GENERALIZED ADDITIVE MIXED MODELS WITH P-SPLINES APPLIED TO SUGARCANE PRODUCTION IN THE STATE OF SAO PAULO

Natalie Verónika Rondinel MENDOZA[1]
Sônia Maria De Stefano PIEDADE[1]

▪ ABSTRACT: In this paper, we apply the generalized additive mixed models with the use of the P-*splines* technique as mixed models, which will be adopted in a problem of the agro-environmental area, in this case on the average levels of sugarcane production, which is influenced by changes in climatic variables such as temperature and precipitation which were measured over 10 years in each mesoregion of the state of Sao Paulo. The reason for using this approach as a smoothing method is that the tendency of these climate covariates is not known on its most, but it is known that they directly influence the response variable. Besides allowing for the inclusion of fixed and random effects in the models to be proposed, these models allow for the inclusion of an autoregressive process AR(1) as a correlation structure in the residuals.

▪ KEYWORDS: P-*splines; B-splines; generalized additive mixed models; autoregressive process AR(1).*

1 Introduction

A common feature of real data is that they often have complexities in the linear or nonlinear relationship between the response variables and their covariates, since in most studies the behavior of predictors or covariates does not follow a certain probabilistic pattern, but the linear regression models are used to model these types of data.

[1] Universidade de São Paulo - USP, Escola Superior de Agricultura Luiz de Queiroz, Departamento de Ciências Exatas, CEP: 13418-900, Piracicaba, São Paulo, Brazil. E-mail: *natalievrm@gmail.com; soniamsp@usp.br*

An alternative to the use of linear regression models is to use the additive models, among them we have the generalized additive mixed models (GAMM) proposed by Lin and Zhang (1999), as a class of models that are an extension of generalized linear mixed models (GLMM) proposed by (BRESLOW and CLAYTON, 1993) and the generalized additive models (GAM) proposed by Hastie and Tibshirani (1990). This class of models uses additive nonparametric functions to model the effect of covariates on the response variable as well as takes into between observations account the effect of the presence of overdispersion and correlation, by adding random effects to the linear predictor of the model (DURBÁN, 2014).

The objective of this paper was to study the average levels of sugarcane production in tons/ha for each mesoregion in the state of Sao Paulo, Brazil, under the influence of changes in the maximum temperature, minimum temperature and precipitation, we proposed using GAMM with the P-splines methodology to asses the effects of climatic variables on the performance of sugarcane production.

The advantage of using the P-splines is that they allow for a non-linear trend of production over time, incorporating also the non-linear behavior of the effects of each of the climatic variables (RONDINEL MENDOZA, 2017).

The remainder of this article is organized as follows. Section 2 presents a brief introduction to the GAMM and a framework brief description of the estimation of the smooth functions of the model using P-splines as mixed models. Section 3 shows the materials and methods used. Section 4 presents an application of this approach in the case study of sugarcane production and the respective results. Finally, in the last section we present the conclusions.

2 Generalized additive mixed models

The generalized additive mixed models (GAMM) were proposed by Lin and Zhang (1999), they estimated the nonparametric functions by using smoothing splines and to jointly estimate the smoothing parameters and the variance components, they used the marginal quasi-likelihood method. However, Chen (2007) estimated nonparametric functions and covariance structures based on penalized marginal likelihood, where he used the maximum likelihood estimation, developing two algorithms the first based on the Newton-Raphson algorithm and the second based on an extension of the Monte Carlo method. Wood (2006) and Zuur *e*t al. (2009) used the GAMMs with smooth functions in terms of tensor products using the thin plate splines as bases of low range. GAMMs can be used in studies with experimental designs whether they are nested or crossed, as well as applied to spatial data, clustered data or hierarchical data. According to Lin and Zhang (1999), the GAMM distinguishes itself from the GLMM, in which the linear predictor is replaced by an additive predictor in the systematic component of the model whose additive predictor involves a sum of smooth functions of the explanatory variables, defined by

$$\eta_i|\boldsymbol{\alpha} = \sum_{j=1}^{p} f_j(x_{ji}) + \sum_{k=1}^{q} z_{ik}\alpha_k, \quad \text{with} \quad i = 1, \ldots, n \tag{1}$$

where $f_j(\cdot)$ are the smooth functions of the covariates $\boldsymbol{x_j}$ with the condition that $f_j^{''}(\cdot)$ is continuous, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_q)^{'}$ is the vector of random effects that has a multivariated normal distribution with q dimension expressed by $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{G})$, the variance-covariance matrix defined by $\boldsymbol{G} = \sigma_\alpha^2 \boldsymbol{I}_\alpha$ and $\boldsymbol{Z} = (z_{1k}, z_{2k}, \ldots, z_{ik})^{'}$ is the random effects design matrix of model. By expression (1), the GAMM for exponential families is basically a hierarchical model, with the following probability density function ,

$$f(\boldsymbol{y}|\boldsymbol{\eta}) = \exp\{\boldsymbol{y'\eta} - \mathbf{1'}b(\boldsymbol{\eta}) + \mathbf{1'}c(\boldsymbol{y})\} \quad , \tag{2}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^{'}$ is the vector of the systematic component, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ is the vector of the observed responses conditionally independent given the vector of random effects $\boldsymbol{\alpha}$ with conditional mean $E[y_i|\boldsymbol{\alpha}] = \mu_i^\alpha$ and conditional variance $Var(y_i \mid \boldsymbol{\alpha}) = \phi(m_i)^{-1}V(\mu_i^\alpha)$. Note that $V(\cdot)$ is the variance function of the conditional mean, m_i is a known prior weight and ϕ is the dispersion parameter as in the case of the generalized linear model (MCCULLAGH AND NELDER, 1989). It is observed that the conditional mean is linked to the additive predictor, given by Equation (1), through the link function $g(\mu_i^\alpha) = \eta_i^\alpha$. The function $g(\cdot)$ is a monotonically differentiable function.

The main objective of GAMM expressed in (1) is that additive smooth functions are used to model the effects of covariates and the random effects are used to model correlation between observations, because they have a flexible covariance structure of random effects $\boldsymbol{\alpha}$.

Among some references where models such as GAMM special cases have been used, Zeger and Diggle (1994) and Zhang *e*t al. (1998) consider expression (1) as a mixed semi-parametric model. The authors assumed a simple non-parametric function as a function of time and of the longitudinal observations, assumed to be normally distributed. Zuur *et al.* (2009) uses GAMM in spatial data and ecological studies to model spatial correlation. Consider the following additive model

$$y_i = f_1(x_{1i}) + f_2(x_{2i}) + \ldots + \boldsymbol{Z}_i\boldsymbol{\alpha} + \epsilon_i, \tag{3}$$

where y_i is the response variable; $f_j(\cdot)$ are the smooth functions of the covariates x_{ji}; $\boldsymbol{Z}_i$ is the matrix ith row of the random effects design matrix; $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{G})$ is the vector of random effects coefficients with unknown positive definite covariance matrix $\boldsymbol{G}$, the vector of errors with ith element ϵ_i given by $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$, with variance-covariance matrix $\boldsymbol{\Lambda}$. To follow, the process of estimation of smooth functions is shown considering as a one-dimensional case for the GAMM model.

Estimation of GAMM smooth functions: We consider the additive model (3), which is composed of smooth one-dimensional functions $f(x)$ for each covariate

x_{ji}. Each smooth function is defined as $f(x) = \boldsymbol{Ba}$, where $\boldsymbol{B}$ is a B-spline basis function and $\boldsymbol{a}$ a vector of B-spline regression coefficients. To estimate the smooth one-dimensional function of model (3) we will use the representation of the P-splines as mixed models. The P-splines were proposed by Eilers and Marx (1996), their name is due to a simple combination of two ideas to fit the curve: regression on the functions of B-splines basis and a penalty of differences on the regression coefficients of B-splines proposed by De Boor (1978). Following the approach of Eilers and Marx (1996), they used a penalty based on the "*d*-order differences" between the adjacent coefficients of the B-splines basis. This type of penalty is more flexible because it independs on the degree of the polynomial used to construct the B-splines basis. This penalty is a good discrete approximation of the second derivative of the integrable square of a function (EILERS and MARX, 1996). Therefore, the penalty is the main part of the estimation of the smooth function; meaning that smoothing of the function is adjusted by changing the weights of the coefficients of the regression.

According to Lee and Durbán (2011), considering a one-dimensional function: $f(x) = \boldsymbol{Ba}$, the P-spline representation approach is used as a mixed model; that is, reparametrization is required for the smooth function in two parts: the first part to be treated as a fixed effect and the other to be treated as a random effect. In general, this can be obtained using the singular value decomposition of the penalty matrix, $\boldsymbol{P} = \boldsymbol{U\Sigma U'}$, where $\boldsymbol{U}$ is an orthogonal matrix whose columns are the eigenvalues of $\boldsymbol{P}$, and $\boldsymbol{\Sigma}$ is a diagonal matrix with the corresponding eigenvalues arranged in descending order on the main diagonal. $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{c-d}}$ denotes the sub matrix of positive eigenvalues of $\boldsymbol{\Sigma}$, where c is the number of columns of the B-splines basis and d is the order of the penalty. Therefore, the new matrix and the new vector of coefficients are

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{BU_0} \qquad\qquad \text{and} & (4) \\ \boldsymbol{Z} &= \boldsymbol{BU_+} & (5) \end{aligned}$$

where, $\boldsymbol{U_0}$ is the sub matrix of null eigenvalues of the singular value decomposition of the penalty matrix, and $\boldsymbol{U_+}$ is the sub matrix that contains the positive eigenvalues of $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{c-d}}$. With new coefficients

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{U_0'a} \qquad\qquad \text{and} \\ \boldsymbol{\alpha} &= \boldsymbol{U_+'a} \qquad\qquad . \end{aligned}$$

A mixed model representation of a smooth function in terms of a linear predictor and a random effects distribution, is given by

$$\boldsymbol{X\beta} + \boldsymbol{Z\alpha}, \quad \text{where} \quad \boldsymbol{\alpha} \sim N\left(\boldsymbol{0}, \frac{\tilde{\boldsymbol{\Sigma}}^{\boldsymbol{-1}}}{\lambda}\right) \quad , \qquad (6)$$

with $\boldsymbol{\beta}$ and λ fixed parameters to be estimated using REML.

3 Material and methods

3.1 Material

The data used in this study were collected from various available statistical sources. Information on the collection of sugarcane production data in tons/ha were obtained from the Municipal Agricultural Production yearbook of the Brazilian Institute of Statistical Geography (IBGE, 2017), data that corresponds to the mesoregion of the state of Sao Paulo from the year 2006 to 2015. Information on meteorological data such as precipitation, maximum temperature and minimum temperature were obtained from meteorological stations of the National Institute of Meteorology (INMET, 2017) located in the state of Sao Paulo, the Integrated Center for Agrometeorological Information (CIIAGRO, 2017), as well as the meteorological station "Professor Dr. Jesús Marden dos Santos" of the Superior School of Agriculture "Luiz de Queiroz - University of Sao Paulo "(ESALQ, 2017). According to the Municipal Agricultural Production, the mesoregions where sugarcane production occurs and which are part of the state of Sao Paulo are: Araçatuba, Araraquara, Assis, Bauru, Campinas, Itapetininga, Litoral Sul Paulista, Marco Metropolitana Paulista, Marilia, Piracicaba, Presidente Prudente, Ribeirão Preto, São José do Rio Preto and Vale do Paraíba Paulista.

3.2 Methods

The aim of this study was to model the sugarcane production of each mesoregion and asses the effects of the average levels of temperature ($^{\circ}C$) and precipitation (mm) recorded annually from 2006 to 2015. The main focus was to explore the behavior of annual effects the change in temperature and precipitation and how they influenced the sugarcane production in each mesoregion of the state of Sao Paulo, applying GAMM with the use of the P-splines technique for each of the climatic variables and year.

As there are multiple observations of each of the mesoregions, the observations within the same mesoregion are more similar than the observations of different mesoregions. However, there is no interest in the effect of the mesoregions, but this variable will be considered as a grouping factor since it allows for a possible correlation between the observations. The general model described to be considered is expressed by

$$\begin{aligned}\text{Yield}_{ij} &= \beta_0 + \beta_1 \times \text{year}_{ij} + \beta_2 \times \text{maximum temperature}_{ij} + \\ &\quad \beta_3 \times \text{minimum temperature}_{ij} + \beta_4 \times \text{precipitation}_{ij} + \epsilon_{ij}, \qquad (7)\end{aligned}$$

where: Yield_{ij} is the jth observation in the ith mesoregion. The variables year_{ij}, $\text{maximum temperature}_{ij}$, $\text{minimum temperature}_{ij}$ and $\text{precipitation}_{ij}$ are continuous, $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 are the regression parameters; and the errors ϵ_{ij} follow a normal distribution with mean 0 and variance σ^2_{ϵ}, with $i = 1, \dots, 14$ mesoregions and $j = 1, \dots, 10$ observations corresponding to each covariate.

Fitting (7), Figure 1 shows the dispersion plots of the observations of the variables yield, precipitation, maximum temperature, minimum temperature and year. The histograms of each of these variables are also presented, which show that each of them presents the behavior of some asymmetric distribution. The behavior of the yield variable as a function of the other covariables shows that there is a possibility of a variability between these observations, some kind of smoothing can be used to take into account the variability of the data and the behavior of the curve.

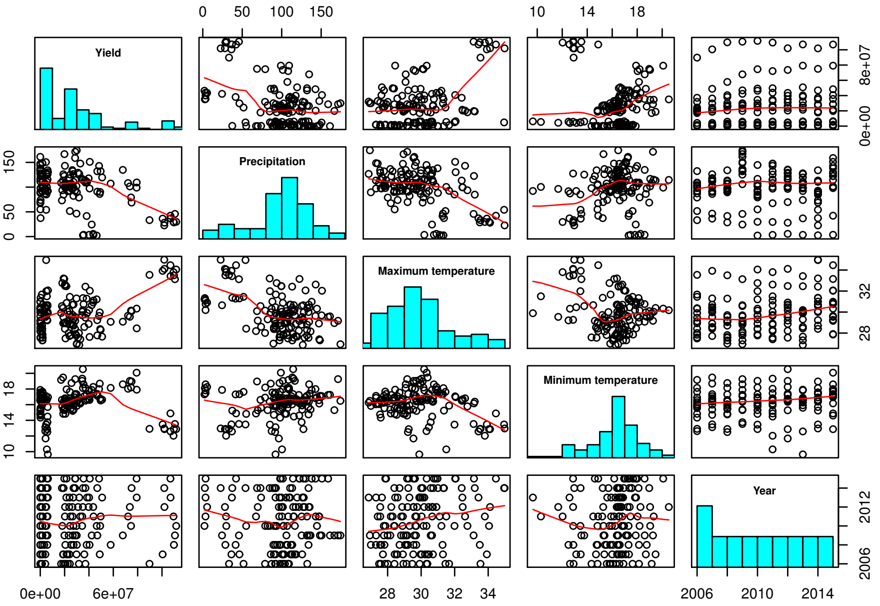


Figure 1 - Dispersion plots of the variables yield, precipitation, maximum temperature, minimum temperature and year.

We fitted model (7) to the data and inspected the half-normal plot with a simulated envelope (ATKINSON, 1985) for the deviance residuals (Figure 2). It indicated a possible lack-of-fit of the model due to the presence of correlation between the observations that was not incorporated, which indicates the need to fit differents models that accommodate these features.

Figure 3 shows the behavior of sugarcane production in each mesoregion in a period of 10 years (2006 to 2015), identified with the following codes for each mesoregion: X1 - São José do Rio Preto, X2 - Ribeirão Preto, X3 - Araçatuba, X4 - Bauru, X5 - Araraquara, X6 - Piracicaba, X7 - Campinas, X8 - Presidente Prudente, X9 - Marilia, X10 - Assis, X11 - Itapetininga, X12 - Macro Metropolitana Paulista,

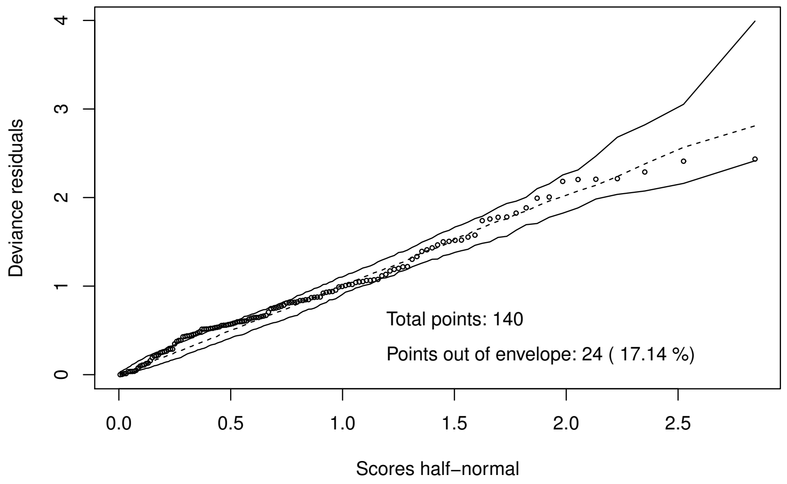


Figure 2 - Half-normal plot with simulated envelope for the residuals of model (7).

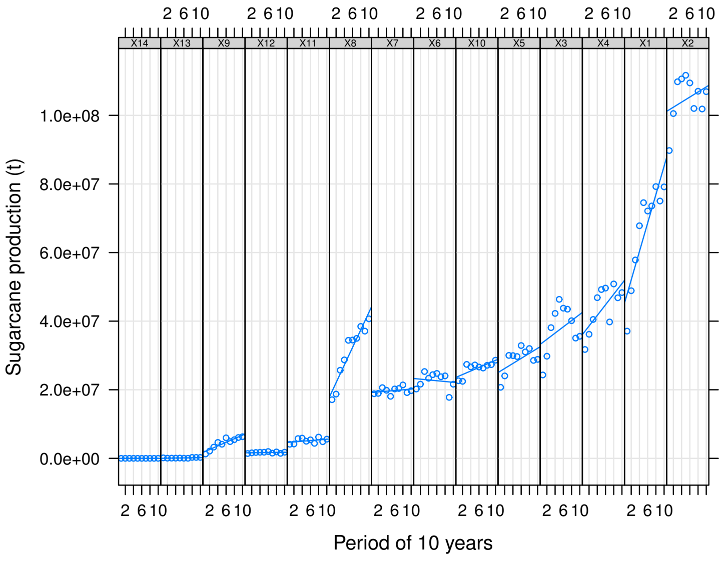


Figure 3 - Graph of the behavior of sugarcane production by mesoregion in a period of 10 years (2006 - 2015).

X13 - Litoral Sul Paulista and X14 - Vale do Paraíba Paulista. We see that the exchange rate of production in each mesoregion shows a great variation between the intercept and slope of the lines that were fitted. This would lead us to consider mixed models with random intercepts for each curve and random slopes with respect to time, in this case considering the covariate year. This will allow us that the curves do not only move in the vertical direction, but that they also oscillate.

Based on this previous analysis of the data and the ignorance of the behavior of each of the covariates that influence the yield variable, the generalized additive mixed model with P-splines was used. The purpose of using this approach is that the possible existence of heteroskedasticity is often not taken into account and one of the possible causes is that the true relationship between the response variable and the covariates is non-linear, as observed in the previous exploratory data analyses.

Therefore, to solve this problem we perform a smoothing on each covariate of the model. The proposed additive model to be used is of the form,

$$\begin{aligned}\text{Yield}_{ij} \quad = \quad & \beta_0 + f(\text{year}_{ij}) + f(\text{maximum temperature}_{ij}) + f(\text{minimum temperature}_{ij}) + \\ & f(\text{precipitation}_{ij}) + \text{a}_{i1} + \text{a}_{i2}\text{year}_{ij} + \epsilon_{ij}, \end{aligned} \tag{8}$$

where $f(\cdot)$ is the unknown smooth function for each climate covariate and for year, which shows the trend of the behavior of these variables for the yield of sugarcane. $f(\cdot)$ is estimated by means of P-splines using the representation as mixed models with the purpose of unifying the structure of the model; a_{i1} is the random intercept that measures the variability between mesoregions following a Normal distribution with mean 0 and variance $\sigma^2_{\text{mesoregion}}$; a_{i2} is the random slope for each curve in relation to the covariate year and follows a Normal distribution with mean 0 and variance σ^2_{year}, and the errors ϵ_{ij} follow a Normal distribution with mean 0 and variance σ^2_{ϵ}, with design matrixes of fixed and random effects respectively,

$$\boldsymbol{X} = \begin{bmatrix} 1 & \text{year}_{11} & \text{maximum temp}_{11} & \text{minimum temp}_{11} & \text{precipitation}_{11} \\ 1 & \text{year}_{12} & \text{maximum temp}_{12} & \text{minimum temp}_{12} & \text{precipitation}_{12} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \text{year}_{1410} & \text{maximum temp}_{1410} & \text{minimum temp}_{1410} & \text{precipitation}_{1410} \end{bmatrix}$$

and

$$\boldsymbol{Z} = \left[\ \boldsymbol{Z}_{\text{year}} : \quad \boldsymbol{Z}_{\text{maximum temp}} : \quad \boldsymbol{Z}_{\text{minimum temp}} : \quad \boldsymbol{Z}_{\text{precipitation}} : \quad \boldsymbol{Z}_{\text{a}}\ \right],$$

with random effects vetor

$$\boldsymbol{\alpha} \sim N\left[\mathbf{0}, \begin{pmatrix} \boldsymbol{I}\sigma^2_{\text{year}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}\sigma^2_{\text{maximum temp}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}\sigma^2_{\text{minimum temp}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{I}\sigma^2_{\text{precipitation}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{I}\sigma^2_{\text{mesoregion}} \end{pmatrix}\right]$$

Considering some notations in the defined matrices: maximum temp is maximum temperature and minimum temp is minimum temperature. The production response variable has Gaussian distribution that belongs to the exponential family of distributions, with function of identity binding and always taking into account that for each covariant to be smoothed, we use P-splines representation as mixed models with function of base polynomial third-degree B-splines, and a second order discrete penalty. In this approach, the autoregressive process is considered in the correlation structure of the errors that is, assuming stationarity, the correlation between residuals ϵ_i and $\epsilon_{i'}$ depends only on the difference between the time points i and i', according to Pinheiros and Bates (2000). A possible correlation structure for the residuals is the autoregressive process of order 1, AR(1). According to Zuur *e*t al. (2009), this structure considers that the error at time i depends on the error of time $i - 1$ and a small error e_i; that is

$$\epsilon_i = \rho\epsilon_{i-1} + e_i, \tag{9}$$

Therefore, the parameter ρ is unknown and needs to be estimated from the observations. Taking into account the correlation structure of the errors, we estimate the curves by smoothing each of the covariates. In each of the smooth functions of model (8) P-splines were used, since with this representation it is possible to estimate to function of each of the covariates and flexibilizes the structure of the errors simultaneously. However, if we do not use the representation of P-splines as mixed models, it would be difficult to estimate these effects.

To fit this type of models, the gamm function of the mgcv package in R (WOOD, 2017) was used. The estimates of the parameters of fixed effects and variance components were be obtained using the REML method, included within the lme function.

4 Results

In this section we describe the results of the proposed approach for this case study. We considered a smoothing function for each of the climatic variables as well as the time variable year within the additive model. The advantage of using these smooth functions is that they allow for a non linear behavior of yield over time as well as the non-linear behavior of the effects of each of the climatic variables. In addition, in this approach we considered the existence of the correlation using the autoregressive order process as expressed in (9).

The Figure 4 shows the trend of the covariate year in two situations: 1) Without considering a correlation structure in the residuals, which causes two effects: an extremely smooth curve that is a straight line (Figure 4a), with the help of a correlogram (Figure 4b), the residuals are highly correlated, that is, a clear violation of the assumption of independence of errors. 2) We use an autocorrelation structure in the model AR(1) in the residuals and we can see that in the Figure 4(c), that the behavior of the covariate year is more flexible and presents a smooth curve with 3.55 degrees of effective freedom, however we can see in the Figure 4(d) using a correlogram with autoregressive structure AR(1), still presents evidence of temporal correlation in the residuals. It can also be seen that in Figure 4(c) the shape of the smoother indicates that sugarcane production grew until 2010 and from that year onwards it was stable until 2015.

Table 1 shows the criteria for selecting models fitted: a model without considering a correlation structure in the errors and other two models using autoregressive structure AR(1) and AR(2), considering a smooth function for the covariate year. The model choice was made using selection criteria AIC and BIC. According to the two criteria used, the chosen model is the model which considers an autoregressive process AR(1).

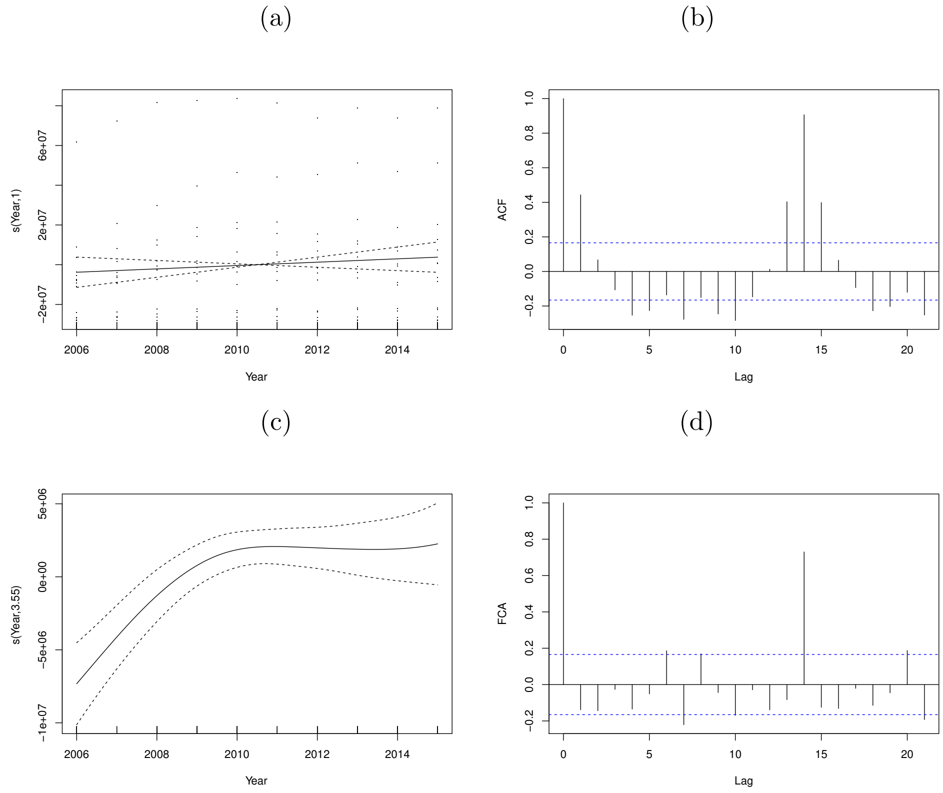


Figure 4 - Forms of the smoothing functions estimated for year with their respective autocorrelation functions.

Table 1 - Criteria for selecting models considering the autoregressive correlation structure in the errors for the covariate Year

Model	AIC	BIC
No correlation	5213.269	5225.035
AR(1)	4662.327	4677.036
AR(2)	4664.420	4682.070

The Figure 5 was obtained with the fitted model of the Equation (8) to validate the normality of the residuals and the homogeneity of variances. The plot (a) of Figure 5 shows evidence of the behavior of residuals following normal distribution. However, the plot (b) shows the standardized residuals versus fitted values and indicates evidence of the presence of heterogeneity of the variance, possibly due to the presence of the temporal correlation in the residuals.

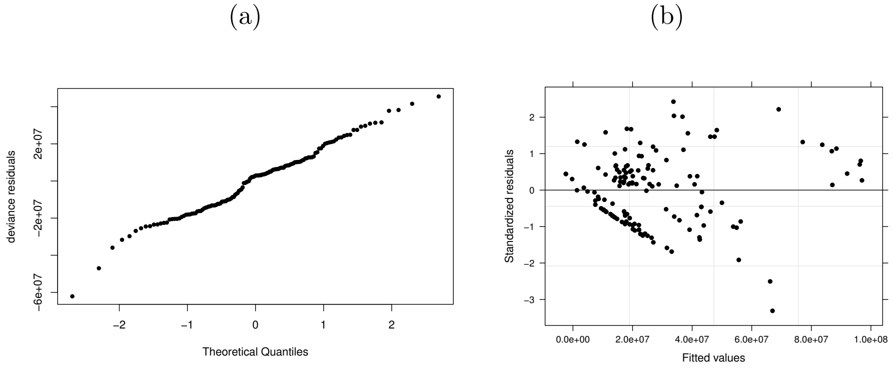


Figure 5 - Graphical validation of the model in Equation (8): (a) Normal Q-Q plot. (b) Standardized residuals versus fitted values.

Based on the choice of the GAMM model considering temporal correlation structure with autoregressive AR(1) process in the residuals, the individual covariates were included in the model.

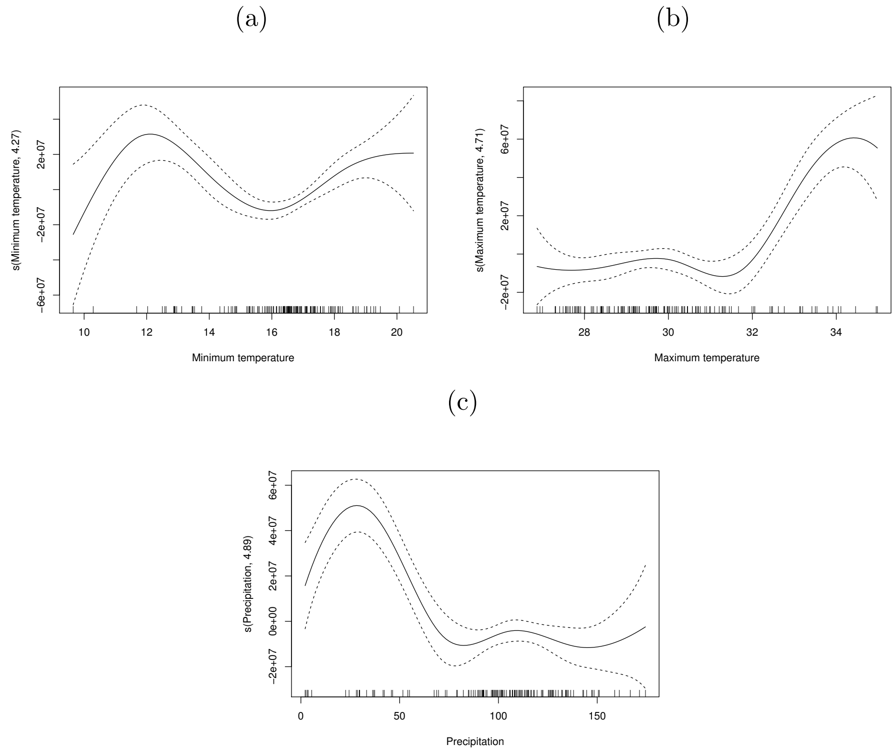
Table 2 shows the estimated smooth functions of each covariate, with their respective effective degrees of freedom (e.d.f), smoothing parameter λ, corresponding F-tests and associated p-values. We observer that the variables of interest were highly significant for the production of sugarcane. In relation to the component of the standard deviation of the random intercept, its estimated value was 1718.599; that means, the variability existing between mesoregions.

Figure 6 shows the estimated of smooth functions for each climate covariate on sugarcane production for GAMM and clearly shows evidence of the effect of each of these covariates following a non-linear pattern. In each plot of the Figure 6, the horizontal axis shows the values of minimum temperature, maximum temperature and precipitation and in the vertical axis, the contribution of function for the fitted values of

Table 2 - Estimation of the smooth function for each covariate in GAMM

Smooth function	e.d.f	λ	F	p-value
s(Maximum temperature)	4.709	1.226891	16.81	2.25e-12
s(Minimum temperature)	4.266	1.809065	9.27	1.61e-06
s(Precipitation)	4.887	0.746584	17.4	2.62e-12
s(Year)	3.555	1317.997	9.884	1.51e-06

this sugarcane production. As the effect of minimum temperature, maximum temperature and precipitation is not linear, such estimated smooth functions are significant on 4.27, 4.71, and 4.89 effective degrees of freedom respectively (a straight line would have only one effective degree of freedom). In Figure 6(a) we observer that the smoother for minimum temperature presents two periods in which the production of sugarcane had a higher production. In Figure 6(b) the estimated smooth function of the temperature variable suggests that there sugarcane production had a higher productivity as the temperature was higher and in Figure 6(c) the rainfall variable shows a very high production when the precipitation values are low.


(a)

(b)

(c)

Figure 6 - Estimated smoothing functions for minimum temperature, maximum and precipitation, the solid line is the estimated smooth function and the dotted lines are the 95% confidence.

Based on all the results obtained, Figure 7 was obtained, which shows a comparison between the predicted values of the model (denoted by the curve) versus the real data (points) for the production of sugarcane throughout the 10 years. Observing this figure, we conclude that the mesoregions Araçatuba, Araraquara, Assis, Bauru, Marilia, Piracicaba, Presidente Prudente, Ribeirão Preto and São José do Rio Preto are those that show significant evidence of higher sugarcane production over 10 years.

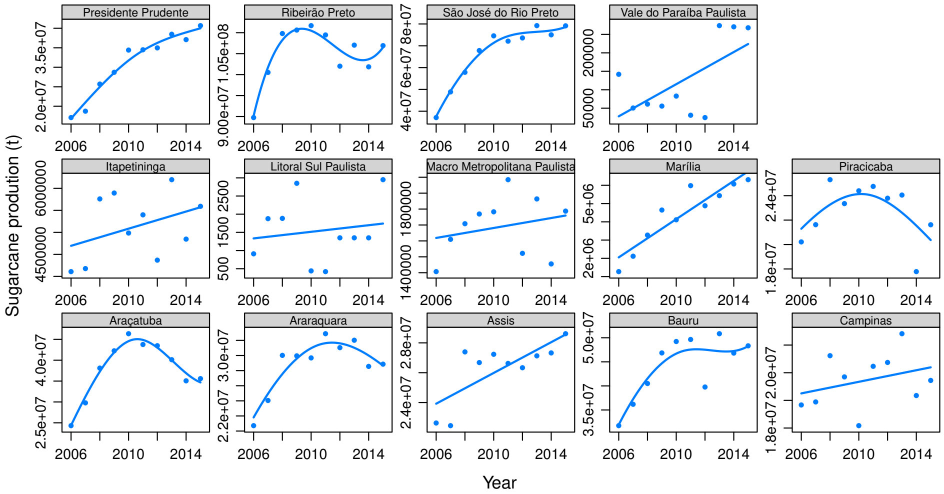


Figure 7 - Predicted values for sugarcane production over 10 years.

Conclusions

In this application, the GAMM with P-splines approach was used to relate sugarcane production variable as a function of time and climatic variables, considering these as unknown functions to be estimated, considering the random effects as independent, also it was used an temporal correlation structure with autoregressive AR(1) process in the residuals (GAMM allows to include these types of correlation in the residuals), this structure improves the adjustment of the curves for climatic variables as for the response variable. For the study case of this present work, the response followed a Normal distribution with identity linking function.

Finally, most studies that relate sugarcane production to their climatic variables use simple linear models or mixed models without taking into account the behavior of covariates, which often leads to erroneous conclusions, however using GAMM with the use of the P-*splines* technique shows excellent results since this approach takes into account the behavior of each of the covariates and how they influence the response variable.

Acknowledgments

We would like to thank CAPES Program for financial support during the development of this work and reviewers and editors for their suggestions.

RONDINEL MENDOZA, N. V.; PIEDADE, S. M. S. Modelos mistos aditivos generalizados com P-*splines* aplicados na produção de cana-de-açúcar no estado de São Paulo. *Rev. Bras. Biom.,* Lavras, v.37, n.1, p.17-31, 2019.

- *RESUMO: Neste trabalho, aplicamos os modelos mistos aditivos generalizados usando a metodologia dos P-splines como modelos mistos, que será adotado em um problema da área agroambiental, neste caso sobre os níveis médios de produção de cana-de-açúcar, que são influenciados por mudanças nas variáveis climáticas, como temperatura e precipitação que foram medidas ao longo de 10 anos em cada mesorregião do estado de São Paulo. O motivo para usar essa abordagem como um método de suavização é que a tendência dessas covariáveis climáticas não é conhecida na maior parte, mas é sabido que elas influenciam diretamente na variável resposta. Além de permitir a inclusão de efeitos fixos e aleatórios nos modelos a serem propostos, esses modelos permitem tambén a inclusão de um processo autorregressivo AR(1) como estrutura de correlação nos resíduos.*

- *PALAVRAS-CHAVE: P-splines; B-splines; modelos mistos aditivos generalizados; proceso autorregresivo AR(1).*

References

ATKINSON, A. C. *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis.* Oxford: Clarendon Press, 1985.

BRESLOW, N. E.; CLAYTON D. G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association.* v.88, p.9-25, 1993.

CIIAGRO, *Centro Integrado de Informações Agrometeorológicas.* Available in
http://www.ciiagro.sp.gov.br, 2017.

CHEN, C. Generalized additive mixed models. *Communications in Statistics - Theory and Methods*, v.29, p.1257-1271, 2007.

DE BOOR, C. *A practical guide to splines.* 1.ed. New York: Springer-Verlag, 1978, 348p.

DURBÁN, M. *Splines con penalizaciones: Teoría y aplicaciones.* Universidad Carlos III de Madrid. 2014, p.1-52. Available in http://halweb.uc3m.es/esp/Personal/personas/durban/esp/web/cursos/Psplines/material/Psplines.pdf

EILERS, P.H.C.; MARX, B.D. Flexible smoothing with B -splines and penalties. *Science*, v.11, p.89-121. 1996.

ESALQ, *Posto Meteorológico "Professor Jesús Marden dos Santos". Escola Superior de Agricultura Luiz de Queiroz - Universidade de São Paulo..* (2017). Available in http://http://www.leb.esalq.usp.br/posto/.

HASTIE, T. J.; TIBSHIRANI, R. J. *Generalized additive models.* 1.ed. London: Chapman & Hall_CRC Monographs on Statistics & Applied Probability, 1990. 352p.

IBGE - Instituto Brasileiro de Geografia e Estatística. *Produção agrícola municipal 2006 e 2015.* Available in http//www.ibge.gov.br, 2017.

INMET - Instituto Nacional de Meteorologia, Ministerio da Agricultura, Pecuária e Abastecimento. *Estações automáticas.* Available in http://www.inmet.gov.br/portal/index.php?r=estacoes/estacoesAutomaticas, 2017.

LEE, D. J.; DURBÁN, M. P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, v.11, p.49-69, 2011.

LIN, X.; ZHANG, D. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society.* Series B, v.2, p.381-400. 1999.

McCULLAGH, P.; NELDER, J. A. *Generalized linear models.* 2.ed. London: Chapman and Hall-CRC Monographs on Statistics and Applied Probability 37, 1989. 511p.

PINHEIRO, J. C.; BATES, D. M. *Mixed-effects models in S and S-PLUS.* 1.ed. New york: Springer-Statistics and computing **1**. 2000. 527p.

RONDINEL MENDOZA, N. V. *Estruturas unidimensionais e bidimensionais utilizando P-splines nos modelos mistos aditivos generalizados com aplicação na produção de cana-de-açúcar*, 2017. 83p. Thesis (Ph.D.) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Brazil, 2017.

WOOD, S. N. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, v.62, p.1025-1036, 2006.

WOOD, S. N. *mgcv: Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation.* R package version 1.8-20. Available at https://cran.r-project.org/web/packages/mgcv/index.html, 2017

ZHANG, D.; LIN, X.; RAZ, J.; SOWERS, M. Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association.* v.93, p.710-719. 1998.

ZEGER, S. L.; DIGGLE P. J. Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics.* v.50, p.689-699. 1994.

ZUUR, A. F.; IENO, E. N.; WALKER, N. J.; SAVELIEV, A. A.; SMITH, G. M. *Mixed effects models and extensions in ecology with R.* 1.ed. New York: Springer, 2009. 580p.

Received on 20.10.2017.

Approved after revised on 07.08.2018.