J-KOMA Journal of Computer Science and Applications

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Open Access

E-ISSN: *2620 - 4872*
Vol.*08*, ***No***.*02*
Doi: *10.21009/JKOMA.082.02*

Recivied: *13 Dec 2025*
Accepted: *18 Dec 2025*
Published: *26 Dec 2025*

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

Correspondence Email:
muhammadarib@unj.ac.id

Handling Missing Data in Bivariate Gamma Generation Data Using the Random Forest Method

Muhammad Arib Alwansyah[1]*, Ramya Rachmawati[2]

[1]Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta, Indonesia.
[2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Bengkulu, Indonesia.

Abstract

Missing data is a common problem in data analysis that can reduce the quality and accuracy of study results if not handled properly. This study aims to evaluate the performance of the Random Forest (RF) imputation method at various levels of missing value proportions, namely 5%, 10%, 15%, and 20%. The data used are Bivariate Gamma data of 200 observations with two variables, generated using RStudio software. Evaluation of imputation performance is carried out by considering the correlation value between the imputed data and the original data, the p-value as an indicator of the significance of the difference, and the error measures Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). The results show that the RF method provides the best results at a missing value level of 10%, characterized by a high correlation and a p-value > 0.05 indicating no significant difference with the original data, accompanied by the lowest MAPE and RMSE values. At missing value proportions of 5%, 15%, and 20%, the imputation results differed significantly from the original data, and the MAPE and RMSE values tended to increase as the proportion of missing data increased. These findings indicate that the RF method performs best at a data loss level not exceeding 10%, while its accuracy decreases at higher missing value proportions.

INTRODUCTION

Missing data, often referred to as missing data, is a crucial problem in research because it can degrade the quality of analysis, reduce estimation efficiency, and create potential bias in the inference process. The presence of incomplete data is often unavoidable, both in empirical data-based research and in simulation studies using generated data. In the context of positive continuous distributions, one model widely used to describe the relationship between two correlated variables is the Bivariate Gamma distribution. This distribution has wide applications in various fields, such as reliability engineering, risk analysis, hydrology, actuarial science, and epidemiology, due to its ability to represent the dependency between two skewed, positive-value random variables [1]. However, when some data in the Bivariate Gamma distribution is missing, further analysis processes such as parameter estimation, hypothesis testing, and dependency structure modeling are hampered and risk producing inaccurate conclusions.

Conventional imputation methods such as mean imputation and single regression imputation are still widely used, but both approaches tend to produce inaccurate estimates because they ignore the natural variability of the data and are unable to represent the structure of non-linear relationships between variables. With the development of machine learning techniques, various modern imputation methods have begun to be applied to obtain more accurate and stable estimates of missing values. One method that has proven effective is Random Forest Imputation (RF) [2].

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

Random Forest Imputation (RF) is an ensemble learning method that randomly constructs multiple decision trees and combines their predictions through an aggregation process. The bagging mechanism and random feature selection within each tree enable RF to capture complex patterns, non-linear relationships, and interactions between variables that traditional parametric methods cannot adequately handle [3]. Several studies have shown that RF Imputation is highly robust to outliers, capable of handling mixed data types, and exhibits superior performance on distributions with high levels of skewness [4].

The application of imputation using the RF method to Bivariate Gamma generation data is important because this method can overcome the limitations of the parametric approach in handling data asymmetry and complex interdimensional dependency structures. Furthermore, the use of RF allows researchers to evaluate the method's ability to produce missing value estimates that are close to the true distribution. Thus, this research is expected to contribute to the development of more effective imputation strategies for handling missing data problems [5].

METHODS

2.1 Bivariate Gamma Distribution

The Bivariate Gamma Distribution is an extension of the univariate Gamma distribution used to model two correlated and positively skewed continuous random variables. This distribution is widely used in insurance, reliability, survival analysis, and hydrology, when two response variables are related but have a right-skewed distribution [6]. For example, if (X, Y) is a pair of random variables that obey the Bivariate Gamma distribution, their dependency is constructed through a shared component model:

$$X = U + V, \qquad Y = W + V \tag{1}$$

Where $U, W, V \sim \text{Gamma}(\alpha, \lambda)$ are independent of each other. With this structure, the V component is the cause of the positive correlation between X and Y [7].

2.2 Characteristics of the Bivariate Gamma Distribution

Characteristics of the Bivariate Gamma Distribution include expected value, variance, covariance, and correlation. If the shared component structure is used, the expected value of the bivariate gamma is as follows:

$$E(X) = \frac{\alpha_U}{\lambda} + \frac{\alpha_V}{\lambda}, E(Y) = \frac{\alpha_W}{\lambda} + \frac{\alpha_V}{\lambda} \tag{2}$$

The variance of the bivariate gamma is as follows:

$$\text{Var}\,(X) = \frac{\alpha_U + \alpha_V}{\lambda^2}, \text{Var}\,(Y) = \frac{\alpha_W + \alpha_V}{\lambda^2} \tag{3}$$

The covariance and correlation of the bivariate gamma are as follows:

$$\text{Cov}\,(X, Y) = \frac{\alpha_V}{\lambda^2} \tag{4}$$

Thus, the correlation is as follows:

$$\rho = \frac{\alpha_V}{\sqrt{(\alpha_U + \alpha_V)(\alpha_W + \alpha_V)}} \tag{5}$$

The correlation is always positive because it depends on the shared gamma component (V) [8].

2.3 Difference in Means Test

The t-test is used to assess whether two population means are significantly different based on sample information. Specifically, the independent sample t-test is applied when the two sample groups are independent. This test compares the estimated means of the two groups, taking into account the level of variation in the data in each sample [9].

The following is the equation for the t-test statistic:

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

$$t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)} \quad (6)$$

The test is performed by comparing the t −statistic value to the t-table at a specific degree of freedom and a predetermined significance level. If the p-value is less than α or the t-statistic value is greater than the t-table, then the null hypothesis that the two population means are equal is rejected [10].

The following is the hypothesis from the one-sample mean difference test:

1. Hypothesis Testing
   $H_0 : \mu = \mu_0$ (Not significantly different)
   $H_1 : \mu \neq \mu_0$ (Significantly different)
2. Required Quantity
   Significance Level $\alpha = 5$
3. Test Statistic
   $t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)}$
4. Rejection Zone
   Reject H_0 if the t_{hit} value $> t_{tabel}$ or $p - value < \alpha = 0.05$
5. Conclusion

2.4 Random Forest Imputations (RF)

Random Forest Imputation is a non-parametric method used to fill in missing data by leveraging the ensemble capabilities of decision trees. This approach models each variable with missing values as a function of the other available variables, then generates predictions for those missing values using a pre-built random forest model. The missForest algorithm works iteratively, starting by assigning initial values to the missing data. Next, for each missing variable, a random forest model is trained using the complete observation as the response variable and the remaining variables as predictors. The missing values are then predicted and updated, and this process continues until convergence is reached. This method offers high flexibility because it can handle mixed data types (continuous and categorical) and can learn non-linear patterns between variables without requiring specific parametric distribution assumptions [11].

The steps of the missForest algorithm are as follows:

1. Determine the order of variables based on the proportion of missing values
2. For each variable X_j that is missing, construct a training dataset $\{(X_{-j}^{(i)}, X_j^{(i)}) : X_j^{(i)}$ "observed"$\}$ and fit a random forest model $\hat{f}_j$ o predict X_j from another variable X_{-j}.
3. Use $\hat{f}_j$ to fill in the missing values of X_j.
4. Repeat steps 2–3 for all variables, and repeat this cycle until the convergence criterion is met.

The practical advantage of this approach is the availability of estimates of the inherent imputation error through the out-of-bag (OOB) mechanism in random forests, allowing researchers to obtain estimates of the root mean squared error (RMSE) for continuous variables and the proportion of falsely classified (PFC) for categorical variables without a separate validation set [12]. Random Forest imputation has shown superior performance compared to other simple parametric methods under conditions where the relationship between variables is non-linear. Comparative studies on observational and simulated data have reported that missForest tends to produce lower RMSE and higher classification accuracy in various missingness scenarios (MCAR/MAR), although the computational cost and execution time are usually greater than simple methods [13]. In the imputation process, Random Forest builds some decision trees using bootstrap sampling, then produces predictions based on the aggregation of models, namely the mode for categorical variables and the average for numeric variables [14]. Mathematically, a forest consists of B

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

regression or classification trees built from bootstrap samples $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_B$. For regression prediction, the imputation of missing values ŷ is expressed as the average of the predictions across the tree [15]:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{7}$$

It is known that $T_b(x)$ is the b-th tree for observation x.

2.5 *Mean Absolute* Percentage *Error* (MAPE)

The Mean Absolute Percentage Error (MAPE) is a popular and easily interpreted measure of forecast error, expressing the average absolute percentage deviation between observed and predicted values. The MAPE equation is defined as follows:

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right|}{n} \times 100 \tag{8}$$

Where y_i is the actual value at the i-th observation, $\hat{y}_i$ is the predicted value, and n is the number of observations [16]. MAPE has several limitations, such as becoming undefined or very large if any actual value of y_i is close to zero, thus biasing the model's assessment of series with small or zero values. MAPE also places a relatively greater weight on errors that occur in small observations [17].

2.6 *Root Mean Square Error* (RMSE)

Root Mean Square Error (RMSE) is a measure of predictive accuracy used in evaluating the performance of statistical and machine learning models. A smaller RMSE value indicates that the model's predictions are closer to the actual values and also shows the variation in predicted values according to variations in the observed data. RMSE is often used because it is more robust to large prediction errors [18]. The RMSE equation can be written as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2} \tag{9}$$

Where $\mathrm{y_i}$ is the actual value, $\hat{\mathrm{y}}_\mathrm{i}$ is the predicted value, and n is the number of observations. RMSE allows for consistent evaluation of model performance, particularly in the context of regression, time series forecasting, and machine learning-based predictive modeling.

2.7 Correlation

Correlation analysis is a statistical method used to measure the strength and direction of a linear relationship between two variables. The correlation coefficient provides a numerical value between -1 and +1. Values close to 1 indicate a very strong positive relationship, close to –1 indicate a very strong negative relationship, and values close to 0 indicate no or very weak linear relationship. The correlation formula can be written as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{10}$$

It is known that x_i and y_i are the values of the i-th observation, $\bar{x}$ and $\bar{y}$ are the sample averages of X dan Y, and n is the number of observations. The r value indicates the degree of strength and direction of the linear association between X and Y [19].

RESULTS AND DISCUSSION

3.1 Descriptive Statistics

This study uses bivariate gamma data generated in R Studio using 200 VGAM datasets with two variables.

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

3.2 Descriptive Statistics

The following is the bivariate gamma data generated in R:

TABLE 1. Bivariate Gamma Data

V1	V2
3.554	9.447
1.818	5.895
⋮	⋮
2.903	11.821

The following are descriptive statistics on the bivariate gamma data that has been generated:

TABLE 2. Descriptive Statistics of Data

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

Based on Table 2, it is known that for variable V1 the minimum value is 0.219, the 1st quartile value is 1.856, the median value is 3.085, the average value is 4.110, the 3rd quartile value is 5.601 and the maximum value is 23.575, while for variable V2 the minimum value is 2.522, the 1st quartile value is 6.716, the median value is 9.166, the average value is 9.925, the 3rd quartile value is 12.473 and the maximum value is 32.948.

3.3 *Missing Value* Data

Before imputation, missing values of 5%, 10%, 15%, and 20% will be added to the data. The following is the data after the missing values were added:

TABLE 3. Data *Missing Value*

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
Maksimum	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 at a missing value level of 5%, variable V1 has 12 missing values and variable V2 has 8 missing values. At a missing value proportion of 10%, the number of missing values in variable V1 increases to 27, while in variable V2 there are 13 missing values. Furthermore, at a missing value level of 15%, variable V1 has 32 missing values and variable V2 has 28 missing values. At the largest missing value proportion of 20%, the number of missing values in variable V1 reaches 50, while variable V2 has 30 missing values.

3.4 *Random Forest Imputations*

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:

Universitas Negeri Jakarta
Phone: (021) 4890856
East Jakarta City
Jakarta, 13220

TABLE 5. MAPE, RMSE and Correlation in Data

MAPE	RMSE	Correlation
0.567461	3.185404	0.7096

The following is a test of the difference in mean correlation in the data:

1. Hypothesis Testing
   $H_0 : \mu = \mu_0$ (Not significantly different)
   $H_1 : \mu \neq \mu_0$ (Significantly different)
2. Required value
   Significance level $\alpha = 5\%$, $n = 200$
3. Test statistic
   $$t = \frac{\bar{X} - \mu}{\left(\frac{SD}{\sqrt{n}}\right)} = \frac{0.7096 - 0.7288161}{\left(\frac{0.01107}{\sqrt{200}}\right)} = -12.202, p - value = 2.2e - 16$$
4. Rejection zone
   Reject H_0 jika if $t_{hit} > t_{tabel}$ or $p - value < \alpha = 0.05$
5. Conclusion
   Since the $p - value = 2.2e - 16 < \alpha = 0.05$, maka H_0 is rejected, meaning the average values between μ and μ_0 are significantly different.

Based on the Classification and Regression Trees Imputation method and using 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	Information	Correlation 0.72881	p-value	MAPE	RMSE
1	5%	RF	Significantly different	0.7096	$2.2e-16$	0.568	3.185
2	10%	RF	Not significantly different	0.7270	0.3387	0.525	3.370
3	15%	RF	Significantly different	0.7381	0.0001	0.593	3.891
4	20%	RF	Significantly different	0.7788	$2.2e-16$	0.827	3.973

The results show that the performance of the Random Forest (RF) imputation method varies according to the proportion of missing values. At a missing data proportion of 5%, the RF method produces a fairly good correlation, although a $p\text{-}value < 0.05$ indicates a significant difference from the original data. A different condition is seen at a missing value proportion of 10%, where RF provides the most optimal imputation results, indicated by the highest correlation value, accompanied by a $p\text{-}value > 0.05$ and the lowest MAPE and RMSE values. This indicates that at moderate levels of data loss, RF is still able to maintain data structure and produce imputations that are not significantly different from the original data. However, at the missing value proportions of 15% and 20%, Despite the increasing correlation value, the imputation results again show significant differences from the original data and are followed by an increase in MAPE and RMSE values. These findings indicate that the greater the proportion of missing data, the lower the accuracy of RF imputation, so its performance is most effective when the data loss level does not exceed 10%.

CONCLUSION

Based on the evaluation results of the Random Forest (RF) imputation method performance at various levels of missing value proportions, namely 5%, 10%, 15%, and 20%, It can be concluded that imputation performance is greatly influenced by the magnitude of the level of data loss. At the missing value proportions of 5%, 15%, and 20%, the RF method produces $p\text{-}values$ smaller than α = 0.05, which indicates that the imputation results differ significantly from the original data. However, the correlation values obtained are still in a relatively high range, namely between 0.7096 and 0.7788, thus indicating that RF is still able to maintain the pattern of relationships between variables. Specifically, a missing value proportion of 10%

provides the most optimal imputation results, indicated by a correlation value of 0.7270 with a $p\text{-}value > 0.05$ so that there is no significant difference with the original data, as well as the lowest MAPE and RMSE values compared to other proportions. These findings indicate that the RF method can be used as an effective imputation approach, especially at data loss levels not exceeding 10%. However, increasing the proportion of missing values tends to reduce imputation accuracy, as evidenced by the increase in MAPE and RMSE values at higher levels of data loss.

REFERENCES

[1] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence 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 to 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. ofStatistical 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] 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.
[12] D. J. Stekhoven, "Nonparametric Missing Value Imputation using Random Forests," *CRAN (manual). (terbaru; dokumentasi paket).*, 2025.
[13] 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.
[14] 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.
[15] 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.
[16] T. Iida, "Identifying causes of errors between two wave-related data using performance metrics," *Appl. Ocean Res.*, vol. 148, no. March, p. 104024, 2024.
[17] 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.
[18] 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.
[19] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," *aournal Anesth.*, vol. 126, no. 5, pp. 1763–1768, 2018.