International Journal of Engineering

Journal Homepage: www.ije.ir

TECHNICAL NOTE

Evaluating Estimation Methods of Missing Data on a Multivariate Process Capability Index

A. Ashuri, A. Amiri*

Industrial Engineering Department, Faculty of Engineering, Shahed University, Tehran, Iran

PAPER INFO

Paper history:
Received 12 March 2014
Received in revised form 10 August 2014
Accepted 18 September 2014

Keywords:
Process Capability Index
Missing Data
Imputation Methods
Response Variable
Main and Interaction Effects

ABSTRACT

Quality of products has been one of the most important issues for manufacturers in the recent decades. One of the challenging issues is evaluating capability of the process using process capability indices. On the other hand, usually the missing data is available in many manufacturing industries. So far, the performance of estimation methods of missing data on process capability indices has not been evaluated. Hence, we analyze the performance of a process capability index when we deal with the missing data. For this purpose, we consider a multivariate process capability index and evaluate four methods including Mean Substitution, EM algorithm, Regression Imputation and Stochastic Regression Imputation to estimate missing data. In the analysis, factors including percent of missing data (k), sample size (m), correlation coefficients (ρ) and the estimation methods of missing data are investigated. We evaluate the main and interaction effects of the factors on response variable which is defined as difference between the estimated index and the computed index with full data by using General Linear Model in ANOVA table. The results of this research show that the Stochastic Regression Imputation has the best performance among the estimation methods and the percent of missing data (k) has the highest effect on response variable. Also, we conclude that the sample size has the lowest effect on response variable among the mentioned factors.

***doi**: 10.5829/idosi.ije.2015.28.01a.12*

1. INTRODUCTION

With increasing the competitiveness of the industry and developing technology, quality of the products is taken into consideration. To achieve this goal, we should have capable processes. Hence, we apply process capability index to evaluate capability of the processes. A process capability index measures the ability of a process in producing the items which conform the customers' needs. Process performance is "what a process actually does". Comparing the process performance with customers' needs leads to "what the process should do"; see Juran's definition in Tano and Vännman [1]. Process capability indices can be considered in two types:

- Univariate process capability indices
- Multivariate process capability indices

In univariate process capability indices, a single quality characteristic of a product is evaluated. Univariate process capability indices were proposed by some authors such as Kane [2]; Pearn and Chen [3]; Noorossana [4] and Chen and Chen [5]. In most cases, the process capability analysis consists of several quality characteristics. In these cases, multivariate process capability indices are used. Some authors have developed multivariate process capability indices including Hubble et al. [6]; Taam et al. [7] and Shahriari et al. [8]. Abu and Sultana [9] did a comparison among multivariate process capability indices. Pan and Lee [10] proposed a process capability index. Moreover, some authors have proposed multivariate process capability indices including Niavarani et al. [11] and Shahriari and Abdollahzadeh [12]. For some recent researches on multivariate process capability indices, the researchers are referred to Ciupke [13], Jalili et al. [14], Gu et al. [15] and Tano and Vännman [16].

*Corresponding Author's Email: amiri@shahed.ac.ir (A. Amiri)

Please cite this article as: A. Ashuri, A. Amiri, Evaluating Estimation Methods of Missing Data on a Multivariate Process Capability Index, International Journal of Engineering (IJE), TRANSACTIONS A: Basics Vol. 28, No. 1, (January 2015) 88-96

In many real applications, some of the observations are missing. Previous studies have shown that missing data is one of the most important issues in statistical sciences; see Little and Rubin [17]. For example, in statistical surveys, sometimes all questions are not answered [18]. Sometimes, measurements for all samples are too costly [19]. Missing data may arise from experimental designs, as a result of insufficient sampling, high costs, and errors in measurements or during data acquisition [18].

Three mechanisms of missing data are presented by Little and Rubin [17].

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

Each of the above models explains the relationship between data and probability of missing data. When data follow the MCAR mechanism, there is no relationship between values of the variables and the probability of missing data. In the MAR mechanism, missingness is related to the observed data and not the values that are missing. In the MNAR mechanism, the probability of missing data is related to the unobserved value of the missing. There are several statistical methods to estimate the missing data. The two most common of these traditional methods are single imputation and deletion; see Peugh and Enders [20]. In single imputation method, one value for each missing element is imputed. Mean Substitution, Regression Imputation and Stochastic Regression Imputation are single imputation methods. In the deletion method, missing values are not considered. Hence, we deal with complete cases. The Maximum likelihood and multiple imputation estimation methods are considered as modern missing data techniques [17]. Despite of many applications of missing data in real cases, the missing data is less considered in the area of statistical process control. Among the few researches in this area, the performance of some estimation methods of missing data on Phase I multivariate control charts was investigated by Mahmoud et al. [21]. Also, Madbuly et al. [22] analyzed the effect of methods for handling missing data on the performance of the multivariate exponentially weighted moving average (MEWMA) control chart. In this paper, the effect of missing data on a process capability index by Taam et al. [7] is evaluated. For this purpose, we consider four methods including Mean Substitution, EM algorithm, Regression Imputation and Stochastic Regression Imputation to estimate missing data. The main aim of this paper is comparing these estimation methods in terms of their ability to reconstruct the original data. In addition, the ANOVA technique is used to analyze the main and interaction effects of the factors on response variable which is defined as difference between the estimated index and computed index with full data. The structure of the paper is as follows: section 2 discusses about the process capability index proposed by Taam et al. [7] and the corresponding assumptions. The estimation methods of missing data are presented in section 3. Numerical simulations are presented in section 4. Concluding remarks and some future researches are given in the final section.

2. A MULTIVARIATE PROCESS CAPABILITY INDEX AND CORRESPONDING ASSUMPTIONS

In many cases, a manufactured product is described by more than one quality characteristic. So, we assume quality characteristics follow a multivariate normal distribution with a mean vector of $\boldsymbol{\mu}$ and a variance–covariance matrix $\boldsymbol{\Sigma}$. To evaluate the capability of the process, we consider the process capability index MC_{pm} proposed by Taam et al. [7]. To compute the MC_p index, at first the tolerance region is converted into a modified tolerance region. Then, process capability index is defined as the ratio of the volumes of the ellipsoids of the modified tolerance region to the process region. For multivariate normal process, process region is an elliptical region represented by the quadratic form $(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\le k(q)$. The MC_p index is defined as follows:

$$MCp = \frac{vol.(\text{modified tolerance region})}{vol.[(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\le k(q)]}, \quad (1)$$

where, $k(q)$ is the 99.73% of the chi-square distribution with υ degrees of freedom. Hence, the MC_p index is rewritten as follows:

$$MCp = \frac{vol.(\text{modified tolerance region})}{(\pi\ \chi^2_{\upsilon,0.9973})^{\upsilon/2}\,|\boldsymbol{\Sigma}|^{1/2}\left[\Gamma(\upsilon/2+1)\right]^{-1}} \quad (2)$$

where, $|\boldsymbol{\Sigma}|$ is determinant of $\boldsymbol{\Sigma}$, and $\Gamma(\cdot)$ is the gamma function. Also, the multivariate process capability index MC_{pm} is defined as:

$$MC_{pm} = \frac{MC_p}{D} \quad (3)$$

where,

$$D = \left[1+\frac{n}{n-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})'(\boldsymbol{\Sigma})^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu})\right]^{1/2} \quad (4)$$

An estimator of the MC_{pm} can be expressed as:

$$\hat{MC}_{pm} = \frac{vol.(\text{modified tolerance region})}{vol.(\text{estimated } 99.73\%\ \text{process region})} \quad (5)$$

where,

$$vol(\text{modified tolerance region}) = \frac{2\pi^{p/2}\prod_{j=1}^{p}(USL_j - LSL_j)/2}{p\,\Gamma(p/2)} \quad (6)$$

and,

$$vol.(estimated\ 99.73\%\ process\ region) = (\pi\ \chi^2_{\upsilon,0.9973})|\mathbf{S}|^{1/2}[\Gamma(\upsilon/2+1)]^{-1} \quad (7)$$

In Equations (4), (6) and (7), n is the sample size, $\bar{\mathbf{x}}$ is sample mean vector, $\boldsymbol{\mu}$ is the process mean vector, p is number of quality characteristics, $\mathbf{S}$ is sample variance-covariance matrix. USL_j and LSL_j are the upper and specification limits of jth quality characteristic.

For a bivariate normal process, the area of modified tolerance region is computed as follows:

$$\text{Area}(\textit{modified}\ \text{tolerance region}) = \pi \times \prod_{j=1}^{2}(\text{USL}_j - \text{LSL}_j)/2 \quad (8)$$

To use the Taam et al.'s MC_{pm} index, it is assumed that the process mean vector is equal to the target value.
As shown in Figure 2, three results can be occurred:
(a) The process is incapable because the process region is greater than the modified tolerance;
(b) The process is capable and 99.73% of the process region falls into the modified tolerance region.
(c) The process is capable because the process region is smaller than the modified tolerance region.

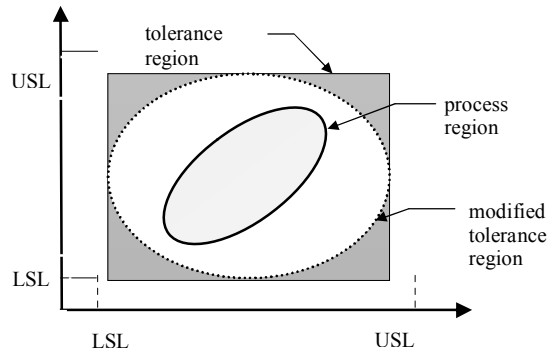

Figure 1. Graphical representation of process region, modified tolerance region and tolerance region for a bivariate normal distribution (Taam et al. [7])

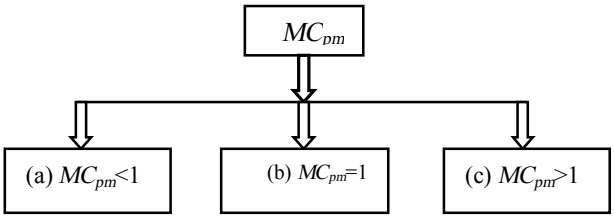

Figure 2. Results obtained from Taam et al.'s index

3. METHODS FOR ESTIMATING MISSING DATA

To evaluate the capability of a process in the presence of missing data correctly, we need first to estimate the missing data accurately. We consider four methods in this paper to estimate the missing data. These methods are explained in the following paragraphs: We also consider the MCAR mechanism among the mechanisms mentioned in the introduction section, which implies that there is no relationship between values of the variables and the probability of missing data [17].

3. 1. Mean Substitution Mean imputation method implies that missing data is replaced with mean of the available data. For example, if we do not have the value of x_{ij}, we replace it with $\bar{x}_j$ (ith observation in jth quality characteristic) based on the available data set. The method is not suitable because replacing missing data with mean of available data leads to underestimating variance and covariance of quality characteristics [22].

3. 2. The EM Algorithm The EM algorithm proposed first by Dempster et al. [23] is a method of estimating missing data which uses the concept of the maximum-likelihood function to estimate the parameters of distribution when the data are missing. The EM algorithm includes the following three steps:
1. At first, the missing data is replaced by an initial estimation of missing data average to achieve complete data
2. In the second step, we estimate the missing data average based on the full data set obtained in the first step.
3. The difference between missing data averages in two successive steps is checked and the algorithm is repeated until the convergence.

Figure 3 explains the EM procedure applied in this paper. The EM algorithm uses the concept of the maximum-likelihood function to estimate distribution parameters. The maximum likelihood estimate of $\boldsymbol{\mu}$ using the complete data is equal to

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n}\mathbf{r}_i}{n} \quad (9)$$

where, $\mathbf{r}_i$ is the ith vector of observations and $\boldsymbol{\mu}$ is the mean vector of full observations. In the case of existing missing data, first we estimate the missing observations with an initial value $\boldsymbol{\mu}^{(t)}$ (see step 1 in Figure 3).

Then, we obtain the vector $\boldsymbol{\mu}^{(t+1)}$ with using the maximum likelihood estimate of mean vector as:

$$\boldsymbol{\mu}^{(t+1)} = \frac{\sum_{i=1}^{n}\hat{\mathbf{r}}_i}{n} \quad (10)$$

in which, $\mathbf{r}_i$ vector is computed by an iterative procedure as

$$\sum_{i=1}^{n}\hat{\mathbf{r}}_i = \boldsymbol{\mu}^{(t)}(n-m) + \sum_{i=1}^{m}\mathbf{r}_i \quad (11)$$

where n is total number of observations, m is the number of observed data and (n-m) is the number of missing data (see step 2 in Figure 3). Then, we check the difference between $\boldsymbol{\mu}^{(t)}$ and $\boldsymbol{\mu}^{(t+1)}$. If the difference is greater than ε, then we replace $\boldsymbol{\mu}^{(t)}$ by $\boldsymbol{\mu}^{(t+1)}$ and repeat the step 2 of the algorithm. Otherwise, we estimate the missing data with vector of $\boldsymbol{\mu}^{(t+1)}$ and stop.

3. 3. Regression Imputation Regression imputation replaces missing data with predicted values from a regression equation. In this method, first we use the full observations in which both elements of vector of observation are known. The second element of observatios vector, X_2 is considered as response variable and the first element X_1 is considered as independent variable. Then, we fit a linear regression model to the full data set and estimate the regression parameters as Equations (12) and (13).

$$\hat{\alpha} = \bar{x}_{2i} - \hat{\beta}\,\bar{x}_{1i} \tag{12}$$

$$\hat{\beta} = \frac{Sx_1x_2}{Sx_1x_1} \tag{13}$$

where, $\hat{\alpha}$ and $\hat{\beta}$ are the estimates of the intercept and the slope and $\bar{y}$ and $\bar{x}$ are the mean of the full observations and Sx_1x_2 and Sx_1x_1 are computed as follows:

$$Sx_1x_2 = \sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \tag{14}$$

$$Sx_1x_1 = \sum_i (x_{1i} - \bar{x}_1)^2, \tag{15}$$

After obtaining the regression parameters, the following equation is used to perdict the missing elements in the vector of observations:

$$\hat{X}_{2i} = \hat{\alpha} + \hat{\beta} X_{1i} \tag{16}$$

It is clear that if X_{1i} is missing, the following equation can be used for estimating the X_{1i} obtained based the Equation (16).

$$\hat{X}_{1i} = \frac{X_{2i} - \hat{\alpha}}{\hat{\beta}} \tag{17}$$

3. 4. Stochastic Regression Imputation Stochastic regression imputation also uses regression model fitted on full vector of observations to predict the missing elements of observations vectors, but it takes the extra step of augmenting each predicted value with a normally distributed residual term. Adding residuals to the estimated values eliminates the biases related to regression imputation scheme [17]. This method is shown schematically in Figure 4. As shown in this figure, there are biases between the real observations and the fitted values by the regression imputation method which is eliminated by the stochastic regression method through adding the random errors to the fitted values.

The step by step procedure of the stochastic regression scheme is explained as follows:

First, we fit a linear regression model to full observations and estimate the regression parameters (similar to what discussed in the regression imputation method).

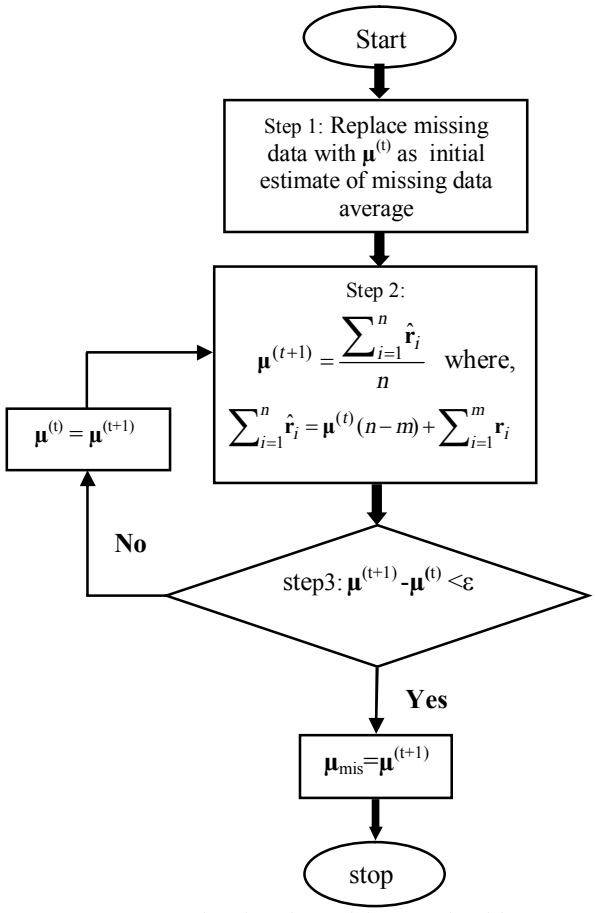

Figure 3. The Flowchart of the EM algorithm

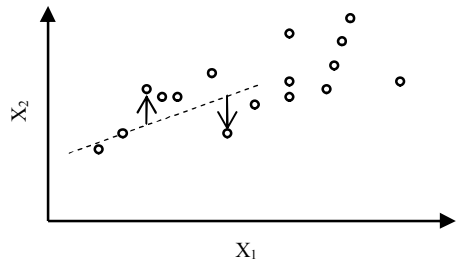

Figure 4. Schematic representation of a simple linear regression model fitted to full observations and the corresponding biases [25].

Then, the residuals of observations are computed and the variance of error term is estimated by using mean square error as follows:

$$MSE = \frac{SSE}{m-2} \quad (18)$$

where, m is the total of full observations, SSE is sum of square error and is computed as:

$$SSE = \sum_{i=1}^{m} e_i^2 \quad (19)$$

where

$$e_i = \hat{X}_{2i} - \hat{\alpha} - \hat{\beta} X_{1i} \quad (20)$$

In the next step, the normal distribution random errors with mean zero and variance equal to MSE are generated through simulation and added to the fitted values as follows:

$$\hat{X}_{2i} = \hat{\alpha} + \hat{\beta} X_{1i} + e_i \quad (21)$$

These random errors compensate for the distance between the real value and fitted values based on the regression imputation method. We use Equation (21) to estimate the missing data.

4. PERFORMANCE EVALUATION

Our simulation study compares the performance of the four methods to estimate the missing data and compute the process capability index by Taam et al. [7].

For this analysis, we have considered three values for the sample sizes (m = 20, 50 and 100), three values for the correlation coefficients (ρ = 0.10, .5 and 0.8) and four values for the percent of missing data (k= 5, 10, 20 and 50). We have assumed the process follows a bivariate normal distribution with the following mean vector of $\boldsymbol{\mu}$ and the variance–covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{\mu} = [2 \quad 4] \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 2.3 \end{bmatrix}.$$

The results of simulation are obtained by 10,000 replications and are summarized in Tables 2, 3 and 4. Also, we investigate the main effect of factors (sample size, correlation coefficients, percent of missing values and the four explained methods) on response variable, the difference between estimated index and computed index with full data, and the results are shown in Figure 5 and Table 6. In addition, the interaction effects between the factors are shown in Table 6 and Figure 9. The specification limits and the target values for considered example are presented in Table 1.

Table 2 shows the estimated process capability index by Taam et al. [7] by using four methods of the Mean Substitution, the EM algorithm, the Regression Imputation and the Stochastic Regression Imputation when $\rho = 0$.

TABLE 1. Specifications and the target values

Quality characteristic	Target	LSL	USL
X_1	3.5	0.1	6.9
X_2	5.5	0.9	10.1

TABLE 2. Estimated process capability index when (ρ = 0)

m	Methods	*K* 5%	10%	20%	50%	Full data
20	MS	0.4534	0.4791	0.5088	0.7140	0.4373
	EM	0.4515	0.4750	0.4996	0.6716	
	RG	0.4472	0.4665	0.4798	0.5834	
	SRG	0.4312	0.4367	0.4318	0.5163	
50	MS	0.4499	0.4568	0.4922	0.6605	0.4259
	EM	0.4488	0.4554	0.4891	0.6466	
	RG	0.4408	0.4453	0.4661	0.5464	
	SRG	0.4159	0.4155	0.4160	0.4788	
100	MS	0.4381	0.4533	0.4873	0.6484	0.4241
	EM	0.4378	0.4526	0.4860	0.6418	
	RG	0.4327	0.4420	0.4618	0.5464	
	SRG	0.4161	0.4114	0.4110	0.4701	

Also, we compute the process capability index when the data is full. The results show that the difference between process capability index with full data and the estimated process capability index for small percent's of missing data, e.g. k =5 and 10%, is smaller than the larger percentages of missing values, e.g. k =20 and 50%. For example, when m = 20 samples and 5% of the data is missing and the SRG method is used, the estimated index is equal to 0.4312, as indicated in Table 2 and the difference between the computed and estimated indices is 0.0061. When the percent of missing data increases to k = 50%, the estimated index is 0.5163 and the difference between computed index with full data and estimated index is 0.0790. We conclude that the method of handling missing data when we deal with the lower percent of missing values has better performance in comparison with high percentages. In Tables 3, 4 and 5, we investigate the effect of the considered methods in estimating the process capability index when quality characteristics are correlated. We consider three different values of (ρ = 0.1, 0.5 and 0.8). As shown in Tables 2-5, when ρ increases the difference between the estimated index and index with full data increases. For example, when $\rho = .1$, m=20 samples, 10% of the data is missing and imputed using the EM algorithm, estimated process capability index is equal to 0.4659. In addition, the difference between estimated and response variable is 0.02. When $\rho = .8$ the index is equal to 0.6198 and the difference is 0.0409. We use the ANOVA technique to analyze the main and interaction effects on the response variable. For this purpose, the General Linear Model in Minitab software is used. To do that, at first we checked normality of the response variable with two replicates using Anderson-Darling test and obtained the p-value less than 0.05. Hence, the normality assumption is

violated. We used different transformation techniques including Johnson transformation and Box-Cox to transform the distribution of the response variable to normal. However, none of them was capable to transform the response variable to normal distribution. Hence, we used the percent of missing value equal to 15% instead of 50% in our analysis. Hence, the levels of the percent of missing values factor changed from (5, 10, 20, 50%) to (5, 10, 15, 20%). Then, we replaced the corresponding response variable. After that, we used Johnson transformation to transform the distribution of the response variable to normal. Finally, the normality assumption is satisfied. Then, we analyzed the main effect of factors including sample size (*m*), correlation coefficient (*rho*), percent of missing values (*k*) and type of method (*type*) on the response variable, difference between process capability index with full data and the estimated process capability index using the General Linear Model. The obtained results are shown in Figure 5. Figure 5 shows that among the considered factors, the percent of missing data (*k*) is the most important factor and has the maximum effect on the response variable. While, the sample size (*m*) has the lowest effect on the response variable. Note that the estimation methods are compared based on the difference between the process capability index with full observations and the process capability index with missing data estimated by each of these methods as follows:

Difference=MC_{pm} (full data)-MC_{pm} (estimated data).

TABLE 3. Estimated process capability index when (ρ = .1)

m	Methods	*K* 5%	10%	20%	50%	Full data
20	MS	0.4630	0.4699	0.5221	0.7337	0.4459
	EM	0.4611	0.4659	0.5128	0.6907	
	RG	0.4568	0.4571	0.4928	0.6089	
	SRG	0.4406	0.4287	0.4426	0.5347	
50	MS	0.4377	0.4823	0.5213	0.7357	0.4479
	EM	0.4372	0.4781	0.5120	0.6074	
	RG	0.4334	0.4695	0.4919	0.6074	
	SRG	0.4201	0.4393	0.4427	0.5341	
100	MS	0.4492	0.4647	0.5008	0.6656	0.4349
	EM	0.4488	0.4640	0.4992	0.6588	
	RG	0.4439	0.4535	0.4760	0.5673	
	SRG	0.4261	0.4210	0.4217	0.4841	

TABLE 4. Estimated process capability index when (ρ = .5)

m	Methods	*K* 5%	10%	20%	50%	Full data
20	MS	0.5254	0.5461	0.5926	0.8409	0.5057
	EM	0.5232	0.5413	0.5820	0.7890	
	RG	0.5198	0.5336	0.5648	0.7117	
	SRG	0.5015	0.4994	0.5058	0.6237	
50	MS	0.5072	0.5306	0.5935	0.8457	0.5059
	EM	0.5066	0.5289	0.5827	0.7933	
	RG	0.5033	0.4198	0.5653	0.7140	
	SRG	0.4868	0.4830	0.5073	0.6259	
100	MS	0.5076	0.5262	0.5674	0.7620	0.4911
	EM	0.5072	0.5254	0.5658	0.7538	
	RG	0.5028	0.5162	0.5450	0.6743	
	SRG	0.4738	0.4781	0.4799	0.5666	

TABLE 5. Estimated process capability index when (ρ = .8)

m	Methods	*K* 5%	10%	20%	50%	Full data
20	MS	0.6015	0.6255	0.6820	0.9649	0.5789
	EM	0.5990	0.6198	0.6693	0.9087	
	RG	0.5953	0.6123	0.6521	0.8304	
	SRG	0.5764	0.5772	0.5925	0.7372	
50	MS	0.6026	0.6251	0.6821	0.9711	0.5781
	EM	0.6001	0.6195	0.6692	0.9108	
	RG	0.5967	0.6120	0.6516	0.8291	
	SRG	0.5775	0.5773	0.5908	0.7377	
100	MS	0.5820	0.6035	0.6523	0.8775	0.5630
	EM	0.5816	0.6027	0.6504	0.8678	
	RG	0.5771	0.5933	0.6296	0.7919	
	SRG	0.5564	0.5547	0.5623	0.6745	

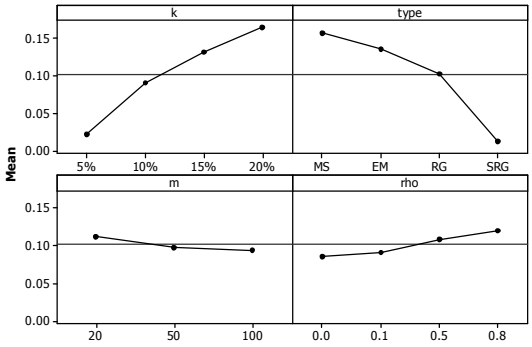

Figure 5. Main effects of the factors on the response variable

As illustrated in the above-right plot of Figure 5, this criterion is computed for each method. Whatever the criterion for a method is closer to the zero value, the performance of the estimation method is better. The plot shows that the stochastic regression imputation has the lowest difference with zero and as a result it is the best estimation method. Based on the criterion which is the difference between the estimated index and the computed index with full data, the regression imputation, the EM algorithm and the Mean substitution have the lowest difference with zero. Hence, these methods are in the next priorities for estimating the missing data. These results are confirmed by the results mentioned by Mahmoud et al. [21] and Madbuly et al. [22] in evaluating the estimatin methods for handling missing data in the T^2 and MEWMA control charts, respectively. The main effect of the correlation coefficient shows that as the correlation coefficient increases, the response variable deteriorates. The main effect of the missing data percentage shows the more percent of missing data, the more difference between estimated process capability index and the process capability index computed by full data

Table 6 shows the significance hypothesis tests of four main effects, the six two-factor interactions, the four three-factor interactions and the one four-factor interaction on response variable. The last column shows

the p-value. When p-value is less than 0.05, the relationship between response variable and factors is significant. As indicated in Table 6, the effect of factors including sample size (*m*), correlation coefficients (*rho*), percent of missing data (*k*) and the type of method (type) on response variable is significant. The three-factor interactions are not significant for interaction between (*type*, *k*,*rho*). The same results are obtained for the two-factor interactions between (*type* and *m*). For example, the p-value for interaction between *type* and *m* is 0.249 and we conclude that the two factor interaction between *type* and *m* is not significant. However, the two-factor interactions between factors (*type*, *k*), (*type*, *rho*), (*rho* and *m*), (*k*, *rho*) and (*k*, *m*) are significant. For example, the p-value of the interaction effect between percent of missing data (*k*) and correlation coefficient (*rho*) is equal to zero and less than 0.05. Also, the four-factor interaction effect on the response variable is not significant. Three assumptions including normality of residuals, constant variance of the residuals and independency of the residuals in ANOVA table are also investigated which are illustrated in Figures 6, 7 and 8. To check the normality assumption of the residuals, we use Anderson-Darling test. The p-value obtained based on the Anderson-Darling normality test is equal to 0.122. So, we conclude that the normality assumption is satisfied. As shown in Figure 7, we use the Bartlett's test in Minitab software to test for equality of variances. The p-values obtained from the Bartlett's test is equal to 0.494. Since this value is greater than 0.05, we fail to reject assumtion of equal variances and this assumption is also satisfied. The indepedendency of the residuals is evaluated by the residual versus observation order plot. If the plot does not reveal any patterns, the independency assumption is satisfied. Based on Figure 8, we conclude that the independency assumption of the residuals is satisfied.

TABLE 6. ANOVA table which shows the effect of factors on the response variable

Source	DF	Seq SS	Adj SS	Adj MS	F	P
type	3	111.3280	97.3243	32.4414	327.54	0.000
k	3	103.3131	96.8970	32.2990	326.10	0.000
rho	3	9.2080	9.1741	3.0580	30.87	0.000
m	2	2.4301	2.8179	1.4089	14.23	0.000
type*k	9	21.1121	18.1001	2.0111	20.30	0.000
type*rho	9	5.3683	4.1249	0.4583	4.63	0.000
k*rho	9	3.3997	3.8258	0.4251	4.29	0.000
type*m	6	1.0519	0.7292	0.1215	1.23	0.294
k*m	6	2.4907	3.2852	0.5475	5.18	0.000
rho*m	6	1.5249	1.8367	0.3061	2.89	0.010
k*rho*m	18	5.2779	5.7104	0.3172	3.20	0.000
type*k*m	18	3.4588	3.2088	0.1783	1.80	0.028
type*k*rho	27	3.5185	3.6752	0.1361	1.37	0.115
type*rho*m	18	3.2435	3.0015	0.1667	1.68	0.046
k*type*m*rho	54	6.1166	6.1166	0.1133	1.14	0.256
error	177	17.5312	17.5312	0.0990		
Total	368	299.6094				

S=0.314716 R-Sq= 94.15% R-Sq(adj) = 87.83%

The two-factor, the three-factor and the four-factor interactions on the response variable are shown in Figure 9. The results similar to the Table 6 are shown in Figure 9. The results of the Figure 9 confirm the results obtained in Table 6. As an example, when we use the MS method and the percent of missing data is *k*=20%, the difference of response value with zero increases with more slope and this means interaction between the mentioned factors has negative influence on response variable. This analysis helps us to understand that in what circumstances, the use of estimation methods is more efficient. It makes the risks and costs arising from the use of these methods significantly lower.

6. CONCLUDING REMARKS AND FUTURE RESEARCHES

In this paper, we evaluated the influence of missing data on the performance of the process capability index proposed by Taam et al. [7].

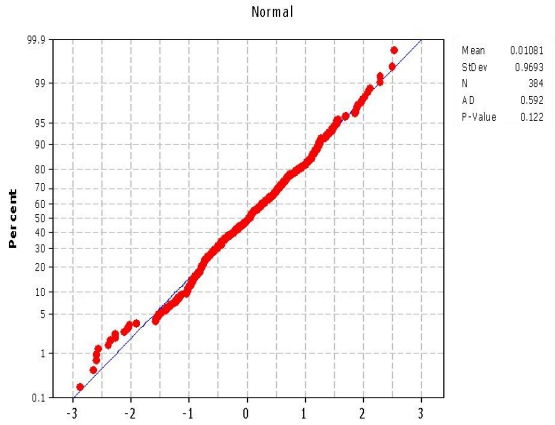

Figure 6. Normality Probability Plot (NPP) to test the normality of the residuals

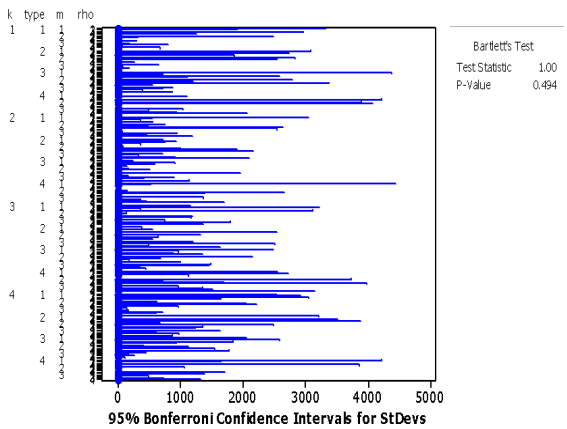

Figure 7. Bartlett's test for equal variances for residuals

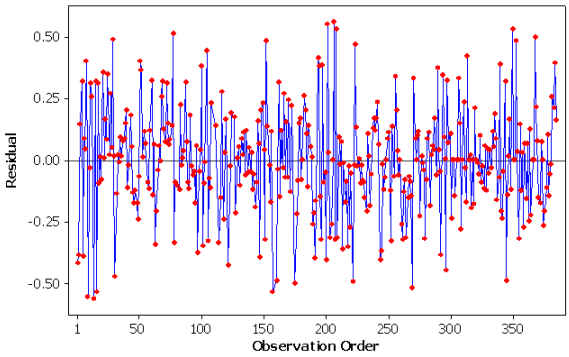

Figure 8. Residuals versus the observation order plot

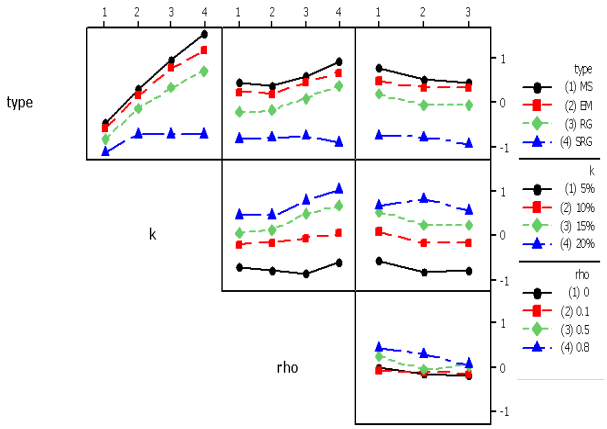

Figure 9. Interaction effects on response variable

For this purpose, we considered four imputation methods including Mean Substitution, EM algorithm, Regression Imputation and Stochastic Regression Imputation to estimate the missing data. We evaluated four factors including the sample sizes (m), the percent of missing data (k), the correlation coefficients between the quality characteristics (ρ) and the methods on response variable which is defined as different between estimated index and computed index with full data. The simulation results showed that the Stochastic Regression Imputation method is the best method to estimate the missing data and as a result the multivariate process capability index under the missing data. The regression imputation, the EM algorithm and the mean substitution methods are in the next priorities for estimation missing data and the last two methods have poor performance. Also, the results showed that the percent of missing values (k) has the highest effect and the sample size has the lowest effect among the mentioned factors on response variable. As a future research, we suggest investigating the effects of missing data on other process capability indices. In addition, one can investigate the effect of missing data on change point estimation [24, 25] or profile monitoring [26, 27] as well.

7. ACKNOWLEDGEMENT

The authors are grateful to the respectful reviewers for precious comments which led to improvement in the paper.

8. REFERENCES

1. Tano, I. and Vannman, K., "Comparing confidence intervals for multivariate process capability indices", ***Quality and Reliability Engineering International***, Vol. 28, No. 4, (2012), 481-495.
2. Kane, V.E., "Process capability indices", ***Journal of Quality Technology***, Vol. 18, No. 1, (1986), 41-52.
3. Pearn, W. and Chen, K., "One-sided capability indices and: Decision making with sample information", ***International Journal of Quality & Reliability Management***, Vol. 19, No. 3, (2002), 221-245.
4. Noorossana, R., "Process capability analysis in the presence of autocorrelation", ***Quality and Reliability Engineering International***, Vol. 18, No. 1, (2002), 75-77.
5. Chen, J.-P. and Chen, K., "Comparison of two process capabilities by using indices: An application to a color stn display", ***International Journal of Quality & Reliability Management***, Vol. 21, No. 1, (2004), 90-101.
6. Hubble, N.F., Shahriari, H. and Cheng, G.S., "A bivariate process capability vector in statistics and design in process control", ***New York: Marcel Dekker***, (1991), 299-310.
7. Taam, W., Subbaiah, P. and Liddy, J.W., "A note on multivariate capability indices", ***Journal of Applied Statistics***, Vol. 20, No. 3, (1993), 339-351.
8. Shahriari, H., Hubele, N. and Lawrence, F., "A multivariate process capability vector", in Proceedings of the 4th Industrial Engineering Research Conference. Vol. 1, (1995), 304-309.
9. Zahid, M.A. and Sultana, A., "Assessment and comparison of multivariate process capability indices in ceramic industry", ***Journal of Mechanical Engineering***, Vol. 39, No. 1, (2008), 18-25.
10. Pan, J.N. and Lee, C.Y., "New capability indices for evaluating the performance of multivariate manufacturing processes", ***Quality and Reliability Engineering International***, Vol. 26, No. 1, (2010), 3-15.
11. Niavarani, M.R., Noorossana, R. and Abbasi, B., "Three new multivariate process capability indices", ***Communications in Statistics-Theory and Methods***, Vol. 41, No. 2, (2012), 341-356.
12. Shahriari, H. and Abdollahzadeh, M., "A new multivariate process capability vector", ***Quality Engineering***, Vol. 21, No. 3, (2009), 290-299.
13. Ciupke, K., "Multivariate process capability vector based on one-sided model", ***Quality and Reliability Engineering International***, Vol., No., (2014).
14. Jalili, M., Bashiri, M. and Amiri, A., "A new multivariate process capability index under both unilateral and bilateral quality characteristics", ***Quality and Reliability Engineering International***, Vol. 28, No. 8, (2012), 925-941.
15. Gu, K., Jia, X., Liu, H. and You, H., "Yield-based capability index for evaluating the performance of multivariate manufacturing process", ***Quality and Reliability Engineering International***, (2013).
16. Tano, I. and Vännman, K., "A multivariate process capability index based on the first principal component only", ***Quality and Reliability Engineering International***, Vol. 29, No. 7, (2013), 987-1003.

17. Little, R.J.A. and Rubin, D.B., "Statistical analysis with missing data (2nd edn), John Wiley & Sons, Inc., New York, NY, Vol. 7, (2002).
18. Schafer, J.L. and Graham, J.W., "Missing data: Our view of the state of the art", ***Psychological Methods***, Vol. 7, No. 2, (2002), 147-177.
19. Grung, B. and Manne, R., "Missing values in principal component analysis", ***Chemometrics and Intelligent Laboratory Systems***, Vol. 42, No. 1, (1998), 125-139.
20. Peugh, J.L. and Enders, C.K., "Missing data in educational research: A review of reporting practices and suggestions for improvement", ***Review of Educational Research***, Vol. 74, No. 4, (2004), 525-556.
21. Mahmoud, M.A., Saleh, N.A. and Madbuly, D.F., "Phase I analysis of individual observations with missing data", ***Quality and Reliability Engineering International***, Vol. 30, No. 4, (2014), 559-569.
22. Madbuly, D.F., Maravelakis, P.E. and Mahmoud, M.A., "The effect of methods for handling missing values on the performance of the mewma control chart", ***Communications in Statistics-Simulation and Computation***, Vol. 42, No. 6, (2013), 1437-1454.
23. Dempster, A.P., Laird, N.M. and Rubin, D.B., "Maximum likelihood from incomplete data via the em algorithm", ***Journal of the Royal Statistical Society. Series B (Methodological)***, (1977), 1-38.
24. Keramatpour, M., Niaki, S.T.A., Soleymanian M.E. and Khedmati, M., "Monitoring and change point estimation of AR(1) autocorrelated of polynomial profiles", ***International Journal of Engineering***, Vol. 26, (2013), 933-942.
25. Fallahnezhad, M., Rasti, B. and Abooie, M., "Improving the performance of bayesian estimation methods in estimations of shift point and comparison with mle approach", ***International Journal of Engineering-Transactions C: Aspects***, Vol. 27, No. 6, (2013), 921-932.
26. Abdella, G., Yang, K. and Alaeddini, A., "Effect of location of explanatory variable on monitoring polynomial quality profiles", ***International Journal of Engineering-Transactions A: Basics***, Vol. 25, No. 2, (2012), 131-140.
27. Niaki, S.T.A., Abbasi, B. and Arkat, J., "A generalized linear statistical model approach to monitor profiles", ***International Journal Of Engineering Transactions A: Basics***, Vol. 20, No. 3, (2007), 233-242.

Evaluating Estimation Methods of Missing Data on a Multivariate Process Capability Index

TECHNICAL NOTE

A. Ashuri, A. Amiri

Industrial Engineering Department, Faculty of Engineering, Shahed University, Tehran, Iran

PAPER INFO

Paper history:
Received 12 March 2014
Received in revised form 10 August 2014
Accepted 18 September 2014

Keywords:
Process Capability Index
Missing Data
Imputation Methods, Response Variable
Main and Interaction Effects

چکیده

کیفیت محصولات یکی از مهم ترین مسائل برای تولیدکنندگان در دهه های اخیر می‌باشد. یکی از مباحث چالش برانگیز در این حوزه بررسی توانایی فرایند با استفاده از شاخص های قابلیت فرآیند است. از طرف دیگردر بسیاری از صنایع تولیدی داده های از دست رفته وجود دارد . تاکنون عملکرد روش های تخمین داده های از دست رفته روی شاخص های توانایی فرایند مورد ارزیابی قرار نگرفته است. بنابراین در این مقاله، ارزیابی عملکرد یک شاخص توانایی فرآیند زمانی که با داده های از دست رفته سروکار داریم مورد تحلیل قرار گرفته است. بدین منظور، یک شاخص توانایی فرایند چند متغیره در نظر گرفته شده و عملکرد چهار روش جایگزینی میانگین، ماکزیمم سازی ارزش انتظاری، رگرسیون و رگرسیون تصادفی برای تخمین داده های از دست رفته بررسی می شود. دراین تحلیل عواملی از قبیل درصد داده های ازدست رفته، اندازه نمونه، ضریب همبستگی و روش های تخمین داده های از دست رفته به عنوان فاکتورهای موثر در نظر گرفته شده است. اثرات اصلی و متقابل این فاکتور ها روی متغیر پاسخ که به صورت اختلاف بین شاخص تخمین زده شده و شاخص بدست آمده با داده های کامل تعریف شده است توسط مدل خطی تعمیم یافته در جدول آنالیز واریانس مورد ارزیابی قرار گرفته است. نتایج این تحلیل نشان می‌دهد که روش رگرسیون تصادفی دارای بهترین عملکرد بین روش های براورد داده های از دست رفته می‌باشد و درصد داده های از دست رفته بیشترین اثررا روی متغیر پاسخ دارد. همچنین اندازه نمونه کمترین اثر را در بین فاکتور های ذکر شده بر روی متغیر پاسخ دارد.

***doi**: 10.5829/idosi.ije.2015.28.01a.12*