Received: 15 September 2021 | Revised: 11 June 2022 | Accepted: 5 October 2022

DOI: 10.1002/widm.1479

OVERVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY

WILEY

Distributional regression modeling via generalized additive models for location, scale, and shape: An overview through a data set from learning analytics

Fernando Marmolejo-Ramos[1] | Mauricio Tejo[2] | Marek Brabec[3] | Jakub Kuzilek[4,5] | Srecko Joksimovic[1] | Vitomir Kovanovic[1] | Jorge González[6] | Thomas Kneib[7] | Peter Bühlmann[8] | Lucas Kook[9,10] | Guillermo Briseño-Sánchez[11] | Raydonal Ospina[12]

Correspondence
Fernando Marmolejo-Ramos, Centre for Change and Complexity in Learning, University of South Australia, Adelaide, Australia.
Email: fernando.marmolejo-ramos@unisa.edu.au

Funding information
Akademie Věd České Republiky, Grant/Award Number: RVO 67985807; Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 305305/2019-0; Deutsche Forschungsgemeinschaft, Grant/Award Number: KN 922/9-1; Fondo Nacional de Desarrollo Científico y Tecnológico, Grant/Award Number: 1201129; Grantová Agentura České Republiky, Grant/Award Number: GJ18-04150Y; H2020 European Research Council, Grant/Award Number: CausalStats—ERC-2017-ADG; Universidad de Valparaíso, Grant/Award Numbers: UVA 20993, 2021

Edited by: Ryszard Kowalczyk, Associate Editor and Witold Pedrycz, Editor-in-Chief

Abstract

The advent of technological developments is allowing to gather large amounts of data in several research fields. Learning analytics (LA)/educational data mining has access to big observational unstructured data captured from educational settings and relies mostly on unsupervised machine learning (ML) algorithms to make sense of such type of data. Generalized additive models for location, scale, and shape (GAMLSS) are a supervised statistical learning framework that allows modeling all the parameters of the distribution of the response variable with respect to the explanatory variables. This article overviews the power and flexibility of GAMLSS in relation to some ML techniques. Also, GAMLSS' capability to be tailored toward causality via causal regularization is briefly commented. This overview is illustrated via a data set from the field of LA.

This article is categorized under:

Application Areas > Education and Learning
Algorithmic Development > Statistics
Technologies > Machine Learning

KEYWORDS

causal regularization; causality; educational data mining; generalized additive models for location, scale, and shape; learning analytics; machine learning; statistical learning; statistical modeling; supervised learning

For affiliation refer to page 16

1 | INTRODUCTION

Most of the data in the field of learning analytics (LA) and educational data mining (EDM) are characterized by being big, second-hand, observational, and unstructured (Motz et al., 2018).[1] The data are big because they come from physical and virtual educational environments with many instructors and thousands of students and for whom several metrics exist (e.g., number of clicks, time stamps, course grades, etc.). The data are second-hand, observational, and unstructured because they are not obtained directly and the type and number of variables are not controlled by the researcher. Although such type of data are amenable to post hoc analyses only and do not allow confident causal inference (i.e. only experimentation enables so by securing first-hand and structured data; see Imai et al., 2008, 2013), they are nonetheless rich and should be statistically treated to extract valuable practical information.

After gathering educational data, the LA/EDM analytical pipeline begins with preprocessing the data so it is amenable to subsequent statistical treatment. Data preprocessing consumes more than 50% of the pipeline and, among other things, it implies selecting and transforming variables of interest (Romero & Ventura, 2020). Given a large chunk of the analytical pipeline is spend on data preprocessing, it is no surprise that unsupervised learning algorithms are heavily relied on in order to make sense of the data (Joksimovic et al., 2018; see also chapter 5 in Brooks & Thompson, 2017). Those algorithms are also used simply because there is no clear dependent variable. Advances in machine learning (ML), however, enable to submit data with no clear dependent variable to a set of unsupervised algorithms (e.g., clustering algorithms). Although unsupervised learning algorithms can meet their intended goal, they somewhat minimize human decision-making in the process of statistical model building. Supervised modeling and ML, on the other hand, are clearly targeted for a response variable of interest (e.g., as for targeted learning [a.k.a, superlearner]; see Van der Laan, 2017).

Once an explanatory model has been identified in the analytical pipeline, it is then tested for its predictive power (e.g., via cross-validation; see Yu & Kumbier, 2020 for a proposal of the place of cross-validation in the analytic pipeline). More importantly, practitioners are mostly in need of interpretable models that can even license causal interpretations. This article has the goal of providing an overview of generalized additive models for location, scale, and shape (GAMLSS); a supervised distributional regression framework that promotes statistical modeling of entire conditional distributions rather than conditional means. It is also argued that such a framework indeed allows for more causal-oriented interpretation and better external validity. The outline of this article is as follows: first, a brief technical review of GAMLSS are provided; second, an LA data set is described; third, the LA data are modeled via GAMLSS; and fourth, a link between GAMLSS and causal regularization is proposed. The discussion section considers distributional regression modeling in the larger context of statistical learning. Related GAMLSS-based analyses such as gradient-based boosting GAMLSS and penalized GAM are provided as Supporting Information along with their R codes at https://cutt.ly/2WuyxXz.

2 | GAMLSS AS A DISTRIBUTIONAL REGRESSION FRAMEWORK FOR STATISTICAL LEARNING

One of the traditional preprocessing practices in LA/EDM research consists of discretising continuous variables in order to enhance their interpretability (see section 3.3 in Romero & Ventura, 2020). Slicing a uniform or normally distributed continuous variable in three quantile-based bins (i.e., high, medium, and low) has been shown to approximate quite well a linear regression (Gelman & Park, 2008). However, in practice, and particularly in the social sciences and education fields, continuous variables tend to follow non-normal shapes (Bono et al., 2017). This fact then suggests that traditional regression models are not optimal and slicing numeric variables will give biased results (see Bennette & Vickers, 2012 for an example of how categorization of continuous data in epidemiology leads to biased estimation). Hence, flexible and interpretable regression techniques are needed to model such type of data. GAMLSS are a regression framework that enables performing comprehensive statistical learning on the distribution of the response variable with respect to the covariates.

GAMLSS are a class of supervised learning tools for semi-parametric regression problems that have led to a growing sophistication in the ML field. From a strict statistical modeling viewpoint (McCullagh, 2002), GAMLSS are used to analyze nonlinear relationships between the distributions of outcomes and covariates (features, in ML parlance) and where the covariates' effects are additively weighted.[2] These models were proposed by Rigby and Stasinopoulos (2001), Akantziliotou et al. (2002), and Rigby and Stasinopoulos (2005) as an improvement and extension to the generalized linear models (GLM) (McCulloch, 2000; Nelder & Wedderburn, 1972) and the GAM (Hastie & Tibshirani, 1990). Key to

GAMLSS are that they enable data analyses that exhibit parsimony, generality, consilience, and predictive capacity (Friedman & Silverman, 1989; Picard & Cook, 1984).

GAMLSS have been used in several fields including high-dimensional regression (De Bastiani et al., 2018; Groll et al., 2019; Hofner et al., 2016; Mayr et al., 2012), psychometrics (Timmerman et al., 2021), neuroimaging (Bethlehem et al., 2022), vision research (Truckenbrod et al., 2020), ecology (Smith et al., 2019), economics (Hohberg et al., 2020), linguistics (Coupé, 2018), hydrology (Dabele et al., 2017), survival analysis (De Castro et al., 2010), clinical management of hearing loss (Hu et al., 2015), insurance (Gilchrist et al., 2009), real-state appraisal of land lots (Florencio et al., 2012), film box-office revenues (Voudouris et al., 2012), among others, and just recently GAMLSS have been proposed as new statistical tool for psychological research (Campitelli et al., 2017). Software-wise, GAMLSS are implemented in R through the gamlss package (Rigby et al., 2020; Stasinopoulos & Rigby, 2007; Stasinopoulos et al., 2017). There are other GAMLSS R packages for extra additive terms (gamlss.add), fitting censored (interval) responses (gamlss.cens), fitting finite mixture distributions (gamlss.mx), fitting nonlinear models (gamlss.nl), fitting truncated distributions (gamlss.tr), among others. Other R packages related to GAMLSS are gamboostLSS (Mayr et al., 2012; Mayr & Hofner, 2018) and BAMLSS (Umlauf et al., 2018), and these allow performing boosting methods for GAMLSS models (suitable for high-dimensional data; see Thomas et al., 2018).

Another appealing feature of GAMLSS are its flexibility for data modeling through estimation algorithms (Cole & Green, 1992; Rigby & Stasinopoulos, 1996) that allow combining ML with statistical modeling (Breiman, 2001b; Stasinopoulos et al., 2018). For example, such algorithms enable fitting the conditional parametric distribution of the response variable with several continuous, discrete, and mixed distributions with different degrees of asymmetry and kurtosis. Therefore, not only the mean, but all of the parameters (i.e. location, scale and shape) can be modeled as parametric and/or additive nonparametric functions of covariates. This feature is quite instrumental to modeling response variables that do not follow an exponential family distribution (some exponential distributions are the Normal, Poisson, Gamma, Beta, Weibull [for fixed shape parameter] and Multinomial distributions) (Barndorff-Nielsen, 1980; Casella & Berger, 2002; McCulloch, 2000). A study by Voudouris et al. (2012) exemplifies the benefits of finding the right distribution for a data set. Film box-office revenue data exhibit a positive skew with a heavy tail and it was traditionally modeled via the Pareto–Levy–Mandelbrot (PLM) distribution (a distribution with infinite variance). However, the PLM distribution could not account for the dispersion, skewness, and kurtosis found in the film revenues data and this was impeding making any stable predictions. Voudouris et al. (2012) demonstrated that the four-parameter Box–Cox power exponential distribution could better fit the data and it allowed correctly predicting, among other things, the price of future contracts indexed by the film's performance. This study thus demonstrates that using a distribution that fits the data well, enables making reliable probabilistic statements. The mechanics of GAMLSS are described next.

Consider a data set $(\mathbf{X}_k, \mathbf{Z}_k, \mathbf{y})_{k \leq p}$ of sample size n, where $\mathbf{y} = (y_1, y_2, ..., y_n)^\top$ is a vector of independent observations on the response variable and $\mathbf{X}_k$, $\mathbf{Z}_k$ are input covariates design matrices for fixed and random effects (a.k.a. features in ML jargon) of size $n \times J'_k$ and $n \times q_{jk}$, respectively. By assuming that the variable of interest follows the probability density function (PDF) $f\left(y_i \,|\, \boldsymbol{\theta}^i \in \mathcal{D}\right)$, a parametric family of distributions (see table 1 in Rigby & Stasinopoulos, 2005) with $\boldsymbol{\theta}^i = (\theta_{i1}, \theta_{i2}, ..., \theta_{ip})^\top$ being a vector of p parameters associated to the explanatory variables and to random effects,[3] each distribution parameter of the GAMLSS model can be written as a function of regressors.

In GAMLSS statistical models, the kth parameter $\boldsymbol{\theta}_k$ is related to an additive predictor η_k through input features and random effects via

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

where $g_k(\cdot)$ is a strictly monotonic link function, $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, ..., \theta_{nk})^\top$ and $\boldsymbol{\eta}_k = (\eta_{1k}, \eta_{2k}, ..., \eta_{nk})^\top$ are $n \times 1$ vectors, $\boldsymbol{\beta}_k = \left(\beta_{1k}, \beta_{2k}, ..., \beta_{J'_k k}\right)^\top$ has dimension $J'_k \times 1$. The random effects parameter vector $\boldsymbol{\gamma}_{jk}$ with length J'_k follows the multivariate Gaussian distribution $N_{q_{jk}}\left(0, G_{jk}^{-1}\right)$, where G_{jk}^{-1} is the inverse of a symmetrical matrix $G_{jk} = G_{jk}(\lambda_{jk})$ of size $q_{jk} \times q_{jk}$ which depends on a λ_{jk} hyperparameter vector. If G_{jk} is singular, then γ_{jk} has a density function proportional to $\exp\left(-1/2 \gamma_{jk}^\top G_{jk} \gamma_{jk}\right)$. A GAMLSS model can be expressed differently by including ML procedures in order to boost its predictive power. For example, if $Z_{jk} = I_n$, where I_n is the identity matrix of type $n \times n$, and $\gamma_{jk} = h_{jk} = h_{jk}(x_{jk})$ for all combinations of j and k expressed in Equation (1), then the GAMLSS model adopts a semi-parametric additive term:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} h_{jk}\left(x_{jk}\right), \tag{2}$$

where function h_{jk} is an unknown function of the independent variable x_{jk} and $h_{jk}(x_{jk})$ is a vector that evaluates function h_{jk} in x_{jk}. Furthermore, smoothers such as cubic splines, penalized splines, fractional polynomials, LOESS curves, terms of variable coefficients, neural networks, kernels, and so on, can be included to deal with nonlinearity, volatility structural changes and other particularities in the data (Wood, 2017; Wood et al., 2016).

The vector of fixed and/or random-effect parameters are estimated within the GAMLSS framework by maximizing the penalized log-likelihood and this can be accomplished by using fast backfitting algorithms and resampling procedures (Groll et al., 2019; Mayr et al., 2012; Rigby & Stasinopoulos, 2005). Model selection is performed by finding the lowest generalized Akaike information criterion [GAIC(k)] for some selected value of k in the same context of AIC (Akaike, 1974) along with cross-validation (Geisser, 1975; Voncken et al., 2019) in order to prevent over-fitting of the data. The GAIC is defined by Voudouris et al. (2012) as GAIC(k) = GD + ($k \times g_l$), where $\mathrm{GD} = -2\ell\left(\widehat{\boldsymbol{\theta}}\right)$ is the global deviance being $\ell\left(\widehat{\boldsymbol{\theta}}\right)$ the maximized log-likelihood function, g_l denotes the total effective degrees of freedom of the adjusted model and k is a constant penalty for each degree of freedom used. If $k = 2$, the GAIC equates to the AIC, and if $k = \ln n$ it equates to the Bayesian information criterion (BIC). For the analysis of residuals, the normalized (randomized) quantile residuals plot can be used (Dunn & Smyth, 1996). In addition, GAMLSS allow examining residuals via probability plots such as the worm plot (van Buuren & Fredriks, 2001) which is an instrumental graphical technique for assessing the overall adequacy of the fitted model (see Fasiolo et al., 2020; Stasinopoulos et al., 2017).

By using family of sets notation, a GAMLSS model can be represented for ML implementation as

$$\mathscr{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}, \tag{3}$$

where $\mathcal{D}$ represents a family of distributions, $\mathcal{G}$ specifies the set of link functions $\left(g_1, ..., g_p\right)$ for parameters $(\theta_1, ..., \theta_p)$, $\mathcal{T}$ specifies the set of predictor terms $\left(\eta_1, ..., \eta_p\right)$, and λ specifies the set of hyperparameters. Thus, the linear regression model, for example, can be written as $y_i \sim \mathcal{D}\left(g_1(\mu(x)), g_2\left(\sigma(x)^2\right)\right)$, where $\mathcal{D}$ is the normal distribution, $\mathcal{G} = \{g_1, g_2\} = \{\mathrm{id}(x), \mathrm{id}(x)\}$, and $\lambda = (\beta, \sigma^2)$, where id($x$) is the identity function. Note that in this case, $\mu(x) = x^\top\beta$ and $\sigma(x)^2 = \sigma^2$. Another important example is logistic regression or softmax regression in the context of neural networks. In such case, a GAMLSS model can be expressed as $y_i \sim \mathcal{D}(g_1(p(x)))$, where $\mathcal{D}$ is the Bernoulli distribution, $\mathcal{G} = \{g_1\} = \{\mathrm{logit}(x) = \log(x/(1-x))\}$ (logistic or softmax function), and $p(x) = P(y_i = 1|x_i) = \exp\left(x_i^\top\beta\right)/\left(1 + \exp\left(x_i^\top\beta\right)\right)$ with $x^\top\beta$ being a linear predictor.

In order to compare two nested competing GAMLSS models $\mathscr{M}_0$ and $\mathscr{M}_1$ based on Equation (3), that is, when one model can be obtained from the others by imposing parametric restrictions, the global deviance or a LASSO approach (Groll et al., 2019) can be used to penalize overfittings and select the best model. When comparing two non-nested GAMLSS models (including models with smoothing terms; Hastie & Tibshirani, 1990), the GAIC (Akaike, 1974) and the J and MJ tests can be used (Cribari-Neto & Lucena, 2017; Davidson & MacKinnon, 1981; Godfrey, 2011; McAleer, 1995).

In a nutshell, GAMLSS are a framework that uses state-of-the-art algorithms for the modeling of continuous responses. As shown above, distributional regression analyses within a GAMLSS framework permit smooth alignment with well-known methods in ML. Although there are methods to compare data's means and standard deviations (Frank & Klar, 2016) and data's kurtosis and skewness (Cain et al., 2017), GAMLSS are a unified framework that promotes going beyond traditional mean regression (Kneib, 2013; Kneib et al., 2021) by considering other moments of the dependent variable's distribution. The GAMLSS framework is thus in line with recent proposals of moving beyond means and standard deviations to, at minimum, data's location and scale (Trafimow et al., 2018).[4] A final aspect to reiterate is that GAMLSS are designed to be a flexible and interpretable regression-based method for statistical learning. This is a beneficial feature to counter "black-box" modeling and instead facilitate models' explainability and applicability (see Yu & Kumbier, 2020). For more details on GAMLSS, see Stasinopoulos et al. (2017) and Rigby et al. (2020).

3 | THE OPEN UNIVERSITY LEARNING DATA SET

The goal of this article is to overview some of the statistical modeling capabilities of GAMLSS through a data set from the field of LA/EDM. The Open University Learning Data Set (OULAD; Kuzilek et al., 2017) is an open-access data set of about 32,593 students in a distance learning setting that relies on a virtual learning environment (VLE; for other data sets in the LA/EDM field see Mihaescu & Popescu, 2021). This data set contains a diverse set of students' attributes obtained from a large sample of university students. The OULAD is thus a suitable data set for testing new approaches to the predictive modeling of students' outcomes, their behaviors in VLEs, and evaluation of new approaches to LA. For example, Alshabandar et al. (2018) used Gaussian mixture models for the prediction of passing the next assessment based on clickstream data. Other models such as k-nearest neighbors, Naive Bayes, Decision Trees, Random Forests (RF), or support vector machines (SVM) have also been used for predicting the results of studied courses (Azizah et al., 2018; Ho & Jin Shim, 2018; Rizvi et al., 2019; Silveira et al., 2019). Finally, the OULAD has also been used for the evaluation of deep learning approaches for estimating students' withdrawal (Hassan et al., 2019), for distinguishing groups of students based on their activities in VLEs via unsupervised learning methods (Heuer & Breiter, 2018; Peach et al., 2019), and for the evaluation of methods of course recommenders based on the detection of learning styles (Li et al., 2019).

The OULAD has been released by the Open University; the largest distance learning institution in the United Kingdom with more than 165,000 students and hundreds of courses. Regular courses take approximately 9 months to study and consist of multiple assignments and a final exam. The assignments can be divided into various categories being the Tutor Marked Assignments (TMAs) the most important as it represents key milestones in a study's schedule. The university employs a Moodle-like online system to deliver the course content to the students. This allows capturing valuable information such as students' demographics, study results, and their behavior within the VLE represented by the summaries of click-stream data.

One particular STEM (science, technology, engineering, and mathematics) course has been selected for the present analysis; FFF and its presentation (semester) *2013J* studied by 2283 students. The course schedule is represented in Figure 1. The course contains five TMAs that represent the milestones for the topic learning periods. TMAs occur in weeks 2, 6, 13, 18, and 24 and at the end of the course an exam is taken. The exam takes place around four or more weeks have passed (in the current data set such information is missing). The present GAMLSS analysis focuses on this last TMA (TMA 5) (in the data set it is labeled `assessment_score`). TMA 5 is thus the dependent variable and it ranges between 0 and 100, such that values over 40 are considered as pass.

The following groups of students were excluded from the data set: actively withdrawn students ($n = 675$) and students who did not submit all TMAs ($n = 500$). Actively withdrawn students were unregistered from the course before its end, and their information regarding VLE activities and assessments is incomplete. The second group did not submit all the assignments in time as required by the course. The resulting data set thus contains data of 1108 students. Table 1 lists and describes the independent variables in the data set. The first column contains the name of the variables and the second column shows a brief description of each variable (more details as to the source data set can be found in Kuzilek et al., 2017). The click-stream information (i.e., "clicks_xy" variables) has been computed for the top five most common activity types in the VLE, and they represent 95% of all student click-stream data.

4 | STATISTICAL LEARNING OF THE OULAD DATA SET VIA GAMLSS

The goal of the following modeling is to illustrate how GAMLSS can be used in practice and has no attempt at making theoretical LA/EDM-related claims based on the OULAD data set. The first step in GAMLSS modeling is to find a set of suitable marginal distributions (i.e., when the dependent variable is not conditioned on any covariates) that approximate well the observed values. The dependent variable was linearly transformed so that its values resided in the [0, 1]

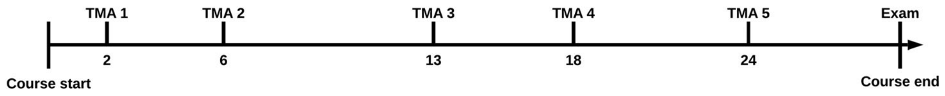

FIGURE 1 Course schedule (timeline occurs in weeks).

TABLE 1 List of independent variables

Attribute	Description
Gender	Student gender
Region	UK region, in which student lives[a]
Highest_education	The highest achieved education of the student
Imd_band	Percentile of the Index of Multiple Deprivation; see Noble et al. (2019) for details
Age_band	Student age band
Num_of_prev_attempts	Indicator whether the student attempted the course in previous years
Studied_credits	Credits studied in parallel by student, serves as the estimation of student workload
Disability	Indicator if student have disability
Cumulative_assessment_results	Weighted sum of all previous Tutor Marked Assignments (TMAs): $a_{\mathrm{sum}} = \sum_{n=1}^{4} w_n a_n$, where $\vec{w}^T = (0.125, 0.125, 0.25, 0.25)$ is vector of corresponding weights
Clicks_forumng	Sum of all clicks/actions student did in the discussion forum
Clicks_homepage	Sum of all clicks on course homepage
Clicks_oucontent	Sum of all views/clicks on TMAs assignments
Clicks_quiz	Sum of all clicks/attempts on nongraded quizzes
Clicks_subpage	Sum of all clicks when browsing the course web-page

[a] The complete list of regions can be found at https://bit.ly/3kKF1zs.

interval; that is, FAS = assessment/score/100; where 100 is the maximum assessment score (GAMLSS modeling of these types of data can be found in Ospina & Ferrari, 2012a, 2012b).

GAMLSS enable to fit several distributions to the target variable via the `histDist()` and `fitDist()` functions and in this study only the latter was used. Given that the marginal distribution is left-skewed and bounded in the [0, 1] interval, the extra arguments `type = "real0to1"` or `type = "realline"` in `fitDist()` can be used to exhaustively search for suitable distributions. The output of the search returns global deviance, AIC, and BIC values that assist in spotting candidate distributions. To avoid numerical problems, zeros and ones were converted to 0.5/100 and to 99.5/100, respectively (see Douma & Weedon, 2019). Note that choosing the distribution, or a set of candidate distributions, is not only a matter of statistical fitness but also of practical interpretability. Distributions with three or more parameters will tend to fit the skewness and kurtosis of the distribution better than distributions with one (e.g., Exponential) or two (e.g., Gamma) parameters. However, the applied researcher should prioritize distributions that parsimoniously explain changes in the values of the dependent variable in relation to the covariates in the context of the topic of the research. Table 2 shows the AIC measures of several distributions fitted to the marginal FAS distribution (GAMLSS have over 100 distributions available in the `gamlss.dist` package, loaded by default with the `gamlss` package).

The PDF and empirical cumulative distribution function (ECDF) plots indicate a negative skewness in FAS (see Figure 2a). It is evident from the CDF plot that the Normal (`NO`) and Reverse Gumbel (`RG`) distributions provide poor marginal fits even though these distributions are encountered in practical work (Rigby et al., 2020). The Beta class distributions (`BE, BEINF`, etc.) are natural candidates (Ospina & Ferrari, 2010) and exhibit reasonable behavior. On the other hand, the generalized beta type 1 (`GB1`) and Skew *t*-type 2 distributions (`ST2`) (Azzalini & Capitanio, 2003; Rigby et al., 2020) gave the best fits.

Figure 3 shows FAS' kernel density estimates conditioned on the covariates gender, disability, and highest education. These PDF plots are exploratory data analysis (EDA) tools (Tukey, 1977) that allow noticing differences in the location, dispersion and shape of the conditional distribution of FAS (i.e., for each combination of covariates). For example, it is evident that the higher the educational qualification, the higher the FAS and that students with disability tend to have lower FASs than nondisability students. Therefore, a useful approach to analyze the relation between FAS and its covariates is via GAMLSS as it allows to learn changes in the location, and other parameters, of the distribution of FAS as influenced by its covariates.

Using Wilkinson and Rogers' notation (Wilkinson & Rogers, 1973), the model to consider is FAS $\sim C + N$, where FAS is the dependent variable (fractional assessment_score), C represents a matrix with categorical covariates (gender, highest_education, age_band, disability, and type_of_click), and N is a numeric covariate (number of clicks). For illustration purposes, parametric fixed-effects-only regression models are specified, the set of distributions {`NO, BE, ST2,`

TABLE 2 Akaike information criterion (AIC) and global deviance (GD) for the fitted models with different probability distributions. The lower the AIC value, the better the goodness-of-fit. For example, generalized Beta type 1 (`GB1`) to Skew *t*-distribution (`SST`) are four-parameter distributions, while Logistic (`LO`) to Reverse Gumbel (`RG`) are two-parameter distributions (except the `exGAUS`, Ex-Gaussian distribution, which is a three-parameter distribution).

Distribution	AIC	GD
GB1	−5593	−5601
ST2	−5556	−5564
ST3	−5556	−5564
ST1	−5556	−5564
SST	−5544	−5564
SN2	−5438	−5564
EGB2	−5395	−5564
SHASH	−5378	−5564
SEP3	−5375	−5564
JSUo	−5362	−5564
JSU	−5362	−5564
SHASHo2	−5361	−5564
SHASHo	−5361	−5564
ST5	−5234	−5564
LOGITNO	−5224	−5601
BE	−5080	−5601
BEo	−5080	−5601
BEOI	−5078	−5601
BEZI	−5078	−5601
BEINF1	−5078	−5601
BEINF0	−5078	−5601
BEINF	−5076	−5601
GU	−4733	−5564
ST4	−4319	−5564
GT	−3474	−5564
NET	−3367	−5564
TF2	−3360	−5564
TF	−3360	−5564
PE	−3218	−5564
PE2	−3218	−5564
LO	−3178	−5564
NO	−2732	−5564
exGAUS	−2729	−5564
SIMPLEX	−2392	−5601
RG	168	−5564

`GB1`} is the response distribution space for the search and the scale and shape parameters are assumed to be constant for all observations; that is, the focus is on the μ location parameter only.[5]

More specifically, the conditional models considered here have the same linear predictor:

$$\eta = \text{gender} + \text{age\_band} + \text{disability} + \text{highest\_education} + \text{type\_of\_click} + \text{number\_of\_clicks}. \tag{4}$$

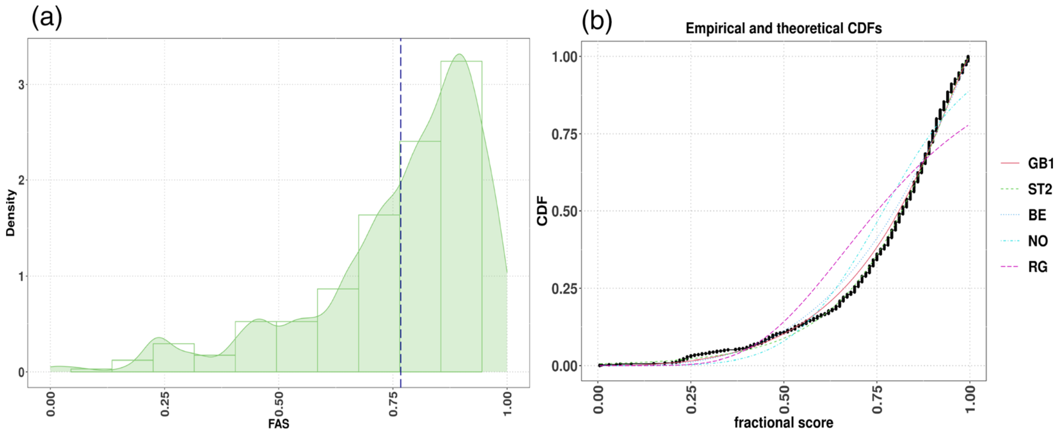

FIGURE 2 FAS′ kernel density estimates superimposed on histogram (a) and FAS′ empirical and theoretical CDFs (b). The vertical dotted line in the left plot indicates the variable's mean. The black line in the right plot shows the FAS′ ECDF and the colored lines represent five theoretical CDFs (ranked from best GB1 to worst fit RG). CDF, cumulative distribution functions; ECDF, empirical CDF; GB1, generalized beta type 1; RG, reverse Gumbel.

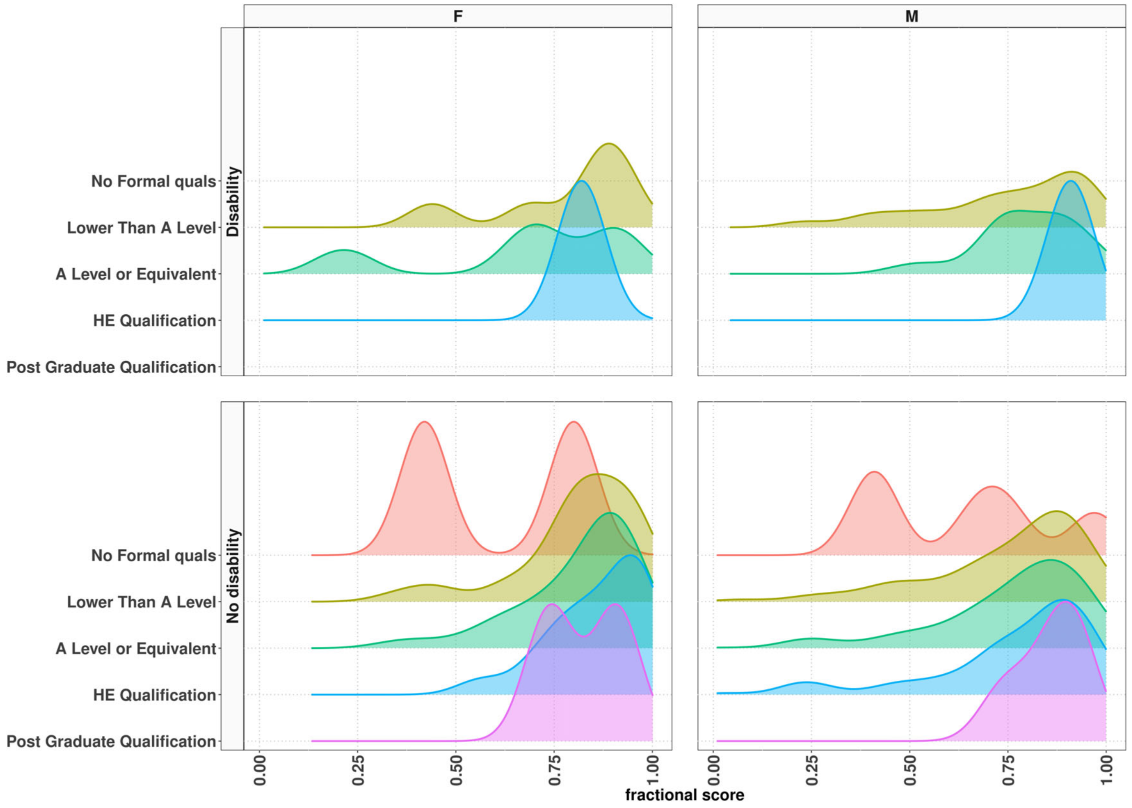

FIGURE 3 FAS′ kernel density estimates conditioned on the covariates gender (with two levels; F = females and M = males), disability (with two levels; first row = disability, second row = no disability) and highest education (with five levels). The graph also indicates the data are imbalanced in that not all combinations of levels of the covariates have values. That is, while there are FAS values for people with nondisability at all education levels, there are FAS values for people with disabilities at three education levels only.

The μ submodel has the following link functions: $g_{\mathrm{NO}}(\mu) = \mathrm{identiy}(\mu) \sim \eta$, $g_{\mathrm{BE}}(\mu) = \mathrm{logit}(\mu) \sim \eta$, $g_{\mathrm{ST2}}(\mu) = \mathrm{identiy}(\mu) \sim \eta$, and $g_{\mathrm{GB1}}(\mu) = \mathrm{logit}(\mu) \sim \eta$, respectively. Note that the Normal distribution (i.e., NO) was included for illustration purposes. Variable selection for the location submodel was made via the stepGAIC method in GAMLSS (Stasinopoulos et al., 2018) (note the drop1() function is also useful for this goal). The selection of the distribution was based on the examination of AIC and quantile residuals via worm plots.

Table 3 presents the comparison of the fitted models with different distributions. As expected, the Normal distribution gave a poor fit (see also Figure 2). On the other hand, and as shown by the ECDF plots (see Figure 2b), the GB1 and ST2 distributions showed the best performance as indexed by the Likelihood, AIC and degrees of freedom estimates.[6] The predictive power of the selected model was assessed via the general pseudo-$R^2_{\text{C\&S}}$ (Cox & Snell, 1968; Nagelkerke, 1991) (implemented in the GAMLSS R function Rsq()) and the pseudo-R^2_{S} (which is given by the square root of Spearman's sample correlation coefficient between the response and the fitted values; this approach is valid only for location submodels). The pseudo-R^2_{S} measures suggest that a model using the GB1 distribution gives the highest predictive power.

Figure 4 shows the gamlss fit output of the GB1 μ model[7] (this output was obtained via the generic function summary()). After a submodel selection step, the resulting μ submodel was:

$$\text{logit}(\mu) \sim \text{gender} + \text{age\_band} + \text{highest\_education} + \text{type\_of\_click} + \text{number\_of\_clicks}. \tag{5}$$

The results indicate that all the covariates, have effects on the μ parameter of the response variable FAS.[8] There was no evidence for an effect of "disability" possibly due to this variable presenting unbalanced information as shown Table 4, such situation could be addressed by obtaining more data or using bootstrap to obtain more confidence in the generality of the model (Branco et al., 2018).

On the other hand, provided all other variables are held constant, an increase in age after 55 is related to a decrease in FAS' location. Likewise, all other variables fixed, an increase in the number of clicks increases the location of the FAS distribution.

The adequacy of the fitted distributions is represented in Figure 4 via worm plots. A lack of fit is displayed by the residuals lying well above and below the value deviation 0.0. Also, the less closer the points of the plot are to the horizontal line at 0.0, the more distant the distribution of the residuals is to a standard normal distribution. In addition, a lack of fit is suggested when more than 5% of the points of the plot lie outside the two elliptic lines (those elliptic lines are point-wise ≈ 95% confidence intervals [CIs]). The results of the Beta, but particularly the Normal, distribution showed lack of fit and their inverted U-shape also signaled negative skewness in the residuals' distribution. This inverted U-shape also indicated that those distributions failed to correctly fit the high left-skewness of the data. Although the worm plot shapes of the GB1 and ST2 distributions suggested good fit, it was not perfect. Both struggled to fit the kurtosis of the marginal distribution and this is evidenced in the distribution of the residuals being leptokurtic in the case of the GB1 distribution (S-shape with left bent down) and platykurtic in the case of the ST2 distribution (S-shape with left bent up). Also, that some of the points in the plots representing the GB1 and ST2 distributions laid outside the ≈ 95% CIs indexes some degree of overdispersion in the data (see chapter 12 in Stasinopoulos et al., 2017 for details as to the interpretation of the worm plot).

The modeling performed here was fully parametric; so smoothers are to be used if a semi-parametric modeling is sought. In this sense, GAMLSS allow adding nonparametric smoothing functions for numeric covariates in order to augment the prediction power. Some of the functions available are: cubic splines, decision trees, locally weighted regression, penalized splines, and neural networks (Rügamer et al., 2020). Generally, it is recommended to use P(enalized)-splines (Eilers & Marx, 1996) in order to include potentially local and nonlinear effects of continuous variables (Wood, 2017; Wood et al., 2016). As illustration, the location parameter μ of the GB1 distribution was modeled using a penalized P-spline function as a way of understanding how the dependent variable FAS is affected

TABLE 3 Goodness-of-fit measures of selected μ submodels

μ submodel	Distribution	Likelihood	AIC	Degrees of freedom	Pseudo-$R^2_{\text{C\&S}}$	Pseudo-R^2_{S}
$g_{\text{GB1}}(\mu)$	GB1	−5629	−5597	5194	0.072	0.126
$g_{\text{ST2}}(\mu)$	ST2	−5282	−5250	5194	0.021	0.122
$g_{\text{BE}}(\mu)$	BE	−5239	−5211	5196	0.089	0.125
$g_{\text{NO}}(\mu)$	NO	−3087	−3087	5196	0.083	0.122

Abbreviations: AIC, Akaike information criterion; BE, Beta class distribution; GB1, generalized Beta type 1; NO, Normal distribution; ST2, Skew t-type 2 distribution.

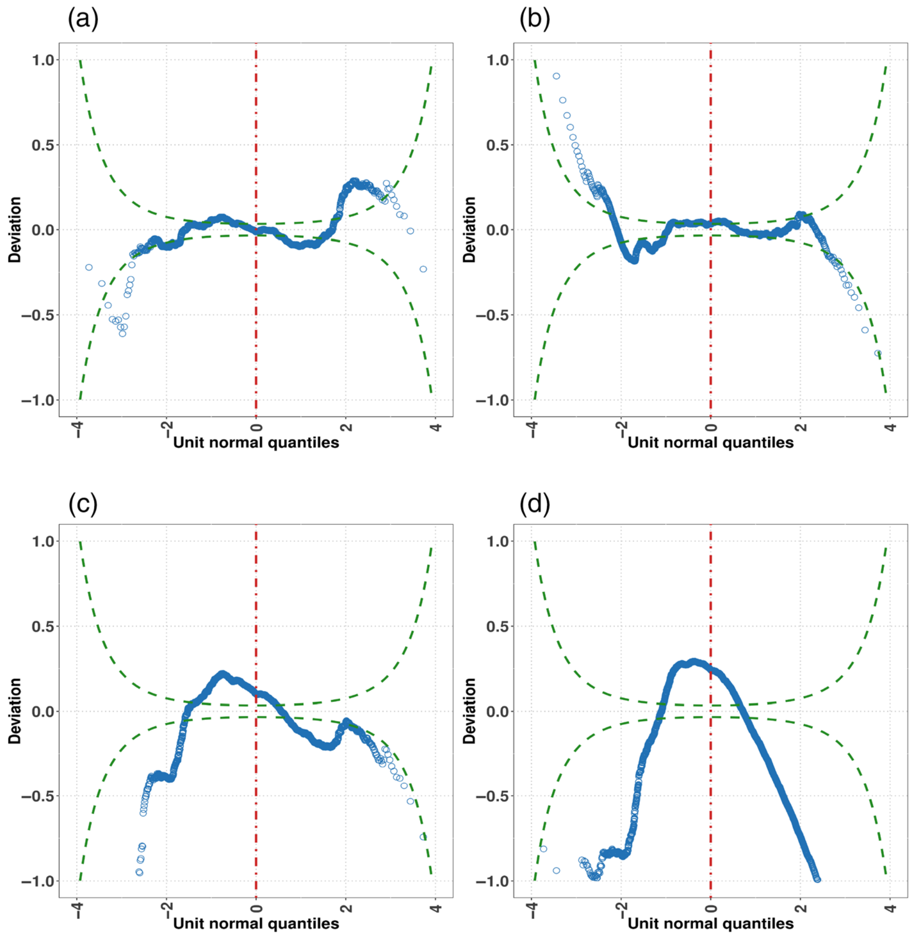

FIGURE 4 Diagnostic worm plots for assessing the fitness of models using the generalized beta type 1 (GB1) distribution (a), Skew *t*-distribution type 2 (ST2) (b), Beta (BE) distribution (c), and Normal (NO) distribution (d) to the FAS variable. A good fit is represented by ≈ 95% of values lying between the two green dotted elliptic lines and close to the deviation value of 0.0. In this example, the GB1 and ST2 distributions fit well most of the data but they struggle to fit the values in the tails of the FAS variable (although the ST2 distribution models better the right tail of the data than the GB1 distribution). However, compared to the GB1 and ST2 models, BE and NO exhibit a poor fit overall.

by the covariates. Specifically, the nonparametric smoother `pb()` in `gamlss` was applied to the covariate `number_of_clicks` to capture local variations in the context of the following model:

$$\eta = \text{gender} + \text{age\_band} + \text{disability} + \text{highest\_education} + \text{type\_of\_click} + \text{pb}(\text{number\_of\_clicks}). \tag{6}$$

To facilitate the interpretation of these predictors, the function `term.plot()` in the `gamlss` package was used. This function produces plots of parameter estimates (in the link function scale) for each covariate in the predictor of each parameter of the population distribution. Point estimates are represented by the trend lines (linear or smooth predictor) in Figure 5a and the shaded areas correspond to the estimates' standard errors. The plot suggests an increasing nonlinear relationship between FAS and the number of clicks. However, the relationship is not monotonic and the change in standard error suggests heteroskedasticity likely due to the presence of groups or clusters (indeed, data sparsity at the upper end of the covariate spectrum could also have played part in this effect). As it was the case of the fully parametric GB1 model, the μ term was not affected by "disability" after the submodel selection procedure. The inclusion of the nonparametric term increased the performance of the AIC (−5917) and the predictive power; pseudo-$R^2_{\text{C\&S}}$ (0.131) and pseudo-R^2_{S} (0.14). The worm plot in Figure 5b suggests that the inclusion of the nonparametric term helped to control the GB1's right tail (compare Figure 4a vs Figure 5b) (see Supporting Information for a deeper discussion of the estimated effect of covariates on the conditional response distribution).

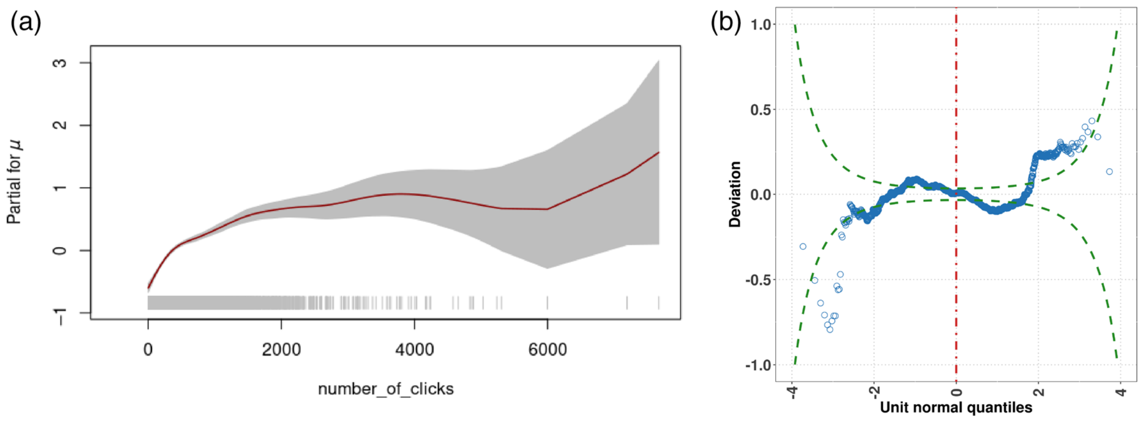

FIGURE 5 Termplot for the μ submodel when it includes a smooth term (P-splines) on the covariate "number of clicks" (a). Plot (b) shows the diagnostic worm plot for assessing the fitness of the GB1 model.

So far, all the GAMLSS modeling has been done only on the location parameter (μ) of the dependent distribution. A way to boost predictive power is by also modeling the other parameters of the dependent variable. A likelihood ratio aimed at determining whether the GAMLSS scale and shape parameters were constant for all observations suggested these parameters were not constant. Thus, the linear predictor given in Equation (4) was applied to the GB1 distribution's parameters through their σ, ν, and τ link functions; that is, $\log(\sigma) \sim \eta$, $\log(\nu) \sim \eta$, and $\log(\tau) \sim \eta$, respectively. As done above for the FAS' location parameter μ, recursive covariate selection based on AIC was performed for the scale and shape parameters. The results of this selection procedure showed that all submodels were distinctively affected by the covariates such that,

$$\begin{aligned}
&\text{logit}(\mu) \sim \text{gender} + \text{age\_band} + \text{type\_of\_click} + \text{number\_of\_clicks} \\
&\log(\sigma) \sim \text{age\_band} + \text{highest\_education} \\
&\text{logit}(\nu) \sim \text{age\_band} + \text{number\_of\_clicks} \\
&\log(\nu) \sim \text{highest\_education}.
\end{aligned}$$

This new model showed a pseudo-$R^2_{\text{C\&S}}$ of 0.12. That is, there was a 15% increase in performance improvement when compared with the model obtained in Equation (5) (model without smoothers). Note also that after the submodel selection procedure, the 'disability' covariate was not part of the predictors once again. Figure 6 displays the residual worm plot for this comprehensive model. This new model indicated that fitting all the parameters of the FAS' distribution led to minimizing the leptokurtosis in the residuals; that is, the points in the left tail of the worm plot are now closer to the ≈95% boundaries (compare Figures 4a vs. 6).

GAMLSS allow using complementary techniques to improve the modeling of the data but it would be prohibitive to attempt to cover them all herein. Thus, some techniques are briefly commented on. Variable selection can be carried out via cross-validation or LASSO in order to control over-fitting by considering different link functions for the covariates (e.g., identity, inverse, reciprocal, etc.). An example of this practice can be found in Cribari-Neto and Lucena (2017). Also, the number of levels in categorical covariates can be reduced in order to improve the fitness of the model (see pcat() function in GAMLSS). GAMLSS also permit to robustify the model's fitness by countering the influence of outliers (via the function gamlssRobust).

Cross-validation is a ubiquitous step in ML. In GAMLSS, *k*-fold cross-validation is attained via the `gamlssCV()` function. If the goal is to fit a `gamlss` model to the training data set and estimate the validation global deviance for the validation data set, the `gamlssVGD()` function can be used. It is important to recall, though, that cross-validation requires relative stability of the structured data and complete observations for each level within each variable in order to obtain reliable estimates (Gronau & Wagenmakers, 2019; Keevers, 2019; Wang & Gelman, 2015). As shown in Figure 3, some levels within the education level variable had missing observations for males and females and with disability/nondisability; this situation thus led to estimation issues when cross-validation was attempted. Finally, missing values can be handled by creating predictive models that include imputation or LASSO-type regularization (Arrieta et al., 2020; Hamzah et al., 2020).

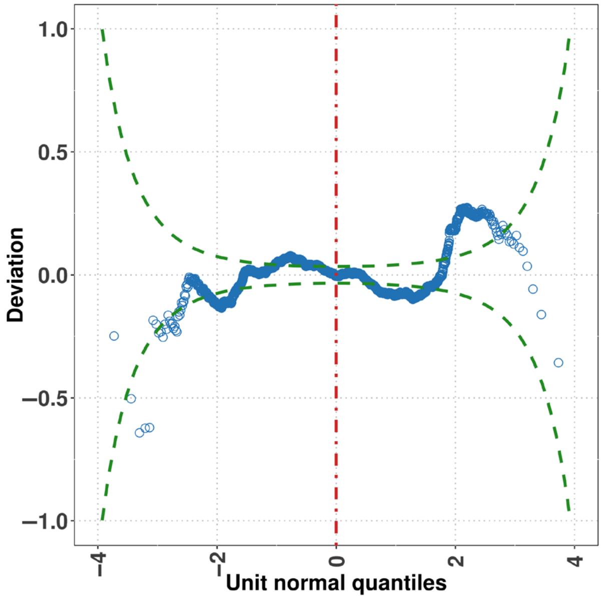

FIGURE 6 Worm plot for the GB1 model when the μ, σ, ν, and τ parameters were modeled.

TABLE 4 Summary results of the generalized Beta type 1 distribution (GB1) when modeling the μ (location) submodel

```
Family: c("GB1," "Generalized Beta type 1")
Call: gamlss(formula = response ~ gender + age_band + disability + highest_education +
type_of_click + number_of_clicks, family = GB1, data = na.omit(data))
Fitting method: RS()
----------------------------------
Mu link function: logit
Mu Coefficients:
```

Variable	Estimate	SE	*t*-value	Pr(>|*t*|)
Intercept	0.8790	0.1437	6.1153	0.0000***
genderm	−0.2211	0.0321	−6.8800	0.0000***
Age_band35-55	0.1283	0.0287	4.4637	0.0000***
Age_band55<=	−0.7917	0.1518	−5.2153	0.0000***
Disability_No.disability	0.0496	0.0460	1.0792	0.2805
Highest_educationHE.Qualification	0.0790	0.0352	2.2438	0.0249*
Highest_educationLower.Than.A.Level	−0.0744	0.0260	−2.8627	0.0042**
Highest_educationNo.Formal.quals	−0.9865	0.1537	−6.4180	0.0000***
Highest_educationPost.Graduate.Qualification	0.0444	0.1794	0.2476	0.8045
Type_of_clickclicks_homepage	0.0125	0.0331	0.3780	0.7054
Type_of_clickclicks_oucontent	−0.2863	0.0385	−7.4392	0.0000***
Type_of_clickclicks_quiz	−0.0051	0.0285	−0.1770	0.8595
Type_of_clickclicks_subpage	0.0886	0.0348	2.5464	0.0109*
Number_of_clicks	0.0005	0.0000	16.2857	0.0000***

4.1 | Comparison of GAMLSS to selected ML methods

A study investigating the performance of GAMLSS against one ML method showed that GAMLSS outperformed artificial neural networks in the modeling of war-fighting combat simulation data (Boutselis & Ringrose, 2013). This section aims at evaluating how GAMLSS perform in relation to other ML algorithms during the modeling of the OULAD data set. Four ML methods were considered:

- **Classification and regression tree (C&RT)** (Breiman et al., 1984): This classic method builds and prunes a decision tree using, for example, Gini's impurity measure. The decision tree itself provides an explanation of each decision and it is simple to understand. However, building such a decision tree is sensitive to the input data, and even small changes in the data can result in a large change in the final model.
- **Random Forest (RF)** (Breiman, 2001a): This is another ensemble learning approach, which uses the ensemble of decision trees for classification and regression.
- **Extreme gradient boosting (EGB)** (Chen & Guestrin, 2016): This is an efficient variant of the ensemble learning proposed by Chen and Guestrin in 2006.
- **Nonlinear support vector machines with radial basis function kernel (nlSVM+k)** (Murphy, 2012): This is an algorithm which constructs the decision boundary based on the structure of the input data. The kernel approach maps data into a higher dimension in order to reduce error caused by nonlinear relationships. This kernel method was used here.

The interest here is not to identify the best model for inference effects (covariate selection) for scientific insight and interpretation. All models with the features used in the linear predictor in Equation (4) were selected and trained in order to make a fair model comparison (Ding et al., 2018; Emmert-Streib & Dehmer, 2019). A 10-fold cross-validation approach was used to validate the models. First, the data were divided into 10 folds and in every step 1 fold was used for the model validation and the rest for the model training. The training set (9 folds) was again divided in 10 folds and cross-validation was used for tuning the models' parameters. Models were compared via the root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) metrics[9]:

$$\begin{aligned}
\mathrm{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\widehat{y}_i\right)^2},\\
\mathrm{MAE} &= \frac{1}{N}\sum_{i=1}^{N}\left|y_i-\widehat{y}_i\right|,\\
R^2 &= 1-\frac{\sum_{i=1}^{N}\left(y_i-\widehat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i-\overline{y}\right)^2},
\end{aligned} \tag{7}$$

where y_i represents the assessment score (FAS), $\widehat{y}_i$ is the predicted FAS, N is the data set's sample size, and $\overline{y}$ is the mean value of the FAS in the OULAD.

TABLE 5 Performance of four ML methods and GAMLSS when applied to the OULAD. The best metrics are shown in bold characters (i.e., the lowest RMSE and MAE and the highest R^2). Means (M) and standard deviations (SD) are estimated across 10-fold cross-validation.

	RMSE		R^2		MAE	
Method	***M***	**SD**	***M***	**SD**	***M***	**SD**
GAMLSS	0.1828	0.0061	0.0685	0.0291	0.1377	0.0038
RF	**0.1803**	0.0061	**0.1061**	0.0223	0.1364	0.0033
C&RT	0.1828	0.0067	0.0655	0.0168	0.1379	0.0043
nlSVM+k	0.1852	0.0070	0.0953	0.0168	**0.1300**	0.0038
EGB	0.1859	0.0075	0.0731	0.0225	0.1395	0.0051

Abbreviations: C&RT, classification and regression tree; EGB, extreme gradient boosting; GAMLSS, generalized additive models for location, scale, and shape; MAE, mean absolute error; ML, machine learning; nlSVM+k, nonlinear support vector machines with radial basis function kernel; OULAD, Open University Learning Dataset; RF, Random Forests; RMSE, root mean square error.

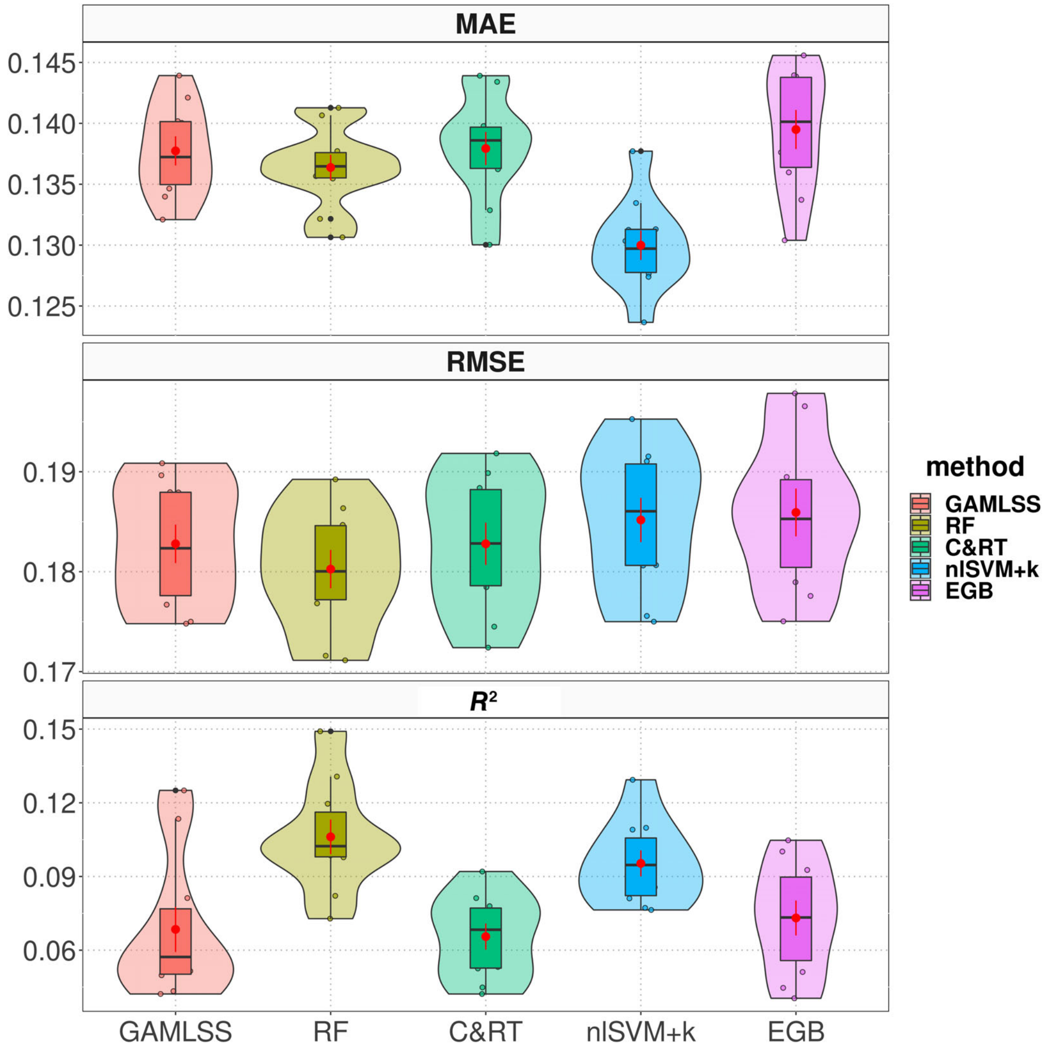

FIGURE 7 Violinplots of the cross-validation results. The mean and its 95% confidence interval (CI) are represented by the red dots and error bars. The overlaid dot plots, on each violin plot, represent the result of each of the 10-fold cross-validation.

The results of the models' performance are shown in Table 5 and in the Figure 7. Overall, there were minimal differences among the models; that is, all methods showed similar predictive power. Although the RF, C&RT, nlSVM+k, and EGB methods tend to be considered as having relatively high flexibility (see figure 2.7 in James et al., 2017) and accuracy (see figure 12 in Arrieta et al., 2020), they also have low interpretability (see those same figures). GAMLSS, however, given its superset relationship to GLM and GAM, can be regarded as also being flexible and accurate but allowing higher interpretability. In other words, while the RF, C&RT, nlSVM+k, and EGB methods could be regarded as "black-box" models, GAMLSS can be regarded as a "white-box" model. Indeed, even if GAMLSS had mid-range flexibility and accuracy, its higher level of interpretability is in line with what future techniques in ML (including artificial intelligence) are striving for (Angelov et al., 2021; Gunning et al., 2019). Thus, the semi-parametric GAMLSS model is an educated and interpretable choice to produce insights into the OULAD data set.

5 | DISTRIBUTIONAL REGRESSION AND CAUSAL REGULARIZATION

One of the ultimate aims of science is to establish causal relationships (Pearl, 2009). Inferring causality from observational or heterogeneous data with unspecific interventions is an overly ambitious task and necessarily requires strong untestable assumptions. Regression models, and distributional regression such as GAMLSS, provide a weaker association measure than a causal one between a response variable Y and some covariates X. However, (distributional) regression models can be regularized toward causality, without claiming to infer causal effects, but leading to a certain kind of invariance, stability, and robustness across experimental settings (Arjovsky et al., 2019; Bühlmann, 2020a, 2020b; Peters et al., 2016; Rothenhäusler et al., 2021). Such additional stability can be very useful for improving generalisability to other settings, and better external validity and statistical replicability of findings.

The idea of causal regularization for enhancing stability and better external validity has been extended to a certain class of distributional regression models (Kook et al., 2022). The file "causal-regularization-supplement" in the repository (see end of Section 6), features the application of causal-regularized distributional regression to the OULAD data set to demonstrate improved worst-case prediction and better external validity.

6 | DISCUSSION AND CONCLUSIONS

This article had the goal of overviewing some of the modeling properties allowed by the GAMLSS framework. In order to do so, an open access data set pertinent to LA/EDM was used. The analyses did not intend to make theoretical claims relating to the data set but simply illustrate how GAMLSS could be used for supervised statistical learning of data. It was then argued and showed that GAMLSS are a flexible and interpretable regression-oriented modeling approach that enables investigating the effect of the covariates on the dependent variable's location, scale, skewness, and kurtosis parameters (more details on GAMLSS in Stasinopoulos et al., 2017; Rigby et al., 2019). The analysis also illustrated that GAMLSS allows building both explanatory and predictive models and producing both types of models is a must in proper statistical learning (Shmueli, 2010; Yarkoni & Westfall, 2017). Likewise, it was shown how distributional regression methods, such as GAMLSS, can be tweaked via causal regularization for inferring causality and thus favoring statistical replicability. The following paragraphs revolve around methodological and statistical issues relating to GAMLSS type analyses and statistical learning in general.

The data were modeled with a `GB1` distribution. Although the traditional two-parameter Beta distribution (`BE`) did not provide a good fit, it does prevent this distribution to be considered for data modeling. It may well be the case that while the `GB1` may not be a good model in a similar data set, the `BE` could be. A short document found in the repository (see link in the last paragraph) provides mathematical arguments in favor of the Beta distribution. It is also worth mentioning that GAMLSS are not the only way to analyze continuous data. As shown in the Supporting Information, a GAM (with penalisation) analysis is also possible. GAM-type analysis is contained within the GAMLSS framework and it has been shown to be instrumental in modeling autocorrelations in experimental data (Baayen et al., 2017). Alternatively, robust regression (Ronchetti, 2021; Rousseeuw & Hubert, 2018) and distributional regression methods such as quantile regression (Waldman, 2018) could have been educated choices (see Kneib et al., 2021 for other distributional regression approaches). An interesting related analytical approach is that of scoring rules (Gneiting & Raftery, 2007). Scoring rules enable to evaluate the predictive ability of distributional regression models, but they require the explicit availability of a predictive distribution which is provided by GAMLSS but not by all predictive approaches. Within a multiverse analysis framework (Steegen et al., 2016), performing several valid statistical analyses is indeed encouraged as they enable to determine patterns in data.

Although it was shown herein that the GAMLSS framework can cater for a supervised statistical learning approach to data analysis, it does not prevent mixing GAMLSS with unsupervised learning techniques. For example, classification algorithms such as the "one rule" (a.k.a. 1R, see Holte, 1993, implemented in R via the `OneR` package) and "Boruta" (Kursa & Rudnicki, 2010, implemented in R via the `Boruta` package) can be used for variable selection and candidate GAMLSS regression models can be built by combining the best subset of variables. Finally, the resulting models' predictive power could be assessed via cross-validation (note though that, besides cross-validation, models should be externally validated). Indeed, there is a recent method called distributional regression forests that blend decision trees (a predictive model popular in ML) and GAMLSS regression (see Schlosser et al., 2019 and the `disttree` R package). These are approaches worth exploring in silico and through real data sets. Ultimately, the goal is to promote statistical learning and modeling and minimize reliance on hypothesis testing. GAMLSS, and the techniques mentioned above, allow precisely this.

Admittedly, the data set featured in the analyses is not high dimensional (i.e., $p < n$); however, GAMLSS can deal with high-dimensional data (i.e., $p > n$) where the estimation of the coefficients via maximum likelihood methods would be intractable. As mentioned above, distributional models can also be fitted using gradient-based boosting methods (Mayr et al., 2012; Mayr & Hofner, 2018). The boosting estimation approach consists of fitting simple submodels by means of gradient descent. In each iteration only the best fitting independent variable is added to the model.

Hence, if a regressor does not improve the model fit, the algorithm retains its partial effect at zero, thus excluding the variable from the model. Thus, the number of fitting iterations becomes the main tuning parameter and it is typically determined using cross-validation. In short, estimating GAMLSS via gradient-based boosting carries out data-driven variable selection, shrinkage of the estimated coefficients, and addresses ill-posed scenarios such as multicollinearity in the covariates and high dimensionality ($p > n$) while retaining interpretability of the estimated partial effects (Hofner et al., 2016). A short tutorial featuring gradient-based boosting modeling via GAMLSS are available in the Supporting Information.

It is also important to point out that recent developments in GAMLSS methodology allow for the inclusion of unstructured or nontabular data into the distributional model, resulting in "semi-structured deep distributional regression" (Rügamer et al., 2020). This recent extension of the distributional regression framework combines advancements in ML and statistics that allow statistical modeling of more complex data structures while retaining the interpretability of the fitted model.

Finally, there has been a growing interest in the topic of causality. It was suggested herein that GAMLSS can be modified to exhibit stability and invariance of regression fits; and these are sensible proxies of causality. That is, GAMLSS are a distributional regression framework "geared toward causality" (Bühlmann, 2020a) in that it can be used to examine stabilization of estimated fits across perturbations (in relation to cross-validation, it is important to note that causal models are not suitable for prediction if there is no distribution shift between training and validation data since including noncausal covariates improves prediction). A recent proposal has demonstrated that combining instrumental variable estimation with GAMLSS is also a fruitful step in this front (Briseño-Sánchez et al., 2020).

R codes, Supporting Information, and data sets used in the analyses can be found at https://cutt.ly/2WuyxXz.

AUTHOR CONTRIBUTIONS

Fernando Marmolejo-Ramos: Conceptualization (lead); investigation (lead); methodology (equal); project administration (lead); software (equal); supervision (lead); visualization (equal); writing—original draft (lead); writing—review and editing (lead). **Mauricio Tejo:** Conceptualization (equal); methodology (equal); software (equal). **Marek Brabec:** Conceptualization (equal); methodology (equal); software (equal). **Jakub Kuzilek:** Data curation (lead); formal analysis (equal). **Srecko Joksimovic:** Writing—review and editing (supporting). **Vitomir Kovanovic:** Writing—review and editing (supporting). **Jorge González:** Writing—review and editing (supporting). **Thomas Kneib:** Writing—review and editing (supporting). **Peter Bühlmann:** Methodology (equal); software (lead); writing—original draft (equal); writing—review and editing (equal). **Lucas Kook:** Methodology (equal); software (equal); writing—review and editing (equal). **Guillermo Briseño-Sánchez:** Methodology (equal); software (equal); writing—review and editing (supporting). **Raydonal Ospina:** Conceptualization (equal); data curation (equal); formal analysis (lead); methodology (lead); software (equal); writing—original draft (equal); writing—review and editing (equal).

AFFILIATIONS

[1]Centre for Change and Complexity in Learning, University of South Australia, Adelaide, Australia
[2]Instituto de Estadística, Universidad de Valparaíso, Valparaíso, Chile
[3]Department of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic
[4]Czech Institute of Informatics, Robotics and Cybernetics, CTU, Prague, Czech Republic
[5]Computer Science Education/Computer Science and Society Research Group, Humboldt University of Berlin, Berlin, Germany
[6]Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago de Chile, Chile
[7]Campus Institute Data Science (CIDAS) and Chair of Statistics, Georg-August-Universität Göttingen, Göttingen, Germany
[8]Seminar for Statistics, ETH Zürich, Zürich, Switzerland
[9]Epidemiology, Biostatistics, and Prevention Institute, University of Zurich, Zurich, Switzerland
[10]Institute of Data Analysis and Process Design, Zurich University of Applied Sciences, Winterthur, Switzerland
[11]Department of Statistics, TU Dortmund University, Dortmund, Germany
[12]Department of Statistics, CASTLab, Federal University of Pernambuco, Recife, Brazil

ACKNOWLEDGMENTS

A preprint version of this article can be found at https://osf.io/preprints/socarxiv/r8azq/. As the present version is an improved and updated version of that pre-print, this present version should be the primary reference. Jakub Kuzilek was supported by a junior grant research project by Czech Science Foundation GACR no. GJ18-04150Y. Marek Brabec was supported by the long-term strategic development financing of the Institute of Computer Science (Czech Republic RVO 67985807). Jorge González acknowledges partial support from Fondecyt Grant 1201129. Raydonal Ospina thanks partial support by the National Council for Scientific and Technological Development (CNPq) in Brazil, through the grant no. 305305/2019-0. Thomas Kneib was supported by the German Research Foundation, DFG, grant KN 922/9-1. Lucas Kook was supported by the Novartis Research Foundation (FreeNovation 2019). Peter Bühlmann was supported by the European Research Council under the Grant Agreement No 786461 (CausalStats—ERC-2017-ADG). Mauricio Tejo was supported by INICI-UV (UVA 20993, 2021) project. Open access publishing facilitated by University of South Australia, as part of the Wiley - University of South Australia agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

we provide a link to an OA repository where data and R codes are available

ORCID

Fernando Marmolejo-Ramos https://orcid.org/0000-0003-4680-1287
Vitomir Kovanovic https://orcid.org/0000-0001-9694-6033

RELATED WIREs ARTICLE

Educational data mining and learning analytics: An updated survey

ENDNOTES

[1] In this article, LA and EDM are referred to indistinctly.

[2] The term "effect" is commonplace in the regression literature (i.e., main effects, interaction effects, fixed effects, and random effects; see chapter 3 in James et al., 2017; Sheskin, 2011) and it stands for relationships between the independent variables and the dependent variable. However, the term "effect" can be interpreted as "cause" only under certain study designs (e.g., chapter 18 in Gelman et al., 2020) or when instrumental variable techniques are used (see section 2.3 in Bell et al., 2019).

[3] For many practical situations and computational implementations, four parameters are required to determine the distribution $f(y_i|(\theta_{i1}, \theta_{i2}, ..., \theta_{ip}))$. The R implementation denotes these parameters as $\mu_i = \theta_{i1}$, $\sigma_i = \theta_{i2}$, $\nu_i = \theta_{i3}$, $\tau_i = \theta_{i4}$, for $i = 1, ..., n$. The first parameter is location (usually the mean or median), the second is scale (usually the standard deviation or precision), and the others are shape parameters (e.g., skewness and kurtosis). For computational implementations of GAMLSS it is desirable that the probability density of y and its first derivatives with respect to each of the parameters must be computable. Also, when the covariates are stochastic (useful in ML for investigating causal relation between variables) the density $f(y_i|\boldsymbol{\theta}^i)$ is taken to be conditional on their values.

[4] A simple and informative metric that combines location and scale is the coefficient of variation (see Arachchige et al., 2022; Ospina & Marmolejo-Ramos, 2019). This measure of relative dispersion is given, in its classic form, by the ratio of the standard deviation to the mean (i.e., σ/μ). This metric is rather underused and just recently has been revived in the fields of data mining and machine learning (Bindu et al. (2020)).

[5] Although statistical analyses and hypotheses tests are traditionally performed on the data's location parameter (e.g., mean or median differences between treatment groups), it is less common to perform tests and analyses on scale parameters (e.g., differences between groups' variances) and even more uncommon on shape parameters (i.e., skewness and kurtosis. In some fields, though, these parameters are investigated; see Ben-David et al., 2015). One reason for this situation is that the interpretation of the shape parameters is challenging due to lack of agreement on what they represent (e.g., kurtosis, traditionally understood as data's "peakedness," has been defined as an index of data's propensity to outliers; see Westfall, 2014). Another reason is that for quality estimation of more delicate features (like higher order moments, various shape parameters) vast amounts of data are needed and that are traditionally not available. Nowadays, big data are increasingly common, allowing for analyses that go far beyond traditional viewpoints and the OULAD data set is just one example

[6] The degrees of freedom is a useful concept for describing model complexity and it is asymptotically equal to the trace of the usual "hat" matrix plus the number of parameters in the error covariance matrix of the model (Hastie & Tibshirani, 1990).

[7] The probability density function of the GB1 is $f(y|\mu, \sigma\nu, \tau) = \frac{\tau\nu^{\beta} y^{\tau\alpha-1}(1-y^{\tau})^{\beta-1}}{B(\alpha,\beta)[\nu+(1-\nu)y^{\tau}]^{\alpha+\beta}}$, where $0 < y < 1$, $\alpha = \mu(1-\sigma^2)/\sigma^2$ and $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$, and $\alpha > 0$, $\beta > 0$. The location is $\mu = \alpha/(\alpha+\beta)$ and the scale is $\sigma = (\alpha+\beta+1)^{-1/2}$.

[8] By default, `gamlss()` chooses the category that comes first alphabetically or numerically (alphanumerically) as the reference category. In this case, `Female` is the reference of the variable `gender` and `A Level or Equivalent` is the baseline of the variable `highest education`. By using the *R* function `relevel()` it is possible to change the reference category for each factor variable.

[9] Note that these are all mean-based (point-prediction performance) measures commonly used in ML work. As the predictive power of GAMLSS is better assessed via out-of-sample log-likelihood, it is thus not surprising that GAMLSS did not outperform the other methods under these three metrics.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Akantziliotou, K., Rigby, R., & Stasinopoulos, D. (2002). The R implementation of generalized additive models for location, scale and shape. In *Statistical modelling in society: Proceedings of the 17th International Workshop on statistical modelling, Statistical Modelling Society, Chania, Crete, July 8-12, 2002* (pp. 75–83).

Alshabandar, R., Hussain, A., Keight, R., Laws, A., & Baker, T. (2018). The application of Gaussian mixture models for the identification of at-risk learners in massive open online courses. In *IEEE congress on evolutionary computation*. IEEE Publishing. http://researchonline.ljmu.ac.uk/id/eprint/8486/

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5), e1424. https://doi.org/10.1002/widm.1424

Arachchige, C. N. P. G., Prendergast, L. A., & Staudte, R. G. (2022). Robust analogs to the coefficient of variation. *Journal of Applied Statistics*, *49*(2), 268–290. https://doi.org/10.1080/02664763.2020.1808599

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:190702893. http://arxiv.org/abs/1907.02893

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Azizah, E. N., Pujianto, U., Nugraha, E., & Darusalam (2018). Comparative performance between c4.5 and naive bayes classifiers in predicting student academic performance in a virtual learning environment. In *2018 4th International Conference on Education and Technology (ICET)* (pp. 18–22). https://doi.org/10.1109/ICEAT.2018.8693928

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 367–389.

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Barndorff-Nielsen, O. E. (1980). *Exponential families*. Wiley Online Library.

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: Making an informed choice. *Quality & Quantity*, *53*, 1051–1074. https://doi.org/10.1007/s11135-018-0802-x

Ben-David, A., von Hausegger, S., & Jackson, A. D. (2015). Skewness and kurtosis as indicators of non-Gaussianity in galactic foreground maps. *Journal of Cosmology and Astroparticle Physics*, *2015*. https://doi.org/10.1088/1475-7516/2015/11/019

Bennette, C., & Vickers, A. (2012). Against quantiles: Categorisation of continuous variables in epidemiological research, and its discontents. *BioMed Central Medical Research Methodology*, *12*(21). https://doi.org/10.1186/1471-2288-12-21

Bethlehem, R. A., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., Adler, S., Alexopoulos, G. S., Anagnostou, E., Areces-Gonzalez, A., Astle, D. E., Auyeung, B., Ayub, M., Bae, J., Ball, G., Baron-Cohen, S., Beare, R., Bedford, S. A., Benegal, V., … Alexander-Bloch, A. F. (2022). Brain charts for the human lifespan. *Nature*, *604*, 525–533. https://doi.org/10.1038/s41586-022-04554-y

Bindu, K., Morusupalli, R., Dey, N., & Rao, C. (2020). *Coefficient of variation and machine learning applications*. CRC Press.

Bono, R., Blanca, M., Arnau, J., & Gómez-Benito, J. (2017). Non-normal distribution commonly used in health, education, and social sciences. A systematic review. *Frontiers in Psychology*, *8*(1602). https://doi.org/10.3389/fpsyg.2017.01602

Boutselis, P., & Ringrose, T. J. (2013). GAMLSS and neural networks in combat simulation metamodelling: A case study. *Expert Systems with Applications*, *40*, 6087–6093.

Branco, P., Torgo, L., & Ribeiro, R. P. (2018). REBAGG: REsampled BAGGing for imbalanced regression. In *Second International Workshop on Learning with Imbalanced Domains: Theory and Applications* (pp. 67–81). PMLR.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modelling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and regression trees*. Taylor & Francis. https://books.google.de/books?id=JwQx-WOmSyQC

Briseño-Sánchez, G., Hohberg, M., Groll, A., & Kneib, T. (2020). Flexible instrumental variable distributional regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(Part 4), 1553–1574.

Brooks, C., & Thompson, C. (2017). *Handbook of learning analytics*. Society for Learning Analytics Research (SoLAR).

Bühlmann, P. (2020a). Invariance, causality and robustness (with discussion). *Statistical Science*, *35*, 404–426.

Bühlmann, P. (2020b). Toward causality and improving external validity. *Proceedings of the National Academy of Sciences United States of America*, *117*, 25963–25965.

Cain, M., Zhang, Z., & Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735.

Campitelli, G., Macbeth, G., Ospina, R., & Marmolejo-Ramos, F. (2017). Three strategies for the critical use of statistical methods in psychological research. *Educational and Psychological Measurement*, *77*(5), 881–895. https://doi.org/10.1177/0013164416668234

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury Pacific Grove.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery* (pp. 785–794). KDD '16. https://doi.org/10.1145/2939672.2939785

Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, *11*(10), 1305–1319.

Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in Psychology*, *9*(513). https://doi.org/10.3389/fpsyg.2018.00513

Cox, D. R., & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *30*(2), 248–265.

Cribari-Neto, F., & Lucena, S. E. (2017). Non-nested hypothesis testing inference for GAMLSS models. *Journal of Statistical Computation and Simulation*, *87*(6), 1189–1205.

Dabele, S. E., Bogdanowicz, E., & Strupczewski, W. (2017). Around and about an application of the GAMLSS package to non-stationary flood frequency analysis. *Acta Geophysica*, *65*, 885–892.

Davidson, R., & MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica: Journal of the Econometric Society*, *49*, 781–793.

De Bastiani, F., Rigby, R. A., Stasinopoulous, D. M., Cysneiros, A. H., & Uribe-Opazo, M. A. (2018). Gaussian markov random field spatial models in GAMLSS. *Journal of Applied Statistics*, *45*(1), 168–186.

De Castro, M., Cancho, V. G., & Rodrigues, J. (2010). A hands-on approach for fitting long-term survival models under the GAMLSS framework. *Computer Methods and Programs in Biomedicine*, *97*(2), 168–177.

Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, *35*(6), 16–34.

Douma, J. C., & Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. *Methods in Ecology and Evolution*, *10*(9), 1412–1430.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–102.

Emmert-Streib, F., & Dehmer, M. (2019). Evaluation of regression models: Model assessment, model selection and generalization error. *Machine Learning and Knowledge Extraction*, *1*(1), 521–551.

Fasiolo, M., Nedellec, R., Goude, Y., & Wood, S. N. (2020). Scalable visualization methods for modern generalized additive models. *Journal of Computational and Graphical Statistics*, *29*(1), 78–86.

Florencio, L., Cribari-Neto, F., & Ospina, R. (2012). Real estate appraisal of land lots using GAMLSS models. *Chilean Journal of Statistics*, *3*(1), 75–91.

Frank, J., & Klar, B. (2016). Methods to test for equality of two normal distributions. *Statistical Methods & Applications*, *25*(4), 581–599.

Friedman, J. H., & Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, *31*(1), 3–21.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*(350), 320–328.

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Gelman, A., & Park, D. (2008). Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician*, *62*(4), 1–8.

Gilchrist, R., Kamara, A., & Rudge, J. (2009). An insurance type model for the health cost of cold housing: An application of gamlss. *REVSTAT–Statistical Journal*, *7*(1), 55–66.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Godfrey, L. G. (2011). Robust non-nested testing for ordinary least squares regression when some of the regressors are lagged dependent variables. *Oxford Bulletin of Economics and Statistics*, *73*(5), 651–668.

Groll, A., Hambuckers, J., Kneib, T., & Umlauf, N. (2019). LASSO-type penalization in the framework of generalized additive models for location, scale and shape. *Computational Statistics & Data Analysis*, *140*, 59–73.

Gronau, Q. F., & Wagenmakers, E. J. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior*, *2*(1), 1–11.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI–Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Hamzah, F. B., MohdHamzah, F., Razali, S. F. M., Jaafar, O., & AbdulJamil, N. (2020). Imputation methods for recovering streamflow observation: A methodological review. *Cogent Environmental Science*, *6*(1), 1745133.

Hassan, S. U., Waheed, H., Aljohani, N. R., Ali, M., Ventura, S., & Herrera, F. (2019). Virtual learning environment to predict withdrawal by leveraging deep learning. *International Journal of Intelligent Systems*, *34*(8), 1935–1952. https://doi.org/10.1002/int.22129

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

Heuer, H., & Breiter, A. (2018). Student success prediction and the trade-off between big data and data minimization. In D. Krömker & U. Schroeder (Eds.), *DeLFI 2018—Die 16. E-Learning Fachtagung Informatik* (pp. 219–230). Gesellschaft für Informatik E.V..

Ho, L. C., & Jin Shim, K. (2018). Data mining approach to the identification of at-risk students. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 5333–5335). https://doi.org/10.1109/BigData.2018.8622495

Hofner, B., Mayr, A., & Schmid, M. (2016). gamboostlss: An R package for model building and variable selection in the gamlss framework. *Journal of Statistical Software, Articles*, *74*(1), 1–31. https://doi.org/10.18637/jss.v074.i01

Hohberg, M., Pütz, P., & Kneib, T. (2020). Treatment effects beyond the mean using distributional regression: Methods and guidance. *PLoS One*, *15*(2), e0226514. https://doi.org/10.1371/journal.pone.0226514

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63–91.

Hu, W., Swanson, B. A., & Heller, G. Z. (2015). A statistical method for the analysis of speech intelligibility tests. *PLoS One*, *10*(7), e0132409.

Imai, K., King, G., & Stuart, E. (2008). Misunderstanding between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, *171*(2), 481–502.

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society; Series A, Statistics in Society*, *176*(Part 1), 5–51.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R*. Springer.

Joksimovic, S., Poquet, O., Kovanovic, V., Dowell, N., Mills, C., Gasevic, D., Dawson, S., Graesser, A., & Brooks, C. (2018). How do we model learning at scale? A systematic review of research on MOOCs. *Review of Educational Research*, *88*(1), 43–86.

Keevers, T. L. (2019). *Cross-validation is insufficient for model validation*. Joint and Operations Analysis Division, Defence Science and Technology Group.

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling*, *13*(4), 275–303.

Kneib, T., Silbersdorff, A., & Säfken, B. (2021). Rage against the mean—A review of distributional regression approaches. *Econometrics and Statistics*. https://doi.org/10.1016/j.ecosta.2021.07.006

Kook, L., Sick, B., & Bühlmann, P. (2022). Distributional anchor regression. *Statistics and Computing*. *32*, 39. https://doi.org/10.1007/s11222-022-10097-z

Kursa, M., & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, *4*, 170171. https://doi.org/10.1038/sdata.2017.171http://10.0.4.14/sdata.2017.171

Li, R., Yin, C., Zhang, X., & David, B. (2019). Online learning style modeling for course recommendation. In S. Patnaik & V. Jain (Eds.), *Recent developments in intelligent computing, communication and devices* (pp. 1035–1042). Springer Singapore.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., & Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data—A flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*(3), 403–427.

Mayr, A., & Hofner, B. (2018). Boosting for statistical modelling: A non-technical introduction. *Statistical Modelling*, *18*(3–4), 365–384.

McAleer, M. (1995). The significance of testing empirical non-nested models. *Journal of Econometrics*, *67*(1), 149–171.

McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics*, *30*(5), 1225–1310.

McCulloch, C. E. (2000). Generalized linear models. *Journal of the American Statistical Association*, *95*(452), 1320–1324.

Mihaescu, M. C., & Popescu, P. S. (2021). Review on publicly available datasets for educational data mining. *WIREs Data Mining and Knowledge Discovery*, *11*(3), e1403. https://doi.org/10.1002/widm.1403

Motz, B., Carvalho, P., de Leeuw, J., & Goldstone, R. (2018). Embedding experiments: Staking causal inference in authentic educational contexts. *Journal of Learning Analytics*, *5*(2), 47–59.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.

Nagelkerke, N. J.(1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Noble, S., McLennan, D., Plunkett, E., Gutacker, N., Silk, M., & Wright, G. (2019). *The English Indices of Deprivation 2019 (research report)*. Ministry of housing, communities and local government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/833947/IoD2019_Research_Report.pdf

Ospina, R., & Ferrari, S. (2012a). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623. https://doi.org/10.1016/j.csda.2011.10.005

Ospina, R., & Ferrari, S. (2012b). On bias correction in a class of inflated beta regression models. *International Journal of Statistics and Probability*, *1*(2), 269–282. https://doi.org/10.5539/ijsp.v1n2p269

Ospina, R., & Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, *51*(1), 111–126.

Ospina, R., & Marmolejo-Ramos, F. (2019). Performance of some estimators of relative variability. *Frontiers in Applied Mathematics and Statistics*, *5*(43). https://doi.org/10.3389/fams.2019.00043

Peach, R. L., Yaliraki, S. N., Lefevre, D., & Barahona, M. (2019). Data-driven unsupervised clustering of online learner behaviour. *npj Science of Learning*, *4*(1), 14. https://doi.org/10.1038/s41539-019-0054-0

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference using invariant prediction: Identification and confidence interval (with discussion). *J Royal Statistical Society, Series B*, *78*, 947–1012.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*(387), 575–583.

Rigby, R., & Stasinopoulos, D. (1996). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, *6*(1), 57–65.

Rigby, R., & Stasinopoulos, D. (2001). The GAMLSS project: A flexible approach to statistical modelling. In *New trends in statistical modelling: Proceedings of the 16th International Workshop on Statistical Modelling* (Vol. 337, p. 345). University of Southern Denmark.

Rigby, R. A., & Stasinopoulos, M. D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., & Bastiani, F. D. (2019). *Distributions for modeling location, scale, and shape*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429298547

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., & de Bastiani, F. (2020). *Distributions for modeling location, scale, and shape using GAMLSS in R*. CRC Press.

Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; a decision tree based approach. *Computers & Education*, *137*, 32–47. https://doi.org/10.1016/j.compedu.2019.04.001

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIRES Data Mining and Knowledge Discovery*, *10*, e1355. https://doi.org/10.1002/widm.1355

Ronchetti, E. (2021). The main contributions of robust statistics to statistial science and a new challenge. *Metron*, *79*, 127–135.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., & Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*(2), 215–246. https://doi.org/10.1111/rssb.12398

Rousseeuw, P., & Hubert, M. (2018). Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery*, *8*, e1236. https://doi.org/10.1002/widm.1236

Rügamer, D., Kolb, C., & Klein, N. (2020). *A unified network architecture for semi-structured deep distributional regression*. arXiv:2002.05777.

Schlosser, L., Hothorn, T., Stauffer, R., & Zeileis, A. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Annals of Applied Statistics*, *13*(3), 1564–1589.

Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures*. CRC Press.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Silveira, P. D. N., Cury, D., Menezes, C., & dos Santos, O. L. (2019). Analysis of classifiers in a predictive model of academic success or failure for institutional and trace data. In *2019 IEEE Frontiers in Education Conference (FIE)* (pp. 1–8). https://doi.org/10.1109/FIE43999.2019.9028618

Smith, A., Hofner, B., Lamb, J. S., Osenkowski, J., Allison, T., Sadoti, G., McWilliams, S. R., & Paton, P. (2019). Modeling spatiotemporal abundance of mobile wildlife in highly variable environments using boosted GAMLSS hurdle models. *Ecology and Evolution*, *9*(5), 2346–2364.

Stasinopoulos, M. D., Rigby, R. A., & de Bastiani, F. (2018). GAMLSS: A distributional regression approach. *Statistical Modelling*, *18*(3–4), 248–273.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & de Bastiani, F. (2017). *Flexible regression and smoothing using GAMLSS in R*. CRC Press.

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

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives in Psychological Science*, *11*(5), 702–712.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., & Hofner, B. (2018). Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, *28*(3), 673–687.

Timmerman, M., Voncken, L., & Albers, C. (2021). A tutorial on regression-based norming of psychological tests with GAMLSS. *Psychological Methods*, *26*(3), 357–373. https://doi.org/10.1037/met0000348

Trafimow, D., Wang, T., & Wang, C. (2018). Means and standard deviations, or locations and scales? That is the question! *New Ideas in Psychology*, *50*, 34–37.

Truckenbrod, C., Meigen, C., Brandt, M., Vogel, M., Wahl, S., Jurkutat, A., & Kiess, W. (2020). Reference curves for refraction in a german cohort of healthy children and adolescents. *PLoS One*, *15*(3), e0230291.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Umlauf, N., Klein, N., & Zeileis, A. (2018). BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *Journal of Computational and Graphical Statistics*, *27*(3), 612–627.

van Buuren, S., & Fredriks, M. (2001). Worm plot: A simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, *20*(8), 1259–1277.

Van der Laan, M. (2017). Targeted learning: The link from statistics to data science. *STAtOR*, *18*(4), 12–16.

Voncken, L., Albers, C. J., & Timmerman, M. E. (2019). Model selection in continuous test norming with GAMLSS. *Assessment*, *26*(7), 1329–1346.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J., & Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics*, *39*(6), 1279–1293.

Waldman, E. (2018). Quantile regression: A short story on how and why. *Statistical Modelling*, *18*(3–4), 203–218.

Wang, W., & Gelman, A. (2015). Difficulty of selecting among multilevel models using predictive accuracy. *Statistics and Its Interface*, *8*(2), 153–160.

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. R.I.P. *The American Statistician*, *68*(3), 191–195.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *22*(3), 392–399.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman & Hall.

Wood, S. N., Pya, N., & Saefken, B. (2016). Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Associatio*, *111*(516), 1548–1575.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Yu, B., & Kumbier, K. (2020). Veridical data science. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(8), 3920–3929.

How to cite this article: Marmolejo-Ramos, F., Tejo, M., Brabec, M., Kuzilek, J., Joksimovic, S., Kovanovic, V., González, J., Kneib, T., Bühlmann, P., Kook, L., Briseño-Sánchez, G., & Ospina, R. (2023). Distributional regression modeling via generalized additive models for location, scale, and shape: An overview through a data set from learning analytics. *WIREs Data Mining and Knowledge Discovery*, *13*(1), e1479. https://doi.org/10.1002/widm.1479