Numerical Prediction of paddy weight of Crop Cutting Survey using Generalized Geoadditive Linear Mixed Model

Muhlis Ardiansyah[1,2] **Anang Kurnia**[2*], **Kusman Sadik**[2], **Anik Djuraidah**[2], **Hari Wijayanto**[2]
[1]*BPS-Statistics of Kotawaringin Timur, Central Kalimantan, Indonesia*
[2] *Department of statistics, IPB University, Bogor, Indonesia*

[*)] Corresponding author: anangk@apps.ipb.ac.id

Abstract. Rice production data is needed to support the information about achieving the second SDGs. Rice production data requires rice productivity data obtained from Crop Cutting Survey by BPS-Statistics Indonesia. The problem is that the measurement of unhulled rice weight in this survey is not always successful. This problem causes the unhulled rice weight data to be missing values. We proposed Geo-GLMM with covariate interaction to estimate the missing values. The proposed methods was compared by GLM, GLMM, and Geo-GLMM. The results showed that seed varieties, TSP/SP36 fertilizer, NPK/ compound fertilizer, urea, organic fertilizer, the number of clumps per plot, pest attack, and climate impacts significantly affected rice productivity. Then, we selected the variables and got the best explanatory variables, namely seed varieties, fertilizer, interaction between urea and KCL fertilizer. Geo-GLMM with fertilizer interaction has better prediction performance than GLM, GLMM, and Geo-GLMM without interaction. Based on the results of the simulations, the Geo-GLMM with covariate interaction produces a smaller bias and RMSE. Therefore, it is recommended that the surveyors of Crop Cutting Survey continue to interview farmers when they fail to take sample plots, so we get covariate data can be used to estimate the unhulled rice weight.

***Keywords*:** *Crop Cutting Survey, Geoadditive*-GLMM, *missing values, rice productivity.*

1. Introduction

Rice productivity data is needed to calculate national rice production. Rice productivity is obtained from Crop Cutting Survey by BPS-Statistics Indonesia. Crop Cutting Survey is a survey that aims to get the rice productivity data. This data is obtained by randomly taking rice plots $2.5 \times 2.5\ m^2$ to be harvested, cleaned, and weighted. In 2019, a sample plot of rice plant is integrated with the observation point of Area Sampling Frame (ASF). ASF method is used to measure the rice harvest area and the phase of rice growth. Meanwhile, BPS officers must take plot samples to get rice productivity data through a Crop Cutting Survey. Multiplication between harvested area from ASF and rice productivity from the Crop Cutting Survey results national rice production data.

The problem of Crop Cutting Survey is that the measurement of unhulled rice weight on selected plots is not always successful, especially in areas with difficult accessibility. BPS officials

cannot take some rice plot samples because its have been harvested by farmers. This problem causes the missing vaules of unhulled rice weight data.

There are three problems due to missing data. First, it can reduce the representativeness of the samples. If the selected sample is in a group with low rice productivity, the estimated results will be lower than the true value (underestimate). On the other case, if the selected sample is in a group with high rice productivity, the estimated results will be overestimated. Second, lack of sample representation causes measurement bias. The missing data can produce biased estimation. Third, the missing data can reduce the statistical power of a survey.

This study is a continuation of research of Ardiansyah and Tofri (2019) which examined whether the measurement methods in the Crop Cutting Survey can be replaced with postharvest interviews. In the post-harvest interview method, BPS officials simply ask how much harvested area and weight of grain in the harvested area to obtain rice productivity data. Based on Ardiansyah and Tofri (2019), rice productivity from postharvest interviews is lower than the measurement results so there is not enough evidence to replace the plot measurement of Crop Cutting Survey with postharvest interviews (farmer recognition).

In this study, a new solution proposed to overcome the problems with the imputation technique. Ardiansyah et al (2020) showed that the imputation technique is better for estimating rice productivity than deleting missing data. Imputation is the replacing missing data with estimated values. This approach preserves all cases by replacing the missing data with a probable value estimated by other available information. After all missing values have been replaced, the dataset is analyzed using the standard techniques for a complete data. The explanatory variables are used to make a prediction, and the predicted value is substituted as if an actual obtained value.

An advanced and acurated statistical model is needed to predict the response variables by utilizing explanatory variables that are easily obtained. The addition of geo-spatial information into the model provides a good estimation performance in predicting rice productivity in the Crop Cutting Survey (Ardiansyah et al. 2018). Some previous studies that applied geoadditive modeling for estimation are Yu et al. (2019) and Muleia et al. (2020). In this study, addition of geo-spatial information uses a Gaussian process basis smooth. The geoadditive model with the Gaussian process provides a flexible regression relationship that will make predictions more accurate than the Generalized Linear Model (GLM) and Generalized Linear Mixed Model (GLMM). This study aims to: (1) get explanatory variables that are able to predict $2.5 \times 2.5\, m^2$ unhulled rice weight and (2) evaluate geoadditive performance in increasing accuracy of predictions.

2. A Review of Methods

The methods used to predict the 2.5×2.5 m^2 unhulled rice weight are the Generalized Linear Models (GLM), Generalized Linear Mixed Model (GLMM), and Generalized Geoadditive Linear Mixed Models (Geo-GLMM).

2.1 GLM

GLM is extension of the linear model. This extension encompasses non-normal response distributions and link functions of the mean equated to the linear predictor. There are three components of a GLM: (1) random component, (2) linear predictor and (3) link function.

The random component of a GLM is the response variable y with independent observations (y_1,…, y_n) from a distribution having probability density or mass function for y_i of the form (Pan *et al* 2019):

$$f_Y(y, \theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \quad (1)$$

The form (1) is called the exponential dispersion family. The parameter θ is called the natural parameter, and ϕ is called the dispersion parameter. Distribution that includes exponential dispersion families: normal, binomial, multinomial, poisson, gamma, exponential, and negative binomial distribution. Suppose y is a random variable that has a normal distribution, then the probability density function (pdf) is

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \quad \text{for} \quad -\infty < y < \infty \tag{2}$$

From equation (2), it will be proven that the normal distribution is included in the exponential family distribution with the following three steps:

i) The natural logarithm on both sides in equation (2):

$$\log\{f(y|\mu,\sigma^2)\} = \log\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}\right\} = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2}$$

ii) Involves exponential on both sides:

$$\begin{aligned}
\exp(\log\{f(y|\mu,\sigma^2)\}) &= \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2}\right\} \\
f(y|\mu,\sigma^2) &= \exp\left\{-\frac{y^2-2\mu y+\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\
&= \exp\left\{\frac{2\mu y-\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\
&= \exp\left\{\frac{\mu y-\frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}
\end{aligned}$$

iii) Comparing with the general form of the exponential family in equation (1):

$$\theta = \mu;\ b(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2;\ a(\phi) = \sigma^2;\ \text{and}\ c(y,\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

From i) to iii) it can be shown that the normal distribution belongs to the exponential family distribution. With the same steps, it can be shown easily that the distribution of binomial, multinomial, poisson, gamma, exponential, and negative binomials is also included an exponential family distribution.

The systematic component (the second component of GLM) is a linear combination of covariates $X_0, X_2, \dots, X_{p-1}$ that can be written in the notation $\eta = \sum_{k=0}^{p-1}\beta_k X_k$ with η as a linear estimator and β_k are constants. The third component is the link function that connects between the random component and the linear predictor. The link function $g(.)$ is monotonously differentiable.

The probability density function included in the exponential family distribution have similarity of expected value $\mathrm{E(Y)} = \frac{\partial \mathrm{b}(\theta)}{\partial\theta}$ and $\mathrm{Var(Y)} = a(\phi)\frac{\partial^2 \mathrm{b}(\theta)}{\partial\theta^2}$. Suppose y follow the normal distribution then $\mathrm{E(Y)} = \frac{\partial \mathrm{b}(\theta)}{\partial\theta} = \frac{\partial\left(\frac{1}{2}\theta^2\right)}{\partial\theta} = \theta = \mu$, the link function can be identified $g\big(\mathrm{E}(Y_i)\big) = g(\mu_i) = \mu_i = \sum_{k=0}^{p-1}\beta_k X_{ki}$ where $i = 1, \dots, \mathrm{n}$ or if y follow the normal distribution then in matrix notation can be written:

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta} \tag{3}$$

where $\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_{11} & \dots & \boldsymbol{x}_{p-1,1} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{1} & \boldsymbol{x}_{1n} & \dots & \boldsymbol{x}_{p-1,n} \end{bmatrix}$ is n × p matrix and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ is p × 1 vector.

If $Y_1, Y_2, \dots\dots, Y_n$ is a collection of random variables that independent and identically distributed and $Y_i \sim N(\mu, \sigma^2)$ then the likelihood function is:

$$\begin{aligned}
L(\mu_i|y_1, y_2, \dots\dots, y_n) &= \prod_{i=1}^{n} f(y_i|\mu_i,\sigma^2) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right\} \\
\log L(\mu_i|y_1, y_2, \dots\dots, y_n) &= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{2\sigma^2} \\
\log L\left(\sum_{k=0}^{p-1}\beta_k X_{ki}\,|y_1, y_2, \dots\dots, y_n\right) &= -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\sum_{k=0}^{p-1}\beta_k X_{ki})^2}{2\sigma^2}
\end{aligned} \tag{4}$$

This function is maximized to get an estimate for the β_k. This maximization process uses the first and second derivatives for each β_k. The specialty of exponential family distribution is that the first derivative and the second derivative of the function have the same form. The first derivative of a likelihood function is called a score function U and the second derivative is called Fisher-information matrix $\boldsymbol{J}$. The maximum likelihood estimator is to maximize the function $L(\mu_i|y_1, y_2, \ldots, y_n)$ or solve score function $\boldsymbol{U} = \boldsymbol{0}$ with:

$$\boldsymbol{U} = \begin{bmatrix} \frac{\partial log(L(\beta_k))}{\partial \beta_0} \\ \vdots \\ \frac{\partial log(L(\beta_k))}{\partial \beta_{p-1}} \end{bmatrix} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{E}(\boldsymbol{Y})) \tag{5}$$

where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{1} & \boldsymbol{x}_{11} & \ldots & \boldsymbol{x}_{p-1,1} \\ \vdots & \vdots & \ldots & \vdots \\ \boldsymbol{1} & \boldsymbol{x}_{1n} & \ldots & \boldsymbol{x}_{p-1,n} \end{bmatrix}$; $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix}$; $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{E}(\boldsymbol{Y}_1) \\ \vdots \\ \boldsymbol{E}(\boldsymbol{Y}_n) \end{bmatrix}$. Fisher-information matrix *is* obtained from second derivative of the function (4):

$$\mathbf{J} = -\begin{bmatrix} \frac{\partial^2 ln(L(\beta_k))}{\partial {\beta_0}^2} & \cdots & \frac{\partial^2 ln(L(\beta_k))}{\partial \beta_0 \partial \beta_{p-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 ln(L(\beta_k))}{\partial \beta_0 \partial \beta_{p-1}} & \cdots & \frac{\partial^2 ln(L(\beta_k))}{\partial {\beta_{p-1}}^2} \end{bmatrix} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} \tag{6}$$

where W is a diagonal matrix with $w_i = \mathrm{Var}(Y_i)$.

The parameter $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ can be estimated using a scoring algorithm:

$$\boldsymbol{\beta}^k = \boldsymbol{\beta}^{k-1} + (\boldsymbol{J}^{-1})^{k-1}\boldsymbol{U}^{k-1}$$
$$\boldsymbol{\beta}^k = \boldsymbol{\beta}^{k-1} + (\boldsymbol{X}'\boldsymbol{W}^{k-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{\mu}^{k-1}) \tag{7}$$

The Invers of Fisher information matrix is a covariance matrix. The elements on the diagonal of the covariance matrix are variance of $\widehat{\boldsymbol{\beta}}$. Standard errors of $\widehat{\boldsymbol{\beta}}$ are obtained from the root element in the diagonal inverse of the Fisher information matrix. The standard error can be used to perform a Wald test, testing if the estimated beta value is significantly different from zero. For $H_0: \beta = \delta$ then the Wald statistic takes the following form:

$$z = \frac{\beta_k - \delta}{Se(\widehat{\boldsymbol{\beta}})} \text{ dengan } Se(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} Se(\hat{\beta}_0) \\ \vdots \\ Se(\hat{\beta}_{p-1}) \end{bmatrix} \tag{8}$$

z^2 approximated by the chi-squared distribution with degrees of freedom 1.

2.2 GLMM

GLMM is an extension to the GLM in which the linear predictor contains random effects and fixed effects. GLMM explicitly include the cluster in the model and describe within-cluster effects. In GLMM, conditional distribution of response variable y_{ij} for j^{th} unit within i^{th} cluster with the random effect $\boldsymbol{u}_i$ is assumed independent and follows exponential family distributions (Handayani et al 2017). GLMM has form:

$$g\left(\mathrm{E}(y_{ij}|\boldsymbol{u}_i)\right) = \sum_{k=0}^{p-1}\beta_k X_{kij} + z_{ij}\boldsymbol{u}_i, \quad i = 1, \ldots, m\,;\, j = 1, \ldots, n_i\,;\, k = 0, \ldots, p-1 \tag{9}$$

where i show the sample order; j denotes i is nested in the suppopulation/ cluster; and k denotes the order of explanatory variables. The constant β_k is the fixed effect of the explanatory variables X_{kij} and

$\{\boldsymbol{u}_i\}$ are random effects which are assumed to have the distribution of $N(\mathbf{0}, \boldsymbol{\sigma}_u^2\boldsymbol{I})$. For y follow the normal distribution with the identity of link function, model (9) can be written in matrix notation:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} \tag{10}$$

where $\boldsymbol{\beta}$ is a p × 1 vector of fixed effects; $\boldsymbol{X}$ is a n × p matrix of explanatory variables. Suppose there are $(p-1)$ explanatory variables and $\boldsymbol{x}_{ij}$ is a vector of fixed effect then $\boldsymbol{x}_{ij} = \left(1, x_{1ij}, x_{2ij}, \ldots, x_{(p-1)ij}\right)$. $\boldsymbol{u}$ is a q × 1 random vector which follow $N(\mathbf{0}, \boldsymbol{\sigma}_u^2\boldsymbol{I})$; $\boldsymbol{Z}$ is a n × q matrix for the random effects; and $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{\sigma}_e^2\boldsymbol{I})$; $\boldsymbol{u}$ and $\boldsymbol{e}$ are independent. The expected value of $\boldsymbol{y}$ is $E(\boldsymbol{y}) = \boldsymbol{X\beta} + \boldsymbol{Zu}$ and the variance of $\boldsymbol{y}$ is $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{Z}^T\mathrm{Var}(\boldsymbol{u})\boldsymbol{Z} + \mathrm{Var}(\boldsymbol{e}) = \boldsymbol{Z}^T\boldsymbol{\sigma}_u^2\boldsymbol{Z} + \boldsymbol{\sigma}_e^2\boldsymbol{I} = \boldsymbol{\Omega}$.

In practice, the variance components σ_u^2 and σ_e^2 are unknown, so they must be estimated based on empirical data. The most used method in estimating the variance matrix in a linear mixed model is the maximum likelihood (ML) method and the restricted maximum likelihood (REML). The REML method produces an unbias estimator, while ML method produces a biased estimator. The breakdown in REML criteria is very complicated which includes the maximization of the likelihood functions of a linear combination of elements $\boldsymbol{y}$ that do not depend on $\boldsymbol{\beta}$. The criteria functions of REML (Verbyla 2019) are:

$$l_{REML}(\boldsymbol{\Omega}) = -\frac{p}{2}log(2\pi) + \frac{1}{2}\,log\left|\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{X}\right| + l_p(\boldsymbol{\Omega}) \tag{11}$$

where $l_p(\boldsymbol{\Omega}) = -\frac{n}{2}\,log(2\pi) - \frac{1}{2}\{log|\boldsymbol{\Omega}| + \mathbf{y}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\left(\mathrm{I} - \mathbf{X}[\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{X}]^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\right)\mathbf{y}\}$.

Variance components will be estimated by the REML method. The parameter estimator in equation (11) cannot be solved partially because each equation still contains other parameters so that the equation can only be solved simultaneously using iteration. Estimation of σ_u^2 and σ_e^2 with REML method is done by Fisher scoring iteration. In REML, $\boldsymbol{\Omega}$ is estimated until iteration $(k+1)$:

$$\boldsymbol{\Omega}^{(k+1)} = \boldsymbol{\Omega}^{(k)} + (F_{REML(\Omega^{(k)})})^{-1}\left(\frac{\partial lnL(\Omega^{(k)})}{\partial \Omega_p}\right) \tag{12}$$

Estimation of the variance component is obtained when iteration at (12) converges. After $\widehat{\sigma_u^2}$ and $\widehat{\sigma_e^2}$ is obtained using the REML method and $\widehat{\boldsymbol{\Omega}} = \boldsymbol{Z}^T\widehat{\boldsymbol{\sigma}_u^2}\boldsymbol{Z} + \widehat{\boldsymbol{\sigma}_e^2}\boldsymbol{I}$, estimators of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are obtained by the formula:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\,\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{y} \text{ and } \widehat{\boldsymbol{u}} = \widehat{\boldsymbol{\sigma}_u^2}\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}). \tag{13}$$

2.3 Geo-GLMM

Geo-GLMM is a combination between the GLMM and Geoadditive model. Suppose $\boldsymbol{S}_i = (s_{i1}, s_{i2})^T$ is the longitude and latitude coordinates at the i-th location, $i = 1, \ldots, n$. Let $\boldsymbol{Y}_i$ be the response variable and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{i2})^T$ is the explanatory variable at the location of $\boldsymbol{S}_i$. We assume that the probability density $(Y \mid x, s)$ belongs to the exponential family with $\mu(\mathbf{x}, \mathbf{s})$ is modeled by the link function $g(.)$ in the following additive form:

$$g\{\mu(\mathbf{x}, \mathbf{s})\} = \sum_{k=0}^{p-1}\beta_k x_k + \alpha(\boldsymbol{s}) \tag{14}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^T$ is an unknown univariate smooth function and $\alpha(.)$ is an unknown bivariate smooth function. If $\mathrm{var}(Y|\mathbf{X} = \mathbf{x}, \mathbf{S} = \mathbf{s}) = \sigma^2 V\{\mu(\mathbf{x}, \mathbf{s})\}$, then estimation of the mean can be achieved by replacing the conditional log-likelihood function $\log\{f_{Y|\mathbf{X},\mathbf{S}}(y|\boldsymbol{x}, \boldsymbol{s})\}$ with the quasi-likelihood function $l(\vartheta, y)$, which satisfies $\nabla_\vartheta l(\vartheta, y) = \frac{y-\vartheta}{\sigma^2 V(\vartheta)}$. This estimation method is based on a nonparametric quasi-likelihood approach (Yu et al 2019).

Suppose the dataset consists of C subpopulations and the c-th subpopulations can be described by the Gaussian distribution (normal) with the mean μ_c and the variance σ_c^2, $c \in \{1, \ldots, C\}$ where C indicates the number of clusters/ subpopulations. When $C = 1$, the model is a normal GLMM with a random Gaussian effect. When C >1, the model is called GLMM with Gaussian mixture random effects (Pan et al 2019).

Suppose $\alpha(.)$ is additive Gaussian Process (GP), then the function α is the sum of k regression functions that are controlled by the parameter set $\emptyset = (\emptyset_1, \dots, \emptyset_k)^T$.

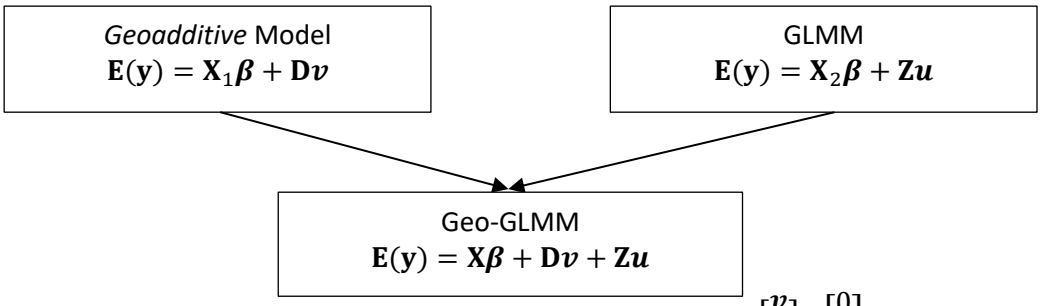
$$\alpha(s_i) = \emptyset_1 f_1(s_1) + \cdots + \emptyset_k f_k(s_k) \tag{15}$$

The Gaussian process provides flexible priors for each component function in $\{f_l, l = 1, \dots, k\}$ where $f_l \sim GP(0, c_l)$ with $c_l(s, s') = \exp\left\{-\sum_{j=1}^{p} K_{ij}\left(s_j - s_j'\right)^2\right\}$. The detailed explanation for GP can be seen in Vo and Pati (2017). Model (14) can be written in the form of an additive model:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{v} + \boldsymbol{e} \tag{16}$$

where $\mathrm{E}\begin{bmatrix} v \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathrm{Cov}\begin{bmatrix} v \\ e \end{bmatrix} = \begin{bmatrix} \sigma_v^2 \mathbf{I}_k & 0 \\ 0 & \sigma_e^2 \mathbf{I}_{K_e} \end{bmatrix}$, $\mathbf{X} = [1, \boldsymbol{x}_{p-1i}]_{1 \le i \le n,}$ $\boldsymbol{\beta} = [\beta_0, \beta_{p-1}]$, $\boldsymbol{v} = [\emptyset_1^s, \dots, \emptyset_k^s]$, $\mathbf{D}$ is $n \times k$ matrix.

Model (10) and (16) can be combined into a Geo-GLMM, so we obtain the following combination of Geoadditive and GLMM (Geo-GLMM models):



(17)

where $\mathbf{X}_1 = [1, \boldsymbol{s}_{ij}^{\mathrm{T}}]_{1 \le i \le n}$, $\mathbf{X}_2 = [1, \mathbf{x}_{ij}^{\mathrm{T}}]_{1 \le i \le n}$, $\mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2]$, $\mathrm{E}\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{u} \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, and

$\mathrm{Cov}\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{u} \\ e \end{bmatrix} = \begin{bmatrix} \sigma_v^2 \mathbf{I}_k & 0 & 0 \\ 0 & \sigma_u^2 \mathbf{I}_i & 0 \\ 0 & 0 & \sigma_e^2 \mathbf{I}_n \end{bmatrix}$.

Furthermore, unknown components of variance can be estimated using REML so that we obtain $\widehat{\sigma_v^2}$, $\widehat{\sigma_u^2}$, and $\widehat{\sigma_e^2}$. The estimated covariance matrix of $\boldsymbol{y}$ is $\widehat{\boldsymbol{\Omega}} = \widehat{\sigma_v^2}\,\mathbf{Z}\mathbf{Z}^{\mathrm{T}} + \widehat{\sigma_u^2}\mathbf{D}\mathbf{D}^{\mathrm{T}} + \widehat{\sigma_e^2}\mathbf{I}_n$ and the estimator for $\boldsymbol{\beta}$, $\boldsymbol{v}$, and $\boldsymbol{u}$ is:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{y} \\ \widehat{\boldsymbol{v}} &= \widehat{\sigma_v^2}\mathbf{D}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ \widehat{\boldsymbol{u}} &= \widehat{\sigma_u^2}\,\mathbf{Z}^{\mathrm{T}}\widehat{\boldsymbol{\Omega}}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \end{aligned} \tag{18}$$

To estimate the response variable can be calculated by the formula:

$$\widehat{\boldsymbol{y}}_i = \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_i\widehat{\boldsymbol{v}} + d_i\widehat{\boldsymbol{u}} \tag{19}$$

The Geo-GLMM model with interaction is multiplying between the covariates $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ then insert in the column matrix element $\mathbf{X}$ in model 17.

2.4 Model Comparison

Dataset is divided into training and testing data. The comparisons between training and testing are built at 90:10 percent; 85:15 percent; 80:20 percent; and 75:25 percent. The models are built based on training data. Then we estimate unhulled rice weight per $2.5 \times 2.5\ m^2$ using explanatory variables in the testing data. Then, the best model is chosen based on the lowest RMSE (root mean square error) and the highest correlation between actual and predicted data. RMSE is obtained based on the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n\_testing}[y_i - \hat{y}_i]^2}{n\_testing}} \tag{20}$$

The correlation used is the Pearson Product Moment correlation to determine the degree of linear relationship between actual and predicted data, and it is calculated using the formula:

$$r_{y_i,\hat{y}_i} = \frac{n\sum_i^{n\_testing} y_i\hat{y}_i - \left(\sum_i^{n\_testing} y_i\right)\left(\sum_i^{n\_testing} \hat{y}_i\right)}{\sqrt{\left[n\sum_{i=1}^{n\_testing} {y_i}^2 - \left(\sum_i^{n\_testing} y_i\right)^2\right]\left[n\sum_{i=1}^{n\_testing} {\hat{y}_i}^2 - \left(\sum_i^{n\_testing} \hat{y}_i\right)^2\right]}} \tag{21}$$

where $-1 < r_{y_i,\hat{y}_i} < 1$. The relationship between y_i and $\hat{y}_i$ can be tested:

$$t_{hit} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

with null hypothesis $H_0: r = 0$ (there is no correlation between y_i and $\hat{y}_i$) and $H_0: r \neq 0$ (there is a correlation between y_i and $\hat{y}_i$). H_0 is rejected if $t_{hit} > t_{table}$ with $n - 2$ degrees of freedom. Their relationship between y_i and $\hat{y}_i$ is targeted to be more than 0.80.

2.5 The Stages of Analysis

The stages of analysis are divided into three stages. First, build the model using all the data and compare the results of the parameter estimation between GLM, GLMM, and Geo-GLMM. Second, the dataset is divided into training and testing data taken randomly. Traning data is used to build the GLM, GLMM, and Geo-GLMM models, and testing data is used to see the goodness of fit. The percentage of training and testing data is set at 90:10 percent; 85:15 percent; 80:20 percent; and 75:25 percent. Third, the simulation data was generated and repeated 100 times. Furthermore, modeling with simulation data is performed to compare the performance between models. Modeling use R software with the mgcv library version: 1.8-31 by Wood (2019). Flowchart about the method can be seen in Figure 1.

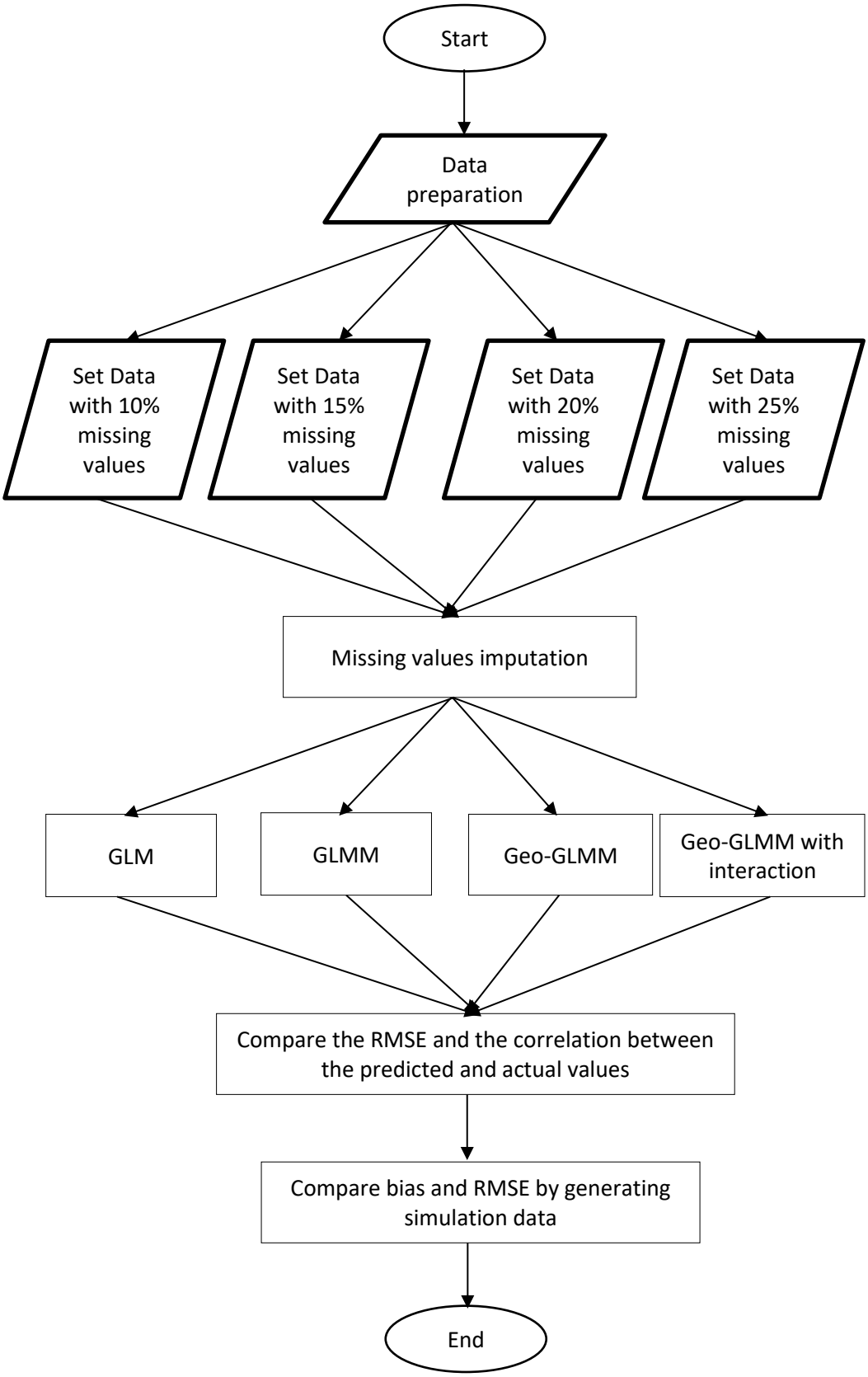


Figure 1. Flow chart comparing performance of GLM, GLMM, Geo-GLMM, and Geo-GLMM with interaction

3. An Empirical Study

This study was conducted in Central Kalimantan-Indonesia based on the Crop Cutting Survey 2019. Central Kalimantan Province was chosen as the research location because there were several sample locations with difficult accessibility, so some samples plots were not successfully obtained. This caused in missing data. The imputation process will be carried out using the explanatory variables on questionnaire of the Crop Cutting Survey to handle the missing data. The distribution of the coordinates of the sample points can be seen in Figure 2.

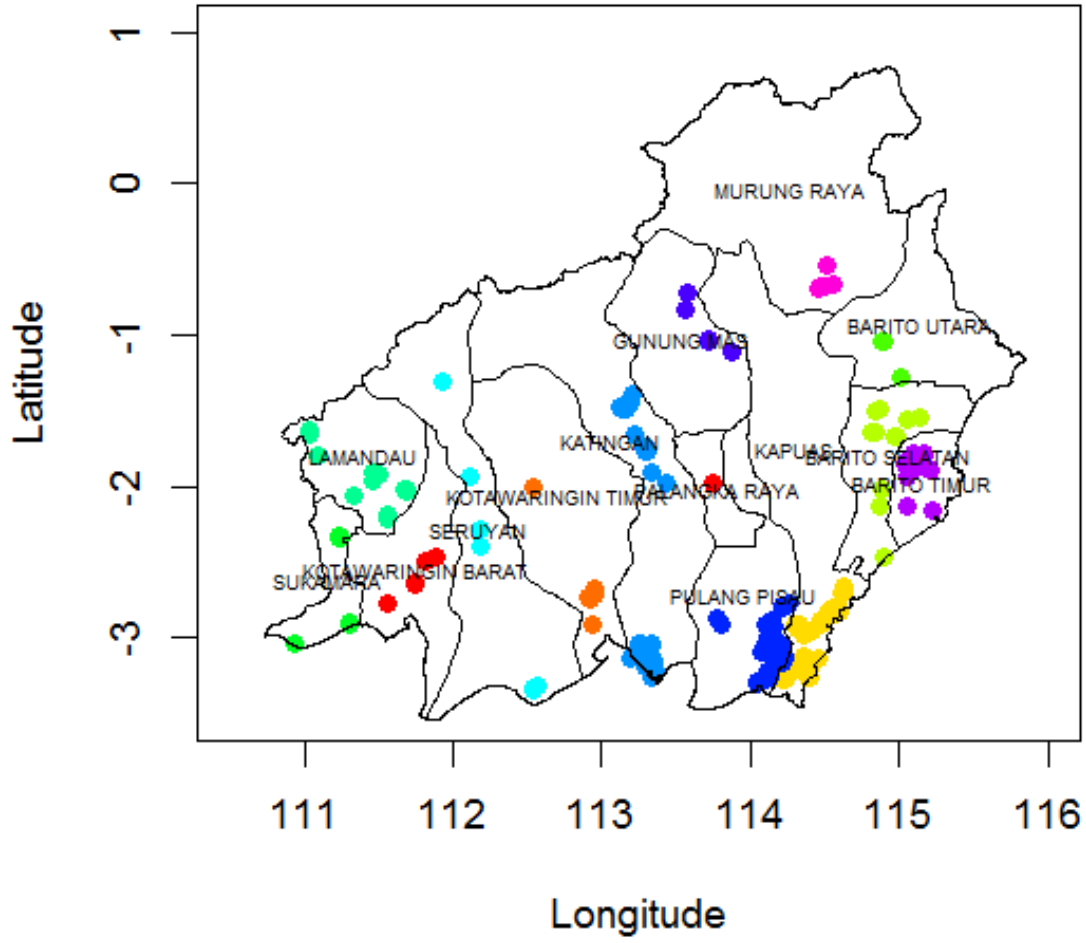


Figure 2. Map of Central Kalimantan Province and Coordinate Point Samples of Crop Cutting Survey 2019

Figure 2 shows the distribution of sample locations in the Crop Cutting Survey in 14 Regency in Central Kalimantan - Indonesia. The vertical axis shows the latitude and the horizontal axis shows the longitude. Astronomically, Central Kalimantan Province is located from 0^0 44′ 55″ (north latitude) to 3^0 47′ 70″ (south latitude), and it is located at longitude east from 110^0 43′ 19″ to 115^0 47′ 36″. As can be seen from Figure 2 that sample plots have spread across each regency. The variables chosen for modeling can be seen in Table 1.

Table 1. Variables used for modeling

Variables	Detail of Variables	Explanation	Question in the Questionnaire
y_i	Weight of harvested dry rice	kg per $2.5 \times 2.5\ m^2$.	Question 701
x_{1i}	Types of Plants	1. Lowland Sawah field 2. Upland field	Question 113
x_{2i}	Method to plant	1. Monoculture 2. Intercropping	Question 605
x_{3i}	Planting system	1. Jajar Legowo 2. Not Jajar Legowo	Question 606a
x_{4i}	government assistance or not	1. Government assistance 2. Non government assistance	Question 607
x_{5i}	Seed varieties	1. Hybrid 2. Non-hybrid	Question 609
x_{6i}	The amount of urea fertilizer used	kg per $10000\ m^2$.	Q610.1 divided by Q604 times $10000\ m^2$
x_{7i}	The amount of TSP/ SP36 fertilizer used	kg per $10000\ m^2$.	Q610.2 divided by Q604 times $10000\ m^2$
x_{8i}	The amount of KCL fertilizer used	kg per $10000\ m^2$.	Q610.3 divided by Q604 times $10000\ m^2$
x_{9i}	The amount of NPK/ Compound fertilizer used	kg per $10000\ m^2$.	Q610.4 divided by Q604 times $10000\ m^2$
x_{10i}	The amount of organic solid / compost fertilizer used	kg per $10000\ m^2$.	Q610.5 divided by Q604 times $10000\ m^2$
x_{11i}	The number of clumps in the plots	Clumps	Question 702
x_{12i}	Pest attack	1. Attacked 2. Not attacked	Q804b (the 1,2 recoding becomes attacked and 3,4 is not attacked)
x_{13i}	Affected by climate (flood and or drought)	1. Affected 2. Not Affected	Question 805b
Regency	Regency cluster	Used for GLMM modeling	Question 102
Long	Longitude	Used for Geo-GLMM modeling	Question 303. Long
Lat	Latitude		Question 303. Lat

As can be seen from Table 1 that y_i is the predicted variable by utilizing additional information of explanatory variables from x_{1i} to x_{13i}. The GLM, GLMM, and Geo-GLMM models have succeeded in identifying factors that have significant and not significant effects on the unhulled rice weight. Then, GLMM used to predict y_i by adding a random regency cluster effect. Next, Geo-GLMM used to predict y_i by adding information on the coordinates. The rice productivity between one region and another is different due to several factors. These factors are used to estimate the unhulled rice weight in $2.5 \times 2.5\ m^2$.

3.1 Estimation of GLM, GLMM, and Geo-GLMM parameters

Estimation of parameters uses all data. It can be identified from the estimation results, which variables have significant effects or not to the response variables. A comparison of the estimated parameters between GLM, GLMM, and Geo-GLMM can be seen in Table 2.

Table 2. Comparison of parameter estimation results and standard errors (se) between GLM, GLMM, and Geo-GLMM

Variables	GLM		GLMM		Geo-GLMM		
	$\hat{\beta}_k$	se	$\hat{\beta}_k$	se	$\hat{\beta}_k$	se	p-value
Intersep	2.2315	0.1641	2.2171	0.1944	2.3296	0.1752	0.000*
x_1	-0.2095	0.1116	-0.1963	0.1109	-0.1032	0.1280	0.4204
x_2	-0.5320	0.2003	-0.5005	0.1995	-0.3316	0.1884	0.0791
x_3	-0.5147	0.0830	-0.4339	0.0866	-0.1933	0.1035	0.0625
x_4	0.2059	0.0747	0.1270	0.0806	0.0337	0.0840	0.6888
x_5	-0.5031	0.1103	-0.5950	0.1083	-0.4619	0.0999	0.0000*
x_6	-0.0001	0.0004	-0.0012	0.0005	-0.0013	0.0005	0.0031*
x_7	0.0021	0.0006	0.0022	0.0006	0.0019	0.0006	0.0032*
x_8	-0.0031	0.0025	-0.0032	0.0024	-0.0018	0.0021	0.3888
x_9	0.0021	0.0004	0.0019	0.0004	0.0012	0.0003	0.0003*
x_{10}	0.0006	0.0005	0.0005	0.0005	0.0008	0.0005	0.0907
x_{11}	0.0026	0.0009	0.0028	0.0009	0.0021	0.0010	0.0407*
x_{12}	0.2672	0.0659	0.2733	0.0633	0.1332	0.0596	0.0259*
x_{13}	0.1993	0.0693	0.1895	0.0683	0.2122	0.0640	0.0010*

Table 2 shows that modeling using GLM, GLMM, and Geo-GLMM gives the similar results of parameters estimation. Variables that significantly influence the rice productivity are x_5, x_6, x_7, x_9, x_{11}, x_{12}, and x_{13} (seed varieties, amount of urea fertilizer, amount of TSP/ SP36 fertilizer, amount of NPK/ compound fertilizer, the number of clumps in the sample plots, the pest attack, the climate impact (flood and or drought). Hybrid varieties produce better rice productivity than non-hybrid varieties. The addition of TSP / SP36 and NPK / compound fertilizers tends to increase rice productivity but not for urea fertilizer. The more clumps in the plots of sample, the heavier unhulled rice weight produced. Pest attack and climate impacts also have a significant effect on decreasing rice productivity. The area affected by climate change will reduce its productivity. For other variables, it does not have a significant effect on the high or low unhulled rice weight in the plots of sample.

3.2 The proposed model

The application of urea and KCL fertilizers (x_6 and x_8) in Table 2 give a negative effect to rice productivity at land types in Central Kalimantan. This indicates that there is an interaction between the two types of fertilizer for land types in Central Kalimantan. We will enter interaction between x_6 and x_8. The type of fertilizer that has a positive effect (x_7, x_9, x_{10}) will be combined into one variable, ie. fertilizer. Variables that did not have significant effect are removed from the model one by one, so the parsimony model is obtained. The proposed model has explanatory variables, namely variety, fertilizer, and fertilizer interaction. This model is called the Geo-GLMM model with interaction where the smooth base used is the Gaussian process (GP). The results of parameter estimation are presented in Table 3.

Table 3. Estimating the parameters of the Geo-GLMM with interaction

Variables	Geo-GLMM with interaction		
	$\hat{\beta}_k$	se	p-value
Intersep	2.5680	0.1008	0.0000*
variety (x_5)	-0.4632	0.0985	0.0000*
fertilizer ($x_7 + x_9 + x_{10}$)	0.0013	0.0003	0.0000*
x_6	-0.0013	0.0004	0.0027*
x_8	-0.0122	0.0049	0.0126*

fertilizer interaction ($x_6 : x_8$)	0.0001	0.0001	0.0106*
α(Long, Lat)	Geoadditive effect		0.0000*

It can be seen from Table 3 that hybrid varieties are better than non-hybrid varieties in increasing the grain weight. TSP/ SP36, NPK, organic fertilizers and the interaction between urea and KCL fertilizers have a positive significant effect on the weight gain. The Geo-GLMM with interaction gives a good result. This is indicated by the high positive correlation between y actual and $\hat{y}$ (y predicted). The pattern of linear relationship between actual y and $\hat{y}$ Geo-GLMM with interaction can be seen in Figure 3.

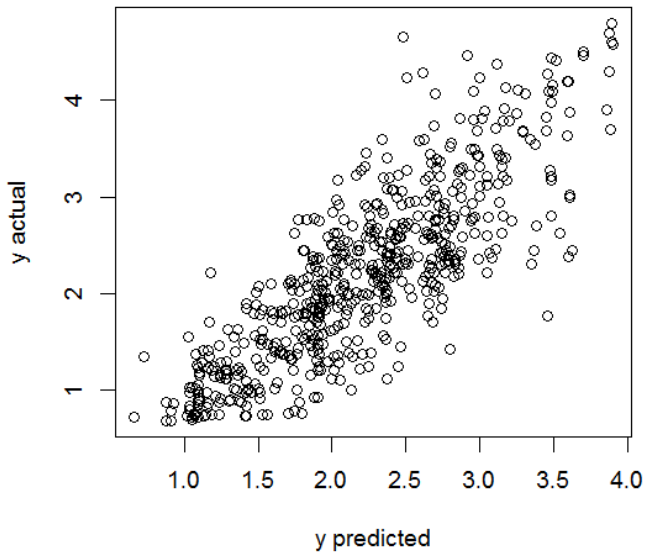


Figure 3. Pattern of linear relationship between y actual and $\hat{y}$ Geo-GLMM with interaction

Figure 3 shows that there is a strong positive relationship between y actual and $\hat{y}$ from Geo-GLMM with interaction. The Pearson correlation value with 95% confidence level is between 0.80 to 0.85 with very small p-value (< 2.2e-16). This result shows that the Geo-GLMM with interaction gives a good prediction result. The performance of this model will be clearly seen when predicting the actual value of y from the randomly selected data. The location of the planting factor also determines the weight of the grain. The pattern of non-linear relationships between coordinate points and response variables can be seen in Figure 3. Overall, the factors that significantly affect rice productivity are seed varieties, fertilizer used, and planting locations.

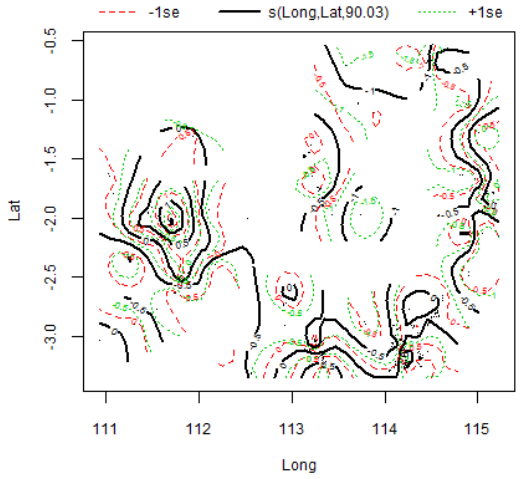


Figure 4. Patterns of non-linear relationships between coordinate points and response variable

3.3 Comparison of RMSE and correlation

The dataset is divided into training and testing data by comparison of 90:10, 85:15, 80:20 and 75:25. Testing data are taken randomly from the 100% dataset. Modeling uses training data. RMSE and correlation are obtained by comparing between testing (actual) data and prediction data. The comparison of RMSE between models can be seen in Table 4.

Table 4. Comparison of RMSE between GLM, GLMM, Geo-GLMM, and Geo-GLMM with interaction

Models	10% testing data	15% testing data	20% testing data	25% testing data	Average
GLM	0.91	0.83	0.85	0.82	0.85
GLMM	0.86	0.79	0.81	0.79	0.81
Geo-GLMM	0.68	0.63	0.67	0.66	0.66
Geo-GLMM with interaction	0.64	0.60	0.62	0.62	0.62

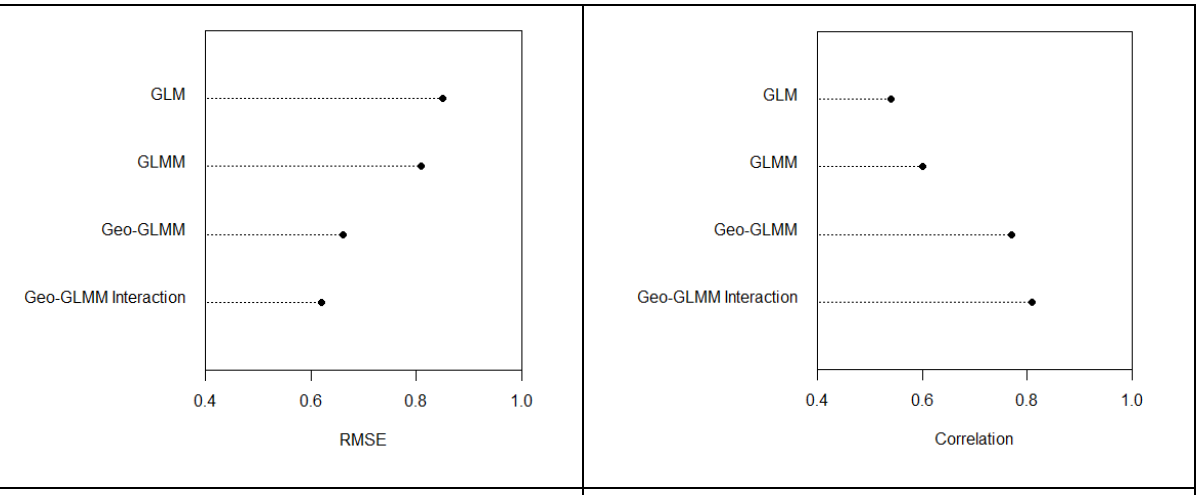
It can be seen from Table 4 that the geo-GLMM with interaction can reduce RMSE in all data testing groups. RMSE using the GLM model of 0.85 and dropped to 0.81 when using the GLMM and then dropped again to 0.66 when using the Geo-GLMM model. When we included the fertilizer interaction factor into the model and reducing several explanatory variables, the RMSE value drops to 0.62. Comparison of correlations between the four models can be seen in Table 5.

Table 5. Comparison of y_i and $\hat{y}_i$ correlations between GLM, GLMM, Geo-GLMM, and Geo-GLMM with interaction

Models	10% testing data	15% testing data	20% testing data	25% testing data	Average
GLM	0.59	0.56	0.50	0.53	0.54
GLMM	0.65	0.62	0.57	0.54	0.60
Geo-GLMM	0.81	0.79	0.74	0.74	0.77
Geo-GLMM with interaction	0.84	0.82	0.79	0.79	0.81

Table 5 shows that the GLMM improved the GLM because it was able to improve the prediction performance. Pearson's correlation increased from 0.54 by GLM to 0.60 by GLMM.

Furthermore, the Geo-GLMM is able to improve the GLMM because it is able to increase the correlation from 0.60 to 0.77. By adding fertilizer interaction factors and reducing several explanatory variables, the Geo-GLMM with interaction was able to improve the Geo-GLMM, for it increase the correlation between the actual and prediction data from 0.77 to 0.81. The change in average RMSE and the correlation between actual and prediction data can be seen in Figure 5.



a) RMSE | b) Correlation

Figure 5. Comparison of average RMSE and correlation between actual and prediction data

As can be seen from Figure 5 that the Geo-GLMM with interaction has the lowest RMSE and the highest correlation. The smooth base used in the Geo-GLMM with interaction is the Gaussian process. Thus, the Geo-GLMM with interaction is the best model for predicting the unhulled rice weight compared to the other three models.

4. Simulation Study

Simulation data is generated based on information of empirical data in Crop Cutting Survey in Central Kalimantan 2019 with 14 groups (regency). The β_i parameters are set at the beginning. Then, we generate data for each variable, namely rice varieties, fertilizer (SP36 + NPK + Organic), urea fertilizer, KCL fertilizer, and latitude-longitude coordinates.

Variables of rice varieties (hybrid and non-hybrid) were generated from the distribution of Binom (n, 1, p = 0.9). The proportion of hybrid varieties is 10 percent based on empirical data. Fertilizer variables ($x_7 + x_9 + x_{10}$) are generated from the normal distribution $\mathrm{N}(\mathrm{n}, 130.9, \sqrt{128.08})$. The urea variable ($x_6$) is generated from the normal distribution $\mathrm{N}(\mathrm{n}, 85.70, \sqrt{83.86})$. The KCL variable ($x_8$) is generated from the normal distribution $\mathrm{N}(\mathrm{n}, 10.18, \sqrt{12.92})$.

The coordinate points are generated from the Uniform distribution. The longitude is generated from the Uniform distribution (n, 110.9,115.2) and latitude are generated from the Uniform distribution (n, 110.9,115.2). The coordinate data generation is based on the coordinate location in Central Kalimantan. Unhulled rice weight (y) is obtained from the addition of $\alpha+\beta_1$*varietas + β_2*pupuk+ β_3*urea+ β_4*kcl+ δ*(urea*kcl) + γ_1*Long+ γ_2*Lat+ random effect area + error. The random effect area is generated from the Normal distribution (rep (rnorm (nclust, 0, sds), each = nplot). The error are generated from the Normal distribution (rnorm (nkab * nplot, 0, sd)).

The dataset is generated several times, then parameters are estimated using GLM, GLMM, Geo-GLMM, and Geo-GLMM with interaction. Comparison of average estimation results between true value (parameter) and parameter estimation can be seen in Table 6.

Table 6. Comparison of the average of parameters estimation of simulation data with 100 replications between GLM, GLMM, Geo-GLMM, and Geo-GLMM with interaction

Models	Variables	Parameters	Average = $\frac{1}{r}\sum_{j=1}^{r}\hat{\beta}_{ij}$
GLM	Varieties	$\beta_1 = -0.4632$	-0.4664
	Fertilizer *)	$\beta_2 = 0.0013$	0.0015
	Urea	$\beta_3 = -0.0013$	0.0143
	KCL	$\beta_4 = -0.0122$	0.1151
GLMM	Varieties	$\beta_1 = -0.4632$	-0.4674
	Fertilizer *)	$\beta_2 = 0.0013$	0.0016
	Urea	$\beta_3 = -0.0013$	0.0144
	KCL	$\beta_4 = -0.0122$	0.1152
Geo-GLMM	Varieties	$\beta_1 = -0.4632$	-0.4601
	Fertilizer *)	$\beta_2 = 0.0013$	0.0012
	Urea	$\beta_3 = -0.0013$	0.0142
	KCL	$\beta_4 = -0.0122$	0.1158
Geo-GLMM with interaction	Varieties	$\beta_1 = -0.4632$	-0.4610
	Fertilizer *)	$\beta_2 = 0.0013$	0.0013
	Urea	$\beta_3 = -0.0013$	-0.0014
	KCL	$\beta_4 = -0.0122$	-0.0153

*) Fertilizer =SP36+NPK+Organic

It can be seen from Table 7 that the Geo-GLMM with interaction is able to produce an average of estimated parameters that have the same direction as the true value (parameters). The other three models produce an average of estimated parameters in the wrong direction due to the data generated by the interaction between x_6 and x_8. Comparison of the bias estimation between the four models can be seen in Table 7.

Table 7. Comparison of the bias estimation of simulation data with 100 replications between GLM, GLMM, Geo-GLMM, and Geo-GLMM with interaction

Models	$\left\|\text{Bias}(\hat{\beta}_i)\right\| = \left\|E(\hat{\beta}_i) - \beta_i\right\|$			
	β_1	β_2	β_3	β_4
GLM	0.0032	0.0002	0.0156	0.1273
GLMM	0.0042	0.0003	0.0157	0.1274
Geo-GLMM	0.0031	0.0001	0.0155	0.1280
Geo-GLMM with interaction	0.0022	0.0000	0.0001	0.0031

As can be seen from Table 7 that the Geo-GLMM with interaction has a lower bias than the bias estimation of GLM, GLMM, and Geo-GLMM. Bias generated from the Geo-GLMM with interaction close to 0. The RMSE comparison is presented in Table 8.

Table 8. Comparison of RMSE estimation of simulation data with 100 replications between GLM, GLMM, Geo-GLMM, and Geo-GLMM with interaction

Models	RMSE of $(\hat{\beta}_i) = \sqrt{\left[\frac{1}{r}\sum_{j=1}^{r}(\hat{\beta}_{ij} - \beta_i)^2\right]}$			
	β_1	β_2	β_3	β_4
GLM	0.1050	0.0024	0.0159	0.1276
GLMM	0.0915	0.0024	0.0160	0.1276

Geo-GLMM	0.0715	0.0019	0.0157	0.1281
Geo-GLMM with interaction	0.0848	0.0019	0.0067	0.0545

It can be seen from Table 8 that the Geo-GLMM with interaction produces an RMSE value which tends to be smaller than the other three models. Comparison of mean, bias, and RMSE values between different numbers of examples can be seen in Table 9.

Table 9. Comparison of mean, bias, and RMSE estimates of parameters in the simulation data according to the number of samples

Number of samples	Measures	β_1	β_2	β_3	β_4
n=140	True	-0.4632	0.0013	-0.0013	-0.0122
	Mean	-0.5416	0.0052	-0.0027	-0.0259
	Bias	0.0784	0.0039	0.0014	0.0137
	RMSE	0.1298	0.0052	0.0127	0.1155
n=700	True	-0.4632	0.0013	-0.0013	-0.0122
	Mean	-0.4680	0.0021	0.0010	0.0002
	Bias	0.0049	0.0008	0.0023	0.0124
	RMSE	0.0596	0.0021	0.0059	0.0403
n=1400	True	-0.4632	0.0013	-0.0013	-0.0122
	Mean	-0.4655	0.0016	-0.0005	-0.0067
	Bias	0.0023	0.0003	0.0008	0.0055
	RMSE	0.0262	0.0011	0.0047	0.0348

It can be seen in Table 10 that the larger the sample, the smaller the estimated bias and RMSE. This informations show that the parameter estimation of Geo-GLMM with interaction is consistent estimator.

5. Conclussion

Factors that significantly affected rice productivity were the seed varieties, TSP/ SP36 fertilizer, NPK/ compound fertilizer, urea fertilizer, organic fertilizer, the number of clumps, pest attacks, and climate impacts (flood and or drought). The explanatory variables was selected to estimate the unhulled rice weight of $2.5 \times 2.5\ m^2$ in the Crop Cutting Survey in Central Kalimantan, namely varieties, fertilizers, interactions between urea and KCL, and coordinate points. The Geo-GLMM model by adding fertilizer interaction was able to estimate the unhulled rice weight better than the GLM, GLMM and Geo-GLMM without fertilizer interaction. The Geo-GLMM with interaction can increase the correlation between actual and predicted data from 0.55 by GLM to 0.81.

The simulation showed that the Geo-GLMM with interaction is the best model because it produces a smaller bias and RMSE. Simulation results with different numbers of samples provide information that the greater the number of samples, the smaller the value of bias and RMSE. This showed that the estimated parameters are consistent estimator. Therefore, it is recommended that the BPS officer keep to interview farmers when failing to take a sample of the plots. The unhulled rice weight can be predicted using other variables in the survey questionnaire. For further research, Geo-GLMM results can be compared with other machine learning.

6. Acknowledgments

The authors thank the Editor and the Associate Editor for their helpful comments and constructive suggestions, which lead to improvement in the quality of the paper. This research was supported by

the BPS-Statistics Indonesia Doctoral Scholarship Program and Department of statistics, IPB University.

References

[1] Ardiansyah M, Djuraidah A, and Kurnia A. 2018. Pendugaan Produktivitas Padi di Tingkat Kecamatan Menggunakan Geoadditive Small Area Model. *Jurnal Penelitian Pertanian Tanaman Pangan.* Vol. 2 No. 2 Agustus 2018: p101-110. DOI: http//dx.doi.org/10.21082/jpptp.v2n2.2018.p101-110.

[2] Ardiansyah M and Tofri Y. 2019. Perbandingan Data Produktivitas Padi Antara Hasil Wawancara Pascapanen dengan Data Survei Ubinan di Kalimantan Tengah. *Jurnal Penelitian Pertanian Tanaman Pangan.* Vol. 3 No. 1 April 2019: p17-22. DOI: http//dx.doi.org/10.21082/jpptp.v3n1.2019.p17-22.

[3] Ardiansyah M, Buana WP, and Kurnia A. (2020). Prediksi Produktivitas Padi Melalui Survei Ubinan Menggunakan Model Linier dan Quantile Regression Forest. Jurnal Penelitian Pertanian Tanaman Pangan. Vol. 4 No. 3 Desember 2020: 135-144. DOI: http//dx.doi.org/10.21082/jpptp.v4n3.2020.p135-144.

[4] Handayani D, Notodiputro KA, Sadik K, and Kurnia A. 2017. A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM). AIP Conference Proceedings 1827, 020033 (2017); https://doi.org/10.1063/1.4979449

[5] Muleia R, Boothe M , Loquiha O, Aerts M, and Faes C. 2020. Spatial Distribution of HIV Prevalence among Young People in Mozambique. *International Journal of Environmental Research and Public Health.* 2020, 17, 885: 1-20. doi:10.3390/ijerph17030885

[6] Pan L, Li Y, He K, Li Y, and Li Y. 2019. Generalized linear mixed models with Gaussian mixture random effects: Inference and application. *Journal of Multivariate Analysis (ELSEVIER).* 175 (2020): p1-19. https://doi.org/10.1016/j.jmva.2019.104555.

[7] Verbyla AP. 2019. A note on model selection using information criteria for general linear models estimated using REML. *Australian & New Zealand Journal of Statistics*. 61(1): p39–50. doi: 10.1111/anzs.12254.

[8] Vo G and Pati D. 2017. Sparse Additive Gaussian Process with Soft Interaction. *Open Journal of Statistics*. 7. 567-588. https://doi.org/10.4236/ojs.2017.74039.

[9] Wood S. 2019. R Package 'mgcv' Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. Version: 1.8-31. Published: 9 November 2019. URL https:// https://cran.r-project.org/web/packages/mgcv/mgcv.pdf.

[10] Yu S, Wang G, Wang L, Liu C, and Yang L. 2019. Estimation and Inference for Generalized Geoadditive Models. *Journal of the American Statistical Association (Taylor & Francis Group)*. 0(0): p1–14. DOI: 10.1080/0162145