Article Citation Format
Eguasa, O. & Odigie, O.O. (2025): Statistical Techniques for Sewage Wastewater Treatment Prediction and Optimization via *Cucumis Melo* Coagulant. Journal of Digital Innovations & Contemporary Research in Science, Engineering & Technology. Vol. 13, No. 1. Pp 34-52 www.isteams.net/digitaljournal
dx.doi.org/10.22624/AIMS/DIGITAL/V13N1P3

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received: 11th January, 2025
Review Type: Blind Peer
Final Acceptance: 22nd March, 2025

Statistical Techniques for Sewage Wastewater Treatment Prediction and Optimization via *Cucumis Melo* Coagulant

[1]*Eguasa O. *[1] & [2]Odigie O. O.
[1]Department of Physical Sciences
Benson Idahosa University
Benin City, Edo State,
E-mails: oeguasa@biu.edu.ng
Phone: +2348028631580

ABSTRACT

The wastewater treatment process involves multiple stages; physical, chemical, and biological designed to remove contaminants such as organic matter, nutrients, pathogens, and heavy metals. Integrating artificial intelligence, statistical modeling, and optimization algorithms enhances prediction and control mechanisms, leading to more sustainable and cost-effective wastewater management solutions. Previous studies optimized multiple responses (turbidity, total suspended solids (TSS), biochemical oxygen demand (BOD), and chemical oxygen demand (COD)) using two operational factors (pH and coagulant dosage) at low and high levels (-1, +1) within a quadratic regression model (QM). However, these efforts were limited by factor selection, narrow variable bandwidths, and the chosen regression model, resulting in insufficient data fitting. This paper introduces an additional critical factor and expands the factor levels to five (-α, -1, 0, 1, +α) to enhance data fitting and optimize responses based on process requirements. An adaptive regression model is applied to improve goodness-of-fit statistics, increase dispersion from the zero-residual line, and enhance optimization outcomes. Comparative analysis, including 3D surface plots, demonstrates that the local linear regression model with variable bandwidths outperformed the QM model in design efficiency, delivering superior goodness-of-fit, reduced residual errors, and better optimization results, with an overall desirability of 100% compared to 95.50% for QM using Cucumis Melo as a natural coagulant.

Keywords: Local linear regression model, Cucumis Melo, Turbidity, Total Suspended Solids, Biochemical Oxygen Demand, Chemical oxygen demand.

1. INTRODUCTION

Sewage wastewater treatment is an essential process that ensures the safe disposal or reuse of wastewater from households, industries, and commercial establishments. With the rapid expansion of urbanization and industrialization, the volume of wastewater generated has grown substantially, posing significant risks to the environment and public health.

Effective sewage treatment is crucial to prevent water pollution, safeguard aquatic ecosystems, and promote public well-being (Yang *et al*., 2023). The wastewater treatment process consists of multiple stages; physical, chemical, and biological designed to remove contaminants such as organic matter, nutrients, pathogens, and heavy metals. The primary objective is to reduce pollutants to levels that comply with environmental regulations before safely discharging treated water into natural water bodies or reusing it for irrigation, industrial applications, or even potable purposes (Joaquin and Nirmala, (2019); Joaquin *et al.*, (2020)).

Unprocessed surface water presents risks to both human health and the environment. Therefore, choosing an effective wastewater treatment method is crucial to safeguarding public well-being and community safety while optimizing efficiency and minimizing costs (Sivaranjani and Rakshit (2017). Advancements in wastewater treatment technologies, including membrane filtration, advanced oxidation processes, and genetic optimization techniques, have significantly enhanced the efficiency and effectiveness of treatment systems. Integrating artificial intelligence, statistical modeling, and optimization algorithms further improves prediction and control mechanisms, resulting in more sustainable and cost-effective wastewater management solutions. With the growing demand for clean water and increasing environmental concerns, ongoing research and innovation in sewage wastewater treatment are essential for ensuring long-term water sustainability.

Response Surface Methodology (RSM) is suitable for optimizing the response variable y based on multiple explanatory variables. ($x_{i1}, x_{i2}, \ldots, x_{ik}$) which can be modeled as:

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

where $\boldsymbol{\varepsilon}_i$ 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 called a response surface Wan and Birch (2011).

The actual response function f is unknown and must be estimated. By applying Response Surface Methodology (RSM), we aim to determine the functional relationship between the response y and the associated explanatory variables ($x_{i1}, x_{i2}, \ldots, x_{ik}$).

The traditional approach to modeling the relationship between the kth explanatory variables and the ith response assumes that the fundamental functional form can be effectively expressed in a parametric manner. A parametric regression model may be more advantageous if the user can accurately determine an appropriate parametric form for the data.

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

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

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 homoscedastic property.

In the literature the Central Composite Design (CCD) and Box-Behnken Design (BBD) are two statistical experimental designs used in Response Surface Methodology (RSM). In this study, CCD was chosen to determine the optimal region for factors influencing the performance of the coagulation-flocculation process due to its efficiency.

The two key operating variables in this study were coagulant dosage and pH. The coded values for pH ($A = x_1$) and coagulant dosage ($B = x_2$) were defined at specific levels: pH (-1 ↔ 5.00) as low and (+1 ↔ 7.00) as high, while coagulant dosage (-1 ↔ 50.00) was set as low and (+1 ↔ 150.00) as high. The response factors analyzed included the percentage reduction in turbidity, TSS, BOD, and COD for *Cucumis Melo* (Joaquin *et al*., (2020)).

.

The existing literature, the second-order regression model (quadratic regression model) were used to fit the data for optimal settings of the two factors pH (A) and coagulant dosage (B). Hence, the quadratic regression model is given as:

The Quadratic model is given as:

$$y_i = \beta_0 + \beta_1 A + \beta_2 B + \beta_{11} A^2 + \beta_{22} B^2 + \beta_{12} AB \qquad (3)$$

which can be written as:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \beta_{ii} x_i^2 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij} x_i x_j + \varepsilon \qquad (4)$$

For which $A = x_1$, $B = x_2$ are the explanatory variables; β_0 is a constant coefficient; the varying coefficients β_1, β_2 and β_{11}, β_{22} are the coefficients of linear, quadratic and interaction terms respectively (Joaquin *et al*., (2020))

The usual method for estimating the parameter vector in Equation (2) is usually based on the Method of OLS. The parameter vector estimates $\widehat{\boldsymbol{\beta}}$ in (2) is given as:

$$\widehat{\boldsymbol{\beta}}^{(OLS)} = \left(\boldsymbol{X}'^{(OLS)} \boldsymbol{X}^{(OLS)}\right)^{-1} \boldsymbol{X}'^{(OLS)} \boldsymbol{y},\ \boldsymbol{X} = \boldsymbol{X}^{(OLS)} \qquad (5)$$

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

$$\widehat{\boldsymbol{y}}_i^{(OLS)} = \boldsymbol{x}_i'^{(OLS)} \widehat{\boldsymbol{\beta}}^{(OLS)} = \boldsymbol{x}_i'^{(OLS)} \left(\boldsymbol{X}'^{(OLS)} \boldsymbol{X}^{(OLS)}\right)^{-1} \boldsymbol{X}'^{(OLS)} \boldsymbol{y}_. \quad i = 1,2, \dots, n \qquad (6)$$

where $\boldsymbol{x}_i'^{(OLS)}$ is the i^{th} row of matrix $\boldsymbol{X}^{(OLS)}$, $n \times (k+1)$ vector, (Carley *et al.,* (2004); Ramakrishnan and Arumugam (2011)).

2. MATERIALS AND METHODS

As specified in the literature, the two operating factors are pH $(A = x_1)$ and coagulant dosage $(B = x_2)$ with the multi-response variables; Turbidity, Total Suspended Solids (TSS), Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) reduction. The goal is to obtain an optimum setting of the factors that would simultaneously maximize the multi-response factors (Joaquin *et al.*, (2020)). The factors used by Joaquin *et al*. (2020) were insufficient for maximizing the optimal settings of the multi-response problem. In this study, an essential factor, temperature (°C), was introduced to enhance the sequence of factors, ensuring a more comprehensive maximization of the multi-response problem.

The rationale behind the local linear regression model lies in its flexibility, allowing it to effectively address boundary bias issues without being constrained to a user-specified data form (Gramato and Calado (2014); Eguasa and Eguasa (2022); Eguasa and Eguasa (2023)). In the literature, the second-order regression model was used. However, in this paper, we apply an adaptive bandwidth for local linear regression (LLR_{AB}) to smooth the data based on location.

The Local Linear Regression (LLR) Model

When a researcher has limited information or only partial knowledge of a model's functional form, a nonparametric regression model serves as the most suitable alternative. In such scenarios, the Local Linear Regression (LLR) model is particularly relevant. LLR, a weighted adaptation of the least squares method derived from first-order Local Polynomial Regression, offers a key advantage over kernel regression by effectively reducing bias, especially at the boundaries of the explanatory variables (Ruppert and Wand (1994); Walker *et al*., (2002); Eguasa and Eguasa (2023).

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

$$\hat{\boldsymbol{y}}_i^{(LLR)} = \boldsymbol{x}_i^{\prime(LLR)}(\boldsymbol{X}^{\prime(LLR)}\boldsymbol{W}_i\boldsymbol{X}^{(LLR)})^{-1}\boldsymbol{X}^{\prime(LLR)}\boldsymbol{W}_i\boldsymbol{y} = \boldsymbol{H}_i^{(LLR)}\boldsymbol{y}_i \qquad (7)$$

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

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

We define *W*, an $n \times n$ diagonal matrix of kernel weights for estimating the response as:

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

(Eguasa and Eguasa (2023)).

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

We can rewrite Equation (7) in terms of hat matrix as:

$$\hat{\boldsymbol{y}}^{(LLR)} = \boldsymbol{H}^{(LLR)}\boldsymbol{y}, \qquad (10)$$

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

$$\boldsymbol{H}^{(LLR)} = \begin{bmatrix} \boldsymbol{x}_1'^{(LLR)}\left(\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_1\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_1 \\ \boldsymbol{x}_2'^{(LLR)}\left(\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_2\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_2 \\ \vdots \\ \boldsymbol{x}_n'^{(LLR)}\left(\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_n\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_n \end{bmatrix} \qquad (11)$$

The limitation of the LLR model is its high bias in regions with curvature, as its model matrix setup does not include quadratic terms (He *et al*., (2009)).

Adaptive Bandwidths

We introduced data-driven locally adaptive bandwidths as given in Eguasa *et al.* (2022):

$$b_{ij} = b \;\; i = 1,2, \;\; ..., n; j = 1,2, ..., k \qquad (12)$$

where, $0 < b_{ij} \leq 1$.

3. EXPERIMENTAL DESIGN

In a Response Surface Methodology (RSM) application, multiple factors are typically involved, making the selection of appropriate levels for explanatory variables critical, as it directly affects the model's accuracy. The Experimental Design phase facilitates the creation of an effective design that accurately estimates the relationship between the response and one or more factors, with the Central Composite Design (CCD) being employed.

In CCD, the number of experimental runs is determined using the formula $2^k + 2k + k_c,$ All factors are evaluated at five levels: $(-\alpha, -1, 0, 1, \alpha)$, where 2^k represents the full factorial design, $2k$ denotes the axial (star) points positioned at a distance $\alpha = \sqrt[4]{2^k}$ from the center point, and k_c is the number of center points. In this study, $k = 2$ (the number of factors used), and $k_c = 5$, resulting in a total of 13 experimental runs for data collection (Eguasa and Eguasa, 2023).

Accordingly, the factors and their corresponding coded levels, as presented in the literature, are shown in Table 1 below:

Table 1: Coded stages and range for the design of experiments (Joaquin *et al*., (2020))

Variables	Factors or Input parameters	-1(Low)	+1(High)
pH	$A = x_1$	5	7
coagulant dosage (mg/L)	$B = x_2$	50	150

Table 2: Experimental design (CCD) for Turbidity, TSS, BOD and COD using *Cucumis Melo* coagulant (Joaquin *et al*., (2020))

Exptal Run	pH x_1	Dosage (mgL^{-1}) x_2	Turbidity y_1	TSS $y_2 Observed$	BOD $y_3 Observed$	COD $y_4 Observed$
1	6	100	37.6	87.7	95.5	71.4
2	5	100	37.6	92.2	97.7	71.4
3	6	50	96.6	86.7	76.1	10
4	6	100	37.6	87.7	95.5	71.4
5	6	150	35.6	76.7	95.5	78.6
6	7	150	40	77.2	88.1	71.4
7	5	50	95.5	93.3	86.4	14.3
8	7	100	34.1	87.7	85.2	71.4
9	6	100	37.6	87.7	95.5	71.4
10	7	50	92.8	85.5	50	34.3
11	6	100	37.6	87.7	95.5	71.4
12	6	100	37.6	87.7	95.5	71.4
13	5	150	36.5	79.5	96.6	81.4

In Table 3, the inclusion of temperature (°C) as a factor, along with the star point (±α), significantly enhances the study's design. This addition improves the sequencing of factors, providing a more comprehensive approach to optimizing the multi-response problem.

Table 3: Coded stages and range for the design of experiments for *Cucumis Melo*

Variables	Factors	-1.682(-α)	-1(Low)	0(Medium)	+1(High)	+1.682(α)
pH	x_1	3	5	0	7	9
Temperature (°C)	x_2	20.5	21	21.5	22	22.5
Coagulant dosage (mg/L)	x_3	25	50	75	150	175

Table 4, explains the choice of CCD in the addition of axial point to the coded factors that can capture curvature and maintain rotatability in the data $\alpha = \pm\sqrt[4]{2^k}$, where k= the number of factors used in the design. Therefore, $\alpha = \pm 1.682$ (Eguasa *et al*., (2022))

Table 4: Experimental design (CCD) for TOC, TN and TSS removal

Exptal Run	pH x_1	Temp. (°C) x_2	Dosage (mgL^{-1}) x_3	Turbidity y_1	TSS $y_2 Observed$	BOD $y_3 Observed$	COD $y_4 Observed$
1	-1	-1	-1	53.2	91.1	95.4	28.6
2	1	-1	-1	30.1	88.9	88.6	11.4
3	-1	1	-1	59.5	87.7	39.8	88.6
4	1	1	-1	53.2	91.1	95.4	28.6
5	-1	-1	1	46.7	89.4	96	50
6	1	-1	1	69.1	77.7	93.3	69.9
7	-1	1	1	41.8	64.4	44	22.9
8	1	1	1	88.2	87.7	94.3	57.1
9	-1.682	0	0	53.2	91.1	95.4	28.6
10	1.682	0	0	68.1	87.7	65.9	71.4
11	0	-1.682	0	53.2	91.1	95.4	28.6
12	0	1.682	0	53.2	91.1	95.4	28.6
13	0	0	-1.682	32.1	88.4	92	48.6

Data transformation to RSM data in the interval of zero and one

The operating factor values are coded within the range of 0 to 1, and the data obtained from the Central Composite Design (CCD) is subsequently transformed using a mathematical relation:

$$x_{NEW} = \frac{Min(x_{OLD}) - x_0}{(Min(x_{OLD}) - Max(x_{OLD}))} \qquad (13)$$

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}, $\mathrm{Min}(x_{OLD})$ 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 4 can be transformed to explanatory variables in Table 5 using Equation (13).

Table 5: Experimental design for Turbidity, TSS, BOD and COD removal using *Cucumis Melo* coagulant

Experimental Run	pH x_1	Temp. (°C) x_2	Coagulant Dosage x_3	Turbidity y_1	TSS $y_2 Observed$	BOD $y_3 Observed$	COD $y_4 Observed$
1	0.2030	0.2030	0.2030	53.2	91.1	95.4	28.6
2	0.7970	0.2030	0.2030	30.1	88.9	88.6	11.4
3	0.2030	0.7970	0.2030	59.5	87.7	39.8	88.6
4	0.7970	0.7970	0.2030	53.2	91.1	95.4	28.6
5	0.2030	0.2030	0.7970	46.7	89.4	96	50
6	0.7970	0.2030	0.7970	69.1	77.7	93.3	69.9
7	0.2030	0.7970	0.7970	41.8	64.4	44	22.9
8	0.7970	0.7970	0.7970	88.2	87.7	94.3	57.1
9	0.0000	0.5000	0.5000	53.2	91.1	95.4	28.6
10	1.0000	0.5000	0.5000	68.1	87.7	65.9	71.4
11	0.5000	0.0000	0.5000	53.2	91.1	95.4	28.6
12	0.5000	1.0000	0.5000	53.2	91.1	95.4	28.6
13	0.5000	0.5000	0.0000	32.1	88.4	92	48.6

Multi-Response Optimization Problem

This entails the simultaneous optimization of two or more responses alongside the associated factors ($x_{i1}, x_{i2}, \ldots, x_{ik}$). The optimization criteria and the desired objectives for the multi-response problem based on experimental result, as presented in (Joaquin *et al*., (2020)), are shown in Table 6 below.

Table 6: Experimental results for two factors and responses using *Cucumis Melo* Coagulant (Joaquin *et al*., (2020))

Criteria	*Goal*	*Lower Limit*	*Upper Limit*
pH	In the range	1	5
Coagulant Dosage (mgL^{-1})	In the range	50	150
Turbidity Reduction (%)	Maximize	-	66.5
TSS Reduction (%)	Maximize	-	92.8
BOD Reduction (%)	Maximize	-	92.1
COD Reduction (%)	Maximize	-	42.9

Based on the type of response, the desirability function transforms the estimated response, $\hat{y}_p(\boldsymbol{x})$ to different individual scalar measure, $d_p\left(\hat{y}_p(\boldsymbol{x})\right)$ namely:

For larger-the-better (LTB) response $d_p\left(\hat{y}_p(\boldsymbol{x})\right)$ is given as:

$$d_p\left(\hat{y}_p(\boldsymbol{x})\right) = \begin{cases} 0, & \hat{y}_p(\boldsymbol{x}) < L \\ \left\{\frac{\hat{y}_p(x)-L}{T-L}\right\}^{t_1}, & L \le \hat{y}_p(\boldsymbol{x}) \le T, \\ 1, & \hat{y}_p(\boldsymbol{x}) > T, \end{cases} \qquad s.t\, x \epsilon\, \varphi\,, \tag{14}$$

where $T$and L are the maximum acceptable value and lower limit, respectively, of the p^{th} response.where ρ is the target value of the p^{th} response. However, for RSM data, the parameters values of t_1 and $t_2$are weights taken to be 1 for linearity (Wan (2007); Castillo (2007); He *et al*., (2009; 2012)).

The overall desirability function

The objective of the desirability function is to maximize the overall desirability D, which is calculated as the geometric mean of the individual desirability functions. The overall desirability D is expressed as:

$$D = \sqrt[r]{(d_1(\hat{y}_1(\boldsymbol{x})).\, d_2(\hat{y}_2(\boldsymbol{x}))\dots\, d_r(\hat{y}_r(\boldsymbol{x}))} \tag{15}$$

where $r = 4$ is the number of response variables, $d_1\hat{y}_1(\boldsymbol{x})$, $d_2\hat{y}_2(\boldsymbol{x})$,..., $d_r\hat{y}_r(\boldsymbol{x})$ are the individual desirability (He *et al*., (2012)). The desirability function $d_m\left(\hat{y}_m(\boldsymbol{x})\right), m = 1,2,\dots,r$ allocate values between 0 and 1 centered on the process requirements such that the most undesirable and desirable values are $d_r\left(\hat{y}_r(\boldsymbol{x})\right) = 0$ and $d_r\left(\hat{y}_r(\boldsymbol{x})\right) = 1$ respectively.

Explanation of Statistical terms

The goodness-of-fit statistics considered in this study are; $PRESS, PRESS^{**}, SSE, MSE, R^2$ and R^2_{adj}. The Prediction Error Sum of Squares (PRESS) statistic, which is often small, serves as a version of the cross-validation criterion given as:

$$PRESS = \sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}(.)\right)^2 \tag{16}$$

$PRESS^{**}$ criterion was derived as a substitute to the $PRESS$, with the tendency to overfit the data. The form of the $PRESS^{**}$ criterion for selecting the bandwidths is given as:

$$PRESS^{**}(\Omega) = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}(.)\right)^2}{n - trace\left(H^{(.)}(\Omega)\right) + (n-k-1)\frac{SSE_{max} - SSE_{\Omega}}{SSE_{max}}} \tag{17}$$

where SSE_{max} is the maximum Sum of Squared Errors obtained as the b_{ij} tend to infinity, SSE_{Ω} is the sum of squared errors associated with a set of bandwidths b_{ij}, $trace\left(H^{(.)}\Omega\right)$ is the trace of the Hat matrix and $\hat{y}_{i,-i}$ is the leave-one-out cross-validation estimated value of y_i with the i^{th} observation left out (Eguasa *et al*., (2022); Eguasa and Eguasa (2023)).

For data emanating from RSM, the vector of optimal bandwidths $\Omega \in [b *_{ij}]$ is derived based on the minimization of the Penalized Prediction Error Sum of Squares, $PRESS^{**}$ (Eguasa *et al*., (2022)). The Sum of Squares Error (SSE) and the Mean Square Error (MSE) evaluate how effectively each regression model estimates f, as defined in Equation (1). The criterion for selecting estimates is to minimize the sum of squared residuals, also referred to as the Sum of Squares Error (SSE). Thus,

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{18}$$

The mean square error, MSE is given as:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k} \tag{19}$$

Here, n represents the sample size, k denotes the number of explanatory variables used in the study, and the degrees of freedom $(n - k)$ significantly influence the value of the Mean Squared Error (MSE). A smaller degree of freedom may result in a larger MSE, while a larger degree of freedom can lead to a smaller MSE (Eguasa and Eguasa, 2022). In Response Surface Methodology (RSM), the coefficient of determination (R^2) and the adjusted coefficient of determination (R^2_{adj}) are used to evaluate the adequacy of a model in representing a real system. R^2 measures the proportion of the observed variability in the experimental data that the model can explain. However, since R^2 tends to increase with the addition of explanatory variables, R^2_{adj} adjusts for this limitation by considering the number of explanatory variables in the model. The expressions for R^2 and R^2_{adj} are provided in Equations (20) and (21), respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{20}$$

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-k)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)} \tag{21}$$

The values of R^2 and R^2_{adj} statistics should be close to unity for statistical significance of the model considered. Thus, k is the number of explanatory variables used in the model, n is the number of sample size, y_i is the raw response, $\hat{y}_i$ is the estimated response and $\bar{y}$ is the mean response.

4. RESULTS AND DISCUSSION

The results in Table 7 clearly demonstrate that the LLR_{AB} method outperformed the second-order Quadratic Model (QM) proposed by Joaquin et al. (2020) in terms of key responses, including turbidity reduction (%), TSS reduction (%), BOD reduction (%), and COD reduction (%). The LLR_{AB} method achieved superior performance statistics in 16 cells, compared to QM, in addressing the multi-response problem.

Cells in bold indicate better performance compared to non-bold cells. Lower values of PRESS, SSE, and MSE signify improved statistical performance, while higher values ofR^2 and R^2_{adj} also indicate better model performance.

Table 7: Model Goodness-of-fits statistics for QM and LLR_{AB} *Cucumis Melo*

Response	*Model*	DF	$PRESS^{**}$	$PRESS$	SSE	MSE	R^2(%)	R^2_{Adj}(%)
y_1	QM	5	-	-	2598.69	519.74	90.14	83.10
	LLR_{AB}	2.3492	57.8794	648.6606	**27.3606**	**11.6468**	**99.05**	**95.15**
y_2	QM	5	-	-	638.48	127.70	92.12	86.49
	LLR_{AB}	0.0111	75.2069	677.6757	**0.0128**	**1.1559**	**100**	**98.00**
y_3	QM	5	-	-	4760.10	952.02	96.11	93.33
	LLR_{AB}	0.1244	800.92	7.3077e+003	**0.1707**	**1.3719**	**100**	**99.67**
y_4	QM	5	-	-	5104.94	1020.99	80.69	66.89
	LLR_{AB}	0.0225	4.8278e+005	4.3557e+006	**0.2292**	**10.1796**	**100**	**98.07**

Table 8, represents the predicted response for Turbidity, TSS, BOD and COD reduction using *Cucumis Melo* as a natural coagulant using LLR_{AB}.

Table 8: Predicted response for Turbidity, TSS, BOD and COD reduction using *Cucumis Melo* coagulant

Exptal Run	Turbidity y_1	Turbidity $\hat{y}_1$	TSS y_2	TSS $\hat{y}_2$	BOD y_3	BOD $\hat{y}_3$	COD y_4	COD $\hat{y}_4$
1	53.2	54.4558	91.1	91.1000	95.4	95.4000	28.6	28.6000
2	30.1	32.3426	88.9	88.9000	88.6	88.6016	11.4	11.4000
3	59.5	56.7592	87.7	87.7000	39.8	39.8000	88.6	88.6000
4	53.2	52.2054	91.1	91.1000	95.4	95.3980	28.6	29.0544
5	46.7	45.3067	89.4	89.4000	96	96.0007	50	50.0000
6	69.1	69.4846	77.7	77.7001	93.3	93.0472	69.9	69.8961
7	41.8	44.2503	64.4	64.5053	44	43.9991	22.9	22.9421
8	88.2	87.9455	87.7	87.6588	94.3	94.6268	57.1	56.9553
9	53.2	53.2088	91.1	91.0999	95.4	95.4000	28.6	28.6000

Exptal Run	Turbidity y_1	Turbidity $\hat{y}_1$	TSS y_2	TSS $\hat{y}_2$	BOD y_3	BOD $\hat{y}_3$	COD y_4	COD $\hat{y}_4$
10	68.1	68.0784	87.7	87.7002	65.9	65.9001	71.4	71.4000
11	53.2	52.2310	91.1	91.1001	95.4	95.4000	28.6	28.6000
12	53.2	54.0469	91.1	91.0994	95.4	95.4000	28.6	28.6045
13	32.1	33.6615	88.4	88.4000	92	92.0000	48.6	48.6000

Table 9: Residual response for Turbidity, TSS, BOD and COD using *Cucumis Melo* as a natural coagulant

Experimental Run	Turbidity r_1	TSS r_2	BOD r_3	COD r_4
1	-1.2558	0.0000	0.0000	-0.0000
2	-2.2426	0.0000	-0.0016	-0.0000
3	2.7408	0.0000	-0.0000	0.0000
4	0.9946	0.0000	0.0020	-0.4544
5	1.3933	0.0000	-0.0007	0.0000
6	-0.3846	-0.0001	0.2528	0.0039
7	-2.4503	-0.1053	0.0009	-0.0421
8	0.2545	0.0412	-0.3268	0.1447
9	-0.0088	0.0001	-0.0000	0.0000
10	0.0216	-0.0002	-0.0001	0.0000
11	0.9690	-0.0001	0.0000	0.0000
12	-0.8469	0.0006	-0.0000	-0.0045
13	-1.5615	0.0000	0.0000	0.0000

In Figure 1, the deep blue straight line is the zero residual plot, green line is the LLR_{AB} plot respectively.

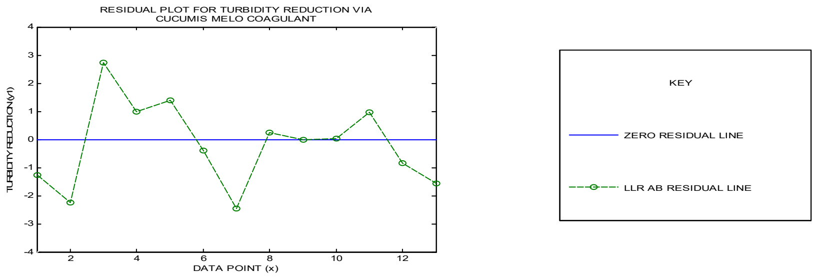

Figure 1: Model Residuals for y_1 (Turbidity reduction (%)) response plotted against the data points for LLR_{AB} model.

In Figure 2, the deep blue straight line is the zero residual plot, green line is the LLR_{AB} plot respectively.

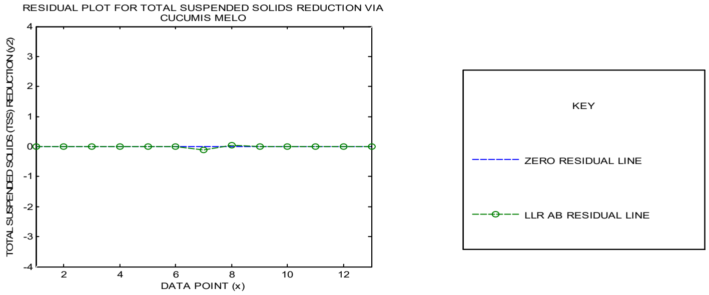

Figure 2: Model Residuals for y_2 (TSS reduction (%)) response plotted against the data points for LLR_{AB} model.

In Figure 3, the deep blue straight line is the zero residual plot, green line is the LLR_{AB} plot respectively.

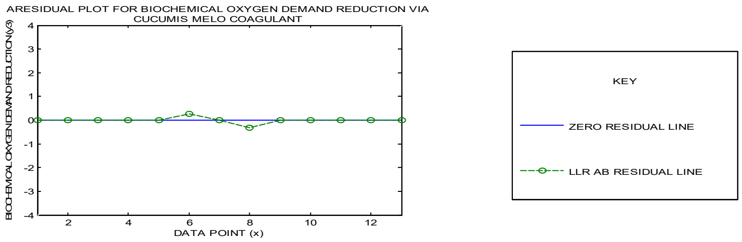

Figure 3: Model Residuals for y_3 (BOD reduction (%)) response plotted against the data points for LLR_{AB} model.

In Figure 4, the deep blue straight line is the zero residual plot, green line is the LLR_{AB} plot respectively.

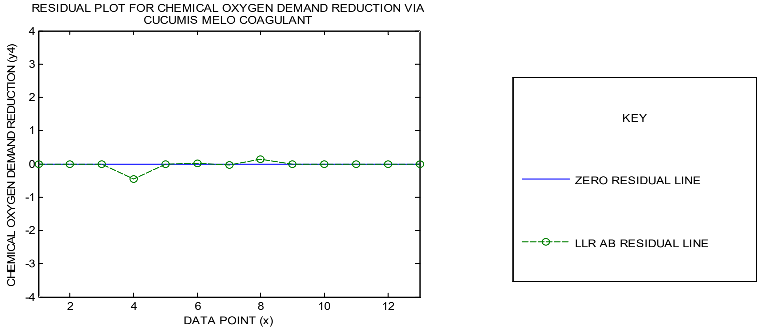

Figure 4: Model Residuals for y_4 (COD reduction (%)) response plotted against the data points for

LLR_{AB} model.

Table 10: Model optimal (maximize) solution based on the multi-response desirability function using *Cucumis Melo* as a natural coagulant

Model	x_1	x_2	x_3	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	$d_1(\hat{y}_1)$	$d_2(\hat{y}_2)$	$d_3(\hat{y}_3)$	$d_4(\hat{y}_3)$	$D(\%)$
Exptal values	5 - 7	-	50-150	78.2	87.7	80.1	64.3	1	1	1	1	100
QM	7	-	76.7	77.0	89.6	86.9	56.9	0.9847	1	1	0.8849	95.50
LLR_{AB}	0.9732	0.8580	0.2601	95.6	94.0	95.4	97.3	1	1	1	1	100

Table 10 shows that LLR_{AB} provide enhanced multi-response optimization for maximizing the reduction of turbidity, TSS, BOD, and COD in sewage wastewater treatment using the natural coagulant *Cucumis Melo*. Compared to QM achieve higher overall desirability for the respective factors: $x_1 = pH\ (\%),\ x_2 = \text{Temperature}(\%)\ \text{and}\ x_3 = \text{Coagulant Dosage}\ (\%)$. Obviously, LLR_{AB} gave a better process condition with 100% overall desirability and with operating factors $x_1 = \mathbf{0.9732} = \mathbf{6.81},\ x_2 = \mathbf{0.8580} = \mathbf{19.0}, x_3 = \mathbf{0.2601} = \mathbf{112.10}$ with the best choice based sewage wastewater treatment via *Cucumis Melo* as a natural coagulant.

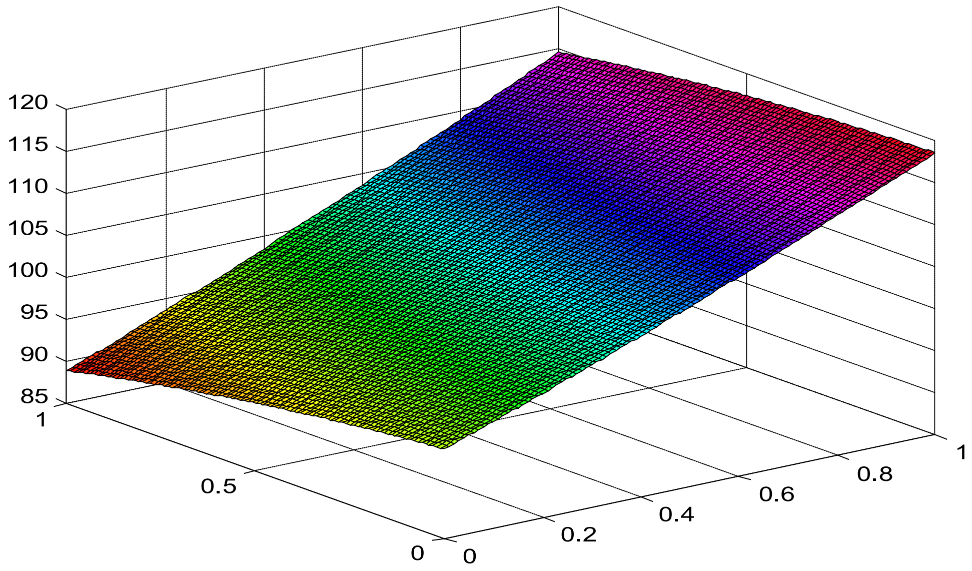

Figure 5: LLR_{AB} surface plot for maximum Turbidity of 90% showing the interactive effect of *pH* and Temperature.

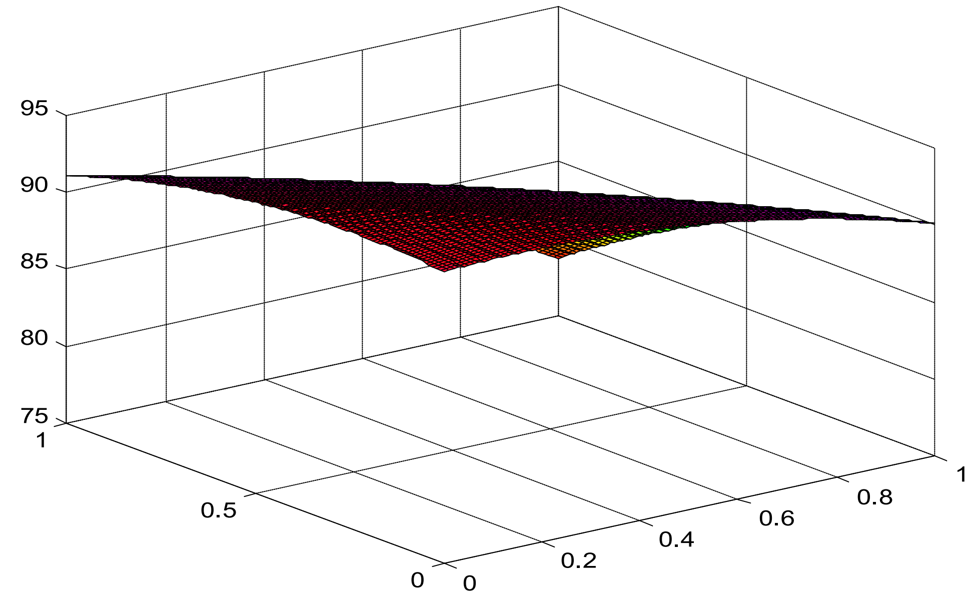

Figure 6: LLR_{AB} Surface plot for maximum Total Suspended Solids (TSS) of 91.0% showing the interactive effect of *pH* and Temperature.

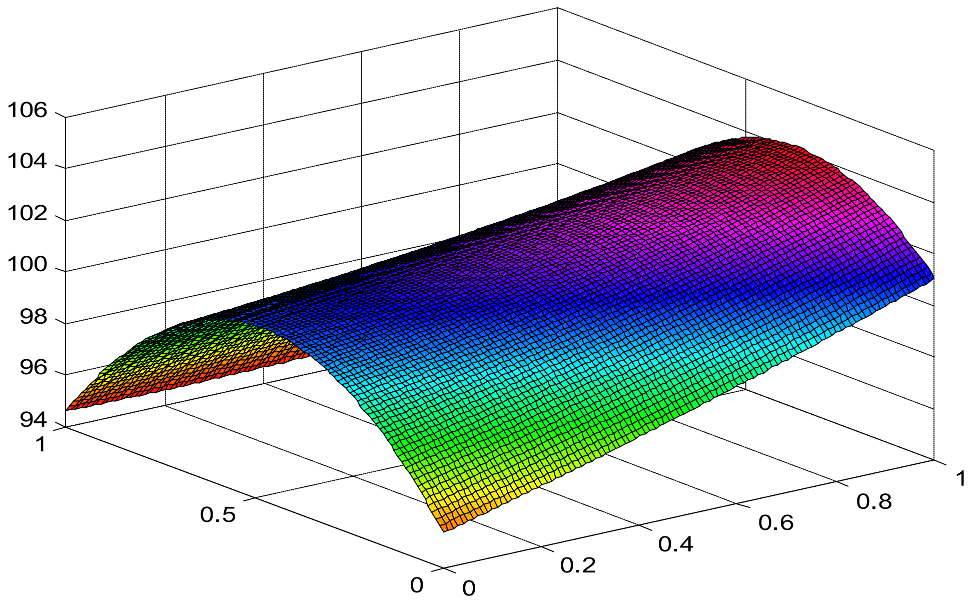

Figure 7: LLR_{AB} surface plot for maximum Biochemical Oxygen Demand (BOD) of 95.4% showing the interactive effect of *pH* and Temperature.

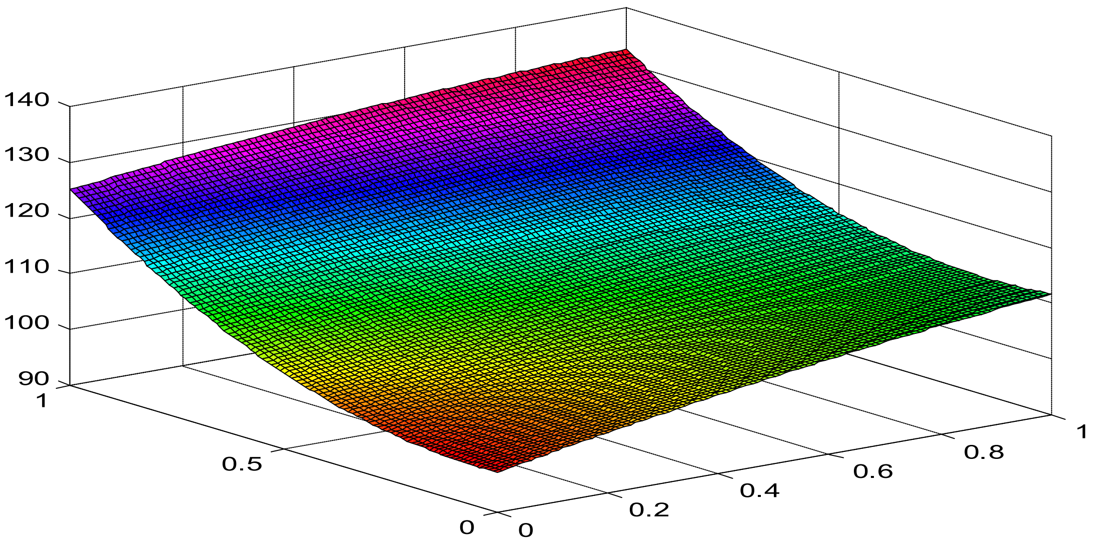

Figure 8: LLR_{AB} Surface plot for maximum Chemical Oxygen Demand (COD) of 125% showing the interactive effect of *pH* and Temperature.

The maximization of the overall desirability functions for *Cucumis Melo* are given in Figure 5, 6, 7 and 8 respectively shows the shapes of the different colour variation in the surface plots representing total desirability of the optimization criteria for Turbidity (mgL^{-1}), TSS (mgL^{-1}), BOD (mgL^{-1}) and COD (mgL^{-1}) reduction respectively. Therefore, in Figure 5, the individual desirability function for LLR_{AB} is 100% with optimal response 95.6% as against the optimization result for QM with individual desirability function of 98.5% with optimal response 77.0% for Turbidity(mgL^{-1}). Whereas, in Figure 6 the individual desirability function for LLR_{AB} is 100% with optimal response 94% as against the optimization result for QM with individual desirability function of 100% with optimal response 89.6% for TSS reduction (mgL^{-1}), while in Figure 7, the individual desirability function for LLR_{AB} is 100% with optimal response 95.4% as against the optimization result for QM with respective individual desirability function of 100% with optimal response 86.9% for BOD reduction (%).Whereas, in Figure 8, the individual desirability function for LLR_{AB} is 100% with optimal response 97.3% as against the optimization result for QM with respective individual desirability function of 88.5% with optimal response 56.9% for COD reduction (%).

5. CONCLUSION

In this paper, we examined the results of the regression models *QM* and LLR_{AB} using the CCD approach, considering three operating factors at five levels and four responses within the RSM data. The main findings indicate that the kernel function, which determines the shape of kernel weights through variable bandwidths, employs x_i as the explanatory variable for location i, j the number of explanatory variables in the design, x_0 as a target point (dummy variable), and b_{ij} as the variable bandwidth or smoothing parameter controlling the smoothness of the estimated function, as defined in Equation (1). This highlights the superiority of the LLR_{AB} model over QM, as it effectively corrects bias at the boundaries and with unevenly spaced explanatory variables. Statistical analysis of the experimental data using the LLR_{AB} model demonstrates its ability to enhance operational characteristics for optimizing multi-response problems (overall desirability of 100% with operating factors $x_1 = \mathbf{0.9732} \ = \mathbf{6.81},\ x_2 = \mathbf{0.8580} = \mathbf{19.0}, x_3 = \mathbf{0.2601} = \mathbf{112.10}$) related to Turbidity, TSS, BOD, and COD reduction. Performance statistics, residual plots, and surface plots further emphasize the significant improvements achieved by LLR_{AB} compared to the existing QM model.

ACKNOWLEDGMENTS

I am deeply grateful to God Almighty for granting us life, strength, and wisdom, making it possible to complete this task.

REFERENCES

1. Castillo, D. E. (2007). Process Optimization: A statistical Method. Springer International Series in Operations Research and Management Science: New York.
2. Eguasa, O., Mbegbu, J. I. & Edionwe, E. (2019). On the Use of Nonparametric Regression Model for Response Surface Methodology (RSM), *Benin Journal of Statistics,* 2:61-75.
3. 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
4. Eguasa, O. & Eguasa, M. E. (2023). An Adaptive Local Linear Regression Model Applied in the Multi-Response Optimization of Oil-in-Water Emulsion for Response Surface Methodology", *Indian Journal of Science and Research,* 3(3): 120-128
5. Eguasa, O. & Eguasa, M. E. (2022) "Optimizing Biodiesel Yield via an Adaptive Local Linear Regression Model with Application to Response Surface Methodology" *Nigerian Journal of Scientific Research, Vol. 21, No.2, pg. 425 – 435*
6. Gramato, D. & Calado, V (2014). The use and importance of design of experiments (DOE) in Process modelling in food science and technology. Mathematical and Statistical Methods in Food Science and Technology, John Wiley and Sons, Ltd. First Edition.
7. He, Z., Wang, J., Oh, J. & Park, S. H. (2009). Robust optimization for multiple responses using response surface methodology, *Applied Stochastic Models in Business and Industry,* 26:157 – 171.
8. He, Z., Zhu, P. F. & Park, S. H. (2012). A robust desirability function for multi-response surface optimization. European *Journal of Operational Research,* 221:241-247.
9. Joaquin, A. A. & Nirmala, G. (2019). Statistical modeling and process optimization of coagulation-flocculation for treatment of municipal wastewater. *Desalination and Water Treatment*, 157: 90-99.
10. Joaquin, A. A., Nirmala, G. & Kanakasabai, P. (2020). Response surface analysis for sewage wastewater treatment using natural coagulants. *Polish Journal of Environmental Studies*, 30(2): 1215-1225.
11. Ramakrishnan, R. & Arumugam, R. (2011). Application to Response Surface Methodology (RSM) for Optimization of Operating Parameters and Performance Evaluation of Cooling Tower Cold water Temperature. *An International Journal of Optimization and Control*: *Theories and Applications*, 1: 39 – 50.
12. Ruppert, D. & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics,* 22:1346-1370.
13. Sivaranjani, S. & Rakshit, A. (2017). A study on removal efficiency of blended coagulants on different types of wastewater. *Nature Environment and Pollution Technology*, 16(1):107-
14. 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.
15. Wan, W. (2007), Semi-parametric techniques for multi-response optimization. Ph.D Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

16. Wan, W. & Birch, J. B. (2011). A semi-parametric technique for multi-response optimization. *Journal of Quality and Reliability Engineering International*, 27: 47-59.
17. Yateh, M., Lartey-Young, G., Li, F., Li, M. & Tang, Y. (2023). Application of response surface methodology to optimize coagulation treatment process of urban Drinking water using polyaluminium Chloride. *Water*, 15(5), 853; https://doi.org/10.3390/w15050853