

Original Research Paper

The Efficiency of Seven-Variable Box-Behnken Experimental Design with Varying Center Runs on Full and Reduced Model Types

Mary Paschal Iwundu and Joan Cosmos

Department of Mathematics and Statistics, University of Port Harcourt, Port Harcourt, Nigeria

Article history

Received: 08-07-2022

Revised: 04-10-2022

Accepted: 15-10-2022

Corresponding Author:

Joan Cosmos

Department of Mathematics and Statistics, University of Port Harcourt, Port Harcourt, Nigeria

Email: cosmosjoane@gmail.com

Abstract: Response Surface Methodology is widely used in the optimization of industrial processes and products that depend on several experimental variables. One of the most tested and efficient second-order Response Surface Methodology designs is the Box-Behnken Design. This research explores the efficiency of a seven-variable Box-Behnken design with varying center runs, on full and reduced quadratic models. Backward Elimination and Forward Selection techniques are employed as the variable selection techniques for obtaining the reduced models, based on the Akaike Information Criterion. Fit Statistics and Design Efficiency values are obtained for the reduced models and are compared with those of the full model. Generally, results show that the reduced quadratic models perform best under one center-point run, thereby making the reduced models the most preferred in terms of the model fit and D-efficiency. Comparative analysis, based on G-efficiency, reveals that the full quadratic model performs better than the reduced models under one center-point Box-Behnken design.

Keywords: Response Surface Methodology, Box-Behnken Design, Variable Selection Techniques, Model Selection Criteria, Fit Statistics, D-Efficiency, G-Efficiency

Introduction

Most industry sectors utilize experimental design to optimize products or processes. A widely used technique for optimizing responses is Response Surface Methodology, also referred to as Response Surface Modelling (RSM). This method investigates the relationship between several explanatory variables (independent variables) and one or more response variables (dependent variables), (Mahallati, 2020). According to Karmoker *et al.* (2019) the Response Surface Methodology in application with appropriate design of experiments has become broadly and generally employed in most industry sectors for formulation optimization. As in George *et al.* (2005) response surface methodology is a collection of mathematical and statistical techniques that is convenient and effective for developing, improving, and optimizing products and/or processes. RSM provides a statistical approach that can be utilized in optimizing the product or response by optimization of the operational factors. A common goal when working with

response surface data is to optimize the response by finding the appropriate settings for the design (Gary, 2000).

When optimizing a response, there may be complications, especially in cases where there are several responses. In the event of this complication, a compromise optimum must be sought so that the responses are good (Gary, 2000). The Box-Behnken Design (BBD) is one of the most efficient and widely used designs for response surface models. A BBD utilizes three levels to each factor with just a few design points in comparison with the classical 3^k factorial design. Specifically, for three factors (i.e., $k = 3$), BBD has 13 distinct design points, which is less than half the number of distinct design points associated with a 3^3 factorial design. For $k = 4$, BBD has 25 distinct design points which are less than one-third of the number of distinct design points associated with a 3^4 factorial design. For a design in seven variables, the Box-Behnken design utilizes 57 distinct design points while a 3^7 factorial design requires 2187 distinct design points. Due to multiple design points, the 3^k factorial design is not regarded as an efficient design. Leiviskä (2013) described the Box-Behnken design as an independent quadratic design simply because there are no factorial or

fractional factorial designs embedded in the BBD. It is recognized as a unique three-level design because it does not contain any points at the vertices of the experimental region. Economically, this could be advantageous when the points on the corners of the cube represent level combinations that are excessively expensive, or impossible to test due to physical limitations. With the inclusion of the mid-level design point, the coefficients of a second-order model can be efficiently estimated (George *et al.*, 2005).

The statistical data due to (Kobyas *et al.*, 2014), on the modeling and optimization of Arsenite removal from groundwater using Al ball anodes by electrocoagulation process, is used as a case study for this study. Kobyas *et al.* (2014) employed the Box-Behnken experimental design to investigate the impact of seven operating parameters on Arsenite removal efficiency. The experiment was conducted using 62 experimental trials, having 57 distinct design points and 5 replicated center-point runs. The seven operating variables are given as: Initial pH, Current, Operating Time, Size of Al Ball Anode, Initial As (III) Concentration, Height of Al in the Reactor, and Air Flow Rate. The relationship between the seven operating variables and the response was established using a 36-parameter full quadratic regression model. In some cases, a 36-parameter model can be too large especially when there is limited time and resources. Hence, it becomes necessary to consider the impact of using a reduced model in describing and optimizing the response surface. This research aims at (i) Fitting more efficient models with fewer parameters, using variable selection techniques. (ii) Comparing the fit statistics of the new models with those associated with the 36-parameter full quadratic model. (iii) Comparing D and G-efficiency values of the designs associated with the different models.

The popularity of Box-Behnken design in experimental sciences has been well-established in different fields of study. Hendrick and Michel (1987) employed the Box-Behnken design in optimization process for direct-current plasma system. The effect of horizontal position, vertical position, nebulizer pressure, and electrode sleeve pressure on precision, drift, and sensitivity of the copper signal were considered. Bosque-Sendra *et al.* (2001) elucidated the benefits of utilizing the Box-Behnken design in modelling second-order response surfaces. Kobyas *et al.* (2014) utilized a seven-variable Box-Behnken experimental plan to model and optimize the removal of Arsenite from ground water. Aziz and Aziz (2018) utilized Box-Behnken Design in studying sound protection and noise absorber. In the optimization of chromatography, (Czyrski and Sznura, 2019) applied the Box-Behnken design in the utilization of High-Performance Liquid Chromatography (HPLC) Separation of Fluoroquinolones. The study aimed at selecting the most significant factors that influenced the responses of the chromatographic separation viz retention time, relative retention time, symmetry of the peaks, tailing factor, several

theoretical plates, Foley-Dorsey parameter, resolution factor, and peak width at half height. The results proved that the design employed was suitable. An interesting fact about the research work is that Box-Behnken design was used to address challenging chromatographic separations.

Omoruyi *et al.* (2019) compared some selection techniques in regression analysis. The study analyzed the performance of four variable selection techniques in developing a model that adequately predicts the dependent variable. The variable selection techniques include the: Direct search method, forward selection, backward elimination, and the stepwise regression method. A 32-year economic data on Real Gross Domestic Product being the dependent variable was used as a measure for economic growth and development. Growth Market Capitalization, All-shares index, the Market turnover, the openness of the Nigerian trade economy, the value of the transaction, the total listing of the Nigeria stock exchange, and total new issues were the independent variables. The result revealed that the backward elimination technique performed best in a variable selection based on the sample collected and it is supported by the use of all possible combination techniques as a control.

Chowdhury and Turin (2020) discussed the concept of variable selection strategies and demonstrated their importance in clinical prediction modeling. All four variable selection techniques namely: Backward elimination, forward selection, stepwise selection, and all subset selection were reviewed and adopting the proper stopping rule/selection criteria (p-values, Akaike information criterion, Bayesian information criterion, and Mallows' Cp statistic) in variable selection. The pros and cons of the techniques were discussed. The paper noted that a crucial part of prediction modeling is the inclusion of appropriate variables since the performance of the model depends on the variables that are ultimately included. Inaccurate results are produced when the proper variables are not included in the model and the true relationship that exist in the data between the selected variables and the outcome will not be reflected in the model. Consequently, researchers ought to be informed and mindful of these crucial aspects of prediction modeling.

A study by Alhajabdalla *et al.* (2021) highlighted the usefulness of the Box-Behnken Design for response surface methodology. A BBD in three variables was used to detect the influence of polymer concentration (ranging from 1-8 vol), fiber concentration (ranging from 0.01-0.08 wt%), and temperature (ranging from 25-80°C) on fibrous suspension stability. Minitab statistical software was used in the analysis of the stability measurements. The BBD performed well in determining the optimum conditions for the stability of the fibrous suspensions. The results predicted by the developed model were in good agreement with the experimental results. The importance of BBD for

response surfaces involving several factors at three levels, coded as 1, 0, 1, was noted by Beg and Akhter (2021) in the development of drug and process optimization in the pharmaceutical industry.

Most research has focused on either the application of the Box-Behnken design in the optimization of processes or products in a different industry sector or the efficiency of the Box-Behnken design in comparison to other response surface designs such as the three-level full factorial design, central composite design amongst others. Little emphasis has been given to investigating the robustness and efficiency of the Box-Behnken design in the face of non-standard (reduced) models under varying center point runs and that is what will be explored in this research work. This study will show how efficient and effective the Box-Behnken design is under reduced models with fewer center point runs. This can be very much applicable in cases where the design has several independent variables. Also with limited resources, it may not be practicable to experiment with the full model parameters and all 3^k design points (experimental runs). Consequently, the application of model reduction techniques for BBD with a fewer number of center-point runs.

Materials and Methods

Box and Behnken (1960) disclosed that for effective fitting of a second-order model, a BBD should utilize the mid-points of the edges as well as the center points. Hence, a BBD uses face points and not the extreme (corner) points, which are often more practicable to implement rather than the corner points used in Central Composite Design (CCD). A Box-Behnken design combined with a central composite design generates a three-level full factorial (Rao and Kumar, 2012). Myers *et al.* (2016); Beg and Akhter (2021), helpful tips are given on the factorial structure of k variable Box-Behnken design. Specifically, balanced incomplete block design forms the basis for the construction of the Box-Behnken designs. The Box-Behnken design can be applied to several factors ranging from three to twenty-one (Beg and Akhter, 2021) and are generally rotatable or nearly rotatable. Although it is not feasible to graphically represent an image for a three-level array of seven-factor Box-Behnken Experimental Design, the experimental design is given by the matrix D where each component represents a vector of design runs. As compared to the full factorial design, Box-Behnken design does not require several points since points and the number of points is chosen so that the variance of the design is about the same in the middle of the design as it is on the outside of the design. Furthermore, Box-Behnken design does not exist in two design variables.

$$D = \begin{pmatrix} 0 & 0 & 0 & \pm 1 & \pm 1 & \pm 1 & 0 \\ \pm 1 & 0 & 0 & 0 & 0 & \pm 1 & \pm 1 \\ 0 & \pm 1 & 0 & 0 & \pm 1 & 0 & \pm 1 \\ \pm 1 & \pm 1 & 0 & \pm 1 & 0 & 0 & 0 \\ 0 & 0 & \pm 1 & \pm 1 & 0 & 0 & \pm 1 \\ \pm 1 & 0 & \pm 1 & 0 & \pm 1 & 0 & 0 \\ 0 & \pm 1 & \pm 1 & 0 & 0 & \pm 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Data for Box-Behnken Design in Seven Variables

For this study, the design used is a seven-variable Box-Behnken design laid out in sixty-two experimental trials (six center point runs inclusive). Secondary data collected from (Koby *et al.*, 2014) to model and optimize the removal of arsenite from groundwater was remodeled (using fewer model coefficients) and analyzed for model fits and efficiencies. In this study, only one of the responses (removal efficiency) is considered. A total of seven independent factors were used and had three levels coded as -1, 0, and +1 (where -1 is the low level, 0 is the mid-level, and +1 is the high level). The seven independent variables are defined as: $X_1 = \text{pH}$, $X_2 = \text{current denoted as } i(\text{A})$, $X_3 = \text{operating time denoted as tEC (min)}$, $X_4 = \text{size of anode denoted as dP (mm)}$, $X_5 = \text{arsenite concentration denoted as Co (mg/L)}$, $X_6 = \text{anode height in the reactor denoted as h (cm)}$, $X_7 = \text{air flow rate denoted as } Q_{\text{air}} \text{ (L/min)}$. The levels of the variables in their natural units (uncoded) are given as follows:

- X_1 : (5.5, 7.0, 8.5)
- X_2 : (0.1, 0.3, 0.5)
- X_3 : (8, 15, 22)
- X_4 : (5.0, 7.5, 10)
- X_5 : (100, 550, 1000)
- X_6 : (2, 5, 8),
- X_7 : (2, 6, 10)

Table 1 gives the experimental data of the independent variables in their natural units and Table 2 gives the experimental data in coded form.

Role of Design Center Points

Center points are experimental runs where the factors X_i 's are set either at the center or mid-point of the factor levels. This implies that the variables are set halfway between the low and high settings. An experiment can have one center point, or the data can be collected at the center point several times. For this research work, the number of center-point runs $n_c \leq 6$.

Table 1: Box-behnken design in seven-variable using natural experimental data (Kobyia *et al.*, 2014)

Run No.	Independent variables							Response Re (%)
	pH (-)	I (A)	tEC (min)	dP (mm)	Co (mg/L)	H (cm)	Qair (L/min)	
1	7.0	0.3	15	5.0	1000	2	6	97.5
2	7.0	0.5	15	7.5	1000	5	10	99.2
3	5.5	0.5	15	10.0	550	5	6	92.4
4	7.0	0.3	15	10.0	1000	8	6	99.4
5	5.5	0.3	15	7.5	550	2	10	97.7
6	8.5	0.5	15	10.0	550	5	6	99.9
7	8.5	0.1	15	7.5	550	2	2	86.3
8	7.0	0.1	15	7.5	550	8	6	93.8
9	7.0	0.3	15	5.0	550	5	10	97.3
10	8.5	0.1	15	5.0	550	5	6	85.8
11	7.0	0.3	15	10.0	550	5	2	98.3
12	7.0	0.5	15	7.5	1000	5	2	99.4
13	7.0	0.3	15	5.0	1000	8	6	99.0
14	7.0	0.3	15	7.5	550	5	6	98.9
15	5.5	0.3	15	7.5	550	8	10	90.7
16	7.0	0.5	22	7.5	550	8	6	99.9
17	5.5	0.1	15	10.0	550	5	6	98.9
18	5.5	0.3	15	7.5	550	2	2	99.8
19	7.0	0.3	8	5.0	550	5	2	97.5
20	7.0	0.1	22	7.5	550	8	6	96.1
21	7.0	0.1	8	7.5	550	2	6	94.3
22	8.5	0.3	15	7.5	550	8	2	99.4
23	7.0	0.1	15	7.5	100	5	2	91.1
24	5.5	0.1	15	5.0	550	5	6	92.2
25	7.0	0.3	22	10.0	550	5	10	99.9
26	7.0	0.3	15	7.5	550	5	6	99.0
27	7.0	0.5	22	7.5	550	2	6	98.3
28	7.0	0.3	15	7.5	550	5	6	99.2
29	7.0	0.5	8	7.5	550	2	6	96.8
30	7.0	0.5	8	7.5	550	8	6	98.7
31	8.5	0.3	15	7.5	550	2	10	98.3
32	7.0	0.3	15	10.0	1000	2	6	99.0
33	5.5	0.5	15	5.0	550	5	6	99.5
34	7.0	0.3	22	5.0	550	5	10	98.2
35	7.0	0.1	22	7.5	550	2	6	93.2
36	8.5	0.1	15	10.0	550	5	6	96.0
37	5.5	0.3	15	7.5	550	8	2	99.9
38	7.0	0.5	15	7.5	100	5	2	99.9
39	8.5	0.3	22	7.5	100	5	6	100.0
40	8.5	0.5	15	5.0	550	5	6	99.7
41	7.0	0.3	15	10.0	100	2	6	96.7
42	7.0	0.3	8	10.0	550	5	10	98.7
43	8.5	0.3	8	7.5	100	5	6	95.3
44	8.5	0.3	8	7.5	1000	5	6	96.5
45	7.0	0.3	15	10.0	100	8	6	98.5
46	7.0	0.3	15	7.5	550	5	6	100.0
47	8.5	0.3	15	7.5	550	8	10	99.9
48	7.0	0.3	15	5.0	100	8	6	100.0
49	8.5	0.3	22	7.5	1000	5	6	99.2
50	7.0	0.3	15	7.5	550	5	6	99.9
51	7.0	0.1	15	7.5	100	5	10	91.7
52	5.5	0.3	22	7.5	1000	5	6	99.0
53	7.0	0.5	15	7.5	100	5	10	100.0
54	7.0	0.1	15	7.5	1000	5	2	95.8
55	7.0	0.3	22	5.0	550	5	2	99.4
56	7.0	0.3	15	5.0	100	2	6	90.3
57	7.0	0.3	8	10.0	550	5	2	96.3

Table 1: Continue

58	7.0	0.3	15	7.5	550	5	6	98.7
59	7.0	0.1	15	7.5	1000	5	10	96.6
60	5.5	0.3	8	7.5	1000	5	6	98.3
61	5.5	0.3	22	7.5	100	5	6	99.8
62	5.5	0.3	8	7.5	100	5	6	99.2

Table 2: Layout of the seven variable box-behnken in coded terms (Kobya *et al.*, 2014)

Rep	X1	X2	X3	X4	X5	X6	X7
1	0	0	0	-1	1	-1	0
1	0	1	0	0	1	0	1
1	-1	1	0	1	0	0	0
1	0	0	0	1	1	1	0
1	-1	0	0	0	0	-1	1
1	1	1	0	1	0	0	0
1	1	-1	0	0	0	-1	-1
1	0	-1	0	0	0	1	0
1	0	0	0	-1	0	0	1
1	1	-1	0	-1	0	0	0
1	0	0	0	1	0	0	-1
1	0	1	0	0	1	0	-1
1	0	0	0	-1	1	1	0
1	0	0	0	0	0	0	0
1	-1	0	0	0	0	1	1
1	0	1	1	0	0	1	0
1	-1	-1	0	1	0	0	0
1	-1	0	0	0	0	-1	-1
1	0	0	-1	-1	0	0	-1
1	0	-1	-1	0	0	-1	0
1	1	0	0	0	0	1	-1
1	0	-1	0	0	-1	0	-1
1	-1	-1	0	-1	0	0	0
1	0	0	1	1	0	0	1
1	0	0	0	0	0	0	0
1	0	1	1	0	0	-1	0
1	0	0	0	0	0	0	0
1	0	1	-1	0	0	-1	0
1	0	1	-1	0	0	1	0
1	1	0	0	0	0	-1	1
1	0	0	0	1	1	-1	0
1	-1	1	0	-1	0	0	0
1	0	0	1	-1	0	0	1
1	0	-1	1	0	0	-1	0
1	1	-1	0	1	0	0	0
1	-1	0	0	0	0	1	-1
1	0	1	0	0	-1	0	-1
1	1	0	1	0	-1	0	0
1	1	1	0	-1	0	0	0
1	0	0	0	1	-1	-1	0
1	0	0	-1	1	0	0	1
1	1	0	-1	0	-1	0	0
1	1	0	-1	0	1	0	0
1	0	0	0	1	-1	1	0
1	0	0	0	0	0	0	0
1	1	0	0	0	0	1	1
1	0	0	0	-1	-1	1	0
1	1	0	1	0	1	0	0
1	0	0	0	0	0	0	0
1	0	-1	0	0	-1	0	1
1	-1	0	1	0	1	0	0

Table 2: Continue

1	0	1	0	0	-1	0	1
1	0	-1	0	0	1	0	-1
1	0	0	1	-1	0	0	-1
1	0	0	0	-1	-1	-1	0
1	0	0	-1	1	0	0	-1
1	0	0	0	0	0	0	0
1	0	-1	0	0	1	0	1
1	-1	0	-1	0	1	0	0
1	-1	0	1	0	-1	0	0
1	-1	0	-1	0	-1	0	0

- I. Center-point runs help in the estimation of pure error thereby providing a lot of information at a minimum cost. This is especially helpful in situations where the experiment is conducted with a single replicate
- II. They help detect the regression model's lack of fit. Lack-of-fit reveals the suitability or non-suitability of the chosen model to the data
- III. When the data collected is not sufficient, increasing the number of center point runs can increase the possibility of detecting significant factors and pure error can be estimated

$$\frac{1}{N} \sum_{i=1}^N x_{iu}^2 \quad ; \quad u = 1, 2, \dots, k$$

Given a p - parameter model for a given N -point design, the moment matrix is defined as $\frac{1}{N} X'X$ and is also called the normalized information matrix where X is the design matrix defined in Eq. (2). The information matrix is the second mixed moment of vector X (it is the covariance matrix of X) defined as:

$$\frac{1}{N} \sum_{i=1}^N x_{iu} x_{iw} \quad ; \quad u \neq w$$

Response Surface Models of Second Order

A second-order quadratic model is defined as:

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^{k-1} \beta_{ii} x_i^2 + \sum_{i(j)}^{k-1} \sum_{j=2}^k \beta_{ij} x_i x_j + \varepsilon \tag{1}$$

where, y denotes the observed response for treatment combinations $x = (x_1, x_2, \dots, x_p)$; β_i 's represents the linear effects of the i^{th} factor; β_{ii} represents the quadratic effect of the i^{th} factor; β_{ij} represents the interaction effects between the i^{th} and the j^{th} factors and ε is the random error term. ε is assumed to be independent and normally distributed with mean 0 and constant variance σ^2 i.e., $\varepsilon \sim N(0, \sigma^2)$.

In the matrix notation, the model is denoted by:

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

where:

- y is an $N \times 1$ vector of observations
- X is an $N \times p$ model matrix consisting of the levels of factors that are laid out in the model form
- β is a $p \times 1$ vector of parameters
- ε is an $N \times 1$ vector of random errors associated with y

The mean of y is the first moment defined by:

$$\frac{1}{N} \sum_{i=1}^N x_{iu} \quad ; \quad u = 1, 2, \dots, k$$

The variance of y is associated with second pure moments defined by:

The normalization is to remove the effect of varying design sizes. The information matrix determines or measures the degree of information about an unknown parameter Θ that can be obtained from a random variable given a specified amount of data. The inverse of the information matrix given as $X'X^{-1}$ is called the variance-covariance matrix. This contains the variances (diagonal elements) and the covariances (off-diagonal elements) of the model coefficients.

The full quadratic model for this study is a 36-parameter model given in Eq. (3):

$$y(\%) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_{11} x_{11}^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{44} x_4^2 + \beta_{55} x_5^2 + \beta_{66} x_6^2 + \beta_{77} x_7^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 + \beta_{15} x_1 x_5 + \beta_{16} x_1 x_6 + \beta_{17} x_1 x_7 + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{25} x_2 x_5 + \beta_{26} x_2 x_6 + \beta_{27} x_2 x_7 + \beta_{34} x_3 x_4 + \beta_{35} x_3 x_5 + \beta_{36} x_3 x_6 + \beta_{37} x_3 x_7 + \beta_{45} x_4 x_5 + \beta_{46} x_4 x_6 + \beta_{47} x_4 x_7 + \beta_{56} x_5 x_6 + \beta_{57} x_5 x_7 + \beta_{67} x_6 x_7 + \varepsilon \tag{3}$$

From Eq. 3:

- $x_1 \dots x_7$ represents the main effects terms
- $x_1^2 \dots x_7^2$ represents the quadratic terms
- $x_1 x_2 \dots x_6 x_7$ represents the cross product or interactions effects terms
- $y(\%)$ represents the response term
- $\beta_1 - \beta_7$ represents the linear effects of the factors
- β_{ii} represents the quadratic effect of the i^{th} factor
- β_{ij} represent the interaction effects between the i^{th} and the j^{th} factors
- ε represents the is the random error i.e., $\varepsilon \sim N(0, \sigma^2)$

Akaike's Information Criterion

Akaike Information Criterion (AIC) is a criterion used to select a model from a set of models. It is an estimator of prediction error and compares the quality of a set of statistical models to each other. The Akaike information criterion was named in 1971 after a Japanese statistician, Hirotugu Akaike. AIC is computed as a function of the total independent variables used to build the models and as a function of the maximum likelihood estimate of the model. The maximum likelihood tells how well the model reproduces the data. The best model following the AIC technique is one that explains the greatest amount of variation with few possible independent variables.

The AIC can be calculated as:

$$AIC = -N \ln(L(\hat{\theta})) + 2k \quad (4)$$

where:

N is given as the sample size

k is given as the number of parameter estimates in the model

$L(\hat{\theta})$ is a measure of the model fit and provides the model's maximum likelihood function value

When comparing multiple AIC values to determine which model is best, the lowest AIC value is selected as the best fit and most preferred model. Even though the AIC chooses the best model from a set, it does not say anything about absolute quality. That is to say that if all the models from the set are poor, it will still choose the best of the poor set of models.

Bayesian Information Criterion

The Bayesian Information Criterion (BIC) also known as Schwarz Criterion (SBIC) is a criterion for model selection among a finite set of models. It is based in part on the likelihood function and closely related to Akaike Information Criterion (AIC). In model fitting, it is possible to increase the likelihood by adding parameters but doing so may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC. The BIC was developed by Schwarz (1978) and gave a Bayesian argument for adopting it. The BIC is closely related to the Akaike Information Criterion (AIC).

Mathematically, BIC is denoted as:

$$BIC = -2 \ln(L(\hat{\theta})) + k \ln(n) \quad (5)$$

where:

N is given as the sample size

k is given as the number of parameters estimated by the model

θ is a set of all parameters

$L(\hat{\theta})$ represents the likelihood of the model being tested when evaluated at maximum likelihood values of θ given the specified data. $L(\hat{\theta})$ can also be understood to be the probability of obtaining the data acquired supposing the model being tested is a given.

The BIC simply reduces to maximum likelihood selection because the number of parameters is equal for the models of interest. When comparing models with the Bayesian information criterion, the BIC for each model is calculated. The model with the lowest BIC is considered the best or most preferred. A lower BIC can imply fewer explanatory variables, a good fit, or a combination of the two and has a dependence on the relative size of n and k .

Variable Selection Techniques

These are techniques used to discover a group of predictor variables to be included in a model which can be as small as possible that fits suitably and gives a good prediction of the response variable. The purpose of variable selection is to prevent over-fitting of the data (this occurs when many predictor variables are in the model) and to avoid loss of much information or underfitting when a lot of predictor variables are deleted from the model. The variable selection techniques include forward selection, backward elimination, stepwise selection, and all subsets. These methods automatically select variables that are significantly important in the model and can be easily performed using statistical software.

The model selection criterion used for the reduced model in this study is the Akaike Information Criterion. AIC was chosen because it exhibits more tolerance in measuring statistical model goodness of fit in the presence of several parameters. This is part of this research work objective- to suitably fit the data to the reduced models of the various variable selection techniques employed. Another reason for the choice of AIC is that BIC is more restrictive than AIC as a result, BIC produces smaller models. According to Hendrick and Michel (1987), BIC is most suitable and recommended for experiments with large sample sizes. That is where the sample size surpasses 100 per independent (predictor) variable. During the analysis carried out with Design Expert statistical software, only two variable selection techniques were adopted (which are forward selection and backward elimination) using the AIC model selection criterion.

Forward Selection

Forward Selection is a variable selection technique performed by running a simple regression analysis on all the independent variables and examining the variables having the greatest F-statistic (and the smallest p-value).

The algorithm for forward selection for a given threshold, α is as follows:

- i. Begin with a model that does not have any predictors (no variable in the model)
- ii. Insert variable that has the largest F-statistic (given P-value less than cut-off α value)
- iii. Refit and repeat step (ii). The F-statistics is recomputed at every step for the remaining variables, then the variable with the highest F-statistic is added to the model. This procedure is continued till there is no remaining variable that is significant at the specified level of α

One of the benefits of forward selection is that it begins with smaller models. This method is also less prone to collinearity (correlation of variables in a model).

The demerit of forward selection as stated in Mantel (1970) is that the forward selection does not examine the effects of the variables simultaneously since it begins with an empty equation. Consequently, in situations where variables in the model are correlated with each other none of the variables will be selected. Another drawback of the forward selection as identified by Chowdhury and Turin (2020) is that the insertion of a new variable may cause an existing variable in the model to be non-significant still the existing variable cannot be deleted from the model.

Backward Elimination

Backward Elimination is a variable selection technique that is the reverse of forward selection.

Given a threshold α , the algorithm for the backward elimination technique is defined below:

- i. Start with all predictor variables in the model
- ii. Remove the variable that has the lowest F-statistics (provided the P-value is greater than α)
- iii. Refit and repeat step (ii). The F-statistic is recomputed after removing variables. Thereafter variables with the lowest F-statistics are deleted. This process is continued till all the variables left in the model are significant at the stipulated level of α

An advantage of the backward elimination technique is its ability to evaluate the joint predictive variables since the procedure begins with all variables incorporated in the model. One of the disadvantages of backward elimination is that after a variable has been eliminated from the model it cannot be re-inserted again (Chowdhury and Turin, 2020). Another demerit of this technique is its non-suitability in cases where the number of variables to be considered is greater than the sample size. This is because it begins with the full model which includes all the predictors.

The quality or efficiency of an experimental design can be measured (Iwundu, 2017). The most common measures of the efficiency of an experimental design are focused on the design's information matrix. The efficiency measures should not be interpreted on their

own but can be used to compare designs. Given a set of designs, the one with the highest efficiency measure is best (with 100% as the maximum efficiency for any criterion). The design efficiency employed in this study is D-efficiency and G-efficiency.

The D-efficiency criterion aims at making the determinant of the variance-covariance matrix for the chosen model parameter estimates as minimum as possible. A design with the highest Deficiency should be the best and most preferred. The D-efficiency values are a function of the number of points in the design, the number of independent variables in the model, and the maximum standard error for prediction over the design points as (Goo's and Jones, 2011):

$$D\text{-efficiency} = \left(\frac{X'X}{n^p} \right)^{\frac{1}{p}} = \frac{X'X^{\frac{1}{p}}}{n}$$

where, $X'X$ is the information matrix of the design and can be denoted as M , p is the number of model parameters and n is the design size. The matrix X is called the model matrix and is defined using the design and the model.

G-efficiency compares the scaled prediction variance maximum value within the design region with regard to its theoretical minimum variance. Iwundu (2017) defines that the G-efficiency of a design can be computed mathematically as $\frac{p}{v(x)_{\max}}$ where, p is the number of model parameters and $v(x)_{\max}$ the maximum scaled variance of prediction thus computed as $\frac{N \text{Var}(\hat{y}(x))}{\sigma^2}$ where, N is given as the number of design points (sample size):

$$\text{Var}(\hat{y}(x)) = \max \{ x'M^{-1} \xi_{N \times 1} \}$$

M is the information matrix $X'X$ is the design

The variance is scaled to make up for the differences in the sample size N . When comparing designs, the quantity becomes scale-free by dividing by σ^2 and when multiplied by N the variance per observation is reflected in the quantity thereby removing the effects of the varying sample size.

Utilized Software

Two advanced statistical software packages namely, Design-Expert version 13 (Trial Version) and Minitab version 17, were utilized for the analysis of the Box-Behnken experimental data used in this research work.

Results and Discussion

The results of the analysis are presented in Tables 3(a-f), Table 4-5. For clarity, the fit statistics, diagnostic properties, and design efficiency values of the full and

reduced quadratic models have been summarized under varying center point runs for easy comparison.

Discussion of Findings

In the design of experiments, a variety of factors can influence how well a particular design performs owing to that fact that the objective of any design is to estimate the model parameters without bias and with the least variance. These factors may include model type, design region, positioning of the design points, and missing design points (Chekwube *et al.*, 2021). The data from (Kobya *et al.*, 2014) on optimizing the removal of arsenite from groundwater was remodeled and analyzed using the Box-Behnken design laid out in sixty-two experimental trials. Analysis of the published model produced a thirty-six-parameter model. Two variable selection techniques namely: Forward selection and backward elimination were employed to generate reduced models at varying design sizes. The design size was achieved by changing the number of center point runs (n_c) ranging from 1 through 6. All models did well in fitting the data.

The results revealed that the reduced models suitably fitted the data. The fit statistics and diagnostic properties values confirm that the forward selection reduced model is the overall preferred model and produced its best models at and 1 center point runs respectively. The backward elimination reduced model came second being highly competitive with its best models at 2 and 3 center point runs. This shows that the sparsity-of-effects principle stands valid since fewer effects are statistically significant. The full quadratic performed the least, with its best model being the design with 6 center point runs.

In terms of the AIC and BIC model selection criteria, the forward selection and backward elimination reduced models had the least AIC and BIC values at $n_c = 1$. While the full quadratic model has its least at $n_c = 6$. As the number of center point runs is increased, the fits and likelihood statistics for the reduced models tend to be closely related, plus at $n_c = 6$, these values are the same. This implies that given a higher number of center point runs for a Box-Behnken design, there may be a tendency for the variable selection methods (forward selection and backward elimination) to produce similar models when employing the AIC model selection criteria.

Comparing the design's D-efficiency values, the 57-point design size (i.e., having $n_c = 1$) for the reduced models (both forward selection and backward

elimination) produced the top choice design. Whereas the 58-point design size (i.e., having $n_c = 2$) design produced the best design for the full quadratic model.

Accessing the G-efficiency, the design with a 57-point design (i.e., having $n_c = 1$) produced the best design for the full quadratic model. The 60-point design (i.e., having $n_c = 3$) produced the best design for the forward selection reduced model, and the 58-point design (i.e., having $n_c = 2$), produced the best design for the backward elimination reduced model.

It is important to note that the forward selection technique as well as the backward elimination technique have the limitation of including insignificant terms in the final model. This limitation was resolved during the analysis by using the Step AIC which selected models with the smallest AIC values.

Following the results of the analysis, it is appropriate to describe the Box-Behnken design as a highly efficient design for fitting second-order response surface models with fewer experimental runs, despite its non-usage of the vertices (corner or axial points). As stated earlier, the Box Behnken experimental design is suited in cases where the corners of the cube points denote levels of combination which may be too expensive and impractical to test. Box-Behnken design does not require so many points since the points on the cube are close to the center point. This explains why the BBD with one center point run seems to produce the best models and Box-Behnken designs with fewer center point run produced the best designs.

The best models produced are summarized below:

1. For the full quadratic (36-parameter) model, the best model produced is that of $n_c = 6$
2. For the forward selection criterion-reduced model, the best model produced is the 19-parameter model with $n_c = 1$
3. For the backward elimination criterion-reduced model, the best model produced is the 18-parameter model with $n_c = 1$
4. The D-efficiency values were obtained for the different design sizes. The reduced model produced higher D-efficiency values. The most preferred design is that of a 57-point design (with $n_c = 1$) for both the full or reduced (forward selection and backward elimination) models
5. Accordingly, the G-efficiency values were obtained for the different model types at varying center runs. The 57-point design (with $n_c = 1$) for the full model produced the overall best design

Table 3a: Summary of ANOVA results for the removal efficiency response of full and reduced models at six center point runs $n_c = 6$

Models	P	R ²	Adj R ²	AIC	BIC	S. D	C. V %	PRESS	F-value	P-value
Full model	36	0.9339	0.8449	334.25	304.26	1.31	1.35	341.37	10.49	<0.0001
Reduced model forward selection	22	0.9173	0.8739	239.49	260.34	1.18	1.22	181.52	21.14	<0.0001
Reduced model backward elimination	22	0.9173	0.8739	239.49	260.34	1.18	1.22	181.52	21.14	<0.0001

Table 3b: Summary of ANOVA results for the removal efficiency response of full and reduced models at five center point runs $n_c = 5$

Models	P	R ²	Adj R ²	AIC	BIC	S. D	C. V%	PRESS	F-value	P-value
Full model	36	0.9338	0.8410	337.04	302.03	1.34	1.37	342.47	10.07	<0.0001
Reduced model forward selection	22	0.9710	0.8723	238.42	258.23	1.20	1.28	182.00	20.52	<0.0001
Reduced model backward elimination	22	0.9101	0.8625	238.46	258.99	1.23	1.27	183.20	20.25	<0.0001

Table 3c: Summary of ANOVA results for the removal efficiency response of full and reduced models at four center point runs $n_c = 4$

Models	P	R ²	Adj R ²	AIC	BIC	S. D	C. V%	PRESS	F-value	P-value
Full model	36	0.9335	0.8366	340.25	299.82	1.36	1.40	343.85	9.63	<0.0001
Reduced model forward selection	22	0.9166	0.8706	237.78	256.10	1.21	1.25	182.50	19.89	<0.0001
Reduced model backward elimination	21	0.9097	0.8634	237.14	256.81	1.25	1.28	183.69	19.64	<0.0001

Table 3d: Summary of ANOVA results for the removal efficiency response of full and reduced models at three center point Runs $n_c = 3$

Models	P	R ²	Adj R ²	AIC	BIC	S. D	C. V%	PRESS	F-value	P-value
Full model	36	0.9336	0.8325	343.30	297.00	1.38	1.42	340.88	9.23	<0.0001
Reduced model forward selection	18	0.8860	0.8387	235.17	255.46	1.36	1.40	192.72	18.74	<0.0001
Reduced model backward elimination	21	0.9100	0.8626	235.08	253.74	1.25	1.29	182.51	19.21	<0.0001

Table 3e: Summary of ANOVA results for the removal efficiency response of full and reduced models at two center point runs $n_c = 2$

Models	P	R ²	Adj R ²	AIC	BIC	S. D	C. V%	PRESS	F-value	P-value
Full model	36	0.9332	0.8268	347.49	294.81	1.41	1.46	342.58	8.78	<0.0001
Reduced model forward selection	18	0.8877	0.8400	232.26	251.81	1.36	1.40	189.51	18.60	<0.0001
Reduced model backward elimination	21	0.9096	0.8607	233.84	251.44	1.27	1.31	182.90	18.60	<0.0001

Table 3f: Summary of ANOVA results for the removal efficiency response of full and reduced models at one center point run $n_c = 1$

Models	P	R ²	Adj R ²	AIC	BIC	S. D	C. V%	PRESS	F-value	P-value
Full model	36	0.9350	0.8266	349.95	290.30	1.42	1.46	349.44	8.63	<0.0001
Reduced model forward Selection	19	0.9024	0.8562	226.45	244.73	1.29	1.33	168.67	19.52	<0.0001
Reduced model backward elimination	18	0.8932	0.8466	227.05	245.83	1.33	1.38	182.06	19.18	<0.0001

Table 4: D-efficiency values of the box-behnken design relative to the design size and model type

		D-efficiency (%)		
Design size	No of center points n_c	Full model	Reduced model (forward selection)	Reduced model (backward elimination)
62	6	17.76	22.28	22.28
61	5	17.96	22.57	23.16
60	4	18.15	22.87	23.46
59	3	18.32	23.84	23.76
58	2	18.44	24.22	24.06
57	1	18.43	24.47	24.59

Table 5: G-efficiency values of the Box-Behnken design relative to the design size and model type

		G-efficiency (%)		
Design size	No of center points n_c	Full model	Reduced model (forward selection)	Reduced model (backward elimination)
62	6	64.97	57.09	57.09
61	5	66.00	58.02	58.02
60	4	67.06	58.97	56.29
59	3	68.12	54.41	57.23
58	2	69.14	55.20	58.04
57	1	69.94	57.40	56.01

Conclusion

The efficiency and robustness of the Box-Behnken design have been thoroughly investigated and demonstrated in this research work, both in the face of standard (full

quadratic) and non-standard (reduced quadratic) models under varying center point runs. This study is practical in situations where physical process (materials/resources and time) limitations make it unrealistic to test the full experimental design. Thus, it is possible to obtain models

that suitably fit the given data and align with the goal of the experimental design and available resources. For effective results, nonstandard (reduced quadratic) models produce the best results when the design size is small (preferably with one center point run). Whereas the standard (full) model performs best when the design size is large (with at least six center point runs or a maximum number of center point runs generated by the statistical software).

Acknowledgment

The authors appreciate the Editors and reviewers for their insights and comments that enhanced the quality of the paper. A special thanks to Femi Oladiran for his moral support and Desmond Bartholomew for his technical advice.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Mary P. Iwundu: Writing-original draft preparation supervision, review and edited.

Joan Cosmos: Writing-original draft preparation, analysis, review and edited.

Ethics

This article is original and contains unpublished materials. The corresponding author confirms that the other author have read and approved the manuscript and there are no ethical issues involved.

References

- Alhajabdalla, M., Mahmoud, H., Nasser, M. S., Hussein, I. A., Ahmed, R., & Karami, H. (2021). Application of response surface methodology and box-behnken design for the optimization of the stability of fibrous dispersion used in drilling and completion operations. *ACS Omega*, 6(4), 2513-2525. <https://doi.org/10.1021/acsomega.0c04272>
- Aziz, A. R. A., & Aziz, S. A. (2018, December). Application of Box Behnken design to optimize the parameters for kenaf-epoxy as noise absorber. In IOP conference series: Materials science and engineering (Vol. 454, No. 1, p. 012001). *IOP Publishing*. <https://doi.org/10.1088/1757-899X/454/1/012001>
- Chekwube, B. D., Paul, O. C., & Joan, I.-C. (2021). Scaled Prediction Variances of Equiradial Design under Changing Design Sizes, Axial Distances and Center Runs. *Asian Journal of Probability and Statistics*, 11(1), 1-13. <https://doi.org/10.9734/ajpas/2021/v11i130256>
- Beg, S., & Akhter, S. (2021). Box–Behnken designs and their applications in pharmaceutical product development. In *Design of Experiments for pharmaceutical product development* (pp. 77-85). *Springer, Singapore*. https://doi.org/10.1007/978-981-33-4717-5_7
- Bosque-Sendra, J. M., Pescarolo, S., Cuadros-Rodríguez, L., García-Campaña, A. M., & Almansa-López, E. M. (2001). Optimizing analytical methods using sequential response surface methodology. Application to the pararosaniline determination of formaldehyde. *Fresenius' Journal of Analytical Chemistry*, 369(7), 715-718. <https://doi.org/10.1007/s002160100751>
- Box, G. E., & Behnken, D. W. (1960). Some new three level designs for the study of quantitative variables. *Technometrics*, 2(4), 455-475.
- Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, 8(1). <https://doi.org/10.1136/fmch-2019-000262>
- Czyrski, A., & Sznura, J. (2019). The application of Box-Behnken-Design in the optimization of HPLC separation of fluoroquinolones. *Scientific Reports*, 9(1), 1-10. <https://doi.org/10.1038/s41598-019-55761-z>
- Gary, W. O. (2000). First Course in Design and Analysis of Experiments, *University of Minnesota*. <http://users.stat.umn.edu/~gary/book/fcdae.pdf>
- George E. P. Box, J. Stuart J. Hunter, and William G. Hunter (2005). *Statistics for Experimenters: Design, Innovation and Discovery, 2nd Edition Wiley, New York*.
- Hendrick, M. S., & Michel, R. G. (1987). Optimization of a direct-current plasma emission echelle spectrometer. *Analytica Chimica Acta*, 192, 183-195. [https://doi.org/10.1016/S0003-2670\(00\)85703-7](https://doi.org/10.1016/S0003-2670(00)85703-7)
- Iwundu, M. P. (2017). Missing observations: The loss in relative A-, D-and G-efficiency. *International Journal of Advanced Mathematical Sciences*, 5(2), 43-9. <https://doi.org/10.14419/ijams.v5i2.7786>
- Karmoker, J. R., Hasan, I., Ahmed, N., Saifuddin, M., & Reza, M. S. (2019). Development and optimization of acyclovir loaded mucoadhesive microspheres by box–Behnken design. *Dhaka University Journal of Pharmaceutical Sciences*, 18(1), 1-12. <https://doi.org/10.3329/dujps.v18i1.41421>
- Kobyas, M., Demirbas, E., Oncel, M. S., Yildirim, Y., Sik, E., Goren, A. Y., & Akyol, A. (2014). Modelling and optimization of arsenite removal from groundwater using al ball anodes by electrocoagulation process. *Journal of Selcuk University Natural and Applied Sciences Online*, 803-811.
- Leiviskä, K. (2013). Introduction to experiment design. University of Oulu Control Engineering Laboratory.
- Mahallati, M. N. (2020). Advances in modeling saffron growth and development at different scales. In *Saffron* (pp. 139-167). *Woodhead Publishing*. <https://doi.org/10.1016/B978-0-12-818638-1.00009-5>

- Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics*, 12(3), 621-625.
- Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). Response surface methodology: Process and product optimization using designed experiments. *John Wiley & Sons*.
- Omoruyi, F. A., Obubu, M., Omeje, I. L., Echebiri, U. V., Onyekwere, K. C., Ngonadi, L. O., and Hamzat, K. I., (2019). Comparison of Some Variable Selection Techniques in Regression Analysis. *American Journal of Biomedical Science & Research*, 6(4).
<https://doi.org/10.34297/AJBSR.2019.06.001044>
- Goos, P., & Jones, B. (2011). Optimal Designs of Experiments: A Case Study Approach, 1st Edition, *Wiley, New York*.
ISBN: 978-0-470-74461-1
- Rao, J. S., & Kumar, B. (2012, September). 3D Blade root shape optimization. In 10th International Conference on Vibrations in Rotating Machinery (pp. 173-188). *Elsevier*.
<https://doi.org/10.1533/9780857094537.4.173>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461-464.
<https://doi.org/10.1214/aos/1176344136>