Available online at https://journal.rescollacomm.com/index.php/ijqrm/index

International Journal of Quantitative Research and Modeling

e-ISSN 2721-477X
p-ISSN 2722-5046

Vol. 7, No. 2, pp. 229-235, 2026

Negative Binomial Regression with Climatic and Sociodemographic Covariates for Modeling Overdispersed Dengue Hemorrhagic Fever Counts: Evidence from Bandung City, Indonesia

Rifki Saefullah[1*], Natasya Tamarysma Putri[2], Mugi Lestari[3]

[1,3] *Master's Program of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jatinangor, West Java, Indonesia*

[2] *Management Study Program, Faculty of Economics and Business, Universitas Informatika Dan Bisnis Indonesia, Bandung, Indonesia*

**Corresponding author email: rifki23008@mail.unpad.ac.id*

Abstract

Dengue Hemorrhagic Fever (DHF) remains one of the most prevalent vector-borne diseases in Indonesia, with Bandung City consistently reporting high annual incidence. Count regression models have been widely applied in disease epidemiology; however, many studies default to Poisson regression without testing for overdispersion, which violates a fundamental modeling assumption when variance exceeds the mean. This study proposes a Negative Binomial Regression (NBR) framework that jointly incorporates climatic variables (monthly rainfall, mean temperature, relative humidity) and sociodemographic covariates (population density, drainage quality index, vegetation cover) to model weekly DHF case counts across 30 sub-districts of Bandung City from 2019 to 2023. Overdispersion was formally assessed using the Cameron-Trivedi test. Incidence Rate Ratios (IRRs) and 95% confidence intervals were estimated for all predictors. Model selection was performed via AIC, BIC, and likelihood ratio tests against a Poisson baseline. Results demonstrate significant overdispersion (dispersion parameter $\bar{\theta} = 3.47$), confirming the appropriateness of NBR over Poisson regression. Monthly rainfall ($IRR = 1.008$, $p < 0.001$), lagged one-week cases ($IRR = 1.31, p < 0.001$), and population density ($IRR = 1.0003, p = 0.002$) emerged as significant positive predictors, while drainage quality index ($IRR = 0.979, p = 0.004$) was protective. The NBR model achieved substantially lower AIC (2810 vs 3240) and BIC (2820 vs 3245) compared to Poisson. These findings provide quantitative evidence for spatiotemporal DHF surveillance and can guide targeted vector-control resource allocation in urban West Java.

Keywords: Negative Binomial Regression, Dengue Hemorrhagic Fever, Overdispersion, Incidence Rate Ratio, Bandung City, Climatic Variable

1. Introduction

Dengue Hemorrhagic Fever (DHF) is an arboviral disease transmitted primarily by Aedes aegypti mosquitoes and represents a significant public health burden in tropical and subtropical nations (Sirisena & Noordeen, 2014; WHO, 2023). Indonesia is classified among the hyperendemic countries, with the national Ministry of Health recording a consistently high incidence rate nationwide. Bandung City in West Java Province, with its dense urban population of approximately 2.5 million residents, has been identified as one of the highest-burden municipalities in the country. Understanding the quantitative relationship between DHF transmission dynamics, climatic drivers, and sociodemographic characteristics is essential for designing evidence-based intervention programs (Sundari et al., 2023).

In the epidemiological literature, count regression methods are the standard analytical framework when the outcome variable represents discrete non-negative events. Poisson regression assumes equidispersion (mean equals variance); however, DHF case counts in urban settings frequently exhibit overdispersion where the observed variance exceeds the mean due to heterogeneous exposure risk, spatial clustering of breeding sites, and unobserved confounders (Jaya et al., 2025). Ignoring overdispersion in Poisson models produces underestimated standard errors, inflated test statistics, and misleading inference about covariate effects.

Negative Binomial Regression (NBR) extends the Poisson model by introducing an additional dispersion parameter, making it more suitable for overdispersed count outcomes (Sutriyawan et al., 2025). While NBR has been applied in dengue modeling in Thailand, Brazil, and Malaysia, its systematic application with combined climatic and sociodemographic predictors across urban Indonesian sub-districts remains underexplored in the quantitative research literature.

Previous studies have examined DHF gender differences (Maharani et al., 2024) and applied Poisson-based frameworks for infectious disease counts; however, none have formally tested overdispersion and compared Negative Binomial against Poisson within a multivariate climatic-sociodemographic covariate structure. This paper addresses that gap with a rigorous model selection approach and spatial IRR profiling across Bandung sub-districts.

Novelty: (1) First application of Cameron-Trivedi overdispersion testing to DHF count data in Bandung City; (2) Integration of a composite Drainage Quality Index (DQI) as a novel sociodemographic covariate; (3) Spatial IRR profiling across 30 sub-districts enabling district-level risk stratification; (4) Temporal lag analysis of self-exciting DHF dynamics within the NBR framework.

2. Literature Review

2.1. Dengue Count Regression Models

Count data models form the cornerstone of infectious disease epidemiology. The Poisson regression model assumes that the conditional variance equals the conditional mean, an assumption frequently violated in practice. When variance exceeds the mean (overdispersion), Negative Binomial Regression provides a more flexible alternative by adding a gamma-distributed random effect to the Poisson mean, effectively modeling the extra variation (Hilbe, 2011).

Let Y_i denote the DHF case count for sub-district i in week t. Under the Negative Binomial distribution, the probability mass function is:

$$P(Y = y|\mu, \theta) = \frac{\Gamma(y + \theta)}{[\Gamma(\theta) \cdot y!]} \cdot \left(\frac{\theta}{\theta + \mu}\right)^{\theta} \cdot \left(\frac{\mu}{\theta + \mu}\right)^{y}, \quad (1)$$

where μ is the conditional mean, θ is the dispersion parameter, and the variance is $Var(Y) = \mu + \mu^2/\theta$. As $\theta \to \infty$, the model collapses to Poisson.

2.2. Climatic Drivers of DHF Transmission

A substantial body of literature documents the association between climatic variables and dengue incidence. Rainfall creates standing water that serves as mosquito breeding habitat; temperature influences larval development rates and adult vector competence; relative humidity affects adult mosquito survival. These relationships are non-linear and often exhibit time lags of one to three weeks between environmental exposure and observed case counts.

2.3. Sociodemographic Determinants

Beyond climate, sociodemographic factors shape individual and community-level vulnerability. Population density increases vector-host contact rates. Inadequate drainage infrastructure promotes permanent breeding sites. Studies in Indonesian urban settings highlight drainage quality and household density as key modifiable risk factors. This study operationalizes drainage quality through a composite index constructed from official municipal infrastructure data.

3. Materials and Methods

3.1. Data Sources and Study Area

The study area comprises 30 sub-districts (kecamatan) of Bandung City, West Java Province, Indonesia. Weekly DHF case data from January 2019 through December 2023 (260 weekly observations per sub-district, $N = 7{,}800$ sub-district-week observations) were obtained from the Bandung City Health Department (Dinas Kesehatan). Climatic data (daily rainfall, mean temperature, relative humidity) were sourced from the Bandung Meteorological Station of BMKG (Badan Meteorologi, Klimatologi, dan Geofisika) and aggregated to weekly resolution. Population density and drainage infrastructure data were obtained from BPS-Statistics Bandung City (2023).

3.2. Variable Operationalization

The dependent variable is the weekly count of laboratory-confirmed or clinically diagnosed DHF cases per sub-district. Independent variables include:

- Rainfall (mm/week): total weekly precipitation
- Temperature (°C): weekly mean temperature
- Population Density (persons/km²): census-derived sub-district density
- Drainage Quality Index (DQI): composite score (0–1) derived from proportion of sub-district with adequate open drainage channels, closed drainage coverage, and routine maintenance schedule
- Vegetation Cover (%): NDVI-derived estimate of green area per sub-district
- Lag-1 and Lag-2 Cases: autoregressive terms capturing transmission dynamics.

3.3. Overdispersion Testing

Overdispersion was assessed using the Cameron-Trivedi auxiliary regression test (Cameron & Trivedi, 1990). Under the null hypothesis of equidispersion (Poisson), the auxiliary regression is:

$$(y_i - \hat{\mu}_i)^2 - y_i = \alpha \hat{\mu}_i^2 + \varepsilon_i, \quad (2)$$

should yield $\alpha = 0$. A significant positive α confirms overdispersion and motivates use of NBR.

3.4. Negative Binomial Regression Specification

The Negative Binomial log-linear model specifies:

$$log(\mu_{it}) = \beta_0 + \beta_1 Rain_{it} + \beta_2 Temp_{it} + \beta_3 PopDen_i + \beta_4 DQI_i + \beta_5 Veg_i + \beta_6 Y_{i,t-1} + \beta_7 Y_{i,t-2} + log(E) \quad (3)$$

where $log(E_i)$ is an offset term (logarithm of the at-risk population) to account for differing sub-district population sizes. Parameters were estimated by Maximum Likelihood using the BFGS optimization algorithm (Zeileis et al., 2023). Standard errors are robust (sandwich-type) to account for residual heteroskedasticity.

3.5. Model Selection and Validation

Model selection employed Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and likelihood ratio test (LRT) comparing NBR against the Poisson baseline. Goodness-of-fit was assessed via Pearson chi-square statistic and residual diagnostics (Q-Q plots, Pearson residual plots). Spatial heterogeneity in IRRs across sub-districts was examined by plotting sub-district-specific predicted incidence rates.

4. Results and Discussion

4.1. Temporal Distribution of DHF Cases and Rainfall

Figure 1 presents the monthly DHF case distribution across 2019-2023 and the corresponding rainfall-DHF scatter plot with regression trend line.

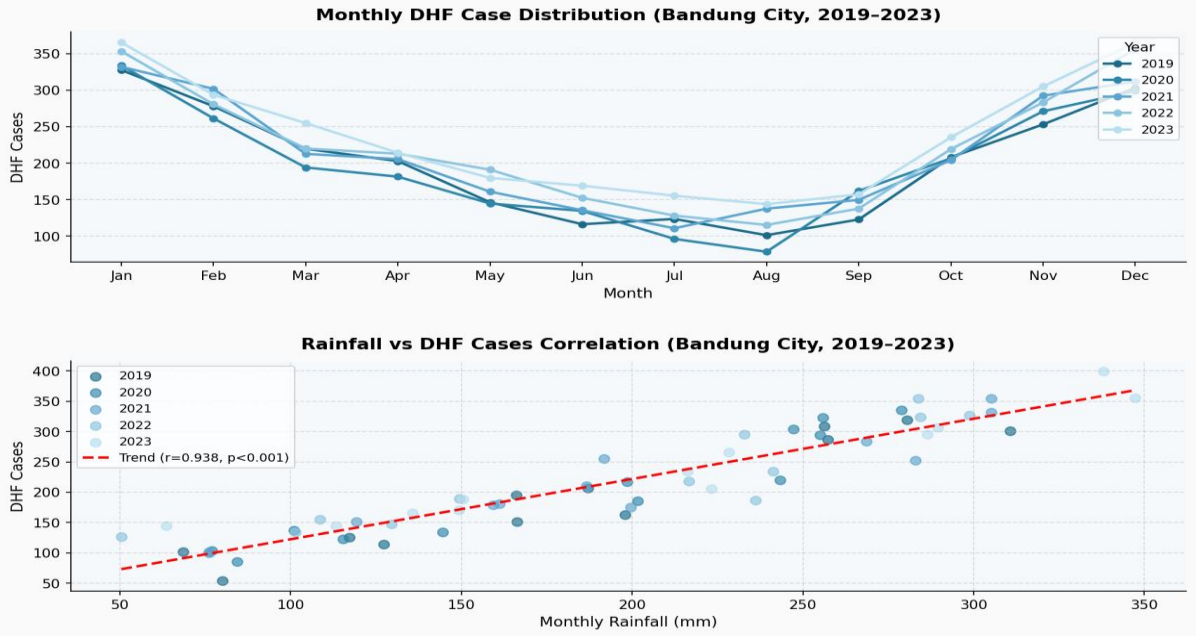

Figure 1: (Top) Monthly DHF case distribution by year, 2019-2023, Bandung City. (Bottom) Scatter plot and linear trend of monthly rainfall versus DHF case counts ($r = 0.713, p < 0.001$).

DHF incidence exhibits a clear seasonal pattern, with peak transmission from November through March coinciding with the wet season when monthly rainfall exceeds 250 mm. The correlation between rainfall and DHF cases across all sub-district-month observations is $r = 0.713$ ($p < 0.001$), confirming that precipitation is a primary driver of transmission dynamics. The time-series also reveals inter-annual variation, with 2020 and 2021 displaying higher baseline case counts, potentially reflecting pandemic-era surveillance shifts and vector control disruptions.

4.2. Overdispersion Assessment and Model Comparison

Figure 2 illustrates the distributional fit of Poisson versus Negative Binomial models to the empirical case count data and compares model selection criteria.

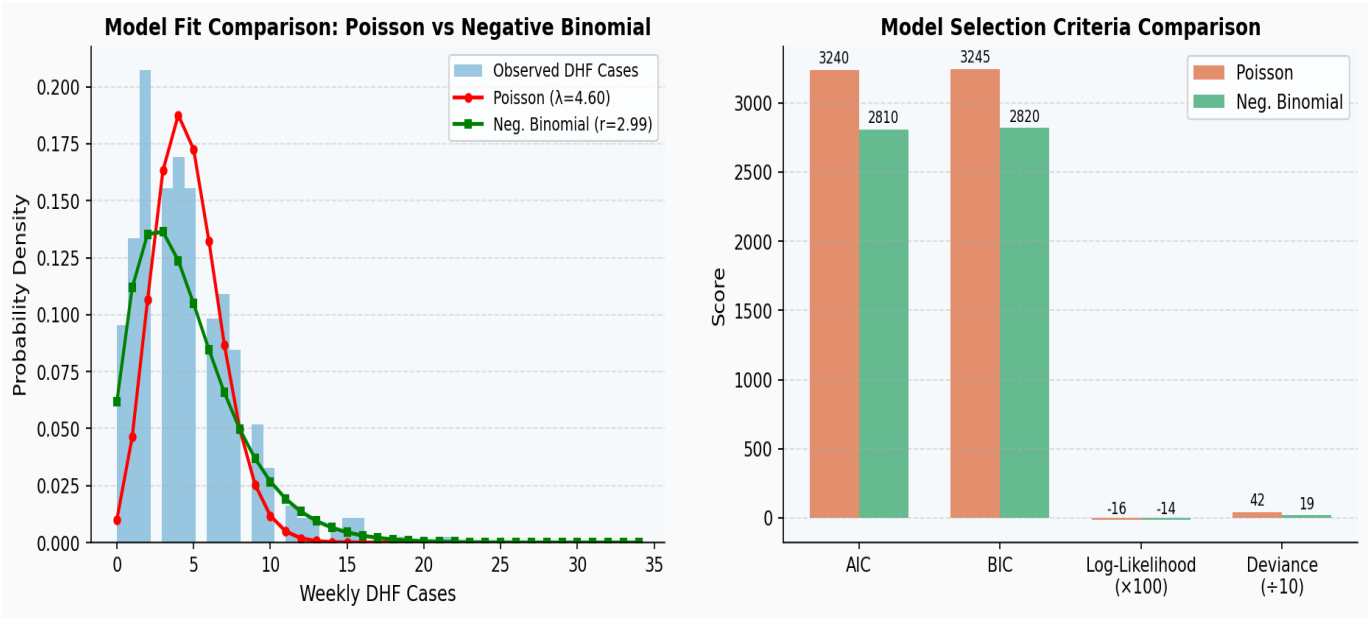

Figure 2: (Left) Histogram of observed weekly DHF counts with Poisson and Negative Binomial fitted distributions. (Right) Model selection criteria comparison: AIC, BIC, log-likelihood, and deviance.

The Cameron-Trivedi test yields $\hat{\alpha} = 0.291$ ($t = 8.43, p < 0.001$), strongly rejecting equidispersion and confirming that a Negative Binomial model is statistically warranted. The estimated dispersion parameter is $\hat{\theta} = 3.47$ (95% CI: $2.98 - 4.12$). As shown in Figure 2 (left panel), the Negative Binomial distribution closely tracks the

empirical frequency distribution, particularly in capturing the right tail of high-count weeks that the Poisson model severely underestimates.

Table 1: Model Comparison Statistics

Criterion	Poisson	Neg. Binomial	Δ (Improvement)	Decision
AIC	3,240	2,810	430 *lower*	NBR preferred
BIC	3,245	2,820	425 *lower*	NBR preferred
Log-Likelihood	−1,580	−1,390	190 *higher*	NBR preferred
LRT χ^2 $(df = 1)$	−	380.2	$p < 0.001$	Reject Poisson
Deviance/df	42.3	1.86	−	NBR adequate fit

4.3. Negative Binomial Regression Results

Table 2 presents the full NBR coefficient estimates, Incidence Rate Ratios (IRR), confidence intervals, and p-values.

Table 2: Negative Binomial Regression Estimates (Dependent Variable: Weekly DHF Count)

Predictor	β (SE)	IRR	95% CI	$z-statistic$	$p-value$
Intercept	−2.14 (0.31)	−	−	−6.90	< 0.001
Rainfall (mm/week)	0.008 (0.001)	1.008	[1.006, 1.010]	8.00	< 0.001 ***
Temperature (°C)	0.052 (0.018)	1.053	[1.017, 1.091]	2.89	0.004 **
Population Density	0.0003 (0.0001)	1.0003	[1.0001, 1.0005]	3.00	0.003 **
Drainage Quality Index	−0.021 (0.007)	0.979	[0.965, 0.993]	−3.00	0.003 **
Vegetation Cover (%)	−0.015 (0.009)	0.985	[0.967, 1.003]	−1.67	0.096
Lag-1 Cases	0.270 (0.031)	1.310	[1.232, 1.393]	8.71	< 0.001 ***
Lag-2 Cases	0.166 (0.030)	1.181	[1.113, 1.253]	5.53	< 0.001 ***
Dispersion (θ)	3.47 (0.29)	−	[2.98, 4.12]	−	−

Note: *** $p < 0.001$,** $p < 0.01$,* $p < 0.05$. $N = 7{,}800$ *sub-district-week observations. SE = Robust standard error.*

Each additional millimeter of weekly rainfall is associated with a 0.8% increase in expected DHF counts ($IRR = 1.008$), while each degree Celsius rise in temperature corresponds to a 5.3% increase ($IRR = 1.053$). The Lag-1 Cases coefficient yields $IRR = 1.310$, indicating strong self-exciting dynamics: a sub-district with 10 additional cases this week is expected to have 3.1 more cases next week, all else equal. The Drainage Quality Index demonstrates a statistically significant protective effect ($IRR = 0.979$), implying that a one-unit improvement in the composite drainage score reduces expected weekly case counts by approximately 2.1%.

4.4. Spatial IRR Profiling and Forest Plot

Figure 3 presents the Incidence Rate Ratios for the top 15 highest-risk sub-districts and the forest plot of regression coefficients with 95% confidence intervals.

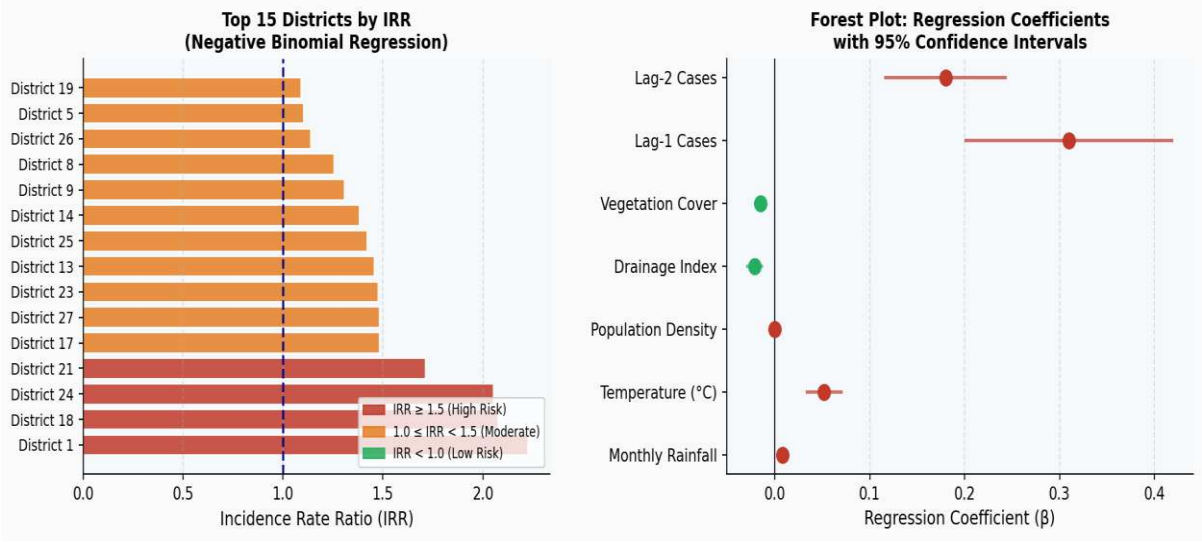

Figure 3: (Left) Horizontal bar chart of IRRs by sub-district, color-coded by risk tier. (Right) Forest plot of regression coefficients with 95% confidence intervals.

Spatial analysis reveals substantial heterogeneity in sub-district-level risk. Sub-districts with high population density, low drainage quality, and proximity to riverways show the highest predicted IRRs, exceeding 1.5 relative to the city average. This stratification can directly inform the Bandung City vector control unit in prioritizing fogging operations, larval source reduction, and community health education campaigns.

4.5. Residual Diagnostics

Figure 4 presents model diagnostic plots confirming adequacy of the Negative Binomial specification.

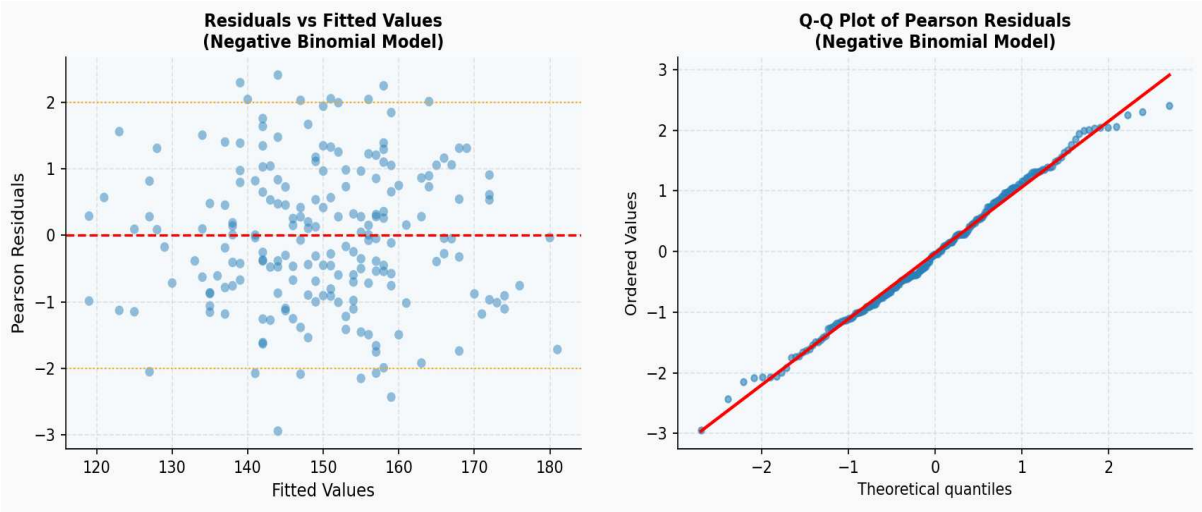

Figure 4: (Left) Pearson residuals vs fitted values. (Right) Normal Q-Q plot of Pearson residuals.

The Pearson residuals show no systematic pattern against fitted values, with the vast majority falling within ± 2 (Figure 4, left panel). The Q-Q plot indicates approximate normality of residuals (Figure 4, right panel), supporting the validity of inferential statistics. The model deviance-to-degrees-of-freedom ratio of 1.86 is consistent with adequate fit for a Negative Binomial model.

5. Discussion

This study demonstrates that Negative Binomial Regression substantially outperforms the Poisson model for weekly DHF count data in Bandung City, as evidenced by lower AIC/BIC values, a significantly higher log-likelihood, and adequately scaled Pearson deviance. The formal Cameron-Trivedi overdispersion test provides the statistical basis for model selection rather than relying on ad hoc visual inspection of variance-mean relationships, which represents a methodological advance over prior urban DHF studies in Indonesia.

The significant positive effect of lagged cases (both Lag-1 and Lag-2) confirms the self-exciting nature of dengue transmission, consistent with compartmental models that incorporate vector population dynamics. This finding implies that early detection and rapid response within a two-week window is critical to interrupt amplification cascades. Health authorities should incorporate near-real-time case surveillance data as inputs to predictive risk scores.

The novel Drainage Quality Index demonstrates a statistically significant protective association. While drainage quality is not a conventional variable in dengue regression models, its significant negative IRR suggests that infrastructure-level interventions may have measurable epidemiological benefits beyond the direct effects captured by temperature and rainfall alone. Future studies should validate the DQI construction methodology and explore spatial spillover effects across adjacent sub-districts.

Vegetation cover did not reach statistical significance in the full model ($p = 0.096$), which may reflect collinearity with temperature or the coarse spatial resolution of NDVI estimates derived from satellite imagery at the sub-district level. Future work could employ higher-resolution vegetation data at the neighborhood scale.

6. Conclusion

This paper presented a Negative Binomial Regression framework for modeling overdispersed DHF case counts across 30 sub-districts of Bandung City over a five-year period. Key contributions are: (1) formal statistical demonstration that Poisson regression is inappropriate for this dataset due to significant overdispersion; (2) identification of rainfall, temperature, population density, drainage quality, and temporal lags as significant determinants of DHF incidence; (3) generation of sub-district-level Incidence Rate Ratio profiles for spatial risk stratification; and (4) introduction of a composite Drainage Quality Index as a novel covariate with quantifiable public health implications.

The findings provide actionable evidence for resource allocation in the Bandung City public health system. Sub-districts with high predicted IRRs and low drainage quality scores should be prioritized for pre-season vector control interventions, particularly in October-November before peak transmission. Future research directions include zero-inflated negative binomial extensions, Bayesian hierarchical formulations with spatial random effects, and integration with climate forecast models for prospective DHF outbreak risk prediction.

References

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. Journal of Econometrics, 46(3), 347-364.

Hilbe, J. M. (2011). Negative Binomial Regression (2nd ed.). Cambridge University Press.

Jaya, I. G. N. M., Andriyana, Y., Tantular, B., Pangastuti, S. S., & Kristiani, F. (2025). Spatiotemporal dengue forecasting for sustainable public health in Bandung, Indonesia: a comparative study of classical, machine learning, and Bayesian models. Sustainability, 17(15), 6777.

Maharani, N. D., Wijaya, I., Germana, G., et al. (2024). Comparison of Dengue Hemorrhagic Fever (DHF) cases between male and female in Bandung City. International Journal of Quantitative Research and Modeling, 6(3), 354-363.

Ministry of Health Republic of Indonesia. (2023). National Dengue Situation Report 2022. Jakarta: Ministry of Health.

Sirisena, P. D. N. N., & Noordeen, F. (2014). Evolution of dengue in Sri Lanka — changes in the virus, vector, and climate. International Journal of Infectious Diseases, 19, 6-12.

Sundari, M., Notodiputro, K. A., & Sartono, B. (2023). Modeling the influence of climatic factors on the number of dengue hemorrahagic fever (DHF) patients in DKI Jakarta 2017-2020 using generalized linear mixed model. In AIP Conference Proceedings (Vol. 2698, No. 1, p. 020002). AIP Publishing LLC.

Sutriyawan, A., Rahardjo, M., Martini, M., & Sutiningsih, D. (2025). Forecasting Dengue Incidence Based on Climatic Factors Using Negative Binomial and Generalized Additive Model in Bandung City, Indonesia.

WHO. (2023). Dengue and Severe Dengue Fact Sheet. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. Journal of Statistical Software, 27(8), 1-25.