OPTIMIZING BIODIESEL YIELD VIA AN ADAPTIVE LOCAL LINEAR REGRESSION MODEL WITH APPLICATION TO RESPONSE SURFACE METHODOLOGY

EGUASA, O.* AND EGUASA, M.E.
Department of Physical Sciences, Benson Idahosa University, Benin City, Nigeria

ABSTRACT

In the estimation of maximum biodiesel yield, a proportion of oil palm and cotton-seed oil of ratio 60% to 40% (P60C40) as given in the literature was collected for the production of biodiesel through ultrasound assisted trans-esterification process. The challenge in the estimation of biodiesel yield, is to find an appropriate regression model that would adequately capture maximum biodiesel yield production that would reflect the experimental yield of biodiesel via three factors namely; reaction time, methanol-to-oil ratio and concentration of catalyst. The existing techniques; second-order linear regression and the Artificial Neural Network (ANN) were utilized but could not capture local variability in the data because the coded factors lack the axial (star) points which allows for the estimation of curvature and maintain rotatability in the data. In other to address the challenge, we introduce an axial points to the coded factors as employed in the experimental design known as the Circumscribed Central Composite Design (CCCD) and the proposed adaptive local linear regression model to improve the goodness-of-fit statistics for Response Surface Methodology (RSM) data. The results obtained show that the proposed adaptive local linear regression model gave the maximum biodiesel yield of 96.31% that is approximately equal to the experimental biodiesel yield of 96.32% over the Ordinary Least Squares (OLS) of 94.95%, Second-order linear regression model of 96.41% and ANN of 96.67% respectively. In addition, the proposed adaptive local linear regression model has the least residual error over the other models utilized in this paper.

Keywords: Adaptive local linear regression model, Second-order regression model, Artificial Neural Network, Biodiesel yield, Response surface methodology, Circumscribed central composite design.
***Correspondence:** oeguasa@biu.edu.ng, +2348028631580

INTRODUCTION

Response Surface Methodology is defined as statistical technique applied by engineers and industrial statistician for empirical model building, with the aim of optimizing the response variables which are influenced by several explanatory variables [1, 2].

RSM is appropriate 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 [3].

The true response function f is usually unknown which must be estimated. Applying RSM, we seek to identify the functional relationship between the responses y and associated explanatory variables $(x_{i1},\ x_{i2},\ \ldots,\ x_{ik})$.

The general parametric regression model in matrix notation can be written as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2}$$

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 common approach 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:

$$\widehat{\boldsymbol{\beta}}^{(OLS)} = \left(\boldsymbol{X}'^{(OLS)}\boldsymbol{X}^{(OLS)}\right)^{-1}\boldsymbol{X}'^{(OLS)}\boldsymbol{y},\ \boldsymbol{X} = \boldsymbol{X}^{(OLS)}, \text{ is the OLS model matrix.} \tag{3}$$

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

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

where $\boldsymbol{x}_i'^{(OLS)}$ is the i^{th} row of matrix $\boldsymbol{X}^{(OLS)}$, $n \times (k+1)$ vector [4].

$\boldsymbol{H}_i'^{(OLS)} = \boldsymbol{x}_i'^{(OLS)}\left(\boldsymbol{X}'^{(OLS)}\boldsymbol{X}^{(OLS)}\right)^{-1}\boldsymbol{X}'^{(OLS)}$ 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:

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

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

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

[4, 5]

This is an experimental design used in fitting a second-order regression model. A second-order regression model is given as:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \sum_{j=1}^{k} \beta_{jj} x_{ij}^2 + \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \beta_{jr} x_{ij} x_{ir} + \varepsilon_i, \quad i = 1,2,\dots,n;\ r = j+1, j+2, \dots, k \tag{7}$$

where x_{ij}, x_{ir} are the explanatory variables; β_0 is a constant coefficient; the varying coefficients β_j, β_{jj} and β_{jr} are the coefficients of linear, quadratic and interaction terms respectively.

The parametric regression model may be superior if the user can adequately specify a parametric form for the data. Otherwise, the parametric regression model is misspecified and the optimal settings of the explanatory variables become miscalculated. Hence, a flexible regression model that does not give reference to a parametric form is preferred over the parametric form, the Local Linear regression (LLR) model.

MATERIALS AND METHODS

The LLR model is a weighted form of the least squares derived from Local Polynomial Regression of order one ($d = 1$) which can adjust to bias both at the boundaries and unequal spacing of the explanatory variables [6, 7].

The Adaptive Local Linear Regression Model (LLR)

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:

$$\widehat{\boldsymbol{y}}_i^{(LLR)} = \boldsymbol{x}_i'^{(LLR)} (\boldsymbol{X}'^{(LLR)} \boldsymbol{W}_i \boldsymbol{X}^{(LLR)})^{-1} \boldsymbol{X}'^{(LLR)} \boldsymbol{W}_i \boldsymbol{y} = \boldsymbol{H}_i^{(LLR)} \boldsymbol{y}, \tag{8}$$

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:

$$\boldsymbol{X}^{(LLR)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \tag{9}$$

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

$$\boldsymbol{W} = \boldsymbol{c_i \delta_{ip}}, \quad i = 1, 2, \dots, n;\ p = 1, 2, \dots, n$$

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

$$\delta_{ip} = \begin{cases} 1, & if\ i = p \\ 0, & otherwise \end{cases} \quad i = 1, 2, \dots, n;\ p = 1, 2, \dots, n \tag{10}$$

Thus,

$$\boldsymbol{W} = \begin{bmatrix} c_1\delta_{11} & c_1\delta_{12} & \cdots & c_1\delta_{1n} \\ c_2\delta_{21} & c_2\delta_{22} & \cdots & c_2\delta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_n\delta_{n1} & c_n\delta_{n2} & \cdots & c_n\delta_{nn} \end{bmatrix} \tag{11}$$

$$\boldsymbol{W} = \begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_n \end{bmatrix}$$

where $c_1 = w_{i1}$, $c_2 = w_{i2}$, …, $c_n = w_{in}$. In terms of location, $\boldsymbol{W} = \boldsymbol{W_i}$

$$\boldsymbol{W_i} = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}, \quad i = 1, 2, \dots, n. \tag{12}$$

[3, 8].

For a single explanatory variable problem used in the diagonal weight matrix $\boldsymbol{W_i}$, the kernel function $K\left(\frac{x_{ij}-x_{1j}}{b_{ij}}\right)$ is a simplified Gaussian kernel when one explanatory variable problem is used in the model matrix $\boldsymbol{X}$, given as:

$$w_{i1} = K\left(\frac{x_{ij}-x_{1j}}{b_{ij}}\right) = e^{-\left(\frac{x_{ij}-x_{1j}}{b_{ij}}\right)^2} \tag{13}$$

In a situation where more than one explanatory variable is used in the model matrix $\boldsymbol{X}$, the kernel weight w_{i1}, is a product kernel given as:

$$w_{i1} = \prod_{j=1}^{k} K\left(\frac{x_{ij}-x_{1j}}{b_{ij}}\right) \Big/ \sum_{p=1}^{n} \prod_{j=1}^{k} K\left(\frac{x_{pj}-x_{1j}}{b_{pj}}\right), p = 1,2,\dots,n, j = 1,2,\dots,k, \tag{14}$$

The values of i in equations 13 and 14 are obtained in equations 17 – 20 [3, 8].

In RSM, the matrix comprising the vector of optimal bandwidths $b_{11}^*, b_{12}, \dots, b_{nk}^*$ is obtained from the

minimization of the Penalized Prediction Error Sum of Squares ($PRESS^{**}$):

$$\text{Minimize } PRESS^{**}\{b_{11}, b_{12}, \dots, b_{nk}\} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}^{(LLR)}\right)^2}{n - trace\left(H^{(LLR)}(\boldsymbol{\omega})\right) + (n-k-1)\frac{SSE_{max} - SSE_{\boldsymbol{\omega}}}{SSE_{max}}}, \quad (15)$$

For $i = 1$ in equation (12), we have:

$$\boldsymbol{W_1} = \begin{bmatrix} w_{11} & 0 & \cdots & 0 \\ 0 & w_{12} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{1n} \end{bmatrix}_{(n \times n)} \quad (16)$$

$$w_{11} = \frac{\prod_{j=1}^{k} K\left(\frac{x_{1j} - x_{1j}}{b_{ij}}\right)}{\sum_{p=1}^{n}\prod_{j=1}^{k} K\left(\frac{x_{pj} - x_{1j}}{b_{pj}}\right)}, \quad p = 1, 2, \dots, n; j = 1, 2, \dots, k. \quad (17)$$

$$w_{11} = \frac{A}{[A + B + \cdots + C]}, \quad A = e^{-\left(\frac{x_{11}-x_{11}}{b_{11}}\right)^2} e^{-\left(\frac{x_{12}-x_{12}}{b_{12}}\right)^2} \dots e^{-\left(\frac{x_{1k}-x_{1k}}{b_{1k}}\right)^2}, B = e^{-\left(\frac{x_{21}-x_{11}}{b_{21}}\right)^2} e^{-\left(\frac{x_{22}-x_{12}}{b_{22}}\right)^2} \dots e^{-\left(\frac{x_{2k}-x_{1k}}{b_{2k}}\right)^2}$$

and $C = e^{-\left(\frac{x_{n1}-x_{11}}{b_{n1}}\right)^2} e^{-\left(\frac{x_{n2}-x_{12}}{b_{n2}}\right)^2} \dots e^{-\left(\frac{x_{nk}-x_{1k}}{b_{nk}}\right)^2}$

$$w_{12} = \frac{\prod_{j=1}^{k} K\left(\frac{x_{2j} - x_{1j}}{b_{ij}}\right)}{\sum_{p=1}^{n}\prod_{j=1}^{k} K\left(\frac{x_{pj} - x_{1j}}{b_{pj}}\right)}, \quad p = 1, 2, \dots, n; j = 1, 2, \dots, k. \quad (18)$$

$$w_{12} = \frac{D}{[E + D + \cdots + F]}, \quad D = e^{-\left(\frac{x_{21}-x_{11}}{b_{21}}\right)^2} e^{-\left(\frac{x_{22}-x_{12}}{b_{22}}\right)^2} \dots e^{-\left(\frac{x_{2k}-x_{1k}}{b_{2k}}\right)^2}, E = e^{-\left(\frac{x_{11}-x_{11}}{b_{11}}\right)^2} e^{-\left(\frac{x_{12}-x_{12}}{b_{12}}\right)^2} \dots e^{-\left(\frac{x_{1k}-x_{1k}}{b_{1k}}\right)^2}$$

and $F = e^{-\left(\frac{x_{n1}-x_{11}}{b_{n1}}\right)^2} e^{-\left(\frac{x_{n2}-x_{12}}{b_{n2}}\right)^2} \dots e^{-\left(\frac{x_{nk}-x_{1k}}{b_{nk}}\right)^2}$

$$\vdots$$

$$w_{1n} = \frac{\prod_{j=1}^{k} K\left(\frac{x_{nj} - x_{1j}}{b_{ij}}\right)}{\sum_{p=1}^{n}\prod_{j=1}^{k} K\left(\frac{x_{pj} - x_{1j}}{b_{pj}}\right)}, \quad p = 1, 2, \dots, n; j = 1, 2, \dots, k. \quad (19)$$

$$w_{1n} = \frac{G}{[J + H + \cdots + G]} \quad (20)$$

$G = e^{-\left(\frac{x_{n1}-x_{11}}{b_{n1}}\right)^2} e^{-\left(\frac{x_{n2}-x_{12}}{b_{n2}}\right)^2} \dots e^{-\left(\frac{x_{nk}-x_{1k}}{b_{nk}}\right)^2}$, $H = e^{-\left(\frac{x_{21}-x_{11}}{b_{21}}\right)^2} e^{-\left(\frac{x_{22}-x_{12}}{b_{22}}\right)^2} \dots e^{-\left(\frac{x_{2k}-x_{1k}}{b_{2k}}\right)^2}$

and $J = e^{-\left(\frac{x_{11}-x_{11}}{b_{11}}\right)^2} e^{-\left(\frac{x_{12}-x_{12}}{b_{12}}\right)^2} \dots e^{-\left(\frac{x_{1k}-x_{1k}}{b_{1k}}\right)^2}$

equation (12) translates to

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

$$\hat{\boldsymbol{y}}^{(LLR)} = \boldsymbol{H}^{(LLR)}\boldsymbol{y}, \quad (21)$$

Equation (8) can be written in terms of hat matrix as:

$$\boldsymbol{H}^{(LLR)} = \begin{bmatrix} \boldsymbol{x_1'}^{(LLR)}\left(\boldsymbol{X'}^{(LLR)}\boldsymbol{W_1}\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X'}^{(LLR)}\boldsymbol{W_1} \\ \boldsymbol{x_2'}^{(LLR)}\left(\boldsymbol{X'}^{(LLR)}\boldsymbol{W_2}\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X'}^{(LLR)}\boldsymbol{W_2} \\ \vdots \\ \boldsymbol{x_n'}^{(LLR)}\left(\boldsymbol{X'}^{(LLR)}\boldsymbol{W_n}\boldsymbol{X}^{(LLR)}\right)^{-1}\boldsymbol{X'}^{(LLR)}\boldsymbol{W_n} \end{bmatrix} \quad (22)$$

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 [5, 9].

Experimental design

In RSM, the number of factors is usually more than one. Hence, if the number of factors is too large, it may directly affect the response (Received signal strength) of interest, and since not all factors are desirable to be included in the experimental design for reason due to cost implication, it required the use of factor screening approach or two-level full factorial design to identify the variables with main effects [1, 8, 10].

Choice of adequate levels to be studied for the explanatory variables is also important as it can affect model accuracy. The Experimental Design phase permits an appropriate design that can provide adequate and considerable estimation relationship between the response and one or more factors. Usually applied Design of Experiments (DOEs) in RSM include: 2^k full factorial design, 3^k full factorial design, and the Central Composite Design (CCD).

In Table 1 gives the different coded levels for the three operating factors namely reaction time, methanol-to-oil ratio and concentration of catalyst.

Table 1: Input process factors for P60C40 yield optimization [11]

Operating Factors	Symbol	Coded Factors	Coded Levels		
			-1(Low)	0(Medium)	+1(High)
Reaction Time	m	A	20	35	50
Methanol-to-oil ratio	Vol. %/Vol. %	B	30	47.5	65
Concentration of catalyst	Wt. %	C	0.5	1	1.5

Table 2 represents the experimental run or the sample size of the factors and both the experimental yield and the predicted yield as given in [11].

Table 2: Experimental design for optimization of P60C40 biodiesel yield [11]

Experimental Run	Time (m)	M: O (v/v %)	Catalyst Concentration (w/w)	Experimental Yield (%)	Predicted Yield (%)
1	20	47.5	1.5	91.43	91.34
2	20	47.5	0.5	95.97	96.05
3	35	30	1.5	88.34	88.19
4	50	47.5	1.5	87.32	87.24
5	35	65	1.5	86.8	87.13
6	35	47.5	1	93.21	92.55
7	35	47.5	1	91.84	92.55
8	35	47.5	1	92.3	93.77
9	20	30	1	93.53	90.19
10	50	30	1	89.95	92.55
11	35	47.5	1	93.01	96.41
12	50	47.5	0.5	96.32	91.27
13	50	65	1	91.51	91.27
14	20	65	1	91.67	91.43
15	35	30	0.5	95.02	94.70
16	35	65	0.5	94.34	94.50
17	35	47.5	1	92.39	92.55

Table 3: Input process factors for P60C40 yield optimization

Operating Factors	**Symbol**	**Coded Factors**	**Coded Levels**				
			($-\alpha$)	**-1(Low)**	**0(Medium)**	**+1(High)**	**($+\alpha$)**
Reaction Time	m	A	5	20	35	50	65
Methanol-to-oil ratio	Vol. %/Vol. %	B	12.5	30	47.5	65	82.5
Concentration of catalyst	Wt. %	C	0.1	0.5	1	1.5	1.9

Table 4 explains the addition of the 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 utilized in the design. Hence, $\alpha = \pm 1.628$ [12].

Table 4: Experimental design for optimization of P60C40 biodiesel yield [11]

Experimental Run	**Time (m)**	**M: O (v/v %)**	**Catalyst Concentration (w/w)**	**Experimental Yield (%)**	**Predicted Yield (%)**
1	-1	-1	-1	91.43	91.34
2	1	-1	-1	95.97	96.05
3	-1	1	-1	88.34	88.19
4	1	1	-1	87.32	87.24
5	-1	-1	1	86.8	87.13
6	1	-1	1	93.21	92.55
7	-1	1	1	91.84	92.55
8	1	1	1	92.3	93.77
9	-1.682	0	0	93.53	90.19
10	1.682	0	0	89.95	92.55
11	0	-1.882	0	93.01	96.41
12	0	1.682	0	96.32	91.27
13	0	0	-1.682	91.51	91.27
14	0	0	1.682	91.67	91.43
15	0	0	0	95.02	94.70
16	0	0	0	94.34	94.50
17	0	0	0	92.39	92.55

The central composite design

A Central Composite Design allows for the building of the second-order regression model in a given response that is frequently used for process optimization [12, 13]. The three types of CCD are based on the locations of the factorial and star points in the design space namely; Circumscribed CCD (CCCD), Faced-Centered CCD and the Inscribed CCD.

The circumscribed central composite design

The most common CCD utilized in RSM is the circumscribed CCD because it allows for the estimation of curvature and the values of star points maintain rotatability which in turn depends on the factorial point of the design [14]. The circumscribed CCD involves three types of trials namely; two levels (2^k) full factorial

designs, $2k$ axial (star) points which are located at distance $\alpha = \sqrt[4]{2^k}$ from the center point and k_c, k^{th} central points [15]. The Circumscribed CCD can express geometrically as:

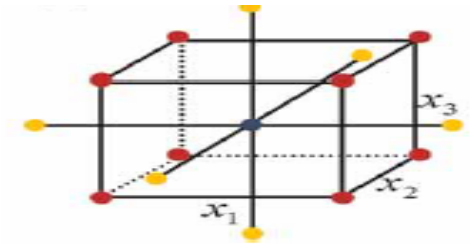

Figure 1: Circumscribed CCD (17 points, when k=3) with factorial design points (8 points), axial points (6 points) and with at least k^{th} central point (3 points). Sources: [1, 16]

The CCCD is utilized because it is cost efficient, maintain rotatability and accommodates small number of experimental runs in the design. The mathematical expression for the CCCD is given as:

$$CCCD = 2^k + 2k + k_c \quad (23)$$

where 2^k is the factorial portion, $2k$ is the axial or star points and k_c is at least kth central points utilized in the design. In this design $k = 3$ and $k_c = 3$ which from equation (23) sum up to 17 experimental runs.

Hereafter, circumscribed CCD shall be referred to as CCCD for easy reference. A CCCD has an advantage over 3^k full factorial design because it reduces the number of experimental runs (e.g. 17points in CCCD as against 27 points in 3^k design for *k*= 3).

Data transformation using central composite design (CCD) to RSM Data

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

$$x_{NEW} = \frac{Min(x_{OLD}) - x_0}{(Min(x_{OLD}) - Max(x_{OLD}))} \quad (24)$$

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, [8].

The natural or coded variables in Table 5 can be transformed to explanatory variables in Table 6 using Equation (24). Target points needed to be transformed from the first row for location 1 under the coded variables are given below:

Target points $x_0 : -1, -1, -1, ;$ $Min(x_{OLD}) : -1.682, -1.682, -1.682, ;$ $Max(x_{OLD}) : 1.682, 1.682, 1.682$

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

$$Explanatory\ variable\ x_1 : x_{11} = \frac{-1.682 - (-1)}{((-1.682) - (1.682))} = 0.2030$$

$$Explanatory\ variable\ x_2 : x_{12} = \frac{-1.682 - (-1)}{((-1.682) - (1.682))} = 0.2030$$

$$Explanatory\ variable\ x_3 : x_{13} = \frac{-1.682 - (-1)}{((-1.682) - (1.682))} = 0.2030$$

Table 5: Experimental design for optimization of P60C40 biodiesel yield

Experimental Run	**Time (m)** x_1	**M: O (v/v %)** x_2	**Catalyst Concentration (w/w)** x_3	**Experimental Yield (%)**	**Predicted Yield (%)**
1	0.2030	0.2030	0.2030	91.43	91.34
2	0.7970	0.2030	0.2030	95.97	96.05
3	0.2030	0.7970	0.2030	88.34	88.19
4	0.7970	0.7970	0.2030	87.32	87.24
5	0.2030	0.2030	0.7970	86.8	87.13
6	0.7970	0.2030	0.7970	93.21	92.55
7	0.2030	0.7970	0.7970	91.84	92.55
8	0.7970	0.7970	0.7970	92.3	93.77
9	0.0000	0.5000	0.5000	93.53	90.19
10	1.0000	0.5000	0.5000	89.95	92.55
11	0.5000	0.0000	0.5000	93.01	**96.41**
12	0.5000	1.0000	0.5000	**96.32**	91.27
13	0.5000	0.5000	0.0000	91.51	91.27
14	0.5000	0.5000	1.0000	91.67	91.43
15	0.5000	0.5000	0.5000	95.02	94.70
16	0.5000	0.5000	0.5000	94.34	94.50
17	0.5000	0.5000	0.5000	92.39	92.55

RESULTS AND DISCUSSION

The performance of the adaptive local linear regression model, the LLR_{AB} over OLS, second-order regression model and ANN of as given in [11] was investigated in terms of maximum biodiesel yield, the goodness-of-fit statistics and residual error for RSM data.

Table 6: The adaptive bandwidths for the three factors and the LLR_{AB} residual

Experimental Run	**Time (m)** x_1, b_1	**M: O (v/v %)** $x_2\ , b_2$	**Catalyst Concentration (w/w)** $x_3\ , b_3$	**Residual for** OLS	**Residual for** LLR_{AB}
1	0.2522	0.3200	0.2847	-0.8408	-0.0018
2	0.1334	0.3200	0.2847	1.0208	0.0121
3	0.2522	0.1686	0.2847	-2.5426	-0.0245
4	0.1334	0.1686	0.2847	-0.4861	-0.0085
5	0.2522	0.3200	0.1599	-0.8647	-0.0289
6	0.1334	0.3200	0.1599	1.1919	0.0020
7	0.2522	0.1686	0.1599	-2.3716	-0.0034

8	0.1334	0.1686	0.1599	-0.5100	0.0100
9	0.3014	0.2383	0.2178	3.2806	0.0281
10	0.1014	0.2383	0.2178	-1.3742	0.0001
11	0.1881	0.3829	0.2178	-0.9528	0.0076
12	0.1881	0.1278	0.2178	2.8592	0.0130
13	0.1881	0.2383	0.3356	1.0406	0.0026
14	0.1881	0.2383	0.1254	0.8658	0.0018
15	0.1881	0.2383	0.2178	0.9980	1.1090
16	0.1881	0.2383	0.2178	0.3180	0.4290
17	0.1881	0.2383	0.2178	-1.6320	-1.5210

Table 7: Experimental design for optimization of P60C40 biodiesel yield

Experimental Run	**Time (m)** x_1	**M: O (v/v %)** x_2	**Catalyst Concentration (w/w)** x_3	**Experimental Yield (%)**	**Predicted Yield (%) See [11]**	**Predicted OLS Yield (%)**	**Predicted Yield (%)** LLR_{AB}
1	0.2030	0.2030	0.2030	91.43	91.34	92.2708	91.4318
2	**0.7970**	**0.2030**	**0.2030**	95.97	96.05	**94.9492**	95.9579
3	0.2030	0.7970	0.2030	88.34	88.19	90.8826	88.3645
4	0.7970	0.7970	0.2030	87.32	87.24	87.8061	87.3285
5	0.2030	0.2030	0.7970	86.8	87.13	87.6647	86.8289
6	0.7970	0.2030	0.7970	93.21	92.55	92.0181	93.2080
7	0.2030	0.7970	0.7970	91.84	92.55	94.2116	91.8434
8	0.7970	0.7970	0.7970	92.3	93.77	92.8100	92.2900
9	0.0000	0.5000	0.5000	93.53	90.19	90.2494	93.5019
10	1.0000	0.5000	0.5000	89.95	92.55	91.3242	89.9499
11	**0.5000**	**0.0000**	**0.5000**	93.01	**96.41**	93.9628	93.0024
12	**0.5000**	**1.0000**	**0.5000**	**96.32**	91.27	93.4608	**96.3070**
13	0.5000	0.5000	0.0000	91.51	91.27	90.4694	91.5074
14	0.5000	0.5000	1.0000	91.67	91.43	90.8042	91.6682
15	0.5000	0.5000	0.5000	95.02	94.70	94.0220	93.9110
16	0.5000	0.5000	0.5000	94.34	94.50	94.0220	93.9110
17	0.5000	0.5000	0.5000	92.39	92.55	94.0220	93.9110

Table 8: Maximum experimental yield as compared with maximum yield of OLS, second-order regression model and ANN of [11] and LLR_{AB}.

S/N	**Model**	**Result Outcome (%)**	**Time (m)** x_1	**M: O (v/v %)** x_2	**Catalyst Concentration (w/w)** x_3
1	Maximum Experimental yield	96.32	0.5000 (50)	1.0000 (47.5)	0.5000 (0.5)
2	Maximum OLS yield	94.95	0.7970 (20)	0.2030 (47.5)	0.2030 (0.5)
3	Maximum RSM See [11]	96.41	0.5000 (35)	0.0000 (47.5)	0.5000 (1.00)
4	Maximum ANN See [11]	96.67	-	-	-
5	Maximum LLR_{AB} yield	96.31	0.5000 (50)	1.0000 (47.5)	0.5000 (0.5)

Table 9: Comparison of the goodness-of-fit statistics for the three model OLS, Second-order linear regression of [11] and the LLR_{AB}

Models	b^*	DF_{error}	Max. Yield %	MSE	SSE	$R^2\%$	$R^2_{adj}\%$	$PRESS$	$PRESS^*$	$PRESS^{**}$
OLS	-	7.0000	94.95	6.2613	43.8292	63.8592	17.3925	312.8207	**44.6887**	44.6887
See [11]	-	9.0000	96.41	13.28	119.52	95.6000	94.97	-	-	-
LLR_{AB}	*	2.0443	**96.31**	**1.8247**	**3.7302**	**96.9200**	**75.9300**	**287.1014**	140.4408	**19.6130**

The results obtained from Table 9, clearly shows that LLR_{AB} outperformed the OLS, Second-order linear regression model and the ANN result in Table 8 via [11] in terms of goodness – of – fit statistics such as PRESS**, PRESS, SSE, MSE, R^2 and R^2Adj and except for PRESS* that OLS turns out to be than other model. Hence, the proposed as given in Tables 8 and 9, the proposed LLR_{AB} in terms of maximum biodiesel yield is 96.31% with optimum conditions for the respective factors (50**m**, 47.5**v/v%** and 0.5**w/w**) which is approximately equal to the experimental biodiesel yield of 96.32% with optimum conditions for the respective factors (50**m**, 47.5**v/v%** and 0.5**w/w**) as compared to OLS with maximum biodiesel yield of 94.95% with optimum conditions for the respective factors (20**m**, 47.5**v/v%** and 0.5**w/w**), Second-order linear regression model of [11] with maximum yield of 96.41% and with optimum conditions for the respective factors (35**m**,

47.5**v/v%** and 1.00**w/w**) and ANN with maximum yield of 96.67% signifying that LLR_{AB} captures more local variability and curvature in the data to any other existing models.

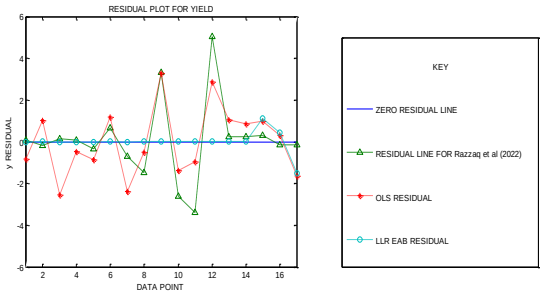

Figure 2: Residual plot for the three regression models OLS, [11] and LLR_{AB}

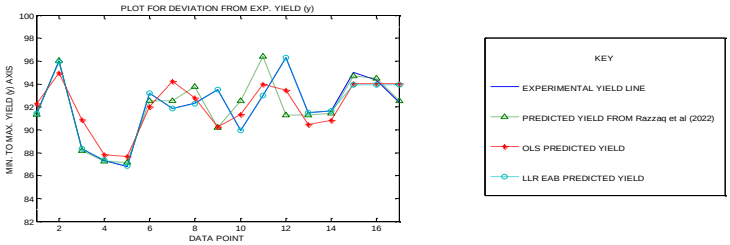

Figure 3: Experimental residual line versus three regression models OLS, [11] and (LLR_{AB})

From Figures 2 and 3, LLR_{AB} provides the smallest residual error over OLS, Second-order linear regression model and the ANN as given in [11] which also justifies the result of the goodness-of-fit statistics obtained as given above. Obviously, the LLR_{AB} performs better as compared with other models as presented.

CONCLUSION

In the estimation of maximum biodiesel yield, a proportion of P60C40) was collected for the production of biodiesel through ultrasound assisted trans-esterification process see [11]. The choice of statistical design to be utilized was prompted by the number of experimental runs for which the circumscribed central composite design was preferred over other design because it is cost efficient, captures local variability and maintain rotatability in the data and thereafter the data were coded to lie between zero and one for RSM data. On the other hand, the LLR_{AB} model is flexible because it is not restricted to user specified form and as such provides better results in terms of goodness-of-fit statistics and with the least residual error. Hence, the LLR_{AB} in terms of maximum biodiesel yield is 96.31% which is approximately equal to the experimental biodiesel yield of 96.32% with an absolute difference of 0.01% as compared to OLS with maximum biodiesel yield of 94.95% (1.37%), Second-order linear regression model as given in [11] with maximum yield of 96.41% (0.09%) and ANN with maximum yield of 96.67% (0.35%) signifying that LLR_{AB} captures more local variability and curvature in the data to any other models.

ACKNOWLEDGEMENTS

I sincerely want to thank my Ph.D. supervisor, Prof. J. I. Mbegbu for his tutelage and Dr. E. Edionwe for introducing me to my Ph.D. research area – Nonparametric Regression Models for Response Surface Methodology (RSM).

REFERENCES

1. 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.
2. YENIAY, O. (2014). Comparative study of algorithm for response surface optimization. *Journal of Mathematical and Computational Applications*, **19**(1): 93 – 104.
3. WAN, W. & BIRCH, J.B. (2011). A semi-parametric technique for multi-response optimization. *Journal of Quality and Reliability Engineering. International*, **27**(1): 47-59.
4. 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*.

5. 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.
6. RUPPERT, D. & WAND, M.P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics, 22: 1346-1370*.
7. 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.
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*, DOI:10.1080/02664763.2022.2026895
9. HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, T.A. (2009). The Elements of statistical learning, data mining, Inference and Prediction. Springer Series in Statistics. Second Edition.
10. MONTGOMERY, D.C. (2009). Introduction to statistical quality control. 7th edition, John Wiley & Sons, New York.
11. RAZZAQ, L., ABBAS, M.M., MIRAN, S., ASGHAR, S., NAWAZ, S., SOUDAGAR, M. E.M., SHUAKAT, N. VEZA, I., KHALIL, S., ABDELRAHMAN, A. & KALAM, M.A. (2022). Response Surface Methodology and Artificial Neural Networks – Based Yield Optimization of Biodiesel Sourced from Mixture of Palm and Cotton Seed Oil. Sustainability, https//doi.org/10.3390/su14106130
12. EGUASA, O. (2020). Adaptive Nonparametric Regression Models for Response Surface Methodology (RSM). PhD Thesis, Department of Mathematics, University of Benin, Benin City, Nigeria.
13. SIVARAO, I., ANAND, T.J.S. & SHUKOR, A. (2010). RSM Based Modeling for Surface Roughness Prediction in Laser Machining. *International Journal of Engineering and Technology IJET-IJENS*, **10**(4): 26 – 32.
14. DUTKA, M., DITARANTO, M. & LOVAS, T. (2015). Application of a Central Composite Design for the study of NO_X Emission Performance of Low NO_X Burner. *Energies*, **8**: 3606-3627.
15. 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**: 965-977.
16. PEASURA, P. (2015). Application of response surface methodology for modeling of postweld heat treatment process in a pressure vessel steel. *The Scientific World Journal, http.//dx.doi.org/10.1155/2015/318475.*