Research, Society and Development, v. 11, n. 6, e31011629160, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i6.29160

# Use of generalized additive mixed models in the comparison of mortar types in different configurations

## Uso de modelos aditivos generalizados mistos na comparação de tipos de argamassa em diferentes configurações

## Uso de modelos aditivos generalizados mixtos en la comparación de tipos de mortero en diferentes configuraciones

Received: 04/12/2022 | Reviewed: 04/20/2022 | Accept: 04/25/2022 | Published: 04/28/2022

**Breno Gabriel da Silva**
ORCID: https://orcid.org/0000-0002-8322-9235
University of São Paulo, Brazil
E-mail: brenogsilva@usp.br
**Willian Luís de Oliveira**
ORCID: https://orcid.org/0000-0002-2941-9804
State University of Maringá, Brazil
E-mail: wloliveira@uem.br
**Terezinha Aparecida Guedes**
ORCID: https://orcid.org/0000-0001-5364-609X
State University of Maringá, Brazil
E-mail: taguedes@uem.br
**Vanderly Janeiro**
ORCID: https://orcid.org/0000-0001-8804-6107
State University of Maringá, Brazil
E-mail: vjaneiro@uem.br
**José Aparecido Canova**
ORCID: https://orcid.org/0000-0003-1750-7572
State University of Maringá, Brazil
E-mail: jacanova@uem.br

**Abstract**
Generalized additive mixed models (GAMM) are useful when the nature of the data is longitudinal, which include the nonlinearity of the individual trajectories of the subjects, associating these with the assumption of the existence of random effects for each individual. In these models it is possible to rewrite the predictor as a sum of smooth nonparametric functions and then use the smooth technique P-spline. Thus, using the generalized additive mixed models with P-splines, this article aims to verify the impact of the types of mortars and the concentration levels of an additive content on the evolution of mortar weight contained in specimens over time, verifying possible differences between combinations of types and concentration. In relation to the results, it was observed through the analyzes that there were significant differences in the estimated curves of the weight evolution in both types of mortar, concluding that the matured mortar has a water absorption speed by capillarity higher than dry mortar. All the necessary assumptions for the validity of the model have been satisfied.
**Keywords:** Generalized additive mixed models; Longitudinal data; Nonparametric regression; P-splines.

**Resumo**
Os modelos aditivos generalizados mistos (GAMM) são úteis quando a natureza dos dados é longitudinal, que inclui a não linearidade das trajetórias individuais dos sujeitos, associando-as à suposição da existência de efeitos aleatórios para cada indivíduo. Nesses modelos é possível reescrever o preditor como uma soma de funções não paramétricas suaves e então usar a técnica de suavização P-spline. Assim, utilizando os modelos mistos aditivos generalizados com P-splines, este artigo tem como objetivo verificar o impacto dos tipos de argamassas e dos níveis de concentração de um teor de aditivo na evolução do peso da argamassa contida nos corpos de prova ao longo do tempo, verificando possíveis diferenças entre combinações de tipos e concentração. Em relação aos resultados, observou-se através das análises que houve diferenças significativas nas curvas estimadas da evolução de massa nos dois tipos de argamassa, concluindo que a argamassa maturada possui uma velocidade de absorção de água por capilaridade superior à argamassa seca. Todas as suposições necessárias para a validade do modelo foram satisfeitas.
**Palavras-chave:** Modelos aditivos generalizados mistos; Dados longitudinais; Regressão não paramétrica; P-splines.

**Resumen**
Los modelos aditivos generalizados mixtos (GAMM) son útiles cuando la naturaleza de los datos es longitudinal, que incluyen la no linealidad de las trayectorias individuales de los sujetos, asociándolas a la suposición de la existencia de efectos aleatorios para cada individuo. En estos modelos es posible reescribir el predictor como una suma de funciones no paramétricas suaves y luego usar la técnica suave P-spline. Así, utilizando los modelos mixtos aditivos generalizados con P-splines, este artículo tiene como objetivo verificar el impacto de los tipos de morteros y los niveles de concentración de un contenido de aditivo en la evolución del peso del mortero contenido en las muestras a lo largo del tiempo, verificando posibles diferencias entre combinaciones. de tipos y concentración. En relación a los resultados, se observó a través de los análisis que existían diferencias significativas en las curvas estimadas de evolución del peso en ambos tipos de mortero, concluyendo que el mortero madurado tiene una velocidad de absorción de agua por capilaridad superior al mortero seco. Se han satisfecho todos los supuestos necesarios para la validez del modelo.

**Palabras clave:** Modelos aditivos generalizados mixtos; Datos longitudinales; Regresión no paramétrica; P-splines.

## 1. Introduction

Regression analysis is a statistical technique used to investigate and to model the possible relationship between a dependent variable and one or more independent variables, which way be linear or nonlinear. In the nonlinear regression models, the observational data is modeled by a function that is a nonlinear combination of the model parameters, and in general, the usual assumption about the normality of error terms is considered in the simplest models, which is sometimes not checked depending on the nature of the dependent variable. Ruppert et al. (2003) describe that in the past, due to the lack of computational resources and statistical approaches that contemplated certain observed characteristics, many of the approaches were used through transformation in the response variable, but, this fact is associated with the loss of information mainly in data that nonlinear effects.

In the literature, it is not difficult to find data with nonlinear and/or longitudinal characteristics, being modeled by a simple regression models. For example in Canova et al. (2009) and Canova et al. (2015), the evolution of mortar weight over time, in different configurations, is modeled using simpler models. However, these models bring with them the loss of information, since they do not consider in their forumation more than one value for each individual i, at time j, being necessary to take the average of the observations in each i, at time j. In these cases, the ideal would be adopt a model that considers the specifics of each individual, which would present an efficiency gain in terms of adjustment.

Given what has been described, considering the hypotheses that the nature of the data is complex, some classes of regression models are available, such as Generalized Additive Models (GAM's) proposed by Hastie and Tibshirani (1990), in which these have been widely used in research involving longitudinal data or data with the need to use this class of models (Benedetti et al. 2009; de Jong et al. 2015; Shadish et al. 2014; Nores & Díaz, 2016; Zhang et al. 2020; Gressani et al. 2021), whereas this class of models has greater flexibility when compared to the usual parametric regression methods, since the parametric methods require the researcher to know the functional form of the trend that may be present in the data.

In recent years, with the advancement of computational techniques and the need to describe nonlinear trends, the functions splines has also been used frequently in the analysis of longitudinal data (Achmad et al. 2018; Andrinopoulou et al. 2018; Garcia-Hernandez & Rizopoulos, 2018; Prawanti et al. 2019; Islamiyati et al. 2019). Toshniwal et al. (2017) defines the functions splines as polynomial functions by parts capable of modeling complex curves, choosing different points in the observation range, which Keele and Keele (2008) denominates as knots, where such polynomial functions are adjusted to each interval to be modeled.

According Ruppert et al. (2003), several techniques are found to represent this adjustment by parts, such as linear spline, cubic spline, B-spline among others. Regarding the existing techniques in the literature to represent this adjustment by parts, according Momen et al. (2019) a function B-spline can be expressed as a set of polynomial functions of order m for each interval limited by the knots, where, the connections in the knots are smoothed which does not occur in the linear spline, thus becoming a single continuous curve. In order to provide greater flexibility to the model and avoid a possible overfitting, Eilers and Marx (1996) proposed a new technique called P-spline, that consists of the union of B-splines with discrete penalty (to control the smoothness of the adjustment) inserted in the log-likelihood function. Over the past decades, progress has been made in the use of P-splines in the context of generalized additive models, due to the possibility of rewriting the linear term by summing up smooth functions not known, not specified in a parametric context, ie, replace $X_i'\beta = \sum x_{ij}\beta$ in the predictor with sum $\sum f_j(x_{ij})$.

With the purpose of introducing in a model the effect of the presence of overdispersion, correlation and the random effect to the additive predictor using nonparametric functions, Lin and Zhang (1999) extended the generalized additive models and proposed a new class of models called Generalized Additive Mixed Models (GAMM), where among the countless potentialities existing in this new class of models, the possibility of these being used in studies with experimental designs, whether nested or cross-checks, hierarchical data, grouped data and spatial data (Durbán et al. 2005; Polansky & Robbins, 2013; McKeown & Sneddon, 2014; Baayen et al. 2018; da Silva et al. 2020; Sudo et al. 2021). It is highlighted in Durbán et al. (2005) that in this class of models it is possible to estimate the curves of individual differences using nonparametric smooth functions, considering the presence of random effects associated with individuals classified between treatment groups.

Therefore, to avoid applying transformations in the response variable and the use of simpler parametric models that cause the loss of information, common in problems involving data with nonlinear and/or longitudinal characteristics, as noted in Canova et al. (2009) and Canova et al. (2015), we propose in this article the use of GAMM with P-splines, to model the relationship between the different configurations of a level of ground rubber powder from waste tires, added to two types of mixed mortar, contained by weight of specimens over time, in order to identify which of these admit a faster reduction of the requirement of water to the standard consistency defined by Rilem (1994). The remaining sections of the article are divided as follows. In Section 2.1, the definitions and specifications of the additive models and the generalized additive mixed models are presented. Section 2.2 describes the P-splines functions. In Section 2.3, the inferential part of the model considered is described. In Section 3 the proposed model is applied to a real dataset that concerns the evolution of the weight of specimens containing mortars of different configurations over time and finally, in Section 4 a conclusion is presented.

## 2. Methodology

### 2.1 Additive models and semiparametric mixed models

The Generalized Additive Model (GAM) is described by Wood (2017) as having a structure

$$g(\mu_i) = \mathbf{X}_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \cdots$$
$$\mu_i \equiv \mathbb{E}(y_i),$$

where $y_i$ is the response variable belonging to the exponential family, $f_j$ are smooth functions of the covariates $x_k$, $\boldsymbol{\theta}$ corresponds to the parameter vector and $\mathbf{X}_i^*$ is defined as the model matrix for any strictly parametric component. Mentions Hastie & Tibshirani (1990) that these models have the attractive feature of modeling the effects of covariates on the response as a sum of individual effects. In this context, it is possible to include random effects peculiar to the subjects, having a new class, called Generalized Additive Mixed Models (GAMM), thus allowing greater flexibility, since among its potentialities, the fact of the relation between the response variable and some explanatory variables admit parametric form, whereas the relationship between the mean of the response variable and the rest of the covariables can be nonlinear, combining the parametric and nonparametric components in a single model, as seen in Hernando & Paula (2016). In general, GAMM's are defined by Wood (2017) as

$$\begin{aligned} g(\mu_i) &= \mathbf{X}_i^{*}\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \cdots + \mathbf{Z}_i\mathbf{u} + \boldsymbol{\epsilon}_i \\ \mu_i &\equiv \mathbb{E}(y_i), \end{aligned} \tag{1}$$

where $y_i$ is the response variable, q is a fixed parameter vector, $\mathbf{X}_i$ are the matrix lines of the fixed effects of the model, $\mathbf{f}_j$ are the smooth functions of covariates $x_k$, $\mathbf{Z}_i$ are the matrix lines of the random effects of the model, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi})$ is the vector of random effect coefficients with a defined positive unknown covariance matrix $\mathbf{\Psi}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$ is the vector of the error terms with covariance matrix $\mathbf{\Lambda}$. The model defined in (1) it admits as particular cases models in which the response has a normal distribution and the link function is identity, for example, the model expressed in (2), which has random intercept and slope. This can be used in situations involving longitudinal data, where Durbán et al. (2005) defines it by:

$$\begin{aligned} y_{ij} &= f(x_{ij}) + U_i + \epsilon_{ij}, \text{ with } U_i \sim \mathcal{N}(0, \sigma_U^2) \text{ and } \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2), \\ f(x_{ij}) &= \beta_0 + \beta_1 x_{ij} + \textstyle\sum_{k=1}^{K} u_k (x_{ij} - \kappa_k)_+, \text{ with } u_k \sim \mathcal{N}(0, \sigma_u^2), \end{aligned} \tag{2}$$

where $y_{ij}$ is the response variable associated with the i-th individual and at the j-th observed moment, $f(x_{ij})$ are smooth functions that describe the behavior of the response variable at the instant, that will be estimated using P-splines, with i = 1,…,m and j = 1,…, $n_i$, $U_i$ is the term of random effect for each individual i, being characterized by containing only a single parameter, $\sigma_U^2$, which is usually called the variance component and $\kappa_k$ represents a set of distinct inside the range of the $x_{ij}$, being fixed and large enough the number of K-knots, chosen as quantiles of $x$ with probability 1/(K+1),…,K/(K+1). Highlights Ruppert et al. (2003), $x_+$ = máx(0, x), ie, for any number $x$, $x_+$ is equal to $x$ if $x$ is positive and is equal to 0 otherwise. Therefore, mentioned authors further define that the $U_i$ are treated as a random sample of distribution $\mathcal{N}(0, \sigma_U^2)$ for some $\sigma_U^2$ > 0, in which in this model the behavior of the response variable of the i-th individual is modeled considering a random intercept, causing the non-description of the individual trajectories. Thus, the idea is to add terms associated with the different inclinations of the trends, making the model more flexible and allowing the curves not to admit parallelism behavior, which is expressed by

$$y_{ij} = f(x_{ij}) + a_{i1} + a_{i2}x_{ij} + \epsilon_{ij}, \text{ with } (a_{i1}, a_{i2})' \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \text{ and } \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{3}$$

where $a_{i1}$ and $a_{i2}$ are defined as the individual random intercept and the individual random slopes of each trajectory respectively. However, the appropriate model for such a situation is one that allows estimating the curves of specific differences between individuals using smooth nonparametric terms, in which the objective is to describe the average individual trajectories and to verify differences in each of these groups, where this model can be described as

$$\begin{aligned} y_{ij} &= f(x_{ij}) + g_i(x_{ij}) + \epsilon_{ij}, \text{ with } \epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2), \\ g_i(x_{ij}) &= a_{i1} + a_{i2}x_{ij} + \textstyle\sum_{k=1}^{K} \nu_{ik} (x_{ij} - \kappa_k)_+, \text{ with } \nu_{ik} \sim \mathcal{N}(0, \sigma_\nu^2), \end{aligned} \tag{4}$$

where $f(x_{ij})$ is a smooth function that explains the trajectory of each individual and $x_{ij}$ indicates the i-th sampling unit observed at the j-th instant. Curves specific to sample units $g_i(.)$, defined in model (4), have parametric $a_{i1} + a_{i2}x_{ij}$ and nonparametric $\sum_{k=1}^{K} v_{ik}(x_{ij} - \kappa_k)_+$ random components, with the linear part of the regression spline also being random and not a fixed effect. The terms $a_{i1}$ and $a_{i2}$ represent the individual random intercept and the individual random slopes for each curve respectively which follows normal distribution with vector of means 0 and matrix of variances and covariance $\boldsymbol{\Sigma}$. However, as we are interested in identifying possible differences between the combinations of treatments over time, it is necessary to adjust a model that describes the mean curves of each of the specimens in their respective treatment, that is, taking into account consideration that there is the presence of the random effect to each individual and that there is interaction between groups of treatments with the continuous predictor, in which the model (4) can be extended to

$$y_{ij} = f_{z_i}(x_{ij}) + g_i(x_{ij}) + \epsilon_{ij},$$

$$f_{z_i}(x_{ij}) = \beta_0 + \beta_1 x_{ij} + \sum_{k=1}^{K} u_k(x_{ij} - \kappa_k)_+ + \sum_{l=2}^{L} z_{il}(\gamma_{0l} + \gamma_{1l}x_{ij}) + \sum_{l=2}^{L} z_{il}\left\{\sum_{k=1}^{K} w_k^l(x_{ij} - \kappa_k)_+\right\}, \tag{5}$$

with $v_{ik} \sim \mathcal{N}(0, \sigma_v^2)$; $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$; $w_k^l \sim \mathcal{N}(0, \sigma_w^2)$ and $u_k \sim \mathcal{N}(0, \sigma_u^2)$, where $z_{il}$ = 1 if $z_i = l$ and $z_{il}$ = 0 otherwise. In this model, we have a common variance parameter Var($w_k^l$) = $\sigma_w^2$ , l = 2,…,L for all curves, so it is assumed that the curves have similar smoothness, however, the random effects are independent of the smooth functions, resulting in the fact that the curves will be different. It should also be noted that the parameter $\gamma_{jl}$ suffers restrictions so that fixed effects are identified, so it is assumed that $\gamma_{0l}$ = $\gamma_{11}$ = 0, that is, $\beta_0 + \beta_1 x_{ij} + \sum_{k=1}^{K} u_k(x_{ij} - \kappa_k)_+$ is the estimated curve for l = 1 and $\gamma_{0l} + \gamma_{1l}x_{ij} + \sum_{k=1}^{K} w_k^l(x_{ij} - \kappa_k)_+$ will be the different curves estimated for the treatment levels. Model defined in (5) can be written in matrix form, in the context of P-splines, as a mixed models as follows

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\epsilon} \tag{6}$$

where the random effects matrix **Z** is defined by

$$\mathbf{Z} = \begin{pmatrix} Z_1 & X_1 & 0 & \cdots & 0 & Z_1 & 0 & \cdots & 0 \\ Z_2 & 0 & X_2 & \cdots & 0 & 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Z_m & 0 & 0 & \cdots & X_m & 0 & 0 & \cdots & Z_m \end{pmatrix},$$

the random effects vector **u** is expressed by

$$\mathbf{u} = (u_1, \dots, u_k, a_{11}, a_{12}, \dots, a_{m1}, a_{m2}, v_{11}, \dots, v_{mk})',$$

and the covariance matrix **G** is written as

$$\mathbf{G} = \mathrm{Cov}(\mathbf{u}) = \begin{pmatrix} \sigma_u^2\mathbf{I} & 0 & 0 \\ 0 & \underset{1 \le i \le m}{\mathrm{blockdiagonal}\,\Sigma} & 0 \\ 0 & 0 & \sigma_v^2\mathbf{I} \end{pmatrix} \tag{7}$$

It should also be noted that for certain values of $\sigma_U^2$ and $\sigma_\epsilon^2$ , estimates of $(\boldsymbol{\beta}, \mathbf{u})$ are obtained by the method of penalized least squares defined by

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{y_{ij} - f(x_{ij}) - U_i\}^2 + \lambda u'u + \frac{\sigma_\epsilon^2}{\sigma_U^2} + U'U,$$

where $\lambda \mathbf{u}'\mathbf{u}$ controls the over parameterization of the regression function by placing a penalty on the smoothness of $\mathbf{u}_k$ and then getting smooth curves and $\lambda$ is the smooth control parameter expressed by $\lambda = \frac{\sigma_\epsilon^2}{\sigma_U^2}$. Mentions Durbán et al. (2005) that smooth through penalized splines corresponds to the ideal predictor in a mixed model structure assuming $\mathbf{u}_k \sim \mathcal{N}(0, \sigma_u^2)$.

### 2.2 P-spline

According to Currie & Durbán (2002), P-spline is defined as a combination of B-splines and a function of penalizing differences of order associated with the estimated coefficients of the bases B-splines, where the use of this tool helps to reduce the flexibility of B-splines thus avoiding over-adjusting the curve. Thus, the use of P-splines allows the researcher to be free to choose the number of knots. The authors Eilers & Marx (1996) present in their article an interesting result regarding modeling involving the combination of B-splines, with respect to the integral of the second squared derivative can be expressed as a quadratic function in the coefficients associated with the sum, where $f(x) = \sum_{j=1}^{n} \boldsymbol{\theta}_j \mathbf{B}_j(x)$, that is, it is a function that composes sums of B-splines. Highlight Eilers & Marx (1996) that for normally distributed data, the model in its matrix form is defined by

$$f(x) = \mathbf{B}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2),$$

where $y = (y_1, \dots, y_n) \in \mathbb{R}^n$, $f(x) = \mathbf{B}\boldsymbol{\theta}$ the curve of model and $\mathbf{B}$ is defined as the regression base function built from the variable $\mathbf{x}$. To estimate the regression coefficients, the function of the penalized sum of squares is minimized. Thus, the function of penalized least squares is expressed by

$$S(f) = (y - \mathbf{B}\boldsymbol{\theta})'(y - \mathbf{B}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}'\mathbf{P}_d\boldsymbol{\theta},$$

where, $\mathbf{P}_d = (\Delta^d)'(\Delta^d)$ is the matrix that penalizes the coefficients, d is defined as the order of the difference's operator $\Delta^d$ for the components $\theta_j$ of B-splines, this being rewritten as a dimension matrix $[(q + 1 + p - d) \times (q + 1 + p)]$, where, $q \in \mathbb{z}$ and $d = p - 1,$ $\lambda$ the smooth control parameter and we still have to $\Delta^d$ is recursively calculated by

$$\Delta^d = \Delta(\Delta^{d-1}\theta_j).$$

One of the advantages of applying a penalty to the coefficients B-splines is the reduction of the dimensionality of the overfitting problem, as it does not depend on the degree adopted for B-splines, in such a way that it is possible to combine any order of the penalty with any order of the bases B-splines. An interesting property inherent in the use of P-splines is that they can conserve the moments of the data, that is, the average and variance of the estimated curve will be similar to those of the data. Durbán (2007) presents a result regarding the degrees of freedom of adjustment, which he calls effective degrees of freedom, which consists of a good approximation of the dimension of the vector of parameters defined by the smooth matrix, in which the trace of the matrix $\mathbf{H}_\lambda$ will depend on the smooth parameter $\lambda$ and more, the positive values of $\lambda$ refers to effective degrees of freedom, that is, whether $\lambda$ is very small, we have that the trace$(\mathbf{H}_\lambda) = \mathrm{d} + 1 + \kappa$, where $\kappa$ is the number of knots. Already is $\lambda$ is very large is trace$(\mathbf{H}_\lambda) = \mathrm{d} + 1$. Durbán et al. (2005) mention that if the smooth parameter $\lambda$ equals zero, then the degrees of freedom of adjustment of the model configure the dimension of the base matrix $\mathbf{B}$ subtracting the number of model constraints, however, if the smooth parameter $\lambda$ is very large, the model is said to be not very flexible and therefore it will have very few degrees of freedom, we still have to define the degrees of freedom of adjustment in this type of model using the matrix trace $\mathbf{H}_\lambda$. Regarding the selection of the smooth control parameter, several criteria are found in the literature, one of which is the Cross Validation Criterion, defined by Craven & Wahba (1978). There is also a result that generalizes the Cross Validation Criterion, denominated Generalized Cross-Validation Criterion, proposed by Friedman et al. (2001)

$$\mathrm{GCV}(\lambda) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left\{\frac{\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}}{1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{H}_\lambda^{\mathrm{ii}}}{\mathrm{n}}}\right\}^2 = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left\{\frac{\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}}{1 - \frac{\mathrm{trace}(\mathbf{H}_\lambda)}{\mathrm{n}}}\right\}^2,$$

with $\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{H}_\lambda^{\mathrm{ii}}$, is the trace$(\mathbf{H}_\lambda)$. The authors Friedman et al. (2001) emphasize that the best $\lambda$ will be the one that minimizes $\mathrm{CV}(\lambda)$ or $\mathrm{GCV}(\lambda)$.

### 2.3 Inference

Corbeil and Searle (1976) describe that the estimates obtained by the maximum likelihood method have bias, since the degrees of freedom that are used to estimate fixed effects are not considered. Thus, instead of using the maximum likelihood method, the restricted maximum likelihood method (REML) of Patterson and Thompson (1971), with regard to this method, the degrees of freedom of estimation of fixed effects are taken into account. Thus, considering the model defined in (4), the restricted log-likelihood function is defined by

$$\mathrm{l_R}(\sigma_\mathrm{u}^2, \sigma_\mathrm{v}^2, \sigma_\epsilon^2) = \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}\mathbf{y}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{y},$$

where $\mathbf{V} = \mathbf{ZGZ}' + \sigma_\epsilon^2\mathbf{I}$ is the estimates of the variance components and $\boldsymbol{G}$ is the covariance matrix defined in (7). Therefore, the parameter vector $\boldsymbol{\beta}$ and the random effects vector $\mathbf{u}$ are obtained by

$$\mathrm{BLUE}(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

$$\mathrm{BLUP}(\mathbf{u}) = \hat{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Mentions Schaeffer (2004) what $\hat{\boldsymbol{\beta}}$ contained in the BLUP of u is replaced by $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and more, the BLUE of $\boldsymbol{\beta}$ is similar to the generalized least squares solution. According Durbán et al. (2005), evaluating the quality of fit of a nonparametric model with a parametric model is not a simple task. Consider the model defined in (2), the idea then is to verify if the function that describes the population average is linear or if there is any evidence of nonlinearity, that is, to compare the dissimilarities between models of this class, being these nested, containing additional terms regarding the observed specificities regarding the behavior of the dependent variable, in which the test hypotheses are given by

$$H_0\colon \sigma_u^2 = 0 \ vs. H_1\colon \sigma_u^2 > 0.$$

The authors Ruppert et al. (2003) mention that one of the problems found when comparing these models is that the parameter of interest is defined in the parametric space [0; ∞), so they define the restricted likelihood ratio statistic as

$$RLRT = \sup_{H_0} REL\ (\beta, \sigma_\epsilon^2, \sigma_U^2, \sigma_u^2) - \sup_{H_1} REL\ (\beta, \sigma_\epsilon^2, \sigma_U^2, \sigma_u^2)$$

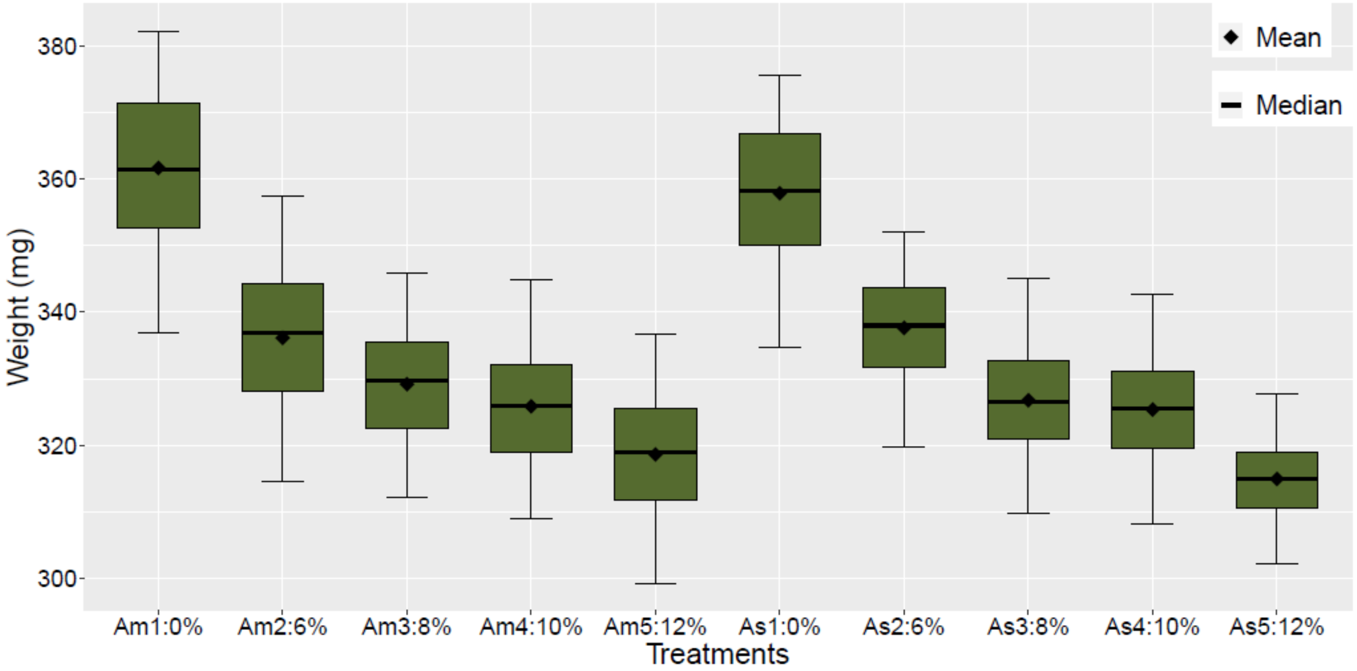
in which it cannot be compared with $\chi_1^2$. Highlights Durbán et al. (2005) that due to the specificities of the model described here and the sample size to be studied, the implementation of the aforementioned method may not be trivial, thus, these same authors suggest that the restricted likelihood ratio statistic be compared with an approximation of the mixture to the chi-square distribution.

## 3. Results and Discussion

The experiment that gave rise to the data set of this work was performed by Canova et al. (2009). The mortars weight in specimens was observed over time, considering 10 levels for the treatment variable (Am1:0%, Am2:6%, Am3:8%, Am4:10%, Am5:12%, As1:0%, As2:6%, As3:8%, As4:10%, As5:12%), this being the combination between two types of mortar and five concentrations of a rubber powder content, where the terms Am and As represent two types of mortar, matured mortar (Am) and kiln-dried mortar (As) respectively, in which the first type was produced containing lime and fine sand washed from the river, which after receiving the addition of cement, became a compound mortar. The second type was produced containing lime and sand that passed through the maturation process, was kiln dried and received the addition of cement addition.

The experiment was based on the procedure of NBR 9779/1995. Thus, the times for the readings of the weight of the specimens were defined below: up to 90 minutes every 10 minutes, from 90 to 150 minutes every 15 minutes, from 150 to 360 minutes every 20 minutes, from 360 to 450 minutes every 45 minutes and 450 to 1350 minutes every 60 minutes. However, it was found that in the first 10 minutes it showed a dissimilar growth compared to the other periods, that is, there was an absorption of water by accelerated capillarity in both types of mortar, this being an atypical behavior for this situation, though that, common in longitudinal planned experiments as described Pinheiro & Bates (2006). Therefore, to avoid possible problems in the adequacy of the models, we have removed the initial evaluation time. The analyzes were performed in the software R (R Core Team, 2022). The Figure 1 presents the box-plot of the response variable for each of the treatments, in which it can be concluded that at the highest concentration levels, the average weight of the specimens decreases.

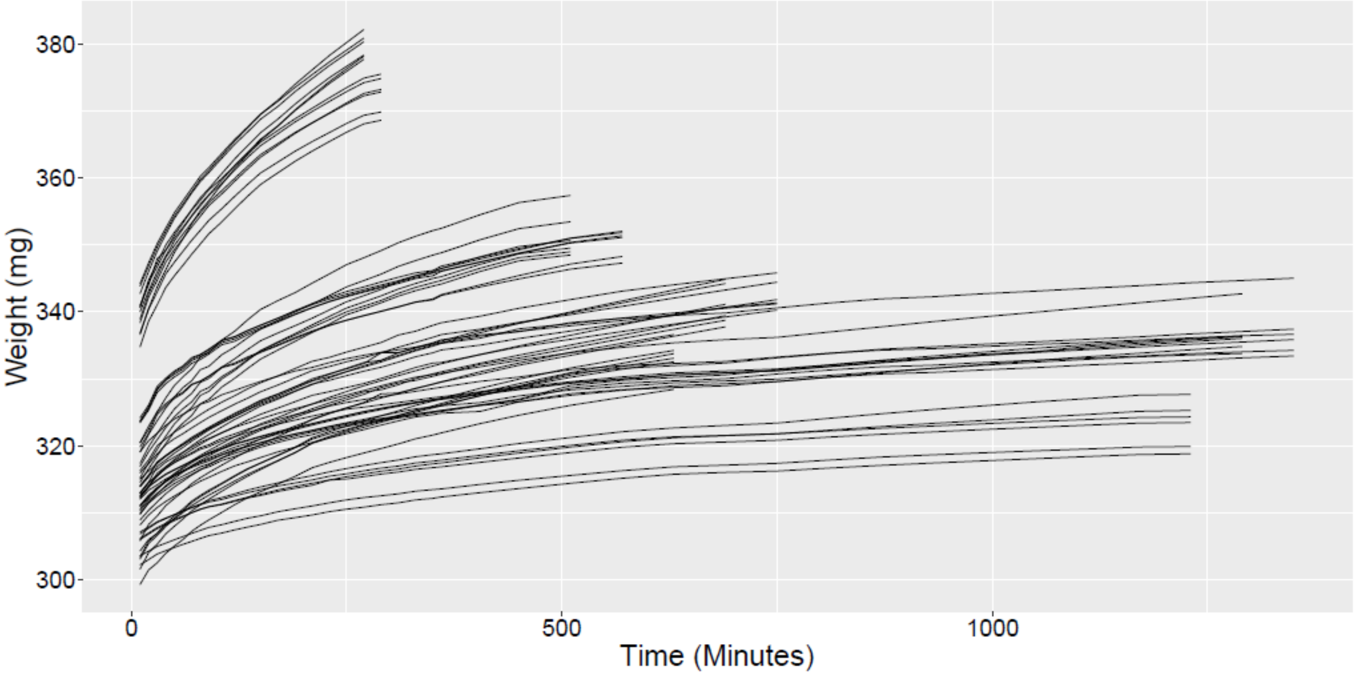
**Figure 1:** Box-plot of mortar weight by treatments.



Source: Authors.

The Figure 2 it presents the graphs of the average profiles of the response variable for each of the specimens, with an increasing behavior as time varies, and moreover, an oscillatory behavior is seen even in the initial instant.

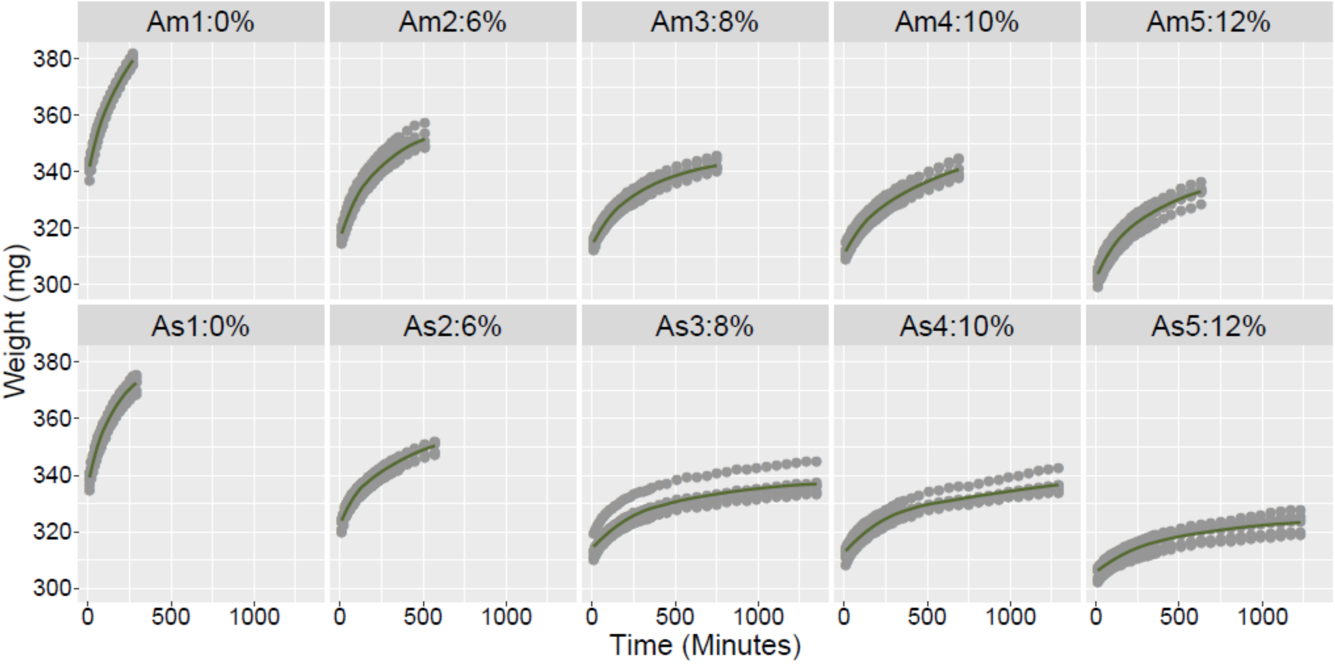
**Figure 2:** Average profile of mortar weight for each sample unit over time.



Source: Authors.

The Figure 3, presents the scatterplot and the fitted global trends by treatments, obtained using the method LOWESS (locally weighted scatterplot smoothing), verifying a number of observations that are outside of the fitted curves, that is, there is a variability between these observations, justifying the fact that you can use some model that takes this into account.

**Figure 3:** Scatterplot of mortar weight by treatments with fitted curve obtained using the method LOWESS (locally weighted scatterplot smoothing).



Source: Authors.

It is also verified that the weight observed after the absorption of water by capillarity in each type of mortar, presents a lot of variation between the intercept and inclination, evidencing the fact that linear models will not present a satisfactory fit, result that was verified but is not being presented in this work. As a consequence, we considered the mixed models with random effects. In the past with the absence of computational resources and methodologies that could solve the problem previously described, this fact was commonly treated by making transformations in the response variable and using simpler models such as the linear and polynomial regression models, but in these cases does not provide a satisfactory fit, leading to mistaken conclusions because the characteristics already mentioned are present in the data set and are not being captured by the adopted methodology.

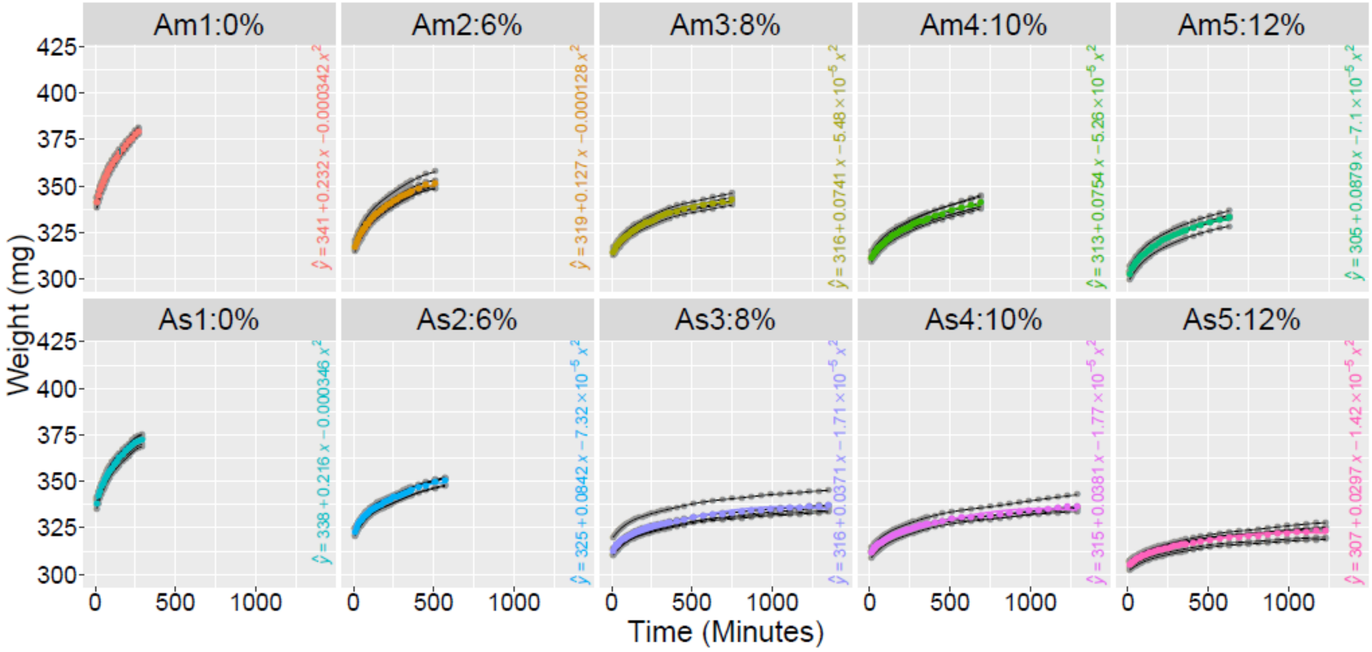
Considering the model defined in (2), it was noted that it is not useful to describe the trajectories of the weights of the specimens containing the mortars, since it is assumed that the trend over time assumes a linear behavior, which in practice does not occur. In order to describe the trajectories over time respecting their specificities, bypassing the problems found previously, Durbán et al. (2005) describes that when using the P-splines smooth functions through the representation of mixed models, these instead of assuming that the trend over time is linear, assume smooth curves that only differ in their intercepts. Therefore, the model defined in (3) was adjusted but, it was noted that even considering the smoothed curves, it was still not possible to describe the individual trajectories.

So, the model (4) is seen as a candidate for such a situation, since this brings as a characteristic the incorporation of a function g(.) nonparametric smooth, however, when including the interaction between treatments with the continuous predictor in this model, it is possible to describe the trajectories of the evolution of the weight of the types of mortar after the absorption of water by capillarity over time considering the possible existence of heteroscedasticity. In view of what has been described, the model (5) describes these taking into account nonlinearity and the associated random effects.

Figure 4, shows the fitted curves obtained from the model (5), considering 10 knots. It can be noted that the model fit was satisfactory in terms of interpolation, and yet, the specimens with matured mortar, added the contents, in a good part of the study admit a superior evolution in the speed of absorption of water by capillarity, being observed that in all the concentrations of contents, in average, the bodies- evidence did not reach the end of the study. It is also noted that when comparing the fitted model (5) with the fitted polynomial models, as used in Canova et al. (2009) and Canova et al. (2015), specifically quadratic polynomial models, we see how much more flexible the model (5) is in relation to the polynomial, since the usual models usually consider the mean of the observations while model (5) the individual effects of each individual, that is, the model brings potentialities in terms of fit in data of this nature.

Regarding specimens of the type of dry mortar, in a good part of the study, they admit a lower evolution in terms of the water absorption speed by capillarity, being observed that in the concentrations of higher contents, specifically in the content of 8%, on average, the specimens reached the end of the study. Therefore, these findings corroborate with the analyzes presented in Canova et al. (2009) and Canova et al. (2015), in which these are associated with the fact that specimens with matured mortar in general admit higher weights compared to specimens with dry mortar.

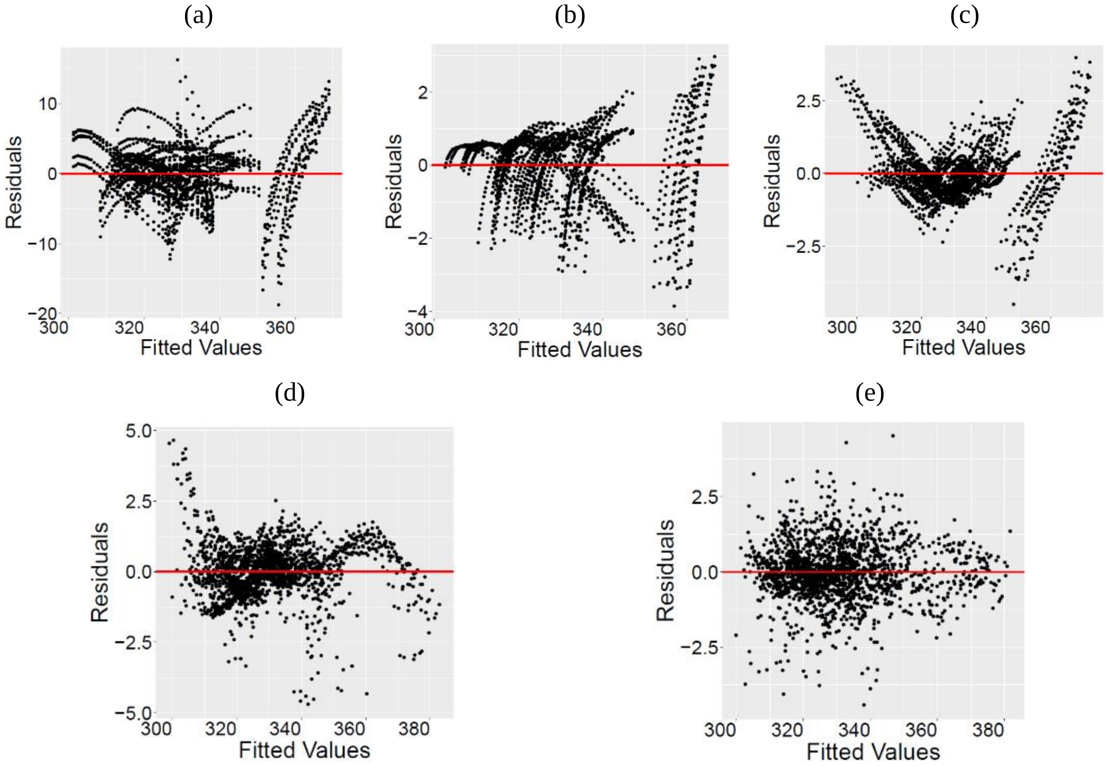
**Figure 4:** Fitted curves of individual subjects obtained from the model (5) and the fitted quadratic polynomial model by treatment.



Source: Authors.

Thus, to choose among the models considered, the Figure 5 the fitted values versus the Pearson residuals and in Figure 6 contains the weight of the specimens versus fitted values of Quadratic polynomial, (2), (3), (4) and (5) models, verifying in the Figures 5(e) and 6(e) that the model defined in (5) presents itself satisfactorily in relation to others.

**Figure 5:** (a), (b), (c), (d) and (e): Pearson residuals of the Quadratic polynomial, (2), (3), (4) and (5) models, respectively.



Source: Authors.

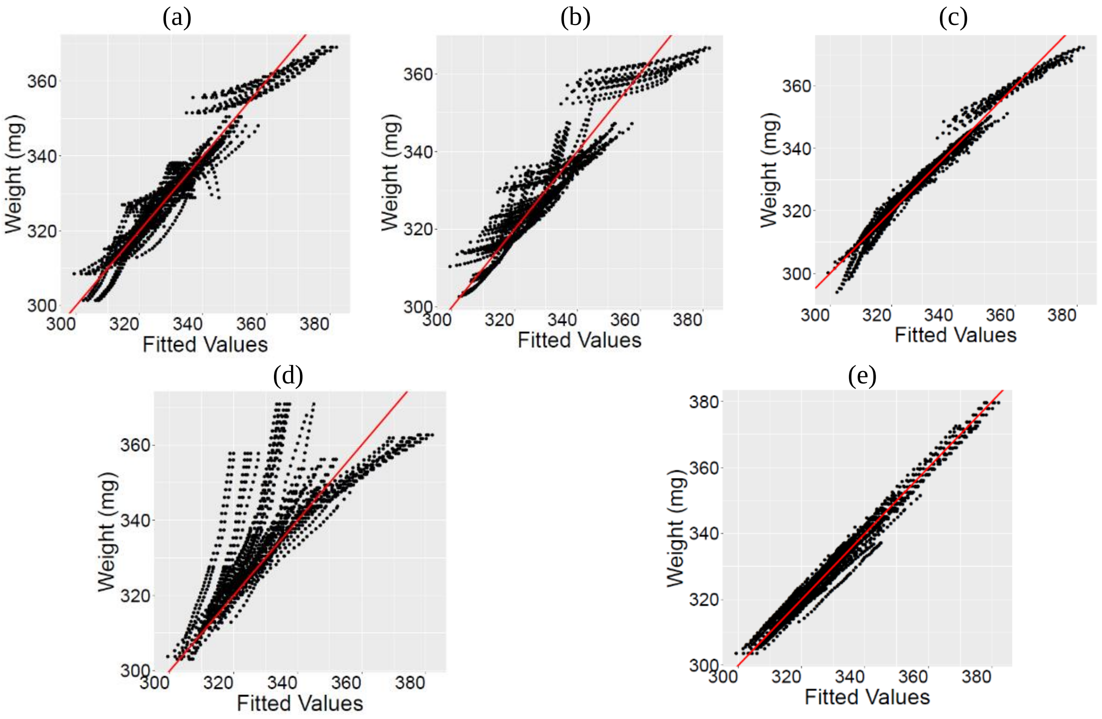
The Table 1 shows the results of the likelihood ratio test for the comparison between the fitted models, verifying that the model (5) presents itself satisfactorily in relation to the others, where this result is an approximation, as already described in Durbán et al. (2005) that suggest the use of simulation to determine the null distribution of the likelihood ratio test statistic.

**Table 1:** AIC, BIC and the approximate likelihood radio test for the fitted models.

| **Models** | **AIC** | **BIC** | **Test** | **L.Ratio** | **p-value** |
|---|---|---|---|---|---|
| Quadratic polynomial | 10334.27 | 10405.89 | - | - | - |
| Model 2 | 11329.58 | 11401.12 | - | - | - |
| Model 3 | 8903.22 | 8980.34 | - | - | - |
| Model 4 | 4746.77 | 4834.91 | Model 3 vs Model 4 | 4160.45 | <.0001 |
| Model 5 | 1618.90 | 1806.20 | Model 4 vs Model 5 | 3163.86 | <.0001 |

Source: Authors.

**Figure 6:** (a), (b), (c), (d) and (e): Relationship between mortar weight data and fitted values of Quadratic polynomial, (2), (3), (4) and (5) models, respectively.



Source: Authors.

## 4. Conclusion

The most common mixed models for longitudinal data represent each individual as the sum of the population average (which varies over time). However, these models may not be appropriate, since they assume that individual trajectories over time have a linear pattern. Regarding the applicability of the models P-splines rewritten as mixed models in the analysis of longitudinal data, with regard to medical and biological areas, these have been shown to be quite flexible, highlighting some potentialities, such as making the linearity hypothesis flexible in many models and the possibility of including complex structures in the usual smooth models. Regarding correlated data, if a curve is adjusted independently of the correlation in the data, it is seen that the methods of selecting smooth parameters will determine a lower value and, therefore, the curve will not be smooth. As for the results of the statistical analysis, it was observed through smooth that there were significant differences in the trajectory of the weight evolution in both types of mortar, concluding that the matured type mortar has a water absorption speed by capillarity higher than the mortar of the kiln dried type, we still have that the necessary assumptions for the validation of the model have been satisfied. For that reason, this methodology proved to be satisfactory for modeling the data of this experiment, whose nature is longitudinal and oscillatory.

## References

Achmad R. F. A., Janssen, P., Sa'adah, U., Solimun, E. A., & Nurjannah A. L. (2018). Comparison of spline estimator at various levels of autocorrelation in smoothing spline non parametric regression for longitudinal data. *Communications in Statistics-Theory and Methods, Taylor & Francis* 47:5265-5285.

Andrinopoulou, E. R., Eilers, P. H., Takkenberg, J. J., & Rizopoulos, D. (2018). Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines. *Biometrics, Wiley Online Library* 74:685-693.

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. Mixed-effects regression models in linguistics, *Springer, Cham* pp. 49-69.

Benedetti, A., Abrahamowicz, M., & Goldberg, M. S. (2009). Accounting for Data-Dependent Degrees of Freedom Selection When Testing the Effect of a Continuous Covariate in Generalized Additive Models. *Communications in Statistics - Simulation and Computation* 38(5):1115-1135.

Canova, J.A., de Angelis N. G., & Bergamasco, R. (2009). Mortar with unserviceable tire residues. *Journal of Urban and Environmental Engineering, JSTOR* 3:63-72.

Canova, J.A., de Angelis N. G., & Bergamasco, R. (2015). Pó de borracha de pneus inservíveis em argamassa de revestimento. *REEC-Revista Eletrônica de Engenharia Civil* 10:41-53.

Corbeil, R. R., & Searle, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics, Taylor & Francis* 18:31-38.

Currie, I. D., & Durbán, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling, Sage Publications Sage CA: Thousand Oaks*, CA 2:333-349.

Craven, P., & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische mathematik, Springer* 31:377-403.

da Silva, B. G., Guedes, T. A., Janeiro, V., Ferreira, E. C., de Araújo, S. M., & Ciupa, L. (2020). Analyzing weight evolution in mice infected by Trypanosoma cruzi. *Acta Scientiarum. Health Sciences* 42:e51437- e51437.

de Jong, R., van Buuren, S., & Spiess, M. (2015). Multiple Imputation of Predictor Variables Using Generalized Additive Models. *Communications in Statistics - Simulation and Computation* 45(3):968-985.

Durbán, M., Harezlak, J., Wand, M.P., & Carroll, R.J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in medicine, Wiley Online Library* 24:1153-1167.

Durbán, M. (2007). *Splines con penalizaciones: Teoría y aplicaciones*. Universidad Pública de Navarra, 60p.

Eilers, P.H.C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science, JSTOR* 89-102.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Vol.1. No 10. New York: Springer series in statistics.

Garcia-Hernandez, A., & Rizopoulos, D. (2018). %JM: A SAS Macro to Fit Jointly Generalized Mixed Models for Longitudinal Data and Time-to-Event Responses. *Journal of Statistical Software* 84:1-29.

Gressani, O., & Lambert, P. (2021). Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines. *Computational Statistics & Data Analysis,* 154, 107088.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Vol. 43. CRC press.

Hernando V. L., & Paula, G. A. (2016). An extension of log-symmetric regression models: R codes and applications. *Journal of Statistical Computation and Simulation* 86:1709-1735.

Islamiyati, A., Fatmawati., & Chamidah, N. (2019). Ability of Covariance Matrix in Bi-Response Multi- Prredictor Penalized Spline Model Through Longitudinal Data Simulation. *International Journal of Academic and Applied Research* 3:8-11.

Keele, L., & Keele, L. (2008). *Semiparametric regression statistic for the science*. Chichester: Wiley. pp 231.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society: Series b* (statistical methodology), Wiley Online Library 61:381-400.

McKeown, G. J., & Sneddon, I. (2014). Modeling continuous self-report measures of perceived emotion using generalized additive mixed models. Psychological methods, *American Psychological Association* 19:155-174.

Momen, M., Campbell, M. T., Walia, H., & Morota, G. (2019). Predicting longitudinal traits derived from highthroughput phenomics in contrasting environments using genomic Legendre polynomials and B-splines. *G3: Genes, Genomes, Genetics* 9:3369-3380.

Nores, M., & Díaz, M. (2016). Bootstrap hypothesis testing in generalized additive models for comparing curves of treatments in longitudinal studies. *Journal of Applied Statistics, Taylor & Francis.* 43:810-826.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, Oxford University Press* 58:545-554.

Pinheiro, J. C., & Bates, D. M. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Polansky, L., & Robbins, M. M. (2013). Generalized additive mixed models for disentangling long-term trends, local anomalies, and seasonality in fruit tree phenology. *Ecology and Evolution, Wiley Online Library* 3:3141-3151.

Prawanti, D. D., Budiantara, I. N., & Purnomo, J. D. (2019). Parameter Interval Estimation of Semiparametric Spline Truncated Regression Model for Longitudinal Data. *IOP Conference Series: Materials Science and Engineering, IOP Publishing* 546:1-10.

R Core Team. (2022). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rilem, T.C. (1994). *RILEM recommendations for the testing and use of constructions materials*. RC 6:218- 220.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. No. 12. Cambridge university press, 386p.

Schaeffer, L.R. (2004). Application of random regression models in animal breeding. *Livestock Production Science, Elsevier* 86:35-45.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of school psychology, Elsevier* 52:149-178.

Sudo, M., Sato, Y., & Yorozuya, H. (2021). Time-course in attractiveness of pheromone lure on the smaller tea tortrix moth: a generalized additive mixed model approach. *Ecological Research*, 36(4), 603-616.

Toshniwal, D., Speleers, H., Hiemstra, R. R., & Hughes, T. J. (2017). Multi-degree smooth polar splines: A framework for geometric modeling and isogeometric analysis. *Computer Methods in Applied Mechanics and Engineering, Elsevier* 316:1005-1061.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC press, 384p.

Zhang, X., Zhong, Q., & Wang, J. L. (2020). A new approach to varying-coefficient additive models with longitudinal covariates. *Computational Statistics & Data Analysis, Elsevier* 145:1-15.