Research Article / Review Article

FUPRE JOURNAL 10(1):120-x (2026)
https://doi.org//10.60787/FJSIR.v10i1.120-133

FUPRE Journal
of
sScientific and Industrial Research

ISSN: 2579-1184(Print) ISSN: 2578-1129 (Online)

http://fupre.edu.ng/journal

An Improved Localized Regression Approach for Bioactive Compound Extraction from Citrullus lanatus Rind via the Central Composite Design

***EGUASA, O.*[1,*], *OSAWE, S. O.*[2]**

[1]*Department of Physical Sciences, Benson Idahosa University, Benin City, Edo State*

[2]*Department of Biological Sciences, Benson Idahosa University, Benin City, Edo State*

ARTICLE INFO

Received: 23/02/2026
Accepted: x/06/2026

Keywords

Central composite design, Citrullus lanatus rind, Improved local linear regression

ABSTRACT

The rind of Citrullus lanatus (watermelon), commonly regarded as agricultural waste, is a rich source of bioactive compounds with significant antioxidant properties, making it a promising raw material for applications in the food, cosmetic, and pharmaceutical industries. This study employed Response Surface Methodology to optimize the extraction conditions for maximizing the recovery of phytochemicals from watermelon rind. A Central Composite Design was adopted to investigate the effects of key extraction variables, namely extraction time and extraction temperature, on the total extraction yield. Previous studies evaluated four conventional regression models Linear, Two-Factor Interaction, Quadratic, and Cubic alongside the adaptive local linear regression model with adaptive bandwidths to determine the optimal extraction yields at low, medium, and high levels of the extraction variables. Building on this framework, the present study incorporated an additional critical extraction factor and expanded the experimental design to five coded levels ($-\alpha$, -1, 0, 1, $+\alpha$), thereby improving model flexibility and enhancing the precision of extraction optimization. Furthermore, an improved adaptive local linear regression model, with variable bandwidth selection was utilized to enhance model goodness-of-fit, minimize deviations from the zero-residual line, and improve prediction accuracy. Comparative analyses, supported by three-dimensional response surface visualizations, demonstrated that the proposed model performed comparably to the Quadratic Model while consistently outperforming the Linear, Two-Factor Interaction, Cubic, and conventional local linear regression models in predicting the extraction of bioactive compounds from Citrullus lanatus rind. The optimization results indicated that the proposed model improved adaptive local linear regression model and local linear regression achieved superior process requirements under the optimal extraction conditions. In addition, simulation studies were conducted to evaluate the robustness of the regression models under a misspecification parameter, γ, within the interval $0 \leq \gamma \leq 1$. The results demonstrated that the simulated improved adaptive local linear regression model remained stable as γ increased from 0 to 1, confirming its robustness against model misspecification. Overall, the findings highlight the effectiveness of the improved local linear regression approach for optimizing bioactive compound extraction.

*Corresponding author, e-mail: aoeguasa@biu.edu.ng
https://doi.org/10.60787/FJSIR.v10i1.120-133
©Scientific Information, Documentation and Publishing Office at FUPRE Journal

1. INTRODUCTION

Citrullus lanatus (watermelon) is a tropical fruit extensively cultivated for its sweet and succulent flesh. Despite its economic importance, the rind, which accounts for approximately 30–40% of the total fruit weight, is commonly discarded as agricultural waste, even though it is a rich source of valuable bioactive compounds, including phenolic compounds, flavonoids, and citrulline (Parmar *et al*., 2017; Luo *et al*., 2019; Eguasa and Okungbowa, 2025). These phytochemicals possess significant antioxidant, anti-inflammatory, and antimicrobial properties, highlighting the potential of watermelon rind as a valuable raw material for functional foods, nutraceuticals, and pharmaceutical products (Alsharabasy *et al*., 2020; Oyeleke *et al.*, 2021).

The efficient extraction of these bioactive constituents is essential for maximizing their recovery and biological efficacy. Conventional extraction techniques often rely on trial-and-error experimentation, which is labor-intensive, time-consuming, and may fail to achieve optimal extraction efficiency. In contrast, Response Surface Methodology (RSM), an integrated collection of mathematical and statistical techniques, provides a systematic and efficient framework for optimizing process variables, evaluating their interactions, and predicting extraction performance (Myers *et al*., 2016). Owing to its ability to minimize experimental runs while maximizing process efficiency, RSM has been widely applied in the optimization of phytochemical extraction from various plant materials, resulting in enhanced yields and improved recovery of target bioactive compounds (Bezerra *et al*., 2008).

The present study employs Response Surface Methodology to optimize the extraction conditions for bioactive compounds from *Citrullus lanatus* rind with the objective of maximizing phytochemical yield. Furthermore, phytochemical profiling and antioxidant activity assays are conducted to assess the functional properties of the extracts. The findings of this study are expected to facilitate the valorization of watermelon rind as a cost-effective, sustainable, and environmentally friendly source of natural antioxidants with promising applications in the food, pharmaceutical, and cosmetic industries.

According to Nair *et al*. (2014) and Yeniay (2014), Response Surface Methodology (RSM) is a statistical and mathematical technique widely employed by engineers and industrial statisticians for empirical model development, with the primary objective of optimizing response variables influenced by multiple explanatory variables.

RSM is suitable for optimizing the response variable $\boldsymbol{y}$ as a function of several explanatory variables $(x_{i1}, x_{i2}, \ldots, x_{ik})$ which can be modeled as:

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{ik}) + \varepsilon_i, \quad i = 1,2, \ldots, n \tag{1}$$

where ε_i is the error term and assumed to have a normal distribution with mean zero and variance $\boldsymbol{\sigma}^2$. The surface represented by $f(x_{i1}, x_{i2}, \ldots, x_{ik})$ is termed a response surface (Wan and Birch, 2011).

The true response function f is typically unknown and must be estimated. In Response Surface Methodology (RSM), the goal is to determine the functional relationship between the response variable y and the corresponding explanatory variables $(x_{i1}, x_{i2}, \ldots, x_{ik})$.

The conventional approach to modeling the relationship between the *k*th explanatory variable and the ith response assumes that the underlying relationship can be adequately represented by a parametric functional form. Parametric

*Corresponding author, e-mail: aoeguasa@biu.edu.ng
https://doi.org/10.60787/FJSIR.v10i1.120-133
©Scientific Information, Documentation and Publishing Office at FUPRE Journal

regression models are effective when an appropriate functional form that accurately reflects the underlying data structure can be correctly specified.

Therefore, the general parametric regression model in matrix notation can be written as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where $\boldsymbol{y}$ is a vector of response, $\boldsymbol{X} = \boldsymbol{X}^{(OLS)}$ is the OLS model matrix, $\boldsymbol{\beta}$ is the unknown parameter vector and $\boldsymbol{\varepsilon}$ is the vector of error term assumed to be normally distributed with zero mean and constant variance property.

The parameter vector in Equation (2) is conventionally estimated using the Ordinary Least Squares (OLS) method. The parameter vector estimates $\widehat{\boldsymbol{\beta}}$ in (2) is given as:

$$\widehat{\boldsymbol{\beta}}^{(OLS)} = \left(\boldsymbol{X}'^{(OLS)}\boldsymbol{X}^{(OLS)}\right)^{-1}\boldsymbol{X}'^{(OLS)}\boldsymbol{y}, \ \boldsymbol{X} = \boldsymbol{X}^{(OLS)} \quad (3)$$

The estimated responses for the i^{th} location can be written as:

$$\widehat{\boldsymbol{y}}_i^{(OLS)} = \boldsymbol{x}_i'^{(OLS)}\widehat{\boldsymbol{\beta}}^{(OLS)} = \boldsymbol{x}_i'^{(OLS)}\left(\boldsymbol{X}'^{(OLS)}\boldsymbol{X}^{(OLS)}\right)^{-1}\boldsymbol{X}'^{(OLS)}\boldsymbol{y}, \quad i = 1,2,\dots,n \quad (4)$$

where $\boldsymbol{x}_i'^{(OLS)}$ is the i^{th} row of matrix $\boldsymbol{X}^{(OLS)}$, $n \times (k+1)$ vector.

$\boldsymbol{H}_i'^{(OLS)} = \boldsymbol{x}_i'^{(OLS)}\left(\boldsymbol{X}'^{(OLS)}\boldsymbol{X}^{(OLS)}\right)^{-1}\boldsymbol{X}'^{(OLS)}$ is the i^{th} row of the OLS "HAT" matrix of dimension $n \times n$, $\boldsymbol{H}^{(OLS)}$. The estimated response in the i^{th} location is given by:

$$\widehat{\boldsymbol{y}}^{(OLS)} = \boldsymbol{H}^{(OLS)}\boldsymbol{y}. \quad (5)$$

where the matrix $\boldsymbol{H}^{(OLS)}$ is given as:

$$\boldsymbol{H}^{(OLS)} = \begin{bmatrix} \boldsymbol{H}_1^{(OLS)} \\ \boldsymbol{H}_2^{(OLS)} \\ \vdots \\ \boldsymbol{H}_n^{(OLS)} \end{bmatrix}, \quad (6)$$

(Carley *et al*., (2004); River, (2009))

2. MATERIALS AND METHODS

When the underlying functional relationship between the response and explanatory variables is partially known or cannot be adequately specified, nonparametric regression provides a more flexible and appropriate modeling framework. Among the available nonparametric techniques, Local Linear Regression (LLR) is particularly effective due to its ability to estimate complex relationships without imposing a predetermined functional form. LLR extends the traditional least squares approach by incorporating kernel-based weighting within a first-order local polynomial regression framework (polynomial degree = 1). Compared with conventional kernel regression methods, LLR offers superior performance by substantially reducing boundary bias, accommodating irregularly spaced explanatory variables, and producing more accurate and stable estimates (Ruppert and Wand, 1994; Walker *et al*., 2002).

The LLR model is derived from standard least squares theory. The LLR estimator $\widehat{\boldsymbol{y}}_i^{(LLR)}$ of $\boldsymbol{y}_i$ is given as:

$$\widehat{\boldsymbol{y}}_i^{(LLR)} = \boldsymbol{x}_i'^{(LLR)}(\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_i\boldsymbol{X}^{(LLR)})^{-1}\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_i\boldsymbol{y} = \boldsymbol{H}_i^{(LLR)}\boldsymbol{y}, \quad (7)$$

where $\boldsymbol{y} = (y_1, \dots y_n)'$, $\boldsymbol{x}_i'^{(LLR)} = (1\ x_{i1} \dots x_{ik})$ is the i^{th} row of the LLR model matrix, $\boldsymbol{X}^{(LLR)}$ given as:

$$\boldsymbol{X}^{(LLR)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad (8)$$

We define $\boldsymbol{W}$, an $n \times n$ diagonal matrix of kernel weights for estimating the response as

In terms of location, $\boldsymbol{W} = \boldsymbol{W_i}$

$$\boldsymbol{W_i} = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}, i = 1, \ \ 2 \ , \ \ . \ \ . \ \ . \ \ , n \, . \quad (9)$$

(Wan and Birch, (2011); Eguasa *et al*., (2019)).

Equation (9) translates to

$\boldsymbol{W_i} = dia(w_{i1}, w_{i2},\ldots, w_{in})$ for each $i = 1, \ \ 2 \ , \ \ . \ \ . \ \ . \ \ , n \, .$

$$\widehat{\boldsymbol{y}}^{(LLR)} = \boldsymbol{H}^{(LLR)}\boldsymbol{y}, \quad (10)$$

Equation (7) can be written in terms of hat matrix as:

$$\boldsymbol{H}^{(LLR)} = \begin{bmatrix} \boldsymbol{x_1'}^{(LLR)}\left(\boldsymbol{X'}^{(LLR)}\boldsymbol{W_1}\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X'}^{(LLR)}\boldsymbol{W_1} \\ \boldsymbol{x_2'}^{(LLR)}\left(\boldsymbol{X'}^{(LLR)}\boldsymbol{W_2}\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X'}^{(LLR)}\boldsymbol{W_2} \\ \vdots \\ \boldsymbol{x_n'}^{(LLR)}\left(\boldsymbol{X'}^{(LLR)}\boldsymbol{W_n}\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X'}^{(LLR)}\boldsymbol{W_n} \end{bmatrix} \quad (11)$$

where the $n \times n$ matrix, $\boldsymbol{H}^{(LLR)}$ is the LLR hat matrix.

The drawback of LLR model is high bias in regions where there is curvature because the structure of model matrix of the LLR model lacks the quadratic terms (Hastie, *et al*., (2009); Rivers, (2009)).

2.1. Locally Adaptive Bandwidths

The bandwidth is the most important parameter in nonparametric regression estimation because of its smoothing properties (Kohler *et al.*, (2014)).

Eguasa *et al.* (2022) presented data-driven locally adaptive bandwidths:

$$b_{ij} = T_{1j}(\tfrac{1}{2} - \tfrac{x_{ij}}{T_{2j}})^2, \ \ i = 1,2, \ \ \ldots, \ n; j = 1,2,\ldots,k \quad (12)$$

where, $0 < b_{ij} \leq 1,, \ \ T_{1j} > 0, \ T_{2j} > 0.$

The $b*_{ij}$ of the locally adaptive optimal bandwidths from Equation (12) is obtained at an optimally selected values of T_{1j}, T_{2j}, (hereafter referred to as T_{1j}^* and T_{2j}^*, respectively), $j = 1,2,\ldots,k$, based on the minimization of the $PRESS^{**}$ criterion.

2.2. Algorithm I: Leave – One – Out Cross Validation Technique for Choosing Adaptive Bandwidths and Fixed Mixing Parameter

Step 1: Obtain the bandwidth b_{ij} for location *ith* and *jth* explanatory variables

$$b_{ij} = T_{1j}(\frac{1}{2} - \frac{x_{ij}}{T_{2j}})^2,\ i = 1,2,\ \dots,\ n; j = 1,2,\dots,k$$

Step 2: Define a set ω of suitable values of bandwidths (for RSM data, $\omega \in (0,1]$) where the locally adaptive bandwidths b_{ij}, $i = 1,2,\dots,n$ are allocated values.

Step 3: Define the leave-one-out cross validation for all likely set of $(b_{1j}, b_{2j}, \dots, b_{nj})$ over the complete range in the set $\omega \in (0,1]$:

$$\hat{y}_{i,-i}\left(b_{1j}, b_{2j}, \dots, b_{nj}\right), i = 1,2,\ \dots,\ n; j = 1,2,\dots,k$$

Step 4: Define $PRESS^{**}$ Criterion:

$$PRESS^{**}(\varphi) = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}^{(.)}(\varphi)\right)^2}{n - trace\left(H^{(.)}(\varphi)\right) + (n-k-1)\frac{SSE_{max} - SSE_{\varphi}}{SSE_{max}}};$$

for choosing locally adaptive bandwidths in the interval (0, 1] and obtain $\hat{y}_{i,-i}(\varphi)$ the estimated response at location *i* , leaving out the *ith* response for the set of locally adaptive bandwidths:

$\varphi = [b_{1j}, b_{2j}, \dots, b_{nj};\ \overset{\circ}{b_{1j}}, \overset{\circ}{b_{2j}}, \dots, \overset{\circ}{b_{nj}};\ b_{1j}^{*}, b_{2j}^{*}, \dots, b_{nj}^{*}]$.

Step 5: obtain SSE_{max} as $\boldsymbol{b}$ tends to infinity, (say $\boldsymbol{b}$ = 1000000000000000000000) in $\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(.)}(\boldsymbol{b})$

$$SSE_{max} = \sum_{i=1}^{n}\left(\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(.)}(\boldsymbol{b})\right)^2$$

Step 6: obtain SSE_{φ} for a specific set of locally adaptive bandwidths:

$$SSE_{\varphi}\left(b_{1j}, b_{2j}, \dots, b_{nj}\right) = \sum_{i=1}^{n}\left(\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(.)}(b_{1j}, b_{2j}, \dots, b_{nj})\right)^2$$

Step 7: obtain the fixed mixing parameter $\boldsymbol{\lambda}$ using $PRESS^{**}$ Criterion:

$$PRESS^{**}(\varphi, \lambda) = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}^{(.)}(\varphi,\lambda)\right)^2}{n - trace\left(H^{(.)}(\varphi,\lambda)\right) + (n-k-1)\frac{SSE_{max} - SSE_{\varphi}}{SSE_{max}}},$$

Step 8: Lastly, outline the optimal locally adaptive bandwidths $b_{1j}^{*}, b_{2j}^{*}, \dots, b_{nj}^{*}$ and optimal fixed mixing parameter $\lambda^{*}\ of\ \lambda$ that minimize $PRESS^{**}$ Criterion (Eguasa *et al*., (2022)).
Step 9: STOP

3. APPLICATION

This study investigates the performance of the adaptive $ImLLR_{AB}$ model relative to the Linear, Two-Factor Interaction (2FI), Quadratic (QM), Cubic, and LLR_{AB} models using Response Surface Methodology (RSM) datasets. The comparison is based on goodness-of-fit statistics and the optimal combinations of the explanatory variables that maximize the response.

3.1. Single response optimization problem

In this study, which considers a single response variable, the objective of the optimization process is to determine the combination of explanatory variable settings that maximizes or minimizes the fitted response (Eguasa *et al*., 2022). Accordingly, the optimization procedure is based on the constrained maximization of the estimated Squared Distance from the Target (SDT), which is defined as follows:

$$\text{Maximize } \widehat{SDT} = (\hat{y}(\boldsymbol{x}) - T)^2 \quad (13)$$

$$s.t\ \boldsymbol{x} \epsilon\ \varphi\ ,$$

where φ is the design space for the study, T denotes the target value set by the researcher, $\hat{y}(\boldsymbol{x})$ is the estimated response at the settings $\boldsymbol{x}$ of the explanatory variables (Pickle, (2006); Najafi *et al*., (2011); Eguasa *et al*., (2022)).

3.2. Genetic Algorithm

Genetic algorithms are iterative optimization techniques that repetitively apply GA operations (such as selection, crossover and mutation) to a set of solutions until some criterion of convergence is reached (Eguasa *et al.,* (2022); Eguasa and Okungbowa, 2025).

3.3. Application: Single response extraction yield for response surface process data

The problem of the study [21] was to relate chemical yield (y) to temperature (x_1) and time (x_2) with the intention to maximize the extraction yield. The raw data is given in Table 1 as:

Table 1: ***Single response extraction yield for CCD response surface process data***

$Exptal.Run$	$Temp.(℃) = x_1$	$Time\ (Hrs) = x_2$	$Actual\ yield = y$	$Predicted\ yield = \hat{y}$
1	30	4	26.32	26.49
2	50	4	28.55	28.81
3	30	6	30.53	30.87
4	50	6	33.58	34.01
5	25.86	5	24.18	23.95
6	54.14	5	28.17	27.81
7	40	3.59	31.01	30.83
8	40	6.41	38.01	37.59
9	40	5	35.73	35.21
10	40	5	34.89	35.21
11	40	5	35.79	35.21
12	40	5	34.02	35.21
13	40	5	35.62	35.21

Source: (Kamal *et al.,* (2023))

The data is obtained using the Central Composite Design is given in Table 1. The problem of the study Kamal *et al*. (2023) was to relate extraction yield (y) to temperature (x_1) and time (x_2) with the intention to maximize the extraction yield. The data is obtained using the Central Composite Design is given in Table 2.

Table 2: Single response extraction yield for CCD response surface process data

$Exptal.Run$	x_1	x_2	y
1	-1	-1	26.32
2	1	-1	28.55
3	-1	1	30.53
4	1	1	33.58
5	-1.414	0	24.18
6	1.414	0	28.17
7	0	-1.414	31.01
8	0	1.414	38.01
9	0	0	35.73
10	0	0	34.89
11	0	0	35.79
12	0	0	34.02
13	0	0	35.62

3.4. Data transformation using central composite design (CCD)

From the nonparametric regression procedures in RSM, the values of the explanatory variables are coded between 0 and 1. The data collected via a CCD is transformed by a mathematical relation:

$$x_{NEW} = \frac{Min(x_{OLD}) - x_0}{\left(Min(x_{OLD}) - Max(x_{OLD})\right)} \quad (14)$$

where x_{NEW} is the transformed value, x_0 is the target value that needed to be transformed in the vector containing the old coded value, represented as x_{OLD}, Min (x_{OLD}) and $Max(x_{OLD})$ are the minimum and maximum values in the vector x_{OLD} respectively, (Eguasa *et al.*, (2022)).

The natural or coded variables in Table 1 can be transformed to explanatory variables in Table 2 using Equation (14)

Target points needed to be transformed for location 2 under the coded variables are given below:

Target points x_0: $1, -1$; $Min(x_{OLD})$: $-1.414, -1.414$; $Max(x_{OLD})$: $1.414, 1.414$

$$x_{NEW} = \frac{Min(x_{OLD}) - x_0}{\left(Min(x_{OLD}) - Max(x_{OLD})\right)}$$

$$Explanatory\ variable\ x_1 :\ x_{21} = \frac{-1.414 - (1)}{((-1.414) - (1.414))} = 0.8536$$

$$Explanatory\ variable\ x_2 :\ x_{22} = \frac{-1.414 - (-1)}{((-1.414) - (1.414))} = 0.1464$$

Table 3: Single response extraction yield for CCD response surface process data

*Corresponding author, e-mail: aoeguasa@biu.edu.ng
https://doi.org/10.60787/FJSIR.v10i1.120-133
©Scientific Information, Documentation and Publishing Office at FUPRE Journal

i	x_1	x_2	y
1	0.1464	0.1464	26.32
2	0.8536	0.1464	28.55
3	0.1464	0.8536	30.53
4	0.8536	0.8536	33.58
5	0.0000	0.5000	24.18
6	1.0000	0.5000	28.17
7	0.5000	0.0000	31.01
8	0.5000	1.0000	38.01
9	0.5000	0.5000	35.73
10	0.5000	0.5000	34.89
11	0.5000	0.5000	35.79
12	0.5000	0.5000	34.02
13	0.5000	0.5000	35.62

Source: [3]

Table 4: Locally Adaptive Optimal Bandwidths for LLR_{AB} and $ImLLR_{AB}$ in the Single Response Process Data

i	***LLR AB*** LLR Bandwidths		***Im LLR AB*** LLR Bandwidths	
	x_1	x_2	x_1	x_2
	b_{i1} T_{11}^{*} =1.6519 T_{21}^{*} = 4.9725	b_{i2} T_{12}^{*} =3.9205 T_{22}^{*} = 4.8439	b_{i1} T_{12}^{*} =0.3862 T_{22}^{*} =2.5185	b_{i2} T_{12}^{*} =0.9747 T_{22}^{*} =4.0849
1	0.1781	0.8652	0.3381	0.3299
2	0.3658	0.8652	0.1503	0.3299
3	0.1781	0.4110	0.3381	0.7486
4	0.4130	0.4110	0.1503	0.7486
5	0.1476	0.6172	0.3450	0.4467
6	0.2636	0.6172	0.0982	0.4467
7	0.2636	0.9801	0.2692	0.3199
8	0.2636	0.3378	0.2692	0.9542
9	0.2636	0.6172	0.2692	0.4467
10	0.2636	0.6172	0.2692	0.4467
11	0.2636	0.6172	0.2692	0.4467
12	0.2636	0.6172	0.2692	0.4467
13	0.1781	0.6172	0.2692	0.4467

Table 5: Comparison of the goodness-of-fit statistics of each method for the process data

METHOD	b^*	DF_{error}	SSE	MSE	LACK OF FIT	R^2	R^2_{adj}	$Pred.R^2$	$PRESS$
OLS	-	5.000	212.47	42.49	-	28.16	13.80	-33.23	287.18
$2F1$	-	4.15	-	-	-	28.24	4.32	-65.88	357.55
QM	-	5	-	-	0.7848	98.57	97.55	95.75	**9.16**
$Cubic\ M$	-	5	**-**	**-**	0.7274	98.60	96.65	77.14	49.28
LLR_{AB}	*	5.7937	3.0093	**0.5194**	0.6024	98.60	**97.11**	92.86	15.38
$Im\ LLR\ AB$	*	4.4835	**2.3793**	0.5307	**0.0738**	**98.90**	97.10	**92.96**	15.18

where * represent the locally adaptive bandwidths as given in Table 5, b_{i1} and b_{i2} respectively. The $Im\ LLR\ AB$ model outperformed the other models in terms of the selected performance statistics, with the bolded entries indicating the best-performing model for each criterion.

Table 6: Comparison of optimization results (process requirement) for the Extraction Process Data

Approach	$x_1\ (Temp\ ℃)$	$x_2(Hours)$	$\hat{y}\ \%$
QM	41.7	6	37.24
LLR_{AB}	0.7272= 39.4	1.0000 = 6.41	36.19
$Im\ LLR\ AB$	0.8165 = 44.3	0.9166 = 6.40	34.00

From Table 6, provides the extraction yield for QM, LLR_{AB}, and $ImLLR_{AB}$ the two settings of the explanatory variables give the desired process satisfaction. Where the QM maximized the process requirements more than LLR_{AB}, and $ImLLR_{AB}$.

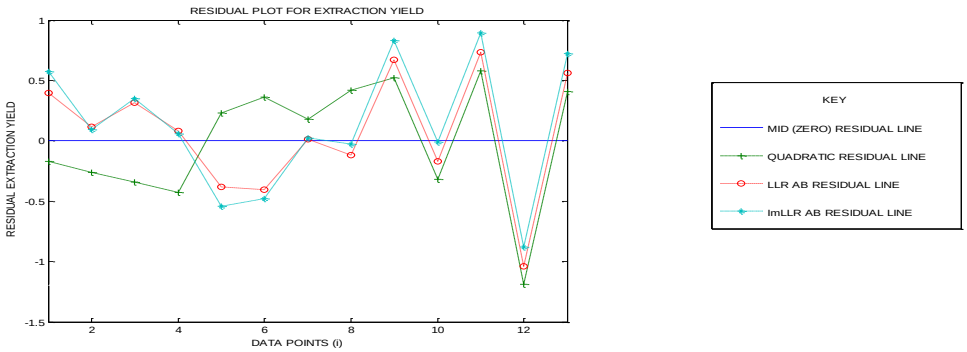


Figure 1. Model Residuals for single response $ImLLR_{AB}$, LLR_{AB} *and* QM Extraction Yield Process Data

As a follow up of the optimization result as given above, the residual result show that $ImLLR_{AB}$ model have the closest residual line to the zero residual line. This also suggests that $ImLLR_{AB}$ model the adaptive bandwidth has a better fit to the Quadratic model and LLR_{AB}.

3.5. Simulation Study

In this section, we compare the performances of the respective regression models, *QM* and LLR_{AB}, and $ImLLR_{AB}$ using data from the optimal responses within the misspecification parameter γ simulated in the range 0, 0.25, 0.50, 0.75, 1.0; where the x_{1i},x_{2i} and x_{ij} are the explanatory variables.

3.5.1. *Simulation Study: $ImLLR_{AB}$ Model*

The estimated response derived from the genetic algorithm was employed to simulate the two explanatory variables, aiming to validate the stability of the regression model under the misspecification parameter (mixing parameter) γ, which was constrained to the interval $0 \leq \gamma \leq 1$ (Eguasa and Okungbowa, 2025).

$ImLLR_{AB}$ Model: `Z=34+MS*(3*sin(3*pi*x1)-2*cos(2*pi*x2)+2*sin(4*pi*x1.*x2))`

$$MS = \gamma = 0.0$$

Therefore, the simulated plot gave optimal response of 34% when $\gamma = 0.0$ as given in Figure (1)

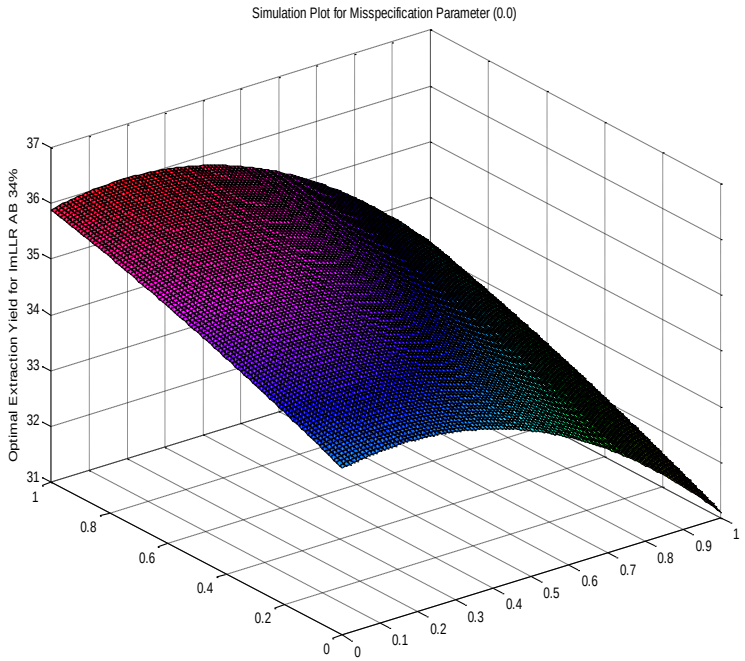


Figure 1: Simulated optimal response (34%) for *ImLLR AB E*xtraction Yield Process Data

$ImLLR_{AB}$ Model: `Z=34+MS*(3*sin(3*pi*x1)-2*cos(2*pi*x2)+2*sin(4*pi*x1.*x2))`

$$MS = \gamma = 0.25$$

Therefore, the simulated plot gave optimal response of 34% when $\gamma = 0.25$ as given in Figure (2)

*Corresponding author, e-mail: aoeguasa@biu.edu.ng
https://doi.org/10.60787/FJSIR.v10i1.120-133
©Scientific Information, Documentation and Publishing Office at FUPRE Journal

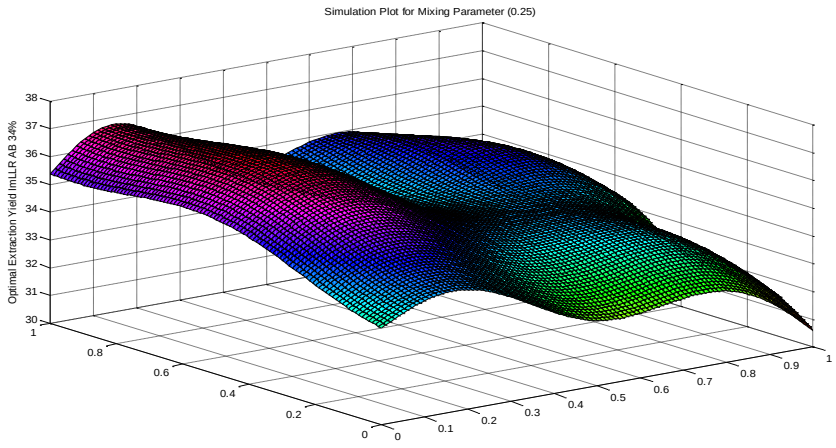


Figure 2: Simulated optimal response (34%) for $ImLLR\ AB$ Extraction Yield Process Data

$\boldsymbol{ImLLR_{AB}}$ **Model:** `Z=34+MS*(3*sin(3*pi*x1)-2*cos(2*pi*x2)+2*sin(4*pi*x1.*x2))`

$$MS = \gamma = 0.50$$

Therefore, the simulated plot gave optimal response of 34% when $\gamma = 0.50$ as given in Figure (3)

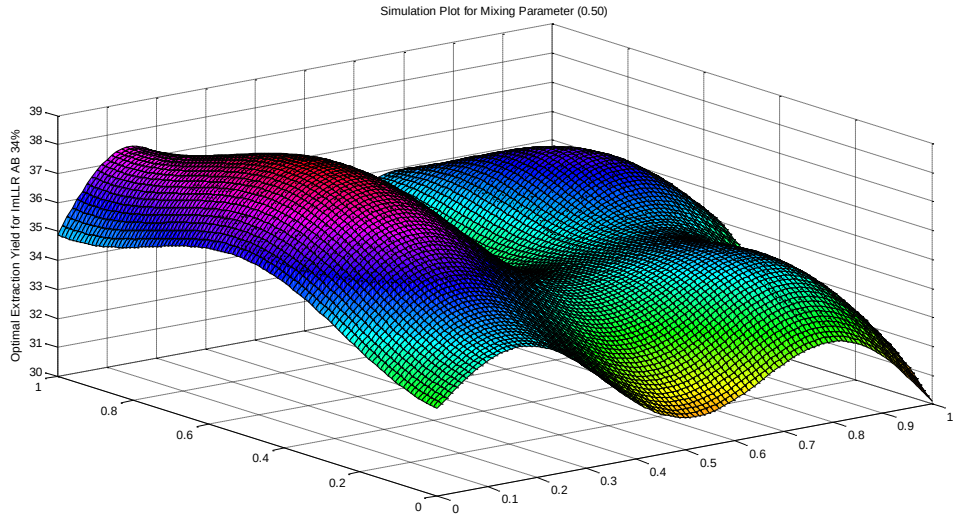


Figure 3: Simulated optimal response (34%) for $ImLLR\ AB$ Extraction Yield Process Data

$\boldsymbol{ImLLR_{AB}}$ **Model:** `Z=34+MS*(3*sin(3*pi*x1)-2*cos(2*pi*x2)+2*sin(4*pi*x1.*x2))`

$$MS = \gamma = 0.75$$

Therefore, the simulated plot gave optimal response of 34% when $\gamma = 0.75$ as given in Figure (4)

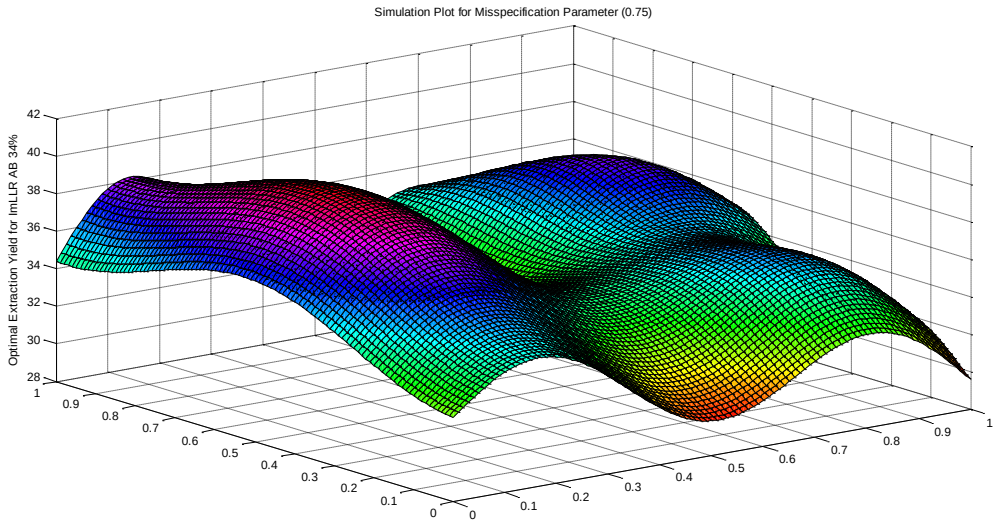


Figure 4: Simulated optimal response (34%) for $ImLLR\ AB$ Extraction Yield Process Data

***ImLLR*$_{AB}$ Model:** `Z=34+MS*(3*sin(3*pi*x1)-2*cos(2*pi*x2)+2*sin(4*pi*x1.*x2))`

$$MS = \gamma = 1.00$$

Therefore, the simulated plot gave optimal response of 34% when $\gamma = 1.00$ as given in Figure (5)

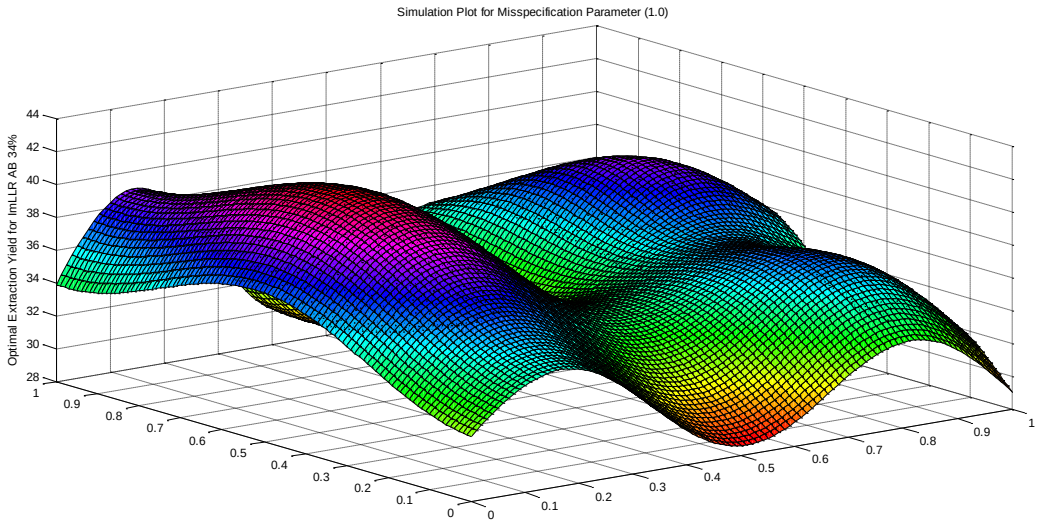


Figure 5: Simulated optimal response (34%) for $ImLLR\ AB$ Extraction Yield Process Data

4. RESULTS AND DISCUSSION

This study investigates the optimization and simulation of bioactive compound extraction from *Citrullus lanatus* (watermelon) rind using Response Surface Methodology (RSM). Although commonly discarded as agricultural waste, watermelon rind is a rich source of valuable phytochemicals, including phenolic compounds and flavonoids, which

possess significant antioxidant properties. To maximize the extraction efficiency of these bioactive constituents, a Central Composite Design (CCD) within the RSM framework was employed to evaluate the effects of key extraction variables, specifically extraction temperature and extraction time, on the extraction yield.

The experimental design data were fitted using a linear, 2F1, cubic, second-order polynomial equation, LLR_{AB}, and $ImLLR_{AB}$. Statistical analysis confirmed the adequacy of these models, showing significant regression results and minimal lack-of-fit. Further optimization using a genetic algorithm (GA) effectively refined the extraction parameters to maximize the yield of bioactive compounds. To evaluate the model's stability and robustness, simulations were performed under parameter misspecification scenarios, using optimal responses obtained from both literature sources and GA estimations. These simulations were conducted within a defined parameter range ($0 \leq \gamma \leq 1$), demonstrating the reliability of the regression models under varying conditions (mixing parameter or misspecification parameter).

5. CONCLUSION

The comparative analyses are driven by three-dimensional response surface plot, confirmed that the improved local linear regression model, $ImLLR_{AB}$, with variable bandwidth selection performed comparably to the QM and LLR_{AB} while outperforming the other regression models evaluated. The regression and optimization results further indicated that the QM achieved slightly superior optimization performance under the optimal extraction conditions. In addition, simulation studies were conducted to evaluate the stability of the regression models under a misspecification parameter γ, within the interval $0 \leq \gamma \leq 1$. The findings revealed that the simulated optimal responses remained stable, exhibiting a systematic reduction as (\gamma) increased from 0 to 1, thereby confirming the robustness of the proposed modeling approach. Overall, the integration of Response Surface Methodology (RSM), Genetic Algorithm (GA)-based optimization, and simulation provide a robust and reliable framework for optimizing bioactive compound extraction and promoting the sustainable valorization of agricultural by-products as functional ingredients for nutraceutical and pharmaceutical applications.

Acknowledgment

I am deeply grateful to God Almighty the giver of life and knowledge.

Conflict of Interest

The authors stated that they have no conflicts of interest to disclose.

REFERENCES

Alsharabasy, A. M., El-Shemy, H. A., & Hussein, M. M. (2020). Phytochemical composition and antioxidant activity of watermelon rind extracts. *Journal of Applied Sciences, 20(5): 123–131*.

Bezerra, M. A., Santelli, R. E., Oliveira, E. P., Villar, L. S., & Escaleira, L. A. (2008). Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta, 76(5): 965– 977*.

Carley, M. K., Kamneva, Y. N. & Reminga, J. (2004). Response surface Methodology. Technical report written to Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research International- Carnegie Mellon University, School of Computer Science, Pittsburgh, PA, 15213.

Eguasa, O., Mbegbu, J. I. & Edionwe, E. (2019): On the Use of Nonparametric Regression

Model for Response Surface Methodology (RSM), *Benin Journal of Statistics, 2(5): 61-75*.

Eguasa, O. Edionwe, E & Mbegbu, J. I. (2022). Local Linear Regression and the problem of dimensionality: A remedial strategy via a new locally adaptive bandwidths selector, *Journal of Applied Statistics,50(6):1283–1309. DOI:10.1080/02664763.2022.2026895*

Eguasa, O. & Okungbowa, A. I. (2025). Response Surface Methodology-Based Optimization and Simulation of Bioactive Compound Extraction from Citrullus lanatus Rind, *Advances in Multidisciplinary and Scientific Research, 11(1):79-96. https://dx.doi.org/10.22624/AIMS/V11N1P6*.

Hastie, T., Tibshirani, R. & Friedman, T. A. (2009). The Elements of statistical learning, data mining, Inference and Prediction. *Springer Series in Statistics*. *Second Edition.*

Kamal, N.A.S.S., Abdullah, A., Ghazali, S.A.I.S.M., Nor, N.M., Dzulkifli, N.N. & Badrudin, S. (2023). Response Surface Methodology for Optimization of Citrullus lanatus Rind Extraction Conditions, Phytochemical Screening and its Antioxidant Activities. *Trend in Sciences, 20(10): 1-11*.

Kohler, M. A., Schindler, & Sperlich, S. A review and comparison of bandwidth selection methods for kernel regression. *International Statistical Review, 82(2014): 243–274*.

Luo, A., He, X., Zhou, S., & Fan, Y. (2019). Chemical composition and antioxidant activity of watermelon rind extract. *Food Chemistry, 285: 321–328.*

Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M. (2016). Response Surface Methodology:Process and Product Optimization Using Designed Experiments (4th ed.). Wiley.

Nair, A. T., Makwana, A. R. & Ahammed, M. M. (2014). The use of Response Surface Methodology for modelling and analysis of water and waste – water treatment processes: A Review. *Water Science and Technology, 69(3): 464 – 478.*

Najafi, S., Salmasnia, A. & Kazemzadeh, R.B. (2011). Optimization of robust design for multiple response problem. *Australian Journal of Basic and Applied Sciences, 5(9): 1566 – 1577*.

Oyeleke, G. O., Olagunju, J. A., & Ogunbanwo, S. T. (2021). Antimicrobial and antioxidant properties of watermelon rind extracts. *African Journal of Biotechnology, 20(2): 45–52*.

Parmar, H. S., Kar, A., & Jha, A. (2017). Watermelon rind: A potential source of natural antioxidants and bioactive compounds. *Journal of Food Biochemistry, 41(6)*: e12334.

Pickle, S. M. (2006). Semiparametric Techniques for Response Surface Methodology. Ph.D. Dissertation. Department of Statistics, *Virginia Polytechnic Institute and State University, Blacksburg, VA.*

Rivers, D. L. (2009). A Graphical Analysis of Simultaneously Choosing the Bandwidth and Mixing Parameter for Semiparametric Regression Techniques, M. Sc. Thesis and Dissertations, *Virginia Commonwealth University.*

Ruppert, D. & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics, 22: 1346-1370.*

Walker, E. L., Starnes, B. A., Birch, J. B. & Mays, J. E., (2002). Model Robust Calibration: Methods and Application to Electrically-Scanned Pressure Transducers, Technical Report, Dept. of Mathematics, Virginia Commonwealth University, Richmond, VA, USA.

Wan, W. & Birch, J. B. (2011). A semi-parametric technique for multi-response optimization. *Journal of Quality and Reliability Engineering International, Vol. 27: 47-59.*

Yeniay, O. (2014). Comparative study of algorithm for response surface optimization. *Journal of Mathematical and Computational Applications, 19(1): 93 – 10*