African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 4, 2023 (pp. 1-11)
www.abjournals.org

THEORY AND APPLICATION OF TWO (2) ITERATIVE IMPUTATION APPROACHES TO NIGERIA ANNUAL RAINFALL DATA REPORTED

Ogbeide E.M.[1], Shuaibu M.[1] and Siloko U.I.[2]

[1]Department of Mathematics, Ambrose Alli University, Ekpoma, Edo State, Nigeria.

[2]Department of Mathematics, Edo State University, Uzairie, Edo State, Nigeria.

Cite this article:

Ogbeide E.M., Shuaibu M., Siloko U.I. (2023), Theory and Application of Two (2) Iterative Imputation Approaches to Nigeria Annual Rainfall Data Reported. African Journal of Mathematics and Statistics Studies 6(4), 1-11. DOI: 10.52589/AJMSS-VDC3AZVN

Manuscript History

Received: 17 June 2023

Accepted: 3 Aug 2023

Published: 24 Aug 2023

ABSTRACT: *This research work is based on missing data statistics. Missing data occur where one or more of the observations in a dataset are completely not available. This work focuses on two (2) iterative imputation approaches. These are the Regression approach and the Expectation Maximization iterative imputation. These approaches were used to analyze the secondary data of the thirty-six (36) states in Nigeria on the rainfall data collected from the Annual Abstract of Statistics 2016. The evaluation criteria and comparison of these two approaches were done based on the error efficiency using the Raw Bias (RB), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and variance. The analysis of the result showed that the Expectation Maximization (EM) method was better for this specific data as reported in the Annual Abstract of Statistics 2016, compared to the other approaches. This was seen in the smaller errors values from the computed cases. It is therefore recommended that this approach should be used for obtaining missing data like other rainfall data in Nigeria. These two imputation approaches are good for making available missing data in observations.*

KEYWORDS: Imputation, Expectation Maximization, Regression, Mean Squared Error.

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

INTRODUCTION

Missing data, also known as missing values, is where some of the observations in a dataset are not available in a dataset. Missing data are issues that most researchers in education, health and agriculture encounter on a daily routine. In survey research, there are many reasons for missing data such as respondents ignoring a few or all questions, or inability of survey administrations to locate the respondent. The most common cause of missing values in surveys is non-response, which is prevalent in any survey and can be severe. Non-response can be refusal to answer the survey completely (unit non response) or refusal to answer specific questions (item non response). There are many different methods of handling missing data which can have profoundly different effects on estimation. For this reason, it is important to select the correct missing data handling method that is suitable to a researcher's particular circumstances (Ogbeide & Osemwenkhae, 2014). So, some outlined approaches of handling missing data according to Carpenter and Kenward (2008) and Ogbeide (2018) are Least Squares (LS) method, the Expectation Maximization (EM) method and the Multiple Imputation method. This study is only restricted to the two iterative approaches of Expectation Maximization approach and the Regression Analysis approach. It is therefore important to use iterative methods in solving missing data problems as it can be implemented on a large size data and they have more efficiency and consistency (Little & Rubin, 2002; Ogbeide, 2018).

This study was undertaken with a view of finding a solution to the problem of missing data from the secondary data obtained from the National Bureau of Statistics of annual rainfall in the thirty-six (36) States in Nigeria between the period of 2011-2015 by using the Expectation Maximization and the Regression approach. This research aims to examine two (2) iterative imputations for obtaining missing data in Nigeria's rainfall data, collected from the National Bureau of Statistics (2016). The objectives are as follow:

i. To apply two iterative approaches to obtain missing data; and

ii. To examine the performance of the two (2) methods based on the criteria of Raw Bais (RB), Mean Squared Error (MSE), Rooted Mean Square Error (RMSE) and Variance.

LITERATURE REVIEW

The missing observation was first popularized by Rubin (1976). Little and Ruben (2002) came up with the classification system that is in use today, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). This mechanism describes the relationship between measured variables and the probability of missing data. While these terms have a precise probabilistic and mathematical meaning, there are essentially three different explanations for why the data are missing from a practical perspective. The mechanisms are assumptions that dictate the performance of missing data techniques, as we give a conceptual description of each mechanism in this section and supplementary resources are available (Schafer & Graham, 2002; Little & Rubin, 2002; Enders, 2010). The problem of missing data is almost omnipresent in both observational studies and randomization traits. Until the advent of sufficiently powerful computers, much of the research in this was focused on the problem of how to handle, in a practical way, the lack of balance caused by incompleteness in data recordings. There are iterative and non-iterative approaches for imputation of missing observation in missing data analysis (Little & Rubin, 2002). This work is based on the iterative

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

approach due to the fact that they are more robust for large size data with efficient comparison via mean square error (Little & Rubin, 2002; Ogbeide, 2018). Ogbeide (2020) gave an example of such development using a key idea of the Expectation Maximization (EM) approach according to Dumpster, Laird and Rubin (2007). Multiple Imputation is another approach.

Missing data occurs when there is no data value that is stored for the variable in an observation or dataset. Missing data occurs commonly and it affects the conclusions that are drawn from a dataset (Ghahramani & Jordan, 1995). The occurrence of missing data can be caused by the non-response of a respondent or respondent does not understand the question, incorrect measurement or human error, among others. Every survey question that has no answer is missing data. There is no perfect way to deal with missing data. Several research studies have concentrated on the impact of missing values in the dataset and its management. Treating the missing values is considered as an important task or step to take in the analysis since it improves the effectiveness of the knowledge discovery process. In the fields that are highly dependent on the data for decision making, missing data is still a problem that needs to be solved (Zha et al., 2013). Abdella and Marwala (2005) illustrate the methods that have been used to handle the missing values in areas such as research in statistics, mathematics and other different disciplines. The good way to handle missing data depends on how the data points have gone missing. The three types of missing data mechanisms are: Missing Completely at Random (MCAR), Missing at Random (MAR), and non-ignorable. MCAR occurs if the probability of missing value for variable X is not related to the value X or any other variable in the dataset. This only happens when the missing data does not depend on the variable of interest. MAR arises if the probability of missing data on a variable X depends on the other variables but not on X itself. Finally, the non-ignorable happens if the probability of missing data X is related to a value of X itself. This is the most difficult mechanism to approximate and model than the other two missing mechanisms.

Recently Vazifehdan et al. (2018) show that in the real world, real datasets often include missing values for various reasons. This is a major challenge when using the machine learning approach. Most of the learning algorithms cannot work with missing data. Imputation of missing values is very useful for unbiased predictions using machine learning tools (Schafer & Graham, 2002; Vazifehdan et al., 2018; Ogbeide, 2018). The types of imputations mentioned in the research above are mean imputation, hot-deck imputation, K-Nearest Neighbours imputation (KNNimpute), regression imputation, Bayesian estimation and Expectation Maximization (EM). Royston et al. (2004) show that the hot-deck imputation may perform poorly when many rows of data have at least one missing value. Troyanskaya et al. (2001) found that iterative approaches appear to yield more robust and sensitive results for missing value estimation in the dataset than non-iterative imputation approaches. The Expectation Maximization (EM) imputation approach is an iterative method. The basic idea of the EM is first to predict the missing values based on assumed values for the parameters, then use these predictions to update the parameters, and repeat until the sequence of parameters converges to maximum likelihood estimates.

Regression Analysis is also used to predict missing values based on the variable's relationship with other variables in the dataset (Royal & Herson, 1973). The first step consists of identifying the independent variables and the dependent variables (Madow & Tepping, 1982). In turn the dependent variable is regressed on the independent variable, the resulting regression equation is then used to predict the missing value. According to Graham et al. (1994), the advantage is that it preserves the variance and covariance structure of variables with missing data.

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 4, 2023 (pp. 1-11)

www.abjournals.org

Multiple Imputation (MI) approach is one of the most applicable methods for dealing with missing values in multivariate analysis (Abdella & Marwala, 2005; Little & Rubin, 2002; Ogbeide, 2020). The EM and the Regression Analysis approach in this work are based on the idea of MI where several optimal choice missing values are obtained. When missing data are encountered in reported data, the statistical inference from such a dataset will be affected, particularly when making predictions from the data, so there is a need for the data cleaning and corrections. A standard method of correcting missing data when data occur and are non-ignorable, according to Little and Rubin (2002), is imputation. This is necessary to correct abnormality in the data. So, this study applies these iterative approaches. For these reasons, it is important to select the correct missing data handling method that is suitable to a researcher in particular circumstances (Ogbeide, 2018). The approaches to be used are Expectation Maximization (EM) and Regression Analysis imputations method. This approach will be used in this study for analyzing the missing data.

METHODOLOGY

This section deals with the two statistical methods that will be employed in this work. These methods are the Regression and Expectation Maximization approaches in estimating missing observations. The secondary data of the thirty-six (36) state and capitals rainfall collected from the Nigerian Annual Abstract of Statistics (2016).

EXPECTATION MAXIMIZATION (EM) METHOD

The Expectation Maximization (EM) approach formalizes a relatively old ad hoc idea for handling missing data. According to Little and Rubin (2002), the process is as follows;

To begin with, replace missing values with estimated values. Further, estimate the parameters, and re-estimate the missing values assuming the new parameter estimates are correct. Then proceed to further re-estimate the parameters, iterating until convergence. Such methods are EM algorithms for models where the complete data log likelihood is computed. Given X_{ij} observed data with $X_{ij} = X_{obs} + X_{mis}$ the *log* likelihood is given by

$\ell(\theta|X_{obs},X_{mis})$. (3.1.1) *w*here, $X_{obs}$is the observed value, X_{mis} is the missing value, θ unknown parameter.

Generally, the log likelihood $\ell(\theta|X)$ itself needs to be estimated at each iteration of the algorithm after estimation of the expectation of the complete available data.

The E step finds the conditional expectation of the ‘‘missing data’’ given the observed data and current estimated parameters, and then substitutes these expectations for the ‘‘missing data.’’ Specifically, let $\theta^{(r)}$ be the current estimate of the parameter θ unknown fromX_{ij}. The E step of Expectation Maximization finds the expected complete-data log likelihood of θ. According to Little and Rubin (2002), the expectation is given by:

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 4, 2023 (pp. 1-11)

$$E\left(\ell(\theta|X_{obs},X_{mis})\right) = E\left(X_{obs} + (n-r)\mu^{(r)}\right)$$

(3.1.2)

The value estimated from this step is substituted into the original data set with missing observation using equation (3.1.1).

The M step is particularly simple to describe Maximum Likelihood (ML) of the "E" step imputation. It performs ML estimation of θ just as if there were no missing data, that is, as if they had been filled in. Thus the M step of EM uses the identical computational method as ML estimation from $\ell(\theta|X)$.The M step of Expectation Maximizations determines $\theta^{(r+1)}$ by maximizing this expected complete-data log likelihood of the data set.

$$Q\left(\theta^{(r+1)}|\theta^{r}\right) \geq Q(\theta|\theta^{r}),$$

(3.1.3)

for all θ.where $Q\left(\theta^{(r+1)}|\theta^{r}\right)$ is the partial derivative of the log likelihood of X_{ij} containing the unknown θ in the current estimated sequence r. The equation (4) is computed repeatedly until convergence is achieved. For the problem of rate-of-convergence, see Dempster et al. (1977), Wu's (1983), and Little and Rubin (2002)

Regression Imputation Method

Regression imputation, also known as conditional mean imputation, replaces missing observations with predicted scores from a regression equation. The basic idea behind this approach is intuitively appealing. It uses information from the complete variables to fill in the incomplete variables. Variables tend to be correlated, so it makes good sense to generate imputations that borrow information from the observed data. In fact, borrowing information from the observed data is a strategy that regression imputation shares with maximum likelihood and multiple imputation, although the latter approaches do so in a more sophisticated manner.

The first step of the imputation process is to estimate a set of regression equations that predict the incomplete variables from the complete variables. A complete-case analysis usually generates these estimates. The second step is to generate predicted values for the incomplete variables. These predicted scores fill in the missing values and produce a complete dataset. The resulting regression equation is

$$\hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 X \qquad (3.2.1)$$

To illustrate this approach, consider a hypothetical dataset with three variables, y_1, y_2, and y_3, all of which have missing data. There are six possible missing data patterns: cases with missing data on (1) only y_1, (2) only y_2, (3) only y_3, (4) y_1 and y_2, (5) y_1 and y_3, and (6) y_2 and y_3. The presence of multiple missing data patterns requires the imputation process for each missing data pattern with a unique regression equation. Another way to construct the equations is to start with an estimate of the mean vector and the covariance matrix because the elements in these matrices define all of the necessary regression coefficients. Again, a complete-case analysis usually generates $\hat{\mu}$ and $\hat{\Sigma}$. Substituting the observed scores into the relevant regression

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 6, Issue 4, 2023 (pp. 1-11)

www.abjournals.org

equations generates predicted values for the incomplete variables, and these predicted scores fill in the missing values and produce a complete data set (Little & Rubin, 2002; Olinsky et al., 2003).

3.3.1 Raw Bias (RB) $RB = E(Q) - Q$ (3.3.1)

3.3.2 Mean Square Error (MSE) $MSE = \Sigma(Y_i - Y)^2 = (E(Q) - Q)^2$ (3.3.2)

3.3.3 Root Mean Square Error (RMSE) $RMSE = \sqrt{(E(Q) - Q)2}$ (3.3.3)

The methods of regression and EM above will be applied to the secondary data in the next chapter. The error estimate will be evaluated for suitability via the RB, MSE, RMSE and Variance.

DATA PRESENTATION AND RESULTS

This section deals with the application of the methods of Regression and Expectation Maximization in analyzing Nigeria's rainfall data reported in the National Bureau of Statistics in the missing value in the annual abstract of Statistics (2016).

Table 4:1 Annual Rainfall in Nigeria by State (2011-2015)

States and capitals	2011	2012	2013	2014	2015
Abia (Umuahia)	136	144	2161	2268	1980
Adamawa (Yola)	827	469	873	930	718
AkwaIbom (Uyo)	2532	2106	2411	2232	1912
Anambra (Akwa)	2027	2057	1545	1790	2273
Bauchi (Bauchi)	1137	1133	1054	965	1621
Bayelsa (Yenagoa)	-	-	2444	2543	-
Benue (Makurdi)	1340	1051	1189	1046	1402
Borno (Maiduguri)	1076	601	669	540	588
Cross River (Calabar)	3428	3061	2986	2761	2522
Delta (Asaba)	1802	1765	1859	1681	1766
Ebonyi (Abakaliki)	-	-	1230	1349	-
Edo (Benin)	2648	2670	2226	2096	2123
Ekiti (Ado-Ekiti)	-	-	2309	1490	-
Enugu (Enugu)	1911	1738	1839	1610	1757
Gombe (Gombe)	834	986	787	879	857
Imo (Owerri)	2362	2818	2386	2117	2738
Jigawa (Dutse)	-	-	670	1051	-
Kaduna (Kaduna)	865	828	1189	1118	1268
Kano (Kano)	113	109	981	931	-
Katsina (Katsina)	704	557	537	562	474
Kebbi (Birnin-Kebbi)	887	1223	668	738	1196
Kogi (Lokoja)	1531	1260	1169	1053	1632
Kwara (Ilorin)	1309	1469	1191	1128	1352

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

Lagos (Ikeja)	1649	1816	1722	1401	1392
Nassarawa (Lafia)	1570	10719	1120	1179	1566
Niger (Minna)	1423	1269	1185	1157	1422
Ogun (Abeokuta)	876	1372	1500	1353	1466
Ondo (Akure)	1406	1466	1424	1430	1310
Osun (Oshogbo)	1422	1598	1341	1336	1278
Oyo (Ibadan)	1219	889	1406	1190	1702
Plateau (Jos)	1357	1260	1244	1270	1237
Rivers (Port Harcourt)	2865	1607	2478	2212	2602
Sokoto (Sokoto)	636	515	617	669	603
Taraba (Jalingo)	1071	1038	-	-	1569
Yobe (Damaturu)	438	320	667	-	367
Zamfara (Gusua)	616	954	884	921	1006
FCT Abuja	1389	1175	1474	1465	1445

Source: *Annual Abstract of Statistics (2016)*

Regression Approach Method

Regression imputation replaces missing values by predicted values from a regression of missing items on items observed for the unit, usually calculated from units with both observed and missing variables present.

Table 4.2: Estimation of missing observation using the Regression approach

States and capitals	2011	2012	2013	2014	2015
Abia (Umuahia)	136	144	2161	2268	1980
Adamawa (Yola)	827	469	873	930	718
AkwaIbom (Uyo)	2532	2106	2411	2232	1912
Anambra (Akwa)	2027	2057	1545	1790	2273
Bauchi (Bauchi)	1137	1133	1054	965	1621
Bayelsa (Yenagoa)	**2430**	**1469**	2444	2543	**2447**
Benue (Makurdi)	1340	1051	1189	1046	1402
Borno (Maiduguri)	1076	601	669	540	588
Cross River (Calabar)	3428	3061	2986	2761	2522
Delta (Asaba)	1802	1765	1859	1681	1766
Ebonyi (Abakaliki)	**1244**	**2057**	1230	1349	**1009**
Edo (Benin)	2648	2670	2226	2096	2123
Ekiti (Ado Ekiti)	**2614**	**1269**	2309	1490	**2531**
Enugu (Enugu)	1911	1738	1839	1610	1757
Gombe (Gombe)	834	986	787	879	857
Imo (Owerri)	2362	2818	2386	2117	2738
Jigawa (Dutse)	**415**	**1372**	670	1051	**556**
Kaduna (Kaduna)	865	828	1189	1118	1268
Kano (Kano)	113	109	981	931	**1814**
Katsina (Katsina)	704	557	537	562	474
Kebbi (Birnin-Kebbi)	887	1223	668	738	1196
Kogi (Lokoja)	1531	1260	1169	1053	1632

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

Kwara (Ilorin)	1309	1469	1191	1128	1352
Lagos (Ikeja)	1649	1816	1722	1401	1392
Nassarawa (Lafia)	1570	10719	1120	1179	1566
Niger (Minna)	1423	1269	1185	1157	1422
Ogun (Abeokuta)	876	1372	1500	1353	1466
Ondo (Akure)	1406	1466	1424	1430	1310
Osun (Oshogbo)	1422	1598	1341	1336	1278
Oyo (Ibadan)	1219	889	1406	1190	1702
Plateau (Jos)	1357	1260	1244	1270	1237
Rivers (Port Harcourt)	2865	1607	2478	2212	2602
Sokoto (Sokoto)	636	515	617	669	603
Taraba (Jalingo)	1071	1038	**1598**	**1991**	1569
Yobe (Damaturu)	438	320	667	**538**	367
Zamfara (Gusua)	616	954	884	921	1006
FCT Abuja	1389	1175	1474	1465	1445

The bold data are the regression imputation values for the rainfall data

Expectation Maximization Method

Expectation-Maximization (EM) algorithm is an iterative method of finding maximum likelihood estimates of the parameters in statistical models, where the model depends on unobserved variables. The technique iteratively goes through the data while preserving the covariance structure of the data. The Expectation Maximization was carried out through the SPSS software.

Table 4.3: Estimation of missing observation using the expectation maximization

States and capitals	2011	2012	2013	2014	2015
Abia (Umuahia)	136	144	2161	2268	1980
Adamawa (Yola)	827	469	873	930	718
AkwaIbom (Uyo)	2532	2106	2411	2232	1912
Anambra (Akwa)	2027	2057	1545	1790	2273
Bauchi (Bauchi)	1137	1133	1054	965	1621
Bayelsa (Yenagoa)	**2096**	**2749**	2444	2543	**2458**
Benue (Makurdi)	1340	1051	1189	1046	1402
Borno (Maiduguri)	1076	601	669	540	588
Cross River (Calabar)	3428	3061	2986	2761	2522
Delta (Asaba)	1802	1765	1859	1681	1766
Ebonyi (Abakaliki)	**938**	**1745**	1230	1349	**1378**
Edo (Benin)	2648	2670	2226	2096	2123
Ekiti (Ado Ekiti)	**4508**	**1038**	2309	1490	**1978**
Enugu (Enugu)	1911	1738	1839	1610	1757
Gombe (Gombe)	834	986	787	879	857
Imo (Owerri)	2362	2818	2386	2117	2738
Jigawa (Dutse)	**294**	**1721**	670	1051	**980**
Kaduna (Kaduna)	865	828	1189	1118	1268

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

Kano (Kano)	113	109	981	931	**1036**
Katsina (Katsina)	704	557	537	562	474
Kebbi (Birnin-Kebbi)	887	1223	668	738	1196
Kogi (Lokoja)	1531	1260	1169	1053	1632
Kwara (Ilorin)	1309	1469	1191	1128	1352
Lagos (Ikeja)	1649	1816	1722	1401	1392
Nassarawa (Lafia)	1570	10719	1120	1179	1566
Niger (Minna)	1423	1269	1185	1157	1422
Ogun (Abeokuta)	876	1372	1500	1353	1466
Ondo (Akure)	1406	1466	1424	1430	1310
Osun (Osogbo)	1422	1598	1341	1336	1278
Oyo (Ibadan)	1219	889	1406	1190	1702
Plateau (Jos)	1357	1260	1244	1270	1237
RiverS (Port Harcourt)	2865	1607	2478	2212	2602
Sokoto (Sokoto)	636	515	617	669	603
Taraba (Jalingo)	1071	1038	**1486**	**1438**	1569
Yobe (Damaturu)	438	320	667	**663**	367
Zamfara (Gusua)	616	954	884	921	1006
FCT Abuja	1389	1175	1474	1465	1445

The bold data are the Expectation Maximization imputation values for the rainfall data.

There are some measures that may inform us about the statistical validity of a particular procedure. These evaluation criteria include the Raw Bias (RB), the Mean Squared Error (MSE), and the Root Mean Squared Error (RMSE).

Table 4.4: Error estimate from the various approaches

Approach	RA	EM
RB	0.82000	0.80000
MSE	0.59744	0.55604
RMSE	0.77294	0.74568
VARIANCE	0.04103	0.03967

The error, performance evaluation using the Regression and the Expectation Maximization approaches are presented in Table 5 above. The missing observation was estimated using Raw Bias (RB), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) and all the statistical procedures were carried out on SPSS software. From the analysis carried out, it shows that the Expectation Maximization approach had the lowest MSE of 0.55604, which is the best performance approach for this specific data when used to evaluate missing observation imputation for the Nigeria's rainfall data collected as reported by the National Bureau of Statistic (2016). There was also a relatively smaller bias of 0.800 for the Expectation Maximization approach compared to the 0.8200 of the Regression. These two iterative imputation approaches are very apt in recovering data in a non-ignorable setting. This further helps in better statistical inference from the dataset with imputed data.

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

CONCLUSION

This work deals with the theory and application of two (2) iterative imputation approaches in analysis of missing data of Nigeria's rainfall of the thirty-six (36) states and their capital as reported in the Annual Abstract of Statistics (2016). We compared the performance of two (2) iterative imputation approaches; that is the Expectation Maximization and the Regression approach. When the error efficiency of the two approaches was measured via Raw bias (RB), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and variance, it was discovered that the Expectation Maximization approach was better suitable for this specific data as it yield the least error as shows in the mean squared error and the variance in the section 4. Since missing data reduces the quality of the inference in the dataset, more attention should be paid to the missing data recovery in the design, collection and analysis of data which affects the performance inference of the data under study. We strongly recommend that the Expectation Maximization approach should be used in obtaining data relating to rainfall as it is a way to find the maximum-likelihood estimate for model parameters when the data is incomplete.

REFERENCE

Abdella, M. and Marwala, T. (2005): The use of genetic algorithms and neural networks to approximate missing data in database. In Computational Cybernetics, 2005. ICCC 2005. IEEE 3rd International Conference. Pages 207–212. IEEE, 2005.

Acock, .A. C. (2005): Working with missing value. Journal of marriage and family 67:1012-1028.

Allison, PG (2001): Missing data-Quantitative applications in the social sciences. Thousand Oaks.CA;Sage. 7(1); 9-19

Allen, F.G and Wishart, J. (2008): A method of estimating the yield of missing plot field experiment J. Agric. Sci 20:399-406.

Alosh, M.(2009): The impact of missing data in a generalized integer value auto regressing model for count data .Journal of Biopharmaceutical statistics 19: 1039-1054.

Aloha, C.F. and Harrell ,F.E. (2006): An introduction to S and the Hmisc and design libraries.

Annual Abstract of Statistics (2016): Table 4: Annual Rainfall in Nigeria by state, 2011-2015. National Bureau of Statistics, Abuja, Nigeria.

Brandi,A.N. and Enders ,C.K.(2010): And introduction to modern missing data analysis.Journal of School psychology 48:(1),5-37 DO1 :1016/j.j. sp.2009.10.

Carpenter,J.R. and Kenward,M.G.(2008): Missing data in clinical trials.-a practical guide. National Health Service Coordinating Centre for Research Methodology : Birmingham,UK.

Carpenter J.R. and Kenward ,M.G.(2013): Multiple Imputations and its Application. Chichester, Wiley .

Collins,L.M;Schafer, J.L. and Cam, C.M(2001): A comparison of inclusive and restrictive strategies in modern missing data procedures.Psychol Meth 6(4):330-351. Doi:10.1037/1082-989X.6.4.330.

Demitasse, H ;Frees, S. and Yucca, R .(2008): Plausiblity of multitative normality assumption when multiply imputing non-Gaussian continuous outcomes. A simulation assessment. Journal of Statistical Computation and simulation 78:69-84.

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN

Dempster. A. P.,Laird, N.M. and Rubin, D. B. (1977): Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological), 6(4); 1-38

Dumpster, A.P. Laird N.M. and Rubin, D.B. (2007): Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the royal statistical society, Series B, 39,1-38..

Enders, C.K .(2001): A primer on Maximum Likelihood Algorithm available for use with missing data.

Enders, C.K. (2010): Applied missing data analysis. New York; Guilford Press.

Grooves. R. and fowler,F.(2004): Survey Methodology,New York: Wiley and sons Inc.

Herel, O. Zhou, X. H. (2007): Multiple Imputation review of theory of implementation and software. Statistics in Medicine (in press).

Hlalele .N U (2009): The impact of missing data imputation on HIV classication. PhD thesis, 2009. (Karangwa et al. 2016)

Little,R.J.A.(1976): Inference about means from incompelete multivariatedata.Biometrika 63.593 -604

Little,R.J.A. and Rubin,D,B (2002): Statistics analysis with missing data, 2nd New York: Wiley and sons Inc.

Nekouie .A and Moattar .M .H 2018): Missing value imputation for breast cancer diagnosis data using tensor factorization improved by enhanced reduced adaptive particle swarm optimization. Journal of King Saud University-Computer and Information Sciences, 2018.

Ogbeide, E.M. (2018):A new imputation based on Expectation Maximization of the data set.Science and Technology journal. 3(1);139-142

Ogbeide, E.M. (2020): Imputation statistics value for missing data in Nigeria crime rate reported. Journal on Demography and Social Statistics. 7(1); 9-16

Ogbeide, E.M. and Osemwenkhae, J.E. (2014): A new imputation method based on Expectation Maximization of the dataset. Proceedings of the statistics Research Group (SRC), National conference, university of Benin, Nigeria.

Peng, C.Y.J.; Harwell, M.; Liou, S.M. and Ehman, L.H. (2006): Advance in missing data methods and implications for education research in sawilowsky ed. Greewhich, CT: information Age. Real Data Analysis.

Petrozziello .A and Jordanov .I (2017): Column-wise guided data imputation. Procedia Computer Science, 108:2282–2286, 2017.

Pigot, T.D. (2001): A Review of methods for missing data, Educational Research and Evaluation 7:353-383.

Raghunathan, T. E. (2004): What do we do with missing data? Some options for analysis of complete data. Annual review of public health 25: 88-117.

Rubin,D.B (2004): 'Basic ideas of multiple imputation on non-respond', survey Methodology, vol.12, No1, pp.37-14.

Schafer, J.L and Graham, J.W (2002): "Missing Data: our view of state of the state of the Art" psychological methods 2: 147-149.

Themes, F. and Enders, C.K. (2007): A structural equation model or testing whether data are missing completely at random. Paper presented at the annual meeting of the American Educational research Association, Chicago.

Article DOI: 10.52589/AJMSS-VDC3AZVN
DOI URL: https://doi.org/10.52589/AJMSS-VDC3AZVN