AIMS Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 9(9): 23158–23181.
DOI: 10.3934/math.20241126
Received: 27 May 2024
Revised: 19 July 2024
Accepted: 23 July 2024
Published: 30 July 2024

Research article

Flexible functional data smoothing and optimization using beta spline

Wan Anis Farhah Wan Amir, Md Yushalify Misro* and Mohd Hafiz Mohd

School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Gelugor, Pulau Pinang, Malaysia

* **Correspondence:** Email: yushalify@usm.my.

Abstract: Functional data analysis (FDA) is a method used to analyze data represented in its functional form. The method is particularly useful for exploring both curve and longitudinal data in both exploratory and inferential contexts, with minimal constraints on the parameters. In FDA, the choice of basis function is crucial for the smoothing process. However, traditional basis functions lack flexibility, limiting the ability to modify the shape of curves and accurately represent abnormal details in modern and complex datasets. This study introduced a novel and flexible data smoothing technique for interpreting functional data, employing the beta spline introduced by Barsky in 1981. The beta spline offers flexibility due to the inclusion of two shape parameters. The proposed methodology integrated the roughness penalty approach and generalized cross-validation (GCV) to identify the optimal curve that best fitted the data, ensuring appropriate parameters were considered for transforming data into a functional form. The effectiveness of the approach was assessed by analyzing the GCV color grid chart to determine the optimal curve. In contrast to existing methodologies, the proposed method enhanced flexibility by incorporating the beta spline into the smoothing procedure. This approach was anticipated to effectively handle various forms of time series data, offering improved interpretability and accuracy in data analysis, including forecasting.

Keywords: functional data analysis; data smoothing technique; beta spline; shape parameters; optimization
Mathematics Subject Classification: 62R10, 65D07, 93E14

1. Introduction

Functional data analysis (FDA) is a statistical technique used to analyze data consisting of functions or data produced by underlying functions. The FDA aims to provide exploratory and inferential tools for analyzing curve and longitudinal data with minimum constraints on the parameters involved [1]. This collection includes techniques such as functional principal component analysis for dimension reduction [2, 3], functional regression [4, 5], and functional clustering and classification [6, 7]. Due to notable progress in methodology and software tools, the development of the FDA has become a firmly established subject in nonparametric statistics. Distinct fields have utilized FDA, such as image analysis [8], studying the transmission of diseases like COVID-19 [9], and growth curve analysis [10].

The application of FDA to climate variables has recently gained significant attention. A study by [11] proposed a functional time series approach for hourly air temperature forecasts, enabling ultra-short period predictions that traditional methods cannot achieve. Similarly, [12] introduced a new spatial functional data analysis approach to evaluate the performance of 18 CORDEX regional climate model (RCM) simulations for the European domain (EURO-CORDEX) in predicting average temperatures in Italy. This approach addresses the limitation of traditional climate model selection, which typically focuses only on average values across time, by considering the overall mean of the function rather than detailed temporal behavior. An innovative method of functional principal component analysis (fPCA) for incomplete space-time data has also been introduced in research by [13], allowing for the identification of main variability patterns in temperature data. According to [14], the initial stage of FDA involves smoothing [15, 16] or interpolation [17], which consists of converting discrete data into a function. Interpolation is applied when discrete values are assumed to be without error, whereas smoothing transforms data into a functional form by removing any observational errors [14].

While interpolation can be performed using beta spline, it is inadequate when derivative information of the data is required [14]. One of the main objectives of the FDA is to study important patterns and variations in the data, which requires accurate derivative information. Derivative information, or the rate of change, can only be extracted when the data is in the form of a function, which is possible only when the data is approximated. Interpolation using beta spline can be done, but typically, it results in a very straight line, which does not effectively represent the functional form of large datasets. This is because the curve is merely interpolated between one data point to another, creating straight lines rather than capturing the data's complexity. Therefore, in the FDA framework, Fourier or B-spline basis are generally used for approximation rather than interpolation, as highlighted by [15, 16, 18, 19]. Thus, the approximation method is preferred to represent and analyze the data adequately.

Smoothing is a technique for identifying a sequence of numbers that accurately represents the trend in a given dataset. This technique is commonly used for time series data with variations or seasonality [20]. Smoothing data eliminates noise or random fluctuations to improve the clarity of patterns and trends. The next step is data visualization, which involves creating visual representations of data to gain insights and identify patterns and trends. The roughness penalty approach (RPA) and generalized cross-validation (GCV) are popular approaches for smoothing discrete data and determining the best parameters for converting it to a functional form. Other than RPA, recent studies have also proposed spline approximation methods for smoothing time series with extreme events. The study by [21] presented a spline Hermite quasi-interpolation method for filling in missing data and smoothing univariate time series. This model can be used for forecasting and detecting anomalies. An entropy-based weighting methodology for determining spline approximations of multivariate time series was introduced by [22]. The method demonstrated to effectively mitigate the impact of outliers and noise even when handling large and highly noisy datasets. The most commonly used basis splines are B-spline and Fourier. The B-spline basis is widely used to fit nonperiodic data due to its characteristics that allow the curve to be more flexible and its efficient modeling of time-varying patterns [23]. In contrast, the Fourier basis is often chosen for its fast processing speed and ability to handle periodic data [24].

Within the FDA framework for analyzing climatic data, Fourier basis functions were utilized in a study by [19] to smooth temperature data due to its periodic structure. In another investigation by [25], monthly maximum temperature variation was explored using B-spline basis functions. Gaussian basis functions were chosen in [26] for their ability to effectively smooth functional data and capture underlying patterns. This approach strikes a balance between overfitting and underfitting, thereby enhancing model performance. Conversely, in a study by [27], Fourier basis functions were initially employed for data smoothing, justified by the periodicity of the air temperature series. However, the use of B-splines yielded slightly better forecast results due to their ability to provide a good balance between model flexibility and overfitting in capturing complex patterns within the data, ensuring accurate forecasts [27].

The choice of basis functions is crucial as it directly impacts the model's ability to represent the data accurately and make reliable analysis, such as in forecasting [27]. When presenting a set of data points, utilizing a spline with greater flexibility is advisable as it enhances the smoothing process, leading to improved forecast analysis performance. Many new basis functions have recently been developed to improve surface and curve flexibility, such as by Ammad et al. [28] and Said Mad Zain et al. [29]. These new basis functions provide new shape parameters to flexibly alter the shape of the curve by retaining the existing control points and conveniently processing shape changes. In 1981, Barsky [30] developed the beta spline, a flexible extension of the B-spline. The beta spline has distinct advantages over other splines as it achieves G^2 continuity while also being characterized by two additional parameters that impact the curve's shape. Its curve and surface shape can be adjusted without altering control points [31]. Beta splines generally create flexible shapes and have a smoother appearance than those generated by Bézier curves [32].

Due to its ability to model curves with flexibility, the beta spline is valuable for study in image processing and machine vision [33]. The beta spline curve provides enhanced capabilities in making 2D graphics, such as digital khat calligraphy, by offering smoother shapes, control over vertex movement, and assuring continuity and flexibility [34]. According to [35], beta spline interpolation is a technique that provides the most accurate and smooth curve fitting by selecting the curve that is closest to the data points. Beta spline surfaces can also be generated by parallel computation, enabling faster processing and handling of large datasets. The parallel beta spline method in surface fitting yields efficient and precise outcomes. According to [36], incorporating the parallel method does not alter the surface structure, guaranteeing the integrity of the reconstructed surface.

This research seeks to integrate cubic beta spline in smoothing climate data observations into a functional form, aiming to develop a new and flexible technique for data smoothing in functional data analysis. In the initial phase, this research integrates spline smoothing using cubic beta spline and RPA to transform the discrete data into a curve. Next, the shape parameters of the spline will be optimized by applying a method known as GCV. This optimization method is applied to help determine the optimal combination of shape parameters to obtain the best-fitted curve. Finally, the rainbow plot is used to represent the best-fit temperature curve of each meteorology station in north Peninsular Malaysia.

This paper is structured as follows: In Section 2, the cubic beta spline basis is defined, the curve is constructed, and the effects of manipulating the shape parameters are explained. Next, the smoothing approach is presented in Section 3, together with the calculation for finding the smoothing parameter value, λ. The formula for selecting shape parameters, GCV, is described in Section 4. Section 5 presents the smoothing results by implementing the shape parameters value for optimal, overfitted, and underfitted shape parameters. Finally, Section 6 provides the conclusion and a few possible directions for further research.

2. Cubic beta spline

This study focused on applying beta spline in the FDA framework, which was developed by Barsky during his doctoral research [30]. The development of the basis was also extensively studied in [37,38]. The expression for the beta spline curve of degree 3 is given by Eq (2.1),

$$F(t) = [T][M][V]^T \tag{2.1}$$

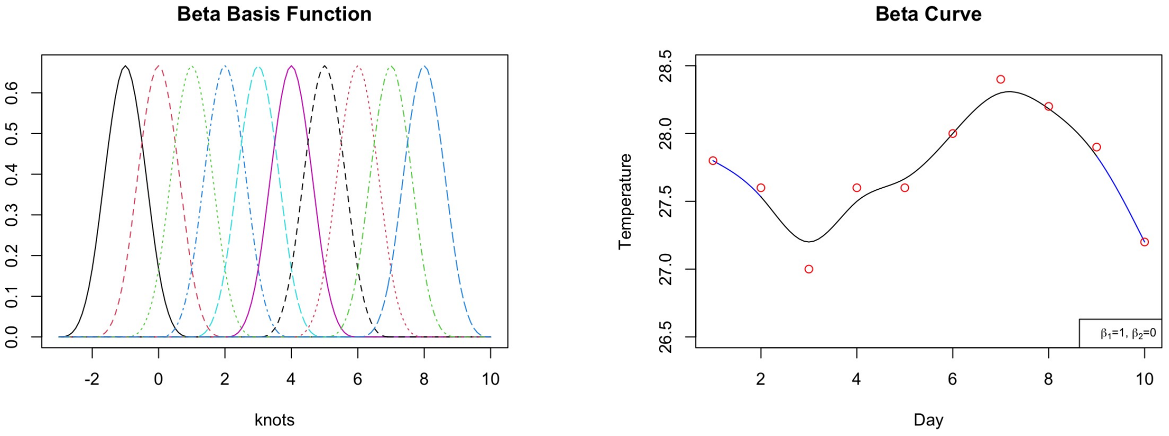
where $[T] = \begin{pmatrix} t^3 & t^2 & t & 1 \end{pmatrix}$ is the polynomial matrix, $[V] = \begin{pmatrix} V_1 & V_2 & V_3 & V_4 \end{pmatrix}$ is the control point vector and $[M]$ is the beta spline basis function matrix that is given as the following Eq (2.2):

$$[M] = \frac{1}{\delta} \begin{pmatrix} -2\beta_1^3 & 2(\beta_2 + \beta_1^3 + \beta_1^2 + \beta_1) & -2(\beta_2 + \beta_1^2 + \beta_1 + 1 & 2 \\ 6\beta_1^3 & -3(\beta_2 + 2\beta_1^3 + 2\beta_1^2) & 3(\beta_2 + 2\beta_1^2) & 0 \\ -6\beta_1^3 & 6(\beta_1^3 - \beta_1) & 6\beta_1 & 0 \\ 2\beta_1^3 & \beta_2 + 4(\beta_1^2 + \beta_1) & 2 & 0 \end{pmatrix} \tag{2.2}$$

where $\delta = \beta_2 + 2\beta_1^3 + 4\beta_1^2 + 4\beta_1 + 2$, with β_1 (bias) and β_2 (tension) are the shape parameters of the spline.

By implementing these shape parameters, one can manipulate the shape without changing the control points while simultaneously introducing G^2 continuity to ensure that the fitted curve maintains accuracy and smoothness. The beta spline curve also satisfies the properties of locality, convex hull, and end conditions. However, end conditions for beta splines are often handled differently than those for other splines, such as B-splines, which will be further discussed in the following subsection. When constructing the beta spline curve, determining an adequate number of data points is not a concern. Each curve segment of the beta spline can be continuously connected between data points without needing intermediate points. In contrast, methods like B-splines or Fourier may encounter issues with having either too many or too few data points. For instance, B-splines and Fourier require an exact number of data points based on the chosen degree of basis functions, while Bézier splines depend on the number of control points. Beta splines, however, rely more on the knots. Consequently, an excessive or insufficient number of data points does not pose an issue when using beta splines, unlike other spline methods.

Figure 1 is a beta spline basis and its curve with $\beta_1 = 1$ and $\beta_2 = 0$, which also has a similar shape with a B-spline without repeated knots. The comparison curves representing B-splines both with repeated knots and without repeated knots and beta splines are depicted in Figure 2. Notably, the curvature of the B-spline with repeated knots differs from the others due to variations in the underlying basis values. When repeated knots are absent, the B-spline curve resembles the beta spline curve, particularly when $\beta_1 = 1$ and $\beta_2 = 0$. However, for different values of β_1 and β_2, the beta spline basis does not maintain C^2 continuity at knots and instead achieves G^2 continuity [39]. Examples illustrating this behavior are presented in the next subsection.



(a) Beta basis functions.

(b) Beta curve.

Figure 1. Beta spline with $\beta_1 = 1$ and $\beta_2 = 0$.

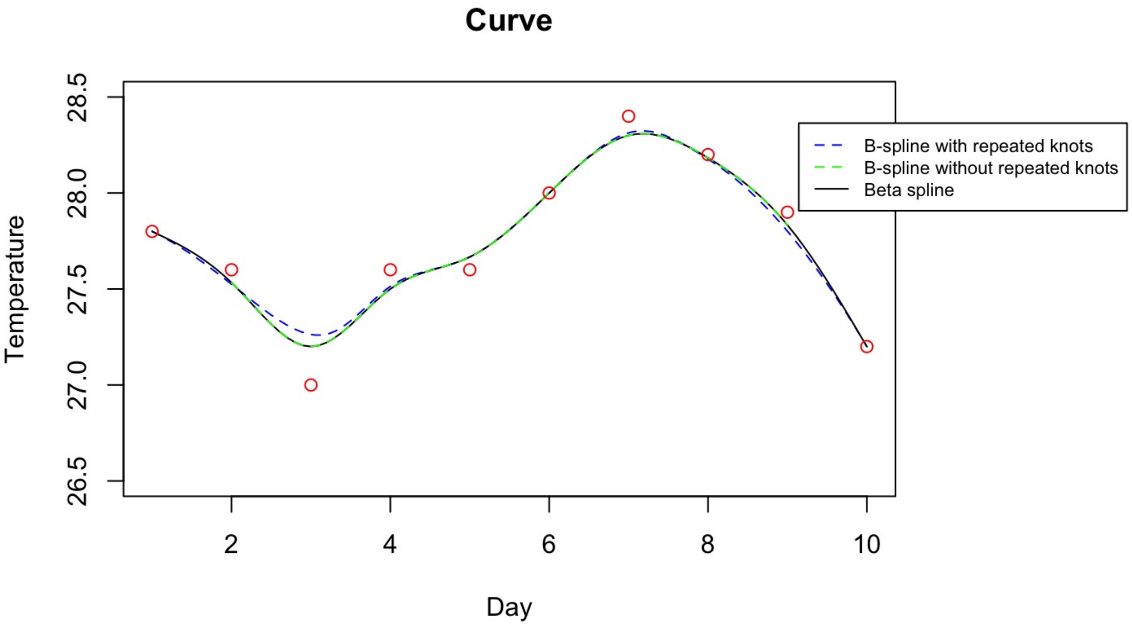


Figure 2. B-spline and beta spline ($\beta_1 = 1$ and $\beta_2 = 0$) curves.

Even periodicity or seasonal variation seems to be one of the salient features of environmental or climate data, such as precipitation and temperature; splines that provide flexibility, like a beta spline or B-spline, are preferred in this study. In a study by [40], the use of the B-spline basis was implied, offering a diverse range of functions to capture the variability in simulated climatic time series. Similarly, [41] employed cubic B-splines as basis functions for weather time series data. B-splines are commonly chosen for their ability to derive higher-order derivative functions, facilitating the analysis of the rate of change in weather variables over time and allowing for a comprehensive analysis of weather variations.

The shape of the beta curve for the first segment, from day 1 until day 4, is shown in Figure 3, illustrating how its parameters vary. Among the curves depicted, the blue one displays the least error compared to the red and black curves. This indicates that the blue curve closely follows the control points of the data. In contrast, the red and black curves have very similar shapes. However, the black curve has a higher error value and is positioned farther from the third control point. Therefore, the red curve, characterized by $\beta_1 = 2$ and $\beta_2 = 1$, emerges as the most suitable representation of the control points in the first segment.

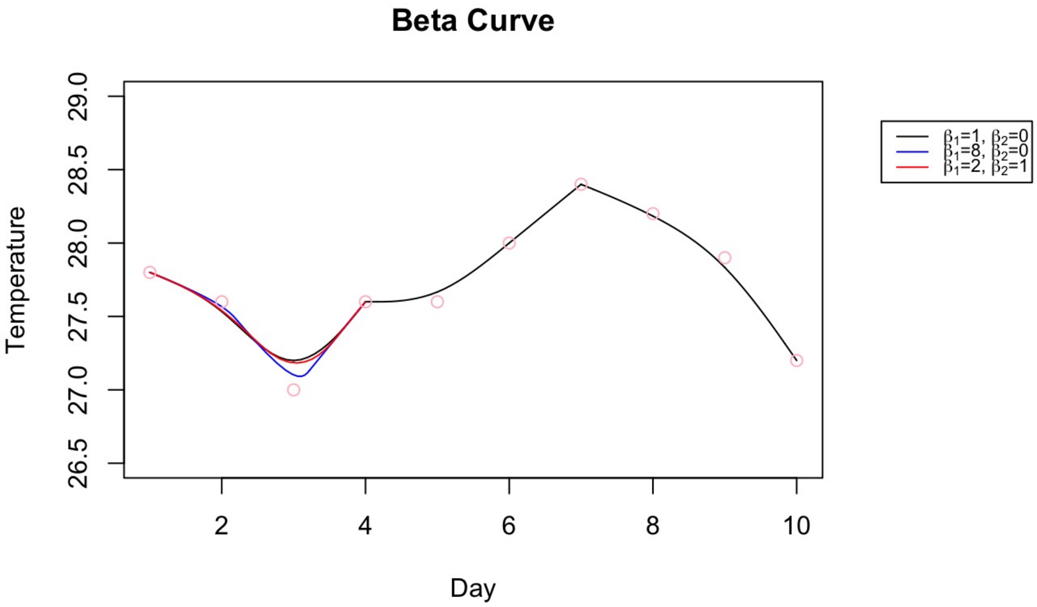


Figure 3. Beta spline curve with different shape parameters at the first curve.

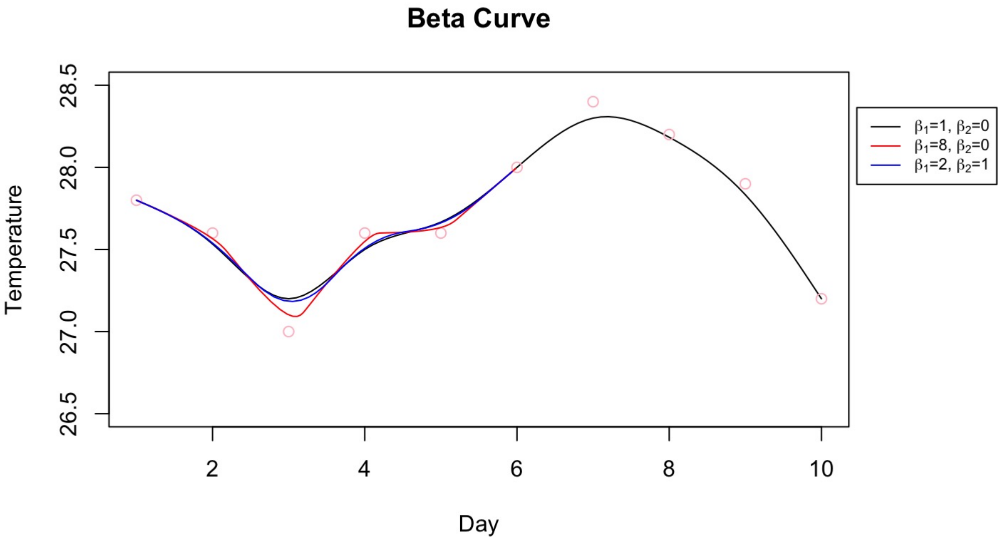


Figure 4. Beta spline curve with different shape parameters.

In Figure 3, sharp curves are observed at the endpoints of each segment, specifically from day 1 to day 4, day 4 to day 7, and day 7 to day 10. These sharp turns exhibit discontinuities at each segment's initial and final points. This discontinuity results from the constant data points at the segment endpoints. While repeated data points at these endpoints are necessary to ensure the beta spline curves reach them, this requirement introduces discontinuities at the segment boundaries. Additionally, another example demonstrates how distinct curves with different shape parameters converge at the same control point. Unlike the previous example, where the entire curve was segmented into four parts, leading to discontinuities at every endpoint, this approach shows that segments with different shape parameters can also achieve good continuity solutions. This is validated in Figure 4.

2.1. *Bias β_1 and tension β_2*

The first shape parameter of the beta spline is β_1, also known as bias. When β_1 is increased, the "velocity" at which one traverses a curve (from left to right, for example) to the right of a joint is greater than the "velocity" just to the left of the joint [37]. This introduces a bias into the curve, where when values exceed one, the unit tangent vector at the joint (which is continuous) exerts a more substantial influence toward the right rather than the left [37].

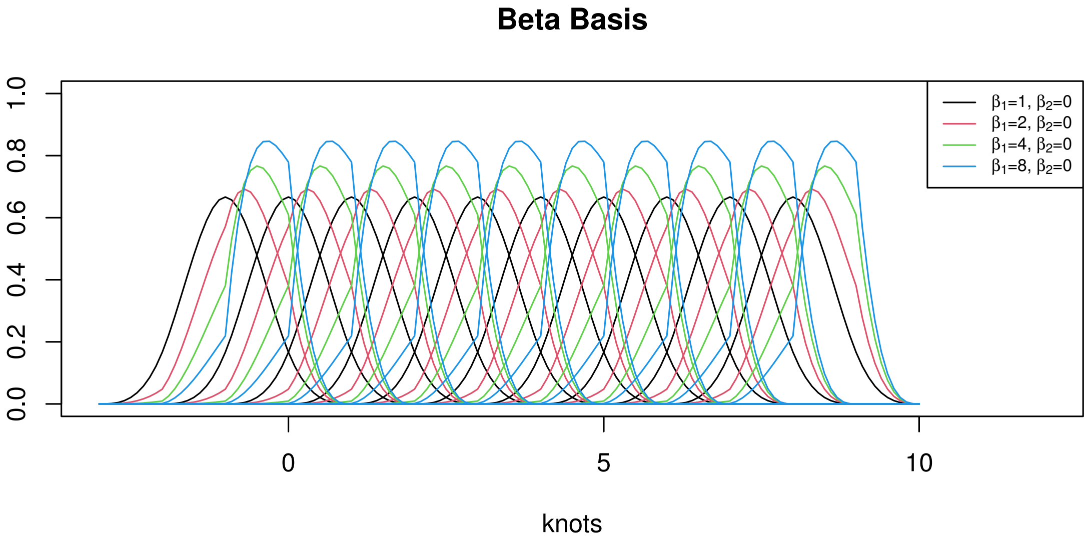


Figure 5. Effect of increasing β_1.

The example can be seen in Figures 5 and 6, where each plot is computed for a distinct value of β_1, which determines the relative magnitude of the slopes to the left and right of each joint. In Figure 5, it can be said that the basis is biased to the right as β_1 increases. The curve also will extend further in the direction of the tangent in the rightmost segment. The effect of increasing the value of β_1 can be seen in how the pink and cyan lines keep expanding to the left with the increasing values of the bias parameter in Figure 6.

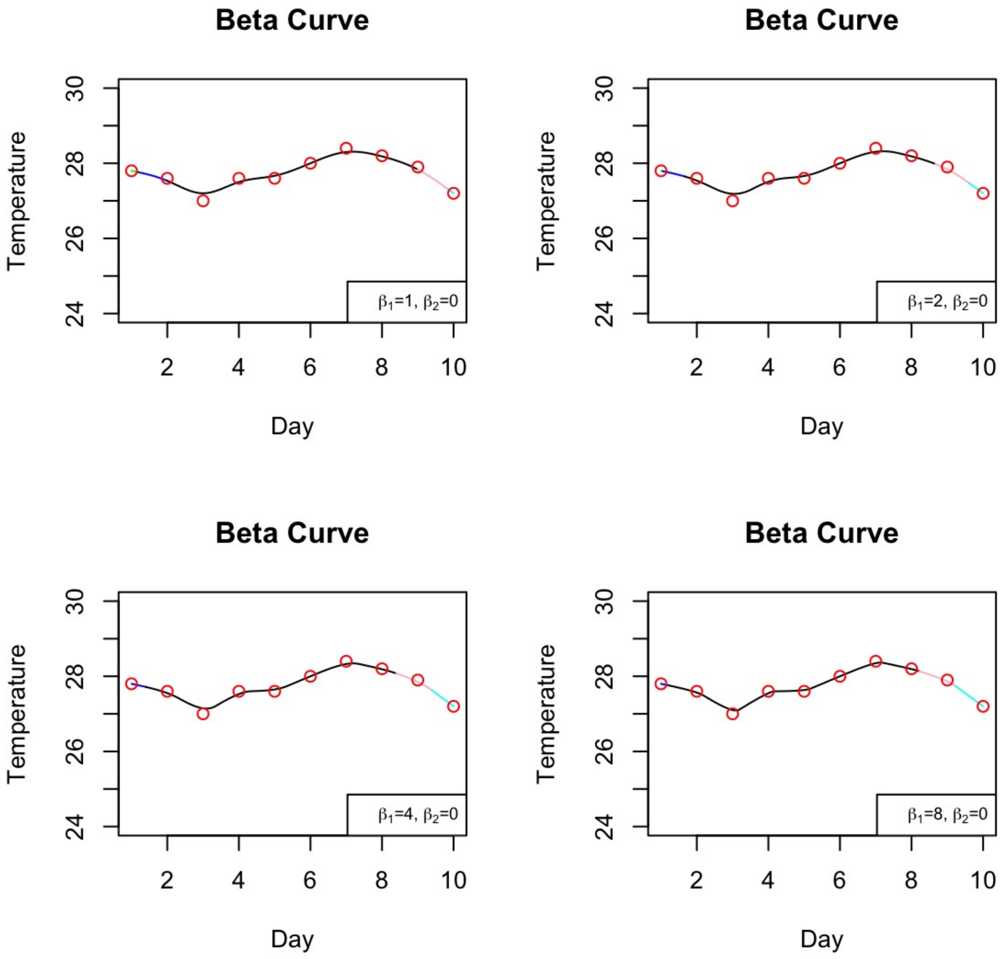


Figure 6. Effect of increasing β_1 on curve.

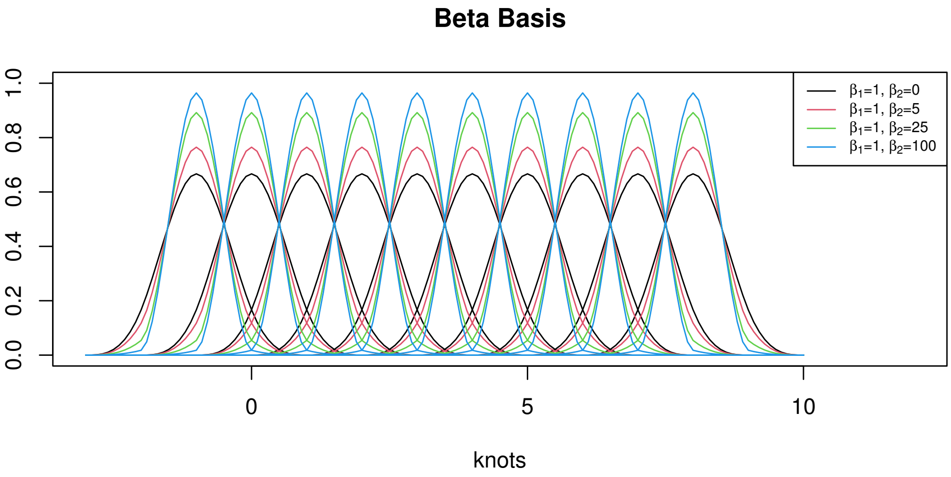


Figure 7. Effect of increasing β_2.

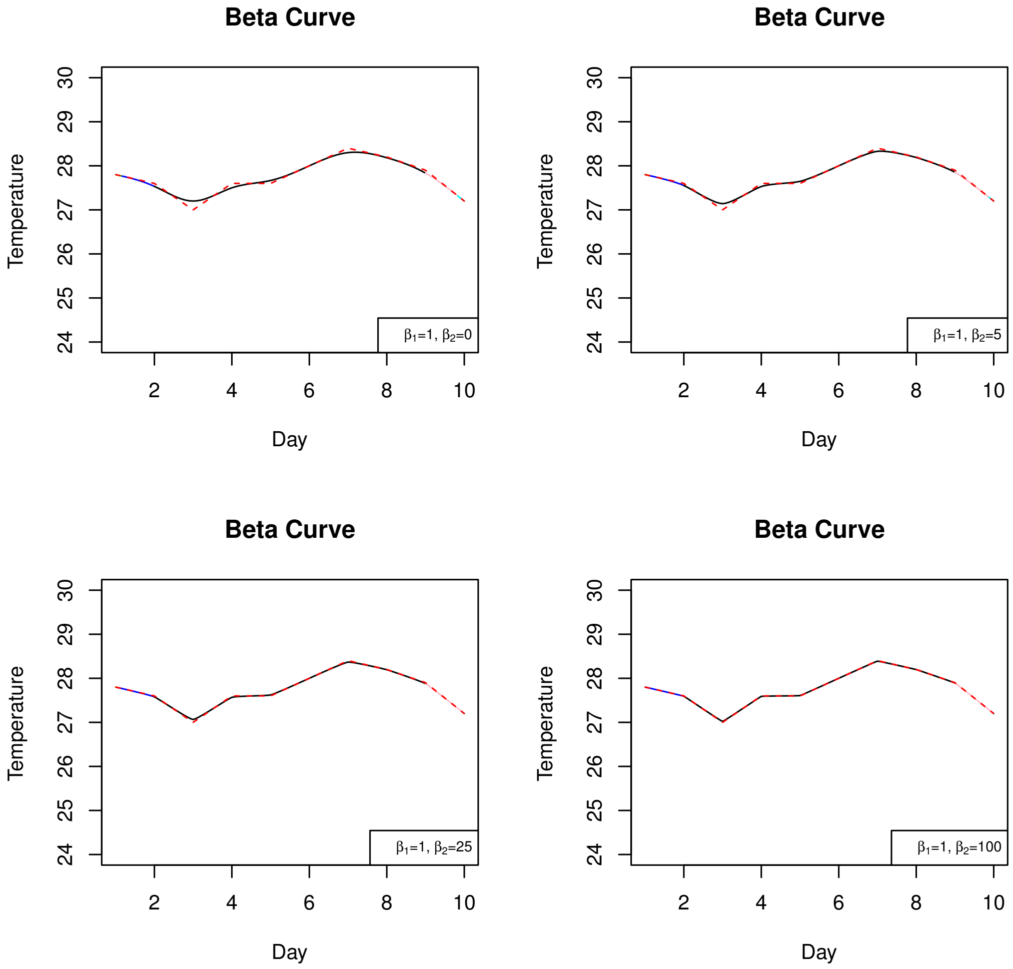


Figure 8. Effect of increasing β_2 on curve.

On the other hand, the second parameter β_2, known as tension, controls the tension inside the curve. By changing the value of β_2, the joint between two segments is moved along a vector that passes through the control vertex. This action is performed simultaneously for all the joints that make up the uniformly shaped curve. As shown in Figure 7, it is apparent that as β_2 increases, the basis function's peak approaches value one, and its "tails," which are located in the support's leftmost and rightmost intervals, approach zero. To illustrate, as the parameter β_2 value increases, each joint is displaced toward its corresponding control vertex, causing the curve to flatten to the control polygon, as shown in Figure 8.

As presented by Barsky, the range of shape parameters encompasses all real numbers. If the values of β_1 are more than 0 and β_2 are greater than or equal to 0, they constitute a basis. This means that they are linearly independent, and every segment of a beta spline curve may be described as a linear combination of these values. Furthermore, the coefficients used to combine these basis functions are distinct since the basis functions do not depend on each other. Thus, each segment of a beta spline curve with $\beta_1 > 0$ and $\beta_2 > 0$ may be expressed uniquely as a linear combination of these basis functions, and the combination coefficients correspond to the control vertices associated with the curve.

The values of negative β can also be used; however, in the context of this study, particularly for data interpolation, negative β values are not feasible. It was found that when applying negative values of β, the basis does not satisfy positivity as reflected on $x = 0$. Furthermore, the curve exhibits an unsuitable irregular shape and fails to depict the data trend accurately. The curve generated from negative values is considered suitable for application in various geometrical contexts but unsuitable for data smoothing purposes.

2.2. End conditions

The beta spline curve typically never starts from a control vertex or any point along the line segment between control points V_0 and V_1. The starting point is inside the convex hull formed by V_0, V_1, and V_2. They are often handled distinctly to exert more precise control over the endpoints. To make the beta spline curves touch the endpoints, the following condition on the vector of control points $[V]$ at the first and second beta spline curves is to be defined as follows: $[V_1] = \begin{pmatrix} v_0 & v_0 & v_0 & v_1 \end{pmatrix}$ and $[V_2] = \begin{pmatrix} v_0 & v_0 & v_1 & v_2 \end{pmatrix}$. The same process goes on to define the second to last of beta spline curves $[V_4] = \begin{pmatrix} v_1 & v_2 & v_3 & v_3 \end{pmatrix}$ and $[V_5] = \begin{pmatrix} v_2 & v_3 & v_3 & v_3 \end{pmatrix}$. The previous definitions are needed to fit five beta spline curve segments in the same control polygon to make them touch both endpoints $[v_0]$ and $[v_3]$. The example of applying and not applying the conditions for endpoints is shown in Figure 9. In the left figure, it can be seen that the curve does not connect the endpoints of the control polygon. The above condition must be fulfilled to make the beta spline curves touch the endpoints; refer to the right figure.

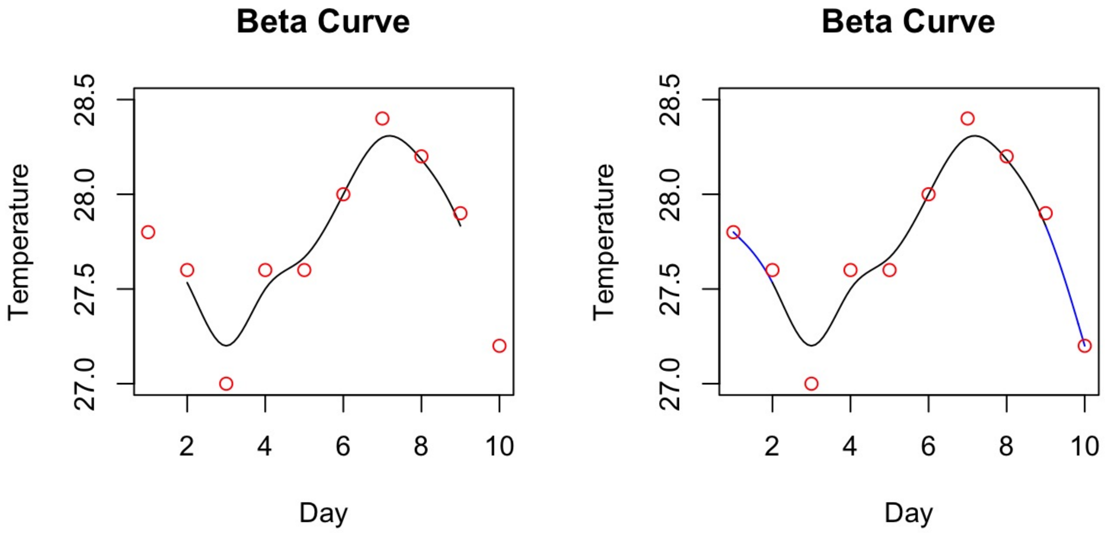


Figure 9. Comparison of beta curve with end conditions.

3. Roughness penalty approach

According to [42], combining the penalized roughness method with the GCV criterion is one strategy for selecting the optimal smoothing parameter. The model's performance can be evaluated using the GCV criterion by comparing the actual and smoothed data. The RPA is one way to control the smoothness of the model by incorporating a penalty term into the objective function. According to [14], this roughness penalty term, R, should be determined and calculated first. Let ϕ be the K-vector of the basis function where K is the total number of basis, and the roughness penalty term, R, can be calculated using Eq (3.1).

$$R = \int D^2\phi D^2\phi'. \tag{3.1}$$

The roughness penalty matrix R defined in Eq (3.1) is composed of the integrals of the outer products of the second derivative D^2 of basis functions ϕ. The notation ϕ' is for the transpose of vector ϕ.

Next, data smoothing was carried out before obtaining the smoothing parameter value, λ, that will be used in the GCV process. Let y_j with $j = 1, \ldots, n$ be the discrete observations, using the model $y_j = x(t_j) + \varepsilon_j$ and basis function expansion of the $x(t)$ in the form of $x(t) = \sum_k^K c_k\phi_k = \Phi' c$, where Φ is the basis function matrix and c is the K-vector of coefficients, and it can be obtained in Eq (3.2).

$$\hat{c} = (\Phi' W \Phi + \lambda R)^{-1}\Phi' W y. \tag{3.2}$$

In Eq (3.2), λ is the smoothing parameter, Φ is the n total observations by K matrix containing the values $\phi_k(t_j)$, and W is the symmetric positive definitive matrix that allows for unequal weighting of squares and products of residuals which assume to be the identity matrix, I if the standard model is assumed. Then, the data-fitting vector $\hat{y}$ is as follows,

$$\hat{y} = \Phi(\Phi' W \Phi + \lambda R)^{-1}\Phi' W y. \tag{3.3}$$

When fitting data with a roughness penalty approach, the smoothing parameter, λ, is employed to regulate the smoothness of the curve. This parameter λ mediates the trade-off between achieving an accurate fit to the data and preserving the smoothness of the function x. To determine the appropriate value of λ, the equation representing the fitted curve using RPA, denoted as $\hat{y}$ in Eq (3.3), is equated with the beta spline curve from Eq (2.1). This beta spline curve is also expressed as a linear combination of the beta spline basis and the control points, where the equation can also be written as $F(t) = \Phi' y = \hat{y}$. The smoothing parameter λ calculation is detailed in Eq (3.4), where its value varies with different values of shape parameters, β_1 and β_2. Let $\hat{y} = F(t)$ and $W = I$, where I is the identity matrix,

$$\begin{aligned} \Phi' y &= \Phi(\Phi'\Phi + \lambda R)^{-1}\Phi' y \\ 1 &= \Phi(\Phi'\Phi + \lambda R)^{-1} \\ (\Phi'\Phi + \lambda R) &= \Phi \\ \lambda R &= \Phi - \Phi'\Phi \\ \lambda &= (\Phi - \Phi'\Phi)R^{-1}. \end{aligned} \tag{3.4}$$

4. Generalized cross-validation

To regulate the model's smoothness and minimize the difference between observed and predicted data, the GCV criteria are used to choose the optimal smoothing parameter. Craven and Wahba developed this algorithm in 1979 [43]. It was initially designed as a simplified alternative to cross-validation, eliminating the necessity for n iterations of smoothing. However, it has demonstrated greater accuracy than cross-validation due to its reduced propensity for under-smoothing. It is also a popular approach for spline smoothing [14]. This study employed the method of GCV to find the best combination values of beta spline curve parameters by optimizing the shape parameters, aiming to achieve a smooth and accurately fitting curve. The equation of GCV is given as follows,

$$GCV = (\frac{n}{n - df(\lambda)})(\frac{SSE}{n - df(\lambda)}). \tag{4.1}$$

In Eq (4.1), $SSE = \sum_n^j (y_j - x(t_j))^2$ and $df(\lambda) = trace[H(\lambda)]$ with $H = \Phi(\Phi'\Phi + \lambda R)^{-1}\Phi'$. Figure 10 depicts the flowchart of the proposed method to improve further understanding. Algorithm 1 outlines the steps required to transform discrete temperature data points into a smooth, continuous curve by applying cubic beta splines, ensuring the optimal representation of the underlying temperature trends.

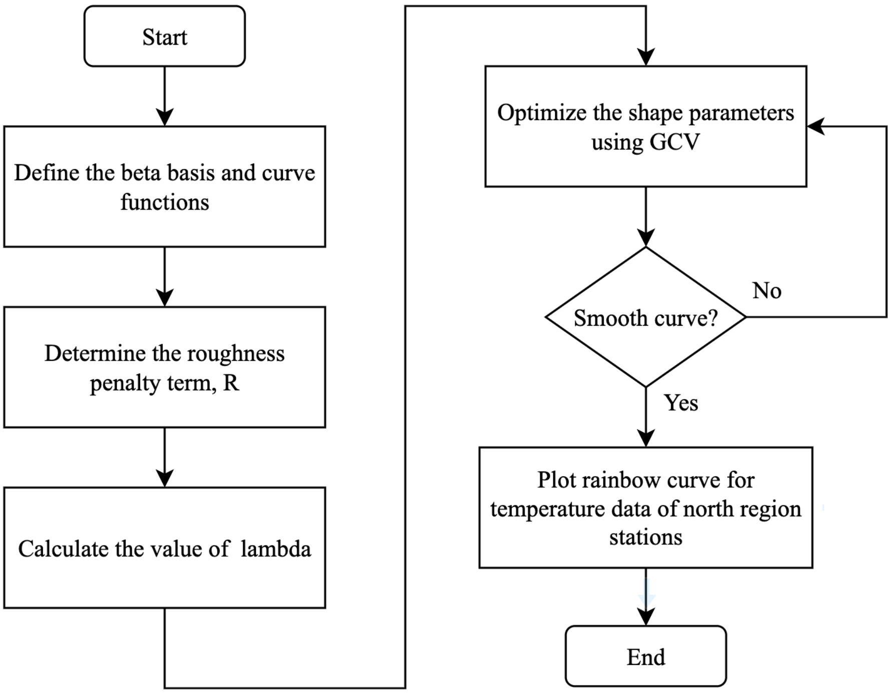

Start
Define the beta basis and curve functions
Determine the roughness penalty term, R
Calculate the value of lambda
Optimize the shape parameters using GCV
Smooth curve?
No
Yes
Plot rainbow curve for temperature data of north region stations
End

Figure 10. Flowchart of the proposed data smoothing technique.

Algorithm 1 Cubic beta curve smoothing process

(i) Input data points in vector $[V]$.

(ii) Compute the values of beta spline basis matrix $[M]$ as discussed in Section 2.

(iii) Calculate the value of roughness penalty term, $R = \int D^2\phi D^2\phi'$.

(iv) Evaluate the values of smoothing parameter, λ as formulated in Eq 3.4.

(v) Compute the GCV errors as in Eq 4.1 for each combination of β_1 and β_2.

(vi) Construct the cubic beta curve, $F(t)$, using the optimal shape parameter values obtained from GCV.

5. Real-life application

The proposed approach was utilized with the temperature data from meteorological stations in northern Peninsular Malaysia for January 2022. This study faced the challenge of determining the optimal combination of parameters as changes in β_1 and β_2 can modify the curve's shape independently of the control vertices. Thus, GCV was utilized to determine the optimal value of the shape parameters. The range was selected for $\beta_1 > 0$ and $\beta_2 \geq 0$ to ensure linear independence. The ranges were adjusted to study their impact on the optimized shape parameter values. The method's approximation errors are compared with those using standard shape parameter values, $\beta_1 = 1$ and $\beta_2 = 0$. Before optimization, underfitting and overfitting can be assessed by evaluating the smoothness of the curve. An over-smoothed curve that fails to capture the underlying pattern of the data points indicates underfitting. Conversely, an under-smoothed curve that lacks smoothness and captures unnecessary data details indicates overfitting. It is crucial to plot the best-fitted curve with low error and optimal smoothness during the transformation process, as overfitting and underfitting curves impact subsequent FDA procedures differently. Overfitting introduces irrelevant details, complicates result interpretation, and increases computational costs, while underfitting removes useful information that cannot be recovered later [44].

For GCV, two values are crossed over, and the procedure consists of exploring several combinations of the crossed parameter values, computing their GCV error for each of them, and choosing the combination that yields the low error with optimal smoothness. Based on the GCV error values, a high error indicates that the curve is over-smoothed and underfitting the data. Conversely, a very low error suggests that the curve is under-smoothed and overfitting the data. Therefore, in this study, the optimal curve is determined to lie between these extremes of overfitting and underfitting. The guidelines followed by this study on the effect of choosing different parameters for smoothing have also been recently practiced by [45] and were well-demonstrated by [44].

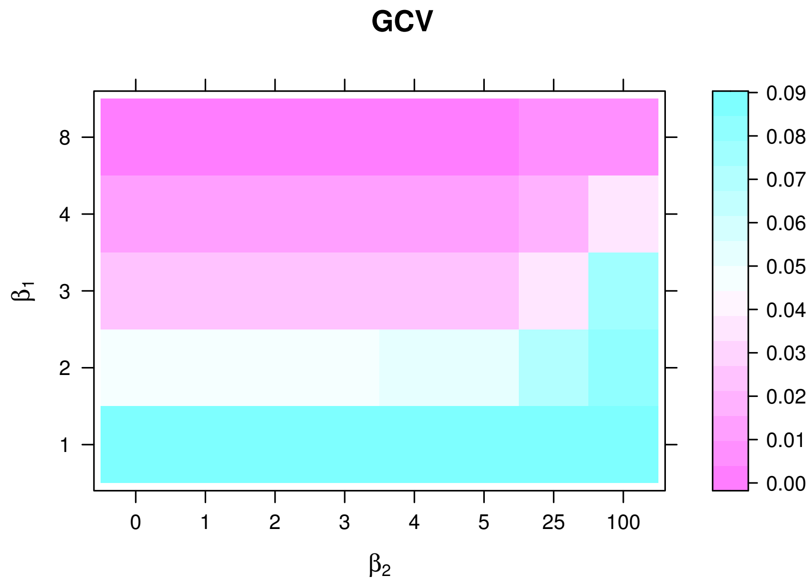


Figure 11. Color grid showing the GCV error for several (β_1, β_2) combinations.

The result of this approach is presented in Figure 11, with the darker pink shade representing the lowest error and the darker blue shade representing the highest error. The color grid displays the GCV values for each combination of shape parameters. As shown in Figure 11, the minimum GCV error in the grid corresponds to $\beta_1 = 8$ and $\beta_2 = 0$, represented by the darkest shade of pink. Figure 12 illustrates a curve derived from solving Eq (2.1) using the provided values. The curve is not smooth, exhibits sharp edges, and closely follows the control polygon, capturing all data details. In the meantime, the highest value of GCV error is produced by the combination of $\beta_1 = 1$ and $\beta_2 = 0$. It is illustrated in Figure 13 that the curve that contains these values is over-smoothed. Because the curve does not cover all the points, it may overlook some significant aspects of the data. Since the objective is not just to create a smooth curve but rather a curve that is best suited to the data, it is unreasonable to represent the data using the aforementioned curves.

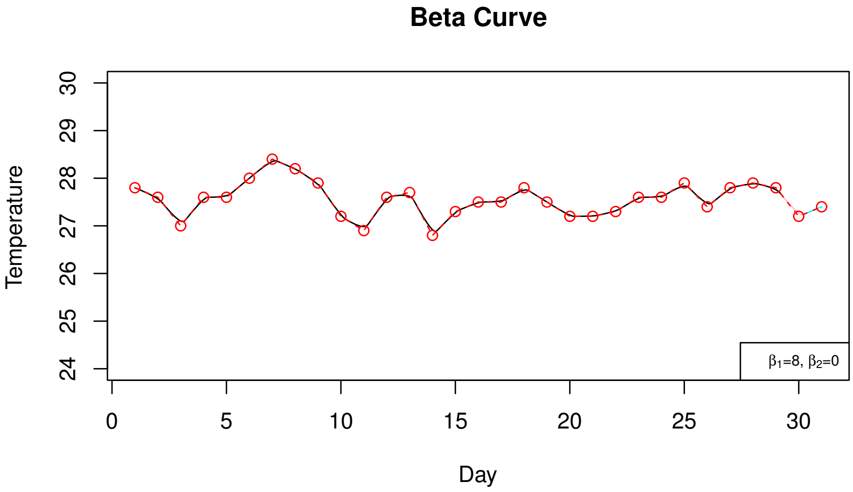


Figure 12. Temperature curve with the lowest GCV error.

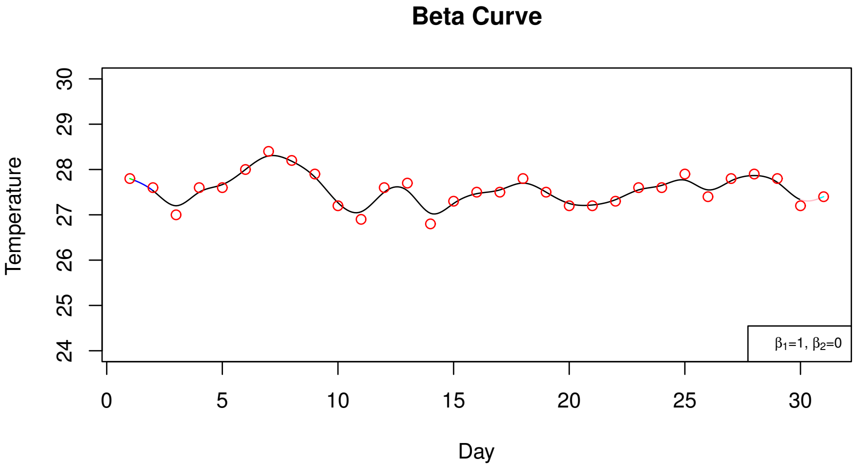


Figure 13. Temperature curve with the highest GCV error.

In Figure 14, the comparison of the curves with shape parameters that yield the highest and lowest errors are plotted. The graph shows that the blue curve does not capture almost all the extreme minimum and maximum temperature data points, such as on days 3, 7, 11, 14, 18, 25, 26, and 30. Meanwhile, the green curve that produces the lowest error almost touches all the extreme points, capturing every data detail. Curve shape adjustments can be made on these segments with extreme data points by altering the value of shape parameters at specific intervals or curve segments.

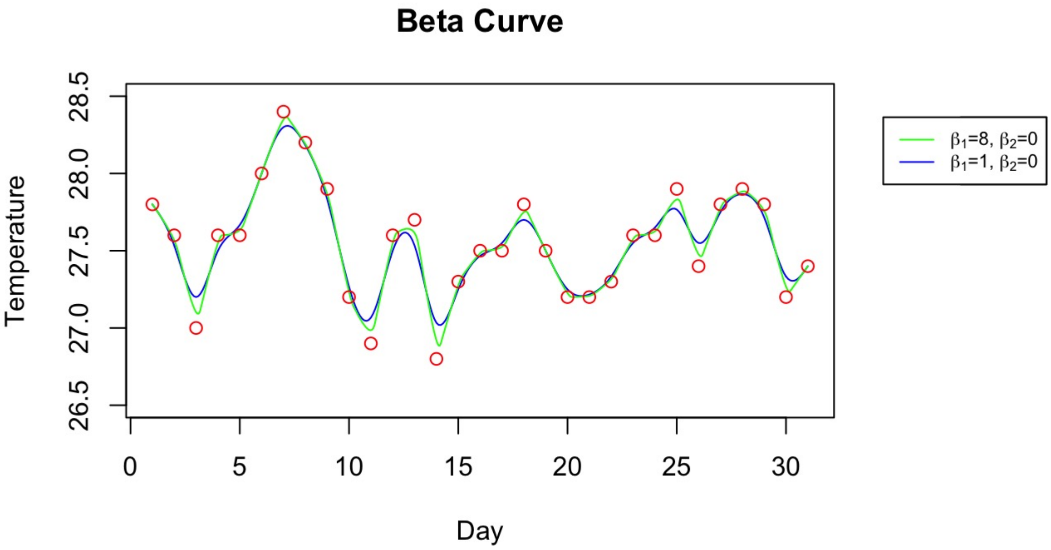


Figure 14. Comparison between the curves with the highest and lowest GCV error.

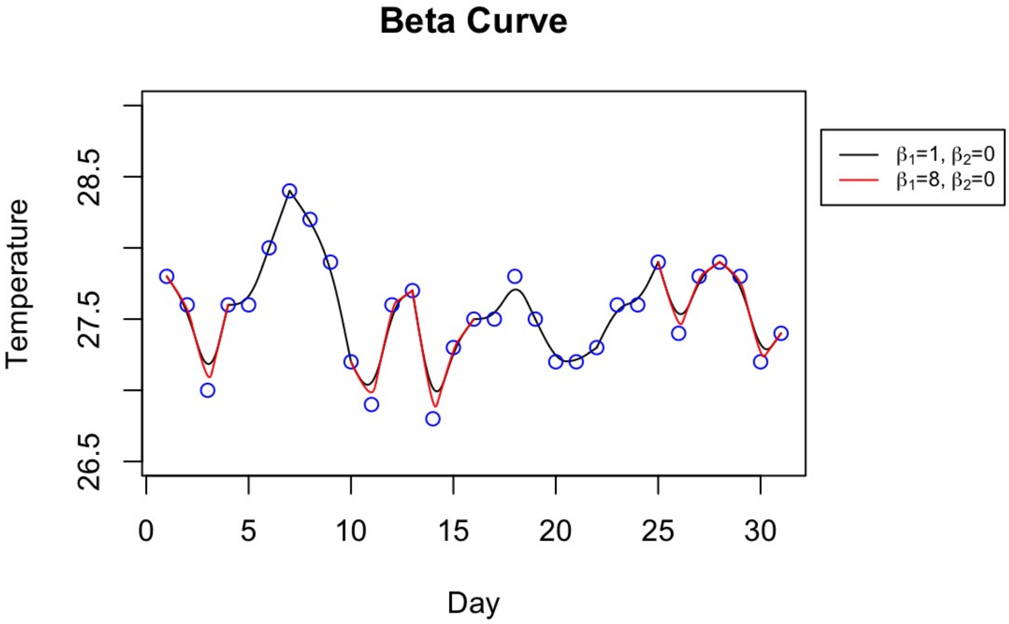


Figure 15. Improved beta curve segments with adjusted shape parameter values.

Figure 15 illustrates how the beta curve enhances specific segments, particularly those with extreme values, by adjusting the shape parameters at those segments. However, the figure reveals discontinuities at each joint between curves. Consequently, an improved curve is presented in Figure 16, which effectively represents the data while also maintaining geometric continuity. This improvement utilizes the restricted form of quintic Hermite interpolation introduced by [37], allowing the definition of distinct shape parameters at each joint without compromising geometric continuity. This method assigns new parameters, α_1 and α_2, to β_1 and β_2, which are the shape parameters at the joint between two curve segments. The beta basis segments are constructed such that β_1 and β_2 are functions of the knot value. These functions interpolate smoothly between α_1 and α_2 at each end of a segment. Consequently, the curve maintains G^2 continuity, ensuring smooth transitions without abrupt changes in direction or curvature at the segment joints. This approach guarantees that the shape parameters transition smoothly from one segment to the next, preserving the overall smoothness of the curve while allowing for specific adjustments at the joints.

Each segment of the curve in Figure 16 is adjusted locally by controlling the values of α_1 and α_2. The legend of the graph shows the value of each α used to control a total of 29 curve segments locally. Additionally, a curve using only one global pair of shape parameter values is also plotted in the figure, as indicated by the black line. The curve generated with global values of $\beta_1 = 2$ and $\beta_2 = 1$, optimized by GCV, shows minimal differences in variation compared to the locally adjusted segments. This demonstrates the competency of the proposed method for data smoothing, highlighting its flexibility and efficiency to automatically smooth curves with large datasets by finding the global value of parameters. Moreover, it can be further enhanced by locally adjusting the value of shape parameters at particular segments.

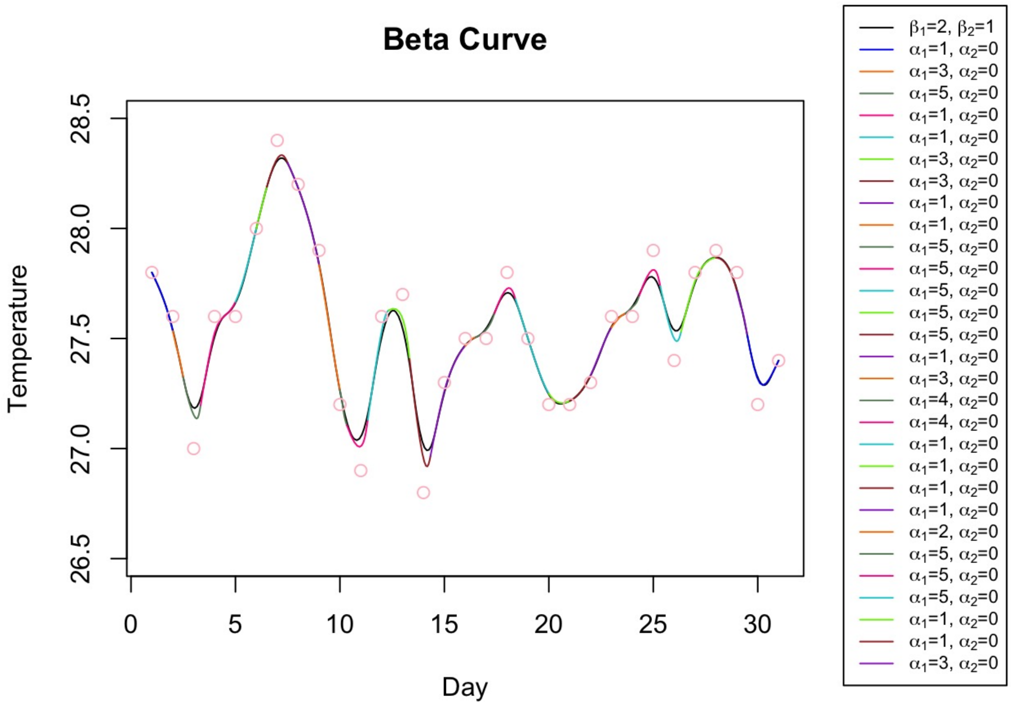


Figure 16. Improved beta curve segments with adjusted shape parameter values while maintaining geometric continuity.

Even though knot insertion and deletion can be performed with standard cubic B-spline fitting to have a local control on the curve, this technique has several disadvantages. Knot insertion in B-spline fitting produces additional control points and necessitates the individual selection of knot values, making the process more tedious and time-consuming. Consequently, beta splines are more efficient in this context. Given that this study does not aim to model overfitted or underfitted curves, finding an optimal solution to yield smoother curves is crucial. A possible combination could be $\beta_1 = 2$ and $\beta_2 = 1$, corresponding to Figure 17, where a good midpoint seems to be reached. Looking at the combination on the grid of Figure 11, it can be seen that by increasing the value of β_1, the GCV error decreases while it increases with the higher value of β_2. The optimal value can be referred to as the light shade of pink and blue color grid. It shows that $\beta_1 = 2$ with β_2 from 0 to 4 has almost the same GCV error and curves. According to the minimum complexity principle, the solutions of $\beta_1 = 2$ and $\beta_2 = 1$ are selected and then applied to smooth all northern stations' temperature data.

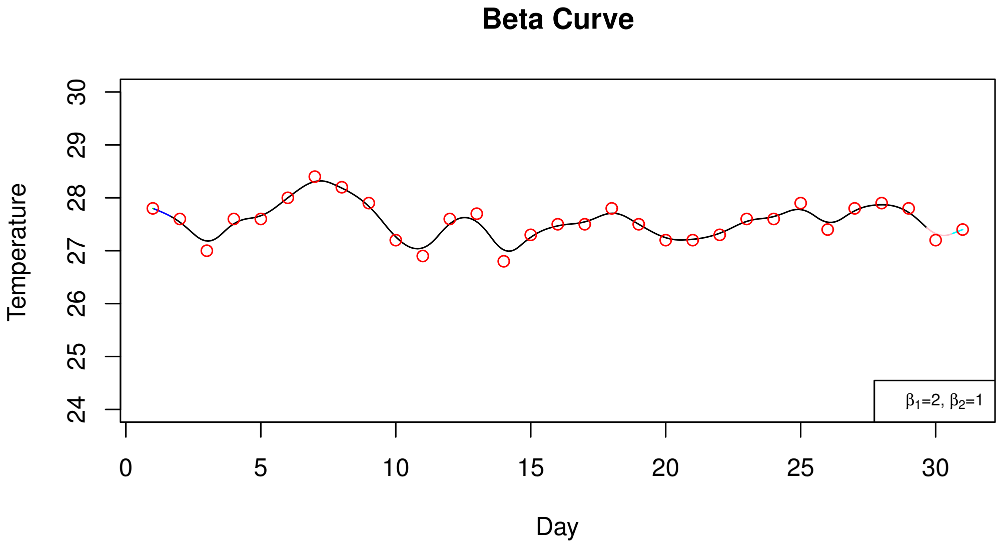


Figure 17. Temperature curve with optimal GCV error.

To evaluate the effectiveness and accuracy of the proposed method, a comparison was conducted with standard smoothing techniques such as moving average, simple exponential smoothing (SES), and double exponential smoothing (DES). In Figure 18, the results for SES with $\alpha = 0.5793$, DES with $\alpha = 0.5573$ and $\beta = 0.0001$, and moving average with an order of 4 are presented. These parameter values were automatically estimated using nonlinear minimization of the observed data using the R built-in function, as it is more robust and objective than choosing parameters based on subjective experience. The figure clearly demonstrates that the proposed method closely follows the shape of the data's control polygon. In contrast, the moving average, SES, and DES methods exhibit some lag in capturing the data's pattern. This comparison highlights the superior performance of the proposed method in accurately reflecting the underlying trends and variations in the data over traditional smoothing techniques.

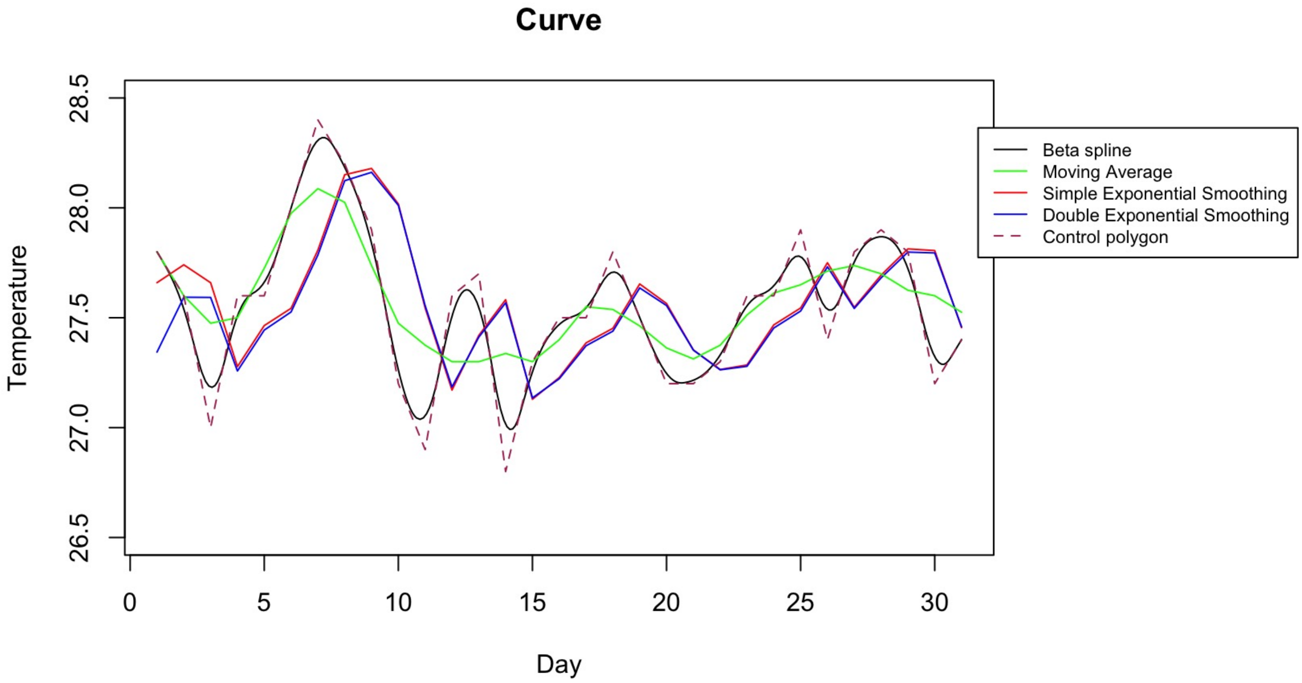


Figure 18. Comparison of the proposed and standard smoothing methods.

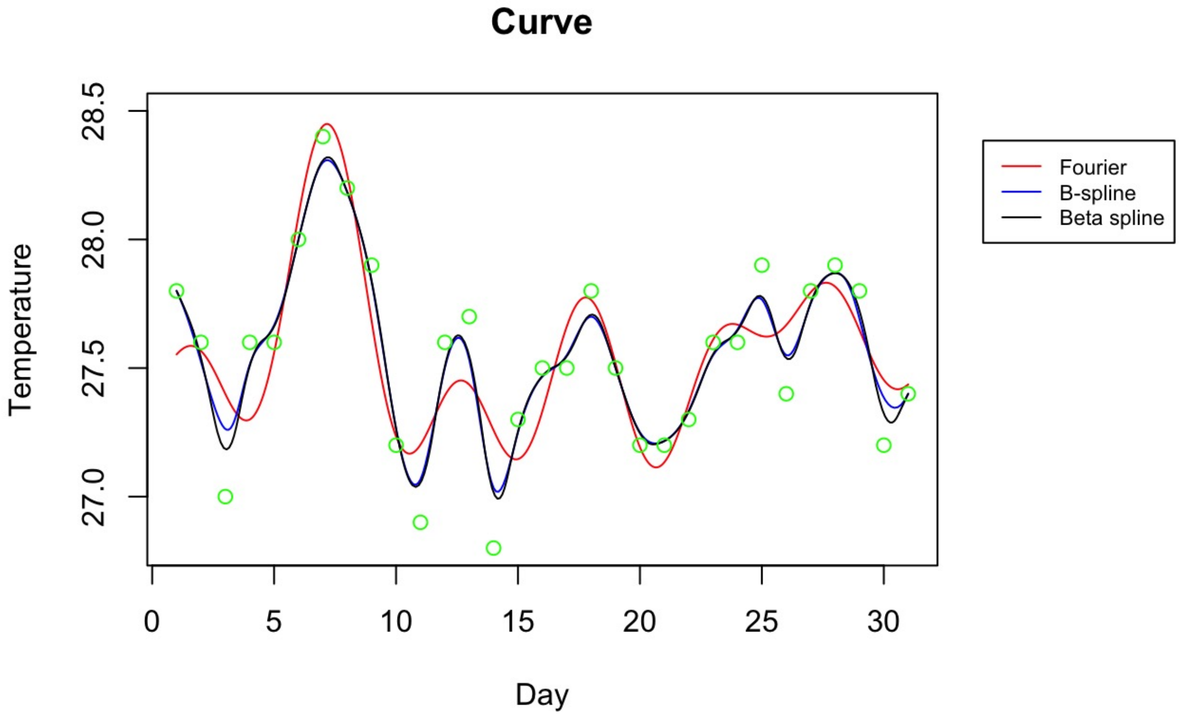


Figure 19. Comparison of the proposed and existing basis functions.

Another comparison of smoothed curves using the proposed method, B-spline, and Fourier is presented in Figure 19. The smooth fit in the figure was achieved with 13 Fourier basis functions, a B-spline basis with repeated knots, and a beta spline basis with optimal parameters $\beta_1 = 2$ and $\beta_2 = 1$. For the Fourier series, the number of basis functions was determined by adding basis functions until the estimated variance ceased to decrease significantly, following the guidelines of [14]. Figure 20 shows how variance estimate decreases to a value by the time 13 Fourier basis functions were used for smoothing the temperature data. Although lower estimated variance values were found in some instances, they were not selected to avoid overfitting the data. The figure demonstrates that the proposed method produced a superior curve compared to other methods, effectively capturing the data pattern without overfitting unnecessary details. The beta spline's ability to adjust the curve by manipulating shape parameter values provides flexibility, surpassing other methods that require adjustments to the number of basis functions or knot values to achieve the desired curve.

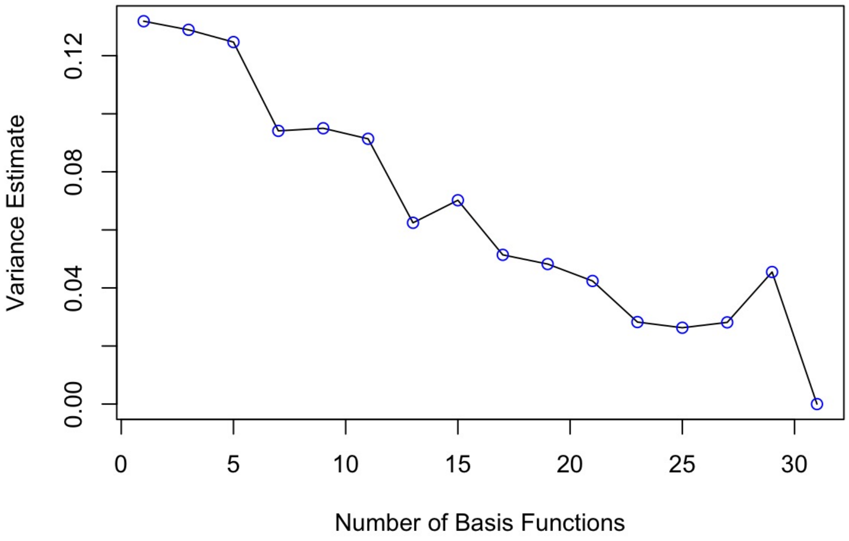


Figure 20. The relation between the number of Fourier basis functions and the unbiased estimate of the residual variance fitting the temperature data.

Despite periodicity being a typical characteristic of weather data, for which the Fourier basis is often chosen, the smoothed curves using B-spline and beta spline yielded better results than the Fourier basis, as demonstrated in Figure 19. This outcome aligns with the findings of [27], where Fourier basis functions were initially used for data smoothing due to the periodic nature of air temperature series. However, B-splines yielded slightly better forecast results because of their ability to balance model flexibility and overfitting, capturing complex patterns within the data to ensure accurate forecasts. Similar methods were observed in the studies by [25, 40, 41], which also employed B-splines for smoothing weather data instead of Fourier basis functions.

Figure 21 shows temperature data from the main meteorological stations in northern Peninsular Malaysia. The curves were smoothed using a cubic beta spline with $\beta_1 = 2$ and $\beta_2 = 1$. The thick black line indicates the mean temperature of the northern region of Malaysia. The GCV method successfully determined the optimal values of the parameters, as the curves show a precisely fitted curve and offer the most optimal representation of the temperature data from each station. Identifying the optimal parameters is essential because they directly influence the accuracy of the curve when fitting the data, therefore improving data analysis such as forecasting. The proposed method can be extended to forecasting using various approaches, including functional regression models.

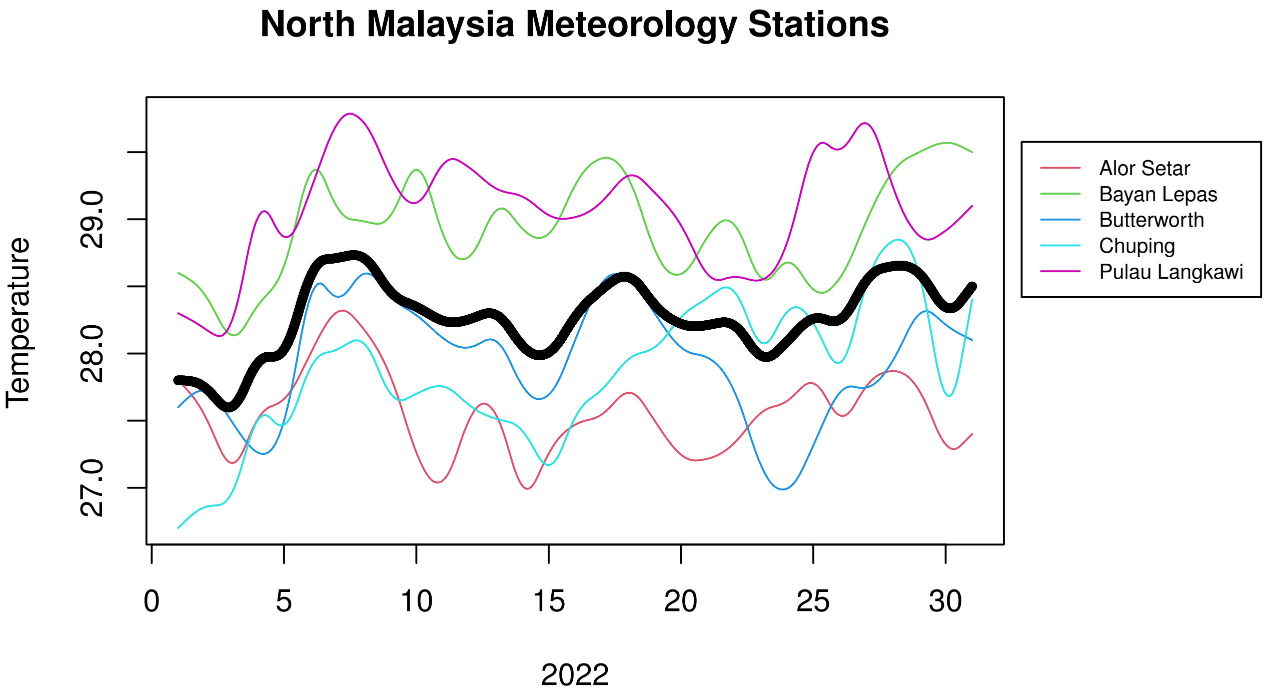


Figure 21. Temperature curves at the north region of Malaysia.

Data smoothing is performed as a preliminary step to prepare the data for forecasting. For example, in the nonparametric FDA model, the continuous or functional term is considered a predictor variable. The smoothed observations generated through the smoothing process are used within the FDA framework for regression modeling to make advanced forecasts. Other than regression, multiple forecasting methods have been proposed within the FDA framework and can be considered for future research. By smoothing the data and constructing functions, the model can better capture underlying patterns and relationships, leading to more accurate forecasts. The problems of underfitted or overfitted curves and the undesirable result of failing to capture essential details of the data or including excessively unnecessary information can be avoided by implementing a flexible smoothing technique.

6. Conclusions

This study presents a novel approach to data smoothing within the FDA framework to improve forecasting. Unlike commonly used basis such as Fourier or B-spline, beta spline, a spline with two shape parameters, was employed to transform discrete data into a functional form. This methodology, previously unexplored in the FDA framework and climate studies, offers several advantages in controlling curve shape by manipulating these parameters. Through experimentation, it was demonstrated that beta spline's flexibility is particularly effective in capturing intricate details in complex climate data, including extreme events. Additionally, an enhanced optimization methodology using generalized cross-validation was developed to determine the optimal combination of shape parameters for data representation. By manipulating these parameters, the proposed method offers a more flexible approach to data smoothing. The GCV color grid facilitates efficient identification of the best parameter combination, enabling the prevention of overfitting or underfitting by observing the error value color indicator. This technique is expected to perform effectively with various types of time series data.

Author contributions

Wan Anis Farhah Wan Amir: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualisation, Writing - original draft; Md Yushalify Misro: Data Curation, Funding acquisition, Project administration, Supervision, Writing - review & editing; Mohd Hafiz Mohd: Supervision, Writing - review & editing. All authors have read and approved the final version of the manuscript for publication.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence tools in the creation of this article.

Acknowledgments

This research was supported by the Ministry of Higher Education Malaysia through the Fundamental Grant Scheme (FRGS/1/2023/STG06/USM/03/4) and the School of Mathematical Sciences, Universiti Sains Malaysia. The authors express their sincere gratitude to the Malaysian Meteorological Department for providing the data used in this study.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. Y. Xu, *Functional Data Analysis*, London: Springer, 2023. https://doi.org/10.1007/978-1-4471-7503-2_4

2. P, Hall, M, Hosseini-Nasab, On Properties of Functional Principal Components Analysis, *J. R. Stat. Soc. Ser. B: Stat. Methodol.*, **68** (2006), 109–126. https://doi.org/10.1111/j.1467-9868.2005.00535.x

3. W. Seo, Functional principal component analysis for cointegrated functional time series, *J. Time Ser. Anal.*, **45** (2023), 320–330. https://doi.org/10.1111/jtsa.12707

4. O. A. Montesinos López, A. Montesinos López, J. Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Cham: Springer, 2022. https://doi.org/10.1007/978-3-030-89010-0

5. H. Hullait, D. S. Leslie, N. G. Pavlidis, S. King, Robust Function-on-Function Regression, *Technometrics*, **63** (2020), 396–409. https://doi.org/10.1080/00401706.2020.1802350

6. J. O. Razo-De-Anda, L. L. Romero-Castro, F. Venegas-Martínez, Contagion Patterns Classification in Stock Indices: A Functional Clustering Analysis Using Decision Trees, *Mathematics*, **11** (2023), 2961. https://doi.org/10.3390/math11132961

7. F. Centofanti, A. Lepore, B. Palumbo, Sparse and smooth functional data clustering, *Stat. Pap.*, **65** (2024), 795–825. https://doi.org/10.1007/s00362-023-01408-1

8. J. A. Arias-López, C. Cadarso-Suárez, P. Aguiar-Fernánde, Computational Issues in the Application of Functional Data Analysis to Imaging Data, *Lect. Notes Comput. Sci.*, **42** (2021), 630–638. https://doi.org/10.1007/978-3-030-86960-1_46

9. C. Tang, T. Wang, P. Zhang, Functional data analysis: An application to COVID-19 data in the United States in 2020, *Quant. Bio.*, **10** (2022), 172–187. https://doi.org/10.15302/J-QB-022-0300

10. C. Zhang, H. Lin, L. Liu, J. Liu, Y. Li, Functional Data Analysis with Covariate-Dependent Mean and Covariance Structures, *Biometrics*, **79** (2023), 2232–2245. https://doi.org/10.1111/biom.13744

11. I. Shah, P. Mubassir, S. Ali, O. Albalawi, A functional autoregressive approach for modeling and forecasting short-term air temperature, *Front. Environ. Sci.*, **12** (2024), 1411237. https://doi.org/10.3389/fenvs.2024.1411237

12. V. Villani, E. Romano, J. Mateu, Climate model selection via conformal clustering of spatial functional data, *Environ. Ecol. Stat.*, **31** (2024), 365–385. https://doi.org/10.1007/s10651-024-00616-8

13. A. Palummo, E. Arnone, L. Formaggia, L. M. Sangalli, Functional principal component analysis for incomplete space-time data, *Environ. Ecol. Stat.*, **31** (2024), 555–582. https://doi.org/10.1007/s10651-024-00598-7

14. J. O. Ramsay, B. W. Silverman, *Functional Data Analysis*, 2 Eds., New York: Springer, 2005. https://doi.org/10.1007/b98888

15. M. A. Hael, Unveiling air pollution patterns in Yemen: A spatial-temporal functional data analysis, *Environ. Sci. Pollut. Res.*, **30** (2023), 50067–50095. https://doi.org/10.1007/s11356-023-25790-3

16. M. Gong, R. O'Donnell, C. Miller, M. Scott, S. Simis, S. Groom, et. al, Adaptive smoothing to identify spatial structure in global lake ecological processes using satellite remote sensing data, *Spat. Stat.*, **50** (2022), 100615. https://doi.org/10.1016/j.spasta.2022.100615

17. R. Raturi, Large Data Analysis via Interpolation of Functions: Interpolating Polynomials vs Artificial Neural Networks, *Amer. J. Intell. Syst.*, **8** (2018), 6–11. https://doi.org/10.5923/j.ajis.20180801.02

18. N. A. Mazelan, J. Suhaila, Exploring rainfall variabilities using statistical functional data analysis, *IOP Conf. Ser.: Earth Environ. Sci.*, **1167** (2023), 012007. https://doi.org/10.1088/1755-1315/1167/1/012007

19. C. Sözen, Y. Öner, The investigation of temperature data in Turkey's Black Sea Region using functional data analysis, *J. Appl. Stat.*, **49** (2021), 2403–2415. https://doi.org/10.1080/02664763.2021.1896683

20. J. Baz, J. Davis, L. Han, C. Stracke, The value of smoothing, *J. Portfolio Manag.*, **48** (2022), 73–85. https://doi.org/10.3905/jpm.2022.1.399

21. A. Falini, F. Mazzia, C. Tamborrino, Spline based Hermite quasi-interpolation for univariate time series, *Discrete Cont. Dyn. Syst. - S*, **15** (2022), 3667–3688. https://doi.org/10.3934/dcdss.2022039

22. L. Brugnano, D. Giordano, F. Iavernaro, G. Rubino, An entropy-based approach for a robust least squares spline approximation, *J. Comput. Appl. Math.*, **443** (2024), 115773. https://doi.org/10.1016/j.cam.2024.115773

23. M. Spreafico, F. Ieva, M. Fiocco, Modelling time-varying covariates effect on survival via functional data analysis: Application to the MRC BO06 trial in osteosarcoma, *Stat. Methods Appl.*, **32** (2023), 271–298. https://doi.org/10.1007/s10260-022-00647-0

24. A. Rahman, D. Jiang, Regional and temporal patterns of influenza: Application of functional data analysis, *Infect. Dis. Modell.*, **6** (2021), 1061–1072. https://doi.org/10.1016/j.idm.2021.08.006

25. M. Rangata, S. Das, M. Ali, Analysing Maximum Monthly Temperatures in South Africa for 45 years Using Functional Data Analysis, *Adv. Decis. Sci.*, **24** (2020), 1–27.

26. U. Beyaztas, S. Q. Salih, K.-W. Chau, N. Al-Ansari, Z. M. Yaseen, Construction of functional data analysis modeling strategy for global solar radiation prediction: Application of cross-station paradigm, *Eng. Appl. Comput. Fluid Mech.*, **13** (2019), 1165–1181. http://doi.org/10.1080/19942060.2019.1676314

27. S. Curceac, C. Ternynck, T. B. Ouarda, F. Chebana, S. D. Niang, Short-term air temperature forecasting using Nonparametric Functional Data Analysis and SARMA models, *Environ. Modell. Software*, **111** (2019), 394–408. http://doi.org/10.1016/j.envsoft.2018.09.017

28. M. Ammad, M. Y. Misro, A. Ramli, A novel generalized trigonometric Bézier curve: Properties, continuity conditions and applications to the curve modeling, *J. Amer. Math. Soc.*, **194** (2022), 744–763. http://doi.org/10.1016/j.matcom.2021.12.011

29. S. A. A. A. Said Mad Zain, M. Y. Misro, K. T. Miura, Generalized Fractional Bézier Curve with Shape Parameters, *Mathematics*, **9** (2021), 2141. https://doi.org/10.3390/math9172141

30. B. A. Barsky, *The Beta-Spline: A Local Representation based on Shape Parameters and Fundamental Geometric Measures*, PhD thesis, The University of Utah, 1981.

31. B. A. Barsky, Rational Beta-splines for representing curves and surfaces, *IEEE Comput. Graph. Appl.*, **13** (1993), 24–32. http://doi.org/10.1109/38.252550

32. N. A. Hadi, A. Ibrahim, F. Yahya, J. M. Ali, A Comparative Study on Cubic Bezier and Beta-Spline Curves, *Mathematika*, **29** (2013), 55–64.

33. B. Sambhunath, C. L. Brian, *Bézier and Splines in Image Processing and Machine Vision*, London: Springer, 2008. https://doi.org/10.1007/978-1-84628-957-6

34. N. A. Hadi, N. S. M. Kamal, H. Nordin, Computational Method for Digital Khat Calligraphy Using Beta-Spline Curve Fitting, *ASM Sc. J.*, **13** (2020). https://doi.org/10.32802/asmscj.2020.sm26(5.8)

35. S. A. Suliman, N. A. Hadi, Optimizing the Shape Parameters of Beta-Spline Using Particle Swarm Optimization, *Int. J. Eng. Technol.*, **7** (2018), 93–97. http://doi.org/10.14419/ijet.v7i4.33.23492

36. M. S. A. Halim, N. A. Hadi, H. Sulaiman, S. Abd Halim, An algorithm for beta-spline surface reconstruction from multi slice CT scan images using MATLAB pmode, *2017 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2017, 1–6. http://doi.org/10.1109/ISCAIE.2017.8074939

37. B. A. Barsky, J. C. Beatty, Local Control of Bias and Tension in Beta-splines, *ACM Trans. Graph.*, **2** (1983), 109–134. http://doi.org/10.1145/357318.357321

38. B. A. Barsky, *Computer Graphics and Geometric Modeling Using Beta-splines*, Berlin, Heidelberg: Springer, 1988. https://doi.org/10.1007/978-3-642-72292-9

39. B. A. Barsky, J. C. Beatty, Varying the Betas in Beta-splines, Technical Report UCB/CSD-83-112, EECS Department, University of California, Berkeley, 1982. Available from: `https://digicoll.lib.berkeley.edu/record/137388/files/CSD-83-112.pdf`.

40. E. Holtanová, T. Mendlik, J. Koláček, I. Horová, J. Mikšovský, Similarities within a multi-model ensemble: functional data analysis framework, *Geosci. Model Dev.*, **12** (2019), 735–747. http://doi.org/10.5194/gmd-12-735-2019

41. D. A. Shah, E. D. De Wolf, P. A. Paul, L. V. Madden, Functional Data Analysis of Weather Variables Linked to Fusarium Head Blight Epidemics in the United States, *Phytopathology®*, **109** (2019), 96–110. http://doi.org/10.1094/PHYTO-11-17-0386-R

42. B. Guo, H. Wu, L. Pei, X. Zhu, D. Zhang, Y. Wang, et al., Study on the spatiotemporal dynamic of ground-level ozone concentrations on multiple scales across China during the blue sky protection campaign, *Environ. Int.*, **170** (2022), 107606. http://doi.org/10.1016/j.envint.2022.107606

43. P. Craven, G. Wahba, Smoothing noisy data with spline functions, *Numer. Math.*, **31** (1978), 377–403. http://doi.org/10.1007/BF01404567

44. M. Gubian, F. Torreira, L. Boves, Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts, *J. Phonetics*, **49** (2015), 16–40. http://doi.org/10.1016/j.wocn.2014.10.001

45. L. Tavi, T. Kinnunen, R. González Hautamäki, Improving speaker de-identification with functional data analysis of f0 trajectories, *Speech Commun,*, **140** (2022), 1–10. http://doi.org/10.1016/j.specom.2022.03.010