Nigerian Journal of Scientific Research (NJSR)
ISSN-0794-0378

A GLOBAL BANDWIDTH SMOOTHING APPROACH IN THE MULTI-RESPONSE OPTIMIZATION FOR OIL-IN-WATER EMULSION

EGUASA, O.[1]* AND ODIGIE, O.O.[2]
[1]Department of Physical Sciences, Benson Idahosa University, Benin City, Nigeria
[2]Department of Biological Sciences, Benson Idahosa University, Benin City, Nigeria

ABSTRACT

Response surface methodology (RSM) data was examined for the ultrafiltration of oil-in-water emulsion procedure via the use of three-phase experimental design methods namely: design of experiment (DOE), modeling phase of the estimated regression model, and the optimization phase. A quadratic regression model (QM) was employed to fit the data, which performed well in areas where the data exhibited a quadratic trend or curvature. Nevertheless, the QM showed limitations near the boundaries with linear trend of the data due to a boundary bias problem. To address this issue, ordinary least squares (OLS) regression and a global or fixed bandwidth (FB) applied to local linear regression (LLR_{FB}) model was considered via an experimental design technique such as the Central Composite Design (CCD). The model aimed to improve goodness-of-fit statistics, the zero-residual line, and optimization results. The comparative analysis, including 3D surface plots, the outcome showed that (LLR_{FB}) performed better than both OLS and QM in terms of the design technique. The (LLR_{FB}) model demonstrated superior goodness-of-fit statistics, minimized residual errors, and better optimization results, achieving an overall desirability of 61.31%, as compared to 56% for OLS and 55% for QM respectively.

Keywords: Global local linear regression model, Chemical oxygen demand (COD), Permeate flux (PF), Trans-membrane pressure, Turbidity removal, Velocity.
***Correspondence**: oeguasa@biu.edu.ng, +2348028631580

INTRODUCTION

Oil in water is a byproduct present in the wastewater generated by various industrial facilities, including petrochemical plants, natural gas processing facilities, nuclear power plants, food processing plants, and agricultural operations. This wastewater contains a complex mixture of water, oil, and additives such as emulsifiers, corrosion inhibitors, anti-foaming agents, and biocides. To either reprocess this wastewater or release it into natural waterways, treatment is necessary [1, 9, 19, 20].

According to [1, 10, 15, 24], several published articles discussed Response Surface Methodology (RSM), which was originally developed by [2]. Subsequent significant contributions to the development of RSM include works by [3, 4, 16]. Additionally, a later paper by [13] explores the practical applications of RSM in the chemical and processing industries. Moreover, the applications of RSM extend across various fields of study [1, 10, 15].

The goal of this paper is to apply a fixed bandwidth via the LLR_{FB} as compared with QM and OLS in order to improve the operating features that would simultaneously optimize the multi-response problems through RSM data for permeate flux (PF) (L/m^{2}h), turbidity removal (%) and chemical oxygen demand (COD) removal (%) with the data from literature [9, 20]. The research gap is addressed by LLR_{FB} which is not affected by the user specified form unlike OLS and boundary biased problem as in QM. The supremacy of LLR_{FB} over OLS and QM is that the hat matrix of LLR_{FB} incorporates Gaussian kernel weight with fixed bandwidth for adequate smoothing of the data. According to [10, 11, 15], defined RSM as statistical technique applied by engineers and industrial statistician for empirical model structure, with the purpose of optimizing the response variables which are influenced by numerous explanatory variables or design variables with a limited number of experiments [7, 8, 9, 21].

RSM is appropriate for optimizing the response variable $\boldsymbol{y}$ as a function of more than one explanatory variable $(x_{i1}, x_{i2}, \ldots, x_{ik})$ 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 called a response surface [23].

The actual response function f is unknown which needs to be estimated. Applying RSM, we try to find the functional relationship between the responses y and related explanatory variables $(x_{i1}, x_{i2}, \ldots, x_{ik})$.

The conventional method suited in modeling the relationship between the *kth* explanatory variables and the *ith* response is to assume that the basic functional form can competently be expressed parametrically. The parametric regression model may be superior, if the user can identify adequately a parametric form for the data.

Hence, 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 homoscedastic property.

The Quadratic model is given as:

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

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 [9, 20].

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:

$$\widehat{\boldsymbol{\beta}}^{(OLS)} = \left(\boldsymbol{X}'^{(OLS)}\boldsymbol{X}^{(OLS)}\right)^{-1}\boldsymbol{X}'^{(OLS)}\boldsymbol{y}, \ \boldsymbol{X} = \boldsymbol{X}^{(OLS)} \quad (4)$$

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

$$\widehat{\boldsymbol{y}}_i^{(OLS)} = \boldsymbol{x}_i'^{(OLS)}\widehat{\boldsymbol{\beta}}^{(OLS)} = \boldsymbol{x}_i'^{(OLS)}\left(\boldsymbol{X}'^{(OLS)}\boldsymbol{X}^{(OLS)}\right)^{-1}\boldsymbol{X}'^{(OLS)}\boldsymbol{y}, \ i = 1,2,\dots,n \quad (5)$$

where $\boldsymbol{x}_i'^{(OLS)}$ is the i^{th} row of matrix $\boldsymbol{X}^{(OLS)}$, $n \times (k+1)$ vector, [5, 17].

MATERIALS AND METHODS

As specified in the literature, the two operating factors are trans-membrane pressure (TMP) and velocity with the multi-response variables; permeate flux (PF), turbidity removal and the chemical oxygen demand (COD) removal. The goal is to obtain an optimum setting of the factors that would simultaneously maximize PF, turbidity removal and COD removal [20]. The impression behind the local linear regression model is because it is flexible and can adjust favourably in addressing boundary bias problem and is not controlled to user specified form for the data [9, 10, 11].

In the literature the local linear regression with adaptive bandwidth (LLR_{AB}) was utilized, whereas in this paper we shall apply a global or fixed bandwidth for a local linear regression (LLR_{FB}) for smoothing of the data by location.

The Local Linear Regression (LLR) Model

An alternate method when the researcher lack information or there is partial knowledge of the functional form of the model, then the use of nonparametric regression model is most appropriate model and as such, the LLR model gain prominence in this situation. The LLR model is a weighted form of the least squares obtained from Local Polynomial Regression of order one, which is an improvement over the kernel regression model because it can adjust to bias both at the boundaries of the explanatory variables [9, 19, 21].

The LLR model is obtained from ordinary least squares theory. The LLR estimator $\widehat{\boldsymbol{y}}_i^{(LLR)}$ of $\boldsymbol{y}_i$ is given as: $\widehat{\boldsymbol{y}}_i^{(LLR)} = \boldsymbol{x}_i'^{(LLR)}\left(\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_i\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X}'^{(LLR)}\boldsymbol{W}_i\boldsymbol{y} = \boldsymbol{H}_i^{(LLR)}\boldsymbol{y}$, (6)

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:

$$\boldsymbol{X}^{(LLR)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad (7)$$

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

$$\boldsymbol{W} = v_i\delta_{ir}, \quad i = 1, 2, \dots, n; \ r = 1, 2, \dots, n$$

where c_i are kernels weight at *ith* location and δ_{ir} is the Kronecker delta function given as

$$\delta_{ir} = \begin{cases} 1, \ if\ i = r \\ 0, \ otherwise \end{cases} \quad i = 1, 2, \dots, n; \quad r = 1, 2, \dots, n \quad (8)$$

Thus,

$$\boldsymbol{W} = \begin{bmatrix} v_1\delta_{11} & v_1\delta_{12} & \cdots & v_1\delta_{1n} \\ v_2\delta_{21} & v_2\delta_{22} & \cdots & v_2\delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_n\delta_{n1} & v_n\delta_{n2} & \cdots & v_n\delta_{nn} \end{bmatrix} \quad (9)$$

$$\boldsymbol{W} = \begin{bmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_n \end{bmatrix}$$

where $v_1 = w_{i1}, v_2 = w_{i2}, \dots, v_n = w_{in}$. In terms of location, $\boldsymbol{W} = \boldsymbol{W}_i$

$$\boldsymbol{W_i} = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}, i = 1, 2, \ldots, n. \quad (10)$$ [8, 9, 23].

Equation (10) can be written as:

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

We can rewrite Equation (6) in terms of hat matrix as:

$$\hat{\boldsymbol{y}}^{(LLR)} = \boldsymbol{H}^{(LLR)}\boldsymbol{y}, \quad (11)$$

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

$$\boldsymbol{H}^{(LLR)} = \begin{bmatrix} \boldsymbol{x_1'}^{(LLR)}\left(\boldsymbol{X'}^{(LLR)}\boldsymbol{W_1}\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X'}^{(LLR)}\boldsymbol{W_1} \\ \boldsymbol{x_2'}^{(LLR)}\left(\boldsymbol{X'}^{(LLR)}\boldsymbol{W_2}\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X'}^{(LLR)}\boldsymbol{W_2} \\ \vdots \\ \boldsymbol{x_n'}^{(LLR)}\left(\boldsymbol{X'}^{(LLR)}\boldsymbol{W_n}\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X'}^{(LLR)}\boldsymbol{W_n} \end{bmatrix} \quad (12)$$

The flaw of LLR model is high bias in areas where there is curvature because the setup of model matrix of the LLR model do not incorporates quadratic terms [12].

Global Bandwidth

We presented data-driven locally adaptive bandwidths:

$b_{ij} = b \;\; i = 1,2, \ldots, n; j = 1,2, \ldots, k$ (13)

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

Experimental design

In RSM application, the factors are usually more than one. Hence, the choice of appropriate levels to be studied for the explanatory variables is also vital as it can affect model correctness. The Experimental Design phase permits an appropriate design that can adequately and substantially estimation relationship between the response and one or more factors. Ordinarily applied DOEs in RSM include: 2^k full factorial design, 3^k full factorial design, and the CCD. In CCD, the number of experimental setup or run can be obtained by the mathematical relation given by; $2^k + 2k + k_c$, and all the factors are studied at five levels given as: $(-, \alpha, -1, 0, 1, \alpha)$, where 2^k is the full factorial design, $2k$ axial (star) points which are located at distance $\alpha = \sqrt[4]{2^k}$ from the center point and k_c. In this case $k = 2$, the numbers of factors utilized in the design and $k_c = 5$. Therefore, the total number of experimental runs is equals thirteen and for the data collection see [9, 10]. Therefore, the factors and coded levels as given in the literature are shown below in Table 1:

Table 1: Coded stages and range for the design of experiments [9]

Variables	Factors or Input parameters	Coded Levels				
		-1.414(-α)	-1(Low)	0(Medium)	1(High)	+1.414(+α)
TMP (bar)	$A = x_1$	0.7	1.0	1.7	2.5	2.8
Velocity (m/s)	$B = x_2$	1.0	1.5	2.6	3.7	4.2

In Table 2, given below is the CCD as given in the literature for two experimental factors and three responses.

Table 2: Coded stages and range for the design of experiments [9, 10, 20]

Exptal. Run	Exptal. Design		Coded Levels		
	A: TMP (code)	B: Velocity (code)	PF (L/m^{2}h)	Turbidity Removal (%)	COD Removal (%)
1	0.7 (-1.414)	2.6 (0)	40	82	78
2	1.7 (0)	1.0 (-1.414)	32	85	82
3	1.7 (0)	2.6 (0)	48	74	79
4	2.5 (+1)	3.7 (+1)	64	72	68
5	2.8 (+1.414)	2.6 (0)	52	70	69
6	2.5 (+1)	1.5 (-1)	42	68	78
7	1.7 (0)	2.6 (0)	50	73	81
8	1.7(0)	2.6 (0)	47	75	78
9	1.7(0)	2.6 (0)	49	77	80
10	1.7(0)	4.2 (+1414)	68	69	68
11	1.0 (-1)	3.7 (+1)	58	80	73
12	1.0 (-1)	1.5 (-1)	33	88	82
13	1.7 (0)	2.6 (0)	49	76	79

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

The values of the operating factors are coded between 0 and 1. The data collected via a CCD is transformed by a mathematical relation:

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

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}, $\text{Min}(x_{OLD})$ and $Max(x_{OLD})$ are the minimum and maximum values in the vector x_{OLD} respectively, [8]. 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 $x_0: -1.414, 0$; $Min(x_{OLD}): -1.414, -1.414$; $Max(x_{OLD}): +1.414, +1.414$

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

$$Explanatory\ variable\ x_1 : x_{11} = \frac{-1.414 - (-1.414)}{((-1.414) - (-1.414))} = 0.0000$$

$$Explanatory\ variable\ x_2 : x_{12} = \frac{-1.414 - (0)}{((-1.414) - (+1.414))} = 0.5000$$

where $x_1 =$ A: TMP , $x_2 = B: Velocity$

Hence, we now make values of the operating factors that lie between zero and one for RSM data as given in Table 3. We shall fit the coded values to the quadratic model of Equation (3) and thereafter fit the OLS and the adaptive LLR model to the RSM data that are coded to lie in the interval of 0 and 1 as given in Table 3.

Table 3: Transformed RSM data in the interval of zero and one

Exptal. Run	Exptal. Design		Coded Levels		
	A: TMP (code)	**B: Velocity (code)**	**PF (L/m^{2}h)**	**Turbidity Removal (%)**	**COD Removal (%)**
1	0.0000	0.5000	40	82	78
2	0.5000	0.0000	32	85	82
3	0.5000	0.5000	48	74	79
4	0.8536	0.8536	64	72	68
5	1.0000	0.5000	52	70	69
6	0.8536	0.1464	42	68	78
7	0.5000	0.5000	50	73	81
8	0.5000	0.5000	47	75	78
9	0.5000	0.5000	49	77	80
10	0.5000	1.0000	68	69	68
11	0.1464	0.8536	58	80	73
12	0.1464	0.1464	33	88	82
13	0.5000	0.5000	49	76	79

Multi-Response Optimization Problem

This involves the optimization of two or more responses simultaneously with the associated factors $(x_{i1}, x_{i2}, \ldots, x_{ik})$. The optimization criteria and the desired goal for the multi-response problem as given in [20] as shown in Table 4 below:

Table 4: Optimization criteria at the desired goal [9]

Criteria	Goal	Lower Limit	Upper Limit
TMP (bar)	In the range	1	2.5
Velocity (m/s)	In the range	1.5	3.7
PF (L/m^2 h)	Maximize	32	68.4
Turbidity Removal (%)	Maximize	68	88
COD Removal (%)	Maximize	68	82

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:

$$d_p\left(\hat{y}_p(\boldsymbol{x})\right) = \begin{cases} 0, & \hat{y}_p(\boldsymbol{x}) < L \\ \left\{\frac{\hat{y}_p(\boldsymbol{x})-L}{T-L}\right\}^{t_1}, & L \le \hat{y}_p(\boldsymbol{x}) \le T, \\ 1, & \hat{y}_p(\boldsymbol{x}) > T, \end{cases}$$
$$s.t\ \boldsymbol{x}\epsilon\ \varphi\ , \qquad (15)$$

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 [6, 12, 13, 22].

In the application, we give the individual desirability of the process requirement for the three responses.

For PF (L/m2 h); $d_p\left(\hat{y}_p(\boldsymbol{x})\right)$ is given as:

$$d_1(\hat{y}_1(\boldsymbol{x})) = \begin{cases} 0, & \hat{y}_1(\boldsymbol{x}) < 32 \\ \left\{\frac{\hat{y}_1(\boldsymbol{x})-32}{68.4-32}\right\}^{t_1}, & 32 \le \hat{y}_1(\boldsymbol{x}) \le 68.4, \\ 1, & \hat{y}_1(\boldsymbol{x}) > 68.4, \end{cases}$$
$$s.t\ \boldsymbol{x}\epsilon\ \varphi\ , \qquad (16)$$

Where $T = 68.4,\ L = 32$ are the maximum acceptable value and lower limit, respectively, of the 1^{st} response.

For Turbidity Removal (%); $d_2\left(\hat{y}_2(\boldsymbol{x})\right)$ is given as:

$$d_2(\hat{y}_2(\boldsymbol{x})) = \begin{cases} 0, & \hat{y}_2(\boldsymbol{x}) < 68 \\ \left\{\frac{\hat{y}_2(\boldsymbol{x})-68}{88-68}\right\}^{t_1}, & 68 \le \hat{y}_2(\boldsymbol{x}) \le 88, \\ 1, & \hat{y}_2(\boldsymbol{x}) > 88, \end{cases}$$
$$s.t\ \boldsymbol{x}\epsilon\ \varphi\ , \qquad (17)$$

Where $T = 88,\ L = 68$ are the maximum acceptable value and lower limit, respectively, of the 2^{nd} response.

For COD Removal (%); $d_3\left(\hat{y}_3(\boldsymbol{x})\right)$ is given as:

$$d_3(\hat{y}_3(\boldsymbol{x})) = \begin{cases} 0, & \hat{y}_3(\boldsymbol{x}) < 68 \\ \left\{\frac{\hat{y}_3(\boldsymbol{x})-68}{82-68}\right\}^{t_1}, & 68 \le \hat{y}_3(\boldsymbol{x}) \le 82, \\ 1, & \hat{y}_3(\boldsymbol{x}) > 82, \end{cases}$$
$$s.t\ \boldsymbol{x}\epsilon\ \varphi\ , \qquad (18)$$

Where $T = 82,\ L = 68$ are the maximum acceptable value and lower limit, respectively, of the 3^{rd} response.

The overall desirability function

The aim of desirability function is to maximize the overall desirability, D, which is the geometric mean of the individual desirability functions. Overall desirability D is given as:

$$D = \sqrt[p]{(d_1(\hat{y}_1(\boldsymbol{x})).\ d_2(\hat{y}_2(\boldsymbol{x}))\ \ldots\ d_v(\hat{y}_v(\boldsymbol{x}))} \qquad (19)$$

where $v = 3$ is the number of response variables, $d_1\hat{y}_1(\boldsymbol{x})$, $d_2\hat{y}_2(\boldsymbol{x})$,…, $d_v\hat{y}_v(\boldsymbol{x})$ are the individual desirability [11, 13, 17]. The desirability function $d_{\boldsymbol{m}}\left(\hat{y}_{\boldsymbol{m}}(\boldsymbol{x})\right), m = 1,2,\ldots,v$ allocate values between 0 and 1 centered on the process requirements such that the most undesirable and desirable values are $d_v\left(\hat{y}_v(\boldsymbol{x})\right) = 0$ and $d_v\left(\hat{y}_v(\boldsymbol{x})\right) = 1$ 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:

$$PRESS = \sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}(.)\right)^2 \qquad (20)$$

$PRESS^{**}$ 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:

$$PRESS^{**}(\Omega) = \frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_{i,-i(.)}\right)^2}{n-trace\left(H^{(.)}(\Omega)\right)+(n-k-1)\frac{SSE_{max}-SSE_{\Omega}}{SSE_{max}}} \qquad (21)$$

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 [8, 9,10]. For data emanating from RSM, the vector of optimal bandwidths $\Omega \in \left[b *_{ij}\right]$ is derived based on the minimization of the Penalized Prediction Error Sum of Squares, $PRESS^{**}$ see, [8, 9, 10]. The Sum of Squares Error, SSE and the Mean Square Error, MSE describes how well each regression model estimates f as given in equation (1). The statistic for selecting estimates is that it should minimize the sum of squares residuals, also known as the sum of squares error given as SSE. Thus,

$$SSE = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (22)$$

The mean square error, MSE is given as:

$$MSE = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-k} \qquad (23)$$

where n, is the number of sample size, k, is the number of explanatory variables used in the study and the degree of freedom, $n-k$ plays a significant role in the value of MSE, since smaller degree of freedom may result to a larger MSE, while larger degree of freedom may also result to smaller MSE see [8, 9,10].

In RSM, the coefficient of determination, R^2 and the adjusted coefficient of determination, R^2_{adj} statistics are also used to validate model adequacy of a real system. R^2 measures the extent of the observed variability in the experimental data the model can explain. The value of R^2 tends to increase as the number of explanatory variables increases and as such

R^2_{adj} was derived to adjust or modify R^2 inadequacy by taking into consideration the number of explanatory variables in the model. The R^2 and R^2_{adj} are expressed in equations (24) and (25) respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (24)$$

$$R^2_{adj} = 1 - \frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \quad (25)$$

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.

RESULTS AND DISCUSSION

The results in Table 5 clearly show that the LLR_{FB} method outperformed OLS and QM in terms of responses such as; permeate flux (L/m²h), turbidity removal (%), and COD removal (%). LLR_{FB} achieved better performance statistics in thirteen cells, compared to four cells for OLS and one cell for QM, for the multi-response problem. A tie between QM and LLR_{FB} occurred in only one cell. The cells in bold indicate better performance compared to those that are not bold. For minimum PRESS, PRESS, SSE, and MSE values indicate better statistical performance, while maximum R² (%) and R²Adj (%) values also show better performance.

Table 5: Model Goodness-of-fits statistics for OLS, QM and LLR_{AB}

Response	*Model*	*DF*	*PRESS***	*PRESS*	*SSE*	*MSE*	$R^2(\%)$	$R^2_{Adj}(\%)$
y_1	OLS	7.0000	6.7789	47.4521	10.7304	1.5329	99.2081	98.6425
	QM	-	-	-	-	-	99.1700	98.5800
	$\boldsymbol{LLR_{FB}}$	6.7976	**3.8129**	**51.4833**	**9.5501**	**1.4049**	**99.3000**	**98.7600**
y_2	OLS	7.0000	62.2476	435.7333	69.0777	9.8682	84.8820	74.0834
	QM	-	-	-	-	-	**97.7000**	**94.5000**
	$\boldsymbol{LLR_{FB}}$	4.4149	**21.7814**	**293.8003**	**10.6993**	**2.4234**	**97.7000**	93.4300
y_3	OLS	7.0000	5.0451	**35.3159**	9.0237	**1.2891**	97.2254	**95.2436**
	QM	-	-	-	-	-	97.2500	94.5000
	$\boldsymbol{LLR_{FB}}$	4.0307	**4.4440**	59.2554	**5.2000**	1.3000	**98.4000**	95.2000

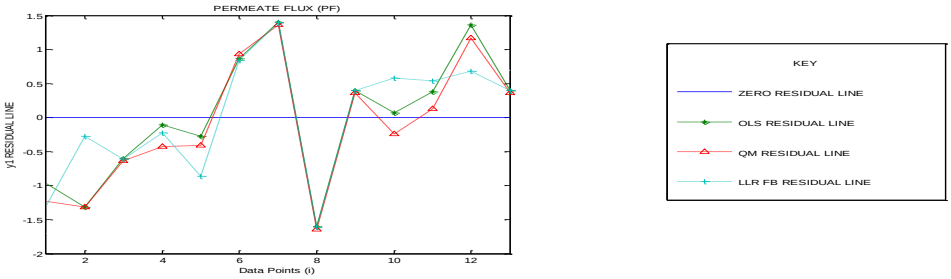

Figure 1: Model Residuals for y1 (Permeate flux) response plotted against the data points for OLS, QM *and* LLR_{FB} models.

In figure 1, the deep blue straight line is the zero-residual plot; green line is the OLS plot; red line is the QM plot and light blue line is the LLR_{FB} plot respectively. The residual plots for the three models as specified in the KEY for multi-response data. Obviously, LLR_{FB} estimated the data better than OLS and QM in terms of all the performance statistics considered.

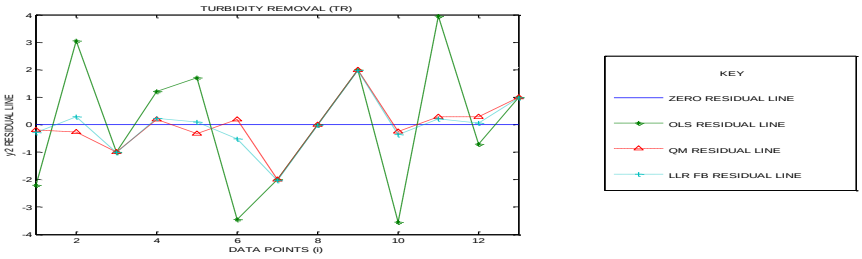

Figure 2: Model Residuals for y2 (Turbidity removal) response plotted against the data points for $OLS, QM\ and\ LLR_{FB}$ models.

In Figure 2, the deep blue straight line represents the zero-residual plot; the green line represents the OLS plot; the red line represents the QM plot; and the light blue line represents the LLR_{FB} plot. The residual plots for the three models are specified in the key for the multi-response data. Clearly, LLR_{FB} estimated the data better than OLS and QM in four of the performance statistics considered.

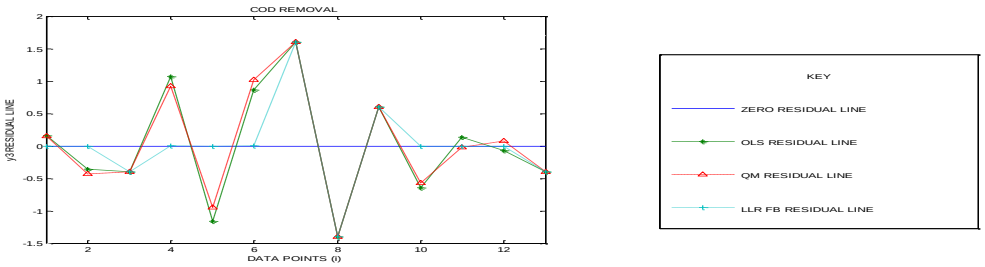

Figure 3: Model Residuals for y3 (COD removal) response plotted against the data points for $OLS, QM\ and\ LLR_{FB}$ models.

In Figure 3, the deep blue straight line represents the zero-residual plot, the green line represents the OLS plot, the red line represents the QM plot, and the light blue line represents the LLR_{FB} plot. The residual plots for the three models are specified in the key for the multi-response data. Clearly, LLR_{FB} estimated the data fairly better than OLS and QM in three of the performance statistics considered.

Table 6: Model optimal solution based on the multi-response desirability function

Model	x_1	x_2	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$d_1(\hat{y}_1)$	$d_2(\hat{y}_2)$	$d_3(\hat{y}_3)$	D**(%)**
OLS	0.2066	0.5396	46.0553	79.4585	79.2505	0.3861	0.5729	0.8036	56.23
QM	0.2727	0.6250	50.6600	78.8300	76.6300	0.5126	0.5415	0.6164	55.51
LLR_{FB}	**0.2835**	**0.3830**	41.9441	80.7240	79.9819	0.4232	0.6362	0.8559	**61.31**

From Table 6, LLR_{FB} offers an improved multi-response in the maximization of the ultrafiltration of oil-in-water emulsion over OLS and QM in terms of overall desirability for the respective factors $x_1 =$ Transmembrane pressure (TMP), $x_2 =$ Velocity.

Obviously, LLR_{FB} gave a better process condition with **61.31%** desirability and with operating factors $x_1 = \mathbf{0.2835}$, $x_2 = \mathbf{0.3830}$ with the best choice based on oil-in-water emulsion.

Table 7: Optimization and process requirement for the actual, predicted factors and responses

Criteria	**Goal**	**Lower limit**	**Target point**	LLR_{FB}	**QM**	**OLS**
TMP (bar)	Range	1	2.5	0.2835	0.2727	0.2066
Velocity (m/s)	Range	1.5	3.7	0.3830	0.6250	0.5396
PF (L/m^2 h)	Maximize	32	68.4	42	**51**	46
Turbidity R (%)	Maximize	68	88	**81**	79	80
COD R (%)	Maximize	68	82	**80**	77	79

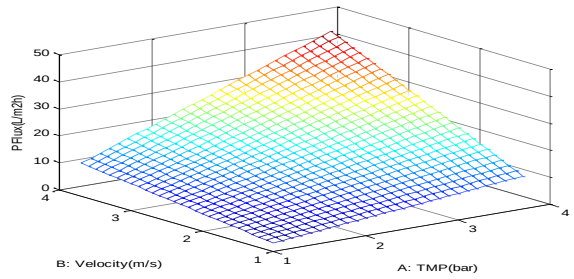

Figure 4: LLR_{FB} surface plot for maximum PF of 42% showing the effect of TMP and Velocity

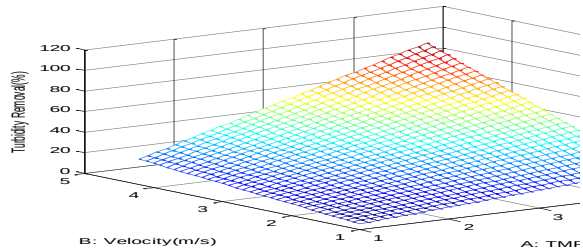

Figure 5: LLR_{FB} surface plot for maximum Turbidity removal of 81% showing the effect of TMP and Velocity

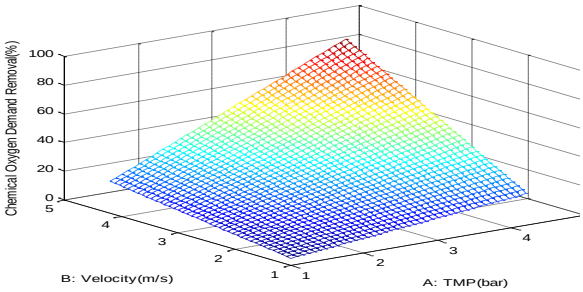

Figure 6: LLR_{FB} surface plot for maximum COD of 80% showing the effect of TMP and Velocity

The maximization of the single desirability functions are given in Figure 4, 5 and 6 respectively shows the shapes of the different colour variation in the surface plots representing single desirability of the optimization criteria for Permeate flux (L/m^{2}h), Turbidity removal (%) and COD removal (%) respectively. Therefore, in Figure 4, the individual desirability function for LLR_{AB} is 42% as against the optimization results for OLS and QM with respective individual desirability functions of 46% and 51% for Permeate flux (L/m^{2}h). Whereas, in Figure 5 the individual desirability function for LLR_{FB} is 81% as against the optimization results for OLS and QM with respective individual desirability functions of 80% and 79% for Turbidity removal (%) and in Figure 6, the individual desirability function for LLR_{FB} is 80% as against the optimization results for OLS and QM with respective individual desirability functions of 79% and 77% for COD removal (%).

CONCLUSION

In this paper, we examined the results from the regression models *OLS*, *QM*, and LLR_{FB} using the CCD approach, considering two operating factors, at five levels, and three responses for RSM data. The key findings and their implications are as follows: the kernel function, which regulates the shape of the kernel weights (e.g., Gaussian kernel) via the use of a fixed or global bandwidth, with x_i as the explanatory variable for location i, j as the number of explanatory variables used in the design, x_0 as a dummy known as the target point, and b as the fixed or global bandwidth or smoothing parameter controlling the smoothness of the estimated or fitted function, as given in equation (1). This illustrates the superiority of LLR_{FB} over OLS and QM models, as it can adjust to bias both at the boundaries and with unequal spacing of the explanatory variables. We have statistically analyzed the experimental data using the LLR_{FB} model which improve the operating features that would simultaneously optimize the multi-response problems through RSM data for permeate flux (L/m²h), turbidity removal (%), and COD removal (%), evaluating performance statistics, residual plots, and surface plots, and comparing LLR_{FB} with significant improvement over the existing OLS and QM models.

ACKNOWLEDGMENTS

I honestly want to thank God Almighty the giver of life, strength and wisdom to enable us accomplished this task.

REFERENCES

1. ANDERSON-COOK, C.M. & PREWITT, K. (2005). Some guidelines for using nonparametric models for modeling data from Response Surface Designs. *Journal of Modern Applied Statistical Models*, **4**: 106-119.
2. BOX, G.E.P. & WILSON, K.B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society Series B*, **13**: 1-45.
3. BOX, G.E.P. & HUNTER, J.S. (1957). Multifactor experimental designs for exploring response surfaces, *Ann Math Statist*, **28**: 195-241.
4. BOX, G.E.P. & DRAPER, N.R. (1959). A basis for the selection of a response surface design. *Journal of American Statistic Association*, **54**: 622-654.
5. 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.
6. CASTILLO, D.E. (2007). Process Optimization: A statistical Method. Springer International Series in Operations Research and Management Science: New York.
7. 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.
8. 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.
9. 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.
10. 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*, **21**(2): 425 – 435.
11. 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.
12. 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.
13. 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.
14. HILL, W.J. & HUNTER, W.G. (1966). A review of response surface methodology: A literature survey. *Technimetrics*, **8**: 571-579.
15. MYERS, R., MONTGOMERY, D.C. & ANDERSON-COOK, C.M. (2009). Response Surface Methodology: Process and Product Optimization Using Designed Experiments Wiley.
16. 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**: 464 – 478.
17. 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.
18. 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.
19. RUPPERT, D. & WAND, M.P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, **22**: 1346-1370.
20. SADEGHIAN, M., SADEGHI, M., HESAMPOUR, M. & MOHEB, A. (2014). Application of response surface methodology (RSM) to optimize operating conditions during ultrafiltration of oil in water emulsion, Desalination and water treatment, DOI:10.1080/19443994.2014.919607.
21. 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.
22. WAN, W. (2007), Semi-parametric techniques for ulti-response optimization. Ph.D Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University, Blacksburg, Virginia.
23. WAN, W. & BIRCH, J.B. (2011). A semiparametric technique for multi-response optimization. *Journal of Quality and Reliability Engineering International*, **27**: 47-59.
24. YENIAY, O. (2014). Comparative study of algorithm for response surface optimization. *Journal of Mathematical and Computational Applications*, **19**: 93 – 104.