doi:10.20944/preprints202105.0390.v1

Robustness of Imputation Methods with Backpropagation Algorithm in Nonlinear Multiple Regression

Castro Gbêmêmali Hounmenou[a,d], Boris Milognon Behingan[b], Christophe Chrysostome[b], Kossi Essona Gneyou[c] and Romain Glèlè Kakaï[d]

[a]Institut de Mathématiques et de Sciences Physiques, Dangbo, University of Abomey-Calavi, (Benin), castrohounmenou@gmail.com (H.G.C.);
[b]Laboratoire de Recherche Avicole et de Zoo-Economie, University of Abomey-Calavi, Abomey-Calavi, Bénin, begboris@yahoo.fr (B.M.B), cchrysostome@gmail.com (C.C.);
[c]Laboratoire de Modélisations Mathématiques et Applications, University of Lome, (Togo), kgneyou@gmail.com (G.E.K.);
[d]Laboratoire de Biomathématiques et d'Estimations Forestières, University of Abomey-Calavi, Abomey-Calavi, (Benin), castrohounmenou@gmail.com (H.G.C.), glele.romain@gmail.com (G.K.R.)

ABSTRACT

Missing observations constitute one of the most important issues in data analysis in applied research studies. The magnitude and their structure impact parameters estimation in the modeling with important consequences for decision-making. This study aims to evaluate the efficiency of imputation methods combined with the backpropagation algorithm in a nonlinear regression context. The evaluation is conducted through a simulation study including sample sizes (50, 100, 200, 300 and 400) with different missing data rates (10, 20, 30 40 and 50%) and three missingness mechanisms (MCAR, MAR and MNAR). Four imputation methods (Last Observation Carried Forward, Random Forest, Amelia and MICE) were used to impute datasets before making prediction with backpropagation. 3-MLP model was used by varying the activation functions (Logistic-Linear, Logistic-Exponential, TanH-Linear and TanH-Exponentiel), the number of nodes in the hidden layer (3 - 15) and the learning rate (20 - 70%). Analysis of the performance criteria (R^2, r and $RMSE$) of the network revealed good performances when it is trained with TanH-Linear functions, 11 nodes in the hidden layer and a learning rate of 50%. MICE and Random Forest were the most appropriate for data imputation. These methods can support up to 50% of missing rate with an optimal sample size of 200.

KEYWORDS

Multilayer perceptron neural network, regression model, backpropagation, missing data, imputation method.

1. Introduction

Let Y be a real random variable revealed mean depends on $\mathbf{x} = (x_1, \cdots, x_p) \in \mathbb{R}^p$, replications of the random vector $\mathbf{X}$, and the dependence may be nonlinear $\mathbb{E}(Y|x_1, \cdots, x_p) = \zeta(x_1, \cdots, x_p)$. This relation is equivalent to : $Y = \zeta(x_1, \cdots, x_p) + \epsilon$

CONTACT Romain Glèlè Kakaï. Email: glele.romain@gmail.com

with $\mathbb{E}(\epsilon) = 0$. Let a parametric nonlinear regression model be represented by : $Y = \zeta(x_1, \cdots, x_p; \theta) + \epsilon$ where ζ is nonlinear with respect to θ, the set of model parameters. This means that, for at least one θ_i, the derivative of ζ with respect to θ_i depends on at least one of the parameters. For example $\zeta(\mathbf{x}; \theta) = \frac{\theta_1 x_1}{1+\theta_2 x_2}$ is used by chemists. Differentiating ζ with respect to θ_1 and θ_2 gives: $\frac{\partial\zeta}{\partial\theta_1} = \frac{x_1}{1+\theta_2 x_2}$ and $\frac{\partial\zeta}{\partial\theta_2} = \frac{x_2 x_1 \theta_1}{(1+\theta_2 x_2)^2}$. One of the nonlinear models that has received great attention last few years is the model based on artificial neural networks (ANNs). They are used in the fields of prediction and classification, fields in which regression models and other related statistical techniques have traditionally been used [1–4]. Multilayer perceptron neural networks (MLPs) are one of the architectures of ANNs acting as a type of regression model, not necessarily parametric, which enables complex functional forms to be modeled [5,6].

In breeding, the knowing of production is necessary for specialists who need simple and accurate techniques to predict the production of meat, eggs, milk etc. Production is influenced by interdependent factors and MLPs offer more flexibility in describing their relationships. But data collected in the case of production are often small (due to the cost of experimentation) and seldom complete. Missing data are one of the most common problems for researchers in breeding [7]. It occurs because of human error, equipment failure, death of animal during the experiment, data collected with difficulty, official statistics not available, etc. [8–12]. Analysis of incomplete datasets results in problems such as biased parameter estimates, inflation of standard errors, loss of information, and weak generalizability of results [8,13–15]. Apart from kohonen network [16], most of statistical analysis methods assume the absence of missing data, and are only able to include observations for which every variables are measured [17]. To overcome this situation, rows with missing values can be deleted (deletion) but it leads to a loss of precision [18,19] with weak sample size. To avoid this situation, imputation methods can be used. Different imputation methods exist based on different approaches: single imputation, multiple imputation, etc. [20,21]. With imputation techniques, researchers can obtain complete data for their prediction.

Despite the success of MLPs in breeding and other disciplines, it exists some factors which can affect its performances like: activation functions, learning rate, number of hidden layers, number of neurons in each hidden layer, etc. [22]. However there are no clear guidelines on which activation function performs better [23] and also about the value of the learning rate [22,23]. Yet, a drawback of this type of network is that it requires a full set of input data. Therefore our study aims to evaluate the empirical robustness of imputation methods in non linear regression with backpropagation algorithm.

The main objective of this study is to analyze the behavior of the imputation methods combined to the backpropagation algorithm for the management of missing data. Specifically we: (i) analyse the effect of imputation methods on the structure of hyperparameters, (ii) determine the best imputation method according to sample size and the missing data rate with the best structure of hyperparameters for the multilayer perceptron neural network.

2. Framework, specification of model and generation of a data population

- **Types of missing data and their management**

Let $X = [x_{ij}]$ be a data matrix of dimension (n, p) of elements $x_{ij} \in \mathbb{R}$, where n and p elements of $\mathbb{N}^*$ are respectively the number of observations, the number of variables and x_{ij} is the value of the variable $j \in [\![1, p]\!]$ for the observation $i \in [\![1, n]\!]$. Let $Z = [z_{ij}]$, an indicator matrix of missing data elements z_{ij}, such that $z_{ij} = 1$ if x_{ij} is missing, and $z_{ij} = 0$ otherwise, then we have: $X = \{X_{obs}, X_{mis}\}$. The matrix Z describes the structure of the missing data and is useful to treat it as a stochastic matrix. The statistical model for missing data is $P(Z|X, \kappa)$, where κ is the parameter of the missing data process and $P(\cdot)$, denotes the conditional distribution of Z given X, of parameters κ. The mechanism of missingness is determined by the dependency of Z on the variables in the data set. According to [12], three categories of missing data can be distinguished: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR).

Definition 2.1. *Missing data are "missing completely at random" (MCAR).*
Missing data are said to be missing completely at random (MCAR) when the fact of not having a value is totally independent of the variables X and we have:

$$\forall\ X, P(Z|X, \kappa) = P(Z|\kappa). \tag{1}$$

When the missing data are not MCAR, we need to know if differences in the characteristics of non-respondents and respondents can be explained by variables common to respondents and non-respondents. We note X_{obs}, the observed part of the data X and X_{mis}, the missing part.

Definition 2.2. *Missing data are "missing at random" (MAR).*
The data are said to be missing at random (MAR) when the distribution of Z given X, depends only on the variables recorded in the database X_{obs}, and we have:

$$\forall\ X_{mis}, P(Z|X_{obs}, X_{mis}, \kappa) = P(Z|X_{obs}, \kappa). \tag{2}$$

Definition 2.3. *Missing data are "Missing Non At Random" (MNAR)*
The data are said to be missing non at random (MNAR) when the distribution of Z given X also depends on X_{mis}, and we have:

$$\forall\ X_{obs}\ and\ X_{mis}, P(Z|X_{obs}, X_{mis}, \kappa) = P(Z|X_{obs}, X_{mis}, \kappa). \tag{3}$$

There are two basic methods for managing data matrices with missing values: (i) the *deletion method* and (ii) the *imputation method* [24]. The first one considers only the individuals for which all the data are available, i.e. to delete any individual having at least one missing value. The second consists to replace the missing values in the data set by estimated values. Two imputation approaches are used: *simple imputation* and *multiple imputation* [25]. Single imputation is to fill in each missing value with a value. The second approach covers methods whose procedures are based on models. This is done by replacing the missing values with several simulated values to properly reflect the uncertainty that is attached to the missing data [26].

- **Factors affecting the predictive performance of a multilayer perceptron neural network and back-propagation algorithm**

A multilayer perceptron neural network (MLP) is a feedforward neural network, consisting of a number of units (called neurons) connected by weight links. The units are organized in several layers, the first one is an input layer, the last one is an output layer and the intermediate one can have one or several hidden layers. The input layer receives an external activation vector, and transmits it via weighted connections to the units of the first hidden layer. These compute their activations and transmit them to the neurons in succeeding layers, see figure 1.

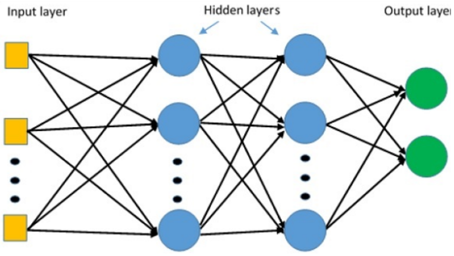

Figure 1. Example of a multilayer perceptron network with two hidden layers

Although multilayer perceptron neural networks have shown good predictive performance compared to classical methods, they are often affected by factors such as: the number of neurons and layers, the choice of transfer functions and the sample size. For more detail, see[27]. The estimation of the network weights is done by minimizing a quadratic cost function. It can be done, among other things, by the backpropagation algorithm (BP), whose procedure is summarized as follows:

1. Initialize all weights to small random values in the interval [-0.9, 0.9];
2. Normalize the training data;
3. Randomly permute the training data;
4. For each training data k:
   (a) Compute the observed outputs by forward propagating the inputs;
   (b) Adjust the weights by backpropagating the observed error from the output layer towards the input layer:

$$\begin{aligned} w_{ij}(k+1) &= w_{ij}(k) + \Delta w_{ij}(k+1) \\ &= w_{ij}(k) + \eta \delta_j(k+1) y_i(k+1) \end{aligned} \tag{4}$$

   with $w_{ij}(k+1)$, the adjusted weight for the neuron j; $w_{ij}(k)$, the previously computed weight for the neuron j; $0 \leq \eta \leq 1$ representing the learning rate; $\delta_j(k+1)$ is the local gradient computed for the neuron j and $y_i(k+1)$ representing either the output of neuron i on the previous layer, if it exists, or the input i otherwise.
5. Repeat steps 3 and 4 up to a maximum number of iterations or until the root mean square error is less than a certain threshold.

- **Specification of model and generation of a data population**

The nonlinear regression model considered is a multilayer perceptron neural network with a hidden layer and its expression is :

$$Y = \zeta_\theta(\mathbf{x}) + \epsilon \tag{5}$$

with $\mathbb{E}(\epsilon) = 0$; $\mathbf{x} \in \mathbb{R}^p$ is a vector of p inputs; $Y \in \mathbb{R}$, $\zeta_\theta(\mathbf{x}) \in \mathbb{R}$ are respectively the observed output and the predicted output.

$$\zeta_\theta(\mathbf{x}) = f_2\left(\mathbf{w}^{(2)}\mathbf{f_1}\left(\mathbf{w}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}\right) + b^{(2)}\right) \tag{6}$$

where $\theta = \left(\mathbf{w}^{(1)}, \mathbf{b}^{(1)}; \mathbf{w}^{(2)}, b^{(2)}\right)$ or $\theta = \left(w_{11}^{(1)}, \cdots, w_{1p}^{(1)}, \cdots, w_{m1}^{(1)} \cdots, w_{mp}^{(1)}; b_{10}^{(1)} \cdots, b_{m0}^{(1)}; w_{11}^{(2)}, \cdots, w_{1m}^{(2)}; b^{(2)}\right)$ are the model's parameters and the total number of parameters is :

$$n_\theta = m(p+2) + 1. \tag{7}$$

with m the number of neurons in the hidden layer, f_2 is a transfer function applied to only neuron of the output layer, $\mathbf{f_1}$ is a vector composed of the same transfer function applied to each neuron in the hidden layer.

In order to have data with multicollinearity, a non-linear relationship between variables and for predictive purposes, we used the results of Insect as Feed for West Africa project [28] which evaluated the effect of maggot meal on the growth and economic performance of guinea fowl. The dependent variable is $y = food\ economic\ efficiency$ and independent variables are $\mathbf{x} = (x_1 = dose\ with\ three\ modality\ (0,\ 50\ and\ 100), x_2 = age, x_3 = food\ consumption, x_4 = weight)$. The predictive model obtained is :

$$\mathbb{E}(y_t) = \sum_{i=1}^{11} \mathbf{w}_i^{(2)} \frac{\exp\left(\langle \mathbf{x}_t, \mathbf{w}_i^{(1)}\rangle + b_i^{(1)}\right) - \exp\left[-\left(\langle \mathbf{x}_t, \mathbf{w}_i^{(1)}\rangle + b_i^{(1)}\right)\right]}{\exp\left(\langle \mathbf{x}_t, \mathbf{w}_i^{(1)}\rangle + b_i^{(1)}\right) + \exp\left[-\left(\langle \mathbf{x}_t, \mathbf{w}_i^{(1)}\rangle + b_i^{(1)}\right)\right]} + b^{(2)} \tag{8}$$

where y_t represents the t^{th} observation ($t \in [\![1, n]\!], n \in \mathbb{N}^*$); $\mathbf{w}_i$ is the weight vector associated with the i^{th} neuron in the hidden layer ($i \in [\![1, 11]\!]$); b_i and b are respectively the bias of the i^{th} neuron in the hidden layer and the bias applied to output neurone of 3-MLP model. The optimal parameters are $\theta = (\mathbf{w}^{(1)}, \mathbf{b}^{(1)}, \mathbf{w}^{(2)}, b^{(2)})$ with

$$\mathbf{w}^{(1)} = \begin{bmatrix} -0.06 & -0.87 & 0.33 & -0.10 & -0.15 & 0.08 & -0.13 & 0.60 & -0.07 & 0.04 & -0.17 \\ -0.06 & 0.41 & -0.38 & 0.29 & -0.18 & -0.31 & 0.22 & -0.44 & -0.12 & 0.16 & 0.28 \\ -0.13 & 0.47 & -0.16 & -0.27 & 0.22 & 0.20 & -0.36 & 1.42 & -0.13 & 0.27 & -0.48 \\ 0.22 & 0.35 & 0.21 & 0.03 & 0.15 & 0.01 & 0.27 & 0.55 & 0.32 & -0.43 & 0.37 \end{bmatrix}$$

$$\mathbf{b}^{(1)} = \begin{bmatrix} 0.31 & 0.65 & 0.45 & -0.82 - 0.39 & -0.38 & -0.86 & -0.01 & -0.79 & 0.79 & -0.43 \end{bmatrix};$$

$$\mathbf{w}^{(2)} = \begin{bmatrix} 0.01 & -0.38 & 0.15 & -0.03 & -0.05 & 0.04 & -0.04 & 0.61 & -0.01 & -0.02 & -0.05 \end{bmatrix};$$

$b^{(2)} = -0.82.$

The total number of parameters, $n_\theta = 67$.

A population of size $N = 10000$ was obtained from equation (8) to which we added the error ϵ of the equation (5) to compute Y. The error was generated according to $\mathcal{N}(\mu = 0, \sigma^2 = 1)$. The input variables X_1 to X_4 related to Y were defined using their respective distributions, X_1 by resampling techniques, $X_2 \sim \mathcal{N}(\mu = 4.5, \sigma^2 = 2.30)$, $X_3 \sim \mathcal{N}(\mu = 29.95, \sigma^2 = 13.04)$ and $X_4 \sim \mathcal{N}(\mu = 239.76, \sigma^2 = 117.11)$.

3. Simulations study

Seven factors were considered in this study. The sample size (5 different sizes), the missingness mechanism (3 different mechanisms), the missing data rate (5 different rates), the imputation methods (4 different methods) and the factors affecting the predictive and explanatory performance of the MLP model: the activation function (4 different functions), the number of hidden neurons (13 different size) and finally the learning rate (6 different rates). For each sample size, we have a combination of $936,000$ items, which is replicated 100 times.

- **Sampling size, simulating missingness and missing data imputation**

Five samples of different sizes n_i (n_i = 50, 100, 300 and 400) were extracted from the population using the bootstrap technique [29]. Three missingness mechanisms were considered, MAR, MCAR and MNAR (see the Section 2) with five missing data rates (MR) (10, 20, 30, 40 and 50%) to generate incomplete datasets. Missingness simulation is conducted on each of the five complete data using *MICE* package [30] from software R 3.3.6 [31]. Each of these previously obtained missing data are imputed with Last Observation Carried Forward (LOCF), Random Forest (RF), Amelia (AMELIA) and Multivariate Imputation by Chained Equation (MICE) methods in R using respectively *zoo* [32], *missForest* [33], *Amelia* [17] and *MICE* package [30].

- **Prediction with 3-MLP in R software**

Before performing the prediction, 75% of each imputed dataset is used to train the neural network and 25% to test trained network concerning its generalization capacity. Before performing the training and testing, the imputed datasets were normalized using min-max normalization technique [34]:

$$new_v = \frac{v - \min_z}{\max_z - \min_z}(new \max_z - new \min_z) + new \min_z \tag{9}$$

where v is an observation of vector z and new_v is a normalized observation.

The function "*mlp*" of *RSNNS* package [35] was used for the prediction. A 3-MLP model (see equation (5)) was used by varying hyper parameters for each sample size of imputed dataset. 4 combinations of activation functions, AF (f_1 and f_2, see equation 5) were used: (i) Logistic-Linear (LL); (ii) Logistic-Exponential (LE); (iii) TanH-Linear (TL) and (iv) TanH-Exponentiel (TE). The expression of activation functions considered are: Linear, $h(x) = x$; Logistic, $h(x) = \frac{1}{1+e^{-x}}$; Exponential, $h(x) = e^x$ and Tangent hyperbolic, $h(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. In additional, 13 numbers of nodes (Node) in the hidden layer were considered: $3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14$ and 15. In addition, 6 learning rates (LR) were considered: $20\%, 30\%, 40\%, 50\%, 60\%$ and 70% and as well as the 5 sample sizes of imputed dataset (Size). The considered learning algorithm is Standard back propagation (see the section 2).

A total of 100 replications was performed on each size of imputed dataset to the analyze performance of the method. Initial weights were generated randomly according to the uniform law in the range -3 and 3. The stopping criteria used are the combination of a fixed number of epochs, NE= 1000 and a sufficiently small training error less than or equal to 10^{-6}.

- **Performance criteria and statistical method comparison**

The performance criteria used are: (i) Coefficient of correlation, r; (ii) Coefficient of determination, R^2 and (iii) Root Mean Squared Error, $RMSE$ [36,37]. In the formula below, y and ζ_θ respectively denote observed outputs and predicted outputs, $\bar{y}$ and $\bar{\zeta}_\theta$, their mean and n the test data size.

$$r = \frac{\sum(y_t - \bar{y}_t)(\zeta_\theta(x_t) - \bar{\zeta}_\theta(x_t))}{\sqrt{\sum(y_t - \bar{y}_t)^2} \times \sqrt{\sum(\zeta_\theta(x_t) - \bar{\zeta}_\theta(x_t))^2}} \tag{10}$$

$$R^2 = \frac{\sum(y_t - \zeta_\theta(x_t)) \times (\sum y_t \times \sum \zeta_\theta(x_t))}{\sqrt{(\sum y_t^2 - (\sum y_t)^2)(\sum \zeta_\theta(x_t)^2 - (\sum \zeta_\theta(x_t))^2)}} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\Big(F_\theta(x_t) - y_t\Big)^2} \tag{12}$$

The appropriate imputation method for a missing data mechanism giving the best configuration of model characteristics (5) with the BP algorithm and for an optimal sample size is the model for which we observe a high correlation between predicted and observed data ($|r| \geq 0.8$) [36], with R^2 close to "1" [38] and with a low value of $RMSE$ [36].

To assess effects of factors (Size, MR, AF, Node and LR) which affect performance of the 3-MLP model, the generalized linear models based on the beta distribution were run on R^2, $|r|$ and the linear fixed effects models on $RMSE$ for each missing data mechanism and by imputation method.

Interaction plot was considered for significant interactions between the MLP hyper parameters by missing data mechanism.

$Mean$, $minimum$, $maximum$ and $coefficient\ of\ variation$ of the criteria considered (R^2, $|r|$ and $RMSE$) were used to compare imputation method performances.

4. Results

- **Effect of imputation methods by missing data mechanism on the performance of the hyper parameter structure of the 3-MLP model**

Table 1 shows the results of the effect of the imputation methods (Amelia, LOCF, RF and MICE) by missing data mechanism (MAR, MACR and MNAR) on the performance of the hyper parameter structure (AF, LR and Node) of the 3-MLP model. The analysis shows that AF, LR and Node significantly affect the performances of the imputation methods whatever the missing data mechanism ($p < 0.05$). However, the second order interaction of these factors (AF:LR:Node) did not impact ($p > 0.05$) the performances of imputation methods for the three missing data mechanism. The predictive performances (R^2 and r) of the imputation methods used were not affected by the interraction between learning rate and the number of neurone in the hidden layer (LR:Node) but had a significant impact on the root mean square error (RMSE) for each missing data mechanism. By considering the interraction between the activation function and the number of neurone in the hidden layer (AF:Node), we observed that from a missing data mechanism to another, the predictive performances of AMELIA and LOCF were not affected by this interaction. However those of RF and MICE were significantly affected. About the RMSE, apart the one of LOCF under MAR assumption, the others were significantly affected by this interraction. Results also revealed a significant effect on the performances of imputation methods concerning the interraction between the activation function and the learning rate (AF:LR) for all missing data mechanism.

Table 1. Effect of imputation methods by missing data mechanism on the structure of hyper-parameters: Results of GLM and linear models

Factors	Amelia			LOCF			RF			MICE		
	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE	r
MAR												
AF	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
LR	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Node	0.002	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
AF:LR	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
AF:Node	0.439	0.001	0.434	0.526	0.092	0.500	0.001	0.001	0.001	0.001	0.001	0.001
LR:Node	0.999	0.001	0.999	0.999	0.098	0.999	0.999	0.001	0.999	0.999	0.001	0.999
AF:LR:Node	0.999	0.001	0.999	0.999	0.001	0.999	0.999	0.001	0.999	0.999	0.001	0.999
MCAR												
AF	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
LR	0.001	0.001	0.001	0.997	0.001	0.994	0.001	0.001	0.001	0.001	0.001	0.001
Node	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
AF:LR	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
AF:Node	0.999	0.001	0.999	0.999	0.001	0.999	0.001	0.001	0.001	0.001	0.001	0.001
LR:Node	0.999	0.001	0.086	0.999	0.098	0.999	0.999	0.203	0.999	0.999	0.001	0.999
AF:LR:Node	0.999	0.001	0.999	0.999	0.001	0.999	0.999	0.001	0.999	0.999	0.001	0.999
MNAR												
AF	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
LR	0.001	0.001	0.001	0.997	0.001	0.994	0.001	0.001	0.001	0.001	0.001	0.001
Node	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
AF:LR	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
AF:Node	0.783	0.001	0.776	0.872	0.001	0.870	0.001	0.001	0.001	0.001	0.001	0.001
LR:Node	0.999	0.001	0.999	0.999	0.036	0.999	0.999	0.001	0.999	0.999	0.002	0.999
AF:LR:Node	0.999	0.001	0.999	0.999	0.001	0.999	0.999	0.001	0.999	0.999	0.001	0.999

Cells contain p-value; AF: activation function, LR: learning rate

- **Effect of imputation methods by missing data mechanism on the performance of activation function and learning rate**

The interaction plots revealed that under MAR assumption, the performances of imputation methods increase with the learning rate when we use activation functions such that TanH-Linear (TL), Logistic-Linear (LL) and Logistic-Exponential (LE), see figure 2. Contrary to those activation functions, TanH-Exponential (TE) starts to decrease after 30% of learning rate. The best values of R^2 and r was obtained with the

TanH-Linear activation function followed by Logistic-Exponential and Logistic-Linear. About the RMSE, the Logistic-Exponential yield the highiest values indicating that the network commit more error with this activation function. TanH-Exponential activation function gave the best RMSE. With this activation function the error vary slightly when learning rate increase contrary to the others function which increased when the learning rate increase. For TanH-Linear and Logistic-Linear, the RMSE was closed but after 40% Logistic-Linear yield a RMSE greater than the other one.

Similar trend have been observed when the missingness mechanism is either MCAR or and MNAR. Thus, highest values of R^2 and r have been observed with TanH-Linear while lowest value have been observed with TanH-Exponential. As observed under MAR assumption, the TanH-Exponential function commit little error when the data is MCAR or MNAR. For the three missingness mechanism, the predictive performances of the imputation methods vary slightly after 50% of learning rate indicating that the neural network can be trained with 50% of learning rate for each activation function. More a little variation of the error has been observed from 50% of learning rate for Tanh-Exponential and TanH-Linear contrary to Logistic-Exponential and Logistic-Linear which continuous to increase.

- **Effect of imputation methods by missing data mechanism on the performance of activation function and Node**

The performances of imputation methods according to the activation function and the number of node in the hidden layer for the three missingness mechanisms revealed almost the same performance, see only figure 3. The predictive performances of imputation methods improve with the increase of the number of nodes for all missing data mechanism. When data is missing at random (MAR), R^2 and r values for the TanH-Linear activation function were greater than the values recorded with the other activation functions. The Logistic-Linear yield the lowest values of R^2 and r. The same trend has been observed under MNAR assumption. However, under MCAR assumption, it is TanH-Exponential activation function which had the lowest values of R^2 and r for LOCF method. Concerning the errors commit by the model, it became more and more lower when the number of hidden neurones increased and this for all the imputation methods for the three missing data mechanism. The model commits more errors with the Logistic-Exponential activation function when the TanH-Exponential functions commit less errors. The errors when the activation functions are Logistic-Linear and TanH-Linear were lower than the one with Logistic-Exponential. For this two activation function, the RMSE was similar from 3 to 7 neurons. After 7 nodes, the model with TanH-Linear was better than the one with Logistic-Exponential. The trend of RMSE observed under MAR assumption was similar to the one observed under MCAR and MNAR assumption.

We also noticed that for the three missing data mechanism a low variation of the performances (R^2, r and RMSE) was observed whatever the imputation method used after 11 nodes in the hidden layer.

- **Effect of imputation methods on size and the missing data rate**

Table 2 shows how sample size, missing rate and their interaction affect the perfor-

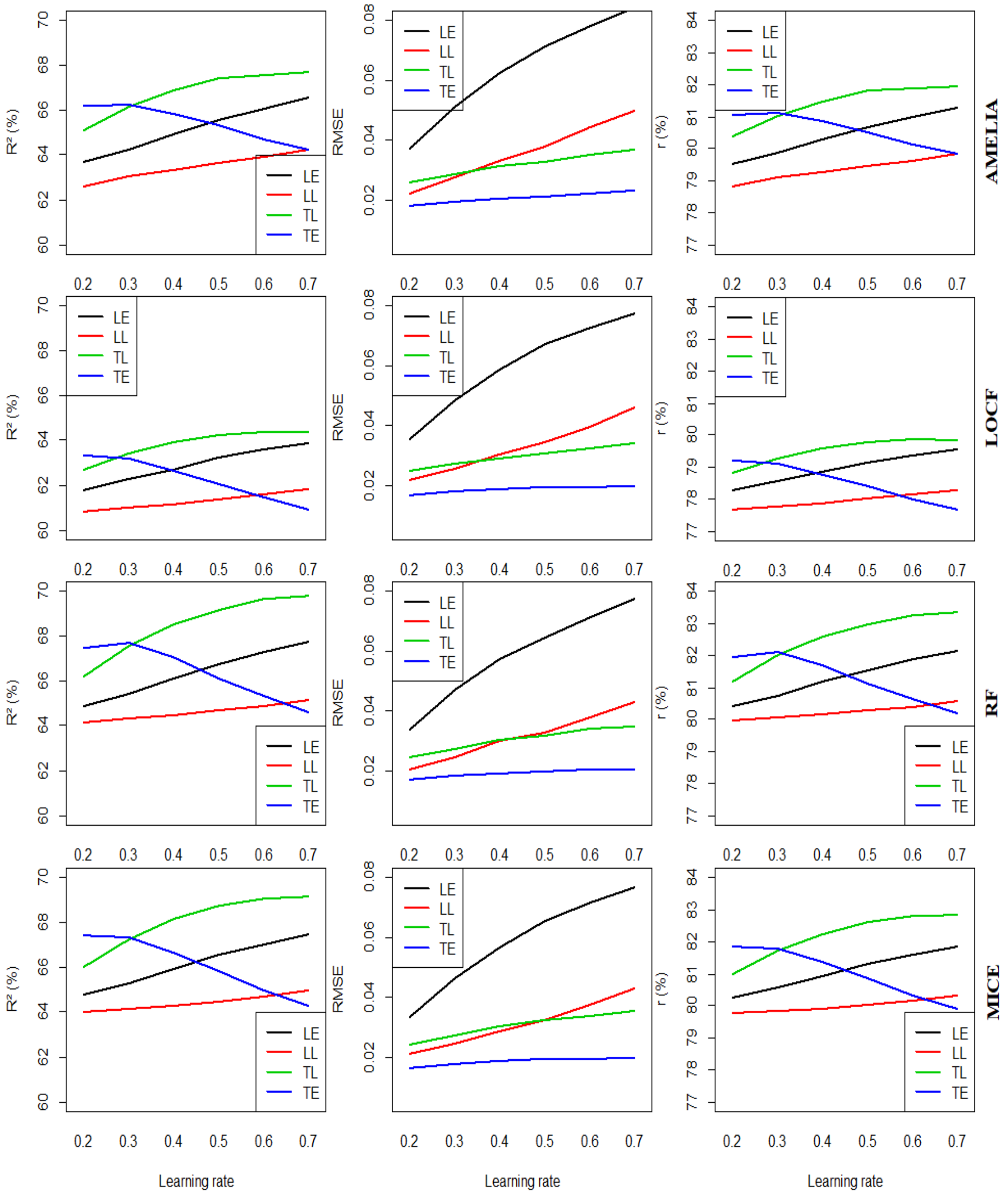

Figure 2. Interaction plot of AF:LR for R^2, RMSE and r under MAR assumption

mances of imputation method according to the missing data mechanism. The interaction between size and missing rate highly significantly affected the performances of imputation methods whatever the missing data mechanism ($p < 0.01$). The interaction plot under MAR, MCAR and MNAR are showed similar trends and only for MAR presented on figure 4.

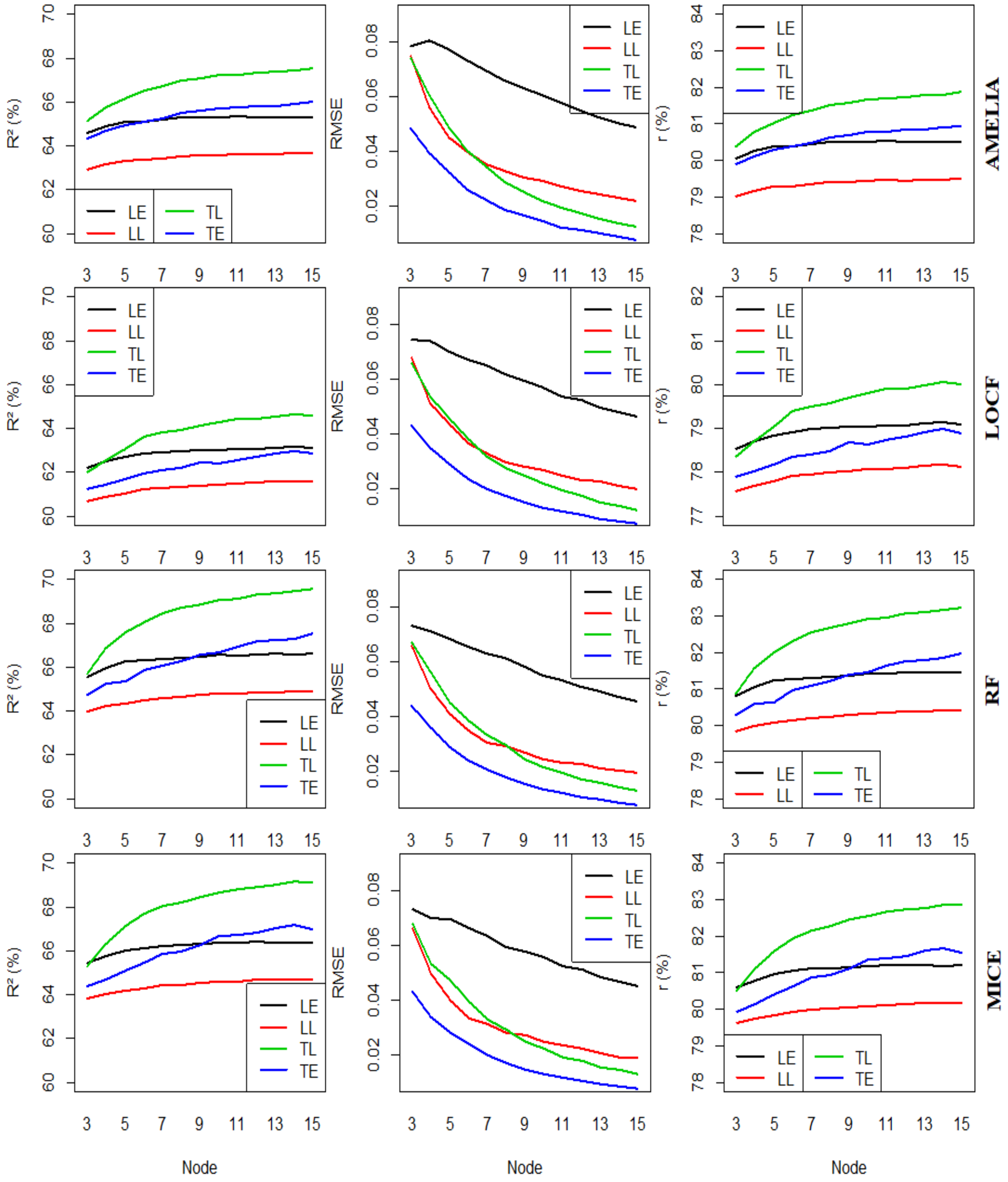

Figure 3. Interaction plot of AF:Node for R^2, RMSE and r under MAR assumption

When the data is missing at random (MAR), an improvement of R^2 and r had been noticed for LOCF method from 50 to 200 sample size whatever the missing data rate considered. But after 200, the predictive performances start to decrease. The RMSE for this method under the same missingness assumptions followed the same trend. It was closed between the missing data rate and it vary slightly from

Table 2. Effect of imputation methods on size and missing data rate: Results of GLM and linear models

Factors	Amelia			LOCF			RF			MICE		
	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE	r	R^2	RMSE	r
						MAR						
Size	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
MR	0.001	0.001	0.001	0.001	0.101	0.001	0.001	0.045	0.001	0.001	0.074	0.001
Size:MR	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
						MCAR						
Size	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
MR	0.001	0.001	0.001	0.001	0.108	0.001	0.001	0.016	0.001	0.001	0.165	0.001
Size:MR	0.001	0.001	0.001	0.001	0.129	0.001	0.001	0.060	0.001	0.001	0.001	0.001
						MNAR						
Size	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
MR	0.001	0.001	0.001	0.001	0.038	0.001	0.001	0.182	0.001	0.001	0.005	0.001
Size:MR	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Cells contain p-value

50 to 200 sample size but increased after 200 sample size. For RF and MICE the predictive performances increased when the sample size is between 50 and 200 whatever the missing data rate. However after 200, predictive performances vary slightly. Under 200, RMSE did not vary greatly but incread from 200. With Amelia, large difference has been noticed among missing rate under 200 sample size for R^2 and r. However, for low missing rate (10 and 20), the performances were better. After 200 sample size, no major difference have been noticed. The trend of RMSE show that the error is less with low missing data rate and increase after 100 sample size.

Under missing completely at random (MCAR) assumption, the error was closed for all missing data rate for LOCF, Random Forest and MICE. It vary slightly under 200 sample size. The predictive performances of LOCF was best with 10% and 40% of missing rate at 200 and 300 sample size respectively. However the RMSE was greater with 40% of missing rate. About Random Forest and MICE R^2 and r values were better with 10% and 20% of missing rate respectively at 200 sample size.
When data is missing not at random (MNAR), the performances obtained for LOCF were similar to what obtained under MAR assumptions. R^2 and r increased between 50 and 200 sample size whatever the missing data rate but decrease after 200. Error was closed between missing rate and was best under 200 sample size. The performances of Random Forest method is best with 40% of missing rate at 200 sample size. However with 20, 30 and 50% of missing rate, values of R^2 and r were closed to the one obtained with 40%. The error did not vary among missing data rate. With MICE method, the error did not vary among missing data rate as observed with Random Forest. Values of R^2 and r were better with 10, 30 and 50% of missing data rate than the values with 20 and 40%. However the difference were not important. Concerning Amelia method,

large variation had been observed among missing data rate for all sample size. The error vary slightly under 100 as sample size and increased beyond 100. R^2 and r were better for 10 and 20% of missing data rate at 100 sample size.

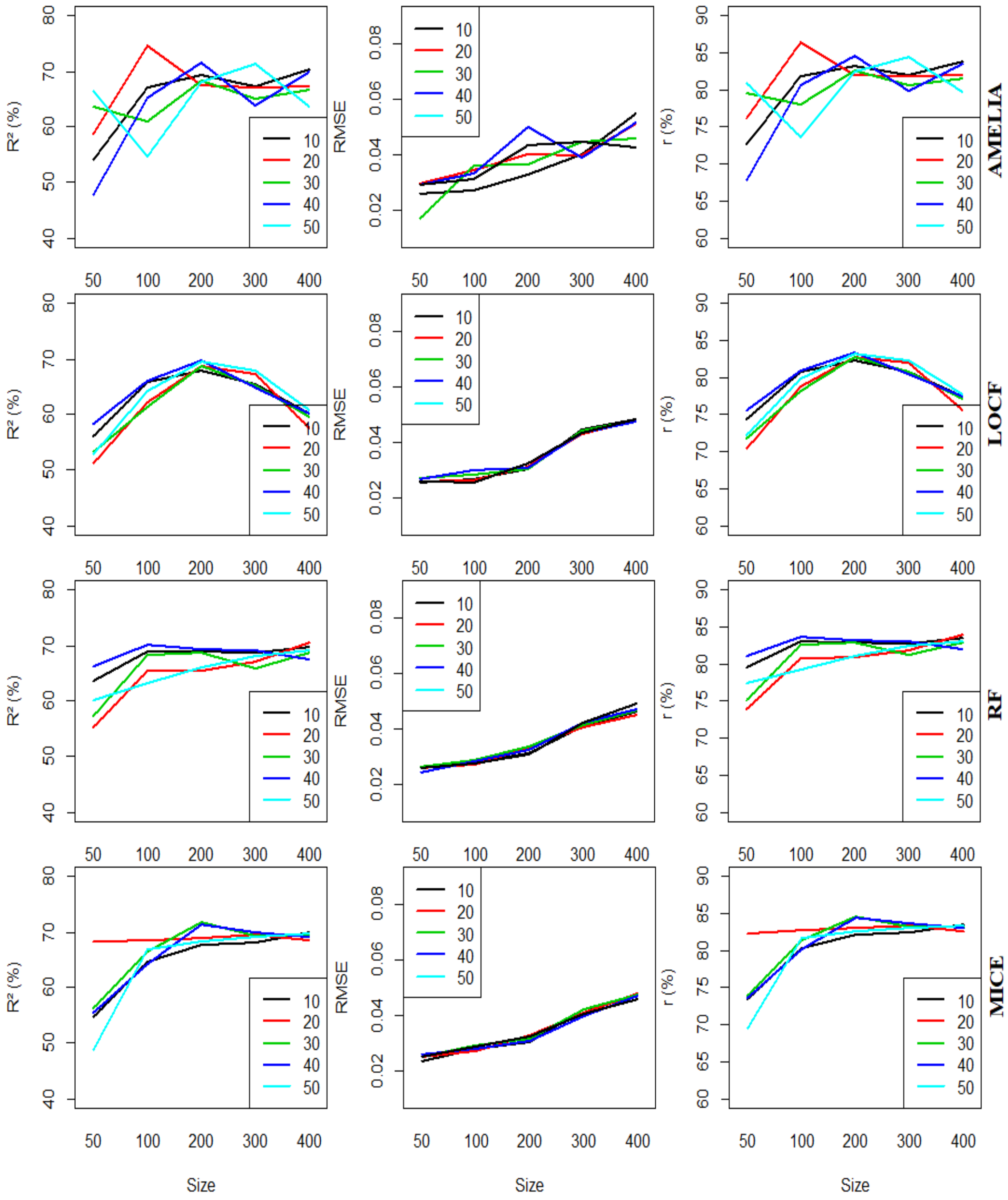

Figure 4. Interaction plot of Size:MR for R^2, RMSE and r under MAR assumption

- **Comparison of imputation methods**

The mean and coefficient of variation the performance criteria according to imputation method and missing data mechanism are presented in Table 3. There is not a great variation among imputation method under the assumption that data is missing at random. However, LOCF method has the lowest value of R^2 (62.59%). For the other imputation methods used in this study, the values obtained were closed (65.22%; 66.46% and 66.19% respectively for AMELIA, RF and MICE). The error commits by the model (RMSE) was similar for all imputation methods. The coefficient of correlation was also low with LOCF (78.76%) compared to the others methods (80.45%; 81.44% and 81.07% respectively for AMELIA, RF and MICE).

When data is missing completley at random (MCAR), a similar trend is observed like under MAR assumption. The lowest R^2 and r were obtained with LOCF method (59.28%) when those of AMELIA, RF and MICE was respectively 66.68%, 65.77% and 66.34% for R^2. About the coefficient of correlation, it also mow (76.64%) when using LOCF for imputation comparatively to AMELIA (81.43%), RF (80.78%) and MICE (81.19%). The RMSE was similar between methods (0.035 for RF and MICE; 0.037 and 0.034 respectively for AMELIA and LOCF).

Under the assumption that data is missing not at random (MNAR), the trend for R^2 and r were different contrary to the values observed when data is MAR and MCAR. R^2 was lower with AMELIA (64.49%) and LOCF (63%) than those of RF (67.35%) and MICE (66.82%). As observed with the other missing data mechanism, RMSE was similar under MNAR assumption. The error was 0.036 for AMELIA, 0.035 for LOCF and RF and 0.034 for MICE. About the coefficient of correlation, it does not vary greatly from a method to another. Thus we recorded 79.96%; 79.03%; 81.84% and 81.55% for AMELIA, LOCF, RF and MICE respectively.

Table 3. Mean and coefficient of variation of performances criterion according to imputation method and missing data mechanism

		MAR	MCAR	MNAR
AMELIA	Rsquare	65.22(15.88)	66.68(14.37)	64.49(16.79)
	RMSE	0.038(88,53)	0.037(87.74)	0.036(88.32)
	r	80.45(8.74)	81.43(7.55)	79.96(9.33)
LOCF	Rsquare	62.59(17.58)	59.28(17.82)	63(16.76)
	RMSE	0.035(89.15)	0.034(91.40)	0.035(89.41)
	r	78.76(9.57)	76.64(9.66)	79.04(9.14)
RF	Rsquare	66.46(12.62)	65.77(15.56)	67.35(13.79)
	RMSE	0.035(88.26)	0.035(87.71)	0.035(88.94)
	r	81.34(6.58)	80.78(8.90)	81.84(7.39)
MICE	Rsquare	66.19(15.46)	66.34(14.85)	66.82(13.03)
	RMSE	0.035(89.17)	0.035(88.59)	0.034(88.86)
	r	81.07(8.35)	81.19(7.98)	81.55(6.81)

5. Discussion

• Effect of imputation method by missing data mechanism on the structure of hyper parameters of 3-MLP models

For each imputation method, the interaction between AF:LR and AF:Node significantly impact the performances of the network for any missing data mechanism. The performance of the 3-MLP models is best when the network is trained with the TanH-Linear activation function, 11 nodes in the hidden layer and a learning rate of 50%. The accuracy of TanH-Linear activation function is much better than the other functions. For the number of node, even if the error continuous to decrease after 11 nodes, the gain of the model in term of prediction has not increased considerably. More, [39], [37] and [38] suggest to set the number of node in the hidden layer to a minimum as possible because a network with large number of nodes increases the computational time needed for training. Our findings about the number of nodes in the hidden layer are in agreement with those of [40] which state that the best approach to set the number of node in the hidden layer is to start with small number of node and increase until no major improvement in the performances is obtained. As regards the learning rate, after 50%, the predictive ability of the network is still increasing. But this increase in the predictive ability is not important and larger learning rate causes network to be more unstable as the error increase. Our results for the optimum learning rate are in agreement with those of [41] who says that if the value of the learning rate is large, the network may show oscillatory response because of the larger changes in the synaptic weight which may cause network to be unstable. However our optimum value of learning rate (50%) is less than 60% suggested by [42]. Another study conducted by [23] set the optimum learning rate as 35% which is less than the one of [42] and the one of our study. This difference can be due to the domain of application which is different.

• Effect of imputation methods on size and missing data rate

Both size and missing data rate affect imputation methods. No matter the mechanism and the method used, the error increased when sample size increase for all missing rate. Apart from Amelia, the optimal size is 200 for the others methods under the three missingness mechanism. For Amelia the optimal sample size is 200 under MAR and MCAR assumption but 100 when the missingness mechanism is MNAR. LOCF can support missing data rate up to 50% with an optimal sample size of 200 under MAR and MNAR assumption. This method perform better with 10% under MCAR assumption. Since differences between missing data rate are not important, Random Forest and MICE can support up to 50% of missing rate at an optimal sample size of 200. Results are not in agreement with those of [43] who found that error decreased when the sample size increase no matter the missing rate. The difference might be explained by the fact that in our study imputed data pass through the network before evaluating the performances.

• Comparison of imputation methods

Four imputation methods have been used in this study and results show that there is not a great variation among imputation method under the assumption that data is missing at random (MAR) and missing not at random (MNAR). However, Random Forest method and MICE seem to perform well than AMELIA and LOCF since they have less error. Our findings are in agreement with the results of [44]; [45]. These authors compare nine imputation methods by considering the three missingness mechanism (MAR, MCAR and MNAR). They found that MICE multiple imputation are overall the best approach. Another research of [33] compared the random forest method to kNN imputation [46], MissPALasso (a method based on EM algorithm, proposed by [47] and MICE [30]). For these authors, random forest could outperform other imputation methods. Results of this authors are similar to our findings. Indeed in this study random forest and MICE yields similar performances.

Under MCAR assumptions, LOCF is not indicated to handle missing data since it gives low R^2. This agree with the conclusion of [48] who states that single imputation and LOCF are not optimal approaches for missing values imputation, as they can cause bias and lead to invalid conclusions. More, [49] states that single imputation are not solidly grounded in mathematical foundations and they exist merely for their ease of implementation. Most of imputation methods assume that data is missing at random. Our results show that even if this assumption is violated they perform well since the performances recorded for AMELIA, RF and MICE do not vary greatly from a missing data mechanism to another. This findings are in agreement with [50] who state that MICE is especially suitable in MAR settings. But [51] and [52] point out that MICE is also capable to deal with MNAR schemes.

6. Conclusions

The possibility to combine imputation methods to multilayer neural network have been accessed in this study through four methods (Amelia, LOCF, Random Forest and MICE) for any missing data mechanism by controlling the hyper-parameters (activation function, number of hidden neurons, learning rate). From our findings, single imputation is not an optimal approach to deal with missing data. However MICE multiple imputation and RF are more appropriate. Even if these methods outperform the two others (Amelia and LOCF), the best solution is to employ maximal efforts to avoid missing data during data collection. With regards to hyperparameters, to learn the model with the backpropagation algorithm, the performance criteria showed that the combination of TanH-Linear activation functions is best suited to implement the network with 11 nodes in the hidden layer with a learning rate of 50%. However, for further studies, most adapted and developed methods have to be compared with the best method found in this study using other learning methods.

Acknowledgement(s)

The authors are very grateful to the Editor-in-Chief, the Associate Editor and the two anonymous Reviewers for the careful review and valuable comments on earlier versions of this paper. He's work is supported by African Centre of Excellence in Mathematics and Applications (CEA-SMA). We thank this institution and its donors.

doi:10.20944/preprints202105.0390.v1

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Khaled AY, Abd Aziz S, Bejo SK, et al. A comparative study on dimensionality reduction of dielectric spectral data for the classification of basal stem rot (bsr) disease in oil palm. Computers and Electronics in Agriculture. 2020;170:105288.

[2] Karthikeyan K, Kumar N, Yousuf A, et al. Land evaluation: A general perspective. Watershed Hydrology, Management and Modeling. 2019;:175.

[3] Zhang C, Pan X, Li H, et al. A hybrid mlp-cnn classifier for very fine resolution remotely sensed image classification. ISPRS Journal of Photogrammetry and Remote Sensing. 2018; 140:133–144.

[4] Ul-Saufie AZ, Yahya AS, Ramli NA, et al. Comparison between multiple linear regression and feed forward back propagation neural network models for predicting pm10 concentration level based on gaseous and meteorological parameters. International Journal of Applied. 2011;1(4):42–49.

[5] Ghazanfari-Hashemi S, Etemad-Shahidi A, Kazeminezhad MH, et al. Prediction of pile group scour in waves using support vector machines and ann. Journal of Hydroinformatics. 2011;13(4):609–620.

[6] Caselli M, Trizio L, De Gennaro G, et al. A simple feedforward neural network for the pm 10 forecasting: Comparison with a radial basis function network and a multivariate linear regression model. Water, Air, and Soil Pollution. 2009;201(1):365–377.

[7] Torres Munguía JA. Comparison of imputation methods for handling missing categorical data with univariate pattern. Revista de Métodos Cuantitativos para la Economía y la Empresa. 2014;17:101–120.

[8] Finch WH. Imputation methods for missing categorical questionnaire data: A comparison of approaches. Journal of Data Science. 2010;8(3):361–378.

[9] Baraldi AN, Enders CK. An introduction to modern missing data analyses. Journal of school psychology. 2010;48(1):5–37.

[10] Graham JW, Taylor BJ, Olchowski AE, et al. Planned missing data designs in psychological research. Psychological methods. 2006;11(4):323.

[11] Sijtsma K, Van der Ark LA. Investigation and treatment of missing item scores in test and questionnaire data. Multivariate Behavioral Research. 2003;38(4):505–528.

[12] Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annual review of public health. 2000; 21(1):121–145.

[13] Dong Y, Peng CYJ. Principled missing data methods for researchers. SpringerPlus. 2013; 2(1):1–17.

[14] De Leeuw ED, Hox JJ, Huisman M. Prevention and treatment of item nonresponse. Journal of Official Statistics. 2003;19:153–176.

[15] Schafer JL. Analysis of incomplete multivariate data. CRC press; 1997.

[16] Cottrell M, Ibbou S, Letrémy P. Traitement des donnees manquantes au moyen de l'algorithme de kohonen. arXiv preprint arXiv:07041709. 2007;.

[17] Honaker J, King G, Blackwell M, et al. Amelia ii: A program for missing data. Journal of statistical software. 2011;45(7):1–47.

[18] Preda C, Duhamel A, Picavet M, et al. Tools for statistical analysis with missing data: application to a large medical database. Studies in health technology and informatics. 2005;116:181–186.

[19] Hox JJ. A review of current software for handling missing data. Kwantitatieve methoden. 1999;20:123–138.

[20] Donzé L. L'imputation des données manquantes, la technique de l'imputation multiple, les conséquences sur l'analyse des données: l'enquête 1999 kof/ethz sur l'innovation. 2001; .

[21] Allison PD. Multiple imputation for missing data: A cautionary tale. Sociological methods & research. 2000;28(3):301–309.

[22] Chaturvedi D. Factors affecting the performance of artificial neural network models. Soft Computing: Techniques and its Applications in Electrical Engineering. 2008;:51–85.

[23] Nagori V. Fine tuning the parameters of back propagation algorithm for optimum learning performance. In: 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I); IEEE; 2016. p. 7–12.

[24] Gregori P, Felip-Bardoll JV, Gras R. Imputation multiple de donnees manquantes par l'analyse statistique implicative. Quaderni di Ricerca in Didattica (Mathematics). 2010; :20–1.

[25] Huisman M. Imputation of missing item responses: Some simple techniques. Quality and Quantity. 2000;34(4):331–351.

[26] Zainuri NA, Jemain AA, Muda N. A comparison of various imputation methods for missing values in air quality data. Sains Malaysiana. 2015;44(3):449–456.

[27] Hounmenou CG, TOHOUN R, GNEYOU KE, et al. Empirical determination of optimal configuration for characteristics of a multilayer perceptron neural network in nonlinear regression. Afrika Statistika. 2020;15(3):2413–2429.

[28] IFWA. Assess the effect of maggot meal flour on the growth and economic performances of guinea fowl. Insect as Feed for West Africa project in Laboratoire de Recherche Avicole et de Zoo-Economie; 2017.

[29] Tibshirani RJ, Efron B. An introduction to the bootstrap. Monographs on statistics and applied probability. 1993;57:1–436.

[30] Groothuis-Oudshoorn K, Van Buuren S. Mice: multivariate imputation by chained equations in r. J Stat Softw. 2011;45(3):1–67.

[31] Team RC. R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria. version 3.3.6. https://wwwR-projectorg/. 2019;.

[32] Zeileis A, Grothendieck G, Ryan J, et al. S3 infrastructure for regular and irregular time series (z's ordered observations) r package version. R Foundation for Statistical Computing. 2015;.

[33] Stekhoven D, Bühlmann P. missforest: Nonparametric missing value imputation using random forest r package version 1.3 ; 2012.

[34] Priddy KL, Keller PE. Artificial neural networks: an introduction. Vol. 68. SPIE press; 2005.

[35] Bergmeir CN, Benítez Sánchez JM, et al. Neural networks in r using the stuttgart neural network simulator: Rsnns. American Statistical Association; 2012.

[36] Kazem HA, Jabar HY. Comparison of prediction methods of photovoltaic power system production using a measured dataset. Energy Conversion and Management. 2017; 148:1070–1081.

[37] Elarabi H, Taha NF. Effect of different factors of neural network on soil profile of khartoum state. American Journal of Earth Sciences. 2014;3(1):62–66.

[38] Shahin MA, Jaksa MB, Maier HR. State of the art of artificial neural networks in geotechnical engineering. Electronic Journal of Geotechnical Engineering. 2008;8(1):1–26.

[39] Macukow B. Neural networks state of art, brief history, basic models and architecture. In IFIP international conference on computer information systems and industrial management. 2016;23(4):3–14.

[40] Nawari LR N O, Nusairat J. Artificial intelligence techniques for the design and analysis of deep foundations. Electronic Journal of Geotechnical Engineering. 1999;(4):1–21.

[41] Uma Rao K. Artificial intelligence and neural networks. ; 2011.

[42] Rajasekaran S, Pai GV. Neural networks, fuzzy logic and genetic algorithm: synthesis and applications (with cd). PHI Learning Pvt. Ltd.; 2003.

[43] Kalkan OK, Yusuf KARA, Kelecoioğlu H. Evaluating performance of missing data impu-

tation methods in irt analyses. International Journal of Assessment Tools in Education. 2018;5(3):403–416.

[44] Durrant GB, et al. Imputation methods for handling item-nonresponse in the social sciences: a methodological review. ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute NCRM Methods Review Papers NCRM/002. 2005;.

[45] Glasson-Cicognani M, Berchtold A. Imputation des données manquantes: Comparaison de différentes approches. In: 42èmes Journées de Statistique; 2010.

[46] Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for dna microarrays. Bioinformatics. 2001;17(6):520–525.

[47] Städler N, Bühlmann P. Pattern alternating maximization algorithm for high-dimensional missing data. arXiv preprint arXiv:10050366. 2010;.

[48] Kang H. The prevention and handling of the missing data. Korean journal of anesthesiology. 2013;64(5):402.

[49] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological methods. 2002;7(2):147.

[50] Janssen KJ, Donders ART, Harrell Jr FE, et al. Missing covariate data in medical research: to impute is better than to ignore. Journal of clinical epidemiology. 2010;63(7):721–727.

[51] He Y, Zaslavsky AM, Landrum M, et al. Multiple imputation in a large-scale complex survey: a practical guide. Statistical methods in medical research. 2010;19(6):653–670.

[52] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine. 2011;30(4):377–399.