

---

# GAMLSS models applied in the treatment of agro-industrial waste<sup>1</sup>

## Modelos GAMLSS aplicados en el tratamiento de residuos agroindustriales

Freddy Hernández<sup>a</sup>  
fhernanb@unal.edu.co

Lina Arteaga<sup>c</sup>  
lina.bacteriologa@gmail.com

Mabel Torres<sup>b</sup>  
mabel.torres@upb.edu.co

Cristina Castro<sup>d</sup>  
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.

**Palabras clave:** GAMLSS models, Gamma distribution, linear regression, parameter estimation.

### Resumen

En este artículo se presenta una aplicación de los modelos GAMLSS (Generalized Additive Models for Location, Shape and Scale) para estudiar la producción de celulosa bacteriana a partir de residuos agroindustriales. El experimento fue realizado para investigar los efectos del pH y el tiempo de cultivo sobre el rendimiento de celulosa bacteriana obtenida a partir de residuos de banano. Varios modelos fueron ajustados a los datos recolectados para determinar expresiones estimadas

---

<sup>1</sup>Hernández, F., Torres, M., Arteaga, L., Castro, C. (2015) GAMLSS models applied in the treatment of agro-industrial waste. *Comunicaciones en Estadística*, **8**(2), 245-254.

<sup>a</sup>Profesor asistente, Universidad Nacional de Colombia, Sede Medellín. Colombia.

<sup>b</sup>Profesora titular, Universidad Pontificia Bolivariana, Medellín. Colombia.

<sup>c</sup>Mágister en biotecnología. Universidad Pontificia Bolivariana, Medellín. Colombia.

<sup>d</sup>Profesora asociada, Universidad Pontificia Bolivariana, Medellín. Colombia.

para la media y la varianza del rendimiento de celulosa bacteriana. Del mejor modelo obtenido se encontró que la media y la varianza del rendimiento de celulosa bacteriana disminuye a medida que el pH se incrementa, mientras al aumentar el tiempo de cultivo tanto la media como la varianza aumentan.

**Keywords:** Distribución gamma, estimación de parámetros, modelo GAMLSS, regresión lineal.

## 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. 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.

Unlike cellulose from plants, bacterial cellulose (BC) is produced with higher purity and exhibits unique mechanical properties (Shoda & Sugano 2005), making it a suitable raw material for high fidelity acoustic speakers, high quality paper, foods, and as a biomaterial in cosmetics, pharmaceuticals and medicine (Raghunathan 2013, Çoban & Biyik 2011, Rani & Appaiah 2013, Chawla et al. 2009).

Relatively high cost of BC production may limit its application to high value-added products (Legge 1990). Significant cost reductions are possible with improvements in fermentation efficiency and economics of scale (Raghunathan 2013). Obtaining bacterial cellulose depends on, among other factors, pH and fermentation time, and therefore, it is important to determine the operational values 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  $f(y_i | \boldsymbol{\theta}_i)$ , where  $\boldsymbol{\theta}_i = (\mu_i, \sigma_i, \nu_i, \tau_i)^T$  corresponds to the parameter vector. The first two elements  $\mu_i$  and  $\sigma_i$  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 (Rigby & Stasinopoulos 2005, Stasinopoulos &

Rigby 2007). GAMLSS models can be summarized as follows:

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

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

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

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

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;  $\mathbf{X}_k$  are known design matrices of order  $n \times J'_k$  associated with fixed effects  $\boldsymbol{\beta}_k$  of  $J'_k \times 1$ ; and  $\mathbf{Z}_{jk}$  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:

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

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 overmature banana, 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.



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

The response variable in the experiment was the bacterial cellulose yield calculated in grams of dry BC and obtained on each experimental unit. 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.

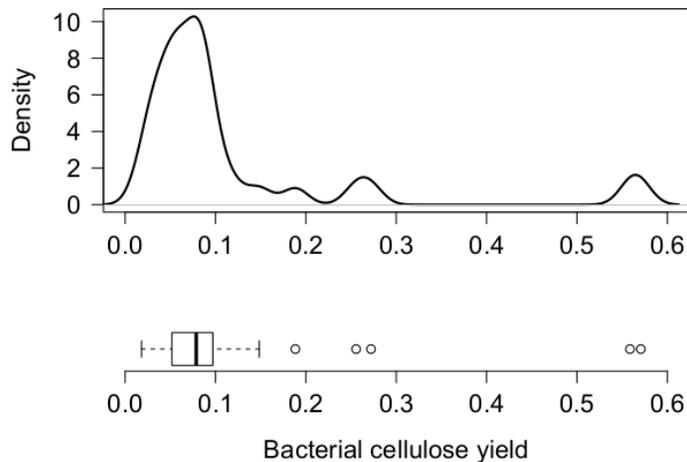


Figure 2: *Density and boxplot for bacterial cellulose yield (g). Source: Own elaboration.*

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.

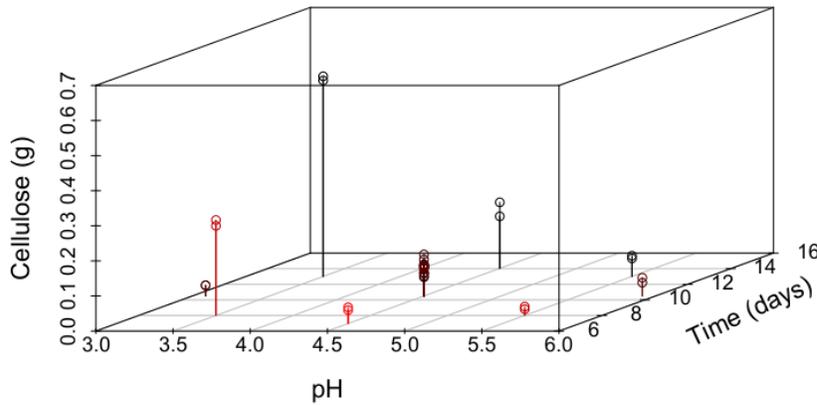


Figure 3: Scatterplot for cellulose (g), pH and cultivation time (days). Source: Own elaboration.

### 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  $\mu$  and  $\sigma$  parameters of each distribution.

Table 1: AIC values for each fitted model.

Model	Distribution	Structure in GAMLSS syntax	AIC
1	Normal	<code>gamlss(y~pH+Time, family=NO())</code>	-46.8
2	Normal	<code>gamlss(y~pH+Time, sigma.fo=~pH+Time, family=NO())</code>	-109.8
3	Normal	<code>gamlss(y~pH*Time + I(pH^2) + I(Time^2), family=NO())</code>	-49.7
4	Gamma	<code>gamlss(y~pH+Time, family=GA())</code>	-102.2
5	Gamma	<code>gamlss(y~pH+Time, sigma.fo=~pH+Time, family=GA())</code>	-118.1
6	Gamma	<code>gamlss(y~pH+Time, sigma.fo=~pH, family=GA())</code>	-119.8
7	log-Normal	<code>gamlss(y~pH+Time, family=LNO())</code>	-95.9
8	log-Normal	<code>gamlss(y~pH+Time, sigma.fo=~pH+Time, family=LNO())</code>	-118.3
9	Inv. Gaussian	<code>gamlss(y~pH+Time, family=IG())</code>	-110.0
10	Inv. Gaussian	<code>gamlss(y~pH+Time, sigma.fo=~pH+Time, family=IG())</code>	-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  $\hat{l} = \hat{l}(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(y_i | \hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$ , 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  $\mu$  and  $\sigma$ .

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

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

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

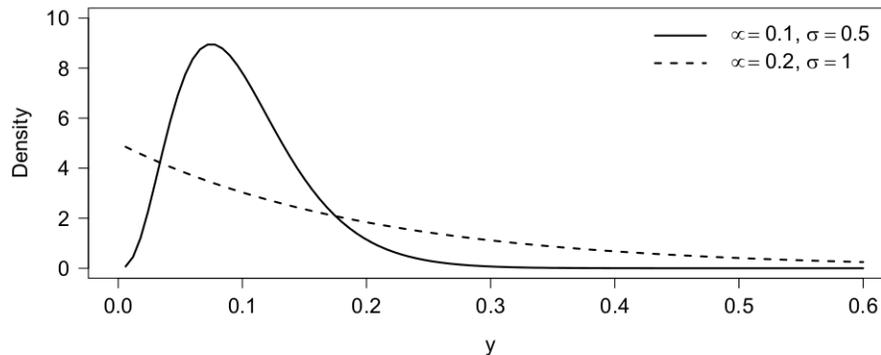


Figure 4: *Density for gamma distribution for two parameter combinations. Source: Own elaboration.*

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  $\mu$  and  $\sigma$  parameters.

From Table 2, estimated expressions can be obtained for the  $\mu$  and  $\sigma$  parameters:

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

Table 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

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

The estimated mean and variance for cellulose yield can be expressed in terms of  $\mu$  and  $\sigma$  as follows:

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

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

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.

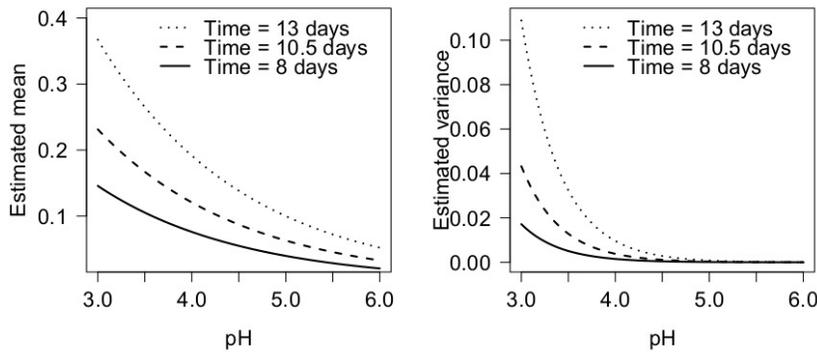


Figure 5: *Estimated mean and variance for three cultivation time values. Source: Own elaboration.*

Figure 6 shows the heat plot for the estimated mean of bacterial cellulose yield

given by the equation 9 and the colors represent the response variable. From this plot we can see that the maximum expected bacterial cellulose yield can be obtain with a low value of pH and a maximum value of cultivation time.

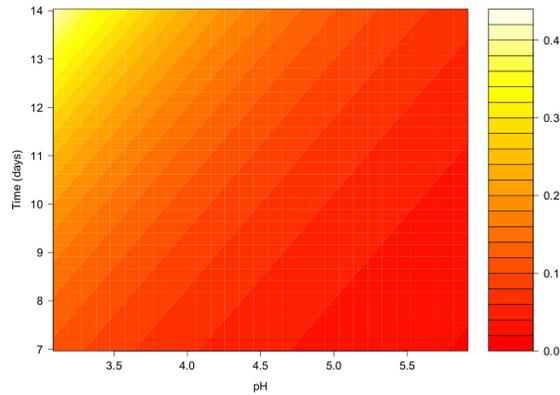


Figure 6: Heat plot for estimated mean of the bacterial cellulose yield  $\hat{E}(Y)$ . Source: Own elaboration.

Figure 7 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, additionally, a Shapiro test for normality was carried out with a  $p$ -value of 0.4027. Despite of in this experiment the sample size was 32, we found that the model 6 explains properly the cellulose yield because the residuals do not violate the normal distribution assumption for residuals.

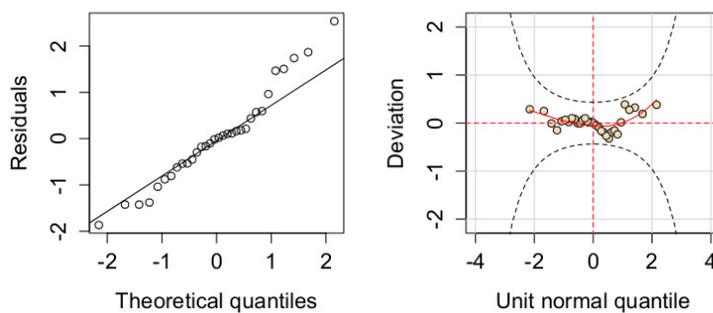


Figure 7: Q-Qplot and worm plot for residuals of model 6. Source: Own elaboration.

## 5 Conclusions

GAMLSS model is a useful statistical technique to model all parameters of a probability density (or mass) function for a response variable using a set of covariates. In this paper we showed an application of GAMLSS to model the bacterial cellulose yield using as covariates pH and cultivation time. The results showed in Figures 5 and 6 point out that the maximum bacterial cellulose yield is obtain for low values of pH and cultivation time close to 14 days, this results agree with the experiment of Castro et al. (2012) that concluded that the optimal bacterial cellulose yield for this type of experiment 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 by researchers to model (or predict) the system behavior under those conditions and to describe the variability of the bacterial cellulose yield.

**Recibido: 27 de agosto del 2015**  
**Aceptado: 15 de octubre del 2015**

## References

- 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.
- Çoban, E. & Biyik, H. (2011), 'Effect of various carbon and nitrogen sources on cellulose synthesis by *Acetobacter lovaniensis* HBB5', *African Journal of Biotechnology* **10**(27), 5346–5354.
- Chawla, P., Bajaj, I., Survase, S. & Singhal, R. (2009), 'Microbial Cellulose: Fermentative Production and Applications', *Food Technol. Biotechnol.* **47**(2), 107–124.
- 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.
- Legge, R. L. (1990), 'Microbial cellulose as a speciality chemical', *Biotechnology Advances* **8**(2), 303 – 319.  
\*<http://www.sciencedirect.com/science/article/pii/073497509091067Q>

- Raghunathan, D. (2013), 'Production of microbial cellulose from the new bacterial strain isolated from temple wash waters', *Int. J. Curr. Microbiol. App. Sci.* (2), 275–290.
- Rani, M. & Appaiah, A. (2013), 'Production of bacterial cellulose by *Gluconacetobacter hansenii* UAC09 using coffee cherry husk', *J Food Sci Technol* **50**(4), 755–762.
- 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.
- Shoda, M. & Sugano, Y. (2005), 'Recent advances in bacterial cellulose production', *Biotechnology and Bioprocess Engineering* **10**(1), 1–8.
- Stasinopoulos, D. & Rigby, R. (2007), 'Generalized additive models for location, scale and shape (GAMLSS) in R', *Journal of Statistical Software* **23**(7), 1–46.