Fitting the rice production model using generalized additive mixed model and generalized estimating equation with shiny web application

A S Darmawan¹, D Anggraeni¹, and I M Tirta¹

¹Department of Mathematics, University of Jember, Indonesia

Email: Antondarmawan717@gmail.com, Tirtaimade@gmail.com

Abstract. Banyuwangi is a district in East Java province in Indonesia which is known as the Rice Barn of East Java. Banyuwangi region is facing challenges such as population growth, increased consumption of rice, reduced agricultural land because there are several agricultural land used as non-agricultural land. Because of there are production it is not enough from objects with just one period, therefore the use of data panels is one solution because it considers the time period and objects needed to determine the factors that influence rice production. In this study the analysis to model panel data from rice production in Banyuwangi was Generalized Additive Mixed Model (GAMM) and Generalized Estimating Equation GEE. In general GAMM as a mixed model, contains random effects so that this model is very useful when inference about individual differences is a major interest. While GEE is a model used to provide the population averaged effect. GEE also has the feature of smoothing natural cubic spline and B-spline so that it can be called GEE (+ NS / BS) which aims to improve the goodness of fitting a model. In this study, we want to see the effect of the total rainfall, harvest area, and population density on rice production. We also develop GAMM in the form of web interface to make users easier to analyze data using GAMM through http://statslab-rshiny.fmipa.unej.ac.id/RProg/MSD/.

1. Introduction
Banyuwangi is a district in East Java province in Indonesia. Banyuwangi has the largest area in East Java, with an area of 5,782.50 m². It is known as Rice Barn of East Java and it has 24 sub-districts. Although Banyuwangi district is one of the largest rice contributors in East Java, it does not mean that there has never been a problem. In the period 2010 - 2011 rice production decreased by 8.71% [1]. Banyuwangi region is facing challenges such as population growth, increased consumption of rice, reduced agricultural land because there are several agricultural land used as non-agricultural land. The data of this study are panel data of rice production in 2013-2017. Research related to Banyuwangi rice production has been carried out by [2] using the Generalized Additive Model Location, Scale and Shape (GAMLSS). Research results [2] with Banyuwangi rice production data have several conclusions: 1. the location model of each predictor: rainfall, agricultural land, students from high school have a significant effect on rice production, 2. The most suitable distribution is Exponential Skew Power (SEP). GAMLSS model [2] was modeled used cross section model data. Different from previous research that uses cross section data, in this study will consider multi-year (panel data) so that the model obtained is more comprehensive. Analysis of time periods can be approached by random effects and correlation structures. In this study, GAMM as a representative of random effect and GEE
as representative of structure correlation are used to get the best model in panel data of rice production.

GAMM is an extension of the Generalized Linear Mixed Model (GLMM). This relative new model class uses additional non parametric functions to model the covariate effect, while calculating overdispersion and correlation by adding random effects to additional predictors. In general GAMM as a mixed model, contains random effects so that this model is very useful when inference about individual differences is a major interest [3]. Another model that is able to analyze panel data is GEE. GEE is one of the statistical methods used to analyze correlated data because of repeated measurements. GEE is a model used to provide the population averaged effect [4] in [3]. In its application, GAMM which contains relatively many types of smoothing, GEE can accommodate natural cubic spline and B-spline smoothing of the model to improve the goodness of the fitting so that it can be called GEE (+ NS / BS).

In this study, GAMM and GEE (+ NS / BS) were used to determine the best effect of the object under study on rice production and then compare it. Moreover, this study provides an opportunity to learn more about GAMM and GEE using R shiny packages. With R Shiny, it is expected to provide for users to analyze GAMM and GEE.

2. Generalized Additive Mixed Model (GAMM)

Suppose that observations of the $i$th of $n$ units consist of an outcome variable $y_i$ and $p$ covariates $x_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}^T$ associated with fixed effect and vector $q \times 1$ from $z_i$ that is associated with random effects. Given a $q \times 1$ vector $b$ of random effects, the observations $y_i$ are assumed to be conditionally independent with means $E(y_i|b) = \mu_i^b$ and variances $\text{var}(y_i|b) = \sigma_i^{-2}v(\mu_i^b)$ where $v(.)$ is a specified variance function, $m_i$ is prior weight and follow a GAM

$$g(\mu_i^b) = \beta_0 + \beta_1(x_{i1}) + \cdots + \beta_p(x_{ip}) + z_i^Tb$$

(1)

where $g(.)$ is monotonic differentiable link function, $f_j(.)$ is a centered-differentiable smooth function, the random effects $(b)$ are assumed to be distributed as $N(0,D(\theta))$ and $\theta$ is a $c \times 1$ vector of variance components. A key feature of the GAMM is that additive non parametric functions are used to model correlation between observations [5].

3. Generalized Estimating Equation (GEE)

There are $n_i$ measurements on subject $i$ and $\sum_{i=1}^n n_i$ total measurements. Correlated data are modeled using the same link function and linear predictor setup (systematic component) as the independence case. The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modeled. Let the vector of measurement on the $i^{th}$ subject be $Y_i = \mu_i = [Y_{i1}, \ldots, Y_{i(n_i)}]$ with corresponding vector of means $\mu_i = [\mu_{i1}, \ldots, \mu_{in_i}]$ and let $V_i$ be an estimate of the covariance matrix of $Y_i$. The GEE estimate approach $\beta$ is an extension of the independence estimating equation to correlated data and is given by [6]

$$\sum_{i=1}^n \frac{\partial}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta)) = 0$$

(2)

4. GEE order 2

GEE2 is a set of estimation equations that are introduced to produce estimates that are consistent with both regression and correlation parameters. This estimation equation is called the second order general equation. Generally, this method has the form $\frac{(r-1)}{2}$ order 2 in the form of $z_i = (z_{i1}, z_{i2}, \ldots, z_{i(r-1)r})$ that can be modeled using $l_i(\alpha) = E(z_i)$ and the corresponding link function [7]. The odds ratio coefficient can be used as a measure of correlation. For example, in continuous data analysis,
regression estimates and correlation parameters can be found by completing the system of equations (3)

\[
U(\alpha) = \sum_{i=1}^{n} \left[ \begin{array}{c} \frac{\partial^2 \alpha}{\partial \beta} \\ \frac{\partial^2 \alpha}{\partial \gamma} \\ \frac{\partial^2 \alpha}{\partial \delta} \\ \frac{\partial^2 \alpha}{\partial \gamma} \\ \frac{\partial^2 \alpha}{\partial \delta} \\ \frac{\partial^2 \alpha}{\partial \gamma} \end{array} \right] \left[ \begin{array}{c} V(\gamma) \\ \text{cov}(\gamma, \gamma) \\ \text{cov}(\gamma, \gamma) \\ \text{cov}(\gamma, \gamma) \\ \text{cov}(\gamma, \gamma) \end{array} \right] \left[ \begin{array}{c} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \end{array} \right] \right]
\]

Where \(l_i(\alpha)\) are correlation coefficients for repeated observations.

5. R Shiny
R-Shiny is a module created by the Rstudio group that can be used to create a web-GUI (Graphical User Interface) menu that uses interactive graphical web interfaces that interact with R [8]. By using shiny, we can create 3 basic models of simulation packages, namely lab modules, instructional materials modules, and online data analysis packages based on [9]. The use of R-shiny has been used by [10] and [11] to create and develop data-based data analyst packages. Components in the R-Shiny have two groups, as follows:

a. User Interface.
   The user interface (UI) can be used as:
   1. The control panel is used to set data input, variables, models etc.
   2. Entry of input data (types of variables needed, selection of models, types, and statistical test)
   3. Presentation of output, output can be displayed in the form of graphs, numbers, and mathematical notation in the latex format.

b. Server
   The server is the center of the program which in its simulation works. Various results from the input data processed then immediately send the results to the output. This section has been supported by various procedures and data analysis which are generally available in the R package.

   With R-Shiny, we will make a prototype with a variety of model analysis options. Each option is used to help Indonesian researchers access R more easily.

6. Research Method
The data used in this study is rice production data in Banyuwangi Regency in 24 sub-districts for 5 years (2013 - 2017). Response variable (\(y\)) is rice production while predictor variables: total rainfall (\(x_1\)), harvest area (\(x_2\)), and population density (\(x_3\)). The steps for analyzing data with GAMM are as follows: 1. Input data, 2. Specifies a suitable distribution, 3. Determine the variables to be modeled parametric or non-parametric, 4. Significance test. The steps for analyzing data with GEE are as follows: 1. Input data, 2. Data exploration (selection of the general structure of working correlation matrix, selection of parametric or non-parametric modeled variables), 3. Estimating parameters, 4. Significance test, 5. Comparing the best model GAMM and GEE.

7. Result
7.1 Distribution Analysis
In this section, we explore the distribution that matches the response variable. In this study, a suitable distribution can be seen through the most appropriate form of the curve and the lowest AIC. In Figure 1, the distribution that is most suitable for data on rice production is shown.

![Figure 1. a. Normal (NO) distribution and b. Inverse Gaussian (IG) distribution](image-url)
In Figure 1, two distributions are relatively suitable for rice production data, but to ensure the best, AIC can be used as a parameter to choose the best distribution.

| Distribution       | AIC     |
|--------------------|---------|
| Normal             | 1003.256|
| Inverse Gaussian   | 1017.249|

According to Table 1, the best distribution used is a normal distribution that has the lowest AIC value. In this study, the selection of variables modeled with parametric and non-parametric will be conducted. In general, exploration of data can be seen in Figure 2.

![Figure 2. The relationship of all predictor variables and y](image)

In Figure 2, it can be seen that $x_1$ and $x_2$ have a fairly good relationship with $y$ but $x_3$ has an relationship that is not as good $x_1$ and $x_2$. The relation of $x_2$ and $y$ are relatively weak. Therefore, we have strong reasons to choose $x_1$ and $x_2$ as parametric modeled variables, while $x_3$ is modeled with non-parametric.

7.2 Fitting the GAMM Models

The selection of random effects tested is the district and year. The selection of random effects chosen is based on predictions that are more specific to each sub-district or per year. The concept is that random effects are more suitable to be modeled with sub-districts than years because if the year is used as a random effect only limited to 2013-2017 where the year has been missed. In Table 2, the best selection of GAMM models is displayed.

| Mod | Par | Non Par | Re   | AIC  |
|-----|-----|---------|------|------|
| A   | $x_1 + x_2$ | $ns(x_3,1)$ | District | 766.82 |
| B   | $x_1 + x_2$ | $s(x_3,4)$  | District | 772.50 |
| C   | $x_1 + x_2$ | $cs(x_3,2)$ | District | 772.48 |
| D   | $x_1 + x_2$ | $ns(x_3,1)$ | Year   | 780.57 |
| E   | $x_1 + x_2$ | $cc(x_3,2)$ | Year   | 788.79 |
| F   | $x_1 + x_2$ | $tp(x_3,1)$ | Year   | 788.66 |

In Table 2 it was found that model A is the best model than the other models because it has the lowest AIC value. The estimation of model A will be shown in Table 3.
Table 3. Estimating of Best GAMM Model (Model A)

| Parameters | Est | Std. Error | p-val | Status    |
|------------|-----|------------|-------|-----------|
| Intercept  | 3.032 | 1.922 | 0.117 | No Significant |
| $x_1$      | -0.001 | 0.001 | 0.014 | Significant |
| $x_2$      | 0.006 | 0.000 | $< 2e-16$ | Significant |
| ns($x_3, 1$) | -3.94 | 4.936 | 0.427 | No Significant |

From Table 3, it can be seen that the effect of total rainfall ($x_1$) and harvest area ($x_2$) of model A have a significant effect on rice production while population density ($x_3$) does not have a significant effect on rice production. In Table 3 also shows that the $r$-square of model A is 0.87422.

7.3 Comparison of fitting GEE1(+NS/BS) and GEE2(+NS/BS)

This study also modeled rice production data with the GEE model (+ NS / BS). Before modeling the production of rice with GEE (+ NS / BS), an analysis of the correlation structure that will be selected through the relationship between rice production in 2013-2017 is needed. The structure of the GEE correlation can be seen in Figure 3.

![Diagram Korelasai: Respon prod. padi, Klasfer x Tahun](image)

Figure 3. Correlation of rice production in 2013-2017

In Figure 3, it can be seen that it is shown that the correlation structure of rice production in Banyuwangi 2013-2017 is relatively constant and decreases so the correlation structure that needs to be considered is uniform (exchangeable) and unstructured.

Table 4. Fitting of GEE(+NS/BS)

| Mod | Par | Non par | Id | St.Corr    | QIC  |
|-----|-----|---------|----|------------|------|
| M   | $x_1 + x_2$ | ns($x_3, 1$) | District | independence | 141.3 |
| N   | $x_1 + x_2$ | ns($x_3, 1$) | District | Uniform | 139.2 |
In Table 4, the best fitting for the GEE1 (+ NS / BS) model is the model N because the model N has the lowest QIC value. The estimation of the model N will be shown in Table 5.

| Mod | Par | Non par | Id | St.Corr | QIC |
|-----|-----|---------|----|---------|-----|
| O   | $x_1 + x_2$ | ns($x_2,1$) | District | Unstructured | 147.5 |
| P   | $x_1 + x_2 + ns(x_2,1)$ | District | Independence | 145.9 |
| Q   | $x_1 + x_2 + ns(x_2,2)$ | District | Uniform | 146.5 |
| R   | $x_1 + x_2 + ns(x_2,1)$ | District | Unstructured | 145.9 |

Table 5. Estimating The Best GEE(+NS/BS) (Model N)

| Mean | Est     | P-value | Status   |
|------|---------|---------|----------|
| Intercept | 3.077806387 | 0.06509549 | Signikan |
| $x_1$ | -0.00115671 | 0.31355810 | No Significant |
| $x_2$ | 0.006112610 | 0.000000000 | Significant |
| ns($x_2,1$) | -3.96079531 | 0.06174061 | Significant |

| Scale | Estimate | P-value | Status   |
|-------|----------|---------|----------|
| Intercept | 29.67254 | 0.1113715 | No Significant |
| Correlation parameters | Estimate | P-value | Status   |
| Intercept | 0.288936 | 0.003468541 | Significant |

Rsq Value 0.8774

In table 5, it can be seen that in the best GEE (+ NS) model the influence of harvest area and population density has a significant influence on rice production but for the effect of total rainfall on rice production does not have a significant effect. The effect of the correlation parameters on the modeling is also significant, this means that the correlation given has a pretty good influence on the model In Table 6, the GEE2 (+ NS / BS) model will be shown through the model N.

Table 6. Estimating GEE2(+NS/BS) (Model N)

| Mean | Est     | P-value | Status   |
|------|---------|---------|----------|
| Intercept | 2.4282433197 | 0.05654995 | Significant |
| $x_1$ | -0.00058001 | 0.29386221 | No Significant |
| $x_2$ | 0.00614026 | 0.000000000 | Significant |
| ns($x_2,1$) | -2.52112489 | 0.61373640 | Significant |

| Scale | Estimate | P-value | Status   |
|-------|----------|---------|----------|
| Intercept | 4.97973348 | 0.3918871 | No Significant |
| $x_2$ | 0.03839751 | 0.2363757 | No Significant |

| Correlation parameters | Estimate | P-value | Status   |
|-------------------------|----------|---------|----------|
| Intercept | 0.3036056 | 0.00509992 | Significant |
In Table 6, GEE2 (+ NS) shows that the overall effect has the same effect except that the GEE2 model (+ NS) involves variance modeling, but it seems that variance modeling does not have a better effect so the best model for GEE is N by using estimates from GEE1 (+ NS). In addition, the table 5 also shows that the r-square of the model N is 0.8774.

7.4 R-Shiny Interface for GAMM

This application is expected to make users easier to analyze panel data using GAMM and GEE. This application can be accessed at http://statslab-rshiny.fmipa.unej.ac.id/RProg/MSD/.

In Figure 4, there are several important things in the initial display of this application, namely 1. Data Input. 2. Data Exploration. 3. GEE1 and GEE2. and 4. GAMM. In part 1, input data is used to enter user data. In part 2, Data Exploration is used to see linearity relationships between variables. In section 3, GEE1 and GEE2 have been developed in the web form by [9]. In section 4, GAMM is web development from before and this section is the focus of this research. The main part of this web is in the input formula section, in this section the user can choose a simple or complex menu. In general, the input formula is illustrated in Figure 5.

In figure 5, users who use a simple input formula can analyze GAMM with the available menu, while users who use complex input formula can analyze GAMM through the script. Display of complex input formulas can be seen in Figure 6.
In Figure 6, the choice of a simple or complex input formula has several output options: General results, GAM attachment, MER fitting (Random Effect Model), Rsq, AIC and plot. In this study also found differences in GAMM and GEE features, in general differences can be seen in Table 7.

| Feature Opt | GAMM | GEE |
|-------------|------|-----|
| Var. Respon | ✓    | ✓   |
| Var. Pred   | ✓    | ✓   |
| Cluster     | hierarki, non-hierarki | hierarki, non-hierarki |
| Dist        | Continuous (normal, gamma, inverse gaussian) | Continuous (normal, gamma, inverse gaussian) |
|             | - discrete (poisson, binomial) | - discrete (poisson, binomial) |
| Data Exploration | ✓ | ✓ |
| Str. Correlation | ar1, exchangeable, independence, unstructured | ar1, exchangeable, independence, unstructured |
| Repeat measurement | ✓ | ✓ |
| Random effect | ✓ | ✓ |
| Smoothing   | s, ns, bs, cs, cr, cc, cp, ds, gp, mrf, re, sf, so, sos, tp, ds, rs | ns, bs |
In Table 7, it was found that GAMM and GEE have advantages and disadvantages. The GAMM feature does not apply the correlation structure and repeat measurement, but GAMM applies the random effects feature and has more smoothing variations so that it can explore the model further. Unlike GAMM, GEE applies a correlation structure and repeat measurement, but smoothing used is limited to NS and BS. The similarity of both is the distribution used and can model scale (linear) parameters. In addition, both of them also have data exploration features and at the analysis stage, both require response variables and predictor variables.

8. Conclusion
In general, this study concluded as follows:
   a. Each of the best GAMM and GEE models (+ NS / BS) is model A and model N using GEE1 (NS). Model A which represents the best GAMM model states that variable: total rainfall, harvest area has a significant effect on rice production while population density has no significant effect on rice production. In contrast to model A, the model N based on GEE1 (+ NS) has the result that the total rainfall has no significant effect on rice production, while the harvest area and population density have a significant effect on rice production and the effect of the correlation parameters on the modeling is also significant. To choose the best rice production model of the two models, R square can be used where the value of R square of the model N is higher than R square GAMM so that the model N is the best model.
   b. GAMM features include: selection of response and predictor variables, clusters, distribution, data exploration, random effects and smoothness that are quite varied. The simple formula input is still limited to the features provided. Solution to overcome the limitations of features, users can choose complex input formulas so users can explore GAMM further. The web can be accessed at http://statslab-rshiny.fmipa.unej.ac.id/RProg/MSD/.

Acknowledgments
We gratefully acknowledge the support from anonymous reviewer for suggestions to improve this paper.

References
[1] Statistics of Banyuwangi Regency 2012 Banyuwangi Regency In Figures 2012 Banyuwangi: BPS of Banyuwangi Regency
[2] Darmawan A S, Anggraeni D and Tirta, I M Application of generalized additive model location, scale and shape (GAMLSS) for rice production in Banyuwangi regency IOP Conference Series: Journal of Physics: Conf. Series 1538 012032

[3] Hu F B, Goldberg J, Hadeker D, Flay B R and Pentz M A 1998 Comparison of Population-Averaged and Subject-Specific Approaches for Analyzing Repeated Binary Outcomes American Journal of Epidemiology 147(7):694-703

[4] Liang K and Zeger S 1986 Longitudinal data analysis for discrete and continuous outcomes Biometrics. 42:121-130

[5] Lin X and Zhang D 1999 Inference in generalized additive mixed models by using smoothing splines JRSSB. 55(2):381-400

[6] Darko I O, Adu I K and Frempong N K Application of Generalized Estimating Equation (GEE) Model on Students’ Academic Performance Applied Mathematical Sciences 8(68):3359-3374

[7] Zayeri F, Bardineshin S, Akbarzadeh-Bagheban A R, Adel M and Asgari S 2012 First and Second Order Generalized Estimating Equations and Their Application in Analyzing Longitudinal Microleakage Data Journal of Islamic Dental Association of IRAN 24 (4): 215-224

[8] Tirta I M 2015 Pengembangan Analisis Respon Item Interaktif Online Menggunakan R untuk Respon Dikotomus dengan Model Logistik (1-PL, 2-PL 3-PL) Seminar Nasional Pendidikan Matematika, FKIP Universitas Jember Jember: Universitas Jember

[9] Tirta I M 2014 Pengembangan E-Modul Statistika Terintegrasi dan Dinamik dengan R-shiny dan mathJax Prosiding Seminar Nasional Matematika, Universitas Jember

[10] Tirta I M, Anggraeni D and Pandutama M 2017 Online Statistical Modeling (regression analysis) for Independent Responses IOP conference Series: Journal of Physics: Conf. Series 855 012054

[11] Tirta I M and Anggraeni D 2018 The Development of Web-based Graphical User Interface for Unified Modeling Data with Multi (Correlated) Responses. IOP Conference Series: Journal of Physics: Conf. Series 1008 012003