The integration of response surface method in microsoft excel with visual basic application

H Sofyan1*, Marzuki1, Marlina2 and B Novrizan1

1 Mathematics Department, Syiah Kuala University, Banda Aceh, Indonesia
2 Chemistry Department, Syiah Kuala University, Banda Aceh, Indonesia

*E-mail: hizir@unsyiah.ac.id

Abstract. The purpose of this research is to create modules Response Surface Method (RSM) in Microsoft Excel by using Visual Basic Application programming language. Microsoft Excel has the facility to calculate the statistical problems. Analysis in Microsoft Excel is also relatively complete and procedure more easily and generally Microsoft Excel is available at almost any computer, as part of Microsoft Office. However, Microsoft Excel has not been able to solve problems related to this Response Surface Method. So the finishing will be the existence of a RSM modules in Microsoft Excel. Modules have been made will be applied to find the optimum conditions of membrane polyurethane (PU) in brackish water desalination process. The optimum conditions were determined by the membrane flux and percent rejection. With the RSM modules, it can be determined stationary point and the Response Surface characteristics so it can be used to determine whether the type of stationary point of maximum, minimum or saddle point.

1. Introduction
Advances in technology and science in the present are very fast. Especially in the field of computer software and hardware. With this development, it is almost certain that every human activity is done not manually. For example in taking a conclusion from a study or data collection.

Statistics is the study of how data collection, data analysis, interpretation of data analysis and drawing conclusions. Currently, there are many software to perform data analysis. For example SPSS, Minitab, Eviews and others. In each software has been available applications to complete various calculation methods in statistics one of them is the method of Response Surface (RSM).

Microsoft is known for its outstanding ability to manipulate data. It also has a simple statistical calculation method facility. However, Microsoft Excel does not provide facilities to calculate advanced statistical methods, such as Response Surface method. Therefore, the use of Response Surface Method in Microsoft Excel and Visual Basic Application have been integrated. The application of Response Surface Method to determine optimum condition of polyurethane membrane of brackish water desalination process is presented.
2. Literature and methods

2.1. Linear regression
The relationship between free variables and matched responses on the experiment data is characterized by prediction called regression equation. In simple linear regression, there is only one free variable $X$ and one random bound variable $Y$. The data can be presented as observation pair [1].

Variation analysis will be very helpful to identify the appropriate factors of tested factors. Regression model will help explain the quantitative correlation of the tested factors and response variable observed.

Prior to modeling, an exploratory analysis of the data is often useful as it may highlight interesting features of the data that can be incorporated into a statistical analysis [2]. Regression analysis can help to identify influencing variables. The mean of response regression analysis can be calculated by entering one independent variable [3].

For example, in two-factor case, the regression model can be identified based on the information obtained from variation analysis to determine linear model with interaction or without interaction among the factors. If in the variation analysis is the influence of treatment, the regression model can be built to explain the level of quantitative influence of the treatment, and it can be used to predict further decision making.

In constructing regression model, it requires an area that shows interval of treatment to make code in each treatment to simplify the calculations:

\[
X_t = \frac{N_t - (N_{\text{min}} + N_{\text{maks}})}{2} \cdot \frac{N_{\text{maks}} - N_{\text{min}}}{2}
\]

2.2. Least square method
The determination of the value of regression coefficient is an initial step that must be done in predicting a regression model. One of the methods that can be used to determine regression coefficient is least square (OLS) or smallest squares method. OLS regression is the most common linear model analysis. OLS models are a standard topic and are better known among a wider audience [4].

OLS is one of the method used in making prediction and estimation. It is used to analyze the relationships among variables. Researchers usually use OLS in Regression Analysis and Correlation Analysis. In OLS, the mean response of each constructs were first computed before proceed to the analysis. [5].

2.3. Response surface method (RSM)
RSM is a collection of statistical techniques and mathematics that can be used to improve, develop, and optimize a process. RSM is used to produce dynamic and optimum condition. RSM is the most effective experimental design to achieve such optimization by improving experimental procedures [6].

In statistics, RSM roams the relationship between several free variables and one or more response variable. This method was introduced by GEP Box and KB Wilson in 1951. The main idea of RSM is to use one experiment unit which is designed to get optimal response.

RSM is one of the studies which designs framework and experiment analysis. It consists of several techniques, including:

1. Treatments of some experiments are based on the right selection settings on free variables, which are arranged as in notation $X_1, X_2, ..., X_k$, that generates a response from $Y$. The option of this setting is done by specific criteria aiming to produce information about sufficient and trusted response. The collection of
several constitutional settings from a matrix D with \( n \times k \) order, where \( n \) is the number of few experiments. Matrix D is used as a response surface design.

2. Determine a mathematical model that is the most suitable from data collection on the draft choice.

3. Determine the optimal operating conditions of free variables with maximum results (or minimum) of value response.

The last aspect of RSM can help to determine the best combination of variables that are instrumental to the expected value of \( Y \) [7].

RSM is a combination of mathematics and statistics that is used to design a model and to analysis a \( Y \) response that is affected by some free variables or \( X \) factor in order to optimize the response.

The correlation of variable \( Y \) and \( X \) are:

\[
Y = f(X_1, X_2, \ldots, X_k) + \varepsilon
\]

where:

- \( Y \) = Response variable
- \( X_i \) = Free variables/ factors ( \( i = 1, 2, 3, \ldots, k \) )
- \( \varepsilon \) = Residual

This research uses response surface method to find value of free variables that generate an optimal response.

2.4. Response surface design
Parameter model can be effectively predicted using an appropriate experimental design for data collection. The draft of RSM is called Response Surface Design [8].

The design was selected based on the experimental object and the number of observed factors. Every experiment has a certain purpose. The purpose of response surface method design is to make the experiment possible to estimate interactions or even quadratic effects. For this reason, it is called response surface method (RSM) designs.

2.5. Experimental design phase I
The first order model uses 2k factorial design coupled with an observation on center point as shown I Table 1. The addition to the focal points is required to obtain an estimation of experimental errors. 2k factorial design is experiment design consisting \( k \) factors, and each of them has two levels. The coding of the level of factors can be done using the following equation [9]:

\[
X_i = \frac{\xi_i - \left( \xi_{low} + \xi_{high} \right)}{2} \left( \frac{\xi_{high} - \xi_{low}}{2} \right)
\]

with:

- \( X_i \) = Code Value
- \( \xi_{low} \) = Lowest free variable value
- \( \xi_{high} \) = Highest free variable value
Table 1. Experiential Design of Phase I (k=3)

| No. | X₁  | X₂  | X₃  |
|-----|-----|-----|-----|
| 1   | -1  | -1  | -1  |
| 2   | 1   | -1  | -1  |
| 3   | -1  | 1   | -1  |
| 4   | 1   | 1   | -1  |
| 5   | -1  | -1  | 1   |
| 6   | 1   | -1  | 1   |
| 7   | -1  | 1   | 1   |
| 8   | 1   | 1   | 1   |
| 9   | 0   | 0   | 0   |
| 10  | 0   | 0   | 0   |
| 11  | 0   | 0   | 0   |
| 12  | 0   | 0   | 0   |

The first step of RSM is finding the relationship between Y response and X factor through a first-order polynomial equation and the linear regression model is used, or better known as the first-order model:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \]

with:
- \( Y \) = Response variable
- \( \beta_k \) = Parameter model coefficient
- \( X_k \) = Free variable coding value
- \( \varepsilon \) = Residual

The estimation for first order is:

\[ \hat{Y} = b_0 + \sum_{i=1}^{k} b_i X_i \]

with:
- \( \hat{Y} \) = Estimation value
- \( b_0 \) = Constanta
- \( b_i \) = Estimation Parameter

2.6. Experimental design phase II

There are several types of designs that can be used in RSM, one of them is Central Composite Design (CCD). In this study, CCD is used because this type of design is rotatable. CCD was developed by Box and Wilson (1951). CCD consists of 2k factorial design and observation on the center point and axial points written as \( \delta \) \[10\].

This design is made as in Table 3, with k=3. Meanwhile, some values of k can be seen in Table 2. Various considerations can be used in selecting the value of \( \delta \). However, in general, researchers determine \( \delta \) value based on the following equation:

For full repetition

\[ \delta = \frac{k}{2} \]

For half repetition

\[ \delta = \frac{k-1}{2} \]
Table 2. Rotatable CCD

| K  | $2^k$ | $n_a$ | $n_0$ | N  | Δ   |
|----|-------|-------|-------|-----|-----|
| 2  | 4     | 4     | 8     | 16  | 1,414 |
| 3  | 8     | 6     | 6     | 20  | 1,682 |
| 4  | 16    | 8     | 12    | 36  | 2,000 |
| 5  | 32    | 10    | 17    | 59  | 2,378 |
| 6  | 64    | 12    | 24    | 100 | 2,828 |

Descriptions:
- $K$ = the number of factors
- $n_a$ = Number of axial points
- $n_0$ = Number of repetitions at center point
- $\delta$ = value of axial point

Table 3. Experimental design phase 2 (k=3)

| No. | X₁ | X₂ | X₃ |
|-----|----|----|----|
| 1   | -1 | -1 | -1 |
| 2   | 1  | -1 | -1 |
| 3   | -1 | 1  | -1 |
| 4   | 1  | 1  | -1 |
| 5   | -1 | -1 | 1  |
| 6   | 1  | -1 | 1  |
| 7   | -1 | 1  | 1  |
| 8   | 1  | 1  | 1  |
| 9   | 0  | 0  | 0  |
| 10  | 0  | 0  | 0  |

| No. | X₁ | X₂ | X₃ |
|-----|----|----|----|
| 11  | 0  | 0  | 0  |
| 12  | 0  | 0  | 0  |
| 13  | 0  | 0  | 0  |
| 14  | 0  | 0  | 0  |
| 15  | -δ | 0  | 0  |
| 16  | Δ  | 0  | 0  |
| 17  | 0  | -δ | 0  |
| 18  | 0  | Δ  | 0  |
| 19  | 0  | 0  | -δ |
| 20  | 0  | 0  | Δ  |

Furthermore, for second order model, usually there is a curvature and second order polynomial model with quadratic function:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \sum_{i=1}^{k} \beta_{ii} X_i^2 + \sum_{i=1}^{k} \sum_{j<i}^{k} \beta_{ij} X_i X_j + \varepsilon$$

With:
- $Y$ = Response Variable
- $X_i$ = Free variable coding value
- $\beta_k$ = Parameter model coefficient
- $\varepsilon$ = Residual

Prediction for second-order models is stated in the following equation:

$$\hat{Y} = b_0 + \sum_{i=1}^{k} b_i X_i^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} b_{ij} X_i X_j, i < j$$

With:
- $\hat{Y}$ = Estimation Value
- $b_0$ = Constanta
- $b_i$ = Estimation Parameter
- $X_i$ = Free variable
2.7. Lack of fit model test
Testing of order one and order two models can be done using Lack of Fit Test. The purpose of this test is to find out the suitability of the model produced. If result Lack of Fit Test is not significant, the first order linear model is correct. Meanwhile, if the result is significant, the first order of linear model is incorrect, and the test requires a linear quadratic model or nonlinear model. The hypotheses are:

- **H0**: there is no lack of fit in the model (appropriate model)
- **H1**: there is lack of fit in the model (inappropriate model)

The statistical test is:

\[
F_{\text{ratio}} = \frac{MS_{\text{LOF}}}{MS_{\text{PE}}}
\]

where:
- \( MS_{\text{LOF}} \) = Mean Square lack of fit
- \( MS_{\text{PE}} \) = Mean Square pure error

rejection of \( H0 \) is rejected if \( F_{\text{ratio}} > F(\alpha,n-k-1,ne) \) meaning that there is a discrepancy between the estimation model and the real model. Another way that can be used to detect Lack of Fit is with the value of \( P \), \( H0 \) is rejected if \( P \) value is less than \( \alpha \) (\( P < \alpha \)).

2.8. Simultaneous parameter testing
The hypotheses used in regression parameter testing are as the following.

- **H0**: \( \beta_1 = \beta_2 = \cdots = \beta_k = 0 \)
- **H1**: at least there is one \( \beta_j \neq 0 \); \( j = 1,2,\ldots,k \)

Statistical testing used is:

\[
F_{\text{hitung}} = \frac{MS_R}{MS_E} = \frac{SS_R}{SS_E} \cdot \frac{k}{N-1-k}
\]

where:
- \( SS_R \) = Sum of squares due to regression
- \( SS_E \) = Sum of squares error
- \( N \) = number of observation
- \( k \) = number of parameter
- \( MS_R \) = Mean squares due to regression
- \( MS_E \) = mean squares due to error

\( H0 \) is rejected at the significant level \( \alpha \) if \( F_{\text{count}} > F(\alpha,k,n-1-k) \). It means that statistically \( X_i \) variables give significant contribution in the model or \( H0 \) is rejected if \( P \) value is less than \( \alpha \) (\( P < \alpha \)).

2.9. Determination of stationary point
The second order model that has been formed will be used to determine a stationary point, response surface characteristics and its optimization model. Stationary point (the point in optimum conditions) can be determined from the second order model that is appropriate. Stationary point can be a maximum point, minimum point, or saddle point. For instance, if we want to get the value of \( X_1,X_2,\ldots,X_k \) will optimize the response predicted [8].

If the optimal values exist, it will be determined by \( X_1,X_2,\ldots,X_k \) which are derived from \( \frac{\partial \hat{Y}}{\partial X_1} = \frac{\partial \hat{Y}}{\partial X_2} = \cdots = \frac{\partial \hat{Y}}{\partial X_k} = 0 \). This point is called \( X_1,s,X_2,s,\ldots,X_k,s \), which is named the stationary point. To determine the position of a stationary point, we firstly should use second-order mathematical models in matrix notation as follows:

\[
\hat{Y} = \hat{\beta}_0 + X'b + X'BX
\]
where:

\[
X = \begin{bmatrix}
X_1 \\
X_2 \\
\vdots \\
X_k
\end{bmatrix}, \quad
b = \begin{bmatrix}
\hat{\beta}_1 \\
\hat{\beta}_2 \\
\vdots \\
\hat{\beta}_k
\end{bmatrix}, \quad
B = \begin{bmatrix}
\hat{\beta}_{11} & \hat{\beta}_{12}/2 & \cdots & \hat{\beta}_{1k}/2 \\
\hat{\beta}_{21}/2 & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2k}/2 \\
\vdots & \vdots & \ddots & \vdots \\
\hat{\beta}_{k1}/2 & \hat{\beta}_{k2}/2 & \cdots & \hat{\beta}_{kk}
\end{bmatrix}_{\text{sym.}}
\]

While \(b\) is the matrix \((k \times 1)\) vector of first-order regression coefficients and \(B\) is symmetric \((k \times k)\) whose diagonal elements are pure quadratic coefficient \((\hat{\beta}_{kk})\) and the rest is half of the quadratic coefficient mix. \((\hat{\beta}_{ij}, i \neq j)\).

The derivative of \(Y\) to vector \(X\) is equal to 0, so it is presented by:

\[
\frac{\partial Y}{\partial X} = b + 2BX = 0
\]

The points of stationary are solution of the above equation, namely:

\[
X_s = -\frac{1}{2}B^{-1}b
\]

Which, \(X_s = (X_1.s, X_2.s, \ldots, X_k.s)\). Substitution of the equation of stationary point to the equation of matrix notation obtains optimal response value which is predicted to occur at the points of stationary, namely:

\[
\hat{Y}_s = \hat{\beta}_0 + \frac{1}{2}X_s'b
\]

2.10 Visual Basic Application (VBA)

Microsoft Visual Basic for Applications (VBA), which was released in 1993, is a derivative of Visual Basic programming language developed by Microsoft or an integrated combination between programming environments (Visual Basic Editor) and programming languages (Visual Basic). It allows users to design and build a Visual Basic program in the Microsoft Office applications intended for specific applications. VBA is designed to perform multiple tasks, such as customizing some applications like Microsoft Office or Microsoft Visual Studio.

The function of VBA is to automate the work which involves repeated and complex operation. VBA is different from Microsoft Visual Basic. Microsoft Visual Basic provides many advanced programming functions so that Microsoft Visual Basic can produce more complex program for the Microsoft Windows operating system or Office. While VBA only be built on the main Microsoft Office applications to control application functions that perform a series of programmed objects. VBA is a programming language implemented in the Microsoft Office environment designed for creating simple applications at the level of desktop applications. It can be used to create user-defined functions or simple macros. [11].

3. Result and discussion

The steps in creating RSM applications using Visual Basic Application in Microsoft Excel can be seen from the following flow diagrams:
Figure 1. Diagram of Work Steps Creating RSM using VBA in Microsoft Excel

4. Conclusions
According to the results and the above discussion, it can be concluded that RSM can be solved using Microsoft Excel 2007 by using VBA module. This RSM module can also be used in Microsoft Excel 2003. Finally, with this extension, RSM problems can be solved easily. In this study, it is applied to determine the value of the independent variables that cause the value of a response variable in brackish water desalination process be optimal.

The results for optimization flux response is obtained optimum condition at reaction temperature of 46,5780 °C, pressure of 2,136 atm and reaction time of 22.326 minutes. Meanwhile, In optimizing
percentage of rejection response, it is obtained that optimum condition is at reaction temperature of 74.7850 °C, pressure of 1.588 atm and a reaction time of 24.634 minutes.

References
[1] Walpole R E and Myers R H 1989 Probability and Statistics for Engineers and Scientists, Fourth Edition (New York: Macmillan)
[2] Kuhnert P and Bill V 2005 An Introduction to R: Software for Statistical Modelling and Computing (CSIRO Australia: Redland City)
[3] Sofyan R and Marzuki 2012 Proc. Int. Conf. on the Symposium of Japanese Society of Computational Statistics 26 pp 161
[4] Pohlman J and Leitner D 2003 The Ohio Journal of Science 103 118
[5] Nazim A and Ahmad S 2013 International Journal of Scientific and Engineering Research 4 717
[6] Silva V, Rouboa A 2012 Applied Mathematics and Computation 218 6733
[7] Khuri A I 2003 Advanced Calculus with Application in Statistics (New Jersey: John Wiley and Sons)
[8] Montgomery D C 2001 Design and Analysis of Experiments 5th edition (New York: John Wiley and Sons)
[9] Khuri A I and Cornell J A 1996 Empirical Model-Building and Response Surface (New York: Marcell Dekker)
[10] Buyske S and Richard T 2009 Advanced Design of Experiments: Response Surface Methodology II (New Jersey: Rutgers University)
[11] Dydowicz P 2016 Journal of Software and Systems Development 2016 1