Available online at www.garmian.edu.krd

Passer Journal of Basic and Applied Sciences

Passer 7 (Issue 1) (2025) 478-494

Original Article

Wavelet Analysis for Outlier Estimation in Multivariate Linear Regression Models

Amira Wali Omer[1], Taha Hussein Ali[*1]

[1]*Department of Statistics and Informatics, College of Administration and Economics, University of Salahaddin, Erbil, Kurdistan Reign, Iraq.*

Received 28 February 2025; revised 15 March 2025;
accepted 14 April 2025; available online 23 June 2025

DOI: 10.24271/PSR.2025.509315.1976

ABSTRACT

Outliers significantly impact parameter estimation accuracy in multivariate linear regression models, so traditional methods, such as the Hampel filter, are used to address this problem. This study proposes a wavelet-based approach to address outliers by estimating these values based on the discrete wavelet transform (DWT) of the wavelets (Daubechies, Coiflets, and Dmey) and calculating the approximation and detail coefficients (Low and High pass filter, respectively) which used in estimating the parameters of the multivariate regression model for the approximation and detail coefficients and through these models the predictive values of these coefficients calculated the inverse of the discrete wavelet transform of the predictive values taken to obtain the filtered observations. Then, the filtered predictive values corresponding to the outliers are the estimated observations that address the outlier problem. The accuracy of the estimated parameters of the multivariate linear regression model using the proposed methods and the Hampel filter was compared through some accuracy criteria of the estimated models, including the mean absolute error (MAE) of the estimated parameters, mean squared error (MSE), and coefficient of determination (R^2)). Simulation experiments and real-world blood test data confirm the superiority of the wavelet-based approach, decreasing MSE by 0.7728 and improving R^2 by 83.19% for Y1, (74.33%) for Y2, and (87.50%) for Y3, compared to the Hampel filter. These results demonstrate that wavelet-based filtering provides a more robust and accurate alternative for handling outliers in multivariate regression models.

Keywords: Multivariate Linear Regression Models, Outliers, Hampel Filter, Wavelets, and Discrete Wavelet Transformation.

1 Introduction

In statistical analysis, multivariate linear regression is a key method used to represent how multiple dependent variables relate to independent ones. However, outliers can distort these relationships significantly, causing predictions and interpretations to be inaccurate. This paper investigates three advanced techniques, including wavelet analysis and the Hampel filter, to address these problems. The main goal is to compare how well these methods improve the reliability of multivariate linear regression models (MLRM). MLRM has been studied and applied in economics, biology, and engineering [1]. Even though it has become a go-to choice for many in practice, the technique remains sensitive to outliers, which can distort results [2].

An outlier is an observation that diverges significantly from the overall distribution of the dataset, often suggesting the influence of an alternative generative mechanism or an anomaly in the data collection process. Outliers are also called abnormalities, discordant values, deviants, or anomalies in data mining and statistics. Often, the data arises from one or more generating processes that show system activity or observations on entities. Unusual behavior in these processes can create outliers. Thus, outliers can provide valuable information about abnormal traits of the systems and entities that affect the data generation process. Recognizing these unusual traits can lead to helpful sights [3].

The Hampel filter is one method for addressing outliers in MLRM. It replaces each data point with a robust estimate of central tendency from a local area, reducing or eliminating outliers. The Hampel filter does not need any data decomposition and can be applied directly to the original dataset [4]. Named after John R. Hampel, who adeptly detected anomalies in datasets, the Hampel filter refines data replacing outliers with more suitable values, preventing extreme data points from distorting the regression model [5].

Wavelet analysis is another effective technique for managing outliers in MLRM analysis. It decomposes the data into various frequency bands using wavelet transforms. Approximate coefficients act as a low-pass filter, while detail coefficients function as a high-pass filter [6].

* Corresponding author
E-mail address taha.ali@su.edu.krd (Instructor).
Peer-reviewed under the responsibility of the University of Garmian.

Both wavelet analysis and the Hampel filter are advanced techniques that help clarify intricate data in MLRM, each playing a unique role in data refinement. These methods have been effectively utilized in analyzing MLRM data [7,8]. Wavelet analysis is adept at estimating outliers while maintaining key localized features, whereas the Hampel filter provides a simpler and more efficient way to remove outliers [9]. Although previous studies have explored both methods, the novelty of this research lies in its comparison of these techniques and its evaluation of their performance within a unified framework, providing a clearer understanding of their relative strengths in outlier detection for MLRM [10].

This article's primary focus is on the issue of outliers in MLRM data, which can damage the accuracy of model parameter estimates and produce anomalously large residuals. Although traditional methods like the Hampel filter are commonly applied to reduce the influence of outliers, this paper presents an innovative approach: employing wavelet analysis as a strong alternative for outlier handling. Specifically, the research examines the effectiveness of different wavelet families (Daubechies, Coiflets, and Demy). The article shows that wavelet analysis provides better quality through simulations and analysis of actual blood test data. Performance compared to the Hampel filter, resulting in more precise parameter estimation in multivariate models and improved outlier management.

2 Methods and Materials

2.1 Multivariate Linear Regression Models

Multivariate regression is a statistical technique to model the relationship between dependent and multiple independent variables [11]. Unlike simple linear regression, which only considers a single predictor, multivariate regression allows for a more comprehensive analysis encompassing several factors [12]. This methodology is crucial in numerous fields, including economics, medicine, and social sciences, where the outcome is often influenced by several variables simultaneously [13]. Despite its widespread applicability, multivariate regression is fraught with challenges. Multicollinearity, where independent variables are highly correlated, can distort the results. Overfitting, where the model becomes too complex and tailored to the training data, reduces its predictive power on new data. Moreover, outliers significantly affect the estimation of regression coefficients, often resulting in biased and unreliable outcomes [1]. Before diving into the specifics of wavelet shrinkage and the Hampel filter, let's quickly connect on multivariate linear regression models. These models are like the bread and butter of regression analysis, allowing us to understand the relationship between multiple input variables and an outcome of interest [14]. They're the workhorses of predictive analytics, helping us make sense of complex data relationships [15]. The multivariate regression model can be expressed mathematically as follows:

$$Y = XB + E \quad (1)$$

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1q} \\ 1 & x_{21} & x_{22} & \dots & x_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nq} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1p} \\ \vdots & \vdots & & \vdots \\ \beta_{q1} & \beta_{q2} & \dots & \beta_{qp} \end{bmatrix} + \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1p} \\ e_{21} & e_{22} & \cdots & e_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{np} \end{bmatrix} \quad (2)$$

Where Y (n × p) is the matrix of dependent variables, X [n × (q + 1)] is the matrix of independent variables, β[(q + 1) × p] is the matrix of coefficients, and E (n × p) is the matrix of error terms. Thus, each row of (Y) contains the values of the (p) dependent variables measured on a subject. Each column of (Y) consists of the (n) observations on one of the (p) variables. Regression Coefficients (β) Each coefficient represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant [16]. The target variable in a regression model should be continuous, and the data must satisfy key assumptions, including linearity, absence of outliers, homoscedasticity, normality of residuals, and lack of multicollinearity. In this case, multicollinearity is absent among independent variables, meaning the coefficients do not require joint interpretation. The primary objective of least squares estimation (LSE) is to determine the coefficient matrix (β) that minimizes the sum of squared residuals:

$$E = Y - X\beta \quad (3)$$

The Sum of Squared Residuals (S) is:

$$S = tr(E^T E) = [(Y - X\beta)^T (Y - X\beta)] \quad (4)$$

Where (tr) denotes the trace of a matrix.

Estimation of (β): To minimize (S), we take the derivative of (S) concerning (β) and set it to zero:

$\frac{dS}{d\beta} = -2X^T(Y - X\beta) = 0$, Solving for β:

$$X^T Y = X^T X\beta$$

$$\beta = (X^T X)^{-1} X^T Y \quad (5)$$

This equation gives the LSE for the coefficient matrix (β).

The least squares estimator (LSE) retains the property of being a statistically unbiased estimator when the expected value of β equals zero, i.e., E(β) = 0. The expected value of (β) corresponds to the actual underlying coefficient matrix and represents the lowest variance among all unbiased linear estimators. The estimator has a minimum variance, a property known as the

Gauss-Markov theorem, and finally, normality. If the Error terms (E) follow a normal distribution, then the LSE for β is also normally distributed [17].

2.2 Outlier Problem

In statistical analysis, particularly in the context of multivariate regression models, outliers are data points that deviate markedly from the general distribution of the data. Such values can significantly distort analytical results, leading to biased parameter estimates, poor model fit, and diminished predictive accuracy [3]. The influence of outliers on regression analysis is a significant concern in this field. Outliers can significantly influence regression coefficients, leading to models that do not accurately represent the underlying data patterns. This issue is particularly critical in multivariate regression, which evaluates the associations among several variables [5]. Identification and Care. There are several ways to find outliers: statistical tests, robust estimating approaches, and graphical tools like boxplots and scatterplots. Outliers can be addressed or managed effectively by employing data transformation, exclusion, or robust statistical techniques to minimize their impact on analytical results. Determining whether a data point is an outlier and how to manage it can pose a significant challenge [18]. Outliers may represent legitimate but rare events and signal potential deficiencies in the data collection process. While previous studies have extensively addressed outlier detection and mitigation, a deeper analysis is needed to understand the specific mechanisms through which outliers distort regression estimates in real-world applications. This study aims to bridge that gap by systematically evaluating the effects of outliers on multivariate regression models and assessing the efficacy of wavelet-based techniques in improving estimation accuracy [19].

Hampel Filter

The Hampel identifier is a variation of the three-sigma rule of statistics that is robust against outliers. Given a sequence $x_1, x_2, x_3, \dots, x_n$ and sliding window of length k, define point-to-point median and standard-deviation estimates using [5]:

$$Local\ median = mi = median\ (x_{i-k}, x_{i-k+1}, x_{i-k+2}, \dots, x_i, \dots, x_{i+k-2}, x_{i+k-1}, x_{i+k}) \quad (6)$$

$$\text{Standard deviation} = \sigma_i = \text{k} \times \text{ median } (|x_{i-k} - m_i|, \dots, |x_{i+k} - m_i|) \quad (7)$$

where $k = \frac{1}{\sqrt{2}erf^{-1}\left(\frac{1}{2}\right)} \approx 1.4826$ the quantity $\frac{\sigma_i}{k}$ is the median absolute deviation (MAD).

If a sample x_i is such that $|x_i - m_i| > n_\sigma \sigma_i$

For a given threshold n_σ then the Hampel identifier declares x_i an outlier and replaces it with m_i. Near the sequence endpoints, the function truncates the window used to compute m_i and σ_i.

- $i < k + 1$

$$m_i = median(x_1, x_2, x_3, \dots, x_i, \dots, x_{i+k-2}, x_{i+k-1}, x_{i+k}) \quad (8)$$

$$\sigma_i = k \times median(|x_1 - m_1|, \dots, |x_{i+k} - m_i|) \quad (9)$$

- $i > n - k$

$$m_i = median(x_{i-k}, x_{i-k+1}, x_{i-k+2}, \dots, x_i, \dots, x_{n-2}, x_{n-1}, x_n) \quad (10)$$

$$\sigma_i = k \times median(|x_{i-k} - m_i|, \dots, |x_n - m_n|) \quad (11)$$

For expressions of erfinv(1-x), use the complementary inverse error function (erfcinv) instead. This substitution maintains accuracy. When x is close to 1, then 1 - x is a small number and may be rounded down to 0. Instead, replace erfinv(1-x) with erfcinv(x).

The effectiveness of the Hampel filter in regression analysis can be demonstrated through simulation and real-world applications. A small simulation study can highlight its ability to detect and correct outliers, thereby improving model accuracy. Applying the filter to contaminated datasets, such as financial or environmental data, further illustrates its practical value by comparing regression results before and after filtering. Additionally, when benchmarked against traditional methods like z-score and IQR filtering, the Hampel filter shows greater robustness, effectively managing outliers while preserving the underlying data structure [4,20].

2.3 Wavelet Analysis

Wavelets refer to short-lived oscillations characterised by amplitudes that begin at zero, increase or decrease, and then return to zero. Scientists have developed a taxonomy of wavelets based on their quantity and polarity [21].

Daubechies Wavelets (DB) are Normal orthogonal wavelets, named after Ingrid Daubechies, which originated in 1988 and enabled discrete wavelet analysis. They are named after her. The wavelet function's vanishing or ephemeral moments are represented by 4d and (DB), while (N) is the candidate's length and (L_1) is the number of transient moments. The second-ranked person in this family is L_1, corresponding to N, $L_1 = N/2$ is a family of small waves of order n, with an anchor on the period [0, 2n-1]. Each wave has n ephemeral moments, and analytes increase with rank.

$$\left|\frac{d^j}{dX^j}\psi(X)\right| < \infty \quad \Rightarrow \quad \int X^j \psi(X) dX = 0\,, \ 1 \le j \le n \quad (12)$$

The family has (rn) continuous derivatives, with a rank of about 0.2. The small wave is a member of this family [21]. Coiflets Wavelets are a family of wavelets designed by Ingrid Daubechies and Ronald Coifman, known for their balance between smoothness and compact support. They are orthogonal wavelets with vanishing moments in both the scaling and wavelet components, making them well-suited for approximating smooth signals. For Coiflets of order N, both wavelet and scaling functions have N vanishing moments. They are nearly symmetric, minimising phase distortion in signal processing [22]. Due to their compact support, Coiflets are computationally efficient and

widely employed across various fields, including signal and image processing, feature extraction in machine learning, and pattern recognition. They are commonly used for denoising and compressing signals while preserving critical features, as seen in JPEG 2000 image compression applications. In machine learning and pattern recognition, these tools extract meaningful features from complex data [23]. Coiflets wavelets are clearly defined by a scaling function $\phi(t)$ and a wavelet function $\psi(t)$, related through the refinement equation:

$$\phi(t) = \sum_k h_k \phi(2t-k) \quad (13)$$

$$\psi(t) = \sum_k g_k \phi(2t-k) \quad (14)$$

h_k and g_k are the filter coefficients related to the scaling and wavelet functions, respectively [24].

The Meyer wavelet (Demy) is widely used in signal processing and data analysis, particularly in image processing and compression, due to its smoothness and good time and frequency localization. Defined by a smooth function in the frequency domain, the Meyer wavelet exhibits high smoothness and often has non-compactly supported wavelets. It enables multi-resolution analysis, a key feature of wavelet transforms, and finds application in image denoising and feature extraction [25]. Its balanced time-frequency localization makes it valuable in various analytical contexts, including use as a bivariate wavelet filter for stationary datasets [26]. The Meyer wavelet in the frequency domain, denoted as ν(ω) where ω represents the angular frequency, is defined by its Fourier transform $\hat{\psi}(\omega)$ as follows:

$$\hat{\psi}(\omega) = (2\pi)^{\frac{-1}{2}} e^{\frac{i\omega}{2}} sin\left(\frac{\pi}{2} v\left(\frac{3}{2\pi}|\omega|-1\right)\right) \; if \; \frac{2\pi}{3} \leq |\omega| \frac{4\pi}{3} \quad (15)$$

$$\hat{\psi}(\omega) = (2\pi)^{\frac{-1}{2}} e^{\frac{i\omega}{2}} cos\left(\frac{\pi}{2} v\left(\frac{3}{4\pi}|\omega|-1\right)\right) \; if \; \frac{4\pi}{3} \leq \frac{8\pi}{3} \quad (16)$$

$$\hat{\psi}(\omega) = 0 \; if \; \notin \left[\frac{2\pi}{3}, \frac{8\pi}{3}\right] \quad (17)$$

where ν(ω) is a smooth function defined as: $\nu(\omega) = 0 \; for \; \omega \leq 0$

$$\nu(\omega) = exp\left(\frac{-1}{\omega^2} exp\left(\frac{-1}{\omega^2-1}\right)\right) \; for \; 0 < \omega < 1 \; and \; \nu(\omega) = 1 \; for \; \omega \geq 1 \quad (18)$$

This function ν(ω) ensures that the wavelet is smooth and has the desired properties in the Fourier domain. Meyer Wavelet in the Time Domain The inverse Fourier transform of $\hat{\psi}(\omega)$ gives the Meyer wavelet $\psi(t)$ in the time domain. However, this time-domain expression is often complex and typically not presented in a closed form due to the intricacies involved in the Fourier transform of the defined function [27]. These wavelets aim to minimize noise and the impact of outliers in multivariate models, as shown by the proposed methodology.

2.4 Proposed Methods

The proposed method relies on wavelet analysis to estimate outliers in MLRM as follows:

Estimation of the regression coefficients matrix of an MLRM using the (OLS) method:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (19)$$

Calculate the residual matrix for the estimated model:

$$E = Y - \hat{Y} \quad (20)$$

Each element E_{ij} represents the residual for the i-*th* observation and the j-*th* dependent variable.

$$E_{ij}^{Standard} = \frac{E_{ij}}{\sigma_j} \quad (21)$$

$E^{Standard}$ represents the standard residual matrix and σ_j is the standard deviation of residuals for j-*th* dependent variable. A standard residual outside the range [−3, 3] suggests an outlier for that dependent variable Y(Outlier). Estimating outliers is done through wavelet analysis as follows:

Suppose Y is a data matrix representing the p variables (y_1, y_2, …, y_p) and n observations. Perform a discrete wavelet transform for each variable (j = 1, 2, …, p) and produce the approximate and detailed coefficients (two partitions) after symmetrically extending the variable vector y ($y_{\{extended\}}$) as in the following formulas:

$$Aj[n] = \sum_{k=0}^{L-1} y_{\{extended\}}[n+k] \times h[k] \quad (22)$$

$$Dj[n] = \sum_{k=0}^{L-1} y_{\{extended\}}[n+k] \times g[k] \quad (23)$$

h (the low-pass filter) and g (the high-pass filter) are the wavelet coefficient values used in analysis with Daubechies, Coiflets, and Dmey wavelets. Their number (length) depends on the optimal order, which is the order that effectively addresses outliers while yielding the lowest mean squared error in the multivariate linear regression model (MLRM). In this study, the fifth-order wavelets (Daubechies5 and Coiflets5) were optimal for achieving the lowest MSE while effectively addressing data contamination of the wavelet used for maximum L.

Where n varies over the appropriate indices of the data, filter length (L), Formula 4 represents the approximation coefficients (first part), scale or father function Aj[n] proportional to the average observations. In contrast, formula 5 represents (second part) the detail coefficients Dj[n], the mother or wavelet function with the same number of approximate coefficients, which are proportional to the differences (variance) of the observations. Calculating the discrete wavelet transform is repeated for each variable in the dependent variable matrix. The matrices of

multivariate approximation coefficients (YA) and detail coefficients (YD) are acquired as follows:

$$YA = [A1[n]\ \ A2[n] \ldots\ \ Ap[n]] \tag{24}$$

$$YD = [D1[n]\ \ D2[n] \ldots\ \ Dp[n]] \tag{25}$$

Estimation of the regression coefficient matrices of an MLRM using the (OLS) method for two matrices YA and YD:

$$\widehat{\beta A} = (X^T X)^{-1} X^T YA \tag{26}$$

$$\widehat{\beta D} = (X^T X)^{-1} X^T YD \tag{27}$$

The estimated regression coefficient matrices are used to compute the approximation $\widehat{YA}$ and detail $\widehat{YD}$ coefficients:

$$\widehat{YA} = X\widehat{\beta A} \tag{28}$$

$$\widehat{YD} = X\widehat{\beta D} \tag{29}$$

Compute the inverse discrete wavelet transform (IDWT) from the approximation and detail estimated coefficients $\widehat{A}j$ and $\widehat{D}j$ for j = 1, 2, …, p. If the DWT decomposes real data into low-pass filter (approximation) and high-pass filter (detail) components, the IDWT uses these components to reconstruct the original data as accurately as possible. The IDWT $\widehat{A}j[n]$ and $\widehat{D}j[n]$ are:

$$\widehat{A}j[n] = \sum_{k=0}^{L-1} \widehat{YA}j\left[\frac{n-k}{2}\right] \times h[k] \tag{30}$$

$$\widehat{D}j[n] = \sum_{k=0}^{L-1} \widehat{YD}j\left[\frac{n-k}{2}\right] \times g[k] \tag{31}$$

The reconstructed data vector Y_j is obtained by summing up the two filtered datasets:

$$Y_j = \widehat{A}j[n] + \widehat{D}j[n] \tag{32}$$

For j = 1, 2, …, p, filtered data matrix YW is

$$YW = [Y_1\ Y_2 \ldots\ Y_p] \tag{33}$$

The previously identified outliers are replaced by their corresponding formula 15 values.

$$\widehat{Y}(\text{Outliers}) = YW(\text{Outliers}) \tag{34}$$

The matrix of dependent variables (YM), whose outliers have been estimated using wavelet analysis, is relied upon to estimate the regression coefficients of MLRM using OLS.

$$\widehat{\beta M} = (X^T X)^{-1} X^T YM \tag{35}$$

3 Results and Discussions

Depending on simulation studies and real data, the proposed method's efficiency in handling outliers is compared with the traditional Hampel filter method in analyzing multivariate linear regression models.

3.1 Simulation Study

Data were generated (using MATLAB program) following a multivariate standard normal distribution representing (1, 3, and 5) independent variables with a sample size of (100 and 300) and assuming the values of the regression coefficients with intercepts for two and three columns shown in Table 1. The model noise was generated randomly following a multivariate standard normal distribution for two and three columns and was added to the MLRM to obtain data for the dependent variables ($Y = XB + E$). E is the random error matrix that has a multivariate standard normal distribution. Also, outliers have generated data using the (randperm (2, 2)) function for the MATLAB program and added to dependent variables (p = 2) and the (randperm (2, 3)) function and added to dependent variables (p = 3) that is, there are two outliers for each dependent variable.

Table 1: Assumed Regression Coefficients for simulation.

Coefficients of Regression with intercepts		
5	-3	8
1.5	-1	-2
-2	1.3	-1.5
0.8	0.6	-0.6
0.5	-0.7	0.7
-1.2	0.9	0.8

Based on the assumed values of the multivariate model regression coefficients in Table 1, the simulation experiments were carried out (1000) times, with the proposed methods applied to the data in conjunction with the traditional Hampel filter and the unfiltered data, allowing for a comprehensive comparison of the results. The Hampel filter method used at a window size and threshold parameter (3) for the case of two dependent variables and (5) for the case of three dependent variables tried to handle outliers and provided better results than the results of unfiltered data, but they were not optimal. The proposed method provided better handling of outliers and lower absolute estimation error than the Hampel filter method. The Demy wavelet gave the best results compared to the rest of the wavelets. To know the efficiency of the estimated models and compare the traditional and proposed methods more clearly, some statistical criteria were calculated (MSE(β), MSE, R^2Y_1, R^2Y_2), and the results are in Tables 2-4:

Table 2: Average of Statistical Criteria for Simulation Data (p = 2, q = 1).

Method	n	MSE(β)	MSE	R^2Y_1	R^2Y_2
Without Filter	100	0.0801	2.9307	43.32%	25.39%
Hampel Filter		0.0942	1.2690	52.40%	38.95%
DB5		0.0016	1.0010	67.67%	49.01%
Coiflets5		0.0016	0.9999	67.70%	49.02%
Dmey		0.0015	0.9966	67.72%	49.14%
Without Filter	300	0.0089	1.6540	57.48%	37.80%
Hampel Filter		0.0043	1.2177	61.71%	45.82%
DB5		0.0001	0.9957	68.79%	49.98%
Coiflets5		0.0001	0.9956	68.79%	49.98%
Dmey		0.0001	0.9956	68.79%	49.98%

Table 3: Average of Statistical Criteria for Simulation Data (p = 2, q = 3).

Method	n	SSE(β)	MSE	R^2Y_1	R^2Y_2
Without Filter	100	0.0799	2.8669	70.30%	51.54%
Hampel Filter		0.2962	2.1522	64.23%	56.46%
DB5		0.0044	1.0524	85.65%	73.86%
Coiflets5		0.0042	1.0459	85.76%	73.92%
Dmey		0.0042	1.0409	85.77%	74.05%
Without Filter	300	0.0093	1.6397	80.66%	64.85%
Hampel Filter		0.0119	1.5118	79.85%	67.40%
DB5		0.0005	1.0102	86.82%	74.98%
Coiflets5		0.0005	1.0106	86.81%	74.98%
Dmey		0.0005	1.0105	86.81%	74.99%

Table 4: Average of Statistical Criteria for Simulation Data (p = 2, q = 5).

Method	n	SSE(β)	MSE	R^2Y_1	R^2Y_2
Without Filter	100	0.0794	2.8038	74.95%	60.64%
Hampel Filter		0.3727	2.4367	66.94%	61.37%
DB5		0.0054	1.0586	88.02%	80.00%
Coiflets5		0.0054	1.0552	88.01%	80.04%
Dmey		0.0051	1.0405	88.12%	80.31%
Without Filter	300	0.0088	1.6299	83.92%	72.74%
Hampel Filter		0.0158	1.6043	82.38%	73.43%
DB5		0.0006	1.0141	89.10%	81.08%
Coiflets5		0.0006	1.0142	89.10%	81.07%
Dmey		0.0006	1.0141	89.10%	81.07%

The results for three variables were summarized in Tables 5-7:

Table 5: Average of Statistical Criteria for Simulation Data (p = 3, q = 1).

Method	n	SSE(β)	MSE	R^2Y_1	R^2Y_2	R^2Y_3
Without Filter	100	0.1710	3.7405	43.22%	25.11%	42.27%
Hampel Filter		0.2097	1.4279	51.78%	39.17%	62.24%
DB5		0.0037	1.0168	67.66%	48.78%	78.24%
Coiflets5		0.0036	1.0144	67.75%	48.79%	78.28%
Dmey		0.0036	1.0126	67.73%	48.94%	78.27%
Without Filter	300	0.0188	1.9282	57.54%	37.60%	61.47%
Hampel Filter		0.0109	1.4003	61.86%	45.54%	68.41%
DB5		0.0003	1.0018	68.88%	49.72%	79.39%
Coiflets5		0.0003	1.0019	68.88%	49.71%	79.39%
Dmey		0.0003	1.0020	68.88%	49.72%	79.39%

Table 6: Average of Statistical Criteria for Simulation Data (p = 3, q = 3).

Method	n	SSE(β)	MSE	R^2Y_1	R^2Y_2	R^2Y_3
Without Filter	100	0.1698	3.6657	70.37%	51.39%	55.39%
Hampel Filter		0.4899	2.2555	64.77%	56.05%	65.51%
DB5		0.0074	1.0597	85.78%	73.85%	85.09%
Coiflets5		0.0073	1.0553	85.77%	73.92%	85.19%
Dmey		0.0073	1.0496	85.85%	74.02%	85.19%
Without Filter	300	0.0188	1.9119	80.63%	64.95%	72.71%
Hampel Filter		0.0208	1.7052	79.78%	67.26%	75.14%
DB5		0.0006	1.0166	86.76%	74.94%	86.35%
Coiflets5		0.0006	1.0166	86.77%	74.93%	86.34%
Dmey		0.0007	1.0168	86.76%	74.94%	86.34%

Table 7: Average of Statistical Criteria for Simulation Data (p = 3, q = 5).

Method	n	MSE(β)	MSE	R^2Y_1	R^2Y_2	R^2Y_3
Without Filter	100	0.1733	3.5785	74.79%	60.72%	60.36%
Hampel Filter		0.6143	2.5382	66.44%	61.11%	67.37%
DB5		0.0106	1.0755	87.93%	79.96%	86.85%
Coiflets5		0.0099	1.0628	88.00%	80.06%	87.09%
Dmey		0.0096	1.0539	88.06%	80.24%	87.13%
Without Filter	300	0.0187	1.9025	83.84%	72.61%	75.83%
Hampel Filter		0.0276	1.7879	82.46%	73.05%	77.14%
DB5		0.0010	1.0171	89.11%	80.89%	88.05%
Coiflets5		0.0010	1.0167	89.12%	80.88%	88.06%
Dmey		0.0010	1.0170	89.11%	80.89%	88.05%

3.1.1 Discussion of Simulation Results

The average statistical criteria for the estimated models in Tables 2-7 showed the following:

1. Without filtering methods, provide the highest MSE and lower determination coefficients, indicating relative performance.

2. The Hampel filter method did not provide a lower average (SSE(β)) compared to the OLS method without data filtering in most cases. However, it resulted in a reduced MSE for the estimated models and a more significant coefficient of determination for both models.

3. Based on the SSE(β) criterion average, the proposed method (Wavelets Analysis) gave the best results, while Demy was the best compared to the rest of the wavelets.

4. Based on the MSE and determination coefficients criteria average, the proposed method gave the best results, while Demy was the best compared to the rest of the wavelets.

5. Increasing the number of independent variables led to a decrease in the SSE(β) and MSE average and an increase in the coefficient of determination.

6. Increasing the number of dependent variables increased the SSE(β) and MSE averages and decreased the coefficient of determination.

7. Increasing the sample size decreased the SSE(β) and MSE average and increased the coefficient of determination.

8. The wavelet analysis gave convergent results for all wavelets and was used in estimating outliers (DB5, Dmey, and Dmey).

The results from the simulation data reveal that wavelet-based filters, particularly the Dmey wavelet, consistently outperform other methods in terms of minimizing the Mean Squared Error (MSE) and maximizing the R^2 values for both dependent variables. The Dmey wavelet showed the lowest MSE and the highest R^2 across all cases, highlighting its effectiveness in remediating contamination in multivariate linear models.

3.2 Real Data

The proposed and conventional methods were applied to datasets for various ordinary patients in Iraq to estimate MLR models. The dataset provides the patients' Cell Blood Count test information to create a Hematology diagnosis/prediction system. The dataset originated from Al-Zahraa Al-Ahly Hospital in 2022. The dependent variables represent Hemoglobin (HGB), with the following normal ranges: 11.0 to 16.0, Unit: g/Dl, Red Blood Cell (RBC), Normal Ranges: 3.50 to 5.50, Unit: 10^12/L, and White Blood Cell (WBC), Normal Ranges: 4.0 to 10.0, Unit: 10^9/L. While the Predictor variables represent 17 tests, in Table 8:

Table 8: Cell Blood Count test.

No.	Symbol	Predictor Variable	Normal Ranges
1	LYMp	Lymphocyte percentage	20.0 to 40.0, Unit: %
2	MIDp	Indicates the percentage combined value of the other types of white blood cells not classified as lymphocytes or granulocytes	1.0 to 15.0, Unit: %
3	NEUTp	Neutrophils (leukocytes)	50.0 to 70.0, Unit: %
4	LYMn	Lymphocyte numbers are a type of white blood cell	0.6 to 4.1, Unit: 10^9/L
5	MIDn	Indicates the combined number of other white blood cells not classified as lymphocytes or granulocytes	0.1 to 1.8, Unit: 10^9/L
6	NEUTn	Neutrophils (leukocytes)	2.0 to 7.8, Unit: 10^9/L
7	HCT	Hematocrit	36.0 to 48.0, Unit: %
8	MCV	Mean Corpuscular Volume	80.0 to 99.0, Unit: FL
9	MCH	Mean Corpuscular Hemoglobin	26.0 to 32.0, Unit: pg.
10	MCHC	Mean Corpuscular Hemoglobin Concentration	32.0 to 36.0, Unit: g/dL
11	RDWSD	Red Blood Cell Distribution Width	37.0 to 54.0, Unit: fL
12	RDWCV	Red blood cell distribution width	11.5 to 14.5, Unit: %
13	PLT	Platelet Count	100 to 400, Unit: 10^9/L
14	MPV	Mean Platelet Volume	7.4 to 10.4, Unit: fL
15	PDW	Red Cell Distribution Width	10.0 to 17.0, Unit: %
16	PCT	The Level of Procalcitonin in the Blood	0.10 to 0.28, Unit: %
17	PLCR	Platelet Large Cell Ratio, Normal Ranges	13.0 to 43.0, Unit: %

Table 9 presents the statistical summary of a random sample of 200 observations drawn from these examinations. The mean level of the dependent variable HGB was (11.175), and it is within the normal range (11-16) with a standard deviation of (2.5288); RBC was (4.5178), and it is within the normal range (3.5-5.5) with a standard deviation (0.6672), and WBC was (7.2835), and it is within the normal range (4-10) with a standard deviation (3.4133). All means of the Predictor variables for blood tests were within the normal range. All dependent and independent variables included observations that fell outside their normal ranges.

Table 9: Descriptive Statistics.

Variable	Minimum	Maximum	Mean	Std. Deviation	Variance
HGB	0.40	16.70	11.1750	2.5288	6.3950
RBC	2.55	5.99	4.5178	0.6672	0.4450
WBC	0.80	24.40	7.2835	3.4133	11.6510
LYMp	6.90	59.20	27.3460	10.9563	120.0400
MIDp	.60	77.00	9.1230	8.4300	71.0650
NEUTp	34.80	85.30	64.6980	10.3914	107.9810
LYMn	.60	6.30	1.8434	.86372	.746
MIDn	.15	7.00	.6352	.58835	.346
MEUTn	1.30	44.00	5.0550	3.90728	15.267
HCT	19.50	316.00	38.1775	20.65926	426.805
MCV	59.10	106.20	81.5935	7.54727	56.961
MCH	15.60	34.20	24.9600	3.24804	10.550
MCHC	20.20	79.60	31.0450	4.98088	24.809
RDWSD	23.30	57.60	37.2320	3.80409	14.471
RDWCV	11.10	124.00	13.6150	7.95190	63.233
PLT	17.00	508.00	173.7550	56.53909	3196.668
MPV	8.10	19.50	9.7590	.99951	.999
PDW	9.50	20.00	13.6165	1.99335	3.973
PCT	.01	.45	.1625	.05263	.003
PLCR	12.30	46.50	26.6594	6.90892	47.733

Outliers in the dependent variables data were diagnosed by plotting the regression standardized residual values against the regression standardized predicted values as in Figures 1-3:

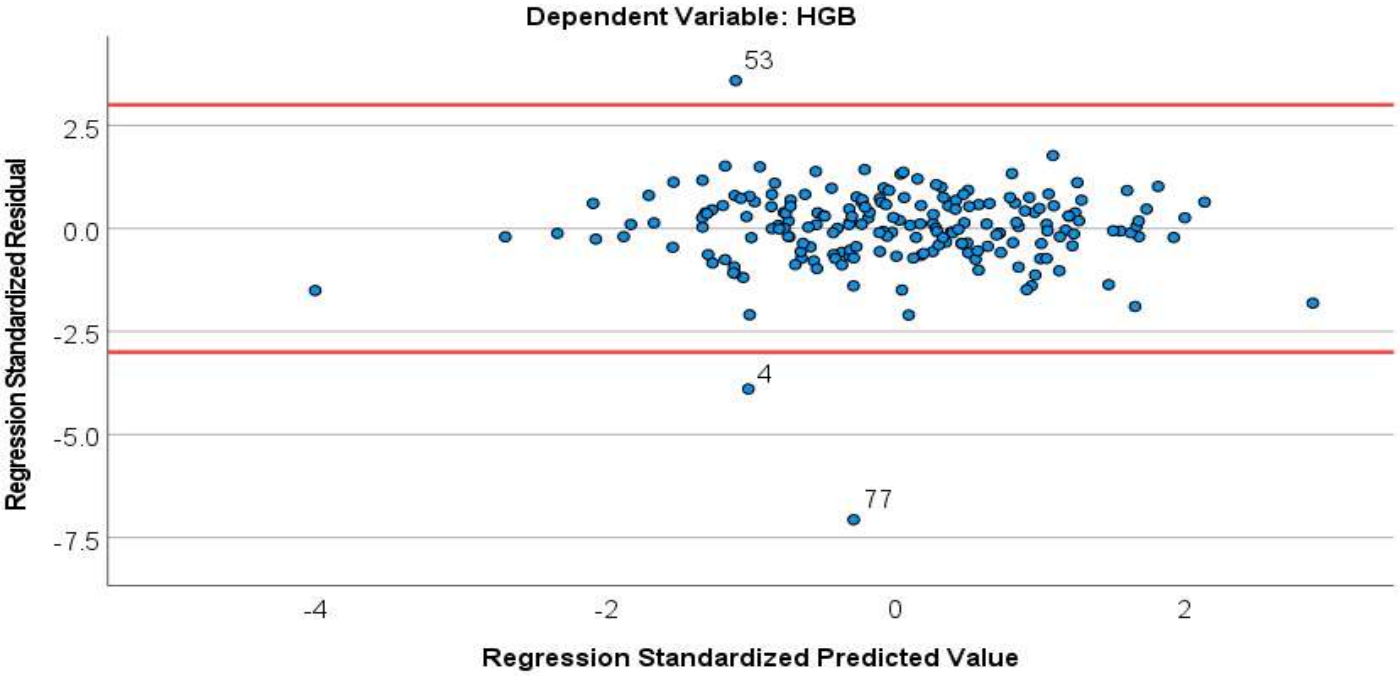

Figure 1: Regression standardized residual for Real Data (HGB).

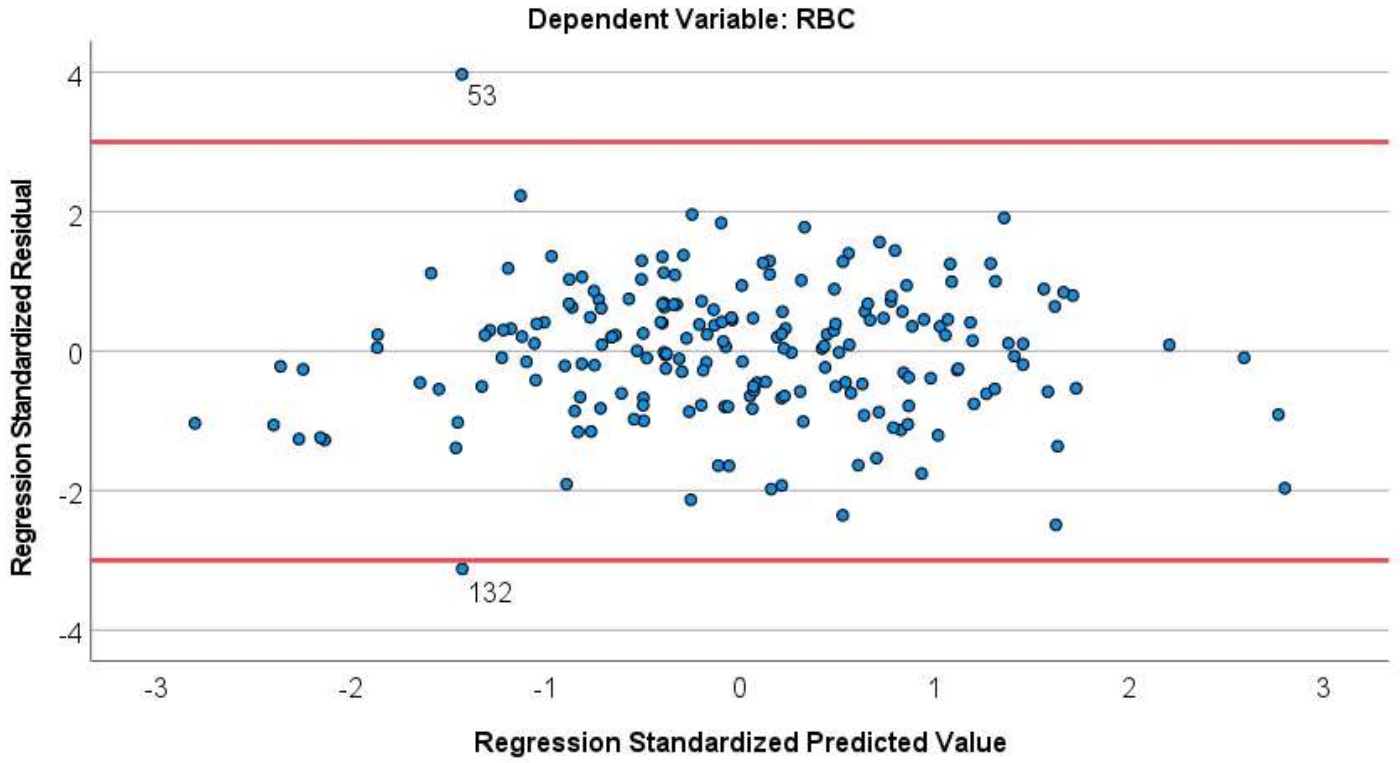

Figure 2: Regression Standardized Residual for Real Data (RBC).

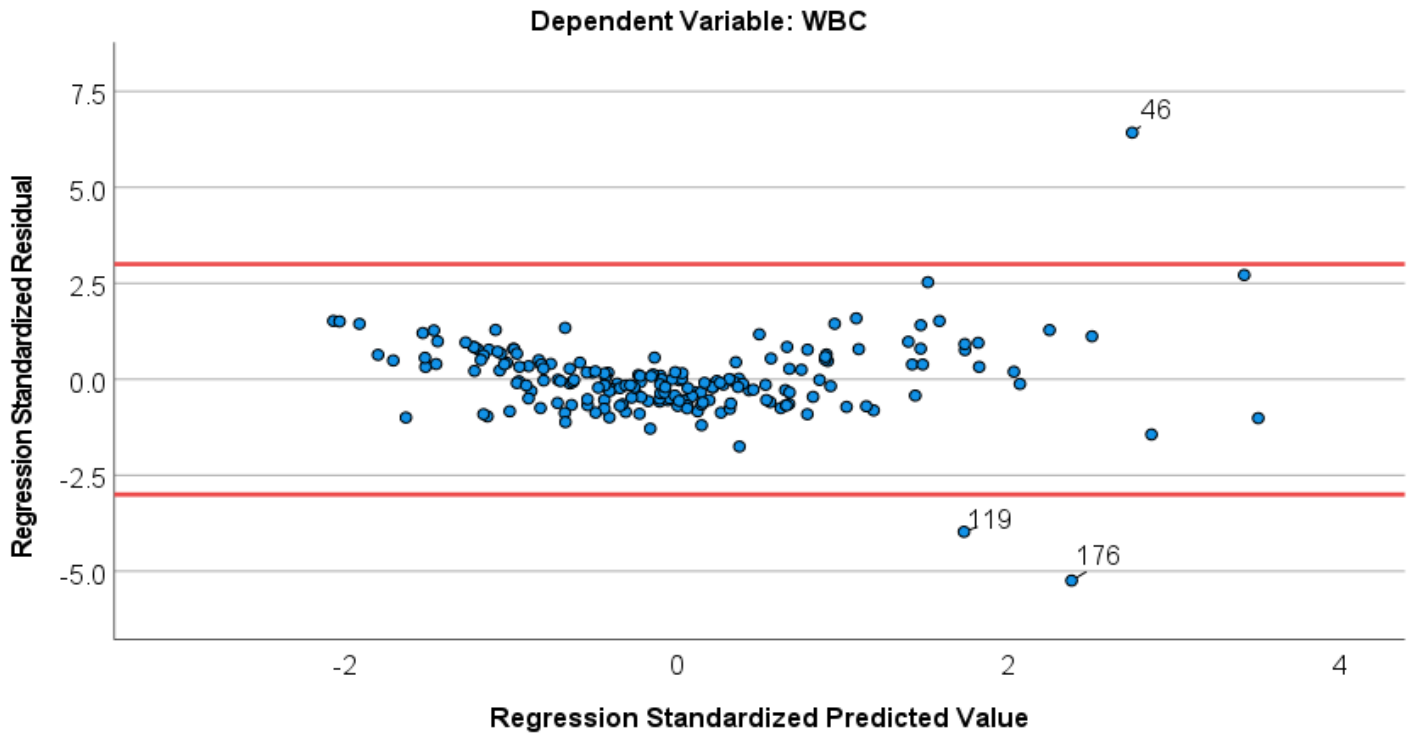

Figure 3: Regression Standardized Residual for Real Data (WBC).

Table 10 shows three outliers in the dependent variable HGB, two in the dependent variable RBC, and three in the dependent variable WBC. Outliers were estimated using the Hampel filter and wavelet analysis of three wavelets. Despite the outliers, each model's MSE values with the coefficient of determination are considered reasonable and acceptable. The overall F test (Table 11) shows the significance of the estimated linear models, i.e., the presence of at least one significant independent variable in the estimated models.

Table 10: Estimated values of outliers for Real Data.

Observations	Outliers	Hampel	DB5	Coiflets	Demy
Y_1(4)	3.80	8.8000	6.6582	9.9102	10.0996
Y_1(53)	13.5	13.500	11.0719	10.8184	11.6527
Y_1(77)	1.20	12.300	9.9453	10.8501	10.9640
Y_2(53)	5.17	5.1700	4.6324	4.4826	4.4198
Y_2(132)	2.56	4.6500	4.0415	4.1860	4.0345
Y_3(46)	24.4	4.6000	9.2476	7.9245	10.6864
Y_3(119)	7.60	7.6000	6.5226	5.6049	7.7347
Y_3(176)	8.00	4.9000	7.0207	7.7240	6.0634

To demonstrate the performance of the proposed methods in managing identified outliers and mitigating noise in real-world data and to facilitate comparison with the conventional Hampel filter, the parameters of the multivariate linear regression model (MLRM) were estimated using several approaches. Subsequently, the mean squared error (MSE) and coefficients of determination (R^2) were computed for each model, as detailed in Tables 11 to 15. The mean squared error (MSE) and the coefficient of determination (R^2) for each model were subsequently calculated and are presented in Tables 11 through 15.

Table 11: Results of MLRM Using Unfiltered Real Data.

Parameters			MSE	Variable	R^2 and Adj.R^2	F	p-value
53.4010	21.6887	5.6828	1.0888	Y_1	75.00% 72.66%	32.1193	0.000
-0.0927	-0.0148	-0.2765					
-0.0020	1.3504	0.0052					
-0.1039	-0.0153	-0.0673					
-0.2339	-0.0747	2.8368					
0.0518	0.0174	0.7860					
-0.0033	-0.0005	0.1832		Y_2	72.10% 69.49%	27.7186	0.000
0.2056	0.0812	-0.0291					
-0.1170	-0.0641	-0.0721					
0.7486	0.1351	0.1848					
-0.0005	-0.0007	0.0071					
-0.0859	-0.0197	-0.0169					
0.0163	0.0058	0.0072		Y_3	86.60% 85.35%	69.2402	0.000
-0.0633	-0.0248	0.0500					
-6.5500	-2.2464	0.9004					
0.2132	0.0695	-0.0376					
69.8338	26.8449	-46.1737					
0.5299	0.1767	0.0015					

Table 12: Results of MLRM Using Hampel Filter Method for Real Data.

Parameters			MSE	Variable	R^2 and Adj.R^2	F	p-value
17.589	9.6507	-0.4036	1.0023	Y_1	33.36% 27.14%	5.3592	0.000
0.0210	0.0265	-0.0808					
0.0010	-0.0024	0.0094					
0.0321	0.0253	-0.0275					
-0.1405	-0.0476	0.8137					
0.1484	0.0297	0.1013					
0.0184	-0.0004	-0.0355		Y_2	32.40% 26.09%	5.1313	0.000
0.0630	0.0332	-0.0218					
-0.1631	-0.0308	0.0351					
0.5486	0.0612	-0.0542					
-0.0079	-0.0058	0.0020					
-0.0298	-0.0023	-0.0201					
0.0009	-0.0046	0.0117		Y_3	25.79% 18.86%	3.7208	0.000
0.0075	-0.0179	0.0459					
-1.8659	-0.9665	0.8773					
0.1542	0.0392	0.0841					
-2.8674	19.3023	-46.3388					
0.1730	0.0579	-0.0400					

Table 13: Results of MLRM Using DB5 Wavelet Filter Method for Real Data.

Parameters			MSE	Variable	R^2 and Adj.R^2	F	p-value
50.7862	21.3764	10.8442	0.7934	Y_1	83.19% 81.62%	52.9645	0.000
-0.0691	-0.0156	-0.3081					
-0.0030	0.0001	0.0023					
-0.0749	-0.0157	-0.0773					

-0.2165	-0.0679	3.0521					
0.0823	0.0276	0.2382					
0.0032	0.0008	0.0895					
0.2052	0.0794	-0.0108		Y_2	73.90% 71.46%	30.3187	0.000
-0.1262	-0.0646	-0.0390					
0.7570	0.1393	0.0776					
0.0003	-0.0008	0.0087					
-0.0820	-0.0211	-0.0121					
0.0164	0.0061	0.0048					
-0.0584	-0.0236	0.0373					
-6.5385	-2.2028	0.3474		Y_3	87.02% 85.81%	71.8032	0.000
0.2326	0.0633	-0.0194					
64.0870	25.8820	-30.4422					
0.5354	0.1772	0.0240					

Table 14: Results of MLRM Using Coiflets5 Wavelet Filter Method for Real Data.

Parameters			**MSE**	**Variable**	**R^2 and Adj.R^2**	**F**	**p-value**
50.1584	21.3350	12.3098	0.7990	Y_1	83.56% 82.02%	54.4151	0.000
-0.0591	-0.0154	-0.3207					
-0.0044	0.0005	0.0021					
-0.0642	-0.0155	-0.0898					
-0.2162	-0.0672	3.0676					
0.0834	0.0292	0.0597					
0.0022	0.0010	0.1005		Y_2	74.06% 71.64%	30.5617	0.000
0.2045	0.0796	-0.0092					
-0.1348	-0.0638	-0.0359					
0.7591	0.1378	0.0674					
-0.0003	-0.0008	0.0088					
-0.0649	-0.0213	-0.0114					
0.0163	0.0061	0.0046		Y_3	86.47% 85.21%	68.3950	0.000
-0.0613	-0.0234	0.0360					
-6.5299	-2.2080	0.3004					
0.2351	0.0625	-0.0180					
66.7080	25.6427	-28.8897					
0.5251	0.1785	0.0256					

Table 15: Results of MLRM Using Dmey Wavelet Filter Method for Real Data.

Parameters			**MSE**	**Variable**	**R^2 and Adj.R^2**	**F**	**p-value**
50.2029	21.3538	8.9872	0.7728	Y_1	83.19% 81.62%	52.9679	0.000
-0.0612	-0.0149	-0.2918					
-0.0043	0.00001	0.0026					
-0.0663	-0.0150	-0.0607					
-0.2167	-0.0678	3.0355					
0.0778	0.0290	0.4659					
0.0018	0.0009	0.0727		Y_2	74.33% 71.93%	31.0009	0.000
0.2016	0.0802	-0.0124					
-0.1434	-0.0626	-0.0424					
0.7763	0.1350	0.0888					
-0.0010	-0.0007	0.0086					
-0.0636	-0.0212	-0.0130					
0.0165	0.0060	0.0051		Y_3	87.50%	74.9119	0.000

-0.0627	-0.0233	0.0388			86.33%		
-6.4452	-2.2239	0.3980					
0.2371	0.0629	-0.0208					
68.0914	25.5653	-32.1642					
0.5131	0.1800	0.0222					

3.2.1 Discussion of Real Data Results

The statistical criteria for the estimated models in Tables 11-15 showed the following:

1. The unfiltered method provides the highest MSE, and lower determination coefficients compared to the proposed methods.

2. The Hampel filter method solved the problem of outliers and provided a lower MSE than the unfiltered data method but provided very weak and unacceptable coefficients of determination.

3. Based on the MSE and determination coefficients criteria, the proposed method (Wavelets Analysis) gave the best results while Demy was the best compared to the rest of the wavelets since MSE is equal to (0.7728).

4. The proposed Dmey wavelet was the best compared with the rest of the wavelets, and determination coefficient values reached 83.19% for Y_1, 74.33% for Y_2, and 87.50% for Y_3, which was better than all the other ways.

5. R^2 = 83.19%, Adj. R^2 = 81.62% for Y_1 indicates that the model explains 83.19% of the variance in Y_1 and the small difference between R^2 and Adj. R^2 (0.40%) suggests that most independent variables are relevant, with minimal overfitting. R^2 = 84.97%, Adj. R^2 = 84.26% for Y_2, explains a slightly lower proportion of variance compared to Model Y_1 and the gap of 0.71% between R^2 and Adj. R^2 suggests a minor influence of irrelevant. R^2 = 94.08%, Adj. R^2 = 93.80 for Y_3, the best-performing model, explaining 94.08% of the variance in Y_3 and the small difference (0.28%) between R^2 and Adj. R^2 shows an efficient model with little overfitting. The minimal differences between R^2 and Adj. R^2 in all models suggest good independent variable selection with minimal overfitting, and high R^2 values (above 90%) indicate strong predictive performance for Models Y_1 and Y_3, while Model Y_2 still performs well with 85% variance explained. The adjusted R^2 values confirm that the models are not excessively complex and are likely generalizable.

6. Based on the results of the Dmey analysis method's preference, it can be relied upon to handle outliers and data noise and the overall F-statistic (52.9679, 31.0009, and 74.9119 for three models, respectively).

After selecting the Dmey method, which yielded the best results compared to the other methods, the significance of the estimated parameters of the (MLRM), was tested using the t-test and P-value calculation, as presented in Table 16:

Table 16: Testing the parameters of an MLRM using a Dmey for real data.

Variable	t-statistic			p-value		
	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
Constant	-7.3561	-2.6185	8.1267	0.000	0.009	0.000
LYMp	11.2167	7.52909	-12.9290	0.000	0.000	0.000
MIDp	-0.7940	-0.9616	2.4730	0.428	0.338	0.010
NEUTp	12.1362	8.6245	-6.8696	0.000	0.000	0.000
LYMn	3.9474	3.3934	28.2771	0.000	0.001	0.000
MIDn	-0.0776	-0.4058	1.5873	0.938	0.685	0.114
MEUTn	0.5185	0.7477	2.4215	0.605	0.456	0.016
HCT	2.2370	-0.3615	1.2722	0.026	0.718	0.205
MCV	-3.0363	-3.6890	-7.8045	0.003	0.000	0.000
MCH	13.8362	5.9653	5.6059	0.000	0.000	0.000
MCHC	-1.5724	0.5527	-1.2419	0.118	0.581	0.216
RDWSD	-17.6602	-17.8475	-1.8896	0.000	0.000	0.060
RDWCV	0.9776	0.6950	-1.8030	0.330	0.488	0.073
PLT	-2.2888	-2.3276	-2.3292	0.023	0.021	0.022
MPV	-2.2221	0.9802	-1.5931	0.028	0.328	0.113
PDW	-13.6741	-15.9027	5.0260	0.000	0.000	0.000
PCT	2.9538	3.0851	4.6514	0.004	0.002	0.000
PLCR	18.7570	16.3709	-6.0096	0.000	0.000	0.000

When using a significance level of $\alpha = 0.01$, a variable is accepted to have a significant effect only if its p-value is less than 0.01. Variables with high significance ($p < 0.01$): LYMp, NEUTp, LYMn, MCV, MCH, PDW, PCT, and PLCR are significant in all models ($p < 0.01$). RDWSD is significant in Y_1 and Y_2 ($p = 0.000$), but non-significant in Y_3 ($p = 0.060$). PLT is significant in all models ($p \approx 0.02$) at the 0.05 level but not significant at 0.01 non-significant variables ($p \geq 0.01$), HCT is significant in Y_1 at the 0.05 level ($p = 0.026$) but not significant at 0.01, and MPV is significant in Y_1 at the 0.05 level ($p = 0.028$), but not significant at 0.01. MIDp and MIDn are non-significant in all models ($p > 0.01$). MEUTn is not significant in Y_1 and Y_2 ($p > 0.01$) and significant in Y_3 only ($p = 0.016$) at the 0.05 level, but not significant at 0.01. MCHC is not significant in all models ($p > 0.01$). RDWCV is not significant in all models ($p > 0.01$). Focus on independent variables that are significant at 0.01 only when building or improving models. Ignore variables that are not consistently significant in all models (MIDp, MIDn, MEUTn, HCT, MCHC, RDWCV, PLT, and MPV); the indicator is red unless they show explanatory significance in a specific context.

Table 17: Results of MLRM Using Dmey Filter Method for Real Data After Removing the Insignificant Independent Variables.

Parameters			MSE	Variable	R^2 and Adj.R^2	F	p-value
-57.9204	-6.0264	42.2095	0.0485	Y_1	91.46% 91.06%	225.9476	0.000
0.7339	0.1759	-0.6230					
0.8814	0.2260	-0.3417					
0.5657	0.2273	3.5164		Y_2	84.97% 84.26%	119.3462	0.000
-0.0543	-0.0282	-0.1097					
0.7025	0.1405	0.2018					
-0.4680	-0.1854	-0.0629		Y_3	94.08% 93.80%	335.2348	0.000
-1.8116	-0.8022	0.4847					
5.2496	2.0924	13.8446					
0.7190	0.2619	-0.1852					

Table 17 provides insights into the significance and practical implications of the coefficients. Independent variables with the largest absolute coefficients as they have the most substantial influence on the dependent variables. Using the hybrid Dmey filter method, briefly explain its role in preprocessing or feature selection. The low value of the MSE of (0.0485) suggests that the models provide predictions for dependent variables close to the observed values, indicating a good fit. $R^2 = 91.46\%$, Adj. $R^2 = 91.06\%$ for Y_1 indicates that the model explains 91.46% of the variance in Y_1 and the small difference between R^2 and Adj. R^2 (0.40%) suggests that most independent variables are relevant, with minimal overfitting. $R^2 = 84.97\%$, Adj. $R^2 = 84.26\%$ for Y_2, which explains a slightly lower proportion of variance compared to Model Y_1 and the gap of 0.71% between R^2 and Adj. R^2 suggests a minor influence of irrelevant. $R^2 = 94.08\%$, Adj. $R^2 = 93.80$ for Y_3, the best-performing model, explaining 94.08% of the variance in Y_3 and the small difference (0.28%) between R^2 and Adj. R^2 shows an efficient model with little overfitting. The small differences between R^2 and Adj. R^2 in all models suggests good independent variable selection with minimal overfitting, and high R^2 values (above 90%) indicate strong predictive performance for Models Y_1 and Y_3, while Model Y_2 still performs well with 85% variance explained. The adjusted values confirm that the models are not excessively complex and are likely generalizable.

The F-Statistic and p-value are crucial for evaluating the overall significance of a regression model. F-statistic (225.9476) for Y_1 indicates a strong relationship between the independent variables and Y_1 (as shown in Figure 4), and the high value reflects the model's ability to explain variance effectively. P-value (0.000) confirms the overall significance of the model. The F-statistic (119.3462) for Y_2 is lower than Model Y_1, suggesting a less robust fit compared to the first model, but still, a firm value indicating the model explains significant variance (as shown in Figure 5), and p-value (0.000) confirming that the model is statistically significant. The F-Statistic (335.2348) for Y_3 is the highest among the three models (as shown in Figure 5), indicating the firmest model fit for Y_3 and the p-value (0.000) again, confirming overall statistical significance. The p-values for all three models are 0.000, firmly rejecting the null hypothesis and confirming that the models are valid. These results indicate that the selected independent variables contribute significantly to the variance explanation of each dependent variable. Although Model Y_2 has a lower F-statistic, it remains statistically robust.

Table 18 provides the results of testing the parameters of an MLRM using a Hybrid Dmey approach on real data after removing insignificant independent variables. The constant (intercept) is statistically significant (p-value < 0.01) in all models. All independent variables (LYMp, NEUTp, etc.) have p-values < 0.01 in all models, meaning they significantly contribute to explaining the dependent variables. Positive t-statistics (e.g., PLCR for Y_1 and Y_2) suggest a positive relationship, and Negative t-statistics (e.g., RDWSD for Y_1 and Y_2) suggest an inverse relationship.

Table 18: Testing the parameters of an MLRM Using a Hybrid Dmey for Real Data After Removing the Insignificant Independent Variables.

Variable	t-statistic			p-value		
	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
Constant	-10.0004	-2.5379	8.0420	0.000	0.0120	0.000
LYMp	14.0238	8.1965	-13.1376	0.000	0.000	0.000
NEUTp	14.8343	9.2764	-6.3456	0.000	0.000	0.000
LYMn	3.9454	3.8668	27.0617	0.000	0.000	0.000
MCV	-3.1010	-3.9303	-6.9186	0.002	0.000	0.000
MCH	14.2517	6.9502	4.5173	0.000	0.000	0.000
RDWSD	-19.3727	-18.7214	-2.8721	0.000	0.000	0.005
PDW	-16.3275	-17.6344	4.8204	0.000	0.000	0.000
PCT	3.3152	3.2231	9.6479	0.001	0.001	0.000
PLCR	20.5450	18.2535	-5.8401	0.000	0.000	0.000

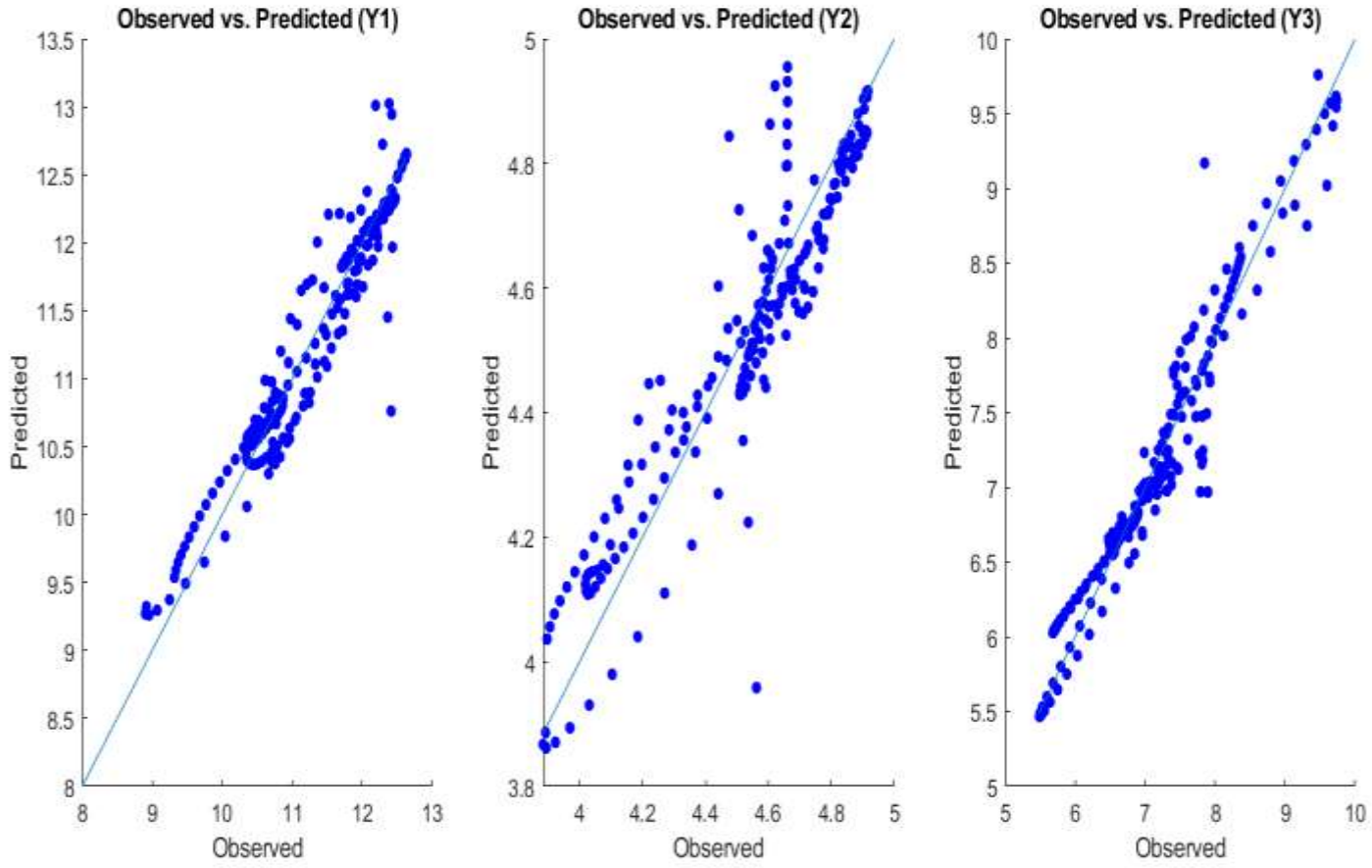

Figure 4: MLRM Dmey Filter for Real Data After Removing the Insignificant Independent Variables.

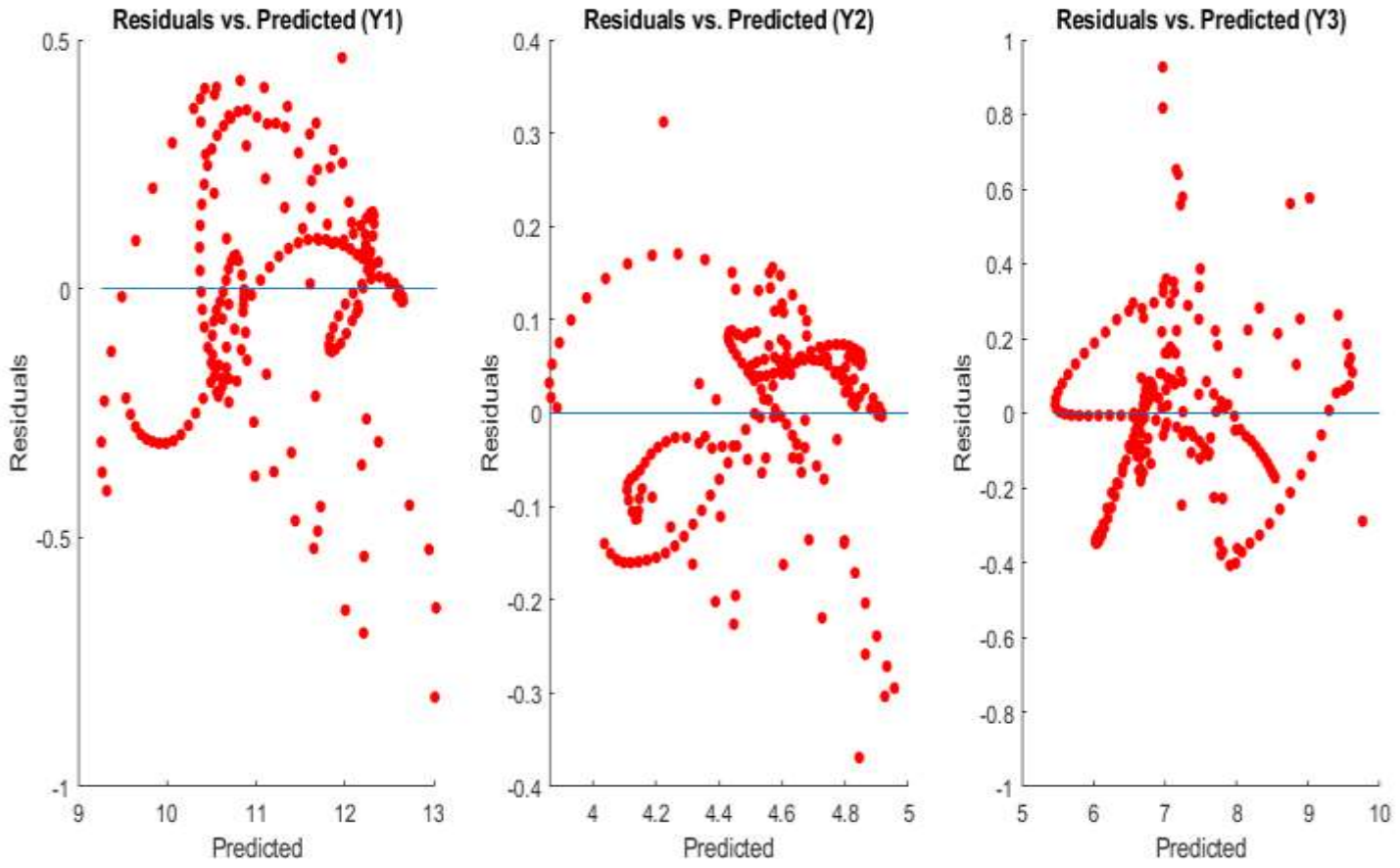

Figure 5: Residuals of MLRM Dmey Filter for Real Data After Removing the Insignificant Independent Variables.

Finally, Figure 5 presents the residuals of the multivariate linear regression model (MLRM) for the Dmey filter method. This effectively addressed the outliers and minimized noise in the dependent variable data, as indicated by the absence of any points outside the interval (-3, 3).

Conclusion

The simulation and real-data results collectively highlight the effectiveness of wavelet-based filtering, particularly the Dmey wavelet, in improving multivariate linear model estimation under data contamination. Simulation studies (Tables 2–7) showed that models without filtering suffered from high MSE and poor determination coefficients, while the Hampel filter, although reducing MSE, did not consistently lower SSE(β) compared to unfiltered models. In contrast, wavelet-based methods, especially Dmey, consistently achieved the lowest MSE (0.9966 for p=2, q=1, n=100) and the highest R^2 values (e.g., R^2Y_1=67.72%, R^2Y_2=49.14%). Model performance improved with an increasing number of independent variables and larger sample sizes, whereas an increase in dependent variables led to a slight reduction in performance.

Real-data analysis (Tables 11–18) further confirmed these findings. The Dmey wavelet method achieved the best performance, with an MSE of 0.7728 and high R^2 values (83.19% for Y_1, 74.33% for Y_2, and 87.50% for Y_3). Adjusted R^2 values closely matched R^2, indicating minimal overfitting and strong model generalizability. Further refinement using a hybrid Dmey filtering and variable selection approach improved results significantly, reducing the MSE to 0.0485 and increasing R^2 values to 91.46% (Y_1), 84.97% (Y_2), and 94.08% (Y_3).

Significance testing at α=0.01 identified key variables (LYMp, NEUTp, LYMn, MCV, MCH, PDW, PCT, PLCR) as consistently influential across all models. High F-statistics (225.95 for Y_1, 119.35 for Y_2, and 335.23 for Y_3) and p-values below 0.01 validated the statistical strength of the models. Additionally, the positive and negative t-statistics provided clear insight into the direction of each variable's effect on the dependent variables.

Overall, the Dmey wavelet filtering approach demonstrated substantial improvements in model accuracy, robustness, and interpretability, proving an effective method for handling multivariate regression models contaminated with outliers and noise.

Recommendations

1. The proposed methods addressed the problem of outliers in MLRM and were more efficient than the Hampel filter method.

2. Based on the SSE(β), MSE, and determination coefficients criteria, the proposed method (Wavelets Analysis) gave the best results. Among all the wavelets, Demy yielded the best results compared to every simulation state. Demy yielded the best results Among all the wavelets compared to every simulation state.

3. The proposed methods gave convergent results for all wavelets and were used in estimating outliers (DB5, Dmey, and Dmey).

4. Increasing the number of dependent variables increased the SSE(β) and MSE averages and decreased the coefficient of determination.

5. Increasing the sample size decreased the SSE(β) and MSE average and increased the coefficient of determination.

6. The proposed MLR models were suitable for the analysis of blood data.

7. The proposed Dmey wavelet performed the best compared with the rest of the wavelets for analyzing blood data.

8. All independent variables (LYMp, NEUTp, LYMn, MCV, MCH, RDWSD, PDW, PCT, and PLCR) have significantly contributed to explaining the dependent variables (HGB, RBC, and WBC).

9. LYMp, NEUTp, LYMn, MCH, PCT, and PLCR for HGB and RBC suggest a positive relationship and MCV, RDWSD, and PDW for HGB and RBC suggest an inverse relationship.

10. LYMn, MCH, PDW, and PCT for WBC suggest a positive relationship and LYMp, NEUTp, MCV, RDWSD, and PLCR for WBC suggest an inverse relationship.

References

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

[2] Kutner, M. H., Nachtsheim, C. J., Wasserman, W. & Neter, J. *Applied Linear Regression Models* (McGraw-Hill, Chicago, IL, 2004).

[3] Aggarwal, C. C. *An Introduction to Outlier Analysis* (Springer International Publishing, New York, 2017).

[4] Hampel, F. R. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* **69**, 383–393 (1974).

[5] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. *Robust Statistics: The Approach Based on Influence Functions* (Wiley & Sons, New York, 2005).

[6] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. & Liu, H. Feature selection: a data perspective. *ACM Comput. Surv.* **50**, 94 (2017).

[7] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).

[8] Unser, M. Sampling—50 years after Shannon. *Proc. IEEE* **88**, 569–587 (2000).

[9] Mallat, S. G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 674–693 (1989).

[10] Syed Abd Mutalib, S. S. S., Satari, S. Z. & Wan Yusoff, W. N. S. Comparison of robust estimators' performance for detecting outliers in multivariate data. *J. Stat. Model. Anal.* **3**, 36–64 (2021).

[11] Omer, A. W., Sedeeq, B. S. & Ali, T. H. A proposed hybrid method for multivariate linear regression model and multivariate wavelets (simulation study). *Mitanni J. Humanit. Sci.* **5**, 112–124 (2024).

[12] Kareem, N. S., Mohammad, A. S. & Ali, T. H. Construction robust simple linear regression profile monitoring. *J. Kirkuk Univ. Adm. Econ. Sci.* **9**, 242–257 (2019).

[13] Tabachnick, B. G. & Fidell, L. S. *Using Multivariate Statistics*, 7th edn (Pearson, 2021).

[14] Cook, R. D. Detection of influential observation in linear regression. *Technometrics* **19**, 15–18 (1977).

[15] Greene, W. H. *Econometric Analysis, Global Edition* (8th edn, Pearson, 2019).

[16] Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. *Multivariate Data Analysis* (Annabel Ainscow, China by RR Donnelley, 2018).

[17] Omer, A. W. & Ali, T. H. Dealing with the outlier problem in multivariate linear regression analysis using the Hampel filter. *Kurdistan J. Appl. Res.* **10** (2025).

[18] Breunig, M. M., Kriegel, H. P., Ng, R. T. & Sander, J. LOF: identifying density-based local outliers. *ACM SIGMOD Int. Conf. Manag. Data*, 93–104 (2000).

[19] Rousseeuw, P. J. & Leroy, A. M. *Robust Regression and Outlier Detection* (Wiley, New York, 2003).

[20] Pearson, R. K. Outliers in process modeling and identification. *IEEE Trans. Control Syst. Technol.* **10**, 55–63 (2002).

[21] Birdawod, H. Q., Khudhur, A. M., Kadir, D. H. & Saleh, D. M. A wavelet shrinkage mixed with a single-level 2D discrete wavelet transform for image denoising. *Kurdistan J. Appl. Res.* **9**, 1–12 (2024).

[22] Nielsen, M. On the construction and frequency localization of finite orthogonal quadrature filters. *J. Approx. Theory* **108**, 36–52 (2001).

[23] Strang, G. & Nguyen, T. *Wavelets and Filter Banks* (Cambridge Press, Wellesley, 1996).

[24] Mallat, S. G. Theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 674–693 (1989).

[25] Guo, T., Xu, Y., Zhang, Y., Li, J., Liu, H. & Wang, H. A review of wavelet analysis and its applications: challenges and opportunities. *IEEE Access* **10**, 58869–58903 (2022).

[26] Abdulqader, Q. M. & Ali, T. H. Comparing the Box-Jenkins models before and after the wavelet filtering in terms of reducing the orders with application. *J. Concrete Appl. Math.* **11**, 190–198 (2013).

[27] Arts, L. P. & Van den Broek, E. L. The fast continuous wavelet transformation (fCWT) for real-time, high-quality, noise-resistant time-frequency analysis. *Nat. Comput. Sci.* **2**, 47–58 (2022).