ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 16, Nomor 1, Tahun 2026, Halaman 200 - 211
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

BURR XII REGRESSION AND OPTIMIZATION FOR DIARRHEA INCIDENCE ANALYSIS IN SURAKARTA CITY

Rizwan Arisandi[1*], Adhe Lingga Dewi[1], Mohammad Fajri[2]
[1] Departement of Computer Science, Faculty of Informatics Engineering, Bina Nusantara University, Semarang – Indonesia
[2]Program Studi Statistika, FMIPA, Universitas Tadulako, Palu - Indonesia
***Email: rizwan.arisandi@binus.ac.id**

DOI: 10.14710/j.gauss.16.1.200-211

Article Info:
Received: 2026-01-25
Accepted: 2026-05-29
Available Online: 2026-06-02
Keywords:
Regression Analysis; Burr XII Distribution; Diarrhea Incidence; BFGS Algorithm.

Abstract: Diarrhea remains a major public health concern in Indonesia, with Surakarta recording 11,434 cases in 2024, an increase of 58% from the previous year. This study applies the Burr XII regression model to identify environmental factors influencing diarrhea incidence and accommodate heavy-tailed variability across urban villages. The Burr XII model was chosen for its robustness in handling skewed data and extreme incidence patterns commonly found in epidemiological studies. Parameter estimation was carried out through the Maximum Likelihood Estimation (MLE) approach, with the BFGS algorithm as the primary optimization method, compared against Genetic Algorithm (GA) and Simulated Annealing (SANN). This research is important due to the increasing trend of diarrhea cases and the need for accurate statistical models to support public health policies. The results indicate that the BFGS method achieved the best fit, with five significant predictors: distance to the nearest hospital, rainfall, distance to waste disposal site, elevation, and distance to the nearest river. Population density and slope were not statistically significant but retained for theoretical relevance. These findings emphasize the importance of environmental and infrastructural factors in diarrhea prevention and support spatially targeted public health interventions.

1. INTRODUCTION

Diarrhea remains a major public health concern in Indonesia. The Ministry of Health defines diarrhea as a condition characterized by an increased frequency of defecation with soft to liquid feces, often followed by indicators such as queasiness, throwing up, stomach cramping, and in some cases, weight loss (KEMENKES RI, 2023). Although it is considered a common illness, diarrhea can become life-threatening, particularly among children and the elderly, if not treated promptly and appropriately.

According to data from Statistics Indonesia, diarrhea was the most prevalent disease in Surakarta City in 2024, with 11,434 reported cases. This marked an increase of approximately 58 percent compared to the previous year, which had recorded 7,209 cases (BPS Surakarta, 2025). This sharp rise highlights the urgency of the issue despite various preventive interventions introduced by the local health department, such as promoting Clean and Healthy Living Behavior (PHBS). These interventions include hand washing with soap, consuming nutritious food, using clean water, maintaining sanitary toilets, and performing daily physical activities (Puskesmas Poncowarno, 2022). Nevertheless, the increasing number of cases suggests that existing efforts have not yet effectively reduced the incidence of diarrhea.

The limitations of statistical modeling in understanding the spread and risk factors of diarrhea are one of the main challenges. Diarrhea cases have been studied several times using

classical statistical approaches. One of them is Wibowo et al. (2021), which examines how environmental and household factors influence diarrhea that occurs in children in Indonesia using binary logistic regression. In addition, Syafitri et al. (2023) also analyzed the survival data of diarrhea patients using Cox extended regression. It can be said that there are still few studies that use more flexible distribution exploration and are able to overcome the problem of non-normal and heavy-tailed data, which these characteristics are often found in health datasets, one of which is diarrheal disease.

The presence of skewness and heavy tails can lead to biased estimates, underestimation of variability, and decreased prediction accuracy if the regression model relies solely on normal-based assumptions (Yasin et al., 2022). Burr XII existence is present to provide a good alternative because it is able to capture the characteristics of the data even though it is skewed and heavy-tailed (de Araújo et al., 2022). Burr XII distribution is a distribution that is often used in various fields, such as survival analysis using two shape parameters (Widiastuti et al., 2023). Because of its advantages, this distribution is suitable for modeling health data such as diarrhea cases that have a non-normal distribution.

Alternative distributions that are more flexible and capable of modeling extreme values should be used to overcome these weaknesses. A concern besides distribution selection is the optimization technique. If robust optimization is not used, there will be a failure to converge when maximum likelihood estimation is used. Widely, the BFGS algorithm is recommended because it can combine fast convergence with numerical stability especially for nonlinear problems (Yang, 2024). Parameter estimation using this optimization is very efficient because it uses a gradient approximation and a hessian approximation so it is very suitable for Burr XII regression (Jin et al., 2024).

Therefore, the purpose of this study is to develop a regression model by utilizing the Burr XII distribution to analyze environmental and infrastructure factors that affect the number of diarrhea cases in Surakarta City. In addition, this study also looks at how to optimize parameter estimation using the BFGS algorithm and make a comparison of the results with other optimization methods such as Simulated Annealing (SANN) and Genetic Algorithm (GA). The result of the study are expected to provide more accurate modeling and provide evidence-based knowledge to support the formulation of policies and interventions for targeted communities.

2. LITERATURE REVIEW

As a developing country, health issues such as diarrhea are still a difficult challenge in Indonesia. This health problem will cause a high mortality rate if there is no good treatment diarrhea that has reached an acute level can even cause a person to lose fluids, experience severe dehydration and other health problems (Purnama et al., 2025; Suparmi et al., 2025). Infectious symptoms in diarrhea are caused by various bacteria, viruses or parasites that are usually contracted through contaminated food or drinking water or from between individuals due to poor hygiene (Winarni, 2021; WHO, 2024).

In new research, it is explained that diarrhea is not only caused by these factors but also influenced by spatial and environmental factors. In a study conducted by Faidah et al. (2023), factors such as the percentage of families implementing Clean and Healthy Living Behavior (PHBS), sanitation, poverty level, population density, exclusive breastfeeding rate for infants under six months old, complete immunization, and access to health services have a significant influence on the prevalence of diarrhea in children under five in West Java.

Another study also conducted by Faidah et al. (2023) provides knowledge that the prevalence of diarrhea in children under five, clean and healthy living behavior, access to healthy latrines, population density, and the proportion of infants who get exclusive breastfeeding can be used in grouping sub-districts in the city of Bandung based on diarrhea

cases. This knowledge provides insight that the complexity of diarrhea transmission requires a model that is able to combine various environmental and infrastructure factors and individual behavior to better understand disease patterns. Therefore, identifying the relationship among these variables is an important preliminary step before constructing a regression model to ensure that the predictors contribute effectively without causing redundancy in the analysis. Evaluation of the strength and direction of the linear relationship between predictor variables is done before entering into the regression model using correlation analysis. When two or more predictor variables have overlapping information or can be said to have a high correlation, this can cause multicollinearity (Fernandes & Solimun, 2016, as cited in Ramadhan et al., 2024).To check for multicollinearity, the Variance Inflation Factor (VIF) value can be used (Zaki et al., 2023). The VIF is formulated as:

$$VIF_k = \frac{1}{1-R_k^2}, \text{ k = 1,2,3,…} \quad (1)$$

Where R_k^2 denotes the coefficient of determination derived from regressing the k-th predictor againts all remaining predictors in the model. A larger R_k^2 produce a higher VIF value, indicating stronger multicollinearity. A VIF values greater than 10 is generally regarded as an issue. Some predictors may exhibit a moderate level of association as a result of common geographic or climatic conditions, particularly in environmental and spatial datasets. To ensure that no strong correlations exist, both correlation coefficients and VIF values were examined. Pearson correlation was employed to assess relationships between each pair of predictors, while VIF values were used to evaluate multicollinearity (Shrestha, 2020).

The Burr Type XII distribution is widely regarded as one of the most versatile continous probability distributions. The distribution under consideration is constructed from three positive parameters, namely β as the scale parameter, and λ and τ as the shape parameters. Public health datasets such as diarrhea disease incidence data, which generally have skewed and heavy-tailed properties, can apply this distribution.

The pdf (probability density function) of the Burr XII model for a stochastic variable $y > 0$ is given by:

$$f_y(y) = \frac{\lambda\beta^{\lambda}\tau y^{\tau-1}}{(\beta+y^{\tau})^{\lambda+1}} \quad \text{for } y > 0 \quad (2)$$

The cdf (cumulative distribution function) can be expressed as:

$$F_y(y) = 1 - \left(\frac{\beta}{\beta+y^{\tau}}\right)^{\lambda} \quad \text{for } y > 0 \quad (3)$$

The shape of the distribution allows the Burr XII distribution to be applied to cases such as the log-logistic and Pareto distributions (Santos & Pescim, 2023). The ability of the Burr XII distribution to cope with heavy-tailed data is very suitable for modeling data with extreme values. In the context of regression, to ensure positive parameter values, an exponential function is used to associate the scale parameter β with the predictor variable. In previous research, Hakim et al. (2021) applied the Burr XII distribution in modeling health data that has positive skewness and heavy tails through regression models. Therefore, understanding the parameterization of the Burr XII distribution becomes essential to facilitate model interpretation and estimation in regression analysis. Beirlant (1998, as cited in Low et al., 2021) posits that the Burr XII distribution is characterized by the utilization of shape parameters β and λ, in conjunction with the scale parameter τ_i. In this study, the notation is rewritten by replacing β with c and λ with d, while τ_i is retained as the scale parameter for each individual. This reparameterization does not alter the primary properties of the Burr Type XII model; however, it facilitates the interpretation of the regression coefficients.

In accordance with the model assumptions, the response variable y_i with predictor x_i adheres to a Burr Type XII distribution, characterized by shape parameter c and d, and an individual scale parameter τ_i, which is expressed as follows:

$$y_i|x_i \sim Burr(c, d, \tau_i) \quad (4)$$

The parameter τ_i, which indicates the scale of each individual, is formulated using the log-link function:

$$\tau_i = \exp(x_i'\theta) \quad (5)$$

where $x_i = [x_{i1}, x_{i2}, \ldots, x_{ik}]$ represents the vector of explanatory variables for the i-th observation, and $\theta = [b_0, b_1, \ldots, b_k]$ denotes the regression parameters awaiting estimation. The pdf (probability density function) of three-parameter Burr Type XII XII model is expressed as:

$$f_Y(y) = \frac{c\, d\, \tau_i y_i^{c-1}}{(\tau_i{}^c + y_i{}^\tau)^{d+1}} \quad \text{for } y > 0 \quad (6)$$

The likelihood function for data with sample size n is given by:

$$L(\Omega) = \prod_{i=1}^{n} f_Y(y_i) \quad (7)$$

Applying the natural log to the likelihood results in the log-likelihood function:

$$\ell(\Omega) = \sum_{i=1}^{n} \log f_Y(y_i) \quad (8)$$

By substituting the pdf from equation (6) and (7) and applying the reparameterization $\tau_i = \exp(b_0 + b_1 x_{1i} + \cdots + b_7 x_{7i})$, the likelihood becomes:

$$L(\Omega) = \prod_{i=1}^{n} \left[cd\tau_i^{-1} \left(\frac{y_i}{\tau_i}\right)^{c-1} \left[1 + \left(\frac{y_i}{\tau_i}\right)^{c}\right]^{-(d+1)} \right] \quad (9)$$

where $\Omega = (c, d, b_0, b_1, \ldots, b_7)$.

The corresponding log-likelihood function is:

$$\log L(\Omega) = \sum_{i=1}^{n} \left[\log c + \log d - \log \tau_i + (c-1)\, log\left(\frac{y_i}{\tau_i}\right) - (d+1) \log\left(1 + \left(\frac{y_i}{\tau_i}\right)^{c}\right)\right] \quad (10)$$

Due to the nested nonlinear form of τ_i, the log-likelihood function cannot be maximized analytically and therefore requires numerical optimization techniques (Al-Shomrani, 2022). In this study, the BFGS algorithm is used as the main optimization method due to its fast convergence and stability in medium-scale problems. For comparison, two global optimization approaches are also evaluated. The first is the Genetic Algorithm (GA), an evolutionary search method inspired by natural selection (Chandan et al., 2022), which operates by developing a set of solutions over generations via selection, recombination, and mutation, allowing it to be efficient in avoiding local optima. The second is Simulated Annealing (SANN), a probabilistic technique that can accepts less optimal solutions in the early stage of the search to avoid local optima, progressively decreasing this probability through iterations (Ma & Yang, 2025).

3. MATERIAL AND METHOD

This study makes use of secondary data on diarrhea cases in Surakarta City in 2024, obtained from the Surakarta City Health Office, Central Statistics Agency (BPS) of Surakarta City, and several publicly accessible geospatial data sources, including Open Street Map, the Open Topography portal, and CHIRPS monthly precipitation data provided by the Climate Hazard Center, University of California. The dataset contains the number of diarrhea cases per sub-district (kelurahan) along with various spatial, environmental, and infrastructural variables relevant to diarrhea cases.

In this research, the variables consist of dependent and independent variables. The description of the research variables is presented in Table 1.

Table 1. Research Variables

Variable	Definition	Unit
Diarrhea cases (Y)	Total number or diarrhea cases recorded in each urban village	cases
Population density (X_1)	Number of people per square kilometer in each urban village	people/km^2
Distance to the nearest hospital (X_2)	Average distance from residential areas to the nearest hospital	kilometers (km)
Rainfall(X_3)	Average annual rainfall	mm/ year
Distance to waste disposal site(X_4)	Average distance to the nearest waste disposal site	kilometers (km)
Elevation (X_5)	Average land surface height above sea level	meters (m)
Slope (X_6)	Average land surface steepness	percent (%)
Distance to nearest river (X_7)	Average distance to the closest river	kilometers (km)

The analysis steps of modeling using Burr XII regression are as follows:

1. Variables in the dataset that exhibit high multicollinearity will be identified and considered for removal. Multicolinearity testing in this study uses Pearson Correlation and Variance Inflation Factor (VIF) with a high collinearity threshold of 9.0 for correlation and VIF > 10. Meanwhile, the Variance Inflation Factor (VIF) is calculated using the following formula, see equation 1
2. The dataset is then prepared by renaming columns into concise formats and checking for missing values and correct data types
3. Standardization is performed on predictor variables using the Z-score method so that the data have a mean value of 0 and a standard deviation of 1. The Z-score standardization formula is given by:
$$Z_i = \frac{X_i - \mu}{\sigma} \tag{11}$$
4. The Burr XII regression model is specified with shape parameters c and d and regression coefficients b_0 to b_7, and likelihood functions are derived accordingly. see equation 6
5. Parameter estimation is conducted using Maximum Likelihood Estimation (MLE) with the BFGS algorithm as the main optimization method, while Simulated Annealing (SANN) and Genetic Algorithm (GA) are used as alternative global optimization methods
   The iterative parameter update is expressed as:
$$\theta_{k+1} = \theta_k - \alpha_k H_k \nabla f(\theta_k) \tag{12}$$
   with the inverse Hessian approximation matrix:
$$H_k = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1p} \\ h_{21} & h_{22} & \cdots & h_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{p1} & h_{p2} & \cdots & h_{pp} \end{bmatrix} \tag{13}$$
   where $p$is the number of estimated parameters.
   The BFGS matrix update formula is written as:
$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T \tag{14}$$
   with the vector forms:

$$s_k = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_p \end{bmatrix} = \theta_{k+1} - \theta_k \quad (15)$$

and:

$$\rho_k = \frac{1}{y_k^T s_k} \quad (16)$$

6. The performance of the model is assessed by comparing the Akaike Information Criterion (AIC) values obtained from each optimization method. he AIC is calculated as follows:

$$AIC = 2k - 2\ln(\hat{L}) \quad (17)$$

7. Convergence of algorithms and stability of parameter estimates are checked to validate the model. Convergence is achieved when the difference between parameter estimates in consecutive iterations becomes sufficiently small, which can be expressed as:

$$\| \theta_{k+1} - \theta_k \| < \varepsilon \quad (18)$$

4. RESULTS AND DISCUSSION

Table 2 summarizes the statistical characteristics of all variables incorporated in the modeling framework. These descriptive measures provide an initial understanding of the central tendency and dispersion of each variable, which is crucial for interpreting the dataset prior to applying the Burr XII regression model.

Table 2. Summary Statistics of Variables

Variable	Description	Mean	Standard deviation	Min	Max
Y	Diarrhea cases	133.43	159.49	3.00	887.00
X_1	Population density (people/km^2)	13,945.43	4,883.15	7,343.93	27,730.43
X_2	Distance to the nearest hospital (km)	0.50	0.26	0.15	1.26
X_3	Rainfall (mm/ year)	166.92	0.97	165.38	169.39
X_4	Distance to waste disposal site (km)	1.85	1.22	0.23	5.49
X_5	Elevation (m)	96.83	5.39	90.14	114.97
X_6	Slope (%)	89.13	0.60	87.24	89.98
X_7	Distance to nearest river (km)	0.00023	0.00025	0.00001	0.00089

Across the 54 urban villages studied, the mean number of diarrhea cases (Y) was 133.43, with a notably high standard deviation of 159.49, reflecting substantial variation in disease prevalence among locations. The minimum recorded value was just 3 cases, while the maximum reached 887 cases in a single village, highlighting significant disparities in health outcomes. Population density (X_1) demonstrated a broad distribution, from 7,344 to 27,730 people/km^2, averaging 13,945.43 people/km^2. This suggests marked differences in settlement concentration across the study area. The average distance to the nearest hospital (X_2) was 0.50 km, with a range between 0.15 km and 1.26 km, indicating uneven access to healthcare facilities.

Rainfall (X_3) was relatively homogeneous, with an average value of 166.92 mm/year and a narrow variation of 0.97 mm, suggesting similar climatic conditions throughout the area. Elevation (X_5) spanned from 90.14 m to 114.97 m, averaging 96.83 m above sea level, while slope (X_6) showed minimal fluctuation around a mean of 89.13%, consistent with a generally flat terrain profile. The average distance to the nearest waste disposal site (X_4) was 1.85 km, with extremes ranging from 0.23 km to 5.49 km, potentially affecting environmental health risk exposure. Finally, distance to the nearest river (X_7) exhibited extremely small values, between 0.00001 km and 0.00089 km (approximately 0.01–0.89 meters), indicating that nearly all villages are situated in immediate proximity to river systems, with negligible variation across the study area. From descriptive statistics, it can be seen that the independent variables have different units and scales. Therefore, data preparation and standardization will be carried out before presenting the model estimates.

Data preparation is done to ensure the dataset is clean and ready for analysis before model estimation. The result show that all variables have no missing values and are numeric. Variable naming is also done to facilitate the modeling process, for example the dependent variable "Jumlah Diare" is named Y, while the independent variables are labeled X_1 to X_7. This step ensures that the dataset to be analyzed is consistent, clear and ready to be used for the modeling stage.

Standardization is performed before estimation because the independent variables have different units and scales. Each independent variable is transformed using the Z-score method with a mean value of zero and a standard deviation of one. This step is used to prevent scale differences that can affect the magnitude of the regression coefficients and ensure that the comparison between independent variables is equivalent. In addition, numerical stability and a smooth convergence process of the optimization algorithm for parameter estimation are also obtained if standardization is performed.

Table 3 shows the Pearson correlation matrix of all predictor variables used in the model. Correlation analysis showed a moderate positive relationship between elevation (X_5) and rainfall (X_3) of 0.639, as well as between elevation (X_5) and distance to the nearest waste disposal site (X_4) of 0.606, patterns commonly observed in higher-altitude areas that tend to receive more rainfall and are located farther from disposal facilities. All other correlations were below 0.6, suggesting a generally low degree of linear association among predictors.

Table 3. Pearson Correlation Matrix of Predictor Variables

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1	0.312	-0.162	-0.181	-0.462	-0.245	-0.148
X_2	0.312	1	-0.055	0.059	-0.115	0.001	0.127
X_3	-0.162	-0.055	1	0.463	0.639	0.1	0.439
X_4	-0.181	0.059	0.463	1	0.606	-0.382	0.386
X_5	-0.462	-0.115	0.639	0.606	1	0.072	0.600
X_6	-0.245	0.001	0.100	-0.382	0.072	1	0.038
X_7	-0.148	0.127	0.439	0.386	0.600	0.038	1

To further assess multicollinearity, the Variance Inflation Factor (VIF) was calculated for each predictor, as shown in Table 4. All VIF values ranged from 1.23 to 3.52, far under the commonly used limit of 10, indicating that multicollinearity does not pose a serious concern in this dataset. The highest VIF was found for elevation (X_5) at 3.52, which consistent with its relatively high correlations with rainfall (X_3), distance to waste disposal site (X_4), and distance to nearest river (X_7).

Overall, both the correlation and VIF results suggest that while some predictors are moderately correlated, the level of multicollinearity is acceptable for regression modeling.

Then it can be said that the estimated coefficients remain stable and can be interpreted in the analysis of Burr XII regression

Table 4. Variance Inflation Factor (VIF) Values

Variable	VIF
X_1	1.594
X_2	1.228
X_3	1.878
X_4	2.428
X_5	3.516
X_6	1.597
X_7	1.649

Parameter estimation is done using the BFGS algorithm with the Maximum Likelihood Estimation (MLE) approach. The algorithm was chosen because it has stable and efficient calculations for the analysis of the Burr XII regression model on the dataset used in the study. For comparative purposes, Genetic Algorithm (GA) and Simulated Annealing (SANN) were also tested. The GA method was used primarily to explore the parameter space and identify suitable starting values, while SANN was applied as an alternative stochastic optimization technique.

Table 5 presents the parameter estimates obtained from the BFGS optimization. Out of the seven covariates, five showed statistical significance at the 5% level: X_2(distance to the nearest hospital), X_4(distance to waste disposal site), and X_7(distance to the nearest river). Positive coefficients for X_2 and X_7 suggest that increases in these variables are associated with higher expected diarrhea incidence, whereas the negative coefficient for X_4 implies that greater distance from waste disposal sites is linked to reduced disease risk. Population density (X_1), rainfall (X_3), elevation (X_5), and slope (X_6) were not statistically significant but retained for theoretical relevance.

Table 5. Estimated Parameters of Burr XII Regression Model

Parameter	Estimate	Standard deviation	z-value	p-value	Significance
c	1.3828	0.3243	4.2640	< 0.001	Significant
d	5.1363	11.0214	0.4660	0.6412	Not Significant
b_0	5.8303	2.0191	2.8876	0.0039	Significant
b_1	0.0976	0.1544	0.6321	0.5273	Not Significant
b_2	0.3027	0.1340	2.2588	0.0239	Significant
b_3	0.1293	0.1759	0.7348	0.4625	Not Significant
b_4	-0.4859	0.2098	-2.3154	0.0206	Significant
b_5	0.4007	0.2453	1.6336	0.1023	Not Significant
b_6	-0.3247	0.1851	-1.7547	0.0793	Not Significant
b_7	0.3749	0.1548	2.4224	0.0154	Significant

Among the tested methods, the BFGS method produced the highest log-likelihood value (−305.038), indicating a better model fit relative to SANN (−305.252) and the GA-based solution (−305.703). Furthermore, the BFGS method achieved the lowest Akaike Information Criterion (AIC) value (630.077), compared to GA (630.505) and SANN (631.406), further incoming its superior model fit. These result are summarized in Table 6

Table 6. Comparison of log-likelihood and AIC values

Optimization Method	Log-likelihood	AIC
BFGS	-305.038	630.077
Genetic Algorithm	-305.252	630.505
Simulated Annealing	-305.703	631.406

The final Burr XII regression model for the count of diarrhea incidence in Surakarta is given by:

$$\hat{y} \sim Burr\ XII(c = 1.3828, d = 5.1363,\ \tau_i = \exp(5.8303 + 0.0976X_1 + 0.3027X_2 + 0.1293X_3 - 0.4859X_4 + 0.4007X_5 - 0.3247X_6 + 0.3749X_7) \quad (19)$$

Most of the predictors show statistically significant relationships with the incidence of diarrhea, except for population density (X_1), rainfall (X_3), elevation (X_5), and slope (X_6), which were not significant at the 5% level. Despite this, these variables were retained in the model due to their theoretical importance and to help control for potential confounding effects.

The significance of distance to the nearest hospital (X_2) likely reflects how access to healthcare influences disease incidence; areas farther from hospitals may experience higher rates of diarrhea due to limited availability of timely medical care. Distance to waste disposal site (X_4) also shows a significant negative effect, aligning with the expectations since living closer to waste sites likely raises exposure to harmful contaminants. Interestingly, greater distance from the nearest river (X_7) is associated with higher diarrhea incidence, which may suggest that communities farther from rivers rely on less clean water sources, thereby increasing their risk.

Population density (X_1) was found to be not significant, possibly because high population density alone does not necessarily increase diarrhea risk if adequate sanitation and clean water access are in place. Rainfall (X_3) and elevation (X_5) may have weaker direct effects on diarrhea incidence, as their impacts could be mediated through other environmental factors. Similarly, slope (X_6) may have a less direct effect on diarrhea incidence because it influences runoff and drainage. Including all predictors in the final model ensures a comprehensive understanding of the various environmental and socio-geographical factors affecting diarrhea incidence, while maintaining consistency with theoretical frameworks and previous literature

All optimization algorithms show well converged results. The BFGS and SANN methods produced a convergence code of 0, which means a stable solution with no numerical problems. For the Genetic Algorithm (GA), convergence was achieved after 142 iterations, with a stable fitness function value at -305.252. The parameter estimates of the three methods showed a consistent pattern in terms of direction and magnitude, further corroborating the stability of the model. Among the three approaches tested, the BFGS algorithm proved to be the most efficient and reliable as it provided faster convergence and the lowest AIC value.

5. CONCLUSION

Based on the analysis results, the Burr XII regression model was successfully applied to analyze diarrhea incidence in Surakarta. The resulting model was able to accommodate positively skewed and heavy-tailed data characteristics, making it suitable for modeling heterogeneous diarrhea incidence across urban villages and capturing extreme observations commonly found in epidemiological data. Parameter estimation using the Maximum Likelihood Estimation (MLE) with the BFGS algorithm achieved the best fit compared to the GA and SANN methods, as indicated by the highest log-likelihood and the lowest AIC value, demonstrating that the BFGS optimization approach contributed to stable, efficient, and convergent parameter estimates that improved the reliability of the regression model. The analysis identified three variables that had a statistically significant effect on diarrhea incidence, namely distance to the nearest hospital, distance to the waste disposal site, and distance to the nearest river, highlighting the important role of healthcare accessibility, environmental sanitation, and exposure to contamination sources in influencing diarrhea

transmission. Meanwhile, population density, rainfall, elevation, and slope were not statistically significant but were retained in the model to control for confounding effects and maintain theoretical relevance. Overall, these findings emphasize that environmental and infrastructural conditions remain key determinants of diarrhea incidence in urban areas and confirm that the integration of Burr XII regression with robust optimization techniques provides a reliable framework for analyzing complex public health data with heavy-tailed distributions, thereby supporting spatially targeted interventions focused on sanitation improvement, environmental management, and equitable healthcare access.

REFERENCES

Al-Shomrani, A. A. (2022). An improvement in maximum likelihood estimation of the Burr XII distribution parameters. *AIMS Mathematics, 7*(9), 17444-17460. https://doi.org/10.3934/math.2022961.

BPS Kota Surakarta. (2025). Jumlah Kasus Penyakit Menurut Kecamatan dan Jenis Penyakit. Badan Pusat Statistik Kota Surakarta. Available at: https://surakartakota.bps.go.id/id/statistics-table/2/MzUwIzI=/jumlah-kasus-penyakit-menurut-kecamatan-dan-jenis-penyakit.html (Accessed: 22 August 2025).

BPS Kota Surakarta. (2025). Jumlah Penduduk Menurut Kelurahan. Badan Pusat Statistik Kota Surakarta. Available at: https://surakartakota.bps.go.id/id/statistics-table/2/NTg4IzI=/jumlah-penduduk-menurut-kelurahan--jiwa-.html (Accessed: 22 August 2025).

BPS Kota Surakarta. (2025). Kepadatan Penduduk per km² Menurut Kelurahan. Badan Pusat Statistik Kota Surakarta. Available at: https://surakartakota.bps.go.id/id/statistics-table/2/NTkwIzI=/kepadatan-penduduk-per-km2-menurut-kelurahan--jiwa-.html (Accessed: 22 August 2025).

BPS Kota Surakarta. (2025). *Kota Surakarta Dalam Angka 2025*. https://surakartakota.bps.go.id/id/publication/2025/02/28/8974f0197a8b96272625a385/kota-surakarta-dalam-angka-2025.html.

BPS Kota Surakarta. (2025). Luas Daerah Menurut Kelurahan. Badan Pusat Statistik Kota Surakarta. Available at: https://surakartakota.bps.go.id/id/statistics-table/2/NTgwIzI=/luas-daerah-menurut-kelurahan.html (Accessed: 22 August 2025).

CHC UCSB. (2025). Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS-2.0), Indonesia Monthly Rainfall Data. *Climate Hazards Center, University of California, Santa Barbara*. Available at: https://data.chc.ucsb.edu/products/CHIRPS-2.0/indonesia_monthly/bils/ (Accessed: 22 August 2025).

Chandan, R., Soni, S., Veeraiah, V., Dhabliya, D., Raj, A., Pramanik, S., & Gupta, A. (2022). Genetic algorithm and machine learning. In *Handbook of Research on Advanced Practical Approaches to Deepfake Detection and Applications* (pp. 168–187). IGI Global. http://dx.doi.org/10.4018/978-1-6684-5656-9.ch009.

de Araújo, F., Guerra, R. R., & Ramírez, F. P. (2022). The Burr XII quantile regression for salary-performance models with applications in the sports economy. *Computational and Applied Mathematics, 41*(6), 1-19. https://doi.org/10.1007/s40314-022-01971-7.

Faidah, D. Y., Darmawan, G., Tantular, B., Pontoh, R. S., Hudzaifa, A. M., & Sain, H. (2023). *Spatial Clusters and Determinants of the High Incidence of Diarrhea in Children*. *Current Medical and Biomedical Research*, 3(2), 35–43. https://doi.org/10.28919/cmbn/8669.

Faidah, D. Y., Hudzaifa, A. M., & Pontoh, R. S. (2023). *Clustering of Childhood Diarrhea Diseases Using Gaussian Mixture Model*. *Current Medical and Biomedical Research*, 3(2), 27–34. https://doi.org/10.28919/cmbn/8365.

Hakim, A. R., Fithriani, I., & Novita, M. (2021). Properties of Burr distribution and its application to heavy-tailed survival time data. *Journal of Physics: Conference Series, 1725(1),* 012016. https://doi.org/10.1088/1742-6596/1725/1/012016.

Jin, Q., Jiang, R., & Mokhtari, A. (2024). Non-asymptotic global convergence rates of BFGS with exact line search. *arXiv.* https://doi.org/10.48550/arXiv.2404.01267.

KEMENKES RI. (2023). *Diare pada Anak*. https://keslan.kemkes.go.id/view_artikel/3028/diare-pada-anak.

Low, V. J. M., Khoo, H. L., & Khoo, W. C. (2021). Quantifying bus travel time variability and identifying spatial and temporal factors using Burr distribution model. *International Journal of Transportation Science and Technology, 11*(1). https://doi.org/10.1016/j.ijtst.2021.07.004.

Ma, H., & Yang, T. (2025). Improved adaptive large neighborhood search combined with simulated annealing (IALNS-SA) algorithm for vehicle routing problem with simultaneous delivery and pickup and time windows. *Electronics, 14*(2), 2375. https://doi.org/10.3390/electronics14122375.

OpenStreetMap. (2025). OpenStreetMap Data: Bangunan, Sungai, Fasilitas Kesehatan, dan Pengolahan Limbah. Available at: https://www.openstreetmap.org/ (Accessed: 22 August 2025).

OpenTopography. (2025). Elevation and Slope Data. Available at: https://portal.opentopography.org/ (Accessed: 22 August 2025).

Purnama, T. B., Wagatsuma, K., & Saito, R. (2025). Prevalence and risk factors of acute respiratory infection and diarrhea among children under 5 years old in low-middle wealth households, Indonesia. *Infectious Disease of Poverty, 14(1),* 13. https://doi.org/10.1186/s40249-025-01286-9.

Puskesmas Poncowarno. (2022). *Yuk Terapkan PHBS untuk Cegah Diare*. https://puskesmasponcowarno.kebumenkab.go.id/index.php/web/post/105/yuk-terapkan-phbs-untuk-cegah-diare.

Ramadhan, R., Fimba, A. B., Fernandes, A. A. R., Solimun, S., Junianto, F. H., Amanda, D. V., & Sumara, R. (2024). Explore The Determinants Of Customers Time To Pay House Ownership Loan On Data With High Multicollinearity With Pca-Cox Regression. *Media Statistika, 17*(2), 117-127. https://doi.org/10.14710/medstat.17.2.117-127.

Santos, M. C. T., & Pescim, R. R. (2023). A new extension of the Burr XII distribution generated by odd log-logistic random variables. *Communications in Statistics - Theory and Methods*, 53(14), 5003–5017. https://doi.org/10.1080/03610926.2023.2200560.

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *Americam Journal of Applied Mathematics and Statistics, 8*(2), 39-42. https://doi.org/10.12691/ajams-8-2-1.

Suparmi, S., Sasman, M. F., Ratnawati, R., & Rustanti, N. (2025). Hygiene and food safety practices among mothers as predictors of diarrhea risk in toddlers in Purwawinangun Village, West Java, Indonesia. *Frontiers in Public Health, 13,* 1530828. https://doi.org/10.3389/fpubh.2025.1530828.

Syafitri, I. N., Rahmawati, R., & Qadrini, L. (2023). Cox Extended Regression Method in Survival Analysis; Case Study of Diarrhea Patients. *Mathematics and Statistics Journal*, *1*(2). https://doi.org/10.35914/mathstat.v1i2.103.

Wibowo, T. P., Melaniani, S., & Salim, L. A. (2021). Modeling of Binary Logistic Regression in the Event of Childhood Diarrhea in Indonesia. *Indian Journal of Forensic Medicine & Toxicology*, *15*(3). https://doi.org/10.37506/ijfmt.v15i3.16020.

Widiastuti, F., Yusuf, S. A., & Iriawan, N. (2023). Survival analysis with Log Burr XII regression on the patient survival time of COVID-19 cases in Jakarta. *AIP Conference Proceedings*, *3165*(1), 040004. https://doi.org/10.1063/5.0215817.

Winarni, T. (2021). Faktor-faktor yang Berhubungan dengan Kejadian Diare pada Balita di Wilayah Puskesmas Pangkalan Balai Kabupaten Banyuasin Tahun 2020.

World Health Organization. (2024). Diarrhoeal disease.

Yang, Y. (2022). A robust BFGS algorithm for unconstrained nonlinear optimization problems. *Optimization, 73*(3), 851-873. https://doi.org/10.1080/02331934.2022.2124869.

Yasin, H., Warsito, B., Hakim, A. R., & Azizah, R. N. (2022). Life Expectancy Modeling Using Modified Spatial Autoregressive Model. *Media Statistika, 15*(1), 72-82. https://doi.org/10.14710/medstat.15.1.72-82.

Zaki, A., Métwalli, A., Aly, M. H., & Badawi, W. K. (2023). 5G and Beyond: Channel Classification Enhancement Using VIF-Driven Preprocessing and Machine Learning. Electronics, 12(16), 3496. https://doi.org/10.3390/electronics12163496.