ORIGINAL ARTICLE
DOI: 10.5433/1679-0375.2021v42n2p163

Building flexible regression models: including the Birnbaum-Saunders distribution in the gamlss package

Construção de modelos de regressão flexíveis: incluindo a distribuição Birnbaum-Saunders no pacote gamlss

Fernanda V. Roquim[1]; Thiago G. Ramires[2]; Luiz R. Nakamura[3];
Ana J. Righetto[4]; Renato R. Lima[5]; Rayne A. Gomes[6]

Abstract

Generalized additive models for location, scale and shape (GAMLSS) are a very flexible statistical modeling framework, being an important generalization of the well-known generalized linear models and generalized additive models. Their main advantage is that any probability distribution (that does not necessarily belong to the exponential family) can be considered to model the response variable and different regression structures can be fitted in each of its parameters. Currently, there are more than 100 distributions that are already implemented in the `gamlss` package in `R` software. Nevertheless, researchers can implement different distributions if they are not yet available, e.g., the Birnbaum-Saunders (BS) distribution, which is widely used in fatigue studies. In this paper we make available all codes regarding the inclusion of the BS distribution in the `gamlss` package, and then present a simple application related to air quality data for illustration purposes.

Keywords: Birnbaum-Saunders distribution. Generalized additive models. Generalized linear models. Smoothing functions. Statistical modeling.

Resumo

Os modelos aditivos generalizados para locação, escala e forma (GAMLSS) são uma estrutura de modelagem estatística extremamente flexível, sendo uma importante generalização dos modelos lineares generalizados e dos modelos aditivos generalizados. Sua principal vantagem é que podemos considerar qualquer distribuição de probabilidade (que não necessariamente deve pertencer à família exponencial) para modelar a variável resposta, e diferentes estruturas de regressão podem ser ajustadas para cada um de seus parâmetros. Atualmente, existem mais de 100 distribuições já implementadas no pacote `gamlss` disponível no software `R`. Não obstante, os pesquisadores podem implementar diferentes distribuições que ainda não estão disponíveis, como, por exemplo, a distribuição Birnbaum-Saunders (BS), largamente utilizada em estudos de fadiga. Neste artigo, disponibilizamos todos os códigos no que tange à inclusão da distribuição BS no pacote `gamlss` e apresentamos uma aplicação relacionada a dados de qualidade do ar para fins de ilustração.

Palavras-chave: Distribuição Birnbaum-Saunders. Modelos aditivos generalizados. Modelos lineares generalizados. Funções de suavização. Modelagem estatística.

[1] Doutoranda em Estatística e Experimentação Agropecuária, UFLA, Lavras, MG, Brasil, E-mail: fer_venturato@yahoo.com.br
[2] Prof. Dr., Depto. Acadêmico de Matemática, UTFPR, Apucarana, PR, Brasil, E-mail: thiagogentil@gmail.com
[3] Prof. Dr., Depto. de Informática e Estatística, UFSC, Florianópolis, SC, Brasil, E-mail: luiz.nakamura@ufsc.br
[4] Chefe de estatística, ALVAZ, Londrina, PR, Brasil, E-mail: ajrighetto@gmail.com
[5] Prof. Dr., Depto. de Estatística, UFLA, Lavras, MG, Brasil, E-mail: rrlima@des.ufla.br
[6] Mestranda, Pós-Graduação em Engenharia Ambiental, UTFPR, Apucarana, PR, Brasil, E-mail: rayneaz@gmail.com

Introduction

Regression models were proposed to identify and quantify the relationship between a given phenomenon (response/target/dependent variable) and a set of independent variables. Although they are still considered, as in, for example, Souza and Raminelli (2013), in their classic form, such models have extremely rigid assumptions, since the response variable must follow a Gaussian distribution with mean given by $\mu = \boldsymbol{X\beta}$ and constant variance σ^2. However, in several practical situations, these assumptions are violated.

In this context, Nelder and Wedderburn (1972) presented the generalized linear models (GLM). Within this framework, we are now able to fit regression models for binary responses, counts, proportions, among others. The main assumption of these models is that the considered distribution necessarily belongs to the exponential family, e.g., the binomial distribution in logistic regression models, e.g., as in Leão and Urbano (2020). Now, the mean is modeled by the explanatory variables through a link function, $g(\mu) = \boldsymbol{X\beta}$. Nonetheless, the previous case is a particular case of the GLM.

A different clear assumption from both previous models is that the relationship between the (function of the) mean of the response and the explanatory variables is strictly linear. However, in certain practical situations, we might think that this is not necessarily true as in, e.g., Fernandes, Bornia and Nakamura (2019). In this sense, Hastie and Tibshirani (1990) pioneered the generalized additive models (GAM), that allow the addition of smoothing functions within the GLM structure, letting the behavior of the explanatory variable guide the relationship with the mean function, i.e., such relationship no longer needs to be linear. Several smoothing functions are available in the literature and, among them, perhaps the most prominent and computationally friendly are the P-splines (EILERS; MARX, 1996; EILERS; MARX; DURBÁN, 2015).

Even though the GLM and GAM represent significant advances in regression model techniques, we may still require greater flexibility to ensure a good fit, e.g., in problems where asymmetry and/or excess of kurtosis are identified and should be explicitly modeled. Rigby and Stasinopoulos (2005) developed the generalized additive models for location, scale and shape (GAMLSS), which present crucial improvements when compared to the above mentioned models. GAMLSS allow to fit models based on any distribution for the response variable, regardless of belonging to the exponential family. In addition, we can now model any and all of the parameters of such distribution as a function of explanatory variables. Mathematically, consider the response variable $Y \sim \mathscr{D}(\theta_k)$, where $\mathscr{D}$ is a probability distribution and $\theta_k = (\theta_1, \ldots, \theta_p)^\top$ is its vector of parameters.

The GAMLSS model can be written as

$$g_k(\theta_k) = \eta_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} s_{jk}(\boldsymbol{x}_{jk}),$$

where $g_k(\cdot)$, $k = 1, \ldots, p$, is a link function relating the parameter θ_k with its predictor η_k, $\boldsymbol{X}_k$ is a known design matrix, $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \ldots, \beta_{J'_k k})$ is a parameter vector of length J'_k and s_{jk} is a non-parametric function (e.g. a P-spline) of an explanatory variable $\boldsymbol{x}_{jk}$.

GAMLSS are being increasingly used in several fields of knowledge, such as, actuarial sciences (NAKAMURA *et al.*, 2017), medical (RAMIRES *et al.*, 2018a; RAMIRES *et al.*, 2018b), agriculture (RIGHETO *et al.*, 2019), sports (NAKAMURA *et al.*, 2019), among others. Furthermore, important global organizations are also considering this framework in their analyses, such as the World Health Organization, the International Monetary Fund, the European Central Bank and the Bank of England, as available at <http://www.gamlss.com>.

Another interesting point is that GAMLSS are fully implemented in the `gamlss` package (which is constantly updated) in `R` software (R CORE TEAM, 2021). There are currently more than 100 different distributions implemented that can be used as basis for these models.

Nevertheless, we can also include new distributions at our discretion. We may cite here some recently implemented models, such as the Weibull cure rate (RAMIRES *et al.*, 2019), the odd log-logistic exponentiated Gumbel (ALIZADEH *et al.*, 2019), the log-sinh Cauchy promotion time (RAMIRES *et al.*, 2018b), among others.

Further, several functions that may assist in the fitting and checking processes are available, such as the `stepGAICAll.A` function (STASINOPOULOS *et al.*, 2017), which applies a quite powerful stepwise-based model selection method (RAMIRES *et al.*, 2021).

Thus, the main objective of this paper is to implement and make available, in the `gamlss` package, the GAMLSS model based on the Birnbaum-Saunders (BS) distribution (BIRNBAUM; SAUNDERS, 1969).

Including the BS distributions in the gamlss package

The BS distribution (BIRNBAUM; SAUNDERS, 1969) occupies a prominent position among models applied in fatigue studies. Despite its great importance it has not been implemented in the `gamlss` package yet.

Mathematically, a response Y which follows a BS distribution with parameters μ and σ, $Y \sim BS(\mu,\sigma)$, has its probability density function (pdf) and cumulative distribution function (cdf) defined as

$$f(y|\mu,\sigma) = \frac{y^{-3/2}(y+\mu)}{2\sigma\sqrt{2\pi\mu}} \exp\left\{-\frac{1}{2\sigma^2}\left(\frac{y}{\mu}+\frac{\mu}{y}-2\right)\right\}$$

and

$$F(y|\mu,\sigma) = \Phi\left[\frac{1}{\sigma}\left\{\left(\frac{y}{\mu}\right)^{1/2} - \left(\frac{\mu}{y}\right)^{1/2}\right\}\right],$$

respectively, where Φ denotes the cdf of the standard normal distribution, $\mu > 0$ is the median of the distribution and $\sigma > 0$ is a shape parameter. In this parameterization, as $\sigma \to 0$, the BS distribution becomes symmetric around μ, while as σ grows, the distribution becomes more positively skewed. It is noteworthy that regardless of the distribution to be implemented in the `gamlss` package, computationally the parameters must necessarily be defined as in Stasinopoulos and Rigby (2007), that is, distributions with up to four parameters must follow the order μ, σ, ν and τ.

Regarding the computational implementation, at first, the name of the new function (distribution) is defined (e.g. `BS`) identifying the number of parameters (two, in this case, μ and σ) and their respective link functions. Analogously, based on some adjustments and generalizations, distributions with up to four parameters can also be included in the `gamlss.dist` package (STASINOPOULOS; RIGBY, 2007; STASINOPOULOS *et al.*, 2017).

In order to facilitate the fitting of the GAMLSS model based on the BS distribution, we provide all implementation codes at <https://git.io/JRi5k>. It is also possible to directly load them into the `R` software through the following command:

```
source("https://git.io/JRi5k")
```

Once implemented codes are loaded, the pdf, cdf, quantile function and the BS random number generator can be accessed through the dBS, pBS, qBS and rBS functions, defined respectively in the code:

```
dBS(x, mu=1, sigma=1, log=FALSE)
pBS(q,mu=1,sigma=1,lower.tail=T,log.p=F)
qBS(p,mu=1,sigma=1,lower.tail=T,log.p=F)
rBS(n, mu=1, sigma=1)
```

Note that we can now use the above mentioned functions to understand the behavior of the BS distribution. For instance, to check some of the density forms of the BS pdf and cdf, for different parameter values, as can be seen in Figure 1, we use the following codes:

```
curve(dBS(x,20,0.5),0.01,100,lwd=2,col=4)
curve(dBS(x,20,1),add=T,lwd=2,col=2)

curve(pBS(x,20,0.5),0.01,100,lwd=2,col=4)
curve(pBS(x,20,1),lwd=2,add=T, col=2)
```

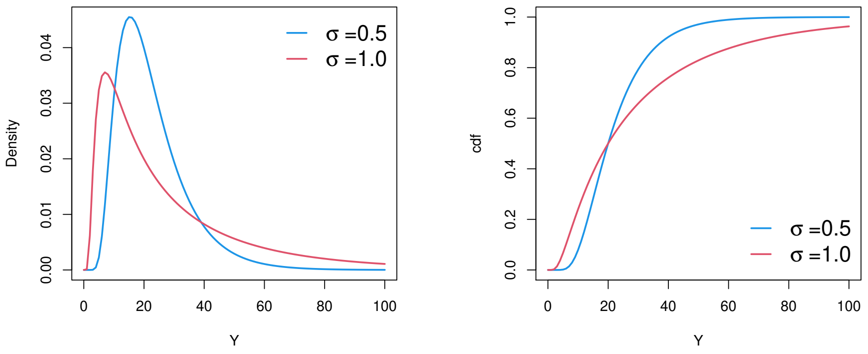
Since we have now implemented the BS distribution into the `gamlss.dist` package, we can fit GAMLSS models based on this distribution. In each regression structure (for both parameter μ and σ), we can consider linear additive terms, such as polynomial, fractional polynomials, piecewise polynomials, B-splines, and others. Conversely, for the smoothing terms, we may consider, for example, P-splines, monotonic smooth, cycle smooth, cubic splines, ridge and lasso regression, among others. For further information, check Stasinopoulos *et al.* (2017).

In order to check the adequacy of GAMLSS models, within the `gamlss` package we may use the normalized (randomized) quantile residuals (DUNN; SMYTH, 1996), given by $\hat{r}_i = \Phi^{-1}(\hat{u}_i)$, where Φ^{-1} is the inverse of the cdf of the standard normal distribution and $\hat{u}_i$ are the quantile residuals. In both cases where y is an observation from a continuous or a discrete random variable, the cdf of the distribution considered in the GAMLSS model must be defined (in our example, the cdf is computationally defined as `pBS`) and the residuals must follow a standard normal distribution for a reasonable fit. For further details, check Stasinopoulos *et al.* (2017).

Application of the implemented model

In this section, we will perform an application on a data set available in the `R` software, illustrating the functionalities of the implemented BS model. It is noteworthy that the main objective here is not to make comparisons with other probabilistic models. The data set consist in the quality of the air in the United States, and can be accessed under the name `usair` in the `gamlss.data` package. The response variable (Y) refers to the annual mean concentration of sulfur concentration in the air (mg.m^{-3}) in 41 cities in the United States. All data relate to means for years 1969–1971 (HAND *et al.*, 1994), and for the purpose

Figure 1 – Birnbaum-Saunders distribution plots considering $\mu = 20$: (a) pdf; and (b) cdf.



(a) (b)

Source: The authors.

of this application, this characteristic will be explained by two available covariates: i) average annual temperature (Fahrenheit), x_1; and ii) average annual wind speed (miles per hour), x_4. The nomenclature x_4 is used here to follow the pattern presented in the `R` software.

Table 1 displays some descriptive statistics of all considered variables. The response variable is positively skewed and presents a leptokurtic shape (skewness and kurtosis values are equal to 1.58 and 2.26, respectively) as can be also seen in Figure 2(a). Hence, the BS may be a reasonable distribution to model such response since a similar shape is observed in Figure 1(a).

Table 1 – Descriptive statistics of the variables

	Y	x_1	x_2
Minimum	8.00	43.50	6.00
Mean	30.05	55.76	9.44
Median	26.00	54.60	9.30
Standard deviation	23.47	7.23	1.43
Maximum	110.00	75.50	12.70
Skewness	1.58	0.82	0.01
Kurtosis	2.26	0.09	0.06

Source: The authors.

Figure 2 (b) and (c) present the relationship between Y and each of the considered explanatory variables. As can be seen in Panel (b), as the average annual temperature increases, sulfur dioxide concentration tends to decrease. Conversely, in Panel (c), for low and high average annual wind speeeds, we have low levels of sulfur dioxide concentration, whereas for speeds around the center of the plot, we have high response concentration. Further, in both scatter plots there is clearly a non-constant variance.

In order to fit this new implemented model, we can use any strategy described in Stasinopoulos *et al.* (2017). In this paper, we fit the GAMLSS model based in the BS distribution through the following code (note that the `pb()` function stands for the P-splines).

```
source("https://git.io/JGv62")
library(gamlss)
data(usair)
mod=gamlss(y~x1, sigma.fo=~pb(x1)+pb(x4),
           family=BS, data=usair)
summary(mod)
term.plot(mod, what="sigma")
```

The fitted GAMLSS model based on the BS distribution, is given by

$$\begin{aligned}\hat{\mu} &= \exp\{6.59 - 0.06x_1\}\\ \hat{\sigma} &= \exp\{-3.09 + s(x_1) + s(x_4)\}.\end{aligned}$$

The average annual temperature (x_1) affects both parameters of the BS distribution. As expected, as this variable increases, the median of the response deacreses. Regarding the parameter σ, which directly affects the shape of the distribution, this relationship was modeled through a P-spline, Figure 3 (a). The sulfur concentration increases up to 55 °F, then decreases up to 70°F and is constant from that point.

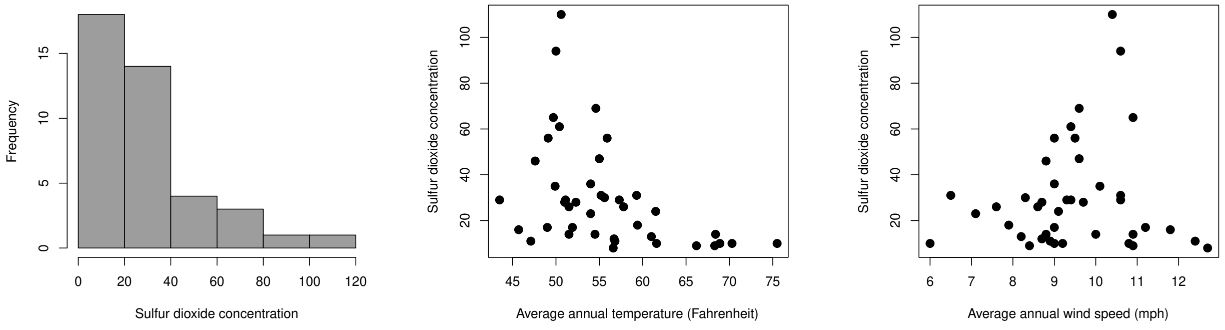
Further, wind speed only affects the shape parameter σ: although this relationship is roughly (positive) linear, Figure 3 (b), the inclusion of the P-spline to explain such variable was important to provide a better fit, which is evaluated through the worm plots Van Buuren and Frederiks (2001) based on the normalized quantile residuals using the following code: (STASINOPOULOS *et al.*, 2017):

```
wp(mod)
```

See Stasinopoulos *et al.* (2017) for more information about the code.

Figure 4 displays the worm plot obtained from the fitted model and since all residuals are within the 95% confidence bands, and present no particular shape, we can say that the GAMLSS model based on the BS distribution provides a reasonable fit to these data.

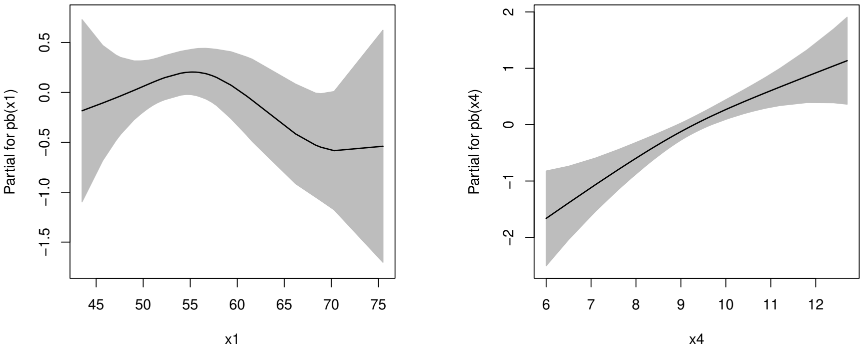
Figure 2 – (a) Sulfur dioxide concentration, (b) Relationship between sulfur dioxide concentration and average annual temperature and (c) Relationship between sulfur dioxide concentration and average annual wind speed.



(a) (b) (c)

Source: The authors.

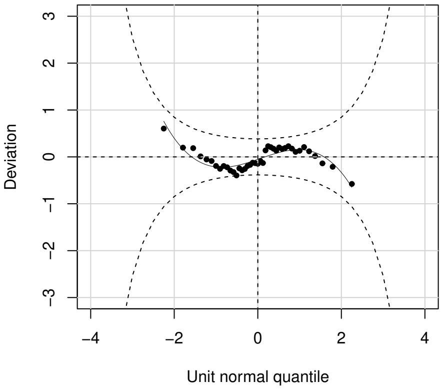
Figure 3 – Relationship between the shape parameter σ against: (a) average annual temperature and (b) average annual wind speed.



(a) (b)

Source: The authors.

Figure 4 – Worm plot of the residuals obtained from the fitted GAMLSS model based on the BS distribution.



Source: The authors.

Final considerations

In this paper we introduced the Birnbaum-Saunders distribution into the `gamlss` package in `R`, which can be straightforwardly used by other researchers. One of the main advantages of including new distributions in this package is that any and all of its parameters can be modeled as a function of independent variables. Hence, we highlight that other different flexible distributions that are not yet implemented in this framework can be easily included, so all available features in the `gamlss` package, such as residual analysis, prediction techniques, estimation, smoothed functions, among others, can be used for the new model.

Acknowledgments

The first author gratefully acknowledge grant from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/Brazil).

References

ALIZADEH, M.; RAMIRES, T. G.; MIRMOSTAFAEE, S. M. T. K.; SAMIZADEH, M.; ORTEGA, E. M. M. O. A new useful four-parameter extension of the Gumbel distribution: Properties, regression model and applications using the GAMLSS framework. *Communications in Statistics-Simulation and Computation*, New York, v. 48, p. 1746-1767, 2019.

BIRNBAUM, Z. W., SAUNDERS, S. C. Estimation for a family of life distributions with applications to fatigue. *Journal of Applied Probability*, Sheffield, v. 6, p. 328–347, 1969.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, Alexandria, v. 5, p. 236-245, 1996.

EILERS, P. H. C.; MARX, B. D. Flexible smoothing with B-splines and penalties. *Statistical Science*, Hayward, v. 11, p. 89-102, 1996.

EILERS, P. H. C.; MARX, B. D.; DURBÁN, M. Twenty years of P-splines. *SORT*, Barcelona, v. 39, p. 149-186, 2015.

FERNANDES, S. M.; BORNIA, A. C.; NAKAMURA, L. R. The influence of boards of directors on environmental disclosure. *Management Decision*, New York, v. 57, p. 2358-2382, 2019.

HAND, D. J.; DALY, F.; LUNN, A. D.; McCONWAY, K. J.; OSTROWSKI, E. *A handbook of small data sets*. London: Chapman and Hall, 1994. 458p.

HASTIE, T. J.; TIBSHIRANI, R. J. *Generalized additive models*. London: Chapman and Hall, 1990. 352p.

LEÃO, A. L. F.; URBANO, M. R. Street connectivity and walking: an empirical study in Londrina-PR. *Semina*: Ciências Exatas e Tecnológicas, Londrina, v. 41, p. 31-42, 2020.

NAKAMURA, L. R.; CERQUEIRA, P. H. R.; RAMIRES, T. G.; PESCIM, R. R.; RIGBY, R. A.; STASINOPOULOS, D. M. A new continuous distribution on the unit interval applied to modelling the points ratio of football teams. *Journal of Applied Statistics*, Abingdon, v. 46, p. 416-431, 2019.

NAKAMURA, L. R.; RIGBY, R. A.; STASINOPOULOS, D. M.; LEANDRO, R. A.; VILLEGAS, C.; PESCIM, R. R. Modelling location, scale and shape parameters of the Birnbaum-Saunders generalized t distribution. *Journal of Data Science*, Taipei, v. 15, p. 221-238, 2017.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. Journal of the Royal Statistical Society. Series A (General), London, v. 135, p. 370-384, 1972.

R CORE TEAM. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna: [*s. n.*], 2021. Available from: <https: //www.Rproject.org/>. Access in: May 01, 2021.

RAMIRES, T. G.; NAKAMURA, L. R.; RIGHETTO, A. J.; ORTEGA, E. M. M.; CORDEIRO, G. M. Predicting survival function and identifying associated factors in patients with renal insufficiency in the metropolitan area of Maringá, Paraná State, Brazil. *Cadernos de Saúde Pública*, Rio de Janeiro, v. 34, p. 1-13, 2018a.

RAMIRES, T. G.; NAKAMURA, L. R.; RIGHETTO, A. J.; PESCIM, R. R.; MAZUCHELI, J.; CORDEIRO, G. M. A new semiparametric Weibull cure rate model: fitting different behaviors within GAMLSS. *Journal of Applied Statistics*, Abingdon, v. 46, n. 15, p. 2744-2760, 2019.

RAMIRES, T. G.; CORDEIRO, G. M.; KATTAN, M. W.; HENS, N.; ORTEGA, E. M. M. Predicting the cure rate of breast cancer using a new regression model with four regression structures. *Statistical methods in medical research*, London, v. 27, n. 11, p. 3207-3223, 2018b.

RAMIRES, T. G.; NAKAMURA, L. R.; RIGHETTO, A. J.; CARVALHO, R. J.; VIEIRA, L. A.; PEREIRA, C. A. B. Comparison between Highly Complex Location Models and GAMLSS. *Entropy*, Basel, v. 23, n. 4, p. 469, 2021.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society*. Series C, Applied Statistics, London, v. 54, p. 507-554, 2005.

RIGHETTO, A. J.; RAMIRES, T. G.; NAKAMURA, L. R.; CASTANHO, P. L. D. B.; FAES, C.; SAVIAN, T. V. Predicting weed invasion in a sugarcane cultivar using multispectral image. *Journal of Applied Statistics*, Sheffield, v. 46, p. 1-12, 2019.

SOUZA, R. C.; RAMINELLI, J. Aplicação do modelo linear na avaliação de dados de estabilidade de medicamento. *Semina*: Ciências Exatas e Tecnológicas, Londrina, v. 34, p. 57-66, 2013.

STASINOPOULOS, D. M.; RIGBY, R. A. Generalized additive models for location, scale and shape (GAMLSS). *Journal of Statistical Software*, [California], v. 23, p. 1-64, 2007.

STASINOPOULOS, M. D.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; BASTIANI, F. *Flexible Regression and Smoothing*: Using GAMLSS. Boca Raton: Chapman and Hall, 2017. 549p.

VAN BUUREN, S.; FREDRIKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, Chichester, v. 20, p. 1259-1277, 2001.

Received: Aug. 31, 2021
Accepted: Oct. 16, 2021
Published: Oct. 29, 2021