The Use of Spline Techniques in the Nonparametric Regression Analysis for the Sequence Data with a Random Walk Process

AUTCHA ARAVEEPORN*, THANRADA CHAIKAJONWAT
Department of Statistics, School of Science,
King Mongkut's Institute of Technology Ladkrabang,
10520, Bangkok,
THAILAND

**Corresponding Author*

Abstract: - This study evaluates and compares various spline techniques in the nonparametric regression analysis, specifically focusing on the smoothing spline regression, the natural spline regression, the B-spline regression, and the penalized spline regression. The dependent variable in this analysis is time series data generated by a random walk process, while the independent variable is represented as sequential data. The simulation data, derived from a random walk process with diverse variances and sample sizes, ensures an absence of fixed patterns in the variable's changes. In addition, real-world data from the monthly trading volume of the SET (Stock Exchange of Thailand) index is used for practical application. The criterion for model efficiency estimation is based on minimizing the average mean square error for the simulation and SET index data. At the same time, predictive performance for future values is assessed through the minimum of average mean absolute percentage error. Among the models tested, the natural spline regression achieved the minimum average mean square error in all simulations due to SET index data estimation, excelling in model fit. However, the B-spline regression proved highly effective for forecasting future values.

Key-Words: - B-spline Regression, Natural Spline Regression, Penalized Spline Regression, Sequence Data, Smoothing Spline regression, Spline Techniques, Nonparametric Regression analysis.

Received: October 22, 2024. Revised: November 29, 2024. Accepted: February 23, 2025. Published: April 8, 2025.

1 Introduction

Regression analysis is a foundational tool in statistics and data science. Analysts can examine relationships between variables, predict future outcomes, and make data-driven decisions by modeling the association between the independent or multiple independent and dependent variables. The regression analysis quantifies the strength and nature of associations and controls for confounding factors, [1]. It forecasts future trends by modeling the relationship between an independent or multiple independent variables and a dependent variable, [2].
While the regression analysis is a powerful tool, certain assumptions underlie its validity. Four assumptions that govern regression results are linearity, Independence, homoscedasticity, and normality, [3]. Multicollinearity significantly impacts the parameter estimation, resulting in unstable and imprecise parameter estimates, [4]. When these assumptions are violated, the reliability of regression results may be compromised, requiring techniques such as transformation [5], nonparametric path analysis [6] and spline techniques [7].

The spline techniques are powerful tools in nonparametric regression analysis that allow for flexible and smooth curve fitting to data, making them ideal for capturing complex, nonlinear relationships, [8]. Instead of assuming a fixed functional form for the entire dataset, splines divide the data range into intervals and fit separate polynomial functions [9] to each interval. These polynomials are then connected at specific points called "knots," ensuring a smooth transition between segments. [10] applied the multivariate adaptive regression splines technique to assess dimensionless parameters' sensitivity to the uplift capacity factor. They proposed an empirical design equation for its efficient prediction.

The spline techniques are among the most popular methods in nonparametric regression analysis for capturing complex, nonlinear relationships within data. By allowing flexible, piecewise polynomial fit that joins smoothly at specific points (knots), splines offer both

adaptability and smoothness in model construction using a smoothing algorithm, [11]. The most widely recognized and influential spline techniques in regression analysis are smoothing spline regression [12], [13], natural spline regression [14], [15], B-spline regression [16], [17], and penalized spline regression [18], [19].

The smoothing spline regression minimizes a tradeoff between data fit and curve smoothness. By introducing a smoothing parameter, they control the amount of flexibility, balancing the fit to the dataset with a penalty function for excessive fluctuations. They are often used when data contain random volatility, as the smoothing parameter allows users to adjust the level of detail captured. [20] explored the smoothing spline regression method as a penalized least squares regression method and applied it to a real-world variational dataset. [21] proposed an adaptive smoothing spline regression estimator for a linear regression model, where the full trajectory of the independent variable influences the dependent variable at each point in the domain.

The natural spline regression is a cubic spline with boundary constraints that force the function to become the linear function at the edges of the data range, reducing boundary effects. The linear boundary constraint ensures the curve remains well-behaved at the edges, making it an attractive option for data with significant variability at the endpoints. [22] introduced a longitudinal data analysis method using the natural spline regression, modeling time as a continuous variable while accounting for testing version effects to capture the mean trajectory over time. Similarly, [23] provided a practical guide for summarizing nonlinear growth patterns of measured continuous outcomes using linear or natural spline regression.

The B-spline regression, or the basis splines, is a family of spline functions that form a basis function as the spline space. They allow for flexible fitting through a piecewise polynomial approach. The B-spline regression divides data into intervals and fits polynomials within each segment, as they only require local control over each segment. [24] proposed a high-order numerical approach utilizing a quintic B-spline regression collocated over the finite elements to numerically solve a class of nonlinear singular boundary value problems. [25] developed a novel differential-recurrence relation for the B-spline functions of a given degree, determining the coefficients of the B-spline functions of various degrees in the Bernstein–Bézier form.

The penalized spline regression, or the P-spline regression, is a smoothing spline regression that uses penalties to control the wiggle curve. A roughness penalty on the differences between coefficients in adjacent intervals is applied. The P-spline regression balances flexibility and control by including many knots while avoiding overfitting. [26] derived the asymptotic distribution of the quantile estimators obtained using the penalized spline regression method. [27] introduced a joint penalized spline regression model, reparametrizing the penalized spline regression as a linear mixed model.

A random walk process is a fundamental concept in probability theory and statistical modeling, with applications in diverse fields such as finance, physics, biology, and economics. It describes a sequence of random steps that occur over time, often used to model phenomena where chance influences outcomes [28]. The time series data plays a pivotal role in a random walk process, providing a framework for understanding and modeling non-stationary behavior in time-dependent data. The current value depends solely on the previous value plus a random disturbance, and it serves as a cornerstone in statistical and econometric modeling, particularly in financial markets. [29] examined the random walk process depending on the time series and proposed a new method for identifying the corresponding distributions of ordinal patterns.

The Stock Exchange of Thailand (SET) index is a key financial indicator reflecting the performance of the Thai stock market. Given its dynamic and often unpredictable nature, the random walk process is a suitable framework for analyzing its time series behavior. The random walk process offers a foundational approach for analyzing the SET Index in time series data. It captures the stochastic nature of price movements while providing insights into market behavior. [30] studied stock price forecasting and demonstrated that sentiment information hidden in corporate annual reports can effectively predict short-run stock price returns.

This research examines the spline techniques in regression analysis, including the smoothing spline regression, the natural spline regression, the B-spline regression, and the penalized spline regression. The simulation study uses random walk-generated time series data to evaluate the average mean square error and the average mean absolute percentage error based on optimal knot selection. Actual data from Thailand's SET index is the dependent variable, with time sequence as the independent variable.

2 Spline Techniques

A spline is a piecewise polynomial function defined over intervals between knots, where each polynomial segment is related smoothly. The degree of the polynomial, location, and number of knots influence the spline's flexibility and smoothness. In the nonparametric regression analysis, the spline techniques exhibit the relationship between an independent (x_i) and the dependent variable (y_i) by fitting a spline function $\mu(x_i)$ that minimizes a chosen objective function, often balancing data fidelity with smoothness. The general form of a spline regression model [31] is:

$$y_i = \mu(x_i)+\varepsilon_i \ , \ i=1,2,3,...,n, \qquad (1)$$

where y_i is the observed value at a point x_i, $\mu(x_i)$ is the unknown spline function, and ε_i is the error term.

In the spline techniques, the function $\mu(x_i)$ is expressed as a weighted sum of the basis functions, each associated with a coefficient estimator.

2.1 Smoothing Spline Regression

The smoothing spline regression is a powerful technique that fits a smooth curve through data points by balancing data fidelity with a smoothness constraint. Unlike the traditional regression models that assume a specific functional form, the smoothing splines adaptively fit the data with a piecewise smooth curve. Write step by step as follows. Knots are the specific points in the domain of a function where piecewise polynomial segments are joined together in spline regression. They serve as breakpoints that control the flexibility of the spline function. The choice of the number and placement of knots affects the smoothness and fit of the spline model. Penalty functions are used in the spline regression to control overfitting by adding a smoothness constraint.

Step 1: Define the smoothing spline regression model

Given data points (x_i, y_i), the main objective of is to find a spline function $\mu(x_i)$ that minimizes the following penalized sum of squares:

$$\sum_{i=1}^{n}\{y_i-\hat{\mu}(x_i)\}^2+\lambda\int\{\hat{\mu}''(x_i)\}^2dx, \qquad (2)$$

where $\lambda\geq 0$ is a regularization parameter that controls the tradeoff between closeness to fit the data and $\hat{\mu}''(x_i)$ is the second derivative of $\hat{\mu}(x_i)$.

Step 2: Represent in matrix form

The roughness penalty [32] is in the matrix form as:

$$\int_a^b\{\hat{\mu}''(x_i)\}^2dx=\hat{\mu}^T A\hat{\mu},$$

where $\hat{\mu}=\left((\hat{\mu}(x_1),...,\hat{\mu}(x_n)\right)^T$ are the fitted values and A is the basis function. The matrix A depends on the configuration of the independent variables as $n\times n$ matrix that is evaluated by $A=\Delta W^{-1}\Delta$, where Δ is the second difference as $(n-2)\times n$ matrix with elements: $\Delta_{ii}=\frac{1}{h_i}$, $\Delta_{i,i+1}=-\frac{1}{h_i}-\frac{1}{h_{i+1}}$, and $\Delta_{i,i+2}=\frac{1}{h_i+1}$. The W is the symmetric tridiagonal matrix as $(n-2)\times(n-2)$ matrix with an element: $W_{i-1,i}=W_{i,i-1}=\frac{h_i}{6}$, $W_{ii}=\frac{(h_i+h_{i+1})}{3}$, and $h_i=\xi_{i+1}-\xi$, the distances between successive knots.

Step 3: Solving for the coefficients

From (2), the penalized sum of squares can be rewritten in matrix form as

$$(\underset{\sim}{y}-\hat{\mu})^{\mathrm{T}}(\underset{\sim}{y}-\hat{\mu})+\lambda\,\hat{\mu}^T A\,\hat{\mu}, \qquad (3)$$

where $\underset{\sim}{y}=(y_1,...,y_n)^{T\cdot}$. By minimizing over $\hat{\mu}$ to differentiate with $\hat{\mu}$ that given the results as

$$-2(\underset{\sim}{y}-\hat{\mu})=2\lambda\,A\,\hat{\mu}=0, \qquad (4)$$

and the smoothing spine estimator is approximated by

$$\hat{\mu}=(I+\lambda A)^{-1}\underset{\sim}{y}. \qquad (5)$$

2.2 Natural Spline Regression

The natural spline regression is a cubic spline regression that not only fits a set of given data points smoothly but also has the additional property that its second derivative is zero at the endpoints, effectively minimizing the curvature at the edges. The following steps outline the process.

Step 1: Define the natural spline regression model

The natural spline as a cubic polynomial is defined on each interval $[x_i, x_{i+1}]$ as

$$\mu_i(x)=a_i+b_i(x-x_i)+c_i(x-x_i)^2+d_i(x-x_i)^3,$$

where a_i, b_i, c_i and d_i are the coefficients to be determined for each interval.

Step 2: Continuity and smoothness conditions

To ensure smooth transitions between intervals, the following conditions must be met:

1. Interpolation: The function passes through all given data points [14] $(x_i \ y_i$ as $\mu(x_i) = y_i$, $\mu(x)$ is a spline function.
2. First-Derivative Continuity: $\mu_i{}'(x)$ is continuous at all knots.
3. Second-Derivative Continuity: $\mu_i{}''(x)$ is continuous at all knots.
4. Natural Boundary Condition: The second derivative is zero at the endpoints as $\mu''(x_1)) = 0$ and $\mu''(x_n)) = 0$.

Step 3: Solving for the coefficients
Using the above conditions, we derive the following system of linear equations in terms of second derivatives $\mu_i''(x_i) = M_i$:

$$h_{i-1}M_{i-1} + 2(h_{i-1} + h_i)M_i + h_i M_{i+1} = 6\left(\frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_i - 1}\right). \quad (6)$$

where $h_i = x_{i+1} - x_i$ represents the interval length.

Step 4: Computing in matrix form
The equations (6) derived in a tridiagonal matrix form as

$$\mathbf{AM} = \mathbf{d}, \quad (7)$$

where $\mathbf{A}$ is an $(n-2)\times(n+2)$ tridiagonal matrix with entries identified the interval lengths as h_i, $\mathbf{M} = [M_2, M_3, ..., M_{n-1}]^T$ is the vector of unknown second derivatives, $\mathbf{d}$ is a vector derived from the values of y_i and h_i. With M_i values can determine the coefficients a_i, b_i, c_i and d_i for each interval $[x_i, x_{i+1}]$ as follows: $a_i = y_i$,

$b_i = \frac{y_{i+1} - y_i}{h_i} - \frac{h_i}{6}(2M_i + M_{i+1})$,

$c_i = \frac{M_i}{2}$, and $d_i = \frac{M_{i+1} - M_i}{6h_i}$.

Step 5: Solved to determine a spline function
The computation of the matrix is approximated by $\mathbf{M} = \mathbf{A}^{-1}\mathbf{d}$. Each segment $\mu_i(x)$ can now be constructed, providing the natural spline regression that smoothly interpolates the given data points.

2.3 B-Spline Regression

The B-spline regression of basis functions [33] is defined over a sequence of knots, $\{t_i\}$, which determine the intervals over which the polynomial pieces apply. Here are the steps to follow.

Step 1: Define the B-Spline basis functions
The B-spline (basis spline) are basis functions used to construct smooth piecewise polynomials. These basis functions are defined recursively. For degree p, the B-spline basis functions are constructed as:

1. For degree $p = 0$:

$$B_{i,0}(x) = \begin{cases} 1 & , \text{ if } t_i < x < t_{i+1} \\ 0 & \text{otherwise} \end{cases}.$$

2. For degree $p > 0$:

$$B_{i,p}(x) = \frac{x - t_i}{t_{i+p} - t_i} B_{i,p-1}(x) + \frac{t_{i+p+1} - x}{t_{i+p+1} - t_{i+1}} B_{i+1,p-1}(x). \quad (8)$$

where t_i are the knots that are predefined points where the polynomial pieces join.

Step 2: Constructing the B-Spline regression model
The function $\mu(x)$ is expressed as a linear combination of B-spline basis functions:

$$\mu(x) = \sum_{k=1}^{m} c_k B_k(x), \quad (9)$$

where c_k are the coefficients to be estimated and $B_k(x)$ are B-spline basis functions.

Step 3: The matrix form of the B-Spline regression
To create a matrix representation, the B-spline regression of basis functions is evaluated by

$$\boldsymbol{\mu} = \mathbf{B}\,\mathbf{c}, \quad (10)$$

where $\boldsymbol{\mu} = [\mu(x_1), \mu(x_2), ..., \mu(x_n)]^T$ is the vector of the spline values at each x_i, $\mathbf{B}$ is an $n \times m$ matrix of the evaluated basis functions $B_{i,p}(x)$ as the j-th B-spline regression of the basis function of the degree p evaluated at x_i, and $\mathbf{c} = [c_1, c_2, ..., c_m]^T$ is the vector of coefficients of each basis function.

Step 4: Solve to the coefficient of the B-Spline
The coefficients $\mathbf{c}$ is the best approximate $\mathbf{y} = [y_1, y_2, ..., y_n]^T$ by minimizing the squared error

$$\min_c \|\mathbf{y} - \mathbf{Bc}\|^2.$$

The solution is given by

$$\mathbf{c} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}\mathbf{y}. \quad (11)$$

2.4 Penalized Spline Regression

The penalized spline regression (P-spline) combines the flexibility of splines with a penalty term to control smoothness. Introduced by [16] and later developed by [34], penalized splines have become a standard tool in regression modeling. The following steps outline the process.

Step 1: Define the penalized spline regression model
The model can be expressed as:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k (x_i - \tau_k)_+ + \varepsilon_i \ , \ \varepsilon_i \sim N(0, \sigma^2), \quad (12)$$

where

y_i is the dependent variable as the observed data,

x_i is the independent variable as the sequence data,

$\beta_0 + \beta_1 x_i$ is the linear component model,

$\sum_{k=1}^{K} u_k (x_i - \tau_k)_+$ is a truncated power basis:

$(x - \tau_k)_+ = \max(0, x - \tau_k)$, τ_k are the knots, u_k are the coefficients for the basis functions and ε_i the random error term.

Step 2: Represent in matrix form
In matrix form, the model is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad (13)$$

where $\mathbf{y}$ is the $n \times 1$ vector of dependent variables, $\mathbf{X}$ is the $n \times p$ design matrix for the fixed effects, $\mathbf{Z}$ is the $n \times K$ matrix of basis functions for the random effects, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effect coefficients, $\mathbf{u}$ is the $K \times 1$ vector of random-effect coefficients, and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of errors.

Step 3: Minimize the penalized residual sum of squares
To estimate the coefficients, the penalized residual sum of squares (RSS) is minimized:

$$\text{RSS} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}\|^2 + \lambda \|\mathbf{u}\|^2 \ , \quad (14)$$

where the first term measures the fit to the data, the second term penalizes the roughness of the curve, and $\lambda = \frac{\sigma^2}{\sigma_u^2}$ is the regularization parameter. The random effects are defined to follow a Gaussian prior as $\mathbf{u} \sim N(0, \sigma_u^2 \mathrm{I})$.

Step 4: Solve the coefficient of penalized spline
To estimate $\boldsymbol{\beta}$ and $\mathbf{u}$, the penalized RSS is minimized by solving the mixed-model equations as

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda I \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}.$$

The coefficients are computed as:

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda I \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{bmatrix}. \quad (15)$$

3 Simulation Study

To visualize the random walk, the time series data is generated from the dependent variables in the formula below

$$y_i = y_{i-1} + \varepsilon_i \ , i = 2, 3, ..., n, \quad (16)$$

where y_i is the value of the time series at the time i, y_{i-1} is the value of the time series at the previous time step, ε_i is a random error term, typically assumed to be white noise. In this case, a random error term is simulated from the normal distribution with mean $\mu = 0$ and standard deviation (S.D.) = 1, 3, 5, and 7. The sample sizes (n) are 100 and 200 as the independent variables in sequence data. Figure 1 and Figure 2 show the time series plot of several standard deviations (S.D.) with 100 and 200 sample sizes, respectively.

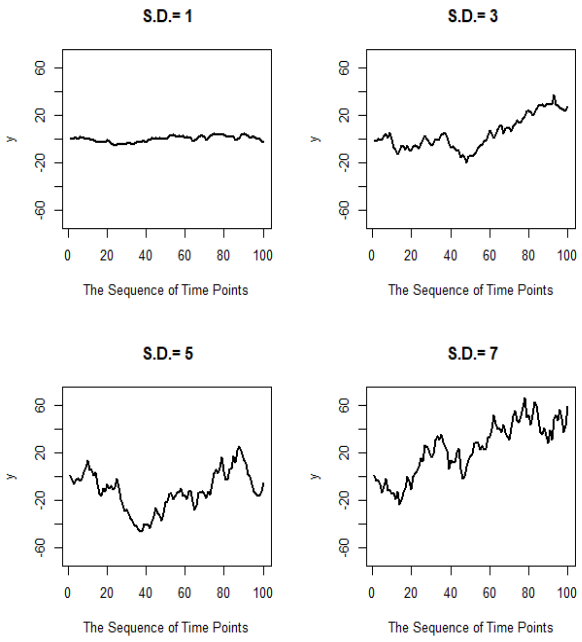

Fig. 1: Time series plot illustrating a random walk process with 100 sample sizes under different standard deviations

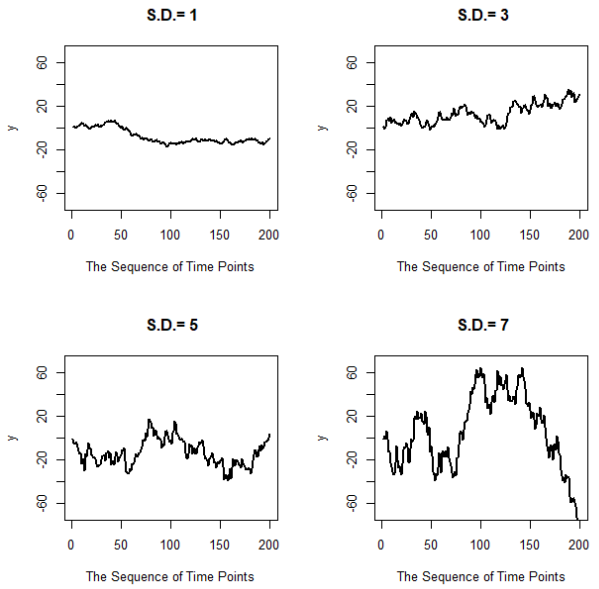

Fig. 2: Time series plot illustrating a random walk process with 200 sample sizes under different standard deviations

The R program generates data and repeats the model fitting 1,000 times. The performance of the estimating spline function is considered the Average Mean Square Error (AMSE) computed from the Mean Square Error (MSE) for each replication. The AMSE and MSE are

$$\text{AMSE} = \frac{\sum_{h=1}^{1,000} \text{MSE}_h}{1,000}, \text{ and}$$

$$\text{MSE}_h = \frac{1}{n-10}\sum_{i=1}^{n-10}(y_i - \hat{y}_i)^2 \quad .$$

$$, h = 1,2,...,1,000 \; ; \; i = 1,2,...,n-10.$$

The effectiveness of the forecasting spline function is assessed using the Average Mean Absolute Percentage Error (AMAPE) calculated as the Mean Absolute Percentage Error (MAPE) across all replications. The AMAPE and MAPE are:

$$\text{AMAPE} = \frac{\sum_{h=1}^{1,000} \text{MAPE}_h}{1,000}, \text{ and}$$

$$\text{MAPE}_h = \frac{1}{10}\sum_{i=1}^{10}\frac{|y_i - \hat{y}_i|}{y} \times 100 \quad .$$

$$, h = 1,2,...,1,000; \; i = 1,2,...,10.$$

After fitting the spline model, Table 1 and Table 2 present the AMSE and the number of knots of various standard deviations via smoothing spline regression (SSR), natural spline regression (NSR), B-spline regression (BSR), and penalized spline regression (PSR) methods for 100 and 200 sample sizes.

Table 1. Comparing AMSE and the number of knots for sample sizes of 100 under different standard deviations

Methods	S.D. = 1	S.D. = 3	S.D. = 5	S.D. = 7
SSR	0.11 (65)	0.95 (65)	2.65 (65)	5.20 (65)
NSR	5.07×10^{-28} (99)	4.51×10^{-27} (99)	1.26×10^{-26} (99)	2.47×10^{-26} (99)
BSR	5.23×10^{-18} (99)	4.70×10^{-17} (99)	1.31×10^{-16} (99)	2.56×10^{-16} (99)
PSR	1.12×10^{-11} (100)	1.01×10^{-10} (100)	2.81×10^{-10} (100)	5.50×10^{-10} (100)

Table 2. Comparing AMSE and the number of knots for sample sizes of 200 under different standard deviations

Methods	S.D. = 1	S.D. = 3	S.D. = 5	S.D. = 7
SSR	0.49 (109)	0.43 (107)	12.3 (106)	24.4 (106)
NSR	7.37×10^{-26} (199)	7.50×10^{-26} (199)	7.85×10^{-26} (199)	8.46×10^{-26} (199)
BSR	2.31×10^{-3} (198)	2.08×10^{-2} (198)	5.78×10^{-2} (198)	0.11 (198)
PSR	2.07×10^{-10} (200)	1.86×10^{-9} (200)	5.17×10^{-9} (200)	1.01×10^{-8} (200)

From Table 1 and Table 2, the natural spline regression (NSR) outperforms the other methods for estimating time series data by a random walk. The number of knots is made accurately close to the sample sizes. The increasing AMSE makes the increasing standard deviation an error. However, when the sample sizes are increased, the performance of these methods is influenced by the rising AMSE. The AMSE can determine which spine techniques can be the best estimator.

The number of knots from small to sample sizes is replaced to approximate the smoothing function, and a minimum mean square error is selected for the optimum knots. Many knots can control the smoothing interpolation and use it to trade off the goodness of fit. After estimating the parameters and knots, the estimated values are used to forecast the following ten values to evaluate the forecasting performance, as shown in Table 3 and Table 4.

Table 3. Forecasting performance evaluation using AMAPE for sample sizes of 100 under different standard deviations

Methods	S.D. = 1	S.D. = 3	S.D. = 5	S.D. = 7
SSR	0.598	0.687	0.974	1.014
NSR	0.024	0.057	0.087	0.098
BSR	0.0001	0.0004	0.0007	0.0009
PSR	0.147	0.269	0.498	0.575

Table 4. Forecasting performance evaluation using AMAPE for sample sizes of 200 under different standard deviations

Methods	S.D. = 1	S.D. = 3	S.D. = 5	S.D. = 7
SSR	0.625	0.783	1.140	1.265
NSR	0.078	0.095	0.168	0.365
BSR	0.002	0.005	0.008	0.018
PSR	0.287	0.336	0.593	0.675

For Table 3 and Table 4, the B-spline regression (BSR) consistently achieves the lowest AMAPE values, indicating the highest accuracy in forecasting future values. The natural spline regression (NSR) and penalized spline regression (PSR) perform moderately, improving over smoothing spline Regression (SSR) but not as accurately as BSR. The SSR shows the highest AMAPE values in all cases, making it the least accurate method. The accuracy of all methods decreases slightly as S.D. increases, but the rankings remain consistent. Overall, BSR is the most reliable method for forecasting across sample sizes.

4 Application on Actual Data

The Stock Exchange of Thailand (SET) index is the primary benchmark index representing the performance of all stocks listed on the Stock Exchange of Thailand. The SET index exhibits the random walk process behavior due to economic and external market factors. We extracted historical monthly volume time series from January 2005 to October 2024 from http://www.set.or.th/th/market/market_statistics.html. Let y_i denote the SET index of the month and x_i where $x_1 = 1$ represents January of 2005 and $x_{238} = 238$ represents October 2024, as shown in Figure 3.

The smoothing spline regression (SSR), the natural spline regression (NSR), the B-spline regression (BSR), and the penalized spline regression (PSR) methods are employed to estimate and forecast the time series of the SET index.

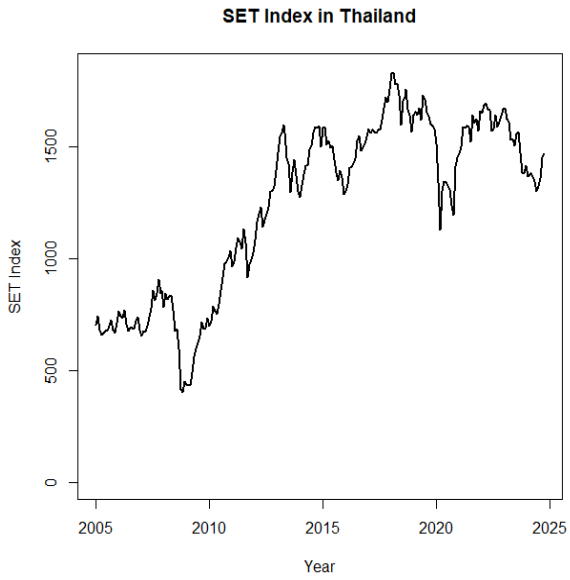

Fig. 3: Monthly trading volume of the Stock Exchange of Thailand (SET) Index from January 2005 to October 2024 under a random walk process used for time series data

The differences between actual and estimated data are computed using the Mean Square Error (MSE) from January 2005 to December 2023. In the next ten months, the Mean Absolute Percentage Error (MAPE) evaluates forecasting accuracy in percentage from January 2024 to October 2024. Both MSE and MAPE measured the precision of the estimated spline function and used the spline function to forecast future values, with their respective formulas provided below:

$$\text{MSE} = \frac{1}{228}\sum_{i=1}^{228}(y_i - \hat{y}_i)^2,\ i = 1,2,...,228$$

and $$\text{MAPE} = \frac{1}{10}\sum_{i=229}^{238}\frac{|y_i - \hat{y}_i|}{y_i}\times 100,\ i = 229,...,238.$$

The actual values are defined y_i, and the estimated and forecasted values are defined $\hat{y}_i$. The actual SET index (y_i) and the estimated values ($\hat{y}_i$) from four methods are shown in Figure 4 and the main result of MSE is approximated in this figure.

As demonstrated in Figure 4, it is hard to see the performance of these methods, and the calculation of MSE can indicate the best estimation method. In computing, Table 5 shows that the natural spline regression method is a minimum of MSE and the optimal knots.

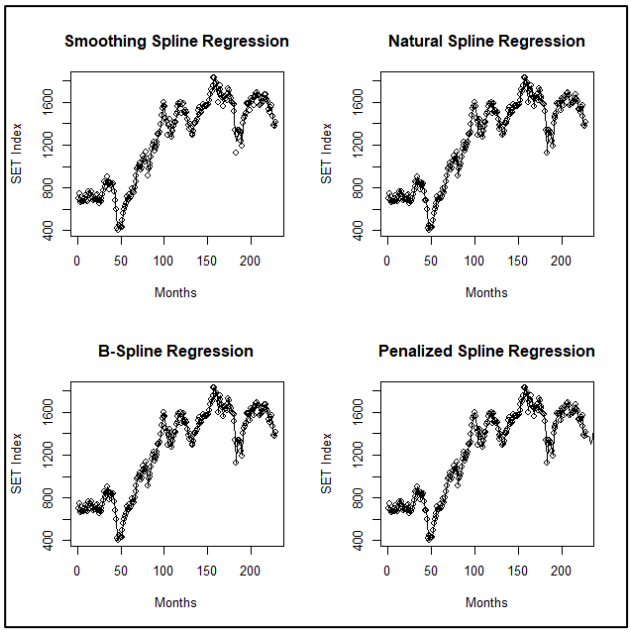

Fig. 4: Comparison of estimated values obtained from different spline regression methods with actual SET Index values from January 2005 to December 2023

Table 5. Comparison of the number of knots and MSE of SET index via spline regression methods

Methods	Knots	MSE
SSR	108	631.154
NSR	227	2.37×10^{-23}
BSR	226	0.322
PSR	228	25.062

The future forecasting values are attractive because they focus on the performance of spline methods in the following ten data points and the forecasting data shown in Figure 5 and Table 6.

Figure 5 shows the forecasting data analysis performed for all methods. Some methods are on the plot of actual SET Index data, so it is challenging to indicate which method to select. Table 6 shows approximate forecasting values and MAPE. The minimum MAPE will be considered the performance method. In Table 6, the B-spline regression obtains the minimum MAPE, the best method to forecast future data. The spline regression techniques offer flexible modeling approaches but come with specific challenges. The smoothing spline requires careful tuning of the smoothing parameter to avoid overfitting or underfitting. The natural spline effectively handles boundary effects but may become unstable if knots are not optimally placed. The B-spline provides adaptability but needs appropriate knots to balance complexity and smoothness. The penalized spline helps control overfitting but can excessively smooth the data if the penalty parameter is misconfigured. Proper parameter selection is crucial to achieving optimal model performance.

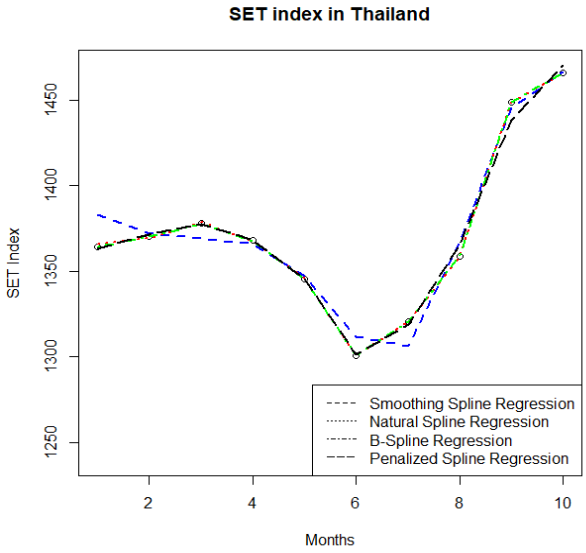

Fig. 5: Forecasted values for the SET Index from January 2024 to October 2024 using various spline regression methods

Table 6. Forecasting the SET index from January 2024 to October 2024 using spline regression techniques with corresponding MAPE

SET Index	SSR	NSR	BSR	PSR
1364.52	1383.00	1366.03	1364.42	1363.30
1370.67	1372.29	1369.57	1370.56	1371.41
1377.94	1369.52	1378.65	1377.82	1377.50
1367.95	1365.80	1367.52	1367.91	1368.20
1345.66	1347.21	1345.88	1345.45	1345.44
1300.96	1311.93	1300.85	1300.92	1301.38
1320.86	1306.53	1320.90	1320.77	1319.24
1359.07	1367.85	1359.05	1359.17	1366.13
1448.83	1445.38	1448.83	1448.72	1438.52
1466.40	1466.73	1466.03	1466.03	1470.59
MAPE	0.5217	0.0304	0.0004	0.1900

5 Conclusion

The spline techniques benefit the nonparametric regression analysis by offering flexibility, smooth transitions, and control overfitting. The spline can model simple and complex patterns without assuming a strict parametric form, allowing them to fit data structures that vary across the data range.

We evaluated the smoothing estimator of the smoothing spline regression, the natural spline regression, the B-spline regression, and the penalized spline regression techniques through simulation studies. Four spline techniques can evaluate the smoothing function on time series data in a random walk process. The results indicate that the natural spline regression performs effectively regardless of sample size or standard deviation. The natural spline regression shows the estimation performance for application in actual data since it is an exciting method for interpolating observed data using spline techniques. Furthermore, the polynomial term can support at some point, making the piecewise interpolation helpful with large data sets. The natural spline regression is appropriately examined in the fitting model on time series data studied by [35]. However, the B-spline regression outperforms the prediction of future values in short-term forecasting due to its ability to provide a flexible yet smooth approximation of data patterns. Its localized control property, where adjustments to one part of the spline do not significantly affect others, makes it particularly effective for capturing short-term trends while avoiding overfitting. Future research will extend the methodology to alternative nonparametric, multivariate time series analysis, apply it to various financial datasets such as commodity prices and cryptocurrency trends, and explore hybrid models that integrate spline techniques with machine learning to enhance forecasting accuracy. These advancements will improve model adaptability and predictive performance in complex financial environments.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the authors used Grammarly for language editing. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

[1] D. C. Montgomery, E. A. Peck, G. G. Vining, *Introduction to Linear Regression Analysis*, Wiley, 2021.

[2] M. H. Kutner, C. J. Nachtsheim, J. Neter, J, W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2004.

[3] J. Fox, *Applied Regression Analysis and Generalized Linear Models*, SAGE Publications, 2016.

[4] I. J. Idowu, A. T. Owolabi, O. J. Oladapo, K. Ayinde, O. A. Oshuoporu, A. N. Alao, Mitigating multicollinearity in linear regression model with two parameter Kibria-Lukman estimators, *WSEAS Transactions on Systems and Control*, Vol. 18, 2023, pp. 612-620. https://doi.org/10.37394/23203.2023.18.63.

[5] G. James, M. Kinnon, L. Magee, Transforming the dependent variable in regression models, *International Economic Review*, Vol. 31, No.2, 1990, pp. 315-339. DOI: 10.2307/2526842.

[6] A. A. R. Fernandes, S. Solimun, L. Muflikhah, A. Alifa, E. Krisnawati, N. M. A. A. Badung, E. C. L. Efendi, Nonparametric path analysis on consumer satisfaction and consumer engagement in PT pertamina, *WSEAS Transactions on Mathematics*, Vol. 21, 2022, pp. 17-22. https://doi.org/10.37394/23206.2022.21.3.

[7] K. Hendrickx, P. Janssen, A. Verhasselt, Penalized spline estimation in varying coefficient models with censored data, *TEST*, Vol. 27, 2018, pp. 871-895. DOI: 10.1007/ s11749-017-0574-y.

[8] S. N. Wood, *Generalized Additive Models: An Introduction with R*, CRC Press, 2017.

[9] G. James, D. Witten, D. T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer, 2021.

[10] T. Jearsiripongkul, V. Q. Lai, S. Keawsawasvong, T. S. Nguyen, C. N. Van, C. Thongchom, P. Nuaklong, Prediction of uplift capacity of cylindrical caissons in anisotropic and inhomogeneous clays using multivariate adaptive regression splines, *Sustainability*, Vol. 14, No.8, 2022, pp. 1-21. DOI: 10.3390/ su14084456.

[11] E. G. Pale-Ramon, Y. S. Shmaliy, L. J. Morales-Mendoza, M. González-Lee, J. A. Ortega-Contreras, R. F. Vázquez-Bautista, Pseudo-ground truth trajectory from contaminated data of object tracking using smoothing algorithms, *WSEAS Transactions on Signal Processing*, Vol. 19, 2023, pp. 67-76. https://doi.org/10.37394/232014.2023.19.8.

[12] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[13] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*, CRC Press, 1999.

[14] D. Ruppert, M. P. Wand, R. J. Carroll, *Semiparametric Regression*, Cambridge University Press, 2003.

[15] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer, 2015.

[16] P. H. C. Eilers, B. D. Marx, Flexible Smoothing with B-Splines and Penalties, *Statistical Science*, Vol. 11, No. 2, 1996, pp. 89–102. DOI: 10.1214/ss/1038425655.

[17] L. L. Schumaker, *Spline Functions: Basic Theory*, Cambridge University Press, 2007.

[18] M. P. Wand, Smoothing and Mixed Models, *Computational Statistics*, Vol. 18, 2003, pp. 223–249. DOI: 10.1007/s001800300142.

[19] L. Fahrmeir, T. Kneib, S. Lang, B. Marx, *Regression: Models, Methods and Applications*, Springer, 2013.

[20] N. Breaz, M. Aldea, On the smoothing spline regression models, *Acta Universitatis Apulensis*, No. 15, 2008, pp. 33-51.

[21] F. Centofanti, A. Lepore, A. Menafoglio, B. Palumbo, S. Vantini, Adaptive smoothing spline estimator for the function-on-function linear regression model, *Computational Statistics*, Vol. 38, 2023, pp. 191-206. DOI: 10.1007/s00180-022-01223-6.

[22] M. C. Donohue, O. Langford, P. S. Insel, C. H. van Dyck, R. C. Petersen, S. Craft, G. Sethuraman, R. Raman, P. S. Aisen, Natural cubic splines for the analysis of Alzheimer's clinical trials, *Pharmaceutical Statistics*, Vol. 22, 2023, pp. 508–519. DOI: 10.1002/pst.2285.

[23] A. Elhakeem, R. A. Hughes, K. Tilling, D. L. Cousminer, S. A. Jackowski, T. J. Cole, A. S. F. Kwong, Z. Li, S. F. A. Grant, A. D. G. Baxter‐Jones, B. S. Zemel, D. A. Lawlor, Using linear and natural cubic splines, SITAR, and latent trajectory models to characterise nonlinear longitudinal growth trajectories in cohort studies, *BMC Medical Research Methodology*, Vol. 22, No. 68, 2022, pp. 1-20. DOI: 10.1186/s12874-022-01542-8.

[24] P. Roul, A high-order B-spline collocation method for solving a class of nonlinear singular boundary value problems, *Journal of Mathematical Chemistry*, Vol. 62, 2024, pp. 1308-1322. DOI: 10.1007/s10910-024-01590-z.

[25] F. Chudy, P. Wozny, Linear-time algorithm for computing the Bernstein–Bézier coefficients of B-spline basis functions, *Computer-Aided Design*, Vol. 154, 2023, 103434. DOI: 10.1016/ j.cad.2022.103434.

[26] T. Yoshida, Asymptotics for penalized spline estimators in quantile regression, *Communications in Statistics - Theory and Methods*, Vol. 52, No. 14, 2023, pp. 4851-4834. DOI: 10.1080/03610926.2013.765477.

[27] D. D. Witte, A. A. Abad, T. Neyens, G. Verbeke, G. Molenberghs, A joint penalized spline smoothing model for the number of positive and negative COVID-19 tests, *PLoS ONE*, Vol. 19, No. 5, 2024, e0303254. DOI: 10.1371/journal. .pone.0303254.

[28] G. Grimmett, D. Stirzaker, *Probability and Random Processes*, Oxford University Press, 2020.

[29] D. DeFord, K. Moore, Random Walk Null Models for Time Series Data, *Entropy*, Vol. 19, No. 11, 2017, pp. 1-18. DOI: 10.3390/ e19110615.

[30] P. Hájek, V. Olej, R. Myšková, Forecasting Stock Prices using Sentiment Information in Annual Reports – A neural network and support vector regression approach, *WSEAS Transactions on Business and Economics*, Vol. 10, No. 4, 2013, pp. 293–305.

[31] R. Andersen, Nonparametric Methods for Modeling Nonlinearity in Regression Analysis, *Annual Review of Sociology*, Vol. 35, 2009, pp. 67-85. DOI: 10.1146/annurev.soc.34.040507 .134631.

[32] P. J. Green, B. W. Silverman, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall/CRC, 1993.

[33] C. De Boor, *A Practical Guide to Splines*, Springer, 1978.

[34] D. Ruppert, R. J. Carroll, Sparse semiparametric regression, *Biometrika*, Vol. 87, No. 3, 2000, pp. 589-602. DOI: 10.1093/biomet/87.3.589.

[35] J. Laipaporn, P. Tongkumchum, Estimating the Natural Cubic Spline Volatilities of the ASEAN-5 Exchange Rates, *The Journal of Asian Finance, Economics and Business*, Vol. 8, No. 3, 2021, pp. 1-10. DOI: 10.13106/ jafeb. 2021.vol8.no3.0001.

Contribution of Individual Authors to the Creation of a Scientific Article

- Autcha Araveeporn: Conceptualized the study, designed the methodology, performed statistical modeling, and wrote the initial draft of the manuscript. Additionally, supervised the research process and coordinated revisions based on reviewer feedback.
- Thanrada Chaikajonwat: Conducted data collection, performed simulations, assisted with statistical analyses, and contributed to interpreting the results. Additionally, supported manuscript editing and formatting to meet journal requirements.

Sources of Funding for Research Presented in a Scientific Article

This work was financially supported by King Mongkut's Institute of Technology Ladkrabang Research Fund (Grant number: KREF186715)

Conflict of Interest

The authors have no conflicts of interest to declare.