Application of generalized additive model location, scale and shape (GAMLSS) for rice production in Banyuwangi regency

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 one of the regencies in East Java that has great potential for rice production. In terms of rice production, Banyuwangi occupies the top 5 positions in all East Java districts. The focus of this research is to get a suitable model for rice production and find out how the influence of the predictor variables on the response variable. The models used in this case is Generalized Additive Model Location, Scale and Shape (GAMLSS) with RS algorithm. GAMLSS is an extension of the GAM and GLM. GAMLSS is also a semi-parametric which has advantages, namely (1) the distribution used is wider, because it contains the distribution of the exponential family and other distributions and (2) all of the parameters (μ, σ, ν, τ) can be modeled depending on the distribution used. The analyzed model contains rice production as a response variable and predictor variables used: total rainfall, harvested area, population in district, students of senior high schools, and health centers. In this study, the best model will be defined by the smallest AIC and SIC from comparison results of all possibilities in vary smoothings, parameters, and degrees of freedom. Furthermore, the best distribution for rice production will also known in this study.

1. Introduction

Banyuwangi Regency is a regency of East Java province in Indonesia. Banyuwangi Regency is the largest regency in East Java, with an area of 5,782.50 km² and Banyuwangi Regency having 24 districts. East Java is one of the provinces nicknamed Lumbung Padi (Rice Barn). Banyuwangi is one of the regencies in East Java that has great potential for rice production. In terms of rice production, Banyuwangi occupies the top 5 positions in all East Java districts[1]. Banyuwangi is one of the regencies known as the Lumbung Padi Jawa Timur (Rice Barn of East Java). Banyuwangi as an East Java rice barn provides many benefits for the country of Indonesia nationally, especially in the food sector. Although Banyuwangi Regency 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 has decreased by 8.71% [2]. The decline caused many interpretations. One interpretation is that every year agricultural land is thought to have experienced a reduction in land, for example, used as residential areas and other uses. The risk of food production (rice production) will decrease in proportion to the reduction in agricultural land. Thus, an effort is needed to control the conversion of food agricultural land to maintain food security. In the sector of agriculture, Banyuwangi faces a challenge, namely the increasing number of population followed by an increasing in consumption of basic needs especially rice, but many agricultural lands have been converted into non-agricultural land. Therefore various efforts are made to increase rice production. Previous research on food crop agriculture was carried out by [3] using the Generalized Linear Model (GLM) model. The results of its research found that
GLM has yielded good fitness on the trajectories of crops production in most cases. In research by [3], it still does not apply to model additive (non-parametric). We will use Generalized Additive Model Location, Scale and Shape (GAMLSS) to model rice production. Besides containing parametric models, GAMLSS also contains non-parametric models so that this model is also called a semi-parametric model and has many alternative distribution options that contain 𝜇, 𝜎, 𝛿, 𝜆 parameters. The focus of this research is to get a suitable model for rice production and find out how the predictor variables influence the response variable.

2. Generalized Additive Model Location, Scale and Shape (GAMLSS)

GAMLSS is a general frame work for regression models where we assume that response variable depends on many predictor variables. Regression can be linear, non linear, or smooth non-parametric. Form example, Linear Model (LM) is the simplest regression model and assumed that there response variable has a normal distribution. In reality, many response variable are not normally distributed. GLM is introduced by [4]. GLM is a monotonic function of the mean, called the linear predictor is a linear function of the predictor variables. Not only that but also GLM accommodates exponential family distribution. However GLM still has a disadvantage that GLM is still not able to model non-parametric relationship. To accommodate non-parametric relationship Generalized Additive Model (GAM) was introduced by [5] and popularized by [6], has made the smoothing techniques within a regression framework available to a wide range of practitioners. GAM can be called a semi-parametric model, because it contains parametric and non-parametric components. GAMLSS is an extension of GAM. GAMLSS is also a semi-parametric which has advantages, namely (1) the distribution used is wider because it contains the distribution of the exponential family and other distributions and (2) all of the parameters (μ, σ, ν, τ) can be modeled depending on the distribution used. The details of the distributions and parameterizations used in GAMLSS models can be found in [7],[8], and [9].

GAMLSS assumes independent observations 𝑦𝑖 = 1, …, 𝑛 with probability density function

\[ f(𝑦|𝜃) \] where \( 𝑣 = ( 𝜇, 𝜎, 𝜋, τ ) \) corresponds to the parameter vector. The first two elements \( 𝜇 \) and \( 𝜎 \) are the location and scale parameters, and the others are shape parameters. In GAMLSS the exponential family distribution assumption is relaxed and replaced by the general distribution family including highly skew distributions [8]. Let \( 𝑦 = ( 𝑦_1, 𝑦_2, …, 𝑦_𝑛 ) \) be the 𝑛 length of vector response variable. For \( k = 1, 2, 3, 4 \) let \( g_k(\.) \) be known montonic link functions relating to the distribution parameters to explanatory variables by

\[
    g_1(\mu) = \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} Z_{j1} \gamma_{j1} \\
    g_2(\sigma) = \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} Z_{j2} \gamma_{j2} \\
    g_3(\nu) = \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} Z_{j3} \gamma_{j3} \\
    g_4(\tau) = \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} Z_{j4} \gamma_{j4}
\]

Where \( g_k(\.) \) is a known montonic link function for \( k = 1, 2, 3, 4 \); \( \mu, \sigma, \nu, \tau \) and \( \eta \) are dimensional vectors \( X_k \) are known design matrices of order \( n \times J_k \) associated with fixed effects \( \beta_k \) of \( J_k \times 1 \); and \( Z_{jk} \) are known design matrices of order \( n \times q_{jk} \) associated with random effects \( \gamma_{jk} \) of \( q_{jk} \times 1 \) with multivariate normal distribution. The quantity \( J_k \) represents the number of covariates used in the fixed effects of \( \eta_k \), while \( J_k \) represents the number of random effects in \( \eta_k \). The model that is given in (1) to (4) can be summarized in a compact form as follows:

\[
    g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} (Z_{jk} \gamma_{jk})
\]
3. The RS Algorithm

One algorithm that can be used in the GAMLSS model is the Rigby & Stasinopoulos (RS) algorithm. Essentially the RS algorithm has an outer cycle which maximizes the penalized likelihood with respect to $\mathbf{\beta}_k$ and $\mathbf{\gamma}_{jk}$, for $j = 1, \ldots, J_k$, in the model successively for each $\theta_k$ in turn, for $k = 1, \ldots, p$ [7]. At each calculation in the algorithm the current updated values of all the quantities are used. The RS algorithm is as follows.

Step 1: start- initialize fitted values $\theta_k^{(1,1)}$ and random effects $\mathbf{\gamma}_{jk}$, for $j = 1, \ldots, J_k$ and $k = 1, \ldots, p$.

Evaluate the initial linear predictors $\eta_k^{1,1} = g_k(\theta_k^{1,1})$ for $k = 1, 2, \ldots, p$.

Step 2: start the outer cycle $r = 1, 2, \ldots$ until convergence. For $k = 1, 2, \ldots, p$:

a. start the inner cycle $i = 1, 2, \ldots$ until convergence
   
   (i) evaluate the current $\mathbf{u}_k^{(r,i)}$, $\mathbf{W}_{kk}^{r,i}$ and $\mathbf{z}_k^{r,i}$;
   
   (ii) start the back fitting cycle $m = 1, 2, \ldots$ until convergence;
   
   (iii) regress the current partial residual $se_{tk}^{r,m} = \mathbf{z}_k^{r,i} - \sum_{j=1}^{J_k} \mathbf{z}_{jk} \mathbf{y}_{jk}^{r,m}$ against design matrix $\mathbf{X}_k$, using the iterative weights $\mathbf{W}_{kk}^{r,i}$ to obtain the updated parameter estimates $\hat{\beta}_k^{r,i,m+1}$
   
   (iv) for $j = 1, \ldots, J_k$ smooth the partial residuals $\mathbf{e}_{jk}^{r,i,m} = \mathbf{z}_k^{r,i} - \mathbf{X}_k \hat{\beta}_k^{r,i,m+1} - \sum_{t=1, t \neq j}^{J_k} \mathbf{z}_{tk} \mathbf{y}_{tk}^{r,i,c}$ using the shrinking (smoothing) matrix $S_{jk}$ given by equation $S_{jk} = Z_{jk}^T (W_{kk} Z_{jk} + G_{jk}^{-1}) Z_{jk} W_{kk}$ to obtain the updated additive predictor term $Z_{jk} \mathbf{y}_{jk}^{r,i,m+1}$;
   
   (v) end the back fitting cycle, on convergence of $\hat{\beta}_k^{r,i,m}$ and $Z_{jk} \mathbf{y}_{jk}^{r,i,c}$ and set $\beta_k^{r,i+1} = \beta_k^{r,i}$ and $\mathbf{y}_{jk}^{r,i+1} = \mathbf{y}_{jk}^{r,i}$ for $j = 1, \ldots, J_k$ and otherwise update $m$ and continue the backfitting cycle;
   
   (vi) calculate the updated $\eta_k^{r,i+1}$ and $\theta_k^{r,i+1}$;

b. end the inner cycle on convergence of $\hat{\beta}_k^{r,i}$ and the additive predictor terms $Z_{jk} \mathbf{y}_{jk}^{r,i}$. And set $\beta_k^{r+1,i} = \beta_k^{r,i}$, $\mathbf{y}_{jk}^{r+1,i} = \mathbf{y}_{jk}^{r,i}$ for $j = 1, 2, \ldots, J_k$. $\eta_k^{r+1,i} = \eta_k^{r,i}$ and $\theta_k^{r+1,i} = \theta_k^{r,i}$; otherwise update $i$ and continue the inner cycle

Step 3: update the value of $k$

Step 4: end the outer cycle if the change in the (penalized) likelihood is sufficiently small; otherwise update $r$ and continue the outer cycle.

4. Research Method

In this paper, we will make a model of rice production in Banyuwangi Regency through GAMLSS with the RS algorithm. Rice production ($y$) as a response variable and we consider predictor variables: total of rain fall ($x_1$), harvested area ($x_2$), population in district ($x_3$), students of senior high school ($x_4$), and health centers ($x_5$). The data is obtained from [10]. After data is ready, we have to know the type of response data used is discrete or continuous data. We do a comparison with several distributions that have the appropriate type of data so we get the most suitable distribution. We need to know whether the suitable distribution has 1,2,3 or 4 parameters. After that we choose which one is selected as a parametric model and non-parametric model in the predictor variables section. Predictor variables that are parametric and non-parametric model are combined so that they are referred to as semi-parametric models. Next we make comparisons to get the best model, with consideration of the smoothing used and provide a model of location ($\mu$), scale($\sigma$), and shape :skewness ($\psi$) and kurtosis ($\tau$). We need to compare each use of the distribution parameters of the model, for example a distribution has 4 parameters meaning that he can model location, scale, and shape so of course we do a comparison by modeling 1,2,3 to 4 parameters. The most suitable model after considerations is
called the best model. Significance test is performed on the best model to see whether each predictor variable has a significant relationship to the response variable on the best model. In general, GAMLSS modeling can be seen in Figure 1.

5. Result
In this section, we explore the distribution that matches the response variable. There are 2 ways to see the suitability of the distribution of the response variable by looking at the shape of the curve and using the Akaike Information Criteria (AIC) and the Schwarz Information Criteria (SIC). In Figure 2, we show 3 distributions that are most suitable for the response variable.
Figure 2. a. Skew Power Exponential (SEP), b. Weibull (WEI) and c. Gamma (GA) distribution

All graph of distribution in Figure 2 are good enough, but if we look at AIC and SIC in Table 1, we can choose SEP and WEI is the best distribution. SEP has the smallest AIC and the WEI has the smallest SIC. In this case we will create a model using these 2 distributions and then compare the AIC and SIC values again.

| Distribution | AIC      | SIC     |
|--------------|----------|---------|
| SEP          | 202.0064 | 206.7186|
| WEI          | 203.9516 | 206.3077|
| GA           | 204.3334 | 206.6895|

For non-parametric components, we explored relationship of y and all predictor variables ($x_1, x_2, x_3, x_4, x_5$) through scatterplot with smoothing, and we find the most appropriate is $x_4$. This consideration is by looking at the weak relationship between $y$ and $x_4$. In Figure 3, it is known that $x_4$ with linear regression (green) does not show a significant effect on $y$. Red curve is a graph that results from loess smoothing. If $x_4$ is forced to become a parametric model, of course the model obtained does not get better. This means that $x_1, x_2, x_3$, and $x_5$ are modeled parametrically.
We must remember that SEP has 4 distribution parameters, while WEI has 2 distribution parameters. In modeling non-parametric models in GAMLSS, we can select some smoothing like \( pb() \), \( cs() \), and \( lo() \) [9]. In Table 2, we get the best model that contains the SEP distribution, smoothing \( ps() \) with \( df = 1 \), and scale model modeled \( x_5 \), while shape \( (\nu \text{ and } \tau) \) models are modeled constant.

**Table 2.** Comparison of AIC and SIC in GAMLSS Estimation (RS algorithm)

| Model | Modeled by | AIC   | SIC    |
|-------|------------|-------|--------|
| SEP DISTRIBUTION | \( \mu \), \( \sigma \), \( \nu \), \( \tau \) |        |        |
| 1     | \( x_1, x_2, x_3, ps(x_4, df=1), x_5 \) | \( x_5 \) | -      | 40.3348 | 58.0057 |
| 2     | \( x_1, x_2, x_3, ps(x_4, df=3), x_5 \) | \( x_1 \) | -      | 40.6247 | 58.1058 |

| Model | Modeled by | AIC   | SIC    |
|-------|------------|-------|--------|
| WEI DISTRIBUTION | \( \mu \), \( \sigma\nu\tau \) |        |        |
| 3     | \( x_1, x_2, x_3, cs(x_4, df=4), x_5 \) | \( x_2 \) | -      | 106.0656 | 120.2025 |
| 4     | \( x_1, x_2, x_3, ps(x_4, df=4), x_5 \) | \( x_2 \) | -      | 108.1939 | 122.2806 |

In Table 3, we can see the estimation results and the p-value of model 1.
Table 3. The Best Model

| μ link function: identity | Estimate | P-value | Status |
|--------------------------|----------|---------|--------|
| Intercept                | -6.461   | 2×10^{-16} | Significant |
| 𝑥₁                       | -2.024e - 02 | 2×10^{-16} | Significant |
| 𝑥₂                       | 6.574e - 03 | 2×10^{-16} | Significant |
| 𝑥₃                       | -1.276e - 05 | 2×10^{-16} | Significant |
| ps(𝑥₄, df = 1)           | 1.994    | 2×10^{-16} | Significant |
| 𝑥₅                       | 3.123e - 03 | 2×10^{-16} | Significant |

| 𝜎 link function: log    | Estimate | P-value | Status |
|-------------------------|----------|---------|--------|
| Intercept               | -6.092   | 0.0226  | Significant |
| 𝑥₅                      | 1.716    | 0.2272  | No Significant |

| 𝜈 link function: identity | Estimate | P-value | Status |
|---------------------------|----------|---------|--------|
| Intercept                 | 0.06482  | 0.103   | No Significant |

| τ link function: log | Estimate | P-value | Status |
|---------------------|----------|---------|--------|
| Intercept           | -3.391   | 7.93×10^{-9} | Significant |

Thus, we can write the best model that is in the form:

\[
\mu = -6.461 + (-2.024e - 02)x₁ + (6.574e - 03)x₂ + (-1.276e - 05)x₃ + ps(x₄, df = 1) + (3.123e - 03)x₅
\]

\[
\log(\sigma) = -6.092 + 1.716x₅
\]

\[
\log(\tau) = -3.391
\]

6. Conclusion

This study concluded that the best results of GAMLSS modeling using the RS algorithm for rice production in Banyuwangi Regency are

\[
\mu = -6.461 + (-2.024e - 02)x₁ + (6.574e - 03)x₂ + (-1.276e - 05)x₃ + ps(x₄, df = 1) + (3.123e - 03)x₅
\]

\[
\log(\sigma) = -6.092 + 1.716x₅
\]

\[
\log(\tau) = -3.391
\]

These components are total rain fall (𝑥₁), harvested area (𝑥₂), population in district (𝑥₃), students of high school (𝑥₄), and health centers (𝑥₅). The distribution used in the best model is SEP. GAMLSS accommodates SEP, but SEP is not in GAM family, so this best model cannot be reduced to GAM model.

Acknowledgments

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

7. References

[1] Statistics of Jawa Timur Province 2017JawaTimurIn Figures 2017(Surabaya: BPS of Jawa Timur Province)

[2] Statistics of Banyuwangi Regency 2012Banyuwangi Regency In Figures 2012 (Banyuwangi: BPS of Banyuwangi Regency)
[3] Zhang W, Liu G and Bai C 2007 A forecast analysis on global production of staple crops (http://www.stats.gov.cn/english/icas/papers/P020071114315023592169.pdf)
[4] Nelder, J A and Wedderburn R W M 1972 Generalized linear models *Journal of the Royal Statistical Society, Series A* **135** 370-384
[5] Hastie T J and Tibshirani R J 1990 *Generalized additive models* (London: Chapman and Hall)
[6] Wood S N 2017 *Generalized additive models: An introduction with R* (London: Chapman and Hall)
[7] Rigby R A and Stasinopoulos D M 2005 Generalized additive models for location, scale and shape (GAMLSS)*Appl Statist* **54** 507-554
[8] Stasinopoulos D M and Rigby R A 2007 Generalized additive models for location scale and shape (GAMLSS) in R*Journal of Statistical Software* **23** 1-46
[9] Stasinopoulos M, Rigby B and Akantziliotou C 2008 *Instructions on how to use the gamlss package in R Second Edition* (http://www.gamlss.com/wp-content/uploads/2013/01/gamlss-manual.pdf)
[10] Statistics of Banyuwangi Regency 2017 *Banyuwangi Regency In Figures* 2017 (Banyuwangi: BPS of Banyuwangi Regency)