JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME 10 (ISSUE 1) JUNE 2026: 29-38
https://doi.org/10.21009/JSA.10103

MISSING DATA IMPUTATION ON BIVARIATE GAMMA GENERATION DATA USING PREDICTIVE MEAN MATCHING AND RANDOM FOREST METHODS

Muhammad Arib Alwansyah[1*], Jose Rizal[2], Ramya Rachmawati[3]

[1]Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta Jln. Rawamangun Muka, Jatinegara Kaum, Pulo Gadung, ADM. East Jakarta, 13220, Indonesia
[2,3]Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Bengkulu, Jln. WR. Supratman Kandang Limun, Muara Bangkahulu, Bengkulu, 38122, Indonesia

Corresponding author's e-mail: * ***muhammadarib@unj.ac.id***

ABSTRACT

Missing data is a common problem in data analysis and can reduce the quality and accuracy of research results if not handled properly. This study aims to compare the Predictive Mean Matching (PMM) and Random Forest (RF) imputation methods in handling missing data with missing levels of 5%, 10%, 15%, and 20% using correlation indicators, p-values, and observing the smallest Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) values. The results show that both methods differ at each level of missing data. At 5% missing data, both methods show significant differences to the original data with a p-value smaller than $\alpha = 0.05$, but the RF method produces smaller MAPE and RMSE values than PMM. At 10% missing data, the PMM method still shows significant differences to the original data, while the RF method does not. At 15% missing data, the PMM method showed results that were not significantly different from the original data and had smaller MAPE and RMSE values than RF. Meanwhile, at 20% missing data, the RF method produced the highest correlation value of 0.7788 compared to PMM at 0.7638. In general, the results of the study indicate that the greater the proportion of missing data, the imputation error rate also tends to increase. Therefore, the selection of imputation methods needs to be adjusted to the characteristics and proportion of missing data to obtain optimal imputation results.

Article History:
Received: May 24, 2026
Revised: June 05, 2026
Accepted: June 16, 2026
Published: June 30, 2026
Available online.

Keywords:
Bivariate Gamma;
Mean Absolute Percentage Error;
Predictive Mean Matching;
Random Forest Imputations;
Root Mean Square Error.

How to cite this article:
Alwansyah, M. A., Rizal, J., Rachmawati, R., "MISSING DATA IMPUTATION ON BIVARIATE GAMMA GENERATION DATA USING PREDICTIVE MEAN MATCHING AND RANDOM FOREST METHODS", Journal Statistika dan Aplikasinya, vol. 10, iss. 1, pp. 29 – 38, June 2026

Journal homepage: *https://journal.unj.ac.id/unj/index.php/statistika*
Journal e-mail: *jsa@unj.ac.id*
Research Article · Open Access

1. INTRODUCTION

Missing data is a common problem in research and can impact analysis results, reduce estimation efficiency, and introduce bias into statistical analysis if not handled properly. Incomplete data can occur in both empirical and simulated data or generated data. One distribution widely used to model the relationship between two correlated variables is the Bivariate Gamma distribution. This distribution has widespread applications in various fields, such as actuarial science, epidemiology, hydrology, reliability engineering, and risk analysis, as it is able to represent the dependency relationship between two positive random variables and exhibits a skewed distribution pattern [1]. However, when some data in the Bivariate Gamma distribution is missing, further analysis processes, such as parameter estimation, hypothesis testing, and modeling relationships between variables, can be disrupted, potentially resulting in inaccurate conclusions.

Various imputation methods have been developed to address the problem of missing data. Conventional imputation methods, such as mean imputation and median imputation, remain widely used due to their simplicity. However, these methods tend to ignore the natural variability of the data and are less able to capture non-linear relationships between variables, resulting in suboptimal estimates. One popular modern imputation method is Predictive Mean Matching (PMM). The PMM method works by matching predicted values from missing data with observational data that have the closest predicted values, then using the actual values from these observations as the imputation results [2].

In addition to PMM, the development of machine learning techniques has also encouraged the use of the Random Forest Imputation (RF) imputation method. The RF method is an ensemble learning approach that randomly constructs some decision trees and aggregates their predictions. The combination of bagging techniques and random feature selection allows RF to capture complex, non-linear relationships and interactions between variables that are difficult to handle with traditional parametric methods [3]. Several studies have shown that the RF method has good resilience to outliers, can handle mixed data types, and provides superior performance on data with a high level of skewness [4]. The application of the PMM and RF methods to bivariate gamma-generated data is important to study because both methods have different approaches in imputing missing data. PMM tends to maintain the characteristics of the original data distribution, while RF is superior in capturing complex, non-linear relationships between variables. Therefore, this study was conducted to compare the two imputation methods in handling missing data in the Bivariate Gamma distribution. The results of this study are expected to contribute to the development of imputation methods that are more effective, accurate, and appropriate to the characteristics of the data being analyzed [5].

2. METHODS

Bivariate Gamma Distribution

The Bivariate Gamma Distribution is an extension of the univariate Gamma distribution used to model two correlated continuous random variables with positively skewed distribution characteristics. This distribution is widely applied in various fields, such as insurance, reliability analysis, survival analysis, and hydrology, especially when the two response variables are dependent, and the distribution pattern is asymmetric to the right [6]. For example, if $(\boldsymbol{X}, \boldsymbol{Y})$ is a pair of random variables that follow the Bivariate Gamma distribution, and their dependency structure can be constructed using a shared component model:

$$\boldsymbol{Y} = \boldsymbol{W} + \boldsymbol{V}, \qquad \boldsymbol{X} = \boldsymbol{U} + \boldsymbol{V}$$

where $\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{V}$ ~Gamma $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$ They are independent of each other. Thus, the $\boldsymbol{V}$ The component is the cause of the positive correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ [7].

Characteristics of the Bivariate Gamma Distribution

The Bivariate Gamma Distribution has several important characteristics, including expected value, variance, covariance, and correlation, which are used to describe the relationship between two continuous random variables. In the shared component model approach, the dependency relationship between variables is established through shared components, allowing each distribution characteristic to be derived based on its constituent parameters.

The expected value of the Bivariate Gamma distribution is expressed as:

$$E(X) = \frac{\alpha_U}{\lambda} + \frac{\alpha_V}{\lambda} \quad (1)$$

$$E(Y) = \frac{\alpha_W}{\lambda} + \frac{\alpha_V}{\lambda} \quad (2)$$

Meanwhile, the variance of each variable is given by:

$$\mathrm{Var}(X) = \frac{\alpha_U + \alpha_V}{\lambda^2} \quad (3)$$

$$\mathrm{Var}\,(Y) = \frac{\alpha_W + \alpha_V}{\lambda^2} \quad (4)$$

The covariance between X and Y is formulated as:

$$\mathrm{Cov}\,(X,Y) = \frac{\alpha_V}{\lambda^2} \quad (5)$$

Based on equation 5, the correlation coefficient in the Bivariate Gamma distribution can be obtained through the equation:

$$\rho = \frac{\alpha_V}{\sqrt{(\alpha_U + \alpha_V)(\alpha_W + \alpha_V)}} \quad (6)$$

The correlation value in equation 6 is always positive because the parameters α_U, α_v, and α_w are parameters with positive values [8].

Mean Difference Test

The t-test is a statistical method used to test whether a sample mean differs significantly from a predetermined population mean. One frequently used form of the one-sample t-test is the independent sample t-test. [9].

The t-test statistic is written as in equation 7 below:

$$t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)} \quad (7)$$

The test decision is made by comparing the calculated t-statistic value with the t-table value at a specific degree of freedom and a predetermined significance level. Additionally, the decision can be based on the $p\text{-}value$. If the $p\text{-}value$ is less than α or the calculated t-statistic value is greater than the t-table value, then the null hypothesis stating that the means of the two populations are not different is rejected [10].

The hypothesis for the one-sample mean difference test is formulated as follows:

1. Hypothesis Testing

$H_0 : \mu = \mu_0$ (There is no significant difference)

$H_1 : \mu \neq \mu_0$ (There is a significant difference)

2. Required Quantity, Significance Level $\alpha = 5\%$
3. Test Statistic

$$t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)}$$

4. Rejection Criteria

Reject H_0 if the t_{hit} value is greater than the t_{table} value or the $p\text{-}value$ is $< \alpha = 0.05$.

5. Conclusion

Predictive Mean Matching (PMM)

One approach used to handle missing data is to take a random sample from the observed data, based on the minimum distance between the missing and observed values. One multiple imputation method that uses this approach is Predictive Mean Matching (PMM). The PMM method consists of three stages [11]:

a. Imputation: Each missing value is filled with a value selected based on the similarity of predictions between the missing and observed data.
b. Analysis: Each imputed dataset is analyzed separately.
c. Combination: Combining the analysis results from multiple imputed datasets to produce more accurate estimates and take into account imputation uncertainty.

PMM is a method that is essentially the same as the regression method; however, the difference is that each missing value is imputed from the closest observed value in the model. The steps are as follows [12]:

1. Obtain the V value using the formula $\boldsymbol{V} = (\boldsymbol{X'X})^{-1}$
2. Estimate the regression parameter $\hat{\beta}$ using the least squares formula, namely $\hat{\beta} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y_{obs}}$.
3. Calculate the variance value $\hat{\sigma}^{2*}$ obtained through the following equation 8:

$$\hat{\sigma}^{2*} = \frac{(Y_{obs} - X\hat{\beta})'(Y_{obs} - X\hat{\beta})}{g} \quad (8)$$

where g is a random variable with a Chi-square distribution with $n - p$ degrees of freedom. The number of observations is n, while p represents the number of independent variables plus one.

4. Estimate the parameter $\hat{\beta}^*$ using the equation $\hat{\beta}^* = \hat{\beta} + \hat{\sigma}^{2*}L'_{ij}z$. where L'_{ij} is an upper triangular matrix resulting from the Cholesky decomposition $V = L_{ij}L'_{ij}$, and z is a $p \times 1$ vector whose elements are distributed normally.
5. Calculate the difference in predicted values for missing data, namely $\eta(i,j) = |X_{obs[i]}\widehat{\beta} - X_{miss[j]}\widehat{\beta}^*|$. With $i = 1, \dots, n$ and $j = 1, \dots, n_0$. X_{obs} predictor matrix with successfully observed data in Y, while X_{miss} is an $n_0 \times p$ predictor matrix with missing data in Y.
6. Collect the difference in predicted values for n_0 sets of missing data. Each set of n_0 contains d data.
7. Select d complete data sets that have the minimum distance between predicted values. The selection of d is based on the minimum result of $|X_{obs[i]}\widehat{\beta} - X_{miss[j]}\widehat{\beta}^*|$.
8. Randomly select one data from the d available data to replace the missing data.

The PMM method offers the advantage that missing data are not only estimated based on model predictions but also imputed using actual, observed values that are similar to the predicted values. This avoids imputation errors that can occur when using a single value to fill in missing data [11].

Random Forest Imputation

Random Forest Imputation is a non-parametric imputation method used to handle missing data using a decision tree-based ensemble learning approach. This method works by modeling each variable with missing values as a function of other variables available in the data, then predicting the missing values using a pre-built Random Forest model. One widely used algorithm in this approach is missForest, which performs the imputation process iteratively. The initial stage is to assign temporary values to the missing data. Next, for each variable containing missing values, a Random Forest model is built using the complete data as the response variable and the other variables as predictors. Missing values are then predicted and updated repeatedly until the process reaches convergence [13].

This method offers a high degree of flexibility because it can handle mixed data types, both continuous and categorical, and can capture non-linear relationships between variables without requiring specific parametric distribution assumptions [14]. The steps in the missForest algorithm are as follows:

1. Determine the order of variables based on the proportion of missing data they contain.
2. For each variable X_j containing missing values, training data is generated using complete observations with the other variables as predictors.
3. Construct a Random Forest model $\widehat{f_j}$ to predict the value of X_j. Missing values are then imputed using the model's prediction results.
4. This is done for all variables and repeated iteratively until the convergence criterion is met.

A practical advantage of this approach is its ability to provide internal imputation error estimates through the out-of-bag (OOB) mechanism in Random Forest. Through this mechanism, researchers can obtain Root Mean Squared Error (RMSE) estimates for continuous variables and Proportion of Falsely Classified (PFC) estimates for categorical variables without requiring additional validation data [15]. Various studies have shown that Random Forest Imputation outperforms simple parametric methods, especially when the relationship between variables is non-linear. Comparative studies on observational and simulation data show that the missForest method tends to produce smaller RMSE values and higher classification accuracy levels for various missingness mechanisms, such as Missing Completely at Random (MCAR) and Missing at Random (MAR), although this method requires relatively greater computational costs and execution time compared to the simple imputation method [16].

In the imputation process, a Random Forest constructs some decision trees using bootstrap sampling techniques, then generates predictions based on the aggregation of all trees. For categorical variables, the prediction result is obtained from the mode value, while for numeric variables, the average of the prediction results is used [17]. Mathematically, a forest consists of B regression or classification trees constructed from bootstrap samples $D_1, D_2, \ldots, D_B$. In the case of regression, the estimated imputation value y^ is obtained by averaging the predictions from all trees as follows [18]:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x) \tag{9}$$

Where $T_b(x)$ denotes the b-th decision tree for observation x.

Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a widely used measure of prediction error because it is easy to understand and interpret. MAPE describes the average absolute percentage difference between the actual value and the predicted value. The smaller the MAPE value, the better the model's predictive ability. The MAPE equation can be expressed as follows:

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right|}{n} \times 100 \tag{10}$$

Where y_i represents the actual value at the i-th observation, $\hat{y}_i$ is the predicted value, and n is the number of observations [19]. Although quite popular, MAPE has several weaknesses, particularly when the actual value is very small or close to zero. This condition can cause the MAPE value to be very large or even undefined, potentially introducing bias in the evaluation of model performance. Furthermore, MAPE tends to assign greater weight to errors in observations with small actual values [20].

Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is a common measure used to evaluate the level of prediction accuracy in statistical and machine learning models. RMSE measures the average prediction error based on the square root of the average of the squared differences between the actual and predicted values. The smaller the RMSE, the closer the model's predictions are considered to be to the actual data. This measure is also known to be more sensitive to large prediction errors and is therefore often used in predictive model evaluation [21]. The RMSE equation is written as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2} \tag{11}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n represents the number of observations. RMSE is widely applied in various analyses, such as regression, time series forecasting, and machine learning-based predictive modeling, because it is able to provide a consistent evaluation of model performance.

Correlation

Correlation analysis is a statistical method used to determine the strength and direction of a linear relationship between two variables. The measurement of this relationship is expressed through a correlation coefficient, which ranges from -1 to $+1$. A coefficient value approaching $+1$ indicates a very strong positive relationship, while a value approaching -1 indicates a very strong negative relationship. Meanwhile, a coefficient value approaching zero indicates a very weak or even non-existent linear relationship between the variables. The correlation coefficient equation can be expressed as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{12}$$

where x_i and y_i are the i-th observation values, $\bar{x}$ and $\bar{y}$ are the sample means of variables X and Y, while n represents the number of observations. The r value is used to describe the strength and direction of the linear relationship between the two variables [22].

3. RESULTS

Descriptive Statistics

This study used bivariate gamma data generated using R Studio software with the VGAM package. The resulting data consisted of 200 observations with two research variables.

Descriptive Statistics

The following is the bivariate gamma data generated using R:

Table 1. Normal Bivariate Data

V1	V2
3.554	9.447
1.818	5.895
⋮	⋮
2.903	11.821

Next, descriptive statistical analysis was performed on the generated data. The results of the descriptive statistics are shown in the following table.

Table 2. Descriptive Statistics of Data

Item	V1	V2
Minimum	0.219	2.522
Quartil 1	1.856	6.716
Median	3.085	9.166
Mean	4.110	9.925
Quartil 3	5.601	12.473
Maximum	23.575	32.948

Based on Table 2, it can be seen that variable V1 has a minimum value of 0.219, the first quartile is 1.856, the median is 3.085, the mean is 4.110, the third quartile is 5.601, and the maximum value is 23.575. Meanwhile, variable V2 has a minimum value of 2,522, a first quartile of 6,716, a median of 9,166, a mean of 9,925, a third quartile of 12,473, and a maximum value of 32,948. These results indicate that both variables have a positive data distribution with a fairly wide range of values.

Missing Value Data

Before imputation, the data were subjected to missing value imputation at 5%, 10%, 15%, and 20% levels. The goal was to evaluate the imputation method's ability to handle missing data at various levels of data loss. The following is the data after data loss.

Table 3. Missing Value Data

Item	5%		10%		15%		20%	
	V1	V2	V1	V2	V1	V2	V1	V2
Minimum	0.219	2.522	0.219	2.522	0.219	2.522	0.219	2.522
Quartil 1	1.806	6.610	1.730	6.683	1.759	6.601	1.650	6.505
Median	3.044	9.120	3.078	9.175	3.064	9.152	2.941	9.062
Mean	4.092	9.874	4.079	9.940	4.115	9.934	4.024	9.794
Quartil 3	5.601	12.473	5.524	12.491	5.735	12.396	5.477	12.508
Maximum	23.575	32.948	23.575	32.948	23.575	32.948	23.575	32.948
NA's	12	8	27	13	32	28	50	30

Based on Table 3, it can be seen that the greater the proportion of missing values, the greater the number of missing data for each variable. At a missing value level of 5%, the variable $V1$ had 12 missing values, while the variable $V2$ had 8 missing values. When the missing value proportion was increased to 10%, the number of missing values in the variable $V1$ increased to 27, while the variable $V2$ had 13 missing values. Furthermore, at a missing value level of 15%, the variable $V1$ had 32 missing values and a variable $V2$ had 28 missing values. At the highest proportion of missing values, at 20%, the number of missing data in the variable $V1$ reached 50 observations, and in a variable $V2$ 30 observations.

Predictive Mean Matching Imputation

Imputation was performed on missing data of 5%, 10%, 15%, and 20%. At this stage, imputation of 5% was performed, resulting in the following MAPE, RMSE, and Correlation results:

Table 4. MAPE, RMSE, and Correlation in Data

MAPE	RMSE	PMM CORRELATION 5%	INITIAL DATA CORRELATION
0.5893	3.3384	0.7104	0.7288161

The following is a test of the difference in the average correlation of the results of the 5% PMM imputation method:

1. Hypothesis Testing
   $H_0 : \mu = \mu_0$ (There is no significant difference)
   $H_1 : \mu \neq \mu_0$ (There is a significant difference)
2. Required Quantity, Significance Level $\alpha = 5\%,\ n = 200$
3. Test Statistic
$$t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)} = \frac{0.7104 - 0.7288}{\left(\frac{0.0111}{\sqrt{200}}\right)} = -10.882, p - value\ = 1.13e - 14$$
4. Rejection Criteria
   Reject H_0 if the t_{hit} value is greater than the t_{table} value or the $p\text{-}value$ is $< \alpha = 0.05$.
5. Conclusion
   Because the $p\text{-}value\ = 1.13e - 14 < \alpha = 0.05$, then H_0 is rejected, meaning that the average value between μ and μ_0 is significantly different.

Random Forest Imputations

Imputation was performed on missing data of 5%, 10%, 15%, and 20%. At this stage, 15% was imputed, resulting in the following MAPE, RMSE, and Correlation results:

Table 5. MAPE, RMSE, and Correlation in Data

MAPE	RMSE	RF CORRELATION 15%	INITIAL DATA CORRELATION
0.5929	3.8909	0.7381	0.7288161

The following is a test of the difference in the average correlation of the results of the 15% RF imputation method:

1. Hypothesis Testing
   $H_0 : \mu = \mu_0$ (There is no significant difference)
   $H_1 : \mu \neq \mu_0$ (There is a significant difference)
2. Required Quantity, Significance Level $\alpha = 5\%,\ n = 200$
3. Test Statistic
$$t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)} = \frac{0.7381 - 0.7288}{\left(\frac{0.0156}{\sqrt{200}}\right)} = 4.2033, p - value\ = 0.00011$$
4. Rejection Criteria
   Reject H_0 if the t_{hit} value is greater than the t_{table} value or the $p\text{-}value$ is $< \alpha = 0.05$.
5. Conclusion
   Because $p\text{-}value\ = 0.0001111 < \alpha = 0.05$, then H_0 is rejected, meaning that the average value between μ and μ_0 is significantly different.

Based on the Classification and Regression Tree Imputation method and the Random Forest Imputation method, the imputation results are as shown in the following table:

Table 6. Conclusions on the Data

No	Missing Value	Imputation Method	Description	Correlation 0.72881	P-Value	MAPE	RMSE
1	5%	PMM	Significantly different	0.7104	$1.1e - 14$	0.589	3.338
		RF	Significantly different	0.7096	$2.2e - 16$	0.568	3.185
2	10%	PMM	Significantly different	0.7240	0.0491	0.537	3.561
		RF	Not significantly different	0.7270	0.3387	0.525	3.370
3	15%	PMM	Not significantly different	0.7303	0.5207	0.584	3.870
		RF	Significantly different	0.7381	0.0001	0.593	3.891

4	20%	PMM	Significantly different	0.7638	$2.2e-16$	0.848	3.951
		RF	Significantly different	0.7788	$2.2e-16$	0.827	3.973

The following is a comparison graph of MAPE and RMSE values based on the percentage level of missing data:

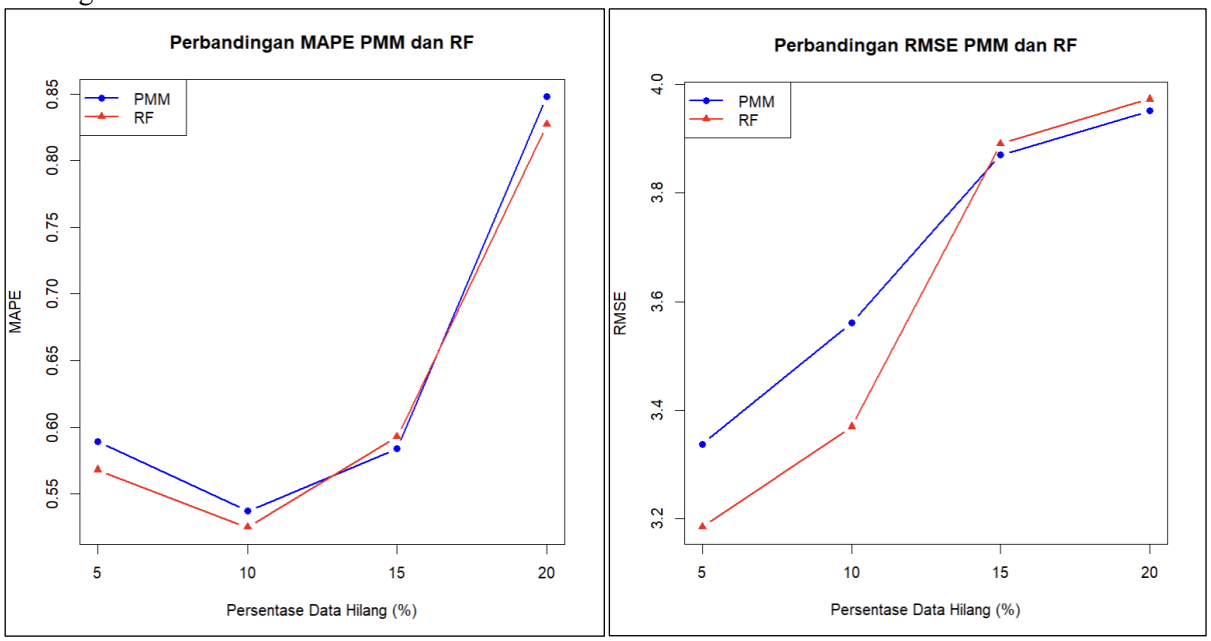

Figure 1. Comparison graph of MAPE and RMSE values

4. DISCUSSIONS

Based on Table 7, the Predictive Mean Matching (PMM) and Random Forest (RF) imputation methods exhibit different characteristics at each level of missing data. At 5% missing data, both methods yielded significant differences from the original data, based on p-values less than 0.05. However, the RF method yielded lower MAPE and RMSE values than PMM. These results indicate that at low levels of missing data, RF is able to impute data more effectively to produce more accurate predictions. RF's ability to analyze data and non-linearities between variables allows this method to produce estimates that are closer to the original data. At 10% missing data, PMM still showed significant differences from the original data, while RF showed no significant differences. These results indicate that RF maintains the characteristics of the original data stably as the proportion of missing data increases. The slightly higher correlation value of RF compared to PMM also indicates that the RF method better maintains relationships between variables. At 15% missing data, PMM showed no significant differences from the original data, while RF showed significant differences. This condition is explained by the fundamental nature of PMM, which performs imputation based on actual observed values. With this mechanism, PMM is able to maintain the original distribution of the data, including the characteristics of the bivariate Gamma distribution used in this study. Therefore, at moderate levels of missing data, PMM tends to be better at maintaining the distribution structure and produces smaller MAPE and RMSE values than RF. At 20% missing data, RF performed better, with the highest correlation value of 0.7788. These results indicate that as the proportion of missing data increases, RF's ability to learn data patterns through the construction of multiple decision trees (ensemble learning) becomes more effective than PMM. RF can utilize information from other variables to estimate missing values, thus maintaining relationships between variables even when the amount of available data decreases. Conversely, PMM becomes more limited as the proportion of missing data increases.

In general, the results show that PMM tends to excel at moderate levels of missing data, while RF excels at higher levels due to its flexibility in handling non-linear relationships and complex interactions between variables. In addition, the increase in MAPE and RMSE values in both methods as the proportion of missing data increases indicates that the greater the percentage of missing data, the greater the uncertainty and error of the imputation results produced.

This research presents a novel approach in the application and evaluation of two imputation methods, Predictive Mean Matching (PMM) and Random Forest (RF), on bivariate Gamma-distributed data. Studies on the performance of imputation methods on the bivariate Gamma distribution, which is characterized by an asymmetric (skewed) distribution and positive correlation between variables, are relatively limited. Furthermore, this study evaluates the ability of both methods to maintain correlation, statistical significance, and imputation error rates at various levels of missing data (5%, 10%, 15%, and 20%). This research contributes to providing empirical evidence regarding the appropriate imputation method for bivariate Gamma data under various missing data conditions.

To the author's knowledge, based on a literature search, no research has been found specifically comparing the performance of the PMM and Random Forest methods on simulated data following the bivariate Gamma distribution with varying levels of missing data. Therefore, this study is expected to fill the research gap related to imputation methods for non-normal data, particularly the bivariate Gamma distribution.

5. CONCLUSION

This study aims to compare the Predictive Mean Matching (PMM) and Random Forest (RF) imputation methods on bivariate Gamma-distributed data with varying levels of missing data: 5%, 10%, 15%, and 20%. The results show that both methods have distinct characteristics and advantages at each level of missing data. At 5% missing data, the RF method performed better than PMM, as indicated by lower MAPE and RMSE values. At 10% missing data, RF better maintained the characteristics of the original data, showing no significant differences from the original data. Meanwhile, at 15% missing data, the PMM method performed better because it maintained the original data distribution, resulting in insignificant differences from the original data and a relatively lower imputation error. At 20% missing data, the RF method again performed superiorly, producing the highest correlation value, demonstrating its ability to maintain relationships between variables even under conditions of greater missing data. In general, the research results indicate that the PMM method is more suitable for moderate levels of missing data due to its ability to maintain the original bivariate Gamma distribution. Conversely, the RF method is superior at higher levels of missing data due to its ability to capture complex and non-linear relationships between variables. Furthermore, the increase in MAPE and RMSE values for both methods as the proportion of missing data increases indicates that the greater the percentage of missing data, the greater the resulting imputation error.

This research contributes to the development of imputation methods for non-normal data, particularly the bivariate Gamma distribution. The results are expected to serve as a reference in selecting an appropriate imputation method according to the data characteristics and proportion of missing data encountered. Furthermore, this study also opens up opportunities for further research considering different missing data mechanisms, more diverse sample sizes, and the application of other imputation methods to the bivariate Gamma distribution.

6. ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the academic institutions that provided facilities and support for computing and data processing during this research. Appreciation is also extended to colleagues and fellow researchers for their valuable suggestions and constructive discussions regarding missing data imputation using Predictive Mean Matching and Random Forest methods. Furthermore, the authors are very grateful to the reviewers for their valuable comments and recommendations, which have significantly contributed to improving the quality of this manuscript.

7. REFERENCES

[1] S. Hong and H. S. Lynn, “Accuracy of random-forest-based imputation of missing data in the presence of interaction,” *J. BMC Med. Res. Methodol.*, vol. 1, pp. 1–12, 2020.

[2] B. O. Petrazzini, H. Naya, F. Lopez-Bello, G. Vazquez, and L. Spangenberg, “Evaluation of different approaches for missing data imputation on features associated with genomic data,”

BioData Min., pp. 1–13, 2021.
[3] M. Kokla, J. Virtanen, M. Kolehmainen, J. Paananen, and K. Hanhineva, "Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data : a comparative study," *J. BMC Med. Res. Methodol.*, pp. 1–11, 2019.
[4] Y. Ge, Z. Li, and J. Zhang, "Open a simulation study on missing data imputation for dichotomous variables using statistical and machine learning methods," *Sci. Rep.*, no. 0123456789, pp. 1–13, 2023.
[5] W. Agwil, D. Agustina, H. Fransiska, and I. A. Hasani, "Meningkatkan Kinerja Model Klasifikasi Curah Hujan Melalui Penanggulangan Missing Value Dengan Imputasi Berbasis Model," *Innov. J. Soc. Sci. Res.*, vol. 4, pp. 11773–11783, 2024.
[6] M. Franco and J. Vivo, "A Generator of Bivariate Distributions : Properties, Estimation, and Applications," *Math. Artic.*, vol. 8, no. 1776, pp. 1–30, 2020.
[7] C. Caamaño-Carrillo and J. E. Contreras-Reyes, "A Generalization of the Bivariate Gamma Distribution Based on Generalized Hypergeometric Functions," *Mathematics*, no. 3, pp. 1–17, 2022.
[8] C. K. Amponsah, T. J. Kozubowski, and A. K. Panorska, "A general stochastic model for bivariate episodes driven by a gamma sequence," *J. of Statistical Distrib. Appl.*, vol. 8, 2021.
[9] D. Curtis, "Welch's t-test is more sensitive to real-world violations of distributional assumptions than Student's t-test, but logistic regression is more robust than either," *Springer, Stat. Pap.*, pp. 3981–3989, 2024.
[10] H. Mustafidah, A. Imantoyo, and S. Suwarsito, "Pengembangan Aplikasi Uji-t Satu Sampel Berbasis Web," *JUITA J. Inform.*, vol. 8, no. 2, p. 245, 2020, doi: 10.30595/juita.v8i2.8786.
[11] G. Vink and L. E. Frank, "Predictive mean matching imputation of semicontinuous variables," *Stat. Neerl.*, vol. 68, no. 1, pp. 61–90, 2014.
[12] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York.John Wiley & Sons. 1989.
[13] M. A. Alwansyah and R. Rachmawati, "Handling Missing Data in Bivariate Gamma Generation Data Using the Random Forest Method," *J. Comput. Sci. Appl.*, vol. 8, no. 02, pp. 9–15, 2025.
[14] D. J. Stekhoven and P. Bühlmann, "MissForest — non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.
[15] D. J. Stekhoven, "Nonparametric Missing Value Imputation using Random Forests," *CRAN (manual). (terbaru; dokumentasi paket)*. 2025.
[16] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE : A CALIBER Study," *Am. J. Epidemiol.*, vol. 179, no. 7, pp. 764–774, 2014.
[17] R. Rachmawati, N. Afandi, and M. A. Alwansyah, "Survival Analysis on Data of Students Not Graduating on Time Using Weibull Regression, Cox Proportional Hazards Regression, and Random Survival Forest Methods," *Barekeng*, vol. 19, no. 3, pp. 2111–2126, 2025.
[18] M. A. Alwansyah, "Survival Analysis of Students Not Graduated on Time Using Cox Proportional Hazard Regression Method and Random Survival Forest Method," *J. Stat. Data Sci.*, vol. 2, no. 1, pp. 13–21, 2023.
[19] T. Iida, "Identifying causes of errors between two wave-related data using performance metrics," *Appl. Ocean Res.*, vol. 148, no. March, p. 104024, 2024.
[20] K. Warneke, S. D. Siegel, J. Afonso, and S. Wallot, "What the mean absolute percentage error (MAPE) should adopt from Bland – Altman analyses," *Ger. J. Exerc. Sport Res.*, 2025.
[21] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci. Model Dev.*, no. 2, pp. 5481–5487, 2022.
[22] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," *Journal Anesth.*, vol. 126, no. 5, pp. 1763–1768, 2018.