Thailand Statistician
July 2022; 20(3): 545-561
http://statassoc.or.th
Contributed paper

Comparison of Listwise Deletion and Imputation Methods for Handling a Single Missing Response Value in a Central Composite Design

Wannaporn Junthopas [a] and Chantha Wongoutong [b]*
[a] Department of Statistics, Faculty of Science, Khon Kaen University, Khon Kaen, Thailand.
[b] Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand.
*Corresponding author; e-mail: fscictw@ku.ac.th

Received: 2 May 2021
Revised: 17 July 2021
Accepted: 25 July 2021

Abstract

This research aims to investigate appropriate methods for handling missing data during analysis, which is one of the most challenging tasks for statistical inference. Our motivation is to replace a missing response value in a central composite design (CCD) and its effect on each particular part (factorial, center and axial) in which the value is missing. Statistical software packages generally set listwise deletion as the default method for dealing with missing data, while imputation methods are also widely used. Hence, we compared listwise deletion and mean and regression imputation. Four test functions were used to examine all possible cases of a single missing response in a CCD with two factors. The performances of the methods for handling a missing response value in each of the three parts of the CCD (factorial, center, or axial) were compared in terms of their optimal responses with complete data by using mean-squared error and correlation coefficient values. Regression imputation and listwise deletion provided similar results for handling the missing value in each of the CCD parts (factorial, center, and axial) and were both superior to mean imputation.

Keywords: Response surface methodology, factorial points, center points, axial points.

Introduction

The central composite design (CCD) is the most popular and widely used design for estimating the second-order response surface method (RSM) introduced by Box and Wilson (1951). CCD is applied to determine the operating variables' optimized values by fitting a second-order model (Kumar et al., 2009). This design is an optimization technique widely used in many fields, such as chemistry, environmental studies, and engineering (Azami et al. 2013; Gano et al. 2015; Momen et al. 2016; Farzadkia et al. 2018; Bagheri et al. 2019) because of the advantage of optimizing multifactor problems with the optimal number of experimental runs.

Missing observations in real-world experiments are not uncommon, even in a well-planned experiment or a well-controlled study. The problem of missing data can significantly affect the statistical power reduction and produce biased estimates, thereby leading to invalid conclusions

(Graham 2009; Ayilara et al. 2019). Most often, listwise deletion (or complete case analysis) is the default method for dealing with missing data in almost all statistical software packages (Briggs et al. 2003), and it may or may not be a bad choice depending on the cause and how many data items are missing (Newman, 2014). Several studies have shown that for specific analyses of a particular dataset, subpopulation analysis can lead to results that diverge from those obtained through listwise deletion (Graubard and Korn 1996; Roth 1994; Roth et al. 1999). Many studies involve missing data, and listwise or pairwise deletion is most often used to handle the missing data when they comprise less than 5% of the total (Rubin 1987; Schafer 1997). Likewise, Bengtsson et al. (2021) who claim that listwise deletion performs best for MCAR data and when the proportion of missingness is not too high. The major disadvantage of listwise deletion is that it regularly removes a large proportion of the sample, thereby leading to loss of the statistical power of the particular test being used (Allison 2001). Listwise deletion works well when the data are missing completely at random (MCAR), which rarely happens in reality (Nakai and Weiming 2011). Meanwhile, missing at random (MAR) allows the probability of missingness to depend on observed variables, which means that multiple imputation and maximum likelihood (ML) methods have a major advantage over listwise deletion in reducing bias. Unfortunately, most researchers do not know that listwise deletion may be less biased than multiple imputation or ML when data are missing on variables in regression analysis under certain circumstances (Allison 2001). Another common approach used to handle the missing values is imputation, which replaces missing values with substituted ones, thereby leading to more accurate analysis. Several imputation methods for imputing the missing values have been used (Engels and Diehr 2003; Saunders et al. 2006; Junger and De Leon 2015) such as mean, regression, hot-deck, K-nearest neighbor, etc.

Even though listwise deletion is the default method for dealing with missing data in most statistical software packages, the effect of using it to handle missing values in CCDs has not yet been clarified. Therefore, we examined the impact of a single missing response value in the various parts of the CCD (factorial, center, axial) and handling it using three methods (mean, regression, and listwise deletion). Whereas the first two methods impute the missing value, listwise deletion deletes it, after which CCD is performed with the rest of the dataset. After that, the performances of the three methods in each part of the CCD were compared.

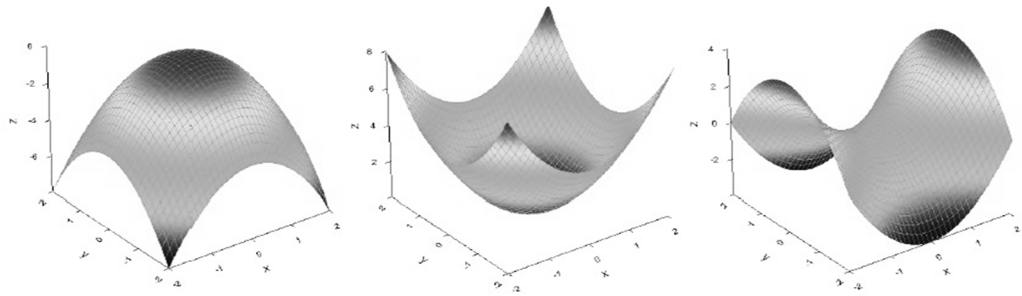
The rest of the paper is organized as follows. The RSM, which is the basic concept of CCD, and the three methods for handling missing values are presented in Section 2. Next, details of a simulation study used to investigate the performances of the three methods are reported in Section 3. After that, the results of the simulation study and a discussion are covered in Section 4. Finally, conclusions and recommendations are presented in Section 5.

2. Methodology and Framework

2.1. RSM

Originally, RSM was developed for experimental model responses (Box and Draper, 1987) and then adapted for numerical experiment modeling. The primary idea is to fit a model for the response variable and explore various settings for the explanatory variables of interest. One of these is maximizing (or minimizing) the mean value of the response variable. In general, the relationship between the response variable of interest and the independent variables is unknown and usually approximated by applying a low-degree polynomial model of the form (Myers and Montgomery, 2002). The first step in RSM is to find a suitable approximation for the true relationship. If curvature is detected near the optimum of the first-order model (suggesting poor model fitting), points are added to obtain a second-order model. CCD is the most popular choice for fitting a second-order model in

RSM. It can fit a second-order prediction equation for the response where the quadratic terms model the curvature of the true response function. Different types of response surfaces are shown in Figure 1.



(a) (b) (c)

Figure 1 The different types of response surfaces: (a) maximum, (b) minimum, and (c) saddle

If a maximum or minimum exists inside the factor region, then RSM can estimate it. A second-order model can be formulated as

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \varepsilon, \tag{1}$$

where $\underline{y} \sim N(\boldsymbol{X}\underline{\beta}, \sigma^2\mathbf{I})$ and $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2\mathbf{I})$.

2.2. CCD

CCDs, the most commonly used models for RSM (Myers and Montgomery 2002) can be efficiently constructed with a second-order model. This design contains a factorial or fractional factorial design with center points augmented with a group of star points (axial points) that allow estimation of the curvature. CCD consists of 2^k factorial runs (coded as the usual ± notation) with 2^k 2^k axial runs $(\pm\alpha, 0, 0, \ldots, 0), (0, \pm\alpha, 0, \ldots, 0), \ldots, (0, 0, \ldots, \pm\alpha)$ and center runs (n_c replicates, $(0, 0, \ldots, 0)$, where the axial point is determined by using α = (the number of factorial runs)$^{1/4}$ for 2^2 factorial design $\alpha = 1.414$. The total number of experiments (N) performed for a CCD is determined when the factorial design is full: i.e., $N = 2^k + 2k + n_c$ points. The CCD for two factors with five center points is provided in Table 1.

The four important steps in executing a CCD are: (1) perform statistically designed experiments. (2) estimate the coefficients. (3) predict the optimal response. (4) check the model's adequacy. After the parameters of the second-order model have been estimated by the ordinary least-squares method, the optimal design point in terms of the coded variables can be written as (Anderson et al. 2009)

$$\mathbf{x}_{opt} = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b}, \tag{2}$$

where $\mathbf{b} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$ and $\mathbf{B} = \begin{pmatrix} \hat{\beta}_{11} & \hat{\beta}_{12}/2 & \cdots & \hat{\beta}_{1k}/2 \\ \hat{\beta}_{12}/2 & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2k}/2 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\beta}_{1k}/2 & \hat{\beta}_{2k}/2 & \cdots & \hat{\beta}_{kk} \end{pmatrix}$.

Table 1 The layout of a CCD for two factors with five center points.

Coded Variables		Response
x_1	x_2	y_i
−1	−1	y_1
1	−1	y_2
−1	1	y_3
1	1	y_4
0	0	y_5
0	0	y_6
0	0	y_7
0	0	y_8
0	0	y_9
−1.414	0	y_{10}
1.414	0	y_{11}
0	−1.414	y_{12}
0	1.414	y_{13}

Meanwhile, the predicted response at the optimal design point is given by

$$\hat{y}_{opt} = \hat{\beta}_0 + \frac{1}{2}\mathbf{x}'_{opt}\mathbf{b} \tag{3}$$

which is converted back in terms of the true variables.

Incomplete data increases the risk of weakening the inference validity and, in the real world, data are often incomplete. For a single missing point in a CCD, the part in which the design point is missing must be considered: (1) a factorial or fractional factorial design, (2) center points, and (3) a group of star points (axial points). Table 2 provides the possible cases for a single missing response value in a CCD with two factors.

Table 2 The possible cases for a single missing response value in a CCD with two factors

Case	Factorial				Center					Axial			
	F1	F2	F3	F4	C1	C2	C3	C4	C5	A1	A2	A3	A4
1	M	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}
2	y_1	M	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}
3	y_1	y_2	M	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}
4	y_1	y_2	y_3	M	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}
5	y_1	y_2	y_3	y_4	M	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}
6	y_1	y_2	y_3	y_4	y_5	M	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}
7	y_1	y_2	y_3	y_4	y_5	y_6	M	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}
8	y_1	y_2	y_3	y_4	y_5	y_6	y_7	M	y_9	y_{10}	y_{11}	y_{12}	y_{13}
9	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	M	y_{10}	y_{11}	y_{12}	y_{13}
10	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	M	y_{11}	y_{12}	y_{13}
11	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	M	y_{12}	y_{13}
12	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	M	y_{13}
13	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	M

y_{ij} are observed data; M is the missing value.

2.3. Missing data mechanisms

Descriptions of missing data mechanisms were first introduced by Rubin (1976), MCAR when the probability of missing is equal for all cases, MAR when the probability of missing is equal only within groups defined in the observed data, and missing not at random (MNAR) when the probability of missing varies for unknown reasons.

2.4. Methods for handling missing values

Imputation is used to replace missing values with substituted ones, thereby leading to a more efficient statistical analysis. An appropriate imputation method for missing data in a CCD depends on which part contains the missing values (factorial, center, axial)). Listwise deletion can be used for a single missing response value in the various CCD design points (factorial, center, axial). The following three methods were used for handling a single missing response value in a CCD.

2.4.1. Mean imputation

As per its name, mean imputation creates a replacement value for the missing value from the mean of the available cases. This method is easy to use but reduces variability in the data, leading to underestimating the standard deviation and variance. The missing response value at can be calculated by using Mean Imputation as follows:

$$\hat{y}_{MI(k)} = \frac{\sum_{\forall j \neq k} y_j}{12}, \text{ where } k \in j = \{1,2,3,...,13\}. \tag{4}$$

2.4.2. Regression imputation

This is used to predict the missing value in a variable by using a regression model. In other words, the available information for complete and incomplete cases is used to predict the missing value for a specific variable. The fitted value is then used to impute the missing value. Ostertagová (2012) described how a polynomial regression model can be useful when the relationship between two variables is curvilinear. Therefore, regression imputation can be applied to impute the missing response value in a CCD by using an estimated regression model of the second-order using the ordinary least-squares method. A suitable polynomial regression model with $k = 2$ predictor variables of the second-order is

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \varepsilon, \tag{5}$$

where $\underline{y} \sim N(\boldsymbol{X}\underline{\beta}, \sigma^2\mathbf{I})$ and $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2\mathbf{I})$. The estimation of the missing response value is calculated as follows

$$\hat{y}_{Rgl(k)} = b_0 + \sum_{i=1}^{2} b_i x_i + \sum_{i=1}^{2} b_{ii} x_i^2 + \sum_{i<j} \sum b_{ij} x_i x_j.$$

2.4.3. Listwise deletion

Listwise deletion (or complete case analysis) is the default method for dealing with missing data in almost all statistical software packages (Briggs et al. 2003). Quite simply, cases with missing data on the variable of interest are deleted; e.g., for a missing response value at design point F1 corresponding to, the data point for F1 is deleted from the dataset and CCD is performed in four steps by using the rest of the dataset. The process to perform CCD by using the handling method by listwise deletion is illustrated in Figure 2.

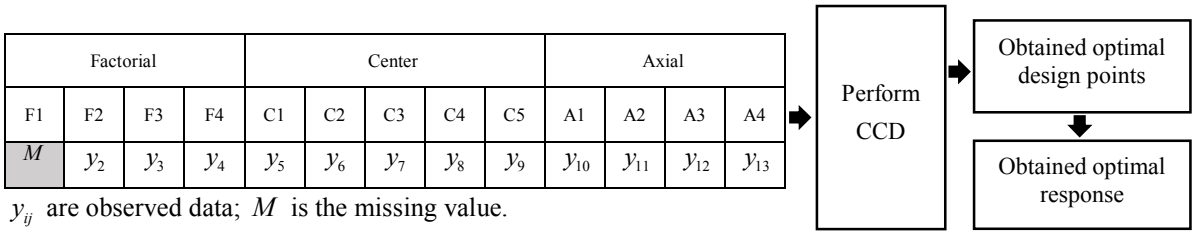


Figure 2 The process to perform CCD by using the handling method by listwise deletion

2.5. The performance metrics

Two performance metrics were used to verify the efficiency of the three methods for handling missing values in a CCD: the mean-squared error (MSE) and correlation (r) respectively defined as

$$MSE = \frac{\sum_{i=1}^{n}\left(y_{opt(i)} - \hat{y}_{opt(i)}\right)^2}{n}, r = \frac{n\sum y_{opt(i)}\hat{y}_{opt(i)} - \sum y_{opt(i)}\sum \hat{y}_{opt(i)}}{\sqrt{\left[n\sum y_{opt(i)}^2 - \left(\sum y_{opt(i)}\right)^2\right]\left[n\sum \hat{y}_{opt(i)}^2 - \left(\sum \hat{y}_{opt(i)}\right)^2\right]}},,$$

where $y_{opt(i)}$ is the optimal response of the complete data at iteration i, $\hat{y}_{opt(i)}$ is the optimal response with the data handling method at iteration i, and n is the number of iteration.

3. Simulation Study

3.1. Test functions

Four test functions from Wongoutong et al. (2017) were used in the simulation to compare the efficiencies of the three methods (mean and regression imputation, and listwise deletion) for handling a single missing value from one of the three parts of a CCD. The four test functions are defined as follows:

$$f_1(x_1,x_2) = -(x_1^2 + x_2 - 11)^2 - (x_1 + x_2^2 - 7)^2 \; ; x_1, x_2 \in [-2,2],$$

$$f_2(x_1,x_2) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4 \; ; x_1 \in [-1,0.5], x_2 \in [0,1],$$

$$f_3(x_1,x_2) = 1431 - 7.81x_1 - 13.3x_2 + 0.0551x_1^2 + 0.0401x_2^2 - 0.01x_1x_2; \; x_1 \in [50,120], x_2 \in [150,200],$$

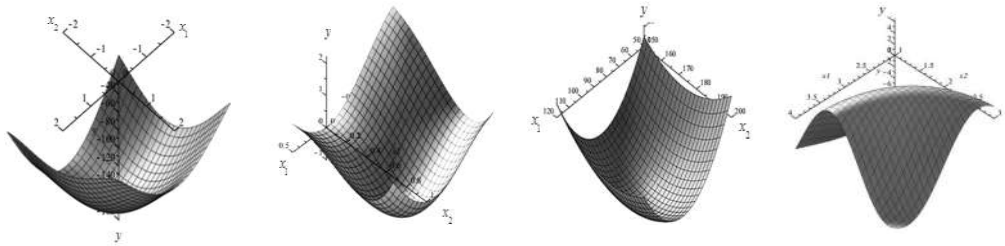
$$f_4(x_1,x_2) = -2 - 0.01(x_2 - x_1^2)^2 - (1 - x_1) - 2(2 - x_2)^2 - 7\sin\left(\frac{x_1}{2}\right)\sin\left(\frac{7x_1x_2}{10}\right) \; ; x_1 \in [2,4], x_2 \in [1,3].$$

The first three test functions are minimized and the last is maximized. The characteristics of the four test functions with minimum and maximum points and true optimum responses are summarized in Table 3. Moreover, the response of the four test functions is illustrated in Figure 3.

Table 3 Characteristics of the test functions

Test Function	True Optimal Point		True Optimal Response
	x_1	x_2	y
f_1	−0.270	−0.920	−181.600
f_2	−0.092	0.713	-1.032
f_3	86.900	176.670	-83.220
f_4	3.200	2.100	6.510

The random generators associated with the corresponding four test functions ($f_i, i = 1,2,3,4$) were $\varepsilon_1 \sim N(0, 0.1^2)$, $\varepsilon_2 \sim N(0, 0.01^2)$, $\varepsilon_3 \sim N(0,1)$, and $\varepsilon_4 \sim N(0, 0.05^2)$, where the variances of the random generators were specified by considering the corresponding ranges of the independent variables in each test function. These can be restricted in different ranges, so the range of independent variables corresponds to $f_i, i = 1,2,3,4$. Therefore, the simulated response was evaluated at a specific point for each test function.



(a) (b) (c) (d)

Figure 3 The response surfaces of test functions (a) f_1, (b) f_2, (c) f_3, and (d) f_4

3.2. The RSM with CCD simulation study

In the simulation study, the four stochastic test functions based on CCD were set with a missing response value and 100 trials using each of the three methods for dealing with missing values (mean and regression imputation, and listwise deletion). For f_1 to f_4, L_{ij} is the low level and H_{ij} the high level of x_i in the j^{th} replication of a simulation experiment generated by using uniform generator $U[a_i, b_i]$, for $i = 1,2;\ j = 1,2,...,100$.

1. For analyzing the factorial design in the CCD, $L(-)$ and $H(+)$ are the low and high levels of the natural variables, respectively, while C_{ij} is the center point in a 2^2 factorial design, where C_{ij} is given by $(L_{ij} + H_{ij})/2$ for $i = 1,2;\ j = 1,2,\ldots,100$. Subsequently, 100 trials of the CCD were performed using L_{ij} and H_{ij} for the two factors in each of the test functions (f_1 to f_4).

2. In each trial, complete data were obtained using the stationary points of the two factors and the optimal response for each test function (f_1 to f_4) by using RSM with CCD simulation.

3. For each test function (f_1 to f_4), a single missing point was set in all possible 13 positions in the CCD by using L_{ij} and H_{ij}. For example, CCD was set as $j = 1$ and the RSM with CCD was conducted with each L_{ij} and H_{ij}. Afterward, a single missing point was set in all possible 13 positions in the CCD.

4. The three data handling methods (mean and regression imputation, and listwise deletion) were used to deal with the missing value. The MSE and correlation values of the optimal responses with complete and treated data were used to measure the performances of the three data handling methods.

After the three methods had been applied, RSM with CCD was conducted to obtain the stationary points of the two factors and the optimal response of each test function (f_1 to f_4). The R programming language version 4.0.3 was used to perform all of the analyses. A flow chart of experimental study steps is shown in Figure 4.

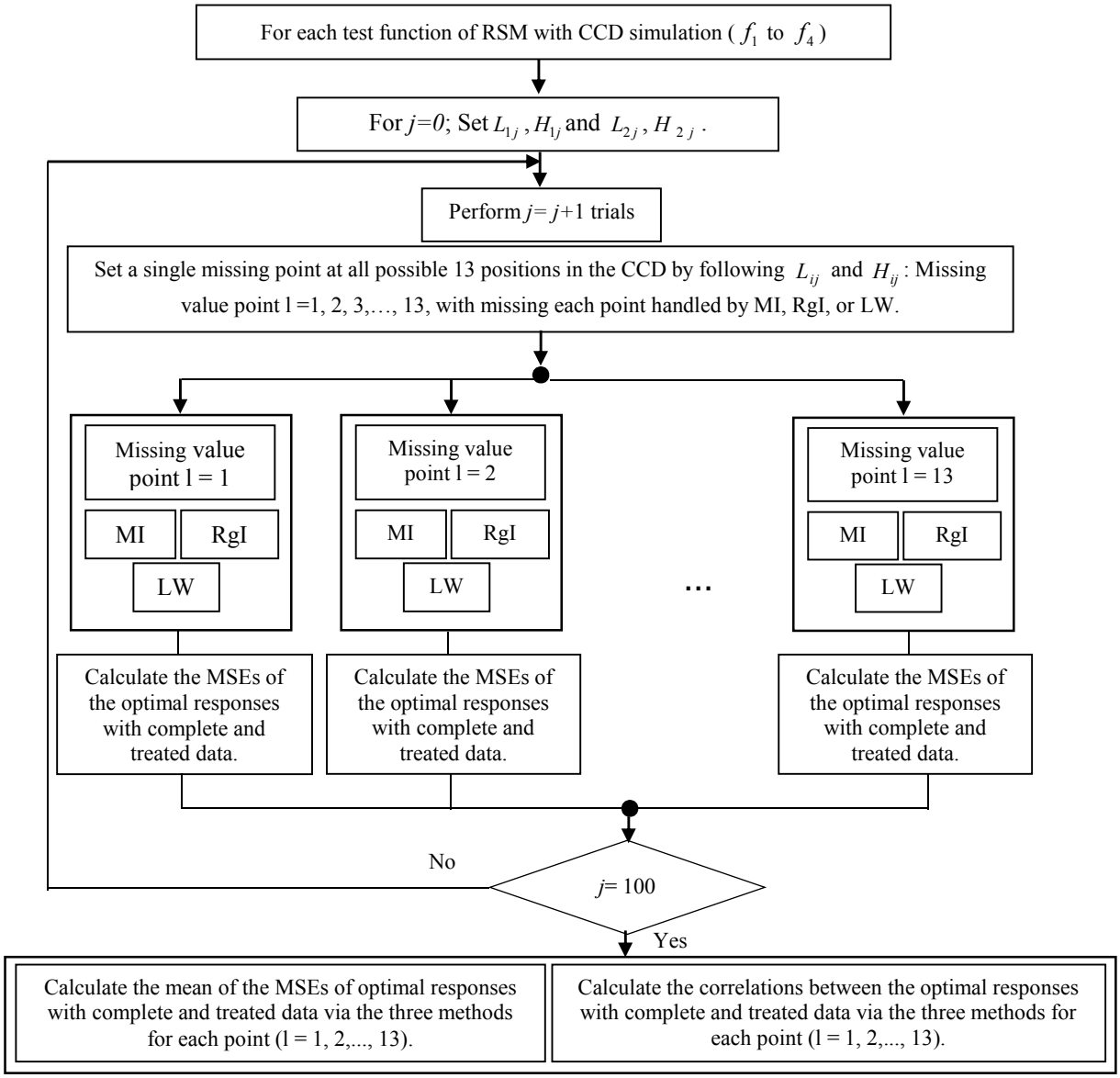


Figure 4 A flowchart of the steps to evaluate the performance of the three methods. MI, mean imputation; RgI, regression imputation; LW, listwise deletion

4. Results and Discussion

4.1. The optimal responses with complete and treated data via the three methods

The performances of the three methods for handling missing values were assessed by comparing their optimal response values with that obtained with complete data. Four test functions were used for the CCD with a single missing point of 13 possible patterns with 100 trials in each. After that, the optimal responses with complete and treated data via the three methods were compared for each case. The results with complete data in the CCD are summarized in Table 4. For example, for f_1, the mean optimal factors of x_1 and x_2 were −0.356 and −0.958 while the mean and standard deviation (SD) of the optimum response were −181.623 and 2.987, respectively. Besides, it was found that the lower

confidence limit (LCL) was −182.215, and the upper confidence limit (UCL) was −179.030 for the 95% confidence interval of the optimal response mean from complete data. Meanwhile, the same trend was evident for f_2 to f_4.

Table 4 The results for the CCD with complete data

Test function	The Mean for the		The Mean and SD of the Optimal Response	95% CI of the Mean Optimal Response
	x_1	x_2		
f_1	−0.356	−0.958	−180.623 (2.987)	[−182.215, −179.030]
f_2	−0.117	0.706	-0.996 (0.063)	[−1.012, −0.980]
f_3	86.877	176.582	-83.186 (0.665)	[−83.753, -82.619]
f_4	3.146	2.113	6.429 (0.085)	[6.420, 6.497]

Note: The number in the parenthesis is the standard deviation of optimum response in compete data, CI, confidence interval.

Table 5 reports the mean optimal responses for the three data handling methods for f_1 to f_4 and each missing point case. Moreover, Figure 5 shows plots between the 95% confidence intervals (CIs) of the mean optimal responses of the three methods for f_1 to f_4 and each missing point case. For example, for f_1 (Figure 5(a)), the 95% CI of the mean optimal response had an LCL of -182.215 and a UCL of −179.030, while the mean optimal responses after using listwise deletion, and regression and mean imputation to handle the missing point at F1 were −183.278, −183.278, and −174.327, respectively, none of which were within the 95% CI of the mean optimal response with complete data. For the missing point at F2, the mean optimal responses after applying listwise deletion, and regression and mean imputation were −180.385, −180.385, and −181.866, respectively. This time, those of listwise deletion and regression imputation were within the 95% CI of the mean optimal response from complete data whereas that with mean imputation was not. This trend was also apparent for missing points at F3 and F4, C1-C5, and A1-A4. Figure 5(a)-(d) clearly shows that both listwise deletion and regression imputation attained similar results for the mean optimal response values for all 13 missing point positions (F1-F4, C1-C5, and A1-A4) of f_1 to f_4.

The percentages of the mean optimal responses for data handling by each method included within the 95% CI of the mean optimal response with complete data in each part of the CCD for f_1 to f_4 are summarized in Table 6. It can clearly be seen that the results for listwise deletion and regression imputation for handling a missing point in all three parts of the CCD were similar. For example, in the center part of f_1, 100% of the optimal responses after data handling by listwise deletion and regression imputation and 0% using mean imputation were within the 95% CI of the mean optimal response with complete data. In the partial factorial of f_1, 75% of the optimal responses for both listwise deletion and regression imputation and 0% for mean imputation were within the 95% CI of the mean optimal response with complete data. In the axial of f_1, 50% of the optimal responses for both listwise deletion and regression imputation and 0% for mean imputation were within the 95% CI of the mean optimal response with complete data. Overall, 76.92%, 76.92%, and 0% of the mean optimal responses in the factorial part by listwise deletion, and regression and mean imputation, respectively, were within the 95% CI of the mean optimal response with complete data.

Table 5 The results with 100 trials of handling a single missing design point in the CCD using the three methods for four test functions

Design Point Missing Value	Methods	Mean Optimal Response			
		f_1	f_2	f_3	f_4
F1	LW	−183.278	−1.048	−83.257*	6.445*
	RgI	−183.278	−1.048	−83.257*	6.445*
	MI	−174.327	−0.871	−98.853	6.524
F2	LW	−180.385*	−0.988*	−83.196*	6.502
	RgI	−180.385*	−0.988*	−83.196*	6.466*
	MI	−181.866	−0.986*	−82.743*	6.503
F3	LW	−180.524*	−1.003*	−83.234*	6.451*
	RgI	−180.524*	−1.003*	−83.234*	6.451*
	MI	−179.417	−0.995*	−82.404	6.453*
F4	LW	−181.007*	−0.967	−83.245*	6.470*
	RgI	−181.007*	−0.967	−83.245*	6.470*
	MI	−164.100	−0.919	−95.271	6.319
C1	LW	−180.644*	−0.996*	−83.221*	6.458*
	RgI	−180.644*	−0.996*	−83.221*	6.458*
	MI	−181.801	−0.907	−81.584	5.883
C2	LW	−180.620*	−0.996*	−83.205*	6.459*
	RgI	−180.620*	−0.996*	−83.205*	6.459*
	MI	−181.725	−0.907	−81.566	5.821
C3	LW	−180.623*	−0.996*	−83.239*	6.457*
	RgI	−180.623*	−0.996*	−83.239*	6.457*
	MI	−181.740	−0.907	−81.608	5.903
C4	LW	−180.600*	−0.996*	−83.206*	6.459*
	RgI	−180.600*	−0.996*	−83.205*	6.459*
	MI	−181.761	−0.907	−81.575	5.894
C5	LW	−180.643*	−0.996*	−83.211*	6.459*
	RgI	−180.643*	−0.996*	−83.211*	6.459*
	MI	−181.812	−0.907	−81.587	5.902
A1	LW	−181.610	−1.004*	−83.233*	6.462*
	RgI	−181.610	−1.004*	−83.233*	6.462*
	MI	−174.110	−1.002*	−81.967	6.531
A2	LW	−181.513*	−1.006*	−83.257*	6.445*
	RgI	−181.513*	−1.006*	−83.257*	6.445*
	MI	−179.542	−1.025	−98.853	6.524
A3	LW	−180.332*	−1.071	−83.196*	6.502
	RgI	−180.332*	−1.071	−83.196*	6.466*
	MI	−190.660	−0.951	−82.743*	6.503
A4	LW	−178.113	−0.975	−83.234*	6.451*
	RgI	−178.113	−0.975	−83.234*	6.451*
	MI	−171.776	−0.864	−82.404	6.453*

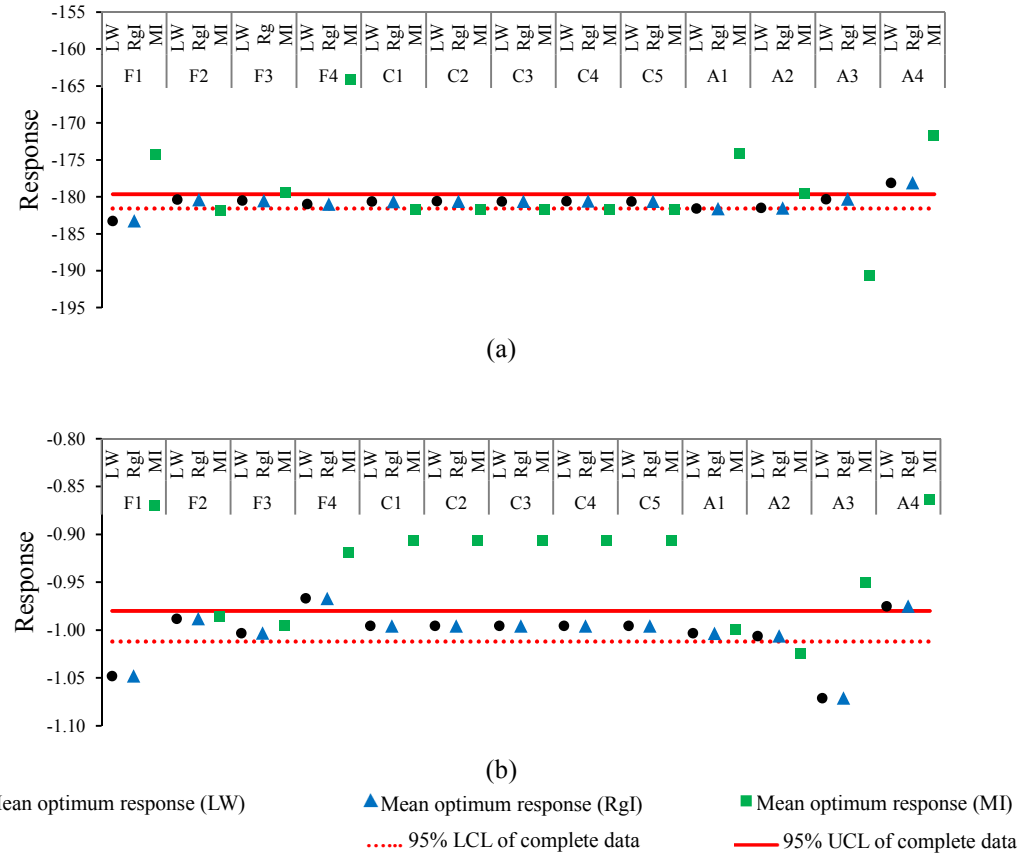
*The mean optimal response is within the 95% CI of the mean optimal response with complete data. MI, mean imputation; RgI, regression imputation; LW, listwise deletion.

The same results trend was apparent for f_2 to f_4. These results support that listwise deletion and regression imputation had similar optimal response values for a single missing value in all three parts (factorial, center and axial) of the CCD and outperformed mean imputation in all cases.

Table 6 The percentages of the mean optimal responses of each method in each part of the CCD included in the 95% CI of the mean optimal response with complete data for f_1 to f_4

Part CCD	Test Function 1 (f_1)			Test Function 2 (f_2)			Test Function 3 (f_3)			Function 4 (f_4)		
	LW	RgI	MI	LW	RgI	MI	LW	RgI	MI	LW	RgI	MI
Factorial	75	75	0	50	50	50	100	100	25	75	100	25
Center	100	100	0	100	100	0	100	100	0	100	100	0
Axial	50	50	0	50	50	25	100	100	25	75	100	25
Overall	76.92	76.92	0	69.23	69.23	23.08	100	100	15.38	84.62	100	15.38

MI, mean imputation; RgI, regression imputation; LW, listwise deletion.



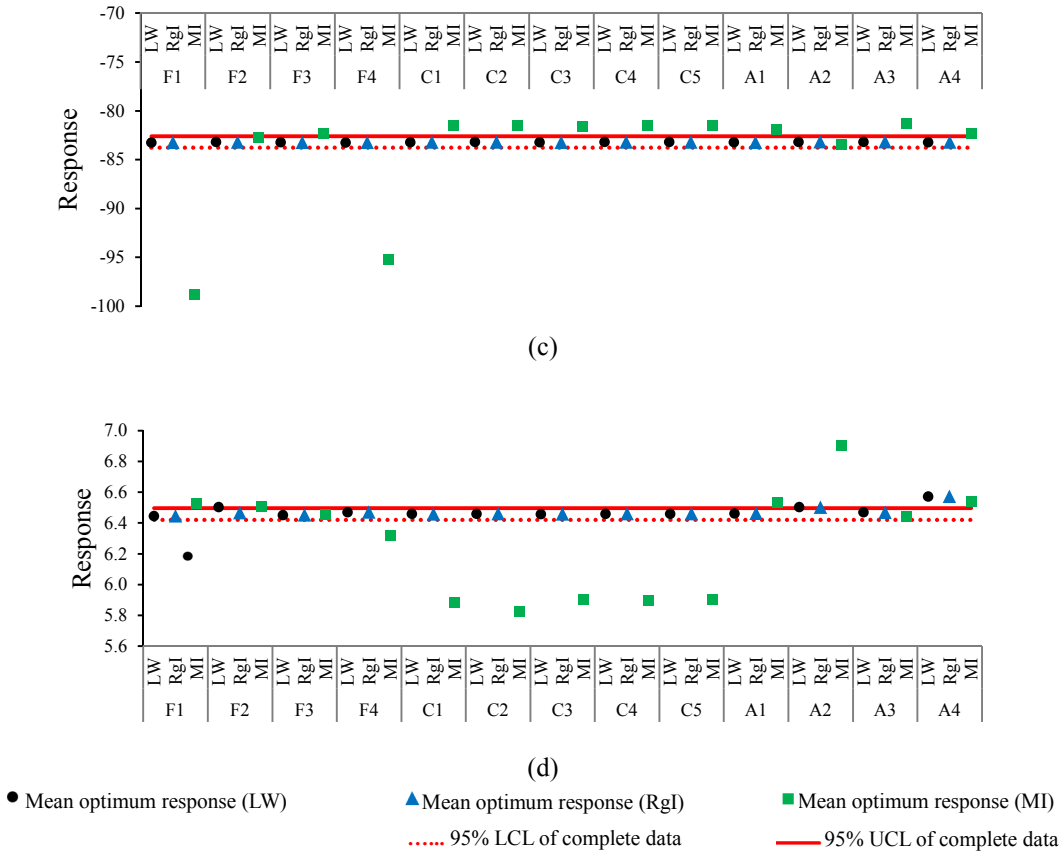


Figure 5 The plots between 95% CI of the mean optimal response with complete data and mean optimal response by three methods for (a) f_1, (b) f_2, (c) f_3 and (d) f_4 in each part of the CCD.

4.2. The performances of the three methods for handling a missing value in the CCD

The performance results for the three methods are summarized for each part of the CCD (factorial, center, and axial) in Table 7. For example, for f_1, the overall of mean and SD of the optimal response with complete data were −180.623 and 2.987, respectively. Meanwhile, the means for the optimal responses after handling the missing value in the CCD factorial part using mean and regression imputation, and listwise deletion were −174.932, −181.299, and −181.299, respectively, and the SDs were 45.667, 3.547, and 3.547, respectively. Thus, regression imputation and listwise deletion for a missing value in the factorial part of the CCD outperformed the mean imputation method. Similarly, in the axial part of the CCD, the overall mean optimal responses after imputing the missing value using the mean, regression, and listwise deletion methods were −183.772, −180.392, and −180.392, respectively, and the SDs were 31.056, 2.934, and 2.934, respectively. Once again, regression imputation and listwise deletion outperformed mean imputation. For a missing value in the center part of the CCD, the mean optimal responses of the mean and regression imputation, and listwise deletion were −181.768, −180.626, and −180.626, respectively, and the SDs were 7.074, 2.993, and 2.993, respectively. In this case, regression imputation and listwise deletion attained almost the same results and outperformed mean imputation, thereby supporting the results in Tables 5 and 6.

Table 7 The results of 100 trials of CCD with handled missing values by three methods

Test function	Optimum response with Complete data: Mean (SD)	Methods	Overall optimum response after handling the missing value					
			Factorial (F1-F4)		Center (C1-C5)		Axial (A1-A4)	
			Mean	SD	Mean	SD	Mean	SD
f_1	−181.623 (2.987)	MI	−174.932	45.667	-181.768	7.074	-183.772	31.056
		RgI	−181.299	3.547*	-180.626	2.993*	-180.392	2.934*
		LW	−181.299	3.547*	-180.626	2.993*	-180.392	2.934*
f_2	−0.996 (0.063)	MI	−0.943	0.186	-0.907	0.110	−0.960	0.187
		RgI	−1.002	0.089*	-0.996	0.063*	−1.014	0.100*
		LW	−1.002	0.089*	-0.996	0.063*	−1.014	0.100*
f_3	−83.186 (0.665)	MI	−86.099	119.32	-81.584	2.524	−82.285	6.047
		RgI	−83.233	0.769*	-83.216	0.689*	−83.207	0.716*
		LW	−83.233	0.769*	-83.216	0.689*	−83.207	0.716*
f_4	6.459 (0.085)	MI	6.450	0.443	5.881	0.596	6.603	1.316
		RgI	6.458	0.092*	6.459	0.086*	6.501	0.139*
		LW	6.458	0.092*	6.487	0.086*	6.501	0.139*

*The best performance in terms of SD. MI, mean imputation; RgI, regression imputation; LW, listwise deletion.

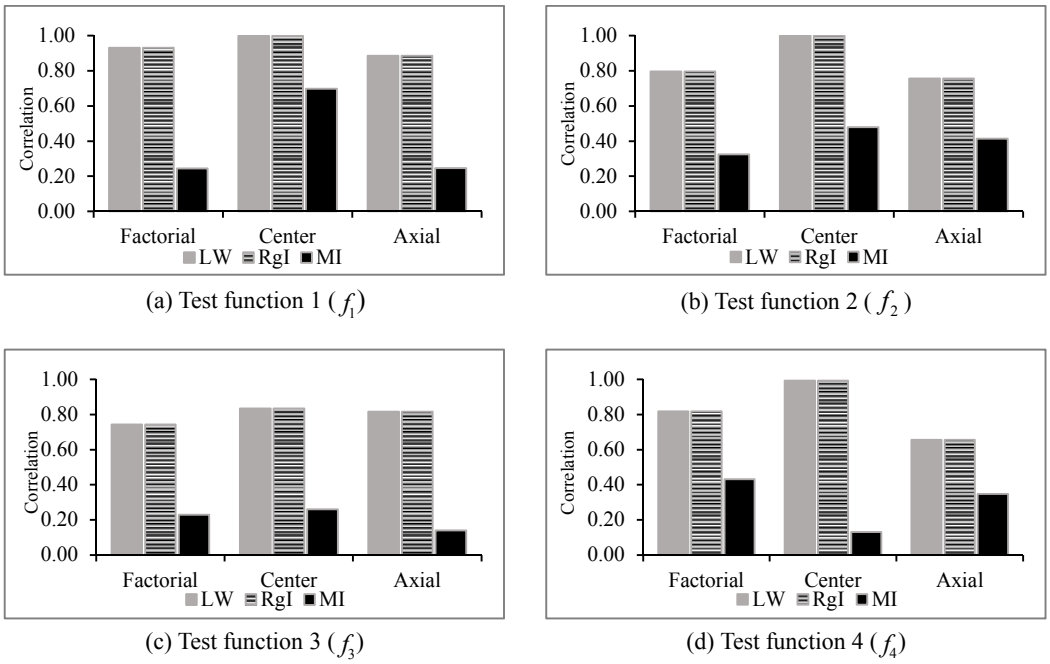
The correlation and MSE values between the optimal response with complete data and those after handling the missing value using the three methods for f_1 to f_4 in each part of CCD (factorial, center or axial) are summarized in Table 8. The lower the MSE value of the MSE, the better the performance of the imputation method, while the higher the correlation coefficient, the stronger the relationship (ranging from 0 for no relationship to 1 for a perfectly predictable relationship). For f_1, regression imputation, listwise deletion, and mean imputation produced mean correlation values of 0.9302, 0.9302, and 0.2445 to handle the missing data point in the in factorial part; 0.8850, 0.8850, and 0.2467 for the axial part; and 0.9984, 0.9984, and 0.6962 for the center part, respectively. Thus, in all parts of CCD, the performances of regression imputation and listwise deletion were similar with high correlation values, and thus quite considerably outperformed mean imputation. The same trend was found for f_2 to f_4.

When considering the mean of the MSE, regression imputation, listwise deletion, and mean imputation produced mean MSE values of 4.5779, 4.5779, and 38.7433 for a missing point with f_1 in the factorial part of the CCD; 4.3387, 4.3387, and 60.3240 for the axial part; and 0.0293, 0.0293, and 30.4907 for the center part, respectively. Thus, in all three parts, regression imputation and listwise deletion achieved the lowest MSE values and quite considerably outperformed the mean method. The same trend was found for f_2 to f_4. The correlation and MSE results are illustrated as bar charts in Figure 6(a)-(d) and Figure 7(a)-(d), respectively, and support that regression imputation and listwise deletion provided similar performances and outperformed mean imputation.

Table 8 The means of the correlation and MSE values for handling a missing data point in a CCD for f_1 to f_4

Part of the CCD	Test Function	Mean of Correlation Values			Mean of MSE Values		
		RgI	LW	MI	RgI	LW	MI
Factorial (F1–F4)	f_1	0.9302	0.9302	0.2445	4.5779	4.5779	38.7433
	f_2	0.7954	0.7954	0.3250	0.0046	0.0046	0.0301
	f_3	0.7438	0.7440	0.2290	0.2889	0.2888	25.7361
	f_4	0.8183	0.8182	0.4321	0.0034	0.0057	0.3991
Center (C1–C5)	f_1	0.9984	0.9884	0.6962	0.0293	0.0293	30.4907
	f_2	1.0000	1.0000	0.4791	5.84E-08	5.84E-08	0.0172
	f_3	0.8337	0.8337	0.2612	0.1529	0.1529	8.4438
	f_4	0.9924	0.9924	0.1307	1.12E-04	1.12E-04	0.7114
Axial (A1–A4)	f_1	0.8850	0.8850	0.2467	4.3387	4.3387	60.3240
	f_2	0.7558	0.7558	0.4141	0.0116	0.0116	0.0459
	f_3	0.8166	0.8166	0.1397	0.1767	0.1768	24.3213
	f_4	0.6567	0.6567	0.3483	0.0171	0.0171	0.0915

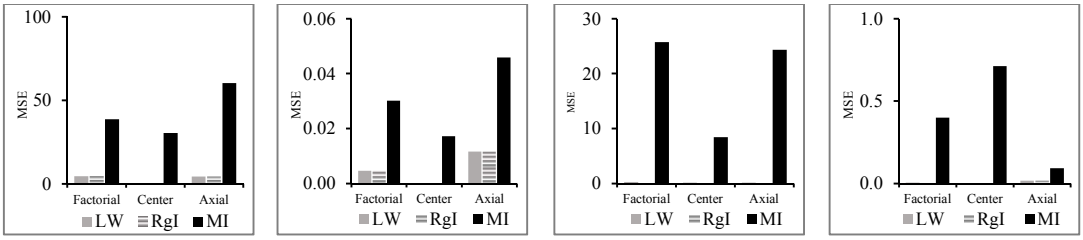
MI, mean imputation; RgI, regression imputation; LW, listwise deletion.



(a) Test function 1 (f_1) (b) Test function 2 (f_2)

(c) Test function 3 (f_3) (d) Test function 4 (f_4)

Figure 6 Mean correlation values of the optimal responses with complete data and after handling the missing response by using the three methods: (a) f_1, (b) f_2, (c) f_3 and (d) f_4.



(a) Test function 1 (f_1) (b) Test function 2 (f_2) (c) Test function 3 (f_3) (d) Test function 4 (f_4)

Figure 7 Mean MSEs of the optimal responses with complete data and after handling the missing response by using the three methods: (a) f_1, (b) f_2, (c) f_3 and (d) f_4.

5. Conclusions

The aim of the study was to compare the performance of listwise deletion with two imputation methods to handle a missing response value in a CCD with two factors in one of the three CCD parts (factorial, center, or axial). Methods for handling the missing value in the CCD for four test functions using mean and regression imputation, and listwise deletion were compared in terms of MSE and correlation coefficient values. One hundred simulation trials for each test function $(f_i; i = 1,2,3,4)$ were conducted by setting the difference between the low and high levels for the two factors and handling the missing value in each part of the CCD (factorial, center, or axial) using the three methods. Regression imputation and listwise deletion performed similarly in terms of the optimal response and were notably superior to mean imputation. For handling missing data by using listwise deletion, the single missing value is deleted from the data set. While regression imputation, the single missing value is estimated by a second-order model from the rest data. Both listwise deletion and regression imputation performed CCD by using the second-order model. According to Allison (2002), listwise deletion may be less biased than multiple imputation or ML when data are missing in regression analysis, and this corresponds with the results of Bengtsson et al. (2021), who claim that listwise deletion performs best for MCAR data and when the proportion of missingness is not too high. Mostly, listwise deletion gives valid inferences for MCAR data even when not using all available information (Allison 2002). Regression imputation and listwise deletion provided similar results for handling the missing value in each of the CCD parts (factorial, center, and axial) and were both superior to mean imputation.

Hence, regression imputation and listwise deletion are both appropriate for handling a single missing value in a CCD. Due to listwise deletion is usually the default method for dealing with missing data in most statistical software packages and has significantly outperformed for missing data when the proportion of missingness is not too high. Consequently, the listwise deletion method is plausible for handling a single missing value in a CCD. Handling the missing values in a CCD with more than two factors is planned for the future as an extension of this study.

Acknowledgments

The authors would like to thank the Basic Research Fund (BRF) of the Faculty of Science from Kasetsart University for supporting the fund in conducting this research and Khon Kaen University for providing the research facilities. In addition, the recommendations from Professor Jirawan Jitthavech and Associate Professor Vichit Lorchirachoonkul were very helpful in completing this study.

References

Allison P. Missing data, Volume 136. Thousand Oaks, CA: Sage Publications; 2002.

Anderson C, Borror CM, Montgomery DC. Response surface design evaluation and comparison. J Stat Plan Inference. 2009; 139: 629-674.

Ayilara OF, Zhang L, Sajobi TT, Sawatzky R, Bohm E, Lix LM. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health Qual Life Outcomes. 2019; 17(1): 1-9.

Azami M, Bahram M, Nouri S. Central composite design for the optimization of removal of the azo dye, Methyl Red, from wastewater using Fenton reaction. Curr Chem Lett. 2013; 2(2): 57-68.

Bagheri R, Ghaedi M, Asfaram A, Dil EA, Javadian H. RSM-CCD design of malachite green adsorption onto activated carbon with multimodal pore size distribution prepared from Amygdalus scoparia: kinetic and isotherm studies. Polyhedron. 2019; 171: 464-472.

Bengtsson F, Lindblad, K. Methods for handling missing values: a simulation study comparing imputation methods for missing values on a Poisson distributed explanatory variable. Bachelor [dissertation]. Sweden: Uppsala University; 2021.

Box GEP, Wilson KB. On the experimental attainment of optimum conditions, J R Stat Soc B. 1951; 13: 1-45.

Box GE, Draper NR. Empirical model-building and response surfaces. New York: John Wiley & Sons; 1987.

Briggs A, Clark T, Wolstenholme J, Clarke P. Missing.... presumed at random: cost-analysis of incomplete data. Health Econ. 2003; 12: 377-392.

Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol. 2003; 56(10): 968-976.

Farzadkia M, Ghorbanian M, Biglari H, Gholami M, Mehrizi EA. Application of the central composite design to optimization of petroleum hydrocarbons removal from oilfield water using advanced oxidation process. Arch Environ Protect. 2018; 44(4): 22-30.

Gano ZS, Mjalli FS, Al-Wahaibi T, Al-Wahaibi Y, AlNashef IM. Extractive desulfurization of liquid fuel with FeCl3-based deep eutectic solvents: experimental design and optimization by central-composite design. Chem Eng Process. 2015; 93: 10-20.

Graham JW. Missing data analysis: making it work in the real world. Ann Rev Psychol. 2009; 60: 549-576.

Graubard BI, Korn EL. Survey inference for subpopulations. Am J Epidemiol. 1996; 144(1): 102-106.

Junger WL, De Leon AP. Imputation of missing data in time series for air pollutants. Atmos Environ. 2015; 102:96-104.

Kumar R, Singh R, Kumar N, Bishnoi K, Bishnoi NR. Response surface methodology approach for optimization of biosorption process for removal of Cr(VI), Ni (II) and Zn (II) ions by immobilized bacterial biomass sp. Bacillus brevis. Chem Eng J. 2009; 146: 401-407.

Momen SB, Siadat SD, Akbari N, Ranjbar B, Khajeh K. Applying central composite design and response surface methodology to optimize growth and biomass production of Haemophilus influenzae Type B. Jundishapur J Microbiol. 2016; 9(6): e25246.

Myers RH, Montgomery DC. Response surface methodology: process and product optimization using designed experiments. New York: John Wiley & Sons; 2002.

Nakai M, Weiming K. Review of methods for handling missing data in longitudinal data analysis. Int J Math Anal. 2011; 5(1): 1-13.

Newman DA. Missing data: Five practical guidelines. Organ Res Methods. 2014; 17(4): 372-411.

Ostertagova E. Modelling using polynomial regression. Proc Eng. 2012; 48: 500-506.

R Core Team (2020). R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. [cited 2021 Jan 20]. Available from: http://www.R-project.org.

Roth PL. Missing data: a conceptual review for applied psychologists. Pers Psychol. 1994; 47(3): 537-560.

Roth PL, Switzer III FS, Switzer DM. Missing data in multiple item scales: a Monte Carlo analysis of missing data techniques. Organ Res Methods.1999; 2(3): 211-232.

Rubin DB. Multiple imputation for non-response in surveys. New York: John Wiley & Sons; 1987.

Rubin DB. Inference and missing data. Biometrika. 1976; 63(3): 581-592.

Saunders JA, Morrow-Howell N, Spitznagel E, Doré P, Proctor EK, Pescarino R. Imputing missing data: a comparison of methods for social work researchers. Soc Work Res. 2006; 30(1): 19-31.

Schafer JL. Analysis of incomplete multivariate data. Boca Raton: CRC Press; 1997.

Wongoutong C, Jitthavech J, Lorchirachoonkul V. An application of Nelder-Mead algorithm in response surface methodology: CCD. Thail Stat. 2017; 15(2): 167-183.