THE EXPERIMENT
INTERNATIONAL JOURNAL OF SCIENCE AND TECHNOLOGY

ISSN-2319-2119

RESEARCH ARTICLE

F.C Eze et al, The *Experiment, 2013 Vol. 14(3), 997-1007*

BALANCING TWO-WAY UNBALANCED DATA WITH REPLICATION

ABSTRACT

Mean Substitution, Regression Estimation, Expectation Maximization and Multiple Imputation techniques were considered for imputing missing data in two-way design under the experimental assumption that the missing data follow monotone distribution with varying missing data densities. The performance, relative efficiency and suitability of the techniques were evaluated and compared using the standard error, the root mean square error and the relative efficiency index as statistical tools. The relative efficiency index is a better measure of efficiency and more powerful than the standard error and the root mean square error. Moreover, the Euclidean Distance is a sufficient measure of relative efficiency and appropriate for evaluation of trend of performance of techniques across levels of missing data. All techniques evaluated performed relatively more efficient at lower percent of missing data; therefore, the efficiency of the data imputation techniques decreases with increasing proportion of missing data. The Regression Estimation technique is the most stable and efficient method for intermediate and high density missing data; while the Mean Substitution and Multiple Imputation technique are most efficient for low density missing data.

KEYWORDS: TWO-WAY UNBALANCED DESIGN, REPLICATION, MISSING DATA, MONOTONE PATTERN OF MISSING DATA, IMPUTATION TECHNIQUES.

2.0 INTRODUCTION

Nowadays, Statistical researches are growing increasingly more multileveled and complex, and expectedly the designs of experiments are often unbalanced: treatments are often not orthogonal to blocks [11]; therefore, not all homogeneous block units have all replicates of all factor levels. Missing data result from random developments which the researcher has little or no control over; they are usually major constraints that may result in outright collapse of researches in spite of expended time and cost. The analyst only has to manage missing data and associated setbacks through the missing data methods to prevent research failure. Missing data problems present, superficially, as loss of some vital information about the study population so that the experimental data lack representation of the population of study resulting, therefore, in bias analysis with a rise in the likelihood of type1 and type II errors [2].

Moreover, major constraints stem from incomplete data: the problems vary in difficulty with the complexity of the design and degenerate with the proportion and the mechanism of the missing data. There is loss of an error degree of freedom for each missing datum [2]. The analysis may likely collapse when the ratio of the missing data to the available data is quite high; or, at least, according to [2] the precision, efficiency and power of the test statistic are lowered with high standard error (bias). Thus, the standard statistical tools are less adequate [13] for unbalanced designs. More so, [6] noted that the different missing data techniques differ in performance efficiency under the different patterns of missing observations: the challenges are relatively easier to manage when and where data miss at random (MAR) or completely at random (MCAR), a pattern described as monotone.

An intricate problem associated with unbalanced design is in connection with the formulation of F-ratio of the analysis of variance. The F-ratio cannot be formulated when the model is either fixed or mixed except for the variance components model according to the Welch-Satterthwaite equation. The question arises therefore what happens when the model is fixed or mixed in spite of the expended time, cost and resources? However, [4] developed a common denominator for unbalanced two-way interactive random model. The non-integer degree of freedom, as a result of using Welch-satterttwaite equation, was resolved to integer value by removing the interaction from the model. This led to the formulation of the F-ratio using the mean square error (MSe) as the common denominator for testing for the main effect.

A reliable approach in handling missing data problems is the use of imputation methods which impute estimates for the missing observations; some of the methods include the single methods: Mean Substitution (MS), Regression Estimation (RE), Last Observation Carried Forward (LOCF), Hot Deck (HD) approach, Nearest Neighbour Imputation (NI) and Expectation-Maximization (EM) algorithm; the Multiple Imputation method and the Model Based Analysis methods comprising Kernel Smoothing and Universal kringing [20]. But they vary in their estimations and efficiencies of estimation of the missing data under the different missing data patterns, and differing sizes of the missing data. Moreover each has unique advantages and drawbacks. Nevertheless, the task of making unbalanced design balanced with minimum estimation error and the imputed dataset thereby representing the study population is of foremost interest in the design and analysis of experiments, hence this research assesses the efficiency of a number of missing data imputation techniques, under the experimental assumption that the missing data followed the monotone pattern, to determine the most suitable imputation methods for handling problems of unbalanced design for a given missing data density.

3.0 METHODOLOGY

In order to meet the theoretical assumption of Normality for Analysis of Variance, normally distributed data were simulated in three replicates per cell. The complete dataset was randomly arrayed to 5%, 10%, 15%, 20% and 25% incomplete datasets, respectively each with monotone missing pattern. Therefore, the data for this study comprises a simulated normal 180-point dataset and five incomplete datasets with three replications of treatments per block-unit and of monotone pattern of 5%, 10%, 15%, 20% and 25% missing values, respectively. Let the observations be represented by the following data matrix

$$Y_{ijk} = \begin{pmatrix} y_{111}, y_{112}, y_{113} & y_{121}, y_{122}, y_{123} & \cdots & y_{1q1}, y_{1q2}, y_{1q3} \\ y_{211}, y_{212}, y_{213} & y_{221}, y_{222}, y_{223} & \cdots & y_{2q1}, y_{2q2}, y_{2q3} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ y_{p11}, y_{p12}, y_{p13} & y_{p21}, y_{p22}, y_{p23} & \cdots & y_{pq1}, y_{pq2}, y_{pq3} \end{pmatrix}$$

i = 1, 2... p; j = 1, 2... q; k = 1, 2, 3

The variance-covariance structure of cell means is estimated from the statistic:

$$\hat{\sigma}_{ii} = \sigma^2_{i.} = \frac{1}{p} \sum_{\alpha=1}^{n} y^2_{i\alpha} - \bar{y}_i^2, \forall i = j \ [12] \qquad \text{.... (3.3)}$$

$$\hat{\sigma}_{ij} = \frac{1}{p} \sum_{\alpha=1}^{n} y_{i\alpha} y_{j\alpha} - \bar{y}_i \bar{y}_j$$

Research methods, which assume the monotone pattern of missing data, employed in estimating and imputing for missing data comprise one ancient, the Mean Substitution (MS); one intermediate, the Linear Regression Estimation (RE) and two modern methods, the Expectation-Maximization (EM) Algorithm and the Multiple Imputation (MI).

3.1 The Mean Substitution (Ms) Technique

The mean of every cell observations was imputed for the missing values of the variable in the cell. Hence, for any variable j in the ijth cell with the kth replicate missing, MS imputes an estimate;

$$\hat{\bar{Y}}_{ijk} = \frac{1}{n-m}\sum_{k=1}^{n-m} y_{ijk}$$

3.2 The Linear Regression Estimation Technique

A regression model is fitted for any samples, m of the replicate 1 of any variable Y_j missing, the replicate with the missing data is taking as the dependent, y_{ij1} and the two co-replicates as the independent (X). According to [14] and [18], the missing data are predicted as follows from the model:

$$\hat{Y}_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + Z_1 \delta_{ij} \qquad \ldots. \quad (3.5)$$

$$Y_j = X\beta$$

Where,

$$\begin{bmatrix} Y_{11k} \\ Y_{21k} \\ . \\ . \\ . \\ Y_{(p-m)1k} \end{bmatrix} = \begin{bmatrix} X_{111} & X_{112} \\ x_{211} & x_{212} \\ . & . \\ . & . \\ . & . \\ x_{(p-m)11} & x_{(p-m)12} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_2 \end{bmatrix}$$

$$\beta = (X'X)^{-1}|(X'Y)$$

The $Z_1\delta_{ij}$ of the model is a random error added to every imputed value to ensure that variable elements with equivalent predictor values would receive different estimates [14], [18].

3.3 The Expectation-Maximization (EM) Algorithm

According to [21], the algorithm is used for computing the maximum likelihood (ML) estimates, in the presence of missing data, of the model parameters for which the observed data are most likely. It consists in two processes, which iterate until convergence occurs and increases the likelihood at all iteration:

The expectation or E-step in which the missing data are approximated based on the conditional expectation given the observed data and the current estimate of the parameter.

The maximization or M-step, which maximizes the likelihood function on the assumption that the missing data are known, already estimated from the E-step.

Let Z be a random vector of missing data variables, z a given variable with missing data, x a vector of the missing values and θ, a model estimate for x. we wish to find the maximum likelihood (ML) estimate θ (optimal value of x, defined L(θ)) for which p(x|θ), the total probability of a most likely estimate for any missing data is a maximum, according to [21] can be defined in terms of Z as,

$$P(x/\theta) = \sum_{1}^{n} P(x/z)P(z/\theta) \qquad \text{..... (3.7)}$$

And the total probability after nth iteration,

$$P(x/\theta_n) = \sum_{1}^{n} P(x/z)P(z/\theta_n)$$

Then,

$$L(\theta) - L(\theta_n) = \ln\sum_{1}^{n} p(x/z)p(z/\theta) - \ln\sum_{1}^{n} p(x/z)p(z/\theta_n)$$

Moreover, applying Jansen's inequality:

$$\ln\sum_{i=1}^{n} \lambda_i X_i \geq \sum_{i=1}^{n} \lambda_i \ln(X_i)$$

Deduces to

$$L(\theta) \geq L(\theta_n) + \iota(\theta/\theta_n)$$

Hence,

$$L(\theta) \geq \iota(\theta/\theta_n)$$

$$\theta_n = \theta_{n+1}$$

Where $L(\theta)$ is the likelihood function at convergence when the algorithm selects θ_{n+1} as the value of θ for which $\iota(\theta/\theta_n)$ is optimized since increasing $\iota(\theta/\theta_n)$ increases $L(\theta)$.

3.4 The Multiple Imputation Method

With the interest to preserving such important characteristics as the mean, the variance, the regression parameters, etc of the dataset as a whole, m >1 linear regression models are set up for every variable with missing data regressed on its covariates to create m plausible datasets which maintain the overall variability in the population while preserving relationships with other variables. According to [21], this restores the natural variability in the missing data and incorporates the uncertainty caused by estimation, as well as achieves maintenance of the original variability of the missing data by creating imputed values based on variables correlated with the missing data and the causes of missing of data. The Process consists in three iterative steps:

Creating different versions of regression model that would produce multiple imputes which are plausible representations of the data.

Performing a chosen statistical analysis on each of these imputed datasets: averaging over these versions guarantees more realistic imputes and reveals how they vary from one to the other to offer an estimate of the extra variance introduced by estimation, and

Combining the results of these analyses (averaging them) to produce a point estimate for each missing datum.

Let Y_i be a continuous variable with missing data, we set up m>1, usually between 3 and 10 regression models in practice, for Y_i as follows

$$\hat{Y}_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q + z_i \delta_{ij}$$
$$\hat{Y}_j = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_q x_q + z_i \delta_{ij}$$
$$\hat{Y}_j = \tau_0 + \tau_1 x_1 + \tau_2 x_2 + ... + \tau_q x_q + z_i \delta_{ij} \quad (20) \qquad .... (3,8)$$

Where

$z_i \delta_{ij}$ is a random error term [21]; α, β *and* τ are parameters for the different versions of regression model that would produce different $\hat{y}_j$ (estimates of Y_J). The point estimate $\hat{y}_j$ from the m>1 model is always more precise than any single estimate obtained from one model: this is an advantage of multiple imputation method over the regression method.

3.5 Assessment of Imputation Techniques

In order to assess satisfactorily the adequacy and comparative efficiency of the techniques investigated in this study, the standard error of the imputes, the Root-Mean-Square Error (RMSE) on the covariance matrix estimate and especially the relative Efficiency index between pairs of methods using their Euclidean Norm ratios were employed as test statistics.

3.5.1The Standard Error

The standard error is an unbiased measure of the dispersion of the estimates ($\hat{y}_{ijk}$) in the imputed dataset as estimated by a given imputation technique from the true observations (y_{ijk}) in the simulated dataset. The variance-covariance composition within the body of a dataset depends on the closeness of the individual observations. There will be no significant difference between estimated variances between the approximated datasets by the different methods if they perform at equal efficiency; otherwise, preference can be made between methods to one with lesser standard error of estimation of the missing data. The standard error of estimation is defined by the statistic

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{ijk} - \hat{y}_{ijk}\right)^2}{n}} \qquad \ldots.\ (3.9)$$

3.5.2 The Root-Mean-Square Error (RMSE)

The Root-Mean-Square Error is used, in this work, to evaluate the performance, that is, the accuracy of the imputation techniques in predicting the missing data with minimum variance. The RMSE for the $\hat{\boldsymbol{\Sigma}}$ is

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{ijk} - \hat{y}_{ijk}\right)^2}{pxp}} = \sqrt{\frac{\sum\left(\sigma_{ij} - \hat{\sigma}_{ij}\right)^2}{p}}\ [19] \qquad \ldots.\ (3.10)$$

where, σ_{ij} is the covariance of the ith and the jth levels of factors A and B in the ij cell of the original data covariance (Σ) matrix, and $\hat{\sigma}_{ij}$ refers to the elements of the covariance matrix, $\hat{\Sigma}$ of the imputed dataset and p is the order of the matrix. The RMSE is used in this study to evaluate the performance of the methods investigated in terms of their accuracy in predicting the elements of the covariance matrix of the complete dataset The method with least RMSE is most efficient among the four investigated methods and the second is next in the order.

3.5.3 The Relative Efficiency Index

The index, a ratio of the Euclidean distances - $\left\|\Sigma - \hat{\Sigma}_1\right\|^2$ and $\left\|\Sigma - \hat{\Sigma}_2\right\|^2$, is a measure of the overall comparative efficiency between a pair of methods (1 and 2) on their prediction of $\sum$, the variance-covariance composition of the original dataset. The Euclidean distance measures the closeness between $\sum$ and its estimate $\hat{\sum}$ for any two of the techniques. Therefore, the efficiency index provides a measure of how one imputation technique performs relative to the other technique in predicting the original $\sum$.

For illustration, the relative efficiency between the MS and the RE for estimation of $\sum$ is

$$Ref\left(\frac{\Sigma_{MS}}{\Sigma_{RE}}\right) = \frac{\left\|\Sigma - \hat{\Sigma}_{MS}\right\|^2}{\left\|\Sigma - \hat{\Sigma}_{RE}\right\|^2} = \frac{trace\left(\Sigma - \hat{\Sigma}_{MS}\right)\left(\Sigma - \hat{\Sigma}_{MS}\right)'}{trace\left(\Sigma - \hat{\Sigma}_{RE}\right)\left(\Sigma - \hat{\Sigma}_{RE}\right)'} \quad (22) \qquad \ldots..(3.11)$$

Ratio less than one shows that the technique in the numerator is more efficient than the other one in the denominator, while a ratio greater than one implies the technique in the denominator is more efficient and preferable and vice visa (21)

THE EXPERIMENT
INTERNATIONAL JOURNAL OF SCIENCE AND TECHNOLOGY

4.0 Data Analysis and Discussion

Table I: Standard Error

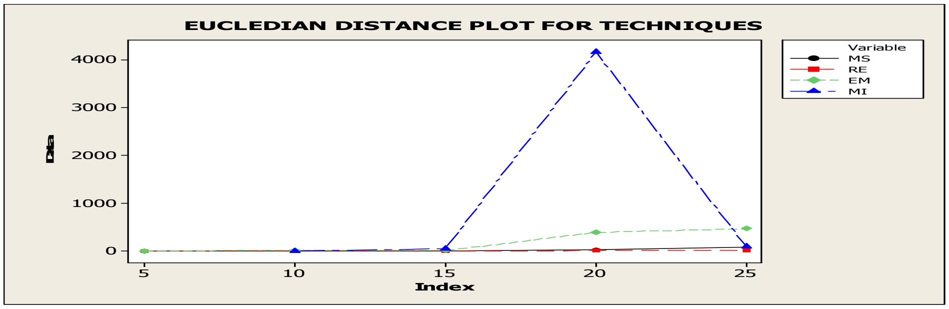
Method	5%	10%	15%	20%	25%
MS	1.78	3.98	5.31	4.56	5.31
Re	5.40	6.09	6.84	6.05	7.21
EM	6.39	4.64	6.48	5.86	5.75
MI	7.44	7.79	7.59	4.56	6.65

Table II: Root Mean Square Error

Method	Root Mean Square Error				
	5%	10%	15%	20%	25%
MS	0.1009	2.9831	1.4677	2.4490	2.1712
Re	0.7015	2.5546	1.6664	2.3883	1.5823
EM	0.2238	5.1643	1.5924	2.7740	2.5601
MI	1.5960	12.3696	2.0776	3.0951	2.3534

Table 111: Euclidean Distance

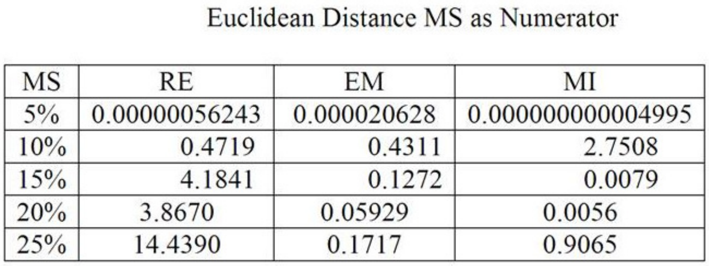
Method	Euclidean Norm				
	5%	10%	15%	20%	25%
MS	0.000000060542	4.23989281	0.364613614	23.15101565	79.97129408
RE	0.107635538	2.00081025	0.087130701	5.9086313201	5.538566917
EM	0.002934952	9.83512321	4.74678253	390.4889171	465.7207287
MI	121211.5531	1.54132225	46.04981569	4163.484733	88.21919577



ISSN-2319-2119

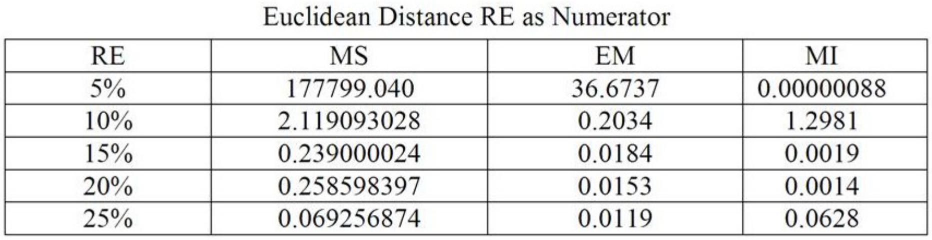
RESEARCH ARTICLE

F.C Eze et al, The *Experiment*, *2013 Vol. 14(3), 997-1007*



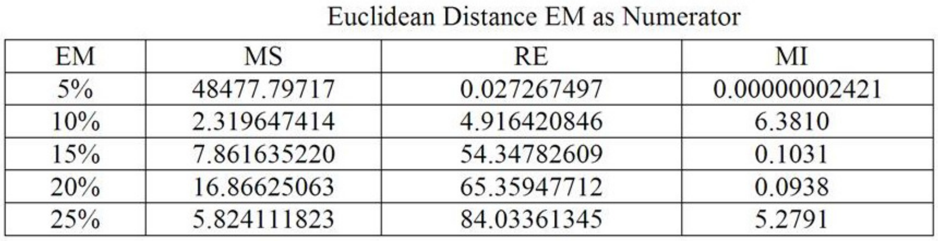
Euclidean Distance MS as Numerator

MS	RE	EM	MI
5%	0.00000056243	0.000020628	0.000000000004995
10%	0.4719	0.4311	2.7508
15%	4.1841	0.1272	0.0079
20%	3.8670	0.05929	0.0056
25%	14.4390	0.1717	0.9065



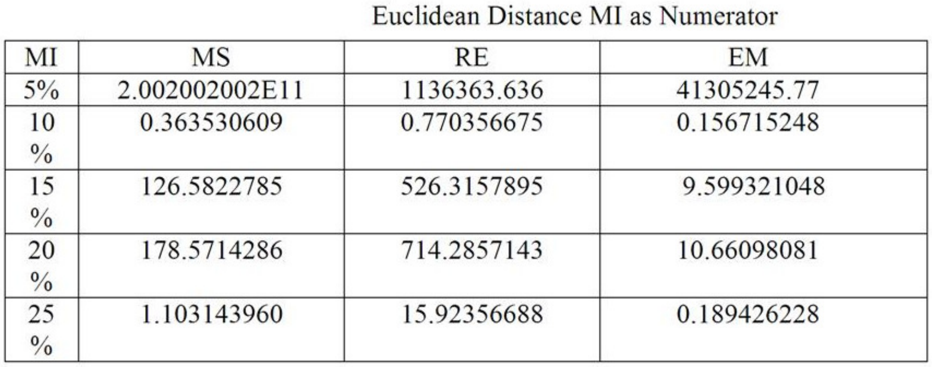
Euclidean Distance RE as Numerator

RE	MS	EM	MI
5%	177799.040	36.6737	0.00000088
10%	2.119093028	0.2034	1.2981
15%	0.239000024	0.0184	0.0019
20%	0.258598397	0.0153	0.0014
25%	0.069256874	0.0119	0.0628



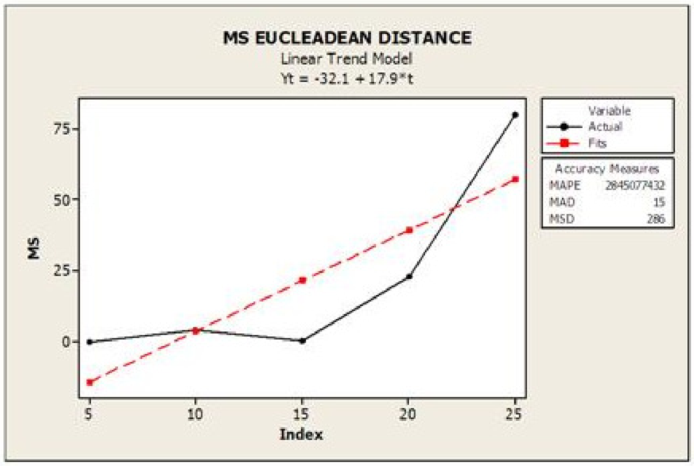
Euclidean Distance EM as Numerator

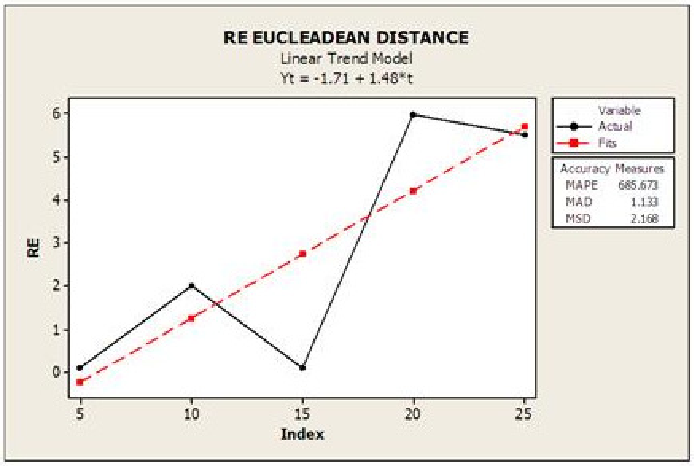
EM	MS	RE	MI
5%	48477.79717	0.027267497	0.00000002421
10%	2.319647414	4.916420846	6.3810
15%	7.861635220	54.34782609	0.1031
20%	16.86625063	65.35947712	0.0938
25%	5.824111823	84.03361345	5.2791

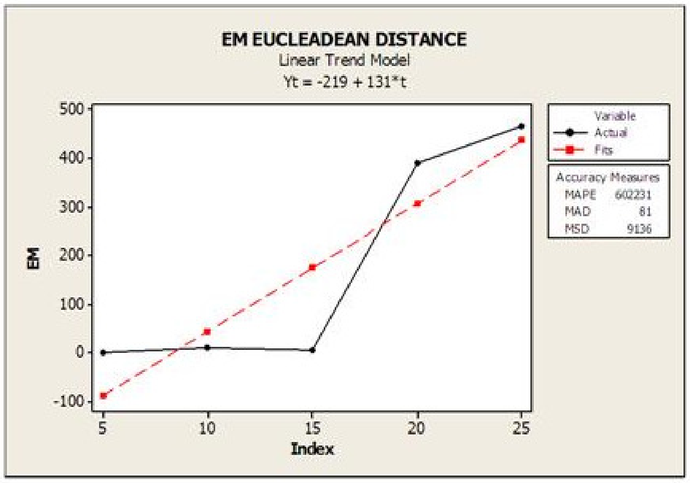


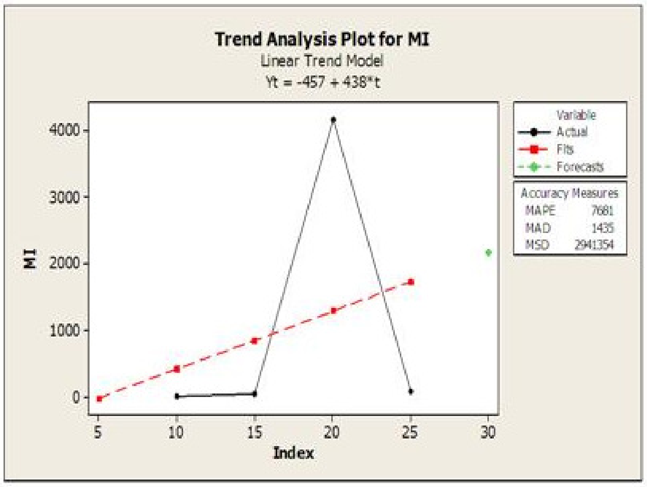
Euclidean Distance MI as Numerator

MI	MS	RE	EM
5%	2.002002002E11	1136363.636	41305245.77
10 %	0.363530609	0.770356675	0.156715248
15 %	126.5822785	526.3157895	9.599321048
20 %	178.5714286	714.2857143	10.66098081
25 %	1.103143960	15.92356688	0.189426228









Summary of Findings

The Mean Substitution method, followed by the Expectation Maximization method, was most efficient at five percent level of missing data.

The Multiple Imputation method, followed by the Regression Estimation was most efficient at ten percent level of missing data.

The Regression Estimation method was most efficient at fifteen, twenty and twenty-five missing data levels.

The Euclidean Distance trend-plot shows that the Mean Substitution, Regression Estimation and Expectation Maximization techniques except the Multiple Imputation fell in their performance efficiency with increase in missing data percentage; therefore, the performance of Multiple Imputation method improved steadily at higher missing data levels.

In concise terms, we conclude that the Relative Efficiency Index is a better measure of efficiency because it uses the ratio of the Euclidean Distances between estimates of the respective methods and that of the balanced data; thus making it more powerful than the standard error and Root Mean Square Error.

Moreover, the Euclidean Distance measure is easier than and as efficient as the Relative Efficiency index in measuring the relative efficiency of missing data techniques and their trend of performance across the levels of missing data.

All techniques evaluated performed relatively more efficient at lower percent of missing data except for the Multiple Imputation; therefore, the efficiency of the data imputation techniques decreases with increasing proportion of missing data.

The Mean Substitution and the Multiple Imputation methods are most efficient for low density missing data.

The Regression Estimation is the most efficient method for intermediate and high density missing data.

5.3 Recommendations

In line with the results of our effort so far, we wish to make the following recommendations:

Scholars and researchers should use the Euclidean Distance measure and the Relative Efficiency Index as measures of the efficiency of data imputation methods.

The Regression Estimation method should be used in preference to other data Imputation techniques for intermediate high density missing data.

For low-density missing data, Mean Substitution and Multiple Imputation techniques are most appropriate; but, the Mean Substitution is preferable for imputing for missing data at low missing data levels.

Interested researchers should use higher percentage of missing data, higher data size and/or different sizes of experimental variables to evaluate the performance of the data imputation techniques.

For a further work on this study, interested researchers may determine a possible division between low-density and high-density missing data to enhance choice of suitable data Imputation Techniques for any given case.

REFERENCES

1. Bernaards, C. A & Sijtsma, K. (2000). Influence imputation and EM methods on factor analysis when item nonresponse in questionnaire data is ignorable. Multivariate Behavioural Research, 35, pp. 321-364
2. Cool, A. L. (2000). A review of methods for dealing with missing data: Paper presented at the Annual Meeting of the Southwest Educational Research Association, Dallas, TX. (ERIC Document Reproduction Service, No. ED 438 311).
3. Eze, F. C. (2003). Fundamentals of Design and Analysis of Experiments (pp. 28-32). Awka, Mega Concepts.
4. Eze, F.C. and Chigbu, P.E. (2013). Presence of Interaction In An Unbalanced Two-Way Random Model. International Journal of Computing Engineering Research (ijceronline.com) vol.3-issue.2, E0320035.
5. Horton, N.J. and Lipchitz, S.R. (2001). Multiple imputation in practice: comparison packages for regression models with missing variables. The American Statistician, 55 (3), 244 -254.
6. Howell, D. C. (2008). The analysis of missing data. In W. Outhwaite & S. Turner. Handbook of Social Science Methodology. London: Sage.
7. Kaarik, E. (2005). Handling dropouts by copulas: WSEAS Transactions in Biology and Biomedicine. In N. Mastorakis (Ed), 1(2), pp. 93-97.
8. Kim, K. H. & Bentler, P. M. (2002). Test of homogeneity of means and covariance matrices for multivariate incomplete data. Psychometrika, 67(4), pp. 609-624.
9. King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American Political Science Review, 95, pp. 49-69.

10. Little, R. J. A. & Hyoggin, A. N. (2004). Robust likelihood-based analysis of multivariate data with missing values. Statistical Sinica, 14, pp. 949-968.
11. Montgomery, D. C. (2001). Design and analysis of experiments, (3ed). New York, John Wiley & Sons.
12. Ogum, G. E. O. (2002). Introduction to methods of multivariate analysis. Aba-Nigeria, Afri Towers Ltd.
13. Peng, C-Y.J., Harwell M., & Ehman, L. H. (2003). Advances in Missing Data Methods and Implications for Educational Research. Chao-Ying Joanne Peng, Department of Counseling and Educational Psychology, School of Education, Room 4050, 201 N. Rose Ave., Indiana University, Bloomington, IN 47405-1006. Retrieved from peng@indiana.edu.
14. Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state-of the-art. Psychological Methods, 7, pp. 147-177.
15. Schneider, T. (2000). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. Journal of climate (in press). Retrieved from http://links. Jstor.org
16. Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics (4th Ed.).Needham Heights, MA Allyn & Baco.
17. Utazi, C. E. (2010). The efficiency of some techniques for estimating covariance and correlation matrices from incomplete data. M.Sc Thesis, Anamdi AzikiweUniversity, Awka-Nigeria.
18. Wayman, J. C. (2003). Multiple imputations for missing data, what is it and how can I use it? A paper presentented at the 2003 Annual Meeting of the American Education Research Association, Chicago, IL. Retrieved from http:// www.csos.jhu.edu.
19. Yongsong, Q., Zhang, S., Zhu, X., Zhang, J. & Zhang, C. (2006). Semi-parametric optimization for missing data imputation. App. Intell. Retrieved from www.staff it.uts.edu.au.
20. Lokupitiya, R. S., Lokupitiya, E., Paustian, K., (2005). Environmetrics 2006; 17: 339–349 published online 7 November 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.773.
21. Borma, S. (2004). The Expectation Maximization Algorithm. A short tutorial. Em-tu@seanbyorman.com
22. Timm, N.H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. Psycometrika, 35 (4), pp.417-439.

***F.C Eze[1], V.O Ezenwoali[2]**

[1,2] Department of Statistics, Nnamdi-Azikiwe University, Awka, Anambra State, Nigeria.