MATEMATIKA, MJIAM, Volume 41, Number 1, 17–36

New *COVRATIO* Statistic for Outlier Detection in Simultaneous Linear Functional Relationship Model and Its Application in Malacca Environmental Dataset

[1]Nur Ain Al-Hameefatul Jamaliyatul, [2]Nurkhairany Amyra Mokhtar*, [3]Basri Badyalina, [4]Adzhar Rambli and [5]Yong Zulina Zubairi

[1,2,3]Mathematical Sciences Studies, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM), Cawangan Johor Kampus Segamat, 85000 Segamat, Johor, Malaysia.

[4] School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.

[5]Institute for Advanced Studies, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

*Corresponding author:nurkhairany@uitm.edu.my

Article history
Received: 23 September 2024
Received in revised form: 4 February 2025
Accepted: 6 March 2025
Published online: 1 April 2025

Abstract Environmental studies, such as monsoon analysis, require examining relationships between multiple linear variables like wind speed, humidity, and temperature, with consideration of errors in each variable. Outliers in the data may affect the accuracy of analysis. This study aims to develop and validate a novel method for detecting outliers in simultaneous linear functional relationship model (LFRM) using the *COVRATIO* statistic. The objectives include deriving cut-off points for outlier detection using Monte Carlo simulations and demonstrating the method's effectiveness on synthetic and real-world environmental datasets from Malacca. The findings confirm that the proposed method accurately identifies outliers, with detection performance improving as the variance of data contamination increases. Application to Malacca's environmental data during the 2020 southwest monsoon season revealed significant outliers in the relationship between wind speed and humidity, while no outliers were found for wind speed and temperature. Removing detected outliers resulted in improved parameter estimates and reduced variance, enhancing the reliability of the LFRM. Data normality was verified through Q-Q plots and the Kolmogorov-Smirnov test statistic, demonstrating the robustness and applicability of the method in environmental studies.

Keywords Simultaneous linear functional relationship model (LFRM); *COVRATIO* statistic; Monte Carlo simulation; outlier detection method; environmental dataset analysis; wind speed-humidity relationship; wind speed-temperature relationship.

Mathematics Subject Classification 62H12, 62F35, 62J05, 62P12, 65C05, 68U20.

1 Introduction

The study of outliers is as old as the field of statistics itself, underscoring its critical importance. An outlying observation, or "outlier," is defined as a data point that deviates significantly from other sample members [1] According to Hampel *et al.* [2] , a typical dataset may contain between one to ten percent outliers, and even the highest-quality data can include some level of outlier presence. The presence of a single outlier can significantly affect parameter estimates, leading to unreliable results [3]. Over the years, researchers have developed numerous techniques to identify outliers in linear data, as demonstrated by Grubbs [4], Sebert *et al.* [5] , Adnan *et al.* [6], and more recently, Arif*et al.* [7, 8] . Outlier detection has practical applications in various fields, including fraud detection, robust statistical analysis, and quality control, where identifying anomalies or irregularities is crucial for maintaining data integrity and ensuring accurate results [1].

Among the various methods for outlier detection, the *COVRATIO* statistic has become a widely used approach, particularly for identifying outliers in linear regression models, due to its ability to assess the influence of data points on the stability of parameter estimates [9]. The application of *COVRATIO* statistic has expanded over time. For instance, it was adapted for bivariate linear functional relationship models (LFRM) involving circular variables, utilising the covariance matrix of parameter estimates [10]. Ghapor *et al.* [11] developed *COVRATIO* statistic specifically for bivariate LFRM involving linear variables. Arif *et al.* [3] further refined the use of *COVRATIO* statistic in linear replicated LFRM to identify influential observations or outliers. The *COVRATIO* statistic, a simple and widely used method, is reliable for outlier detection in LFRM [7]. However, existing methodologies have not yet developed an outlier detection method in simultaneous LFRM for linear variables.

To address these gaps, this study presents a novel derivation of the *COVRATIO* statistic specifically designed to detect outliers in simultaneous LFRM. The importance of this advancement lies in its ability to improve the accuracy of parameter estimates by effectively identifying and mitigating the influence of outliers, thereby enhancing model reliability and applicability. The significance of this new method is demonstrated through Monte Carlo simulations and applications to synthetic and real-world environmental datasets. By establishing cut-off points for outlier detection and validating the method across diverse scenarios, this study bridges the gap between existing outlier detection techniques and the unique challenges posed by simultaneous LFRM. These advancements have critical implications for various fields, including environmental monitoring, econometrics, and engineering, where accurate modeling of interdependent variables with measurement errors is essential [3].

2 Methodology

2.1 Linear Data

Linear data refers to a datasets where data points follow a linear trend, suggesting a straight-line relationship between variables. These datasets typically follow a normal distribution, which aids in analysing underlying patterns and trends [12]. In this study, we assess the data distribution by applying the Q-Q plots and Kolmogorov-Smirnov test statistic, which helps reveal any deviations from normality.

2.2 Simultaneous Linear Functional Relationship Model

The error-in-variables model (EIVM) takes into account errors in both X and Y variables [13]. The functional relationship model, part of the broader EIVM category, accounts for measurement inaccuracies in both independent variable, X and dependent variable, Y [1, 12, 14, 15]. Unlike traditional linear regression, which assumes no error in X , the EIVM corrects for biases and inconsistencies [16]. These errors are common in fields like econometrics, environmental sciences, and engineering [17].

Ghapor *et al.* [11,18] identify the outliers in bivariate LFRM using the *COVRATIO* statistic. Jamaliyatul *et al.* [19] extends the bivariate LFRM to simultaneous LFRM for linear variables, allowing for the exploration of multiple linear relationships while accounting for measurement errors. This study will derive COVRATIO statistic to detect the outlier in simultaneous LFRM.

Let the variable $Y_{ji}(j = 1, ..., q; i = 1, ..., n)$ and $X_i = (i = 1, ..., n)$ related by simultaneous LFRM of $Y_j = \alpha_j + \beta_j X$, where n is the number of data point in dataset and q is the number of response variables [19]. Consider a data point represented as (x_i, y_{ji}) , where it aligns with the assessment of the actual values of (X_i, Y_{ji}) with some random error [19]. The random error δ_i and ϵ_{ji} are presumed to follow a normal distribution with $\delta_i \sim N(0, \sigma_i^2)$ and $\epsilon_{ji} \sim N(0, \tau_j^2)$, respectively [19].

The model of simultaneous LFRM based on [19], can be written as follows:

$$Y_j = \alpha_j + \beta_j X, \tag{1}$$

where $x_i = X + \delta_j$ and $y_{ji} = Y_{ji} + \epsilon_{ji}$ for $j = 1, ..., q; i = 1, ..., n$.

The ratio of error variances is known, denoted as $\lambda = \frac{\tau_j^2}{\sigma^2}$ for all findings across both variables [12, 19]. Consequently, there are $(p + q + 1)$ parameters to be estimated, namely $\alpha_j, \beta_j, \sigma^2$ and X_i . The log-likelihood function equation of the model is as follows.

$$\log \quad L = -n\log(2\pi) - \frac{n}{2}\log\lambda - n\log\sigma_i^2 - \frac{1}{2\sigma_i^2}\left\{ \begin{array}{l} \sum_{i=1}^{n}(x_i - X_i)^2 \\ +\frac{1}{\lambda}\sum_{j=1}^{q}\sum_{i=1}^{n}(y_{ji} - \alpha_j - \beta_j X_i)^2 \end{array} \right\} \tag{2}$$

The equation for $\hat{\alpha_j}, \hat{\beta_j}, \hat{\sigma_i^2}$ and $\hat{X_i}$ given by

$$\hat{\alpha_j} = \overline{y_j} - \hat{B_j}\overline{x}, \tag{3}$$

where $\overline{y_j} = \frac{1}{n}\sum_{i=1}^{n} y_{ji}$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\hat{X_i} = \frac{\lambda\sum_{i=1}^{n} x_i + \sum_{j=1}^{q}\beta_j\sum_{i=1}^{n}(y_{ji} - \alpha_j)}{\lambda + \sum_{j=1}^{q}\beta_j^2} \tag{4}$$

$$\hat{\beta_j} = \frac{(S_{yy} - \lambda S_{xx}) + \sqrt{(\lambda S_{xx} - S_{yy}^2 + 4\lambda S_{xy}^2}}{2S_{xy}}; \tag{5}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2;$$

$$S_{yy} = \sum_{i=1}^{n} (y_{ji} - \overline{y_j})^2.$$

and

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \\ \hat{\sigma}^2 &= \frac{1}{n-2} \left\{ \sum_{i=1}^{n} (x_i - X_i)^2 + \frac{1}{\lambda} \sum_{j=1}^{q} \sum_{i=1}^{n} (y_{ji} - \alpha_j - \beta_j X_i)^2 \right\} \end{aligned} \tag{6}$$

2.3 Covariance Matrix using Fisher Information Matrix

The covariance matrix of $\hat{\alpha_j}$ and $\hat{\beta_j}$ is derived using Fisher Information Matrix [19]. The Fisher information matrix is as follows,

$$F = \begin{pmatrix} \frac{p}{\lambda\sigma^2} & \frac{1}{\lambda\sigma^2} \sum_{i=1}^{p} \hat{X}_i \\ \frac{1}{\lambda\sigma^2} \sum_{i=1}^{p} \hat{X}_i & \frac{1}{\lambda\sigma^2} \sum_{i=1}^{p} \hat{X}_i^2 \end{pmatrix}. \tag{7}$$

Thus, the variance and covariance of $\hat{\alpha_j}$ and $\hat{\beta_j}$ are

$$V\hat{a}r(\hat{\alpha_j}) = \frac{\left(\lambda + \hat{\beta}_j^2\right) \hat{\sigma}^2 \hat{\beta}_j}{S_{xy}} \left\{ \bar{x}^2 (1 + \hat{T}) + \frac{S_{xy}}{p\hat{\beta}_j} \right\}, \tag{8}$$

$$V\hat{a}r(\hat{\beta_j}) = \frac{\left(\lambda + \hat{\beta}_j^2\right) \hat{\sigma}^2 \hat{\beta}_j}{S_{xy}} \{1 + \hat{T}\}. \tag{9}$$

$$C\hat{o}v(\hat{\alpha_j}, \hat{\beta_j}) = \frac{\left(\lambda + \hat{\beta}_j^2\right) \hat{\sigma}^2 \hat{\beta}_j \bar{x}}{S_{xy}} \{1 + \hat{T}\}, \tag{10}$$

where

$$\hat{T} = \frac{p\lambda\hat{\beta}_j\hat{\sigma^2}}{(\lambda + \hat{\beta}_j^2) S_{xy}}.$$

In this study, we propose that the determinant of the covariance matrix for the simultaneous LFRM is expressed as follows:

$$|COV|_j = V\hat{a}r(\hat{\alpha_j}) V\hat{a}r(\hat{\beta_j}) - C\hat{O}V(\alpha_j \hat{,} \hat{\beta}_j)^2. \tag{11}$$

where where Equation (8), (9) and (10) are utilised to obtain covariance matrix for the simultaneous LFRM, that is to be used in finding the cut-off point in outlier detection. Therefore, the covariance matrix is shown in Equation (12).

$$|COV|_j = \left(\frac{n}{\lambda\sigma_i^2} - \left(\frac{1}{\lambda\sigma_i^2}\sum_{i=1}^{n}\hat{X}_i\right)\left(\frac{1}{\lambda\sigma_i^2}\sum_{i=1}^{n}\hat{X}_i^2\right)^{-1}\left(\frac{1}{\lambda\sigma_i^2}\sum_{i=1}^{n}\hat{X}_i\right)\right)^{-1}$$
$$\frac{\left(\lambda+\hat{\beta}_j^2\right)\hat{\sigma}_i^2\hat{\beta}_j}{S_{xy}}\{1+\hat{T}\} - \left(\frac{(\lambda+\hat{\beta}_j^2)\hat{\sigma}_i^2\hat{\beta}_j\bar{x}}{S_{xy}}\{1+\hat{T}\}\right)^2 \tag{12}$$

2.4 Development of a Cut-off Formula for Detecting Outliers in Simultaneous Linear Functional Relationship Model

The $COVRATIO$ statistic for the i^{th} observation, where $j = 1, ...q$, is expressed as

$$|COVRATIO_{(-1)} - 1|_j = \frac{|COV|_j}{|COV_{(-1)}|_j} \tag{13}$$

- Numerator ($|COV|_j$): This represents the determinant of the covariance matrix for the complete dataset, which includes all observations.
- Denominator ($|COV_{(-1)}|_j$): This represents the determinant of the covariance matrix after removing the i^{th} observation.

The formula calculates the relative change in the determinant of the covariance matrix when a specific observation, i^{th} is removed [11,18]. If the calculated value exceeds a predefined cut-off threshold, it indicates that the i^{th} observation is an outlier [11, 18]. These cut-off values are determined through simulation studies to ensure their reliability and effectiveness in detecting outliers [11, 18].

2.5 Simulation Study for Developing of Cut-off Formula for Outlier Detection

From the Monte Carlo simulation method, the cut-off points for the $COVRATIO$ statistic are derived to detect outliers in the simultaneous LFRM when $q = 2$ [18]. For the simulation, sample sizes of $n = 30, 50, 100, 150, 200, 250,$ and 500 are used, along with five different standard deviation of error, $\tau_j = 0.2, 0.4, 0.6, 0.8$ and 1.0 , respectively. A set of normal random errors is generated from the normal distribution with mean 0 and error variance, τ_j^2 , respectively, for each sample size of n and τ_j [18]. Assuming the variance of the error term of δ_1, ϵ_1 and ϵ_2 to be equal, the following procedure is carried out:

First, generate a random X_i of size$n$, with $i = 1, 2, 3, ..., n$, where n is the sample size [11,18]. Without loss of generality, the slope and y-intercept parameters of simultaneous LFRM are fixed at $\alpha_1 = 0, \alpha_2 = 0, \beta_1 = 1$ and $\beta_2 = 1$, respectively. Then, generate three random errors δ_1, ϵ_1 and ϵ_2 from $\delta_1 \sim N(0, \sigma_i^2)$ and $\epsilon_j \sim N(0, \tau_j^2)$, respectively. Assuming the errors are equal. Next, compute the values of x, y_1 and y_2 . Afterward, fit the generated data to the parametric simultaneous LFRM and calculate $|COV|_j$ by using Equation (11).

The next step involves excluding the i^{th} row from the generated sample for y_1 and y_2 , where $i = 1, 2, 3, ..., n,$. Then, for all i, repeat step 4 till step 6 to obtain $|COV_{(-i)}|_1$ and $|COV_{(-i)}|_2$. After that, calculate $|COVRATIO_{(-i)}|_1$ and $|COVRATIO_{(-i)}|_2$ use equation (12) for all i [18].

Finally, identify the maximum value of $|COVRATIO_{(-i)} - 1|_1$ for y_1 and $|COVRATIO_{(-i)} - 1|_2$ for y_2 for all i.

This process is repeated 10,000 times for each combination of sample size n and standard deviation of error, τ_j [18]. Following this, the upper percentiles of the maximum values of $|COVRATIO_{(-i)} - 1|_1$ and $|COVRATIO_{(-i)} - 1|_2$ for all i are computed [11,18]. Power series equations derived from the graph of these percentiles are used as cut-off points for detecting outliers in the simultaneous LFRM for y_1 and y_2 [18].

3 Results and Discussion

3.1 Power Series in Finding the Cut-off Points for Outlier Detection

The cut-off points for y_1 and y_2 are determined at 5% upper percentile by graphing the power series curves [18]. The 5% upper percentiles of the highest values of $|COVRATIO_{(-i)} - 1|_1$ and $|COVRATIO_{(-i)} - 1|_2$ for each τ_j are averaged [11, 18]. The power series graph curves of 5% upper percentile are shown as illustrated in Figure 1 and Figure 2. The R^2 value, or the coefficient of determination, in a power series graph indicates the quality of fit for the regression model. It reflects the proportion of variance in the dependent variable explained by the independent variables. The graphs reveal that R^2 values are nearly 1, indicating a strong fit.

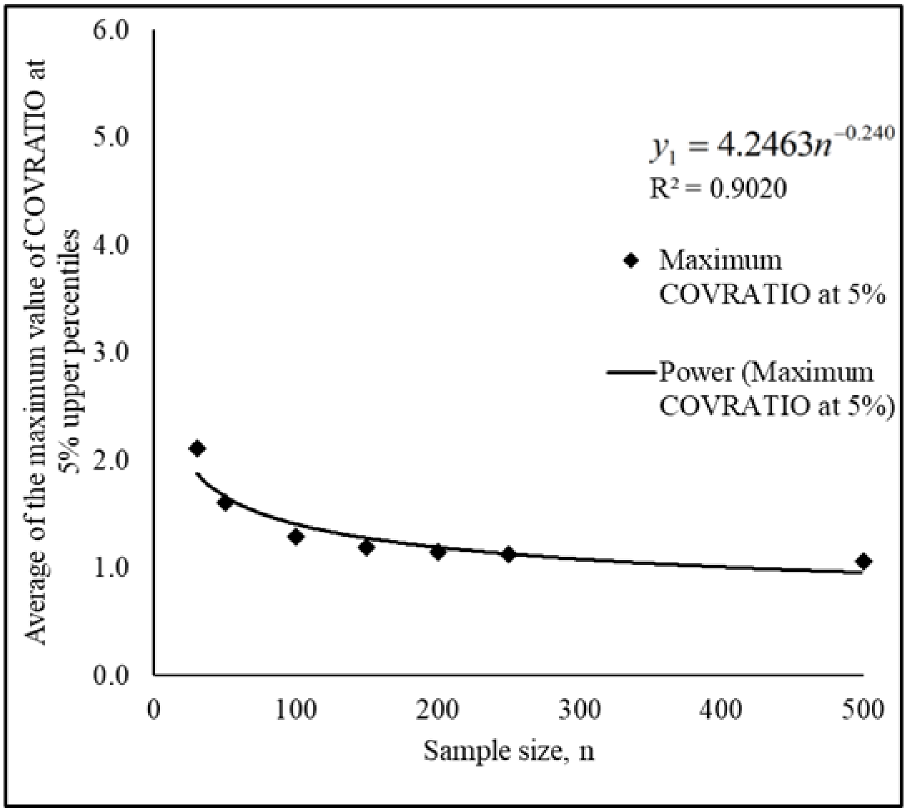

Figure 1: Power series graph used to establish the cut-off points formula for the 5% upper percentiles of y_1.

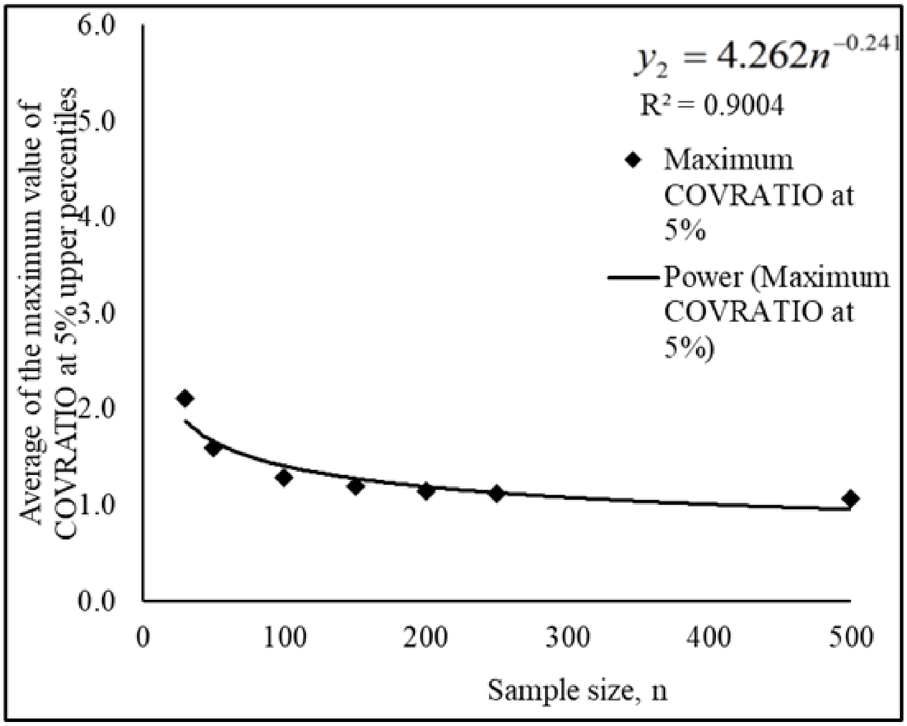

Figure 2: Power series graph used to establish the cut-off points formula for the 5% upper percentiles of y_2.

The curves are fitted using the power series formula through the least squares method [18]. The cut-off points formula is obtained as $y_1 = 4.2463n^{-0.24}$ and $y_2 = 4.262n^{-0.241}$ for 5% upper percentiles show that the average of the maximum values decreases as the sample size, n increase. From Figure 1 and Figure 2, the R^2 values are closer to 1, which indicate that the power series models provide a strong fit for the data, confirming the reliability of the formulas in capturing the cut-off trends. These findings are critical for identifying appropriate thresholds in statistical analyses, particularly for detecting outliers.

3.2 Power of Performance of *COVRATIO* statistic

The power of performance of highest values of $|COVRATIO_{(-i)} - 1|_1$ for y_1 and $|COVRATIO_{(-i)} - 1|_2$ for y_2 is examined using the Monte Carlo simulation method to evaluate the effectiveness of identifying an outlier in the simultaneous LFRM. Three distinct sample sizes, namely $n = 50, 70, 150$ and 500 are employed accordingly. Outliers are randomly introduced at a specific observation, such as the d^{th} observation in assessing the performance of the *COVRATIO* statistic [11,18]. In this study, we introduce the outlier during the 20^{th} observation. For the outlier at the 20^{th} observation, we generate the data of y_1 and y_2 from the normal distribution with mean zero and variance of contamination, σ_δ^2, where $\sigma_\delta^2 = 0, 6, 8, 10, 12, 14$ and 16, respectively [8, 11, 20].

The simulated data is then fitted into the simultaneous LFRM using Equation (1), and then the $|COV|_1$ and $|COV|_2$ are computed using equation (12). Later on, the i^{th} row is subsequently removed from the sample of y_1 and y_2, where $i = 1, ..., n$. The removed data is refitted by computing $|COVRATIO_{(-i)} - 1|_1$ and $|COVRATIO_{(-i)} - 1|_2$ using equation (13) [18]. The highest values of the $|COVRATIO_{(-i)} - 1|_1$ statistic of y_1 and $|COVRATIO_{(-i)} - 1|_2$ statistic of y_2 are determined by estimating the percentage of correctly detecting the contaminated observation at the 20^{th} observation [11,18].

The power of performance for a fixed τ_j is explored, and the sample sizes, n are varied [18]. Table 1 presents the power of performance of $|COVRATIO_{(-i)} - 1|_1$ for y_1 and $|COVRATIO_{(-i)} - 1|_2$ for y_2 when $n = 50, 70, 150$ and 500.

Table 1: Power of performance for highest values of $|COVRATIO_{(-i)} - 1|_1$ and $|COVRATIO_{(-i)} - 1|_2$ when $\tau_i = 0.2$ for y_1 and y_2 .

n	Variance of contamination,σ_δ^2	Power of performance for y_1	Power of performance for y_2
50	0	99.75	99.76
	2	99.98	99.98
	4	99.99	99.99
	6	100	100
	8	100	100
	10	100	100
	12	100	100
70	0	99.76	99.77
	2	99.99	99.99
	4	99.99	99.99
	6	100	100
	8	100	100
	10	100	100
	12	100	100
150	0	99.8	99.85
	2	99.99	99.99
	4	100	100
	6	100	100
	8	100	100
	10	100	100
	12	100	100
500	0	99.85	99.86
	2	100	100
	4	100	100
	6	100	100
	8	100	100
	10	100	100
	12	100	100

Table 1 presents the power of performance for highest values of $|COVRATIO_{(-i)} - 1|_1$ and $|COVRATIO_{(-i)} - 1|_2$ when $\tau_j = 0.2$ for y_1 and y_2 when $n = 50, 70, 150$ and 500. This outcome suggests that the power of performance increases as the variance of contamination, σ_δ^2

increasing suggests that the ability to detect outliers improves as the variance of contamination increases. When $\tau_j = 0.4, 0.6, 0.8,$ and 1.0 , all the results consistently showed similar outcomes to those for $\tau_j = 0.2$.

3.3 Malacca Environmental Datasets

The datasets, obtained from the Malaysian Meteorological Department (MDD) and organised in Microsoft Excel, consist of daily maximum wind speed, 24-hour average relative humidity, and 24-hour average temperature recorded during the southwest monsoon season in Malacca from May 18, 2020, to September 22, 2020. With the sample size, n of 128, the wind speed of Malacca at time t is addressed as the variable x_t , the mean relative humidity of Malacca at time as , and the mean temperature data of Malacca at time t as $y_{2,t}$. This study seeks to explore the relationship between daily maximum wind speed, 24-hour average humidity, and 24-hour average temperature in Malacca during the southwest monsoon season and to model this relationship using a simultaneous Linear Functional Relationship Model (LFRM) for linear variables. Table 2 summarises the descriptive statistics for these variables, including their mean, standard deviation, and range (minimum and maximum values). These statistics highlight the variability in environmental conditions during the monsoon season.

Table 2: Descriptive statistics summarising wind speed, humidity, and temperature during the southwest monsoon season of 2020 in Malacca.

Type of data	Mean	Standard deviation	Minimum	Maximum
Wind speed (m/s)	11.3532	2.2589	5.3	16.6
Humidity (%)	74.0071	5.9506	64.7	92.0
Temperature (°C)	27.7214	1.0667	23.3	29.9

The descriptive statistics reveal distinct characteristics of the environmental data during the southwest monsoon season in 2020 for Malacca. Wind speed shows moderate variability, with an average of 11.35 m/s and a range of 5.30 to 16.60 m/s. Humidity exhibits significant fluctuations, averaging 74.01% and ranging from 64.70% to 92%. In contrast, temperature remains relatively stable, with a mean of 27.72 °C and a narrower range of 23.30 to 29.90 °C. These patterns highlight the dynamic nature of wind speed and humidity compared to the steadiness of temperature, emphasizing the need for robust outlier detection methods in such datasets.

3.4 Statistical Test for Normality using Kolmogorov-Smirnov Test Statistic for Malacca Wind Speed, Humidity, and Temperature Data

The normality of Malacca's wind speed, humidity, and temperature data is evaluated using the Kolmogorov-Smirnov test [19]. The Kolmogorov-Smirnov test statistic, D is defined as Equation (14), where F is the theoretical cumulative distribution.

$$D = \max_{1 \le i \le N} \left(F\left(Y_i\right) - \frac{i-1}{N}, \frac{i}{N} - F\left(Y_i\right) \right) \tag{14}$$

The equation critical value of D when $\alpha = 0.05$, and the sample size is over 35, is $\frac{1.36}{\sqrt{n}}$ [19, 21, 22]. Insert the value of the sample size, 128, into the equation $D = \frac{1.36}{\sqrt{128}}$. Hence, the critical value is 0.1202 [19]. H_0 is rejected if D exceeds the critical value [12,19]. The following are the null, H_0, and alternative, H_A, hypotheses used in a Kolmogorov-Smirnov test statistic, respectively:

- H_0:The data is normally distributed.
- H_A:The data is not normally distributed.

Table 3 presents the Kolmogorov-Smirnov test statistics for wind speed, humidity, and temperature during the southwest monsoon season of 2020 in Malacca.

Table 3: Kolmogorov-Smirnov test statistic (D) for wind speed, humidity, and temperature throughout the southwest monsoon season in 2020 for Malacca.

Type of data	Kolmogorov-Smirnov test statistic (D)
Wind speed	0.1186
Humidity	0.0480
Temperature	0.0713

As shown in Table 3, all D values for Malacca are below the critical value, indicating that the null hypothesis (H_0) cannot be rejected. This suggests that the wind speed, humidity, and temperature data for Malacca during the southwest monsoon in 2020 can be assumed to follow a normal distribution. Consequently, the simultaneous LFRM employed in this study effectively represents the relationship between wind speed, humidity, and temperature during this period.

3.5 Graphical Tool for Normality using Q-Q plot for Malacca Wind Speed, Humidity and Temperature Data

Q-Q plots are constructed for wind speed, humidity, and temperature data in Malacca during the southwest monsoon of 2020 to assess their goodness-of-fit to the normal distribution [19]. These plots visually represent the data distribution, where points aligning with the reference line indicate conformity to a normal distribution. The Q-Q plots for wind speed, humidity, and temperature are displayed in Figures 3, 4, and 5, respectively.

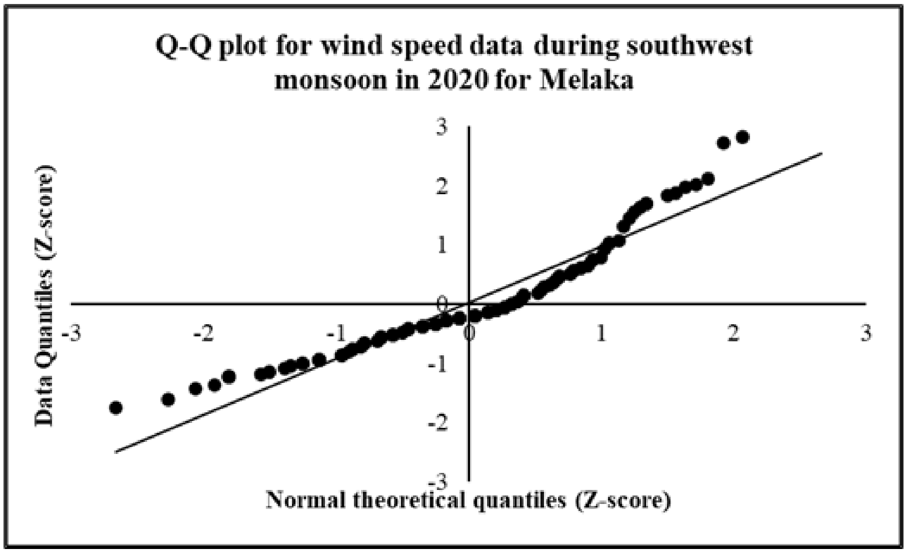

Figure 3: Q-Q plot for wind speed data during southwest monsoon in 2020 for Malacca.

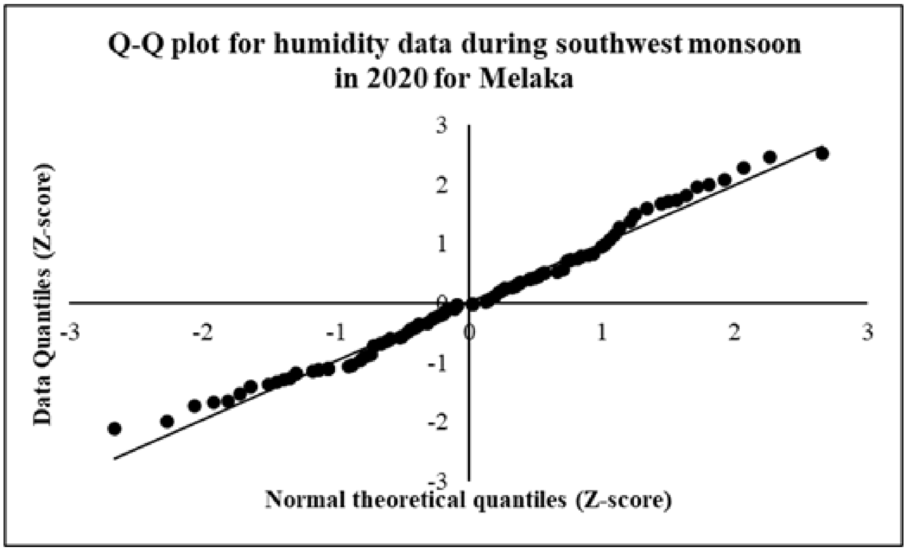

Figure 4: Q-Q plot for humidity data during the southwest monsoon in 2020 for Malacca.

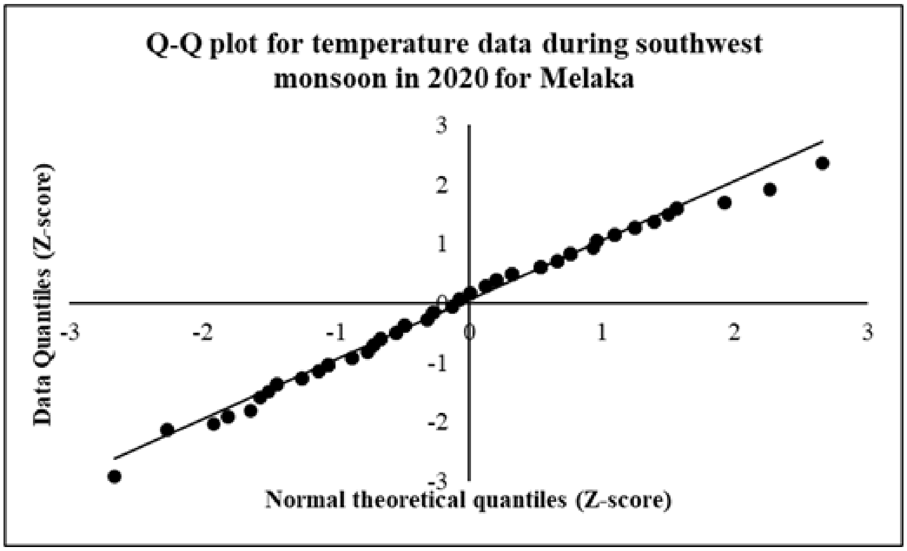

Figure 5: Q-Q plot for temperature data during the southwest monsoon in 2020 for Malacca.

Both the Kolmogorov-Smirnov test statistics and the Q-Q plots confirm that the wind speed, humidity, and temperature data for Malacca during the southwest monsoon in 2020 can be reasonably modelled as following a normal distribution [19].

3.6 Application to Real Data

The relationship between the wind speed (x_t), humidity ($y_{1,t}$), and temperature ($y_{2,t}$) is presented in Table 4.

Table 4: The simultaneous LFRM employed to real data set.

Type of relationship	Simultaneous LFRM
Humidity with wind speed	$y_{1,t} = \alpha_1 + \beta_1 x_t$
Temperature with wind speed	$y_{2,t} = \alpha_2 + \beta_2 x_t$

Table 5 presents the parameter estimates for wind speed, humidity, and temperature recorded in Malacca during the southwest monsoon of 2020, as obtained by fitting a simultaneous functional relationship model for linear variables [19].

Table 5: The simultaneous LFRM employed to real data set.

Parameter estimate	Value of parameter estimate	
$\hat{\alpha_j}$	$\hat{\alpha_1} = 13.9429$	$\hat{\alpha_2} = 28.9451$
$\hat{\beta_j}$	$\hat{\beta_1} = 6.7536$	$\hat{\beta_2} = -0.1217$
σ_j^2	$\sigma_1^2 = 4.3346$	$\sigma_2^2 = 0.7793$
$\text{Var}(\hat{\alpha_j})$	$\text{Var}(\hat{\alpha_1}) = 1113.2165$	$\text{Var}(\hat{\alpha_2}) = 0.2005$
$\text{Var}(\hat{\beta_j})$	$\text{Var}(\hat{\beta_1}) = 11.5343$	$\text{Var}(\hat{\beta_2}) = 0.0020$
$\text{Cov}(\hat{\alpha}_j, \hat{\beta}_j)$	$\text{Cov}(\hat{\alpha}_1, \hat{\beta}_1) = 0$	$\text{Cov}(\hat{\alpha}_2, \hat{\beta}_2) = 0$

According to Table 5, the simultaneous LFRM for wind speed, humidity, and temperature recorded in Malacca during the 2020 southwest monsoon is given by $y_{1,t} = 13.9429 + 6.7436x_t$ and $y_{2,t} = 28.9451 - 0.1217x_t$. The positive relationship between wind speed, x_t and humidity, $y_{1,t}$ suggests that as wind speed increases, humidity also tends to increase, which could be due to various meteorological factors such as the wind bringing in more moist air from other areas. The negative relationship between wind speed, x_t and humidity,$y_{2,t}$ suggests that an increase in wind speed is associated with a decrease in temperature, possibly due to the wind chill effect or the dispersion of heat through increased air movement. This theoretical investigation aims to illustrate the potential utility of simultaneous LFRM involving humidity, wind speed, and temperature variables. While empirical data-driven analyses are commonly employed in environmental studies, this study focuses on demonstrating the conceptual framework and methodological approach rather than conducting exhaustive data analysis. Next, the variance of $\hat{\alpha_1}$ and $\hat{\beta_1}$ are relatively high. A notably high variance may warrant a more detailed review of the data to ensure the accuracy and reliability of the parameter estimation. Outliers or measurement errors could be responsible for the unusually high variance values, potentially affecting the reliability of the parameter estimates. The variance of $\hat{\alpha_2}$ and $\hat{\beta_2}$ are relatively small and indicates good estimation for $\hat{\alpha_2}$ and $\hat{\beta_2}$.

3.7 COVRATIO statistic to Detect Outlier in Synthetic Data for Simultaneous Linear Functional Relationship Model

Synthetic data is artificially created to resemble real-world data, though it is not generated from actual observations. It is produced using statistical models, algorithms, or other techniques to simulate the behavior of real data. In this study, we created synthetic data. For demonstration, we generate synthetic data points from simultaneous LFRM. The parameters $\alpha_1 = 0, \alpha_2 = 0, \beta_1 = 1, \beta_2 = 1, \lambda = 1, \mu = 0$ and $\sigma_\delta^2 = \sigma_\epsilon^2 = 0.4^2$. The 20^{th} data point of the synthetic dataset was intentionally contaminated by creating the contamination using $\epsilon_i \sim N(0, 16)$ [18]. The scatter plots of the synthetic datasets for $y_{1,t}$ and $y_{2,t}$ including the contaminated observation, as shown in Figure 6 and Figure 7.

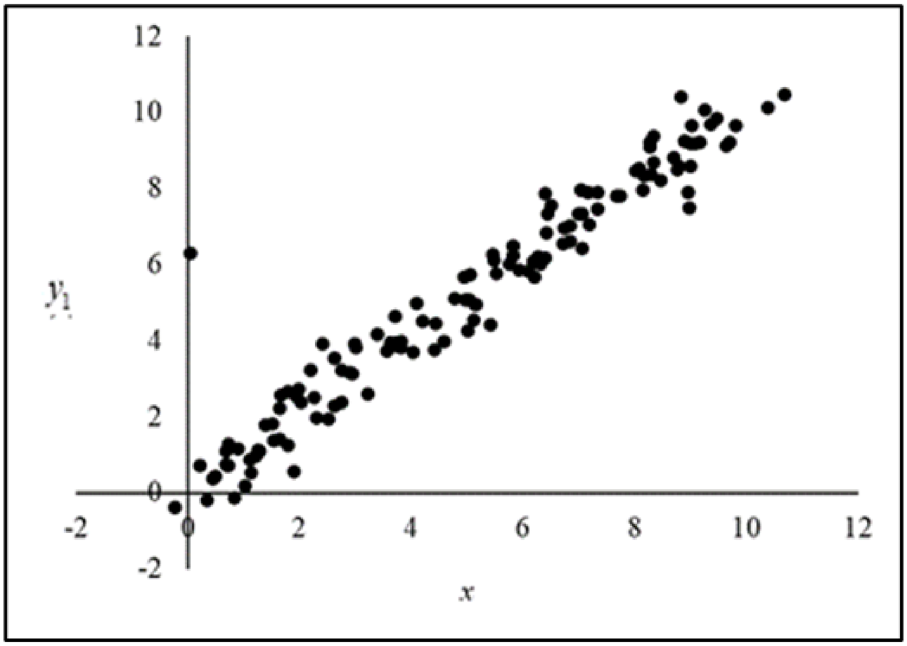

Figure 6: The scatter plot for the synthetic data for $y_{1,t}$.

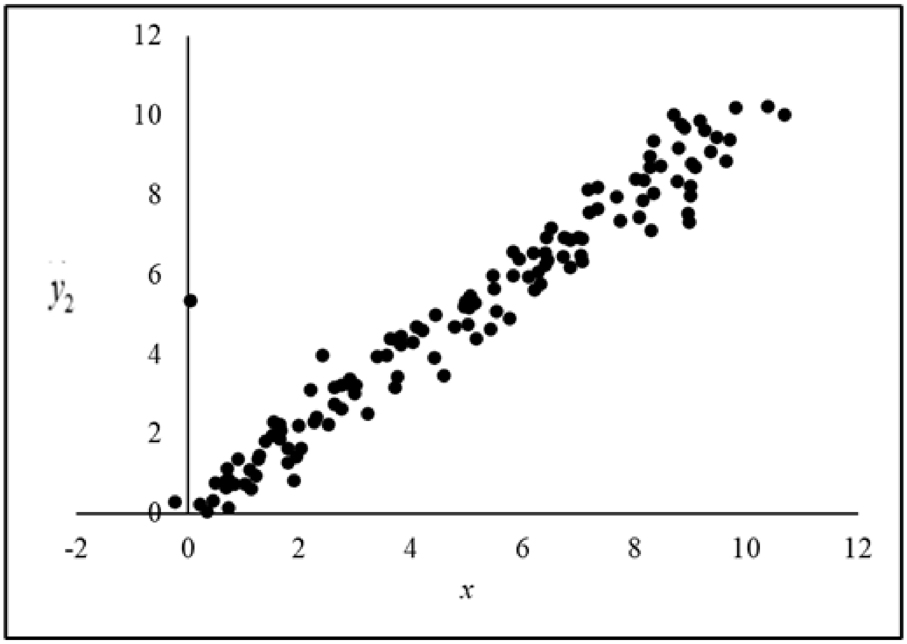

Figure 7: The scatter plot for the synthetic data for $y_{2,t}$.

The highest values of $|COVRATIO_{(-i)} - 1|_1$ and $|COVRATIO_{(-i)} - 1|_2$ are calculated. Based on the formula in Table 1, the cut-off point for $y_{1,t}$ is 1.3203, while for $y_{2,t}$ is 1.3187, corresponds to the 5% upper percentile when $n = 130$. The $COVRATIO$ statistic is then used to identify outliers by plotting these highest values of $|COVRATIO_{(-i)}-1|_1$ and $|COVRATIO_{(-i)}-1|_2$ against the index. Figure 8 and Figure 9 illustrate that the 20^{th} observation exceeds the cut-off points for both $y_{1,t}$ and $y_{2,t}$. In conclusion, the formulated $COVRATIO$ statistic is effective in detecting outliers within the simultaneous LFRM datasets.

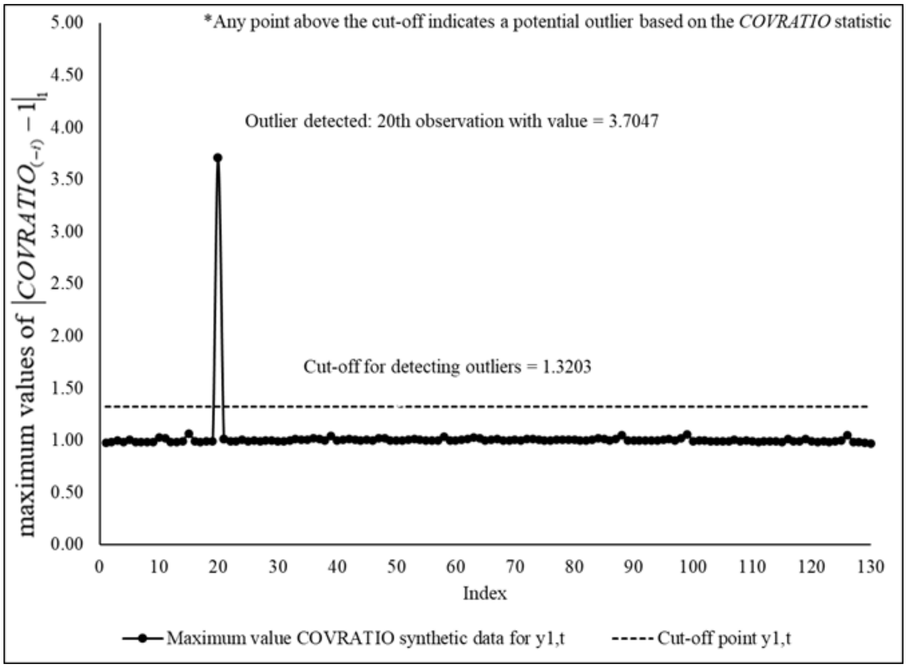

Figure 8: Graph of maximum values of $|COVRATIO_{(-i)} - 1|_1$ for synthetic data for $y_{1,t}$.

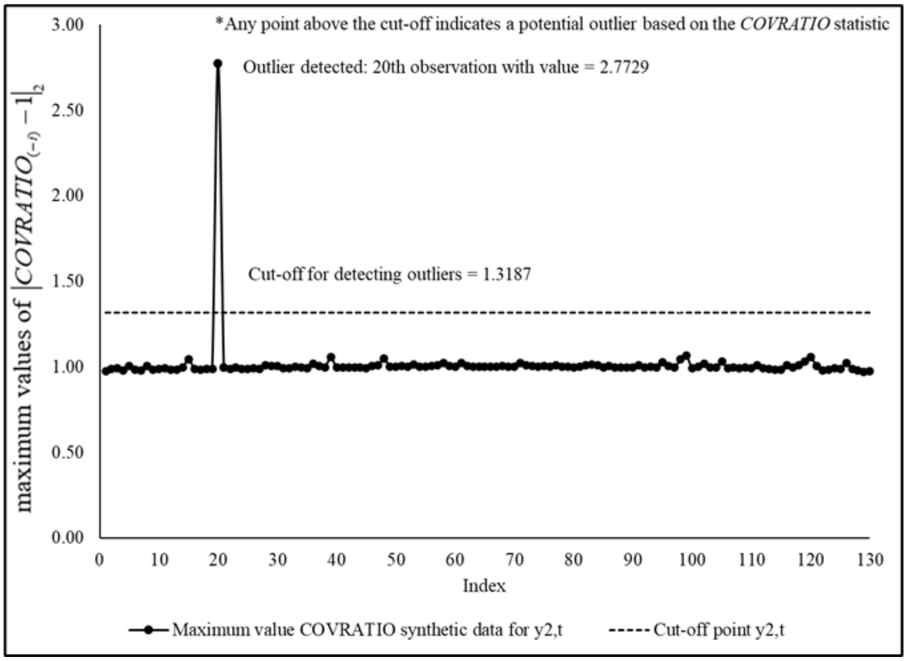

Figure 9: Graph of maximum values of $|COVRATIO_{(-i)} - 1|_2$ for synthetic data for $y_{2,t}$.

3.8 Detection of Outlier using New *COVRATIO* statistic in Simultaneous Linear Functional Relationship Model and Its Application in Malacca Environmental Datasets

The *COVRATIO* statistic is employed to detect outliers in environmental datasets from Malacca within the simultaneous LFRM. To verify the presence of an outlier, the COVRATIO statistic is used by plotting the highest values of $|COVRATIO_{(-i)} - 1|_1$ and $|COVRATIO_{(-i)} - 1|_2$ against the data point or index, as shown in Figure 10 and Figure 11. Based on formula in Table 1, the cut-off point at the 5% upper percentiles for $y_{1,t}$ and $y_{2,t}$ are 1.3252 and 1.3237, respectively, for a sample size of $n = 128$.

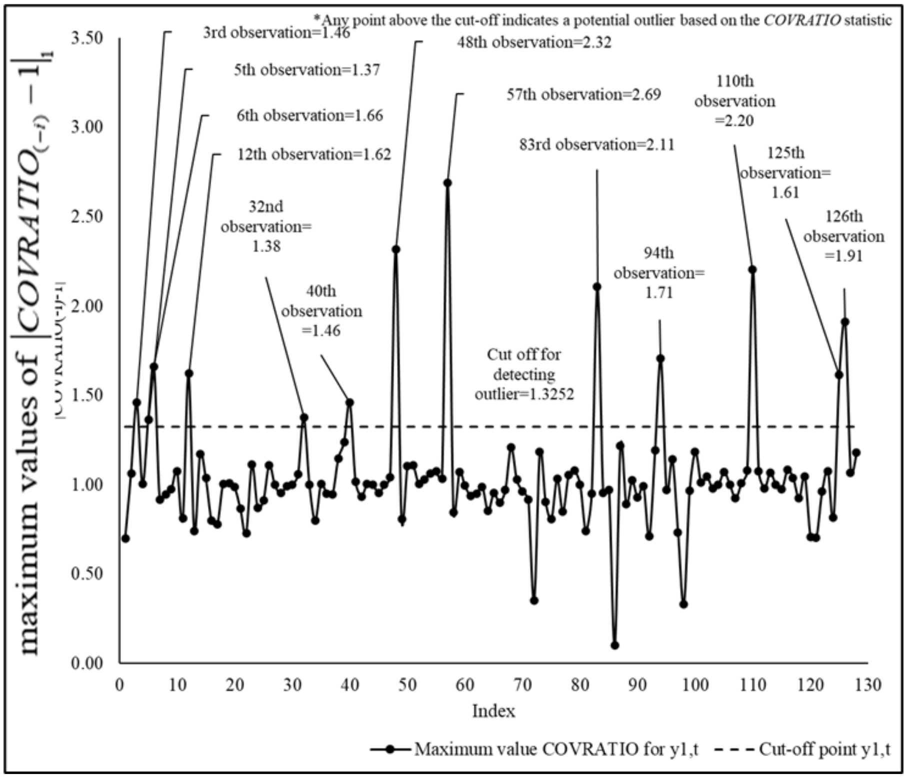

Figure 10: Graph of maximum values of $|COVRATIO_{(-i)} - 1|_1$ for wind speed and humidity.

From Figure 10, it can be seen that the $3^{rd}, 5^{th}, 6^{th}, 12^{th}, 32^{nd}, 40^{th}, 48^{th}, 57^{th}, 83^{rd}, 94^{th}, 110^{th}, 125^{th}$ and 126^{th} observation has highest values of $|COVRATIO_{(-i)} - 1|_1$ surpass the cut-off point of 1.3252. Therefore, its indicate the observations are the outliers for wind speed and humidity.

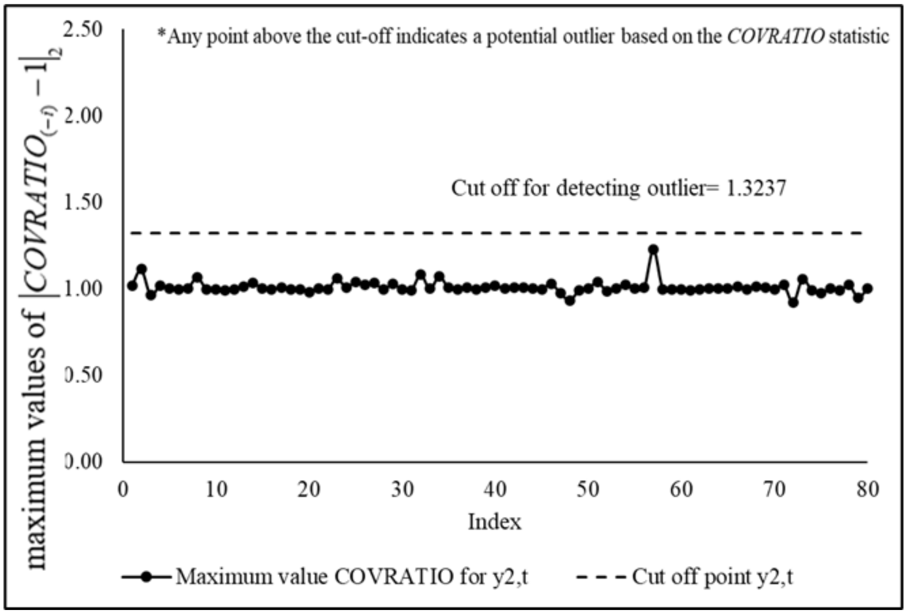

Figure 11: Graph of maximum values of $|COVRATIO_{(-i)} - 1|_2$ for wind speed and temperature.

From Figure 11, it can be seen that no observation has highest values of $|COVRATIO_{(-i)} - 1|_2$ exceeding the cut-off point of 1.3237. Therefore, no observations are outliers for the relationship between wind speed and temperature. A comparison of the parameter variance is conducted with and without the presence of outliers as shown in Table 6.

Table 6: : Parameter estimates and variance of the estimates when outliers are included in the data and when outliers are removed from the data of Malacca Environmental Datasets.

Outliers are included in the data, $n = 128$			Outliers are removed from the data, $n = 115$		
Parameter estimation		Variance	Parameter estimation		Variance
$\hat{\alpha}_1$	13.9429	1113.2165	$\hat{\alpha}_1$	54.5828	22.0871
$\hat{\beta}_1$	6.7436	11.5343	$\hat{\beta}_1$	2.5472	0.2252
$\hat{\alpha}_2$	28.9451	0.2005	$\hat{\alpha}_2$	29.9439	0.1661
$\hat{\beta}_2$	-0.1217	0.0020	$\hat{\beta}_2$	-0.2142	0.0017

The results in Table 6 reveal the substantial impact of outliers on parameter estimates and their variances when analyzing the Malacca environmental datasets. Parameter estimates, such as $\hat{\alpha_1}$ and $\hat{\beta_1}$, show significant shifts after outlier removal, with increasing from 13.9429 to 54.5828 and $\hat{\beta_1}$ decreasing from 6.7436 to 2.5472. Additionally, the variances of these estimates are dramatically reduced, such as the variance of α_1 , which drops from 1113.2165 to 22.0871, reflecting a notable increase in precision and reliability. Similarly, for α_2 , the estimate remains relatively consistent (28.9451 versus 29.9439), but its variance decreases markedly from 0.2005 to 0.1661, further reinforcing the reliability of the estimate. For β_2 , the estimate slightly shifts from -0.1217 to -0.2142 after outlier removal, while its variance decreases significantly from 0.0020 to 0.0017, indicating improved stability and precision. These findings indicate that outliers severely distort parameter estimates and increase variability, compromising model stability and predictive power. Removing outliers using the *COVRATIO* statistic enhances the robustness of the model, yielding a more reliable representation of environmental dynamics during the monsoon season. This highlights the critical role of outlier detection in improving model accuracy and ensuring informed decision-making for environmental monitoring and policy development. Therefore, the new and more appropriate simultaneous LFRM of Malacca environmental datasets after the outliers are removed are $y_{1,t} = 54.5828 + 2.5472x_t$ and $y_{2,t} = 29.9439 - 0.2142x_t$.

4 Conclusions

This paper introduces the *COVRATIO* statistic as a method to detect outliers in simultaneous LFRM. The cut-off points for the *COVRATIO* statistic are established through the Monte Carlo simulation method. Any data point exceeding these cut-off points is classified as an outlier. The simulation study and its application to real-world datasets demonstrate that the cut-off values at the 5% upper percentiles are obtained by $y_1 = 4.2463n^{-0.24}$ and $y_2 = 4.262n^{-0.241}$. These cut-off point effectively outliers in linear data, which can be modelled by simultaneous LFRM. The extended model's performance was assessed, showing that as the standard deviation of error, τ_j decreases, the ability to accurately identify outliers improves. We demonstrate the applicability of the simultaneous LFRM through the analysis of real-world environmental datasets, incorporating variables such as wind speed, humidity, and temperature. In both synthetic and real-world data applications, the parameter estimates become more accurate with smaller variance when the outliers are detected and excluded. Hence, it is crucial to recognise outliers to facilitate their exclusion from the dataset, thereby enhancing the reliability of the proposed method. Beyond environmental datasets, this method could be applied to fields such as economics, engineering, and medicine, where simultaneous linear relationships between multiple variables are commonly modelled, and outlier detection is crucial for accurate predictions. This study assumes normality of the data and relies on simulation-based cut-off points, which may limit its applicability to other datasets. Specifically, non-normal datasets may exhibit skewness or heavy tails, which could impact the accuracy of outlier detection and parameter estimation. One possible strategy to address this challenge is to develop robust methods that adjust for data distribution characteristics or to use transformation techniques to approximate normality. Future research could focus on extending the method to accommodate non-normal datasets by investigating specific types of distributions such as skewed, heavy-tailed, or multimodal distributions. Additionally, exploring robust estimation techniques that are less sensitive to distributional assumptions could prove valuable. The method could also be applied to more complex models, including multivariate time series models and hierarchical models, to assess its effectiveness in capturing simultaneous linear relationships in dynamic and nested data structures.

Acknowledgments

The authors would like to thank the Ministry of Higher Education Malaysia for supporting this study through the Fundamental Research Grant Scheme, code: FRGS/1/2021/STG06/UITM/02/5.

References

[1] Mokhtar, N. A., Zubairi, Y. Z., & Hussin, A. G.. A Clustering Approach to Detect Multiple Outliers in Linear Functional Relationship Model for Circular Data. *Journal of Applied Statistics*. 2018. 45(6), 1041-1051.

[2] Hampel FR, Ronchetti EM, Rousseeuw P, Stahel WA. Robust Statistics: The Approach Based on Influence Functions. *John Wiley & Sons*. 2011. 536.

[3] Arif AM, Zubairi YZ, Hussin AG. Maximum Likelihood Estimation of Replicated Linear Functional Relationship Model. *Appl Math Comput Intell*. 2021b.10(1):3018.

[4] Grubbs FE. Procedures for Detecting Outlying Observations in Samples. *Technometrics*. 1969.11(1):121.

[5] Sebert DM, Montgomery DC, Rollier DA. A Clustering Algorithm for Identifying Multiple Outliers in Linear Regression. *Comput Stat Data Anal*. 1998;27(4):46184.

[6] Adnan R, Nor Mohamad M, Setan H. Multiple Outliers Detection Procedures in Linear Regression. *Matematika*. 2003.19(1):2945.

[7] Arif AM, Zubairi YZ, Hussin AG. COVRATIO Statistic for Replicated Linear Functional Relationship Model. *J Phys Conf Ser*. 2021a.1988(1).

[8] Arif AM, Zubairi YZ, Hussin AG. Outlier Detection in Balanced Replicated Linear Functional Relationship Model. *Sains Malays*. 2022;51(2):599607.

[9] Mokhtar NA, Zubairi YZ, Hussin AG, Moslim NH. An Outlier Detection Method for Circular Linear Functional Relationship Model Using Covratio Statistics. *Malays J Sci*. 2019;38(2):4654.

[10] Mokhtar NA, Badyalina B, Chang KL, Yaacob FF, Ghazali AF, Shamala P. Error-in-Variables Model of Malacca Wind Direction Data with the Von Mises Distribution in Southwest Monsoon. *Appl Math Sci*. 2021b;15(9):4719.

[11] Ghapor AA, Zubairi YZ, Mamun ASMA, Imon AHMR. On Detecting Outlier in Simple Linear Functional Relationship Model Using COVRATIO Statistic. *Pak J Stat*. 2014;30(1):12942.

[12] Jamaliyatul NAAH, Badyalina B, Mokhtar NA, Rambli A, Zubairi YZ, Abdul Ghapor A. Modelling Wind Speed Data in Pulau Langkawi with Functional Relationship. *Sains Malays*. 2023;52(8):241930.

[13] Mokhtar NA, Zubairi YZ, Hussin AG, Yunus RM. On parameter estimation of a replicated linear functional relationship model for circular variables. *Matematika*. 2017;15963.

[14] Arif AM, Zubairi YZ, Hussin AG. Parameter estimation in replicated linear functional relationship model in the presence of outliers. *Malays J Fundam Appl Sci*. 2020a;16(2):15860.

[15] Jamaliyatul, N. A. A. H., Mokhtar, N. A., Badyalina, B., Rambli, A., & Zubairi, Y. Z. . Statistical model of Malacca wind speed data with functional relationship and error terms. In *AIP Conference Proceedings*.2024, Vol 3150, No. 1). `https://pubs.aip.org/aip/acp/article-abstract/3150/1/040008/3312423`.

[16] Ghapor AA, Zubairi YZ, Mamun ASMA, Imon AHMR. A robust nonparametric slope estimation in linear functional relationship model. *Pak J Stat*. 2015;48(1):33950.

[17] Doganaksoy N, Van Meer H. An Application of the Linear Errors-in-Variables Assessment an Application of the Linear Errors-in-Variables Model in Semiconductor Device Performance Assessment. *Qual Eng*. 2015;27(4):50011.

[18] Ghapor AA. Parameter Estimation and Outlier Detection in Linear Functional Relationship Model [PhD. Thesis]. Universiti of Malaya; 2017.

[19] Jamaliyatul NAAH, Mokhtar NA, Badyalina B, Rambli A, Zubairi YZ. Modelling Wind Speed, Humidity, and Temperature in Butterworth and Melaka during Southwest Monsoon in 2020 with a Simultaneous Linear Functional Relationship. *Malays J Fundam Appl Sci*. 2024;20(2):30219.

[20] Mamun ASMA, Zubairi YZ, Hussin AG, Imon AHMR, Rana S, Carrasco J. Identification of Influential Observation in Linear Structural Relationship Model with Known Slope. *Commun Stat Simul Comput*. 2019;51(1):7283.

[21] Massey FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. *J Am Stat Assoc*. 1951;46(253):68.

[22] Hawkins DI, Kanji GK. 100 Statistical Tests. *J Mark Res*. 1995;32(1):112.