BAREKENG: Journal of Mathematics and Its Applications
December 2024 Volume 18 Issue 4 Page 2611–2622
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss4pp2611-2622

FACTORS AFFECTING INDONESIAN PADDY HARVEST FAILURE: A COMPARISON OF BETA REGRESSION, QUASI-BINOMIAL REGRESSION, AND BETA MIXED MODELS

Dian Kusumaningrum[1], Agus Sofian Eka Hidayat[2*], Khairil Anwar Notodiputro[3], Anang Kurnia[3], Bagus Sartono[3], I Made Sumertajaya[3]

[1]School of Applied STEM, Universitas Prasetiya Mulya
Jln. BSD Raya Utama, Tangerang, 15339, Indonesia

[2]Actuarial Study Program, Universitas Presiden
Jln. Ki Hajar Dewantara, Jababeka City, Cikarang, 17530, Indonesia

[3,4,5,6]Department Statistics and Data Science, IPB University
Jln Meranti, Babakan, Dramaga, Bogor, 16680, Indonesia

*Corresponding author's e-mail: ** ***agus.eka@president.ac.id***

ABSTRACT

Article History:

Received: 21st, May 2024
Revised: 7th, July 2024
Accepted: 12th,August 2024
Published:19th, October 2024

Keywords:

Insurance;
Failure-Rate;
Beta Regression;
Quasi-Binomial;
Mixed Model.

The Paddy harvest failure rate is one of the key aspects in determining the total number of claims in a crop insurance policy. It is also an important factor indicating the fulfillment of targeted total production. Therefore, we proposed Beta Regression, Quasi Binomial Regression, and Beta Mixed Models which can be used to analyze significant variables affecting paddy harvest failure rates. Model selection and evaluations indicated that the Nested Beta Mixed Model is the best. Previous research has shown four significant fixed effect variables: drought, flood, pests, and disease risks. Pests and other types of risks also affect the variability of loss rate. All variables have positive effects, indicating higher values cause a higher possibility of a higher average harvest failure rate. High variability was shown for province, municipality, and farmers' random effects. Hence, to prevent a more significant loss rate, MoA should consider more intensive and innovative participatory activities in farmer groups to enhance good farming practices, especially for farmers who suffer from certain risks. These activities should also consider the local characteristics of each province or municipality. As for AUTP development and improvement, farmers with lower failure risks could be given a discounted premium to make it more appealing.

How to cite this article:
D. Kusumaningrum1, A. S. E. Hidayat, K. A. Notodiputro, A. Kurnia, Bagus Sartono and I. M. Sumertajaya., "FACTORS AFFECTING INDONESIAN PADDY HARVEST FAILURE: A COMPARISON OF BETA REGRESSION, QUASI-BINOMIAL REGRESSION, AND BETA MIXED MODELS," *BAREKENG: J. Math. & App.*, vol. 18, iss. 4, pp. 2611-2622, December, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Agriculture, one of the main fields of occupation for the rural population in Indonesia, has been faced with the risk of uncertainty. Since 2015, the Ministry of Agriculture (MoA) has implemented agriculture insurance in Indonesia to shield farmers from crop losses caused by floods, drought, pests, and diseases. MoA selected Jasindo as a state insurance provider to administer an indemnity-focused crop insurance plan, also called Multi-Peril Crop Insurance (MPCI). This policy is well known as *Asuransi Usaha Tanam Padi* (AUTP), and it is heavily subsidized by the government (80%) to help poor farmers who have small land (maximum 2 Ha) afford to pay the premium.

Applying the MPCI policy in Indonesia is a type of insurance where insured land consisting of several plots shares risks caused by perils and natural disasters. A farmer who owns the policy can propose a claim when at least 70% of its insured plot suffers damage mainly caused by floods, drought, pests, and diseases. MoA extension workers do field visits to ensure that the farmer is eligible to process the claim. Afterward, the indemnity of the MPCI will be calculated based on the percentage of loss multiplied by the sum insured, where the sum insured is IDR6 million per hectare for one planting season [1].

eports have shown that after more than 5 years of implementation, an average increase of 73% in the amount of claim value indicates poor farm performance. Flood, drought, and pest and disease infestations were probably uncontrollable, specifically in 2018-2019 [2]. This condition differs from the national objective, namely, increasing rice production. The higher the claim, the worse the achievement of the production target. A better understanding of what causes low rice production or high loss rate would be important to improve the current mechanism. Much research has been done to create AUTP insurance premiums, such as using areas yield-based index [2], rainfall index using burn analysis [3] or copula [4], and claim index [5]. The method used in this research uses more than one aspect of the index to create a better understanding of the loss rate on the impact of variables floods, drought, pests, and diseases in several regions in Indonesia.

The loss rate is one of the key aspects in determining the total amount of claim AUTP insurance policy. It is also an important indicator for the Government of Indonesia (GoI) to fulfill the total production target, leading to self-sufficiency. The higher the rate in a certain area, the more likely it is that the area will not reach the target of total production. Therefore, it will be important to develop a model that can be used to analyze what variables significantly affect the rate of paddy harvest failure. The model can also predict other farmers' harvest failure rates based on selected significant variables.

In this paper, we propose a Beta Regression, Quasi Binomial Regression model, and Beta Mixed Model which are known to be suitable for the proportion data used in this paper. The loss rate modeled in this paper is categorized as proportion data, and it is known that proportion data have a Beta distribution. A Beta distribution is a continuous distribution with values strictly within the (0, 1) interval. Beta Regression and Beta Mixed Models are well known for modeling the bounded data. Nevertheless, Beta Mixed Models have an advantage over Beta Regression when the data has a hierarchical or clustered structure because it incorporates random effects to account for variability within clusters/areas. For example, in our study, we want to examine whether the variability of loss rate among municipalities is apparent. Thus, we compare the two models. Meanwhile quasi binomial regression is more suitable for proportion data when overdispersion is present. In an ideal condition, we expect that loss rates are consistent across fields. However, we may observe that some fields have higher loss rates due to bad soil conditions or higher pest occurrence. This variability between fields could cause the variance of loss rates to be higher and cause overdispersions. Hence, taking into account Quasi Binomial Regression model is also important.

2. RESEARCH METHODS

2.1 Beta Distribution and Quasi Binomial Distribution

The loss rate is the proportion of the damaged area of each plot owned by a certain farmer. Theoretically, if farmers in an area apply good farming practices, the occurrence of a large proportion of loss is minimal, and it is more likely that most farmers have a low rate of losses. Therefore, it is assumed that the loss rate in most areas should have a beta distribution.

The beta distribution is a family of continuous probability distributions defined on the interval [0,1]. It is parameterized by two positive shape parameters, denoted by p and q, where p is a location parameter, and the scale parameter is (q-p). The PDF of the Beta distribution can be written as:

$$f(y;p;q) = \frac{\Gamma(p+q)}{\Gamma p \Gamma q} y^{p-1}(1-y)^{q-1}, 0 < y < 1 \qquad (1)$$

where p, q> 0, and Γ(·) is the gamma function. The expected value or mean of this distribution and variance can be formulated as follows:

$$Mean = E(y) = \mu = \frac{p}{p+q} \text{ and } Var(y) = \frac{pq}{p+q} 2(p+q+1) \qquad (2)$$

The beta distribution is extremely flexible (i.e., it can take on many shapes), depending on the combination of parameter values, including left- and right-skewed or the flat shape of the uniform density (which is a special case of the more general beta density). Still, outcomes must be 0<y<1. There will be problems if we have many 0's in the outcome. When it occurs, we need to add a "zero-inflation" factor. Problems will also be faced if there are many 1's in the outcome, and we need to add "one inflation [6].

The quasi-binomial distribution(QBD), while similar to the binomial distribution, has an extra parameter p_2 (limited to $|p_2| \leq \min\{\frac{p_1}{n}, \frac{1-p_1}{n}\}$) that attempts to describe the additional variance in the data that cannot be explained by a Binomial distribution alone [7]. Research by [8], with the help of an urn model, obtained a three-parameter binomial distribution, and [9]obtained QBD II given by the probability function of

$$P(X = x) = \binom{n}{x} \frac{p_1(1-p_1-np_2)}{(1-np_2)} (p_1 + xp_2)^{x-1}(1-p_1-xp_2)^{n-x-1} \qquad (3)$$

In case $p_2 = 0$, it can easily be seen that QBD's reduce to the classical binomial distribution In QBD's, as the probability of success increases or decreases with the number of successes, the distribution is supposed to be more realistic than the binomial distribution for many practical situations [4].

2.2 Beta Regression

A researcher suggested a regression model for continuous variables within the standard unit interval, such as rates, proportions, or concentration indices [10]. The model assumes beta distribution for the response variable and is referred to as the beta regression model. In their model, the regression coefficients can be understood based on the average of y (the variable under study). A more general model by [11] accounted for the parameter of the precision of the data, which is not assumed to be constant across observations or is allowed to vary, leading to the variable dispersion beta regression model. The densities are parameterized in terms of the mean μ and the precision parameter φ. The flexibility makes the beta distribution an attractive candidate for data-driven statistical modeling.

For the first model Let $y_1, \ldots, y_n$ be a random sample such that $y_i \sim B(\mu_i, \varphi_i)$, for $i = 1, \ldots, n$. The beta regression model is defined as $\beta = (\beta_1, \ldots, \beta_k)^T$ is a k × 1 vector of unknown regression parameters (k < n), $x_i = (x_{i1}, \ldots, x_{ik})$ is the vector of k regressors and η_i is a linear predictor (i.e., $\eta_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$; usually $x_{i1} = 1$ for all i so that the model has an intercept) Here, g(·) : (0, 1):→ IR is a link function, the main motivation for using a link function in the regression structure is

- Initially, when a link function is used on μ_i, both parts of the regression equation take values on the real number line.
- Additionally, there is increased flexibility as the practitioner can select the function that provides the optimal fit.

Some common link functions are logit g(μ) = log(μ/(1 − μ)); probit g(μ) = Φ−1(μ), where Φ(·) is the standard normal distribution function. Other link function includes complementary log-log where g(μ) = log{− log(1 − μ)}; log-log where g(μ) = − log{− log(μ)}; and Cauchy where g(μ) = tan{π(μ − 0.5)}[12].

Note that the variance of y is a function of μ which renders the regression model based on this parameterization naturally heteroskedastic. In particular,

$$Var(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi} = \frac{g^{-1}(x_i^T\beta)[1-g^{-1}(x_i^T\beta)]}{1+\phi} \qquad (4)$$

The log-likelihood function is $\ell(\beta, \varphi) = \sum_{i=1}^{n} \ell_i(\mu_i, \varphi)$, where

$$\ell(\mu_i, \phi) = log\Gamma(\phi) - log\Gamma(\mu_i\phi) - log\Gamma((1-\mu_i)\phi) + (\mu_i\varphi - 1)log y_i + \{(1-\mu_i)\phi - 1\}\log(1-y_i) \quad (5)$$

Notice that $\mu_i = g^{-1}(x_i^T \beta)$ is a function of β, the vector of regression parameters. Parameter estimation is performed by maximum likelihood (ML). An extension of the beta regression model above, which was employed by [13] and formally introduced by [14], is the variable dispersion beta regression model. In this model, the precision parameter is not constant for all observations but instead modeled in a similar fashion as the mean parameter. More specifically, $y_i \sim B(\mu_i, \varphi_i)$ independently, $i = 1, \dots, n$, and where $\beta = (\beta_1, \dots, \beta_k)^T$, $\gamma = (\gamma_1, \dots, \gamma_h)^T$, $k + h < n$, are the sets of regression coefficients in the two equations, η_{1i} and η_{2i} are the linear predictors, and x_i and z_{inc} are regressor vectors. As before, both coefficient vectors are estimated by ML, simply replacing φ by φ_i in Equation 3. This feature is already provided in the betareg package in R [12].

2.3 Quasi Binomial Regression

The beta regression model mentioned above was created so that professionals could accurately represent continuous variables that fall within the unit interval, such as rates, proportions, and inequality or concentration indices [6]. Nevertheless, beta regressions can also represent data types involving proportions of "successes" from multiple trials as long as enough trials support a continuous model. In this scenario, beta regression resembles a binomial generalized linear model (GLM).

To allow for unobserved extra-binomial variation, a continuous variables P_i is introduced. It is independently distributed on (0, 1) with $E(P_i) = \theta_i$, $var(P_i) = \emptyset\theta_i(1-\theta_i)$, and assume that, conditional on $P_i = p_i$, R_i is binomial (m_i, p_i). Unconditionally$E(R_i) = m_i\theta_i$, and $var(R_i) = v_i w_i^{-1}$ where $v_i = m_i\theta_i(1-\theta_i), w_i{}^{-1} = 1 + \emptyset(m_i - 1)$.

Now consider the estimation of β in quasi binomial model when the value of ∅ is known. Maximum likelihood cannot be used because the distribution of the R_i is not fully specified, but the relationship between the expectation and variance of R_i allows the definition of a quasi-likelihood [15] which is maximized with respect to the parameters β by iterative use of the weighted least squares equations

$$\mathbf{X}^T\mathbf{W}\mathbf{V}^*\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{W}\mathbf{V}^*\mathbf{Y}^* \quad (6)$$

where $\mathbf{W} = diag(w_i)$.

An alternative derivation of these equations is obtained by using a Taylor series expansion about an initial estimate β* to obtain the approximating linear regression

$$E(R_i) = m_i\theta_i \cong m_i\,\theta_i^* + v_i^* \sum_s x_{is}\,(\beta_s - \beta_s^*), Var(R_i) \cong w_i^{-1}v_i^*$$

which can be written as

$$E(\boldsymbol{Y}^*) \cong \boldsymbol{X\beta}\,, Var(\boldsymbol{Y}^*) \cong (\boldsymbol{WV}^*)^{-1}$$

The weighted least squares estimate of i in this approximating linear regression is given by Equation (6). The weights w_i, needed if Equation (6) is to be used, depending on ∅ which is usually unknown. If the weights w_i are calculated from an initial estimate of ∅, and β is estimated iteratively from Equation (6) [16]. Beta Regression allows for increased flexibility, especially in situations where the trials are not independent, and the standard binomial model may be too rigid. In this scenario, the fixed dispersion beta regression is similiar to the quasi-binomial model but completely parametric [17].

2.4 Beta Mixed Regression

The Beta Mixed Model was selected as the best model and will be used for further analysis and predictions. This model can be expanded from the beta regression model introduced by [10], including nested random variables shown in [18]. Hence, it can be written as

$$Y_{ijk} \sim \mathrm{B}(\mu_{ijk}, \varphi) \text{ and } \eta_{ijk} = \mathbf{X\beta} + \mathbf{P}\boldsymbol{u_i} + \mathbf{M}\boldsymbol{v_{j(i)}} + \mathbf{F}\boldsymbol{w_{k(j(i))}} \tag{7}$$

where,

$$\boldsymbol{u_i} \sim iid\ N(0, \sigma^2_{Prov}), \boldsymbol{v_{j(i)}} \sim iid\ N\left(0, \sigma^2_{Muncip(Prov)}\right), and\ \boldsymbol{w_{k(j(i))}} \sim iid\ N\left(0, \sigma^2_{Farmer(Muncip(Prov))}\right)$$

and Y_{ijk} denoted the paddy harvest failure rate for farmer k in municipality j nested in province i and $\eta_{ijk} = g\left(E(Y_{ijk})\right) = \mathrm{g}(\mu_{ijk}) = \log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right)$ for the logit link function. Fixed effect estimates of the selected variables are denoted by $\boldsymbol{\beta}$ and $\boldsymbol{u_i}, \boldsymbol{v_{j(i)}}, \boldsymbol{w_{k(j(i))}}$ are the nested random effect of province, municipality, and farmers.

2.5 Data and Variable

This research was based on primary data taken from surveys of 414 farmers, which were held in West Java, East Java, Central Java, Yogyakarta, West Nusa Tenggara, and North Sumatra. The survey was done in 2019 and sponsored by the READI Project. In general, Harvest Failure Rate is the response variable (Y), Eighteen explanatory variables were chosen from aspects including Farmer Identity, Farming Area, Farm Economics, Risks, Farmers Group and Other Activity, and Experience and Satisfaction of crop insurance. The list of variables is given in Table 1.

Table 1. List of Dependent and Independent Variables used in the Model

Aspect	Variable	Name of Variable	Unit of Measurements/Categories
Harvest Rate	Harvest Failure Rate	Y	Percentage (%)
Basic Farmer Identity Information	Age	X1	Years
	Length of Education	X2	Years
Land ownership and Farming Area	Farming Area	X3	Hectares
Farmers Social activity and Experience	Planting Experience	X4	Years
	Interact with extension workers	X5	1= Yes, 0 = No
	Farmer Groups	X6	1= Yes, 0 = No
Farmer Well being	Total Income	X7	Rupiahs (Rp)
	Poverty Level	X8	1= Poor, 0 = Not Poor
Risks Past 3 Years	Drought	X9	Weight between 0-1
	Flood	X10	Weight between 0-1
	Pests	X11	Weight between 0-1
	Plant Disease	X12	Weight between 0-1
	Tornado	X13	Weight between 0-1
	Others	X14	Weight between 0-1
	Total Case of Failure	X15	Total Case
Crop Insurance Policy	Experience with crop insurance policy	X16	1= Yes, 0 = No
Farm Income and Farming Costs	Total Farming Cost	X17	Rupiah (Rp)
	Total Farming Income	X18	Rupiah (Rp)

2.6 Research Process

This study aims to define what variables significantly affect paddy harvest failure rate, and the research process that has been conducted can be seen below.

1. Primary data collection was done through surveys and data preparation of the variables mentioned in Table 1. Data preparation included inputting missing values, excluding unreliable data, and preparing needed tabulations.

2. We were testing the assumptions needed in the model. We used a Kolmogorov-Smirnov test to check whether the harvest failure rate (Y) assumptions meet a beta distribution fulfilled for each municipality.

3. Test results indicated that the municipalities in East Java, North Sumatra, and NTB have a beta-distributed loss rate. Municipalities that don't have beta distribution may have a uniform distribution, indicating that the loss percentage is approximately the same for all farmers in the town. In this article, we will specifically concentrate on modeling regions with a loss rate that has been confirmed to follow a beta distribution.

4. We determined the proportion of training data (80%) and testing data (20%) to be used. Thus, we divide all the data into training and testing data.

5. Using the training data set, develop and apply the purposed Beta Regression, Quasi Binomial Regression, and Beta Mixed Regression. R Packages used include **betareg** package [12] for the Beta Regression, **glm package** for Quasi Binomial Regression models [15], and **glmmTmb** package [19] for Beta Mixed Models. The models' parameters estimates were based on likelihood estimators, and the link functions selected were logit, probit, complementary loglog (cloglog), and cauchit.

6. Independent variable selection for all models, defining variance condition for Beta Regression, and determining the most appropriate random effect for this study in the Beta Mixed Regression. The variable selection process was conducted by simultaneously removing insignificant variables. Next, for defining variance conditions, there is a fixed and flexible variance option. At the same time, the selection of random effects included testing Province, Municipalities, Farmer random effects, and nested or cluster random effects in the Beta Mixed Models. The nested random effect was Municipalities Nested in Province along with Farmer random effect. In contrast, the cluster random effect was based on a pre-clustering paddy harvest failure rate process.

7. Thus, there is more than one model built and compared, which includes:

   i. Beta Regression Models that use all independent variables (Full Model) with Fixed Variance.
   ii. Beta Regression Models that use selected independent variables (Reduced Model) with Fixed Variance.
   iii. Beta Regression Models that use all independent variables (Full Model) with Flexible Variance.
   iv. Beta Regression Models that use selected independent variables (Reduced Model) with Flexible Variance.
   v. Quasi Binomial Regression Models that use all independent variables (Full Model).
   vi. Quasi Binomial Regression Models that use selected independent variables (Reduced Model).
   vii. Beta Mixed Models that use all independent variables (Full Model) with Province, Municipalities, Farmer random effects.
   viii. Beta Mixed Models that use all independent variables (Full Model) nested or cluster random effects.
   ix. Beta Mixed Models that use selected independent variables (Reduced Model) with Province, Municipalities, Farmer random effects.
   x. Beta Mixed Models that use selected independent variables (Reduced Model) with nested or cluster random effects.

8. Define the best fit Beta Regression, Quasi Binomial Regression, and Beta Mixed Regression model based on the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) Value formulated as:

$$AIC = -2 * log(L) + 2 * k \tag{8}$$

   Where L represents the maximized likelihood of the model and k represents the number of parameters in the model. The selected best-fit model will have the lowest AIC value. We will also check the assumption of the models based on their residuals. We will verify indications of heterogeneity and large residual values indicating that the model does not have a good fit.

9. Predict harvest failure rate using the testing data set based on the best fit model selected above. Evaluate the prediction accuracy results based on:
   i. Mean Absolute Error (MAE), formulated as:

$$MAE = \frac{\sum|y_i - \hat{y}_i|}{n} \tag{9}$$

ii. Root Mean Square Error (RMSE), formulated as:

$$RMSE = sqrt\left[\frac{\sum(y_i - \hat{y}_i)}{n}\right] \tag{10}$$

iii. R-Square between actual and predicted values, formulated as:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} \tag{11}$$

where $SS_{RES} = \sum(y_i - \hat{y}_i)^2$ and $SS_{TOT} = \sum(y_i - \bar{y}_i)^2$

The best model with the best predictions will have the lowest MAE and RMSE value. It will also have the highest R^2 value.

10. Analyze and deliver conclusions and recommendations. When the best model is found, we can finalize this paper by providing findings and recommendations that are beneficial for improving the MPCI policy selection process and deliver insights on how the GoI can decrease individual losses to meet the national total production.

3. RESULTS AND DISCUSSION

3.1 Harvest Failure Rate

The histogram of harvest loss rate of all municipalities with a beta distribution **Figure 1** is skewed to the right, indicating that more farmers have low rate harvest losses. We can estimate the two positive shape parameters, p, and q, using R's maximum likelihood parameter estimation method, which is equivalent to p = 0.426285 (location parameter). While q =2.301574, the scale parameter will be (q-p) = 1.875289. Based on these parameters, we can estimate the mean and variance of the harvest loss rate using **Equation (2)**. Hence, this led to an average of 0.16 and a variance of 0.09. This indicates that the average land loss in the selected areas is 16%.

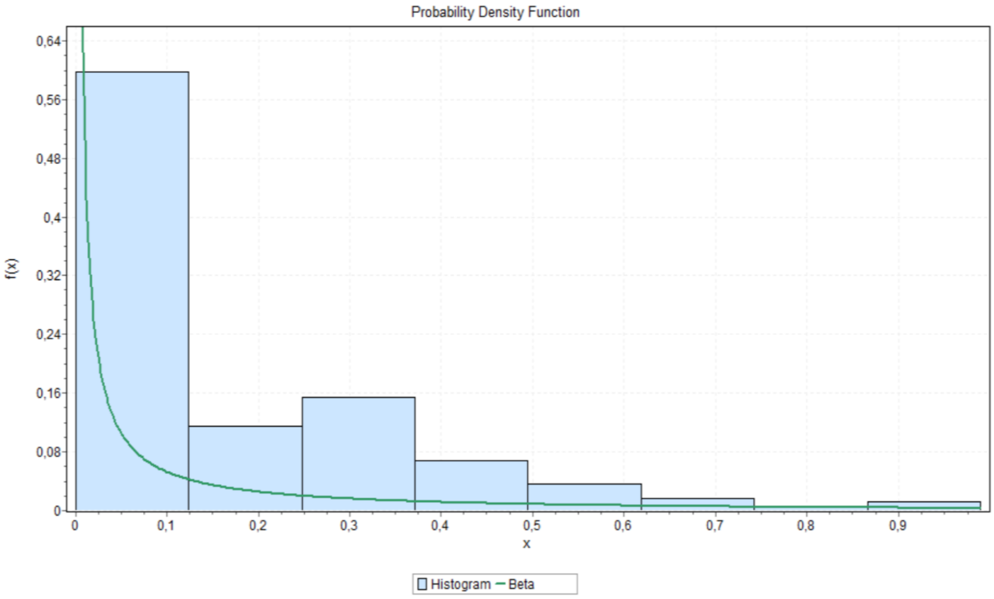


Figure 1. Distribution of paddy harvest loss rate

Table 2 shows the Beta distribution Maximum Likelihood parameter estimates of loss harvest rate conditions for each chosen municipality. Location and shape parameter estimates differ among municipalities, causing different averages. By using **Equation (2)**, the minimum average paddy harvest loss rate is found in Malang (6%), and the maximum is in Kulon Progo (47.4%). The wide range of mean values shows the diversity in areas of loss. Areas of research for developing a more specific local model for each municipality, which includes land quality, climate, spatial effect, and other local variables, are apparent.

Table 2. Maximum Likelihood Parameter Estimates of Beta Distribution of Paddy Harvest Loss Rate

Municipality	Location (p)	q	Shape (q-p)	Mean	Var
All	0.426	2.302	1.875	0.156	0.088
Kulon Progo	1.463	1.623	0.160	0.474	0.174
Lamongan	0.580	3.752	3.173	0.134	0.090
Langkat	1.011	3.491	2.480	0.225	0.137
Lombok Barat	0.371	4.129	3.758	0.082	0.060
Lombok Tengah	0.333	2.711	2.378	0.109	0.068
Malang	0.341	5.130	4.789	0.062	0.048
Simalungun	0.676	4.953	4.277	0.120	0.087

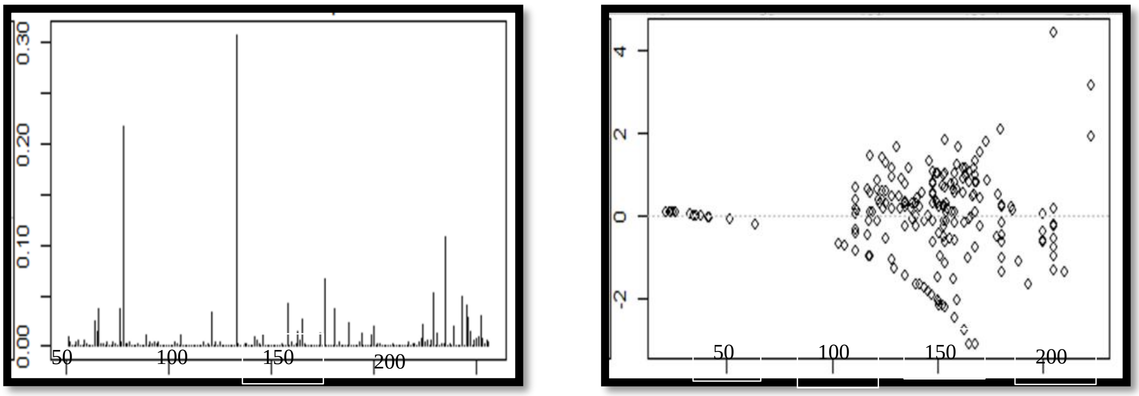
At first, we will find the best fit Beta Regression and Quasi Binomial Regression based on its AIC values that can be calculated by using **Equation (8)**. In total, twenty beta regression models were developed, and eight Quasi Binomial Regression developed. Hence, we selected the five best models for simplicity, as shown in **Table 3**. We made the predictions based on the testing data sets and compared the evaluation of the fitted values based on training data sets side by side in **Table 3**. The calculation of MAE, RMSE, and the R^2 between actual and predicted values were based on **Equation (9)**, **Equation (10)**, and **Equation (11)** which were needed to assess the prediction performance of each model. There were only slight differences between the five models. The performance between training and testing data sets was also very similar.

Table 3. Evaluation of the Predictions Based on Selected Beta Regression Quasi Binomial Regression

Model	Type of Model	Variables in the Model	Link Function	MAE		RMSE		R^2	
				Train Data	Test Data Set	Train Data	Test Data Set	Train Data	Test Data Set
model2_2a	Beta Regression	Reduced Model	cloglog	0.115	0.092	0.009	0.028	0.640	0.517
model1_2a	Beta Regression	Full Model	cloglog	0.114	0.095	0.008	0.026	0.654	0.492
model1_1	Beta Regression	Full Model	logit	0.117	0.102	0.013	0.033	0.643	0.452
model3_2a	Quasi Binomial	Reduced Model	Probit	0.107	0.084	0.000	0.025	0.647	0.565
model3_1a	Quasi Binomial	Full Model	Probit	0.106	0.088	0.000	0.026	0.660	0.530

Quasi Binomial Regression models slightly outperform Beta Regression, but the difference is insignificant. Therefore, choosing between these models will be similar; we can use both models in this case. Nevertheless, when data collected is less and productivity loss shows indications of a Beta distribution, Beta Regressions are suggested due to their flexibility in modeling data. We can choose between fixed variance models and even flexible variance models. We can improve the model by modeling the variance based on chosen explanatory variables, especially in cases where the variance differs among areas or over time. For example, in this case, it is known that the paddy loss ratio is different in mean and variances among municipalities.

Risk Aspects (Drought, Flood, Pests, and Disease) are the most significant variables found in the best-fit Beta regression. Pests and other risks significantly affect the variability of the loss rate (**Table 4**). All these variables have positive effects; the higher the value of these variables, the higher the possibility of a higher average harvest failure rate. We also found that farmers' group membership, participatory activities, and experience owning crop insurance policies do not significantly affect loss rates. It was expected that farmers having these experiences would have higher knowledge or awareness of how to prevent failure. These findings suggest that these aspects' direct impact on the paddy harvest loss rate might be limited due to various practical, behavioral, or contextual factors, and further evaluations and adjustments will be made. Even though the Beta Regression model results are promising, in general, this model still shows insufficient results. Therefore, further assumption evaluation on the residuals should be carried out. **Figure 2** shows indications of the existence of high-leverage outliers and heteroscedasticity. Henceforth, the Beta Mixed model was developed based on the significant variables obtained to improve the model fit accuracy.



(a) (b)

Figure 2. Beta regression residual diagnostics based on (a) Cooks distance plot (b) Residuals and predictor plot

Next, we developed the Beta Mixed model and defined the best one. The challenge in determining the best fit Beta Mixed models is finding the most suitable random effects. As mentioned above, we have considered Province, Municipalities, Farmer, nested, and cluster random effects in the Beta Mixed Models. Therefore, there were fifteen models. The sufficient random impact for this case study was found by adding the farmers' random effect, which is nested in the municipalities' random effect and nested in the province's random effect. The structure of the selected Beta Mixed model can be seen in Equation (7). Findings show that the nested Beta Mixed model is sufficient to model harvest loss rates (Table 4) because it has a lower AIC value. Therefore, we can conclude that the Beta Mixed model is the best overall.

Table 4. Evaluation of the Best Fit Beta Regression and Beta Mixed Regression

Variables	**Beta Regression**			**GLM M Beta Mixed Regression**		
	Estimate	**Std.Error**	**P-Value**	**Estimate**	**Std.Error**	**P-Value**
Model for Mean						
Intercept	-2.365	0.376	0	-2.537	0.437	0
X2	0.029	0.018	0.106	0.02	0.017	0.25
X8	0.19	0.138	0.168	0.162	0.144	0.261
X9	1.356	0.357	0	1.491	0.381	0
X10	1.044	0.342	0.002	1.278	0.339	0
X11	0.883	0.298	0.003	0.94	0.306	0.002
X12	0.975	0.373	0.009	1.458	0.379	0
X13	1.449	0.698	0.038	0.451	0.309	0.144
X14	0.517	0.302	0.087	-2.537	0.437	0
Kulon Progo	0.999	0.277	0	**Random Effects**		
Lamongan	-0.709	0.228	0.002	**Groups**	**Variance**	**Std.Dev.**
Langkat	-0.47	0.258	0.069	Province	0.30217	0.1078
Lombok Barat	-0.879	0.273	0.001	Municipality	0.01162	0.5497
Lombok Tengah	-0.656	0.284	0.021	Farmers	0.05858	0.242
Malang	-1.049	0.243	0	[redacted]	[redacted]	[redacted]
Simalungun	-0.77	0.218	0	[redacted]	[redacted]	[redacted]
Phi Coeff	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Model for Variance						
(Intercept)	1.075	0.169	0	10.889	0.2453	9.02E-06
X11	0.766	0.246	0.002	0.6984	0.2464	0.0046
X14	0.634	0.246	0.01	0.7513	0.2427	0.00196
Model Evaluation						
AIC	-470.983			-546.000		

By choosing the best model we can further analyze the impact of each variable on paddy harvest loss ratios. Through further analysis, we can deliver recommendations to the MoA to boost productivity and suppress the harvest-loss ratio. On the other hand, Jasindo and MoA can also use this model to develop a selection model for crop insurance policies. Based on the information gathered, this model can predict whether farmers have high or low potential harvest loss rates. Hence, it will be helpful to indicate which farmers need attention and guidance and which are good to go on their own. Also, it can filter farmers with a higher loss ratio and give them higher premiums to prevent low-profit levels. We explained further discussions on this issue below.

The mixed beta regression strengthens the idea that risk aspects (Drought, floods, pests, tornadoes, and other risks) dominate the paddy harvest loss rate. By calculating the odds ratio from the fixed effect parameters in **Table 4**, drought and plant disease were the most significant threats causing high average loss rates. A percentage increase in both risks causes more than four times the chance of an increase in the average paddy harvest loss rates. Floods and pests then follow these risks. Pest and other risks not only have an impact on the average loss rates but also the variance of loss rates. Knowing that the current AUTP is already covering most of these risks is quite assuring for farmers. Meanwhile, other types of risks should also be considered, such as tornadoes, pollution, fire, or other local risks. As for the random effects, the variance of Province random effects had the highest variability, followed by the variance of nested farmers random effect (**Figure 3**).

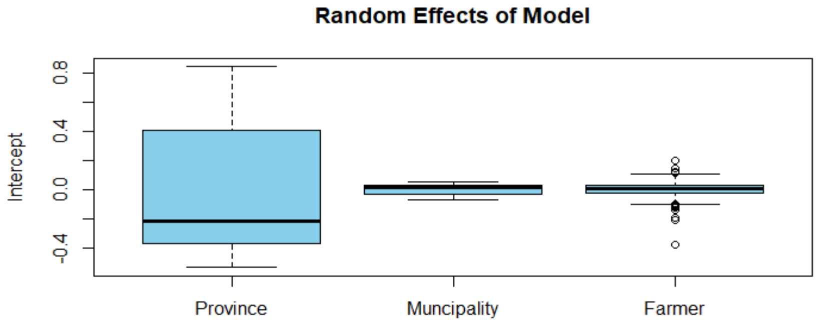


Figure 3. Boxplot of the variance of nested beta mixed model random effects

It has been known that when significant risks cause the harvest loss rate, it will be correlated at a farm level. This means that when the dominant cause of failure, such as pest or plant disease, happens on one farm, it will spread quickly throughout other farms. Therefore, high loss rates are caused not only at the farm level but also at the municipality and province levels. This can be justified by testing the variances of the nested random effects. For example, by using an alternative hypothesis of correlation among farmers within a municipality, we can prove the fact mentioned earlier **[18]**. All the random effects are significant when the significance level is set at 0.05. It implies a considerable correlation among farmers within a municipality, as most of them come from the same farmer groups. There is also a correlation among municipalities within the province. This suggests that the risks that cause harvest failure are directly transmitted from farmer to farmer and can upscale to municipalities within provinces. Henceforward, early mitigation of farmers having areas prone to drought, plant disease, flood, and pests is crucial. Socializing state-of-the-art methods to prevent or handle this situation will also be beneficial. Last, further research on factors causing these risks and developing early warning systems to avoid massive upscaled impact will also be apparent.

4. CONCLUSIONS

To overcome the high occurrence and amount of claims in crop insurance and also as an effort to reach the targeted productivity set by MoA, modeling loss harvest rate is important. Three candidate models were chosen: Beta Regression, Quasi Binomial Regression, and mixed Beta Regression. The model selection process and evaluation based on AIC values point out that Beta Mixed Regression outperforms the other two models. This model shows that the most significant fixed effect variables are the Risk Aspects Drought, Flood, Pests, Tornado, and Other Risks). Drought and plant disease are estimated to have caused the highest impact on average loss rates. Pests and other types of risks also significantly affect the variability of the average loss ratio. Length of Education and Poverty Level also significantly affect paddy harvest loss rates but at a lower significance level. All these variables have positive effects, indicating the higher the value of these variables, the higher the possibility of having a higher average harvest loss rate. Another interesting aspect to investigate is that farmers' groups, participatory activities, and experience in owning a crop insurance policy have no direct effect on loss ratios. It was expected that farmers having these experiences would have higher knowledge or awareness of how to prevent loss. As for the random effects, it was proven that there is a significant correlation among farmers within a municipality, as most of them come from the same farmer groups. A correlation among municipalities within the province followed them. We suggest that the risks that cause harvest failure are directly transmitted from farmer to farmer and can be upscale to municipalities within provinces.

Henceforward, to prevent more significant loss rates. First, early mitigation of farmers in areas prone to drought, plant disease, flood, and pests is crucial. Secondly, socializing state-of-the-art methods to prevent or handle this situation will also be beneficial. The third aspect is more intensive participatory activities throughout farmer groups to enhance good farming practices of farmers who often suffer from the above risks. These activities should consider the local characteristics of each province and municipality. Innovations in delivering knowledge to farmers also need to be upgraded. Last, further research on factors causing these risks and developing early warning systems to prevent massive upscaled impact will also be apparent. Specifically, for AUTP development and improvement, insurance companies can have prior knowledge and estimates of the farmer's loss rates based on further predictions of this model. Farmers estimated to have lower loss rates could be considered given a reduced premium. Therefore, this enhances the policy's appeal to farmers.

ACKNOWLEDGMENT

The writer extends gratitude to Prasetiya Mulya University for enabling the pursuit of a PhD degree, SEARCA for offering a full PhD scholarship, The Ministry of Agriculture for sharing data and knowledge, and The READI Project of Global Affairs Canada for funding and supporting data collection and training.

REFERENCES

[1] S. M. Pasaribu and A. Sudiyanto, “Agricultural risk management: Lesson learned from the application of rice crop insurance in Indonesia,” *Climate change policies and challenges in Indonesia*, pp. 305–322, 2016.

[2] R. and S. V. A. and T. K. S. Kusumaningrum Dian and Anisa, “Alternative Area Yield Index Based Crop Insurance Policies in Indonesia,” in *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, M. and P. C. and P. C. and S. M. Corazza Marco and Gilli, Ed., Cham: Springer International Publishing, 2021, pp. 285–290.

[3] A. S. E. Hidayat and A. C. Sembiring, “Application of the Historical Burn Analysis Method in Determining Rainfall Index for Crop Insurance Premium Using Black-Scholes,” *Journal of Actuarial, Finance, and Risk Management*, vol. 2, no. 2, pp. 1–9, 2023.

[4] A. S. E. Hidayat and G. Gunardi, “Calculation of crop insurance premium based on dependence among yield price, crop yield, and standard rainfall index using vine copula,” in *AIP Conference Proceedings*, 2019.

[5] W. Estiningtyas, “Asuransi pertanian berbasis indeks iklim: opsi pemberdayaan dan perlindungan petani terhadap risiko iklim,” *Jurnal Sumberdaya Lahan*, vol. 9, no. 1, 2015.

[6] F. Cribari-Neto and A. Zeileis, “Beta Regression in *R*,” *J Stat Softw*, vol. 34, no. 2, 2010, doi: 10.18637/jss.v034.i02.

[7] A. Mishra, D. Tiwary, and S. K. Singh, “A class of quasi-binomial distributions,” *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 67–76, 1992.

[8] P. C. Consul, “A simple urn model dependent on predetermined strategy, Sankhya, B, 36,” 1974.

[9] P. C. Consul and S. P. Mittal, “A new urn model with predetermined strategy,” *Biom Z*, vol. 17, no. 2, pp. 67–75, Jan. 1975, doi: 10.1002/bimj.19750170202.

[10] S. Ferrari and F. Cribari-Neto, “Beta regression for modelling rates and proportions,” *J Appl Stat*, vol. 31, no. 7, pp. 799–815, 2004.

[11] S. L. P. Ferrari, P. L. Espinheira, and F. Cribari-Neto, “Diagnostic tools in beta regression with varying dispersion,” *Stat Neerl*, vol. 65, no. 3, pp. 337–351, 2011, doi: https://doi.org/10.1111/j.1467-9574.2011.00488.x.

[12] F. M. Bayer and F. Cribari-Neto, “Model selection criteria in beta regression with varying dispersion,” *Commun Stat Simul Comput*, vol. 46, no. 1, pp. 729–746, Jan. 2017, doi: 10.1080/03610918.2014.977918.

[13] M. Smithson and J. Verkuilen, “A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables.,” *Psychol Methods*, vol. 11, no. 1, pp. 54–71, Mar. 2006, doi: 10.1037/1082-989X.11.1.54.

[14] A. B. Simas, W. Barreto-Souza, and A. V. Rocha, “Improved estimators for a general class of beta regression models,” *Comput Stat Data Anal*, vol. 54, no. 2, pp. 348–366, Feb. 2010, doi: 10.1016/j.csda.2009.08.017.

[15] R. W. M. WEDDERBURN, “Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method,” *Biometrika*, vol. 61, no. 3, pp. 439–447, 1974, doi: 10.1093/biomet/61.3.439.

[16] P. McCullagh, *Generalized linear models*. Routledge, 2019.

[17] D. A. Williams, “Extra-Binomial Variation in Logistic Linear Models,” *Appl Stat*, vol. 31, no. 2, p. 144, 1982, doi: 10.2307/2347977.

[18] C. E. McCulloch, “AN INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS,” *Conference on Applied Statistics in Agriculture*, Apr. 1996, doi: 10.4148/2475-7772.1314.

[19] B. Bolker, “Getting started with the glmmTMB package,” *Cran. R-project vignette*, vol. 9, 2019.