Lithuanian Journal of Statistics
Lietuvos statistikos darbai
Vol. 61, 2022, pages 16–31
https://doi.org/10.15388/LJS.2022.31573

Vilnius
University
Press

Filling the gap in food and nutrition security data: What imputation method is best for Africa's food and nutrition security?

Adusei Bofa, Temesgen Zewotir

School of Mathematics, Statistics, and Computer Science, University of KwaZulu Natal
Westville campus, Durban, 4041, South Africa

E-mail: 221119873@stu.ukzn.ac.za; zewotir@ukzn.ac.za

Received October 26, 2022; revised December 5, 2022; published online December 20, 2022

Abstract. Our study presents the methods adopted to produce accurate imputed values for Africa's food security and nutrition (FSN). We focused primarily on the following five imputation methods for handling missing data: Mean Imputation; Multiple Imputed values using a Chained Equation (MICE); imputation based on the Conditional Distribution of a variable Diagnostics (mi); Additive Regression and Predictive Mean Matching (Hmisc); and Random Forest (missForest). We describe each method, including how they performed under MAR and MNAR using RMSE and MAE as a measure of accuracy. After these methods of imputation were examined for nonignorable missing values in the context of accurate and unbiased estimates for food and security analysis, we found that MissForest handled nonignorable missing values more effectively and with less bias, increasing the precision of the data by imputing the closest data values within the dataset. Hence the missForest is the best alternative for handling missing values for food security and nutrition concerning Africa. This study adds to the current body of knowledge on food and nutrition insecurity and provides useful information to policymakers, particularly about the imputations of missing values aimed at food security and nutrition concerning Africa, which has significant economic and social ramifications.

Keywords: food security and nutrition; missForest; missing value; multiple imputations; Missing at Random (MAR); Missing Not at Random (MNAR); Africa

1 Introduction

Quality monitoring of Food Security and Nutrition (FSN) has enormous fiscal and public health benefits, thereby contributing toward the success of Goal 2 of sustainable development to end hunger, achieve food security and improve nutrition, and promote sustainable agriculture [29]. Once again, food security is a critical outcome of sustainable agricultural and food systems. Despite this, approximately one half of the globe's populace remains affected by malnutrition and food insecurity; a warning sign of the food system's current dysfunctions [11]. FAO [14] explained how food security and nutrition (FSN) are critical for many individuals such as farm households in developing countries, mainly those encountering insufficiency in water and increased climate variability.

More often the usual challenge encountered when analyzing data is missing values. Efron [10] specified that missing data is something that causes trouble or difficulty because of the absence of some data elements in the informed data composition. Liu *et al.* [24] indicated that missing values in a dataset enormously affect the outcome inferred from the dataset. Morris *et al.* [26] described that when the percentage of missing data is more than 60 such problems become challenging, and prevailing approaches have substantial difficulty in dealing with such situations.

A major challenge is how to deal with missing values on food security and nutrition (FSN) data concerning Africa. The findings on FSN can be biased because of the incidence of missing data, irrespective of how uncommon it is. Researchers aspire to get more complete data without missing values and hence must plan to remove or replace (i.e. impute for) missingness within the data. Missing data may lead to biased estimates of parameters and increase their standard error estimates. Also, the statistical power of tests might be weakened, and the sample may not be a true reflection of the population [5]. Hence It is essential to identify the most effective method for estimating missing values.

The authenticity of FSN research could be jeopardized if data is missing. When only cases with complete records are included in the analysis, power is lost and the results can be misleading. If nonignorable missing data patterns are not well-addressed, parameter estimates can be distorted, limiting the representativity of the data and the application of rigorous statistical analysis [17].

The missing data mechanism and the percentage of missing values may alter over time in food and nutrition security analyses. The appropriate multiple imputation method for handling missing values in food and nutrition security data is still up for debate. In Africa, a major difficulty is the lack of a high-quality large-scale dataset with no missing values to benchmark food and nutrition security, allowing for more comprehensive policy and systematic evaluation. To the best of our knowledge most existing literature used simulated missing values in evaluating the performance of methods that handle missing values, and few works are related to food and nutrition security problems (see [9, 26]). To perform efficient multiple imputations with a high level of missing data, one must take into account the mechanisms and extent of the missing data.

The challenge now is to know if the missing values within Africa's food security and nutrition data are nonignorable or ignorable, and how purportedly efficient multiple imputation methods perform with real missing values. Again, which multiple imputation method is suitable to handle nonignorable and ignorable missing values and hence improve the statistical power for analyzing food and nutrition security? And so, our primary aim for this study was to identify the pattern of missing values and their missingness mechanism, and the optimum model for analyzing missing values for a measured variable related to food security and nutrition in Africa, leading to better analysis or outcomes (food and nutrition security) for effective planning. Our study will provide researchers with a framework for imputation, allowing them to use a modern strategy for handling missing values to avoid inaccurate results when evaluating missing data in a variety of disciplines, including FSN.

2 Data

This study seeks to examine a case of handling missing values in the area of food security and nutrition. The United Nations Food and Agriculture Organization (FAO) is tasked with the obligation of making available variables for assessing global food security and nutrition. The data set compiled by FAO from all African countries for monitoring nutrition and food security contains missing data. The data accessed from FAO contains 42 variables measuring Africa's food security and nutrition from 2000 to 2019, representing 19 years. The dataset has been categorized into six groups of indicators namely accessibility, availability, stability, utilization, featured, and others (Fig. 1). For simplicity, these 42 variables were assigned names (Table 2 in Appendix A).

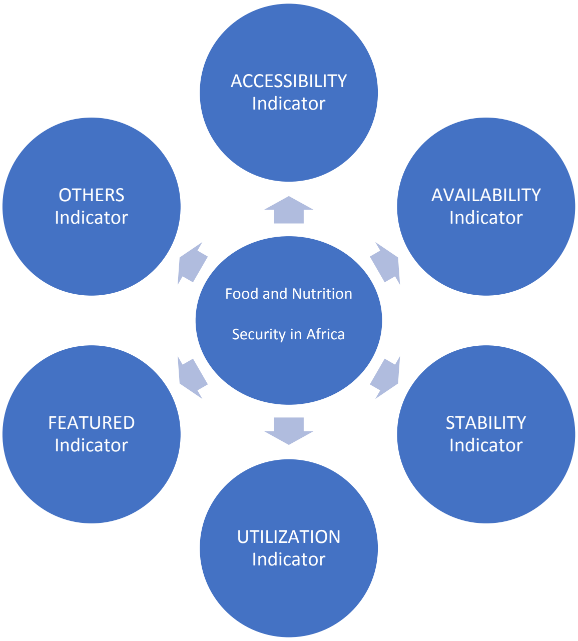

Fig. 1. Predictors of food security and nutrition classification from FAO.

3 Methods: theoretical concept

3.1 Missing data

In statistical analyses, it is essential to identify indicators that potentially lead to missing data. Efron [10] describes missing data as an issue created by the nonappearance of a familiar data structure. Little and Rubin [23] categorized missing data into three categories based on the plausibility of missing data mechanisms.

Missing Completely at Random (MCAR) means that the probability of an observation being missing does not depend on the data [30]. It is measured as a special case of 'Missing at Random' once there is no regular difference among the variables with complete data and missing data. Bache-Mathiesen *et al.* [5] defined MCAR as data with the same probability of missing all variables in the dataset. We presume that the missing values are not systematically different from the observed values. The positions of missing values in the dataset are totally random and do not depend on any other data. Given that ω is an $n \times p$ matrix that includes n cases with all p variables in the sample dataset. Let denote observed values as (ω_{obs}), while the missing values are denoted as (ω_{miss}). The missing values location in ω are indicated by the matrix R. The observations of ω and R are denoted as ω_{ij} and r_{ij}, respectively. Therefore, $r_{ij} = 1$ when ω_{ij} is observed, whereas $r_{ij} = 0$ once ω_{ij} is missing. R depends on $Y = (\omega_{obs}, \omega_{miss})$. If the missing values within the dataset are presumed to be MCAR, we can write $Pr(R \vee \omega_{miss}, \aleph)$ as

$$Pr(R = 0 \,|\, \omega_{obs}, \omega_{mis}, \aleph) = Pr(R = 0 \vee \aleph), \tag{1}$$

where $\aleph$ comprises the parameters of the missing data in the model. This means that the probability of missing a data value is solely determined by the model's estimated parameters.

Missing At Random (MAR) means the probability of an observation being missing may depend on observed information but not on unobserved information [30], i.e. determinedly linked to observed data and not to unobserved data. MAR is also well-defined as a condition where the probability of missingness remains the same within observed data variables [31]. The observed data is influenced by the positions of missing values in the dataset. That is,

$$Pr(R = 0 \,|\, \omega_{obs}, \omega_{mis}, \aleph) = Pr(R = 0 \vee \omega_{obs}, \aleph). \tag{2}$$

Missing Not at Random (MNAR) means the probability of an observation being missing is conditional on the unobserved data. Mack *et al.* [25] nicely defined MNAR by way of missing data not linked to any computable factors or events. The position of missing values in the dataset affects the missing values themselves. That is

$$Pr(R = 0 \,|\, \omega_{obs}, \omega_{mis}, \aleph). \tag{3}$$

Especially after 2005, several studies on missing value management for the application of statistical models have been undertaken over the years. Newer breakthroughs were developed when statisticians (researchers) progressed in their knowledge of this topic. For decades, missing value handling using imputation has been the standard approach for data samples with one or more missing attribute values. Lin and Tsai

[20] evaluated 111 journal papers published between 2006 and 2017 and identified various technical concerns encountered during the Missing value imputation procedure, missingness rates, and the missingness mechanism from an experimental design perspective (see [20]).

3.2 Imputation methods

Imputation is the procedure of substituting meaningful estimates for missing values. Researchers are often tempted to delete variables or observations with missing values; however, this can result in information loss, which can have an adverse effect on the results. Another possibility is pairwise deletion, in which complete cases of relevant observations or variables are analyzed; consequently, the sample size varies for different explanatory variables.

The incomplete dataset (the dataset having missing values) is copied several times in the first step. In each dataset copy, imputed values replace the unobserved (missing) values in the next step. Due to random variation, slightly different values are imputed in each copy. As a result, numerous imputed datasets exist. In the third phase, the imputed datasets are independently assessed, and the study results are then merged to provide the final study result.

Van Ginkel *et al.* [34] pointed out the misconception about the missing values assumptions in their research. They made it clear that multiple imputations are always preferable over other methods, such as listwise deletion, independent of the missingness mechanism. Multiple imputations are favored in MCAR because it provides higher statistical power; in MAR because, in addition to providing more power, it provides unbiased results, whereas other methods may not, and in MNAR, since missingness is based on data that is not observed, unbiased outcomes of imputation cannot be assured. Little [22] developed a standard case designed for lowering Rubin's requirements for frequentist maximum likelihood conjecture with accuracy based on observable data and argued that the missingness mechanism can be ignored even if MAR does not hold in some instances. Hence, all three assumptions of the missing value mechanism (MCAR, MAR and MNAR) were examined in our work using mean, mi, MICE, Hmsic, and missForest. These five techniques were chosen because they have been cited as the most appropriate for longitudinal data [6].

3.2.1 Mean imputation

Column-based means are used to replace missing values, i.e. the mean value for each variable. This is the simplest method, but it is not very precise. It disregards the correlation between dependent variables. Moreover, if the data are skewed, one may take into account mode or median replacement.

3.2.2 Multiple Imputed values using a Chained Equation (MICE)

Imputations are performed one by one for each variable in MICE iterations. Azur *et al.* [4] demonstrated the technique. To begin, substitute individual predictor variables for all missing values, using the mean imputation as a "placeholder". Second, to impute a variable, y, mean values are substituted for missing values; y is now a dependent variable and is regressed against the other variables acting as independent variables;

missing values are then replaced by the regression model's predictions. Third, the imputed values of variable y will be used as the independent variable to impute other missing variables, while the remaining variables will be used with mean substitution. The fourth, second, and third steps are repeated until each variable has predictions imputed. These are the steps in a single iteration or cycle. The number of iterations/cycles is increased, and the associated imputations are updated.

3.2.3 Multiple imputation diagnostics (mi)

Su *et al.* [33] stated that mi imputation is similar to MICE except for one significant difference: it imputes from the conditional distribution of a variable, whereas other variables are either imputed or observed. mi has an advantage over MICE in that it can handle data irregularities such as multicollinearity within a dataset. Four steps comprise the procedure for imputing a variable [33]. To begin, the setup analyzes missing data patterns to diagnose data structure issues, preprocesses the data, and identifies conditional models. Second, using a conditional model it iterates over MICE-based imputations and verifies the conditionality, acceptability, and convergence of imputed values. Thirdly, analysis collects and pools multiple imputed complete datasets to perform a complete case analysis. Fourthly, validation establishes sensitivity, performs cross-validation, and confirms compatibility.

3.2.4 Harrell miscellaneous (Hmisc)

According to Harrell and Dupont [18], the algorithm is capable of performing both simple imputations using the mean/mode/median and multiple imputations using additive regression, bootstrapping, and predictive mean matching approaches. It begins by identifying missing values in each variable using a randomized sample of non-missing values of size 'm'. Second, by fitting the flexible additive model to the transformation, it optimizes it. Additionally, an identity transformation can be forced. Third, it makes predictions for the observed values that are not missing using the flexible fitted model. Fourth, it uses the predicted transformed value that is closest to the missing value's predicted transformed value to replace the missing value with the observed value. Fifth, to impute other variables, it chooses the imputed values at random. The first set of 'x' iterations is a burn-in set for n iterations. Fifth, to impute other variables it randomly selects the variable's imputed values.

3.2.5 missForest

Stekhoven and Bühlmann [32] defined this as a non-parametric approach based on variables' pairwise independence. The algorithm is based on the random forest approach developed by Breiman [7]. MissForest constructs a random forest for each variable based on observed values, and forecasts missing values. The algorithm is repeated until the specified number of iterations is reached, the specified number of iterations is maximized, or the stopping criterion is met. Oshiro *et al.* [27] recommend that a forest contains between 64 and 128 trees.

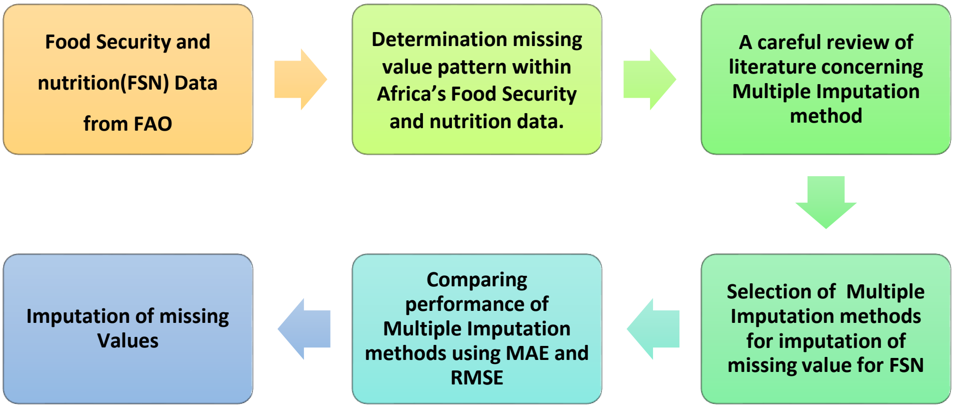
Food Security and nutrition(FSN) Data from FAO
Determination missing value pattern within Africa's Food Security and nutrition data.
A careful review of literature concerning Multiple Imputation method
Selection of Multiple Imputation methods for imputation of missing value for FSN
Comparing performance of Multiple Imputation methods using MAE and RMSE
Imputation of missing Values

Fig. 2. Flow chart of the methodology.

3.3 Experimental design

Our experiment was planned to determine the missing mechanism that is existing in FAO's Africa FSN data, as well as the optimal solution for dealing with the missingness found in the dataset. The experiment is based on a data set with 42 variables for each of Africa's 54 countries from FAO. Following a comprehensive review of the literature, we decided on five potential imputation techniques: mean imputation, multiple imputations via multiple imputations with diagnostics (mi), chained equations (MICE), Harrell miscellaneous (Hmsic), and missForest, a random forest-based iterative imputation. We applied Little's test for MCAR on the dataset to study the existing missing mechanisms [21]. The RMSE and MAE were calculated to see how well the five-imputation approach handled missing values. The greater the value of the error indicator (MAE or RMSE), the greater the error and less improvement of precision in the data. The result is a description of the most effective methods for dealing with missingness within the data set vital for food security and public health in Africa. The five imputation methods (Mean, MICE, mi, Hmisc, missForest) were used to impute missing values using 20 iterations, except for the mean imputation where the iteration remained constant for each iteration. In their study, van Ginkel *et al.* [34] brought up a misconception about the missing values assumptions. As a result, all three missing value mechanism assumptions, MCAR, MAR, and MNAR, were tested on all five imputation methods in our study. Figure 2 provides a summary of our research's methodology.

3.3.1 Measure of accuracy criteria for imputation methods

To find the optimal imputation approach, two model performance tests were used: root mean square error (RMSE) and mean absolute error (MAE), which are calculated in the following way:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\theta_i - \hat{\theta}_i\right)^2}, \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i}^{n} |\theta_i - \hat{\theta}_i|, \tag{5}$$

where $\hat{\theta}_i$ and θ_i respectively, are the ith observations for the comparison and reconstructed data sets. The difference between the estimated and observed values were used to calculate the error. For both RMSE and MAE tests the smaller the value obtained, the better the estimation method.

4 Results

Our study was designed to discover the mechanism of missingness that is present and the imputation algorithm that best fits the food and security in Africa. The study was conducted using 42 predictor variables from FAO. The aim of this work was to obtain the most effective method to deal with missing observations concerning food security and nutrition data from Africa. RMSE and MAE metrics for all the assumptions of missing values (MAR, MCAR, MNAR) were used to check the effectiveness of single and multiple imputation methods.

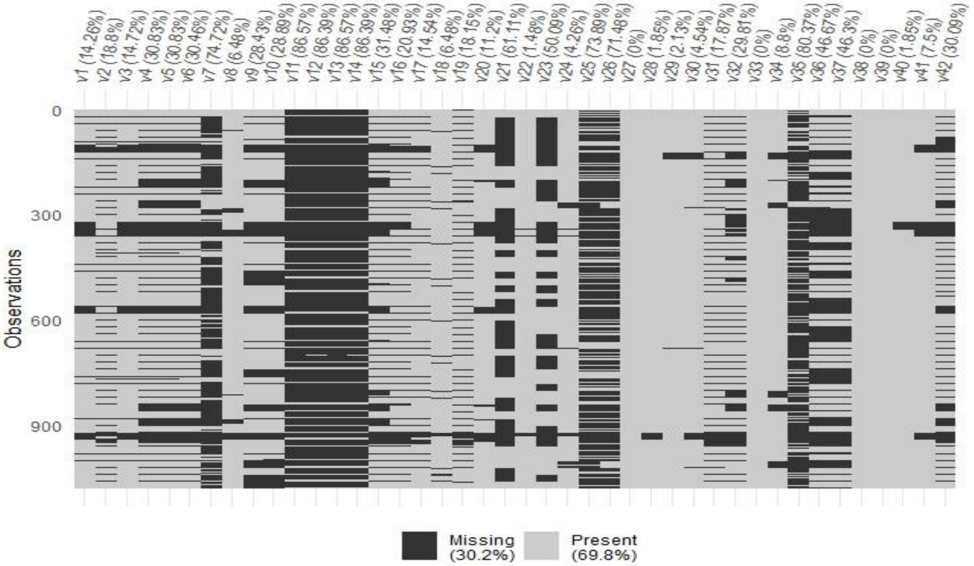

Fig. 3. Distribution of missing value pattern across variables.

During the missing value exploration stage, Fig. 3 displays the occurrence of missing values in each variable. The outcome of the missing situation present in the data revealed that there are undeniably missing data in the food and nutrition security dataset for Africa, accounting for about 30.2 percent of the values ($n = 1082$) for all 42 FAO measured variables. Except for three variables (v39, v38, v27), each variable has a significant number of missing values, as seen in Fig. 3.

We next visualized the locations (pattern) of missing values among all variables in the dataset to acquire a better idea of the distribution of missing values in the data

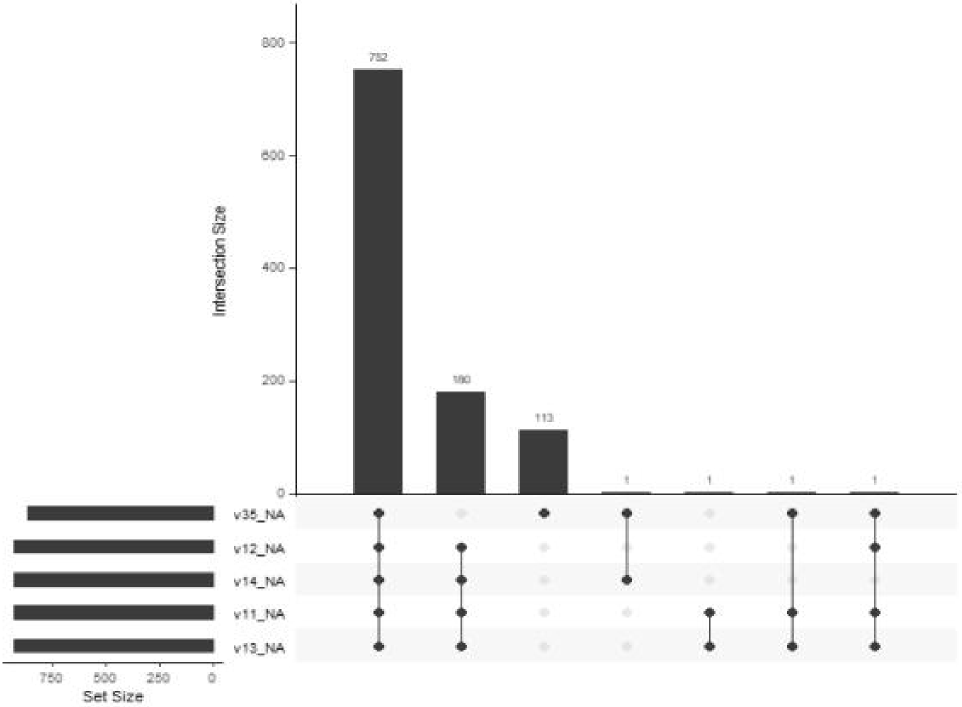

Fig. 4. Distribution of variables missing together.

(Fig. 3). The distribution indicates that the missing value patterns for food security and nutrition from 2000 to 2019 concerning Africa follow a non-monotonic pattern. In the plot, grey indicates observed values and black indicated missing values.

We have now seen where the missing values are clustered, and it appears to verify our prior findings of the occurrence of missing values for each variable. Is there a link between missing values in one variable and missing values in other variables? By answering this question, we will be able to figure out what mechanism is at work in our data. We then looked at which variables were missing when they were combined (Fig. 4). The findings are consistent with the observation that there are a significant number of cases where specific variables have missing values (e.g. v13, v11, v34). This indicates that data is not missing at random completely, paving the way for the MAR and MNAR, and hence imputed missing values are recommended. We used Little's test for MCAR and obtained a p value (0.000) that is statistically significant at alpha 0.05. Here again, the MCAR assumption is violated which is in support of Fig. 4. This indicates that the missing mechanism in Africa's food and nutrition security data is MNAR or MAR, rather than MCAR, and hence the missing values are non-ignorable. This is in support of van Ginkel *et al.* [34] that the missing values can be imputed from the observed information about the variables if it is not MCAR.

To visualize the post-imputation diagnostics for each imputation method concerning the nonignorable missing values identified in our dataset, a scatter plot for the correlation coefficient of the imputed and original data was employed as shown in Fig. 5. For all the five methods of imputation there is a positive relationship between

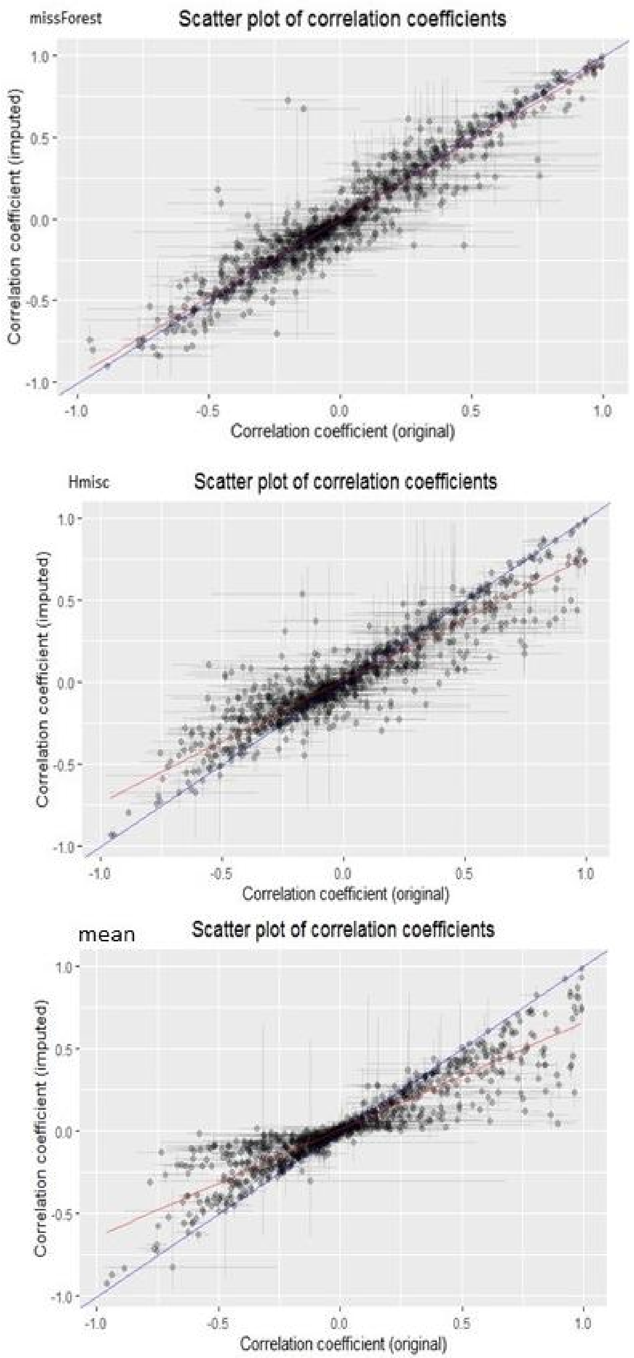

Fig. 5. Bootstrapped correlation coefficients from the original data and the imputed data.

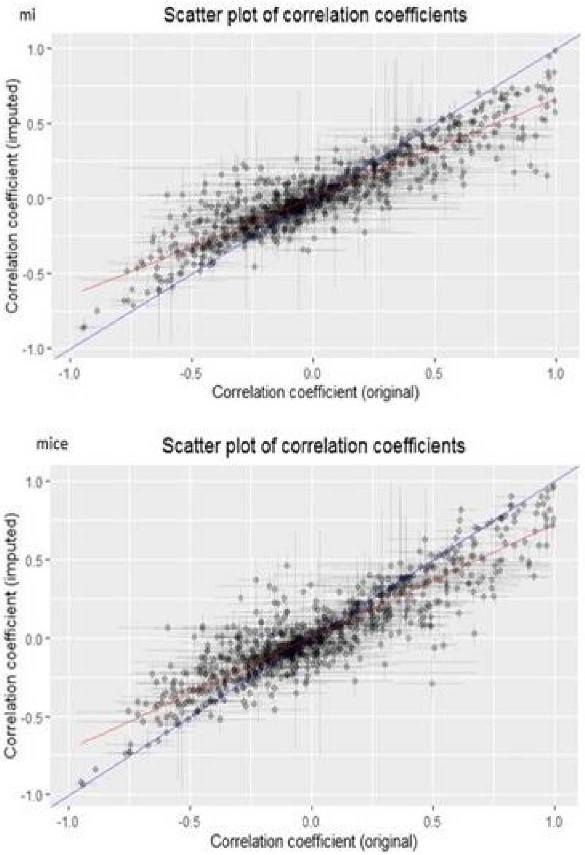

Fig. 5. (*continued*).

the initial dataset and its imputed dataset. This potentially emphasizes the role of imputation [19] and the significance these five imputation methods play in handling nonignorable (MAR and MNAR) missing values [1].

In the single imputation approach, the mean substitution method was the worst method that did not give account for imputation uncertainty after the imputation was complete. Whereas the missForest gave us a good account of imputation (the red line almost aligns with the blue line), even outperforming its parametric competitors concerning MI methods, this is evident in the bootstrapped correlation coefficients diagram (Fig. 5). This is similar to what Erhan *et al.* [13] reported concerning multiple imputations.

Our results revealed that missForest was the best-performing imputation technique. It even outperformed the MICE method which is perceived to be more advantageous under the MAR assumption. For the nonignorable (MAR and MNAR) missing mechanism identified in our study, missForest consistently produced the lowest imputation errors concerning the RMSE and MAE values (Table 1), indicating that missForest imputation gave the closest imputed values of food security and nutrition. This demonstrates the potency of missForest which has previously been linked to handling nonignorable missingness in longitudinal measures including emotional

Table 1. Measures of imputation error (RMSE and MAE) for food security and nutrition in Africa.

Imputation method		MAR		MNAR	
	Computation time	RMSE	MAE	RMSE	MAE
Mean imputation	0	1	0.79	1	0.79
MICE imputation	0.97	0.94	0.62	0.93	0.61
mi imputation	3.33	1.04	0.7	0.99	0.69
missForest imputation	5.2	0.8	0.61	0.8	0.61
Hmsic imputation	0.5	1.01	0.7	0.99	0.66

and psychophysical outcomes and air quality monitoring [3, 16]. Nevertheless, the time used by this method was the highest among the five methods which is likely to be its drawback. Mean, Hmisc, and mi performed similarly, suggesting that imputations built on either of these methodologies would be close to each other, which is utterly the same as what Erhan *et al.* [13] reported in their study. Again, concerning non-ignorable missingness, the so-called mi methods consistently produced the highest imputation errors regarding the RMSE and MAE values (Table 1), indicating that mi imputation gave the least close imputed values for Africa's security and nutrition (FSN). This was also found elsewhere see [28]. In what follows, we focus on the missForest imputation method for the nonignorable (MAR and MNAR) missing systems present in our data.

With five different imputed methods which included the mean-based imputation, missing data were imputed using twenty iterations for all methods. The method that performed best for the nonignorable missing data mechanisms (MAR and MNAR) in terms of RMSE and MAE is identified in this paper. From our study concerning African food security and nutrition data, missForest was the most effective and accurate method for imputing missing values. Among the five methods, missForest proved to be the one with the least RMSE and MAE. This corroborates Alkabbani's *et al.* [2] finding. We found that missForest can cope with mixed-type data and has a reputation for excelling in difficult situations such as complicated interactions and non-linear data structures. Even with high dimensional data where the number of variables is likely to outnumber the number of observations by a significant margin, missForest imputation results are still good and this supports earlier research [3, 12, 15]. MissForest is resistant to noisy data and multicollinearity. MissForest performed well for our data (food and nutrition security) for both assumptions of missingness. This may have been its significant abilities, giving a case for itself (irrespective of the missingness mechanism in the data) why it had the least values for RMSE and MAE in this work. MissfForest imputations solve the nonignorable missing value difficulties of unreliable statistical modeling and biased inferences, as evidenced by their lowest RMSE and MAE making it effective and flexible. This agrees with previous work [8].

5 Conclusion

A major facet in looking into previous, current, and upcoming scenarios has become very reliant on physical data, with food and nutrition security not being an exception. Missingness within a dataset is a challenge to researchers. Nonetheless, a suitable scheme of imputation can help provide the best remedy for the situation at hand.

Making the most appropriate imputation technique for food security data available is critical for determining the interdependence of variables in food and nutrition security data, as well as estimating the potential effect, which is a critical step. Our findings imply that when missForest imputations are employed for nonignorable missing values, the best accurate estimates can be generated from nonignorable missing values, even when the missing data processes change over time. We can now hypothesize that missForest does not assume randomness (practically at all forms of data) concerning missing values, because we have demonstrated that missForest is nearly impervious to randomness (i.e. MAR and MNAR) even though the identified missingness mechanism in our data was MAR. Hence missForest remains the best panacea for high standard errors and biases associated with nonignorable (MAR) missing value parametrization to fill the gap to improve statistical accuracy, especially food and nutrition security. MissForest will therefore provide informed evidence for researchers and policymakers in modeling food security and nutrition for Africa.

Conflict of Interest Statement

All authors certify that we have no financial or non-financial interests in the subject matter or materials discussed in this manuscript and we have no affiliations with or involvement in any organization or entity that has a financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Appendix A

Table 2. Food security and nutrition predictors (42) variables by FAO.

Variable name	Definition of variables
v1	Average dietary energy supply adequacy (percent) (3-year average)
v2	Average value of food production (constant 2004-2006 I$/cap) (3-year average)
v3	Dietary energy supply used in the estimation of prevalence of undernourishment (kcal/cap/day) (3-year average)
v4	Share of dietary energy supply derived from cereals, roots and tubers (kcal/cap/day) (3-year average)
v5	Average protein supply (g/cap/day) (3-year average)
v6	Average supply of protein of animal origin (g/cap/day) (3-year average)
v7	Rail lines density (total route in km per 100 square km of land area)
v8	Gross domestic product per capita, PPP, dissemination (constant 2011 international $)
v9	Prevalence of undernourishment (percent) (3-year average)
v10	Number of people undernourished (million) (3-year average)
v11	Prevalence of severe food insecurity in the total population (percent) (3-year average)
v12	Prevalence of moderate or severe food insecurity in the total population (percent) (3-year average)
v13	Number of severely food insecure people (million) (3-year average)
v14	Number of moderately or severely food insecure people (million) (3-year average)
v15	Cereal import dependency ratio (percent) (3-year average)
v16	Percent of arable land equipped for irrigation (percent) (3-year average)
v17	Value of food imports in total merchandise exports (percent) (3-year average)

continued on next page

Table 2. (*continued*)

Variable name	Definition of variables
v18	Political stability and absence of violence/terrorism (index)
v19	Per capita food production variability (constant 2004–2006 thousand int$ per capita)
v20	Per capita food supply variability (kcal/cap/day)
v21	Percentage of population using safely managed drinking water services (percent)
v22	Percentage of population using at least basic drinking water services (percent)
v23	Percentage of population using safely managed sanitation services (percent)
v24	Percentage of population using at least basic sanitation services (percent)
v25	Percentage of children under 5 years affected by wasting (percent)
v26	Number of children under 5 years affected by wasting (million)
v27	Percentage of children under 5 years of age who are stunted (modelled estimates) (percent)
v28	Number of children under 5 years of age who are stunted (modeled estimates) (million)
v29	Percentage of children under 5 years of age who are overweight (modelled estimates) (percent)
v30	Number of children under 5 years of age who are overweight (modeled estimates) (million)
v31	Prevalence of obesity in the adult population (18 years and older)
v32	Number of obese adults (18 years and older) (million)
v33	Prevalence of anemia among women of reproductive age (15–49 years)
v34	Number of women of reproductive age (15–49 years) affected by anemia (million)
v35	Prevalence of exclusive breastfeeding among infants 0–5 months of age
v36	Prevalence of low birthweight (percent)
v37	Number of newborns with low birthweight (million)
v38	Minimum dietary energy requirement (kcal/cap/day)
v39	Average dietary energy requirement (kcal/cap/day)
v40	Coefficient of variation of habitual caloric consumption distribution (real number)
v41	Incidence of caloric losses at retail distribution level (percent)
v42	Average fat supply (g/cap/day) (3-year average)

References

[1] O. Akande, J.P. Reiter. Multiple imputations for nonignorable item nonresponse in complex surveys using auxiliary margins. In *Statistics in the Public Interest*, pp. 289–306. Springer, Cham, 2022. https://doi.org/10.1007/978-3-030-75460-0_16.

[2] H. Alkabbani, A. Ramadan, Q. Zhu, A. Elkamel. An improved air quality index machine learning-based forecasting with multivariate data imputation approach. *Atmosphere*, **13**(7):1144, 2022.

[3] A.R. Alsaber, J. Pan, A. Al-Hurban. Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018). *Int. J. Environ. Res. Public Health*, **18**(3):1333, 2021.

[4] M.J. Azur, E.A. Stuart, C. Frangakis, P.J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *Int. J. Methods Psychiatr. Res.*, **20**(1):40–49, 2011. https://doi.org/10.1002/mpr.329.

[5] L.K. Bache-Mathiesen, T.E. Andersen, B. Clarsen, M.W. Fagerland. Handling and reporting missing data in training load and injury risk research. *Sci. Med. Footb.*, **6**(4):452–464, 2022.

[6] N. Boukichou-Abdelkader, M.Á. Montero-Alonso, A. Muñoz García. Different routes or methods of application for dimensionality reduction in multicenter studies databases. *Mathematics*, **10**(5):696, 2022.

[7] L. Breiman. Random forests. *Mach. Learn.*, **45**:5–32, 2001. https://doi.org/10.1023/A:1010933404324.

[8] E. Casiraghi, R. Wong, M. Hall, B. Coleman, M. Notaro, M.D. Evans, J.S. Tronieri, H. Blau, B. Laraway, T.J. Callahan. A methodological Framework for the Comparative Evaluation of Multiple Imputation Methods: Multiple Imputation of Race, Ethnicity and Body Mass Index in the US National COVID Cohort Collaborative, 2022. Preprint arXiv:2206.06444.

[9] H. Chakraborty, H. Gu. *A Mixed Model Approach for Intent-to-Treat Analysis in Longitudinal Clinical Trials with Missing Values.* RTI Press, 2019.

[10] B. Efron. Missing data, imputation, and the bootstrap. *J. Amer. Statist. Assoc.*, **89**(426):463–475, 1994. https://doi.org/10.1080/01621459.1994.10476768.

[11] H. El Bilali. Research on agro-food sustainability transitions: where are food security and nutrition? *Food Secur.*, **11**(3):559–577, 2019.

[12] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona. A survey on missing data in machine learning. *J. Big Data*, **8**(1):1–37, 2021.

[13] L. Erhan, M. Di Mauro, A. Anjum, O. Bagdasar, W. Song, A. Liotta. Embedded data imputation for environmental intelligent sensing: a case study. *Sensors*, **21**(23):7774, 2021.

[14] FAO. *The State of Food Security and Nutrition in the World 2018. Building Climate Resilience for Food Security and Nutrition.* Food and Agriculture Organization of the United Nations, Rome, 2018.

[15] S. Feng, C. Hategeka, K.A. Grépin. Addressing missing values in routine health information system data: an evaluation of imputation methods using data from the Democratic Republic of the Congo during the COVID-19 pandemic. *Popul. Health Metr.*, **19**(1):1–14, 2021.

[16] C. Fernández-de Las-Peñas, M. Palacios-Ceña, J.A. Valera-Calero, M.L. Cuadrado, A. Guerrero-Peral, J.A. Pareja, L. Arendt-Nielsen, U. Varol. Understanding the interaction between clinical, emotional and psychophysical outcomes underlying tension-type headache: a network analysis approach. *J. Neurol.*, **269**(8):4525–4534, 2022.

[17] S.J. Hadeed, M.K. O'Rourke, J.L. Burgess, R.B. Harris, R.A. Canales. Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Sci. Total Environ.*, **730**:139140, 2020.

[18] F.E. Harrell, C. Dupont. Package 'Hmisc': Harrell Miscellaneous, 2016. `https://hbiostat.org/R/Hmisc/`. R topics Documented.

[19] T.F. Johnson, N.J. Isaac, A. Paviolo, M. González-Suárez. Handling missing values in trait data. *Glob. Ecol. Biogeogr.*, **30**(1):51–62, 2021.

[20] W.-C. Lin, C.-F. Tsai. Missing value imputation: a review and analysis of the literature (2006–2017). *Artif. Intell. Rev.*, **53**(2):1487–1509, 2020.

[21] R.J. Little. A test of missing completely at random for multivariate data with missing values. *J. Amer. Statist. Assoc.*, **83**(404):1198–1202, 1988.

[22] R.J. Little. Missing data assumptions. *Annu. Rev. Stat. Appl.*, **8**:89–107, 2021.

[23] R.J. Little, D.B. Rubin. *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons, 2019.

[24] Y. Liu, T. Dillon, W. Yu, W. Rahayu, F. Mostafa. Missing value imputation for industrial IoT sensor data with large gaps. *IEEE Internet Things J.*, **7**(8):6855–6867, 2020.

[25] C. Mack, Z. Su, D. Westreich. *Managing Missing Data in Patient Registries.* Agency for Healthcare Research and Quality (US), 2018.

[26] T.P. Morris, I.R. White, M.J. Crowther. Using simulation studies to evaluate statistical methods. *Stat. Med.*, **38**(11):2074–2102, 2019.

[27] T.M. Oshiro, P.S. Perez, J.A. Baranauskas. How many trees in a random forest? In *Machine Learning and Data Mining in Pattern Recognition*, volume 7376 of *Lect. Notes Comput. Sci.*, 2012.

[28] T.H. Ruggles, D.J. Farnham, D. Tong, K. Caldeira. Developing reliable hourly electricity demand data through screening and imputation. *Sci. Data*, **7**:155, 2020.

[29] V. Saravanakumar, U. Malaiarasan, R. Balasubramanian. Sustainable agriculture, poverty, food security and improved nutrition. In *Sustainable Development Goals*, pp. 13–39. Springer, 2020.

[30] R.M. Schouten, G. Vink. The dance of the mechanisms: How observed information influence the validity of missingness assumptions. *Sociol. Methods Res.*, **50**(3):1243–1258, 2021.

[31] A. Sportisse, C. Biernacki, C. Boyer, J. Josse, M.M. Lourdelle, G. Celeux, F. Laporte. Model-Based Clustering with Missing not at Random Data, 2021. Preprint arXiv:2112.10425.

[32] D.J. Stekhoven, P. Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**(1):112–118, 2012. https://doi.org/10.1093/bioinformatics/btr597.

[33] Y.S. Su, A. Gelman, J. Hill, M. Yajima. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *J. Stat. Softw.*, **45**(2):1–31, 2011. https://doi.org/10.18637/jss.v045.i02.

[34] J.R. van Ginkel, M. Linting, R.C. Rippe, A. van der Voort. Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *J. Pers. Assess.*, **102**(3):297–308, 2020.

REZIUMĖ

Maisto ir mitybos saugumo duomenų spragų užpildymas: koks metodas yra geriausias Afrikos maisto ir mitybos saugumui?

A. Bofa, T. Zewotir

Šiame tyrime pristatomi metodai, naudojami siekiant parinkti tikslesnes procedūras priskiriant reikšmes trūkstamiems Afrikos maisto saugumo ir mitybos (MSM) duomenims. Pagrindinis dėmesys skiriamas šiems penkiems trūkstamųjų reikšmių duomenyse priskyrimo metodams: vidurkio priskyrimas; kartotinis priskyrimas naudojant grandininę lygtį (Multiple Imputation using a Chained Equation, MICE); priskyrimas su diagnostika, pagrįstas kintamojo sąlyginiu skirstiniu (mi); adityvioji regresija ir vidurkio prognozės atitikimas (Hmisc); bei priskyrimas naudojant atsitiktinius miškus (MissForest). Aprašant kiekvieną metodą aptariama, kaip jis veikė atsitiktinai trūkstamųjų (Missing At Random, MAR) ir neatsitiktinai trūkstamųjų (Missing Not At Random, MNAR) reikšmių atvejais naudojant vidutinės kvadratinės paklaidos kvadratinę šaknį ir vidutinę absoliučiąją paklaidą kaip tikslumo matą. Ištyrus, ar nėra įvertinių tikslumo ir nepaslinktumo prasme MSM analizėje neignoruotinų trūkstamųjų reikšmių, buvo nustatatyta, kad MissForest metodas efektyviau ir su mažesniu poslinkiu priskyrė trūkstamąsias reikšmes tuo pagerindamas duomenų kokybę. Vadinasi, MissForest metodas yra tinkamiausia alternatyva Afrikos MSM duomenų trūkstamosioms reikšmėms priskirti. Šis tyrimas papildo dabartines žinias apie maisto ir mitybos nesaugumą ir suteikia naudingos informacijos politikos formuotojams, ypač apie galimas trūkstamųjų reikšmių priskyrimo MSM duomenyse ekonomines ir socialines pasekmes Afrikoje.

Raktiniai žodžiai: maisto saugumas ir mityba; missForest; trūkstamoji reikšmė; kartotinis priskyrimas; atsitiktinai trūkstamosios reikšmės; neatsitiktinai trūkstamosios reikšmės; Afrika