Comunicaciones en Estadística
Diciembre 2014, Vol. 1, No. 3

GAMLSS models applied in the treatment of agro-industrial waste

Modelos GAMLSS aplicados en el tratamiento de residuos agroindustriales

Freddy Hernández Barajas[1]
fhernanb@unal.edu.co

Mabel Torres Taborda[2]
mabel.torres@upb.edu.co

Lina Arteaga Valencia[3]
lina.bacteriologa@gmail.com

Cristina Castro Herazo[4]
cristina.castro@upb.edu.co

Abstract

In this paper, we present an application of GAMLSS (Generalized Additive Models for Location, Shape and Scale) to study bacterial cellulose production from agro-industrial waste. An experiment was conducted to research the effects of pH and cultivation time on bacterial cellulose yield obtained from discarded bananas. Several models were fitted to the collected data to determine an estimated expression for the mean and variance of bacterial cellulose yield. We found that the mean and variance of cellulose yield decrease as pH increases, while the opposite occurs as cultivation time increases.

***Key words*:** Gamma distribution, linear regression, parameter estimation.

Resumen

Abstract in the second language

***Palabras clave*:** Distribución gama, regresión lineal, estimación de parámetros.

1. Introduction

The problems of the massive exploitation of natural resources and environmental pollution have motivated the building of an economy based on renewable materials.

[1] Profesor asistente, Universidad Nacional de Colombia, Medellín
[2] Profesora asociada, Universidad Pontificia Bolivariana, Medellín
[3] Mágister en biotecnología
[4] Profesora asociada, Universidad Pontificia Bolivariana, Medellín

For this reason, polymers obtained from renewable resources such as polysaccharides, proteins, and lignin, among others, are attracting considerable attention (Jaramillo et al. 2013). It has been found that valuable products such as bacterial cellulose can be obtained from agro-industrial waste through suitable processing. Obtaining bacterial cellulose depends on, among other factors, pH and fermentation time, and therefore, it is important to determine the combination of these factors that maximizes the bacterial cellulose yield.

2. GAMLSS

Rigby & Stasinopoulos (2005) proposed the GAMLSS models (Generalized Additive Model for Location Scale and Shape), which assume that the response variables y_i (with $i = 1, \dots, n$) are independent with a probability density function $f(y_i \mid \boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i = (\mu_i, \sigma_i, \nu_i, \tau_i)^T$ corresponds to the parameter vector. The first two elements μ_i and σ_i are the location and scale parameters, and the others are shape parameters. GAMLSS models allow each parameter to be a function of a set of explanatory variables, and the distribution of random variable y_i is not limited to the exponential family (Stasinopoulos & Rigby 2007). GAMLSS models can be summarized as follows:

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1} \tag{1}$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2} \tag{2}$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3} \tag{3}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4} \tag{4}$$

where $g_k(\cdot)$ is a known monotonic link function for $k = 1, \dots, 4$; $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$ and $\boldsymbol{\eta}_k$ are n-dimensional vectors; $\boldsymbol{X}_k$ are known design matrices of order $n \times J_k^{'}$ associated with fixed effects $\boldsymbol{\beta}_k$ of $J_k^{'} \times 1$; and $\boldsymbol{Z}_{jk}$ are known design matrices of order $n \times q_{jk}$ associated with random effects $\boldsymbol{\gamma}_{jk}$ of $q_{jk} \times 1$ with multivariate normal distribution. The quantity $J_k^{'}$ represents the number of covariates used in the fixed effects of $\boldsymbol{\eta}_k$, while J_k represents the number of random effects in $\boldsymbol{\eta}_k$. The model given in (1) to (4) can be summarized in a compact form as follows:

$$g_k(\boldsymbol{\theta_k}) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk} \tag{5}$$

The GAMLSS model considers both continuous and discrete distributions with different parameterizations for the same distribution. The details of the distributions and parameterizations used in GAMLSS models can be found in Rigby & Stasinopoulos (2010, page 199). Another advantage of GAMLSS models is that these models allow the use of fixed effects, random effects and non-parametric smoothing functions to model all parameters of the assumed distribution for the response variable.

3. Experiment description

An experiment was conducted to study the effect of pH and cultivation time (days) on the production of bacterial cellulose using the microorganism *Gluconacetobacter medellinensis*. Each sample unit corresponded to 100 grams of banana peel, which was cut into smaller pieces and homogenized with 400 mL of water using a blender. This mixture was filtered using a cloth membrane. The juice obtained from each sample was analyzed to determine the pH. After completion of the fermentation time, the obtained bacterial cellulose membrane was removed and placed in a solution of KOH at 5 % (p/p) for 14 hours at a temperature between 28 and 30 degrees Celsius. The cellulose membranes were then washed successively with water until the pH was neutral, and the washed membranes were dried in a convection oven at 60 degrees Celsius for 24 hours and then at 105 degrees Celsius for 2 hours or until constant weight was reached. At the end of this process, the amount of bacterial cellulose was measured; see Figure 1.

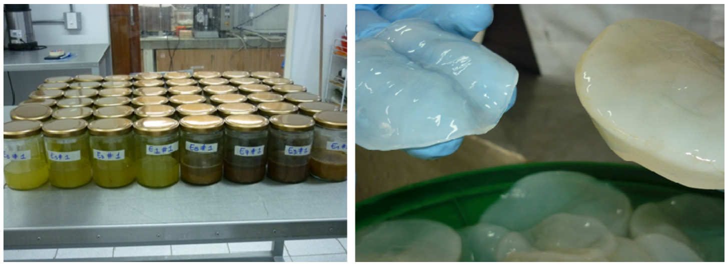

Figura 1: Experiment illustration. Left, unit samples. Right, cellulose membrane.

The response variable in the experiment was the bacterial cellulose yield. Figure 2 shows the density plot and boxplot for bacterial cellulose yield, revealing that the response variable is right-skewed with a minimum value of 0.0181, median of 0.0787, maximum of 0.5707 and 5 observations of 32 that appear to be outliers. For these reasons, it seems reasonable to use a skewed distribution to model the cellulose yield.

Figure 3 shows the scatterplot for bacterial cellulose yield, pH and cultivation time. We observe that the maximum bacterial cellulose yield was obtained at pH 3.5 with 13 days of cultivation; it was noted that the yield decreases with increasing pH and tends to increase with cultivation time.

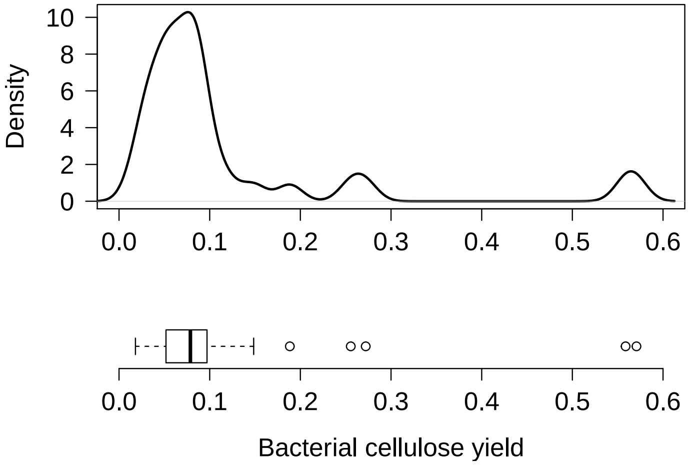

Figura 2: Density and boxplot for bacterial cellulose yield (g).

4. Results

In this section, we present the results of using GAMLSS models to explain bacterial cellulose yield (y) with the explanatory variables pH and cultivation time. In Table 1, we present the models considered: models 1 to 3 assume a response variable with normal distribution (only as a reference point), and models 4 to 10 consider asymmetric distributions for the response variable. The third column of the table shows the structure in GAMLSS syntax to model the μ and σ parameters of each distribution.

Tabla 1: *AIC* values for each fitted model.

Model	Distribution	Structure in GAMLSS syntax	*AIC*
1	Normal	`gamlss(y~pH+Time, family=NO())`	-46.8
2	Normal	`gamlss(y~pH+Time, sigma.fo=~pH+Time, family=NO())`	-109.8
3	Normal	`gamlss(y~pH*Time + I(pH^2) + I(Time^2), family=NO())`	-49.7
4	Gamma	`gamlss(y~pH+Time, family=GA())`	-102.2
5	Gamma	`gamlss(y~pH+Time, sigma.fo=~pH+Time, family=GA())`	-118.1
6	Gamma	`gamlss(y~pH+Time, sigma.fo=~pH, family=GA())`	-119.8
7	log-Normal	`gamlss(y~pH+Time, family=LNO())`	-95.9
8	log-Normal	`gamlss(y~pH+Time, sigma.fo=~pH+Time, family=LNO())`	-118.3
9	Inv. Gaussian	`gamlss(y~pH+Time, family=IG())`	-110.0
10	Inv. Gaussian	`gamlss(y~pH+Time, sigma.fo=~pH+Time, family=IG())`	-116.7

The last column of Table 1 shows the Akaike information criterion (AIC) proposed by Akaike (1973), which is a measure of the relative quality of a statistical model for a given data set. The expression to obtain AIC is given by $AIC = -2\hat{l} + 2df$, where $\hat{l}$ corresponds to the estimated log-likelihood function defined by $\hat{l} =$

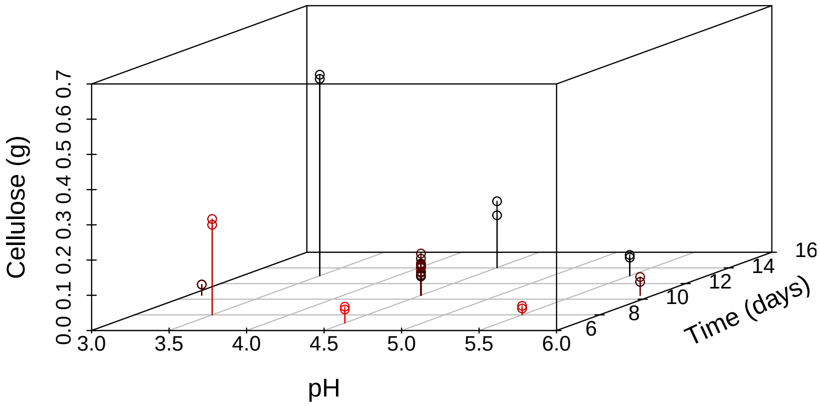

Figura 3: Scatterplot for cellulose (g), pH and cultivation time (days).

$\hat{l}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log f(y_i \mid \hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$, and df corresponds to the number of estimated parameters. Different models can be compared using their global deviances, $GD = -2\hat{l}$ (if they are nested), or using the generalized Akaike information criterion, $GAIC = -2\hat{l} + \# df$ with # as a required penalty; when $\# = 2$, the $GAIC$ corresponds to the usual Akaike information criterion AIC. The preferred model is the one with the minimum AIC value. Table 1 shows that model 6 has the lowest AIC. This model considers a gamma distribution for cellulose yield with $\log(\cdot)$ as the link function to model μ and σ.

The probability density function for the gamma distribution with μ and σ parameters ($\mu > 0$ and $\sigma > 0$) is given by

$$f_Y(y \mid \mu, \sigma) = \frac{1}{(\sigma^2\mu)^{\frac{1}{\sigma^2}}} \frac{y^{\frac{1}{\sigma^2}-1} e^{-\frac{y}{\sigma^2\mu}}}{\Gamma\left(\frac{1}{\sigma^2}\right)} \tag{6}$$

where $E(Y) = \mu$ and $Var(Y) = \sigma^2\mu^2$. Figure 4 shows the density for two combinations of parameters μ and σ. The gamma distribution is suitable for modeling skewed variables such as bacterial cellulose yield.

Table 2 presents the estimated parameters for model 6, which considers the gamma distribution for the response variable. From this table, we can see that each variable is significant at 5 % in explaining the μ and σ parameters.

From Table 2, estimated expressions can be obtained for the μ and σ parameters:

$$\log(\hat{\mu}) = -1.45 - 0.65\,\text{pH} + 0.18\,\text{Time} \tag{7}$$

$$\log(\hat{\sigma}) = 1.58 - 0.56\,\text{pH} \tag{8}$$

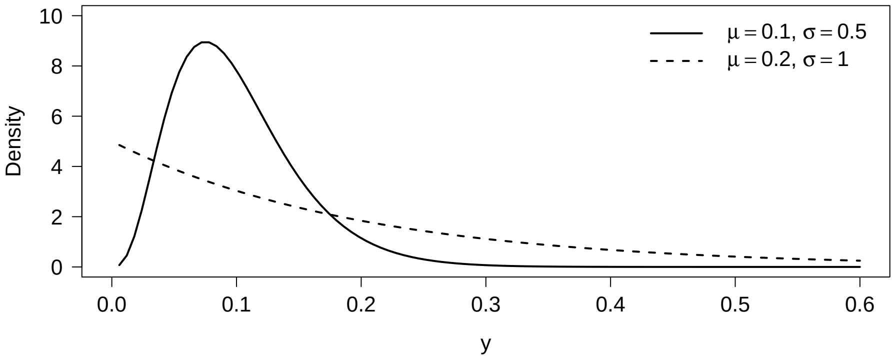

Figura 4: Density for gamma distribution for two parameter combinations.

Tabla 2: Estimated parameters for model 6.

$\log(\mu)$ model	Estimate	Std. Error	t value	P-value
Intercept	-1.45	0.58	-2.52	1.785e-02
pH	-0.65	0.09	-7.43	5.468e-08
Time	0.18	0.03	5.36	1.173e-05
$\log(\sigma)$ model	Estimate	Std. Error	t value	P-value
Intercept	1.58	0.57	2.79	0.0095142
pH	-0.56	0.12	-4.54	0.0001058

The estimated mean and variance for cellulose yield can be expressed in terms of μ and σ as follows:

$$\hat{E}(Y) = \hat{\mu} = e^{-1.45-0.65\,\mathrm{pH}+0.18\,\mathrm{Time}} \tag{9}$$

$$V\hat{a}r(Y) = \hat{\mu}^2\hat{\sigma}^2 = e^{0.26-2.42\,\mathrm{pH}+0.36\,\mathrm{Time}} \tag{10}$$

From the above expressions, we note that for each additional day of cultivation time, at a fixed value of pH, the mean cellulose yield increases by 19.72 % (obtained from $e^{0.18} = 1.1972$); similarly, for fixed cultivation time, the variance decreases by 91.11 % for each additional unit of pH (obtained from $e^{-2.42} = 0.0889$). Figure 5 plots the estimated mean and variance for several cultivation time values. From this figure, we observe that the mean and variance for cellulose yield decrease as pH increases. The opposite occurs for mean and variance as cultivation time increases.

Figure 6 presents the residual analysis for model 6. The distribution of the residuals is not far from the normal distribution, which indicates that this model is appropriate for the data. A Shapiro test for normality was carried out with a p-value of 0.4027.

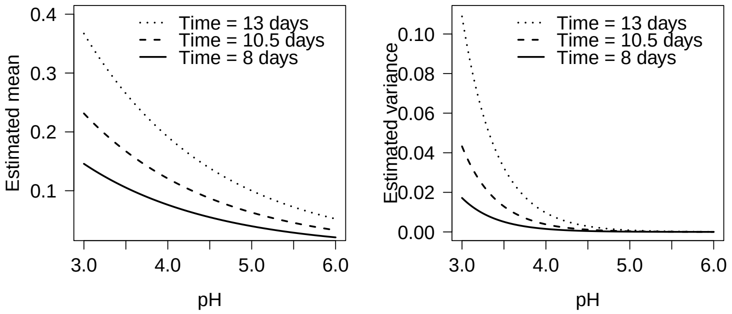

Figura 5: Estimated mean and variance for three cultivation time values.

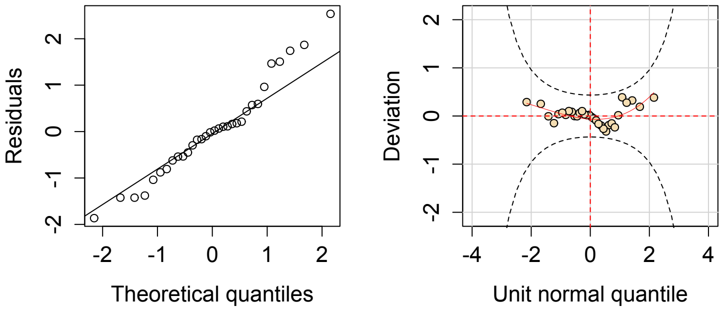

Figura 6: QQplot and worm plot for residuals of model 6.

5. Conclusions

The results agree with Castro et al. (2012), concluding that the optimal bacterial cellulose yield is found near pH 3.5. The two explanatory variables used in the model were significant in explaining the mean and variance of bacterial cellulose yield; the equations 9 and 10 could be used to model the system behavior under those conditions and to describe the variability of the bacterial cellulose yield.

Referencias

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrov & F. Csaki, eds, '2nd International Symposium on Information Theory', Budapest: Akademia Kiado, pp. 267–281.

Castro, C., Zuluaga, R., Álvarez, C., Putaux, J.-L., Caro, G., Rojas, O. J., Mondragón, I. & Ganán, P. (2012), 'Bacterial cellulose produced by a new acid-resistant strain of gluconacetobacter genus', *Carbohydrate Polymers* **89**(4), 1033 – 1037.

Jaramillo, L., Perna, M., Benito-Revollo, A., Arrieta, M. & Escamilla, E. (2013), 'Efecto de diferentes concentraciones de fructosa sobre la producción de celulosa bacteriana en cultivo estático', *Rev. Colombiana Cienc. Anim* **5**(1), 116–130.

Rigby, B. & Stasinopoulos, M. (2010), 'Instructions on how to use the gamlss package in R'.
*http://gamlss.org/images/stories/papers/gamlss-manual.pdf

Rigby, R. & Stasinopoulos, D. (2005), 'Generalized additive models for location, scale and shape', *Applied Statistics* **54**(3), 507–554.

Stasinopoulos, D. & Rigby, R. (2007), 'Generalized additive models for location, scale and shape (GAMLSS) in R', *Journal of Statistical Software* **23**(7), 1–46.