximations and approximate fekete points," *Math. Comput.*, vol. 80, no. 275, pp. 1623–1638, 2011, doi: 10.1090/S0025-5718-2011-02442-7.

[19] B. Ghanbari and J. F. Gómez-Aguilar, "Modeling the dynamics of nutrient–phytoplankton–zooplankton system with variable-order fractional derivatives," *Chaos, Solitons and Fractals*, vol. 116, pp. 114–120, 2018, doi: 10.1016/j.chaos.2018.09.026.

[20] J. E. Solís-Pérez, J. F. Gómez-Aguilar, and A. Atangana, "Novel numerical method for solving variable-order fractional differential equations with power, exponential and Mittag-Leffler laws," *Chaos, Solitons and Fractals*, vol. 114, pp. 175–185, 2018, doi: 10.1016/j.chaos.2018.06.032.

[21] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *International Workshop on Acoustic Signal Enhancement, IWAENC 2012*, 2012.

[22] Y. X. Zou, S. L. Zhang, Y. C. Lim, and X. Chen, "Timing mismatch compensation in time-interleaved ADCs based on multichannel lagrange polynomial interpolation," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 4, pp. 1123–1131, 2011, doi: 10.1109/TIM.2010.2085291.

[23] C. Harder, D. Paredes, and F. Valentin, "A family of Multiscale Hybrid-Mixed finite element methods for the Darcy equation with rough coefficients," *J. Comput. Phys.*, vol. 245, pp. 107–130, 2013, doi: 10.1016/j.jcp.2013.03.019.

[24] M. A. Chkifa, "On the Lebesgue constant of Leja sequences for the

complex unit disk and of their real projection," *J. Approx. Theory*, vol. 166, no. 1, pp. 176–200, 2013, doi: 10.1016/j.jat.2012.11.005.

[25] X. Tang, F. Peng, R. Yan, Y. Gong, Y. Li, and L. Jiang, "Accurate and efficient prediction of milling stability with updated full-discretization method," *Int. J. Adv. Manuf. Technol.*, vol. 88, no. 9–12, pp. 2357–2368, 2017, doi: 10.1007/s00170-016-8923-7.

[26] H. Abdi and D. Valentin, "Multiple Correspondence Analysis," *Encycl. Meas. Stat.*, 2007.

[27] S. Boorboor, H. Jafari, and S. A. H. Feghhi, "Development of a novel approach for precise pulse height extraction using Lagrange interpolation," *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, vol. 919, pp. 82–88, 2019, doi: https://doi.org/10.1016/j.nima.2018.12.028.

[28] S. De Marchi, F. Marchetti, E. Perracchione, and D. Poggiali, "Polynomial interpolation via mapped bases without resampling." 2019.

[29] T. H. Cheong and T. Gaik, "Newton's divided difference interpolation using scientific calculator," 2016, vol. 1775, p. 30102, doi: 10.1063/1.4965222.

[30] B. Das and D. Chakrabarty, "Newton's Divided Difference Interpolation formula: Representation of Numerical Data by a Polynomial curve," *Int. J. Math. Trends Technol.*, vol. 35, pp. 197–203, 2016, doi: 10.14445/22315373/IJMTT-V35P528.

[31] X. Qin and Q. Xu, "C1 Rational Cubic/Linear Trigonometric Interpolation Spline with Positivity-preserving Property," *Eng. Lett.*, vol. 25, no. 2, pp. 152–159, 2017.

[32] K. S. Sim, F. F. Ting, J. W. Leong, and C. P. Tso, "Signal-to-noise Ratio Estimation for SEM Single Image using Cubic Spline Interpolation with Linear Least Square Regression," *Eng. Lett.*, vol. 27, no. 1, pp. 151–165, 2019.

[33] T. A. Tabet and F. A. Aziz, "Modeling Microfibril Angle and Tree Age in Acacia Mangium Wood using X-Ray Diffraction Technique," in *Proceedings of The World Congress on Engineering 2012*, 2012, pp. 1687–1691.

[34] X. Qin, L. Qin, and Q. Xu, "C1 Positivity-preserving Interpolation Schemes with Local Free Parameters," *IAENG Int. J. Comput. Sci.*, vol. 43, no. 2, pp. 219–227, 2016.

[35] H. Li, L. Li, and D. zhao, "An improved EMD method with modified envelope algorithm based on C2 piecewise rational cubic spline interpolation for EMI signal decomposition," *Appl. Math. Comput.*, vol. 335, pp. 112–123, 2018, doi: https://doi.org/10.1016/j.amc.2018.04.008.

[36] P. Roul, K. Thula, and R. Agarwal, "Non-optimal fourth-order and optimal sixth-order B-spline collocation methods for Lane-Emden boundary value problems," *Appl. Numer. Math.*, vol. 145, pp. 342–360, 2019, doi: https://doi.org/10.1016/j.apnum.2019.05.004.