# A new continuous distribution on the unit interval applied to modelling the points ratio of football teams

**Luiz R. Nakamura, Pedro H. R. Cerqueira, Thiago G. Ramires, Rodrigo R. Pescim, R. A. Rigby & Dimitrios M. Stasinopoulos**

# A new continuous distribution on the unit interval applied to modelling the points ratio of football teams

Luiz R. Nakamura [a], Pedro H. R. Cerqueira[b], Thiago G. Ramires [b], Rodrigo R. Pescim[c], R. A. Rigby[d] and Dimitrios M. Stasinopoulos[d]

[a]Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Florianópolis, Brazil; [b]Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, São Paulo, Brazil; [c]Departamento de Estatística, Universidade Estadual de Londrina, Londrina, Brazil; [d]Centre of Communications Technology and Mathematics, London Metropolitan University, London, UK

ABSTRACT

We introduce a new flexible distribution to deal with variables on the unit interval based on a transformation of the sinh–arcsinh distribution, which accommodates different degrees of skewness and kurtosis and becomes an interesting alternative to model this type of data. We also include this new distribution into the generalised additive models for location, scale and shape (GAMLSS) framework in order to develop and fit its regression model. For different parameter settings, some simulations are performed to investigate the behaviour of the estimators. The potentiality of the new regression model is illustrated by means of a real dataset related to the points rate of football teams at the end of a championship from the four most important leagues in the world: Barclays Premier League (England), Bundesliga (Germany), Serie A (Italy) and BBVA league (Spain) during three seasons (2011–2012, 2012–2013 and 2013–2014).

## 1. Introduction

During the past few years, statistics had been widely incorporated to sports in order to develop better strategies, increase any individual performance using specific training methods for each athlete or improve some fitness aspects. Even though there is a high use of statistical results in sports, football generally uses only basic information such as descriptive analyses. In the past decades, several works were developed in order to predict championship results or evaluate team quality such as [15] who used a Poisson regression model to explain the number of goals from football team using data from The English Premier League (1995–1996 Season). Brillinger [3] considered a trinomial model to estimate the win, draw and loss probabilities of any particular team of the 2006 Brazilian Championship. Karlis and Ntzoufras [14] proposed a Bayesian approach for predicting match outcomes for fitting the goal difference (away and home) of a football team for the English

Premier League (2006–2007 Season). More recently, Louzada et al. [16] introduced a Poisson distribution for the number of goals scored by a football team in a match using a linear regression model to predict the quality of a team using data from the English Premier League (2008–2009 Season) and the UEFA (Union of European Football Association) Champions League (2008–2009; 2009–2010), Cerqueira et al. [5] used causal models in order to evaluate team quality on European football teams and Schauberger et al. [24] proposed an extension of the Bradley–Terry model in order to identify the on-field variables that are connected to the sportive success or failure of single matches in the German Bundesliga. However, we can note that those works do not consider the points rate (the ratio of points scored to the maximum points possible in a championship) of a football team in league as response variable on the unit interval $(0, 1)$. Moreover, the points rate plays an important role in sports, mainly in football, because it summarises all information related to the possible outcomes (win, draw or loss) of a team in a football match. So, in this sense, we consider the points rate of the football teams at the end of a championship from the four most important leagues in the world: English Premier League (England), Bundesliga (Germany), Serie A (Italy) and BBVA league (Spain), during three seasons ( 2011–2014).

In several areas of research such as engineering, reliability, life testing experiments, finance, econometrics and also in sports, various types of data are modelled by finite range distributions. Since the interval used is the standard unit interval $(0, 1)$, the data can be interpreted as rates or proportions.

The beta distribution is one of the most important models to account for a response variable which produces results in the range $(0, 1)$. This distribution has been used extensively in theoretical and applied statistics for over a century and it can be fitted practically to any data representing a phenomenon in almost any field of application [17]. For more details about applications of the beta distribution, see [4,12].

Although the beta distribution is the main family of continuous distributions on unit interval bounded support, since it has only two parameters, this distribution and its regression model [10] may present a lack of flexibility when modelling both skewness and kurtosis. In this sense, we present a new very flexible model to deal with variables between zero and one based on a transformation of the sinh–arcsinh distribution (SHASHo) [13] which has four parameters and is able to model different degrees of skewness and kurtosis, thus it may be a really interesting alternative to model this type of data. In order to model the points rate of all football teams from the four above cited leagues as a function of other variables, we used the generalised additive models for location, scale and shape (GAMLSS) [22]. GAMLSS is a very general class of univariate regression models in which all parameters of a distribution (that does not necessarily belong to the exponential family) can be modelled as parametric and/or additive nonparametric smooth functions of explanatory variables. Using this approach, we can note which variables affect more the points rate obtained by a football team at the end of a league.

This paper has two main aims. First to introduce a new distribution called the logitSHASHo distribution into the GAMLSS framework in order to provide a very flexible regression model for situations where the response variable is considered as rates and proportions, i.e. on the support (0,1). Second to study the variables which may potentially affect a football team in a championship using this new sophisticated method based on the GAMLSS framework, and not just consider football as a simple illustrative example. This data was already introduced in a short conference paper by the authors [18], but here we

develop the model with some simulation studies and provide a thorough analysis of the data. Please note that form causal conclusions cannot be made as this is an observational study.

This paper is organised as follows. Section 2 summarises the GAMLSS framework, showing procedures to obtain the final model, i.e. GAMLSS estimation processes and selection of the response variable distribution and explanatory variables, present the new model based on the SHASHo distribution and also present some Monte Carlo simulations for the proposed model. In Section 3, we consider a discussion regarding the full data set, showing some descriptive statistics and presenting the results of some fitted models based on the GAMLSS methodology including our new model. A detailed discussion regarding the results is presented in Section 4. Finally, Section 5 ends the paper with some concluding remarks.

## 2. GAMLSS framework

GAMLSS is a very flexible class of semi-parametric regression models which involves a distribution, that does not necessarily belong to the exponential family, for the response variable and may involve parametric and/or non-parametric smoothing terms when modelling any or all of the parameters of the distribution as functions of a set of explanatory variables [22]. This methodology is already implemented in the `gamlss` package [26] in `R` [20] that includes several continuous and discrete distributions with up to four parameters (conveniently denoted by $\mu$,$\sigma$, $\nu$ and $\tau$), including a few highly skewed and kurtotic ones. This framework is being widely used in many different fields, such as economics [23], natural sciences [30] and medical field [21], among others.

Mathematically, a GAMLSS model assumes that independent observations $Y_i$ have probability (density) function $f_Y(y_i|\boldsymbol{\theta})$ conditional on $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})^\top$ a vector of four distribution parameters. The GAMLSS model is defined as

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

for $k = 1, 2, 3, 4$, where $g_k(\cdot)$ is a known monotonic link function (usually determined by the range of parameters) relating the distribution parameter $\theta_k$ to the predictor $\eta_k$, $\boldsymbol{X}_k$ is a known design matrix, $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \ldots, \beta_{J'_k k})$ is a parameter vector of length $J'_k$, $h_{jk}$ is a smooth non-parametric function of an explanatory variable $\boldsymbol{x}_{jk}$. If $\sum_{j=1}^{J_k} h_{jk}(\boldsymbol{x}_{jk}) = 0$, Equation (1) reduces to the GAMLSS parametric model.

### 2.1. The logitSHASHo distribution

As presented in [11], any distribution on the range $-\infty < Z < \infty$ can be transformed to a restrictive range $0 < Y < 1$ using an inverse logit transformation, i.e.

$$Y = \frac{1}{1 + \exp(-Z)}. \tag{2}$$

Based on the skew student $t$ (SST) [9] distribution, Hossain et al. [11] presented the logitSST distribution, i.e. the inverse logit transformation of the skew student $t$ distribution,

as a competitive and alternative model to the beta distribution. In this paper, we propose a new very flexible distribution to model a response variable (e.g. points rate) on the unit interval (0, 1), with four parameters, using the transformation given in (2).

If $-\infty < Z < \infty$ follows a sinh–arcsinh distribution [13], denoted by $Z \sim \text{SHASHo}(\mu, \sigma, \nu, \tau)$, with probability density function (pdf) given by

$$f_Z(z|\mu,\sigma,\nu,\tau) = \frac{\tau c}{\sigma\sqrt{2\pi}} \frac{\exp\left(-r^2/2\right)}{\left(1+w^2\right)^{1/2}},$$

where $c = \cosh(\tau \operatorname{arcsinh}(w) - \nu)$, $r = \sinh(\tau \operatorname{arcsinh}(w) - \nu)$ and $w = (z-\mu)/\sigma$, for $-\infty < z < \infty, -\infty < \mu < \infty, \sigma > 0, -\infty < \nu < \infty$ and $\tau > 0$, then $Y = 1/(1+\mathrm{e}^{-Z})$ follows a logit sinh–arcsinh distribution, denoted by $Y \sim \text{logitSHASHo}(\mu, \sigma, \nu, \tau)$ for $0 < Y < 1$. The pdf and cumulative distribution function (cdf) of the logitSHASHo distribution are given by

$$f(y|\mu,\sigma,\nu,\tau) = \frac{\tau\, c^\star}{\sigma\sqrt{2\pi}\,(y-y^2)} \frac{\exp\left(-r^{\star 2}/2\right)}{\left(1+w^{\star 2}\right)^{1/2}}$$

and

$$F(y;\mu,\sigma,\nu,\tau) = \Phi(r^\star), \tag{3}$$

respectively, where $c^\star = \cosh(\tau \operatorname{arcsinh}(w^\star) - \nu)$, $r^\star = \sinh(\tau \operatorname{arcsinh}(w^\star) - \nu)$, $w^\star = (-\log(1/y-1) - \mu)/\sigma$ and $\Phi(.)$ denotes the cdf of the standard normal distribution.

In this parametrisation, $\mu$ and $\sigma$ are the location and scale parameters, respectively, $\nu$ is a skewness parameter (skewness increases with increasing $\nu$) where $\nu > 0$ and $\nu < 0$ present positive and negative skewness respectively, and $\tau$ is a parameter related to the tails of the distribution where the distribution presents heavier and lighter tails than the normal distribution if $\tau < 1$ and $\tau > 1$, respectively.

The GAMLSS parametric model based on the logitSHASHo distribution, i.e. $Y \sim \text{logitSHASHo}(\mu, \sigma, \nu, \tau)$, can be expressed as

$$\begin{aligned} g_1(\mu) &= \eta_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1, \\ g_2(\sigma) &= \eta_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2, \\ g_3(\nu) &= \eta_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3, \\ g_4(\tau) &= \eta_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4, \end{aligned} \tag{4}$$

in which we use, in the next sections, the identity link function for $g_1$ and $g_3$, and the logarithmic link function for $g_2$ and $g_4$. These link functions were chosen since in the GAMLSS framework they are usually determined by the range of the parameters of the response variable distribution [7,27,29].

### 2.2. Estimation

The total log-likelihood function for the GAMLSS parametric model (4) under the assumption that observations of the response variable are independent is given by

$$l = \sum_{i=1}^{n} \log f_Y(y_i|\mu, \sigma, \nu, \tau). \tag{5}$$

Within GAMLSS, we maximise the log-likelihood function (5) in order to obtain the estimates for $\boldsymbol{\beta}_k$ in (4). Three different options, based on an iteratively reweighted (penalised) least squares algorithm, are available in `gamlss` package to do this task: (i) `CG` algorithm, which is a generalisation of the algorithm proposed by [6]; (ii) `RS` algorithm proposed by [22] and (iii) a combination of both methods denoted by `mixed` in `gamlss`. `CG` algorithm jointly updates all parameters of the response variable distribution since it uses all cross derivatives in its estimation process. `RS` algorithm maximises the likelihood over each of the parameters in turn, cycling until convergence and should be preferred in most cases, since it is computationally more stable than the first algorithm. The `mixed` procedure should be taken into account in cases where distributions with highly correlated parameters are used. Further information about those methods is given in [22] and complete flow charts of the algorithms' mechanism are provided in [27].

### 2.3. Selecting the distribution

The first step to fit a GAMLSS model to a data set is to choose an appropriate distribution for the response variable. Basically, this is achieved in two different stages: fitting and diagnostic. During the fitting stage, we fit and compare different fitted models using a generalised Akaike information criterion (GAIC), given by $GAIC(k) = -2 \times \hat{l} + k \times df$, where $\hat{l}$ is the fitted log-likelihood function, $df$ are the effective degrees of freedom of the fitted model and $k$ if a penalty for each degree of freedom in the model [29]. More common measures of goodness of fit are the Akaike information criterion (AIC) [1] and the Schwarz Bayesian criterion (SBC) [25], which are special cases of the GAIC when $k = 2$ and $k = \log n$, respectively.

The diagnostic stage is based on the normalised (randomised) quantile residuals [8]. The main advantage of this type of residual is that, whatever the distribution of the response variable their true values $r_i$, $i = 1, \ldots, n$, always have a standard normal distribution given the assumption that the model is correct. Mathematically, the normalised quantile residuals are given by $\hat{r}_i = \Phi^{-1}(\hat{u}_i)$, where $\Phi^{-1}$ is the qf of a standard normal variable and $\hat{u} = F_Y(y|\hat{\boldsymbol{\theta}})$ is the fitted cdf. Given those residuals, we provide the residuals diagnostic plots, such as the normal probability plot with envelope [2].

Another alternative available in `gamlss` is the worm plots [28]. Worm plots are detrended normal Q–Q residual plots that allow the detection of inadequacies in the model globally or within a specific range of an explanatory variable. Basically, if a vertical shift, a slope, a quadratic or a cubic shape is observed, it may indicate misfits in the mean, variance, skewness and excess kurtosis of the residuals, respectively. Further details can be obtained in [27].

### 2.4. Selecting the explanatory variables

Some approaches are available in `gamlss` package [26] in order to select which variables are statistically significant to the model. Here, we used the function `StepGAICAll.A` that is basically a set of forward and backward procedures and is described in [19,27] as follows.

(1) Fix $\sigma$, $\nu$ and $\tau$, and perform a forward GAIC selection procedure to select an appropriate model for $\mu$.
(2) Fix $\nu$ and $\tau$, and perform a forward GAIC selection procedure to select an appropriate model for $\sigma$, given the model for $\mu$ in (1).
(3) Fix $\tau$ and perform a forward GAIC selection procedure to select an appropriate model for $\nu$, given the models for $\mu$ and $\sigma$ obtained in (1) and (2), respectively.
(4) Perform a forward GAIC selection procedure to select an appropriate model for $\tau$, given the models for $\mu$, $\sigma$ and $\nu$ obtained in (1), (2) and (3), respectively.
(5) Given the models for $\mu$, $\sigma$ and $\tau$ obtained in (1), (2) and (4), respectively, perform a backward GAIC selection procedure to select an appropriate model for $\nu$.
(6) Given the models for $\mu$, $\nu$ and $\tau$ obtained in (1), (5) and (4), respectively, perform a backward GAIC selection procedure to select an appropriate model for $\sigma$.
(7) Given the models for $\sigma$, $\nu$ and $\tau$ obtained in (6), (5) and (4), respectively, perform a backward GAIC selection procedure to select an appropriate model for $\mu$.

The resulting model may contain different explanatory variables for each of the parameters $\mu, \sigma, \nu$ and $\tau$. Please note that as in most model selection procedures, the order of the steps can affect the final chosen model. In GAMLSS, we start by choosing a model for $\mu$ because this is usually related to the location (and usually related to the first moment), and we do not want to select a complicated model for $\sigma$ (which may be related to the second moment) and for the remaining two parameters (which may be related to the third and fourth moments) as a substitute for the $\mu$ model.

### 2.5. Simulation study

In this section, we conduct two Monte Carlo simulation studies. The objective of the first study is to assess the finite sample behaviour of the maximum likelihood estimates (MLEs) of the parameters for different sample sizes $n$ and parameter settings. In the second study, suggested by a reviewer, we investigate the behaviour of the selecting explanatory variables method presented in Section 2.4.

We simulate logitSHASHo random variables using the quantile function (qf), which is obtained by inverting the cdf $F(y) = u$ using (4) to give $y = F^{-1}(u) = Q(u)$. The qf of $Y \sim \text{logitSHASHo}(\mu, \sigma, \nu, \tau)$ is given by

$$y = Q(u) = \left(\exp\left\{-\sigma \sinh\left[\frac{\nu}{\tau} + \frac{1}{\tau}\text{arcsinh}\left(\Phi^{-1}(u)\right)\right] - \mu\right\} + 1\right)^{-1}, \qquad (6)$$

where $\Phi^{-1}(\cdot)$ denotes the qf of a standard normal distribution. Equation (6) is used to simulate random variables by fixing $\mu, \sigma, \nu, \tau$ and setting $u$ as an uniform random variable on the $(0, 1)$ interval.

**Table 1.** The biases, MSEs and PVC based on 1000 simulations of the logitSHASHo GAMLSS for sample size $n = 50$, 100 and 300.

| | $n = 50$ | | | $n = 100$ | | | $n = 300$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameters | Bias | MSE | PVC | Bias | MSE | PVC | Bias | MSE | PVC |
| $\beta_{01}$ | 0.147 | 0.207 | 0.89 | 0.079 | 0.095 | 0.92 | 0.019 | 0.018 | 0.93 |
| $\beta_{11}$ | 0.255 | 0.515 | 0.91 | 0.109 | 0.238 | 0.94 | 0.035 | 0.074 | 0.94 |
| $\beta_{02}$ | 0.046 | 0.522 | 0.95 | 0.007 | 0.154 | 0.95 | 0.016 | 0.049 | 0.95 |
| $\beta_{12}$ | 0.169 | 0.953 | 0.95 | 0.131 | 0.339 | 0.95 | 0.084 | 0.099 | 0.95 |
| $\beta_{03}$ | 0.182 | 0.216 | 0.96 | 0.086 | 0.066 | 0.97 | 0.017 | 0.008 | 0.96 |
| $\beta_{13}$ | 0.161 | 0.454 | 0.94 | 0.055 | 0.198 | 0.94 | 0.018 | 0.057 | 0.93 |
| $\beta_{04}$ | 0.125 | 0.111 | 0.94 | 0.063 | 0.038 | 0.96 | 0.022 | 0.010 | 0.95 |
| $\beta_{14}$ | 0.133 | 0.222 | 0.94 | 0.079 | 0.100 | 0.95 | 0.046 | 0.032 | 0.94 |

- First study – asymptotic and convergence

Here, we consider the logitSHASHo GAMLSS model by modelling the parameters using the explanatory variable $x_i$, namely $\mu_i = \beta_{01} + \beta_{11}x_i$, $\sigma_i = \exp(\beta_{02} + \beta_{12}x_i)$, $\nu_i = \beta_{03} + \beta_{13}x_i$ and $\tau_i = \exp(\beta_{04} + \beta_{14}x_i)$. The sample sizes are generated by taking $n = 50$, 100 and $n = 300$. The response variable values, denoted by $y_1, \ldots, y_n$, are generated from the logitSHASHo distribution using the qf (6), in which the $\boldsymbol{\beta}$ parameters were fixed and the values of the explanatory variable $x_i$ were generated randomly from a binomial $(n, 0.5)$ distribution. For each scenario, all results are obtained from 1000 Monte Carlo replications and the simulations are carried out using the R programming language. For each replication, a random sample of size $n$ is drawn from the logitSHASHo GAMLSS (4) and the RS algorithm is used for maximising the total log-likelihood function (5).

For the GAMLSS model, the true parameter values used in the data-generating processes are $\mu_i = -1 + 1x_i$, $\sigma_i = \exp[\log(0.5) + \log(2)x_i]$, $\nu_i = 0.5 + 0.3x_i$ and $\tau_i = \exp[\log(0.5) + \log(2)x_i]$. With these parameter settings, we are assuming that $Y|x_i = 0 \sim$ logitSHASHo$(-1, 0.5, 0.5, 0.5)$ and $Y|x_i = 1 \sim$ logitSHASHo$(0, 1, 0.8, 1)$. For each fit, the biases, mean squared errors (MSEs) and the parameter value coverage (PVC) at 95% level of confidence are calculated and the results are reported in Table 1.

In Figure 1, the estimated and generated densities of the logitSHASHo GAMLSS are presented using the average estimates (AE) of the $\beta$ parameters, for $x = 0$ and 1 and for sample sizes $n = 50$, 100 and 300. The results of the Monte Carlo study in Table 1 indicate that the MSEs of the MLEs of the parameters decay toward zero as the sample size increases, as expected. As $n$ increases, the AEs of the parameters tend to be closer to the true parameter. Finally, the PVC is substantially high even for small sample values.

- Second study – Validation of the variable selection method

The purpose of this study is to investigate the behaviour of the variable selection method described in Section 2.4, i.e. when the regression model is not specified, being necessary to identify which explanatory variables will compose it. Here we consider the following regression structure for the four parameters of the logitSHASHo distribution: $\mu = \beta_{01} + \beta_{11}x_1 + \beta_{21}x_3$, $\sigma = \exp[\beta_{02} + \beta_{12}x_2]$, $\nu = \beta_{03} + \beta_{13}x_4$ and $\tau = \exp[\beta_{04}]$, where the true parameter values of the logitSHASHo GAMLSS model used in the data-generating processes are $\mu = -0.3 + 0.6x_1 + 0.2x_3, \sigma = \exp[\log(0.1) + \log(1.5)x_2], \nu = 0.2 - 0.6x_4$

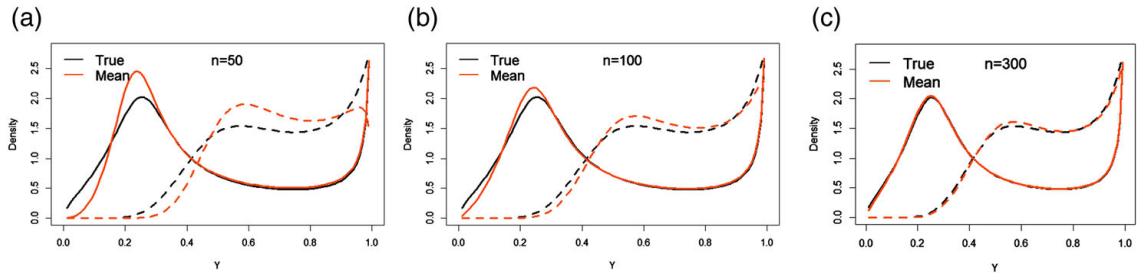


**Figure 1.** The logitSHASHo density for the true $\beta$ parameter values and at the average estimates of the $\beta$'s for (a) $n = 50$, (b) $n = 100$ and (c) $n = 300$, for $x = 0$ (–) and $x = 1$ (- -).

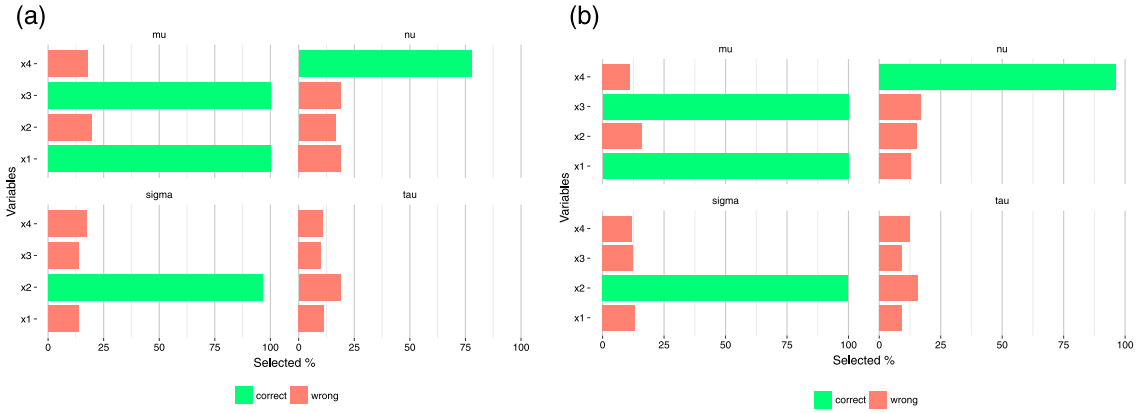


**Figure 2.** The correct/incorrect specification percentages of the explanatory variables selected by the `StepGAICAll.A` method for (a) $n = 100$ and (b) $n = 200$.

and $\tau = \exp[\log(0.5)]$, and the explanatory variables were generated based on the following distributions: $x_1 \sim \text{Binomial}(n, 0.5)$, $x_2 \sim \text{Binomial}(n, 0.5)$, $x_3 \sim U(1, 10)$ and $x_4 \sim U(0, 1)$.

The sample sizes are generated by taking $n = 100$ and $n = 200$ and, for each scenario, all results are obtained from 1000 Monte Carlo replications. For each replication, the `StepGAICAll.A` method is applied to select the model parameters, then, for each replication the estimates are noted and, at the end of all replications, the correct/incorrect specification percentage of each parameter is calculated. Figure 2 displays the percentage of correct/incorrect variables selected for each model parameters. We may conclude that the `StepGAICAll.A` method is able to correctly identify the explanatory variables for all parameters of the logitSHASHo distribution, since as we can see in Figure 2 in almost 100% of the replications an explanatory variable that should be considered in a parameter, in fact was selected by the procedure. Furthermore, explanatory variables that should have not been selected by the described method were considered in a very few number of replications (in the worst case, the variable $x_2$ was selected for $\tau$ only 18.8% for $n = 100$).

## 3. Data analysis

In this section, we introduce the data set used in this paper. We also present some results by fitting and comparing different distributions on support (0,1) using the GAMLSS framework in order to find the best model to explain the response variable, points rate, using the available explanatory variables.

### 3.1. Data set description

In a football game, external interferences, such as referee mistakes or the absence of one of the best players, can interfere dramatically in the final results, especially in championships organised in the format of cups, in which losing a game means for a team, the elimination of the championship. In order to minimise these problems, we used data from leagues instead of cups, since in leagues all football teams play against each other twice and so this type of external interferences can be reduced.

We collected all information from the four most important football leagues which are affiliated to the UEFA. The leagues were Barclays Premier League from England, Bundesliga from Germany, Serie A from Italy and BBVA league from Spain, and there was considered three different seasons (2011–2012, 2012–2013 and 2013–2014). The data set comprised the following variables the response variable points rate using the explanatory variables: league, season, goals for, goals against, goals difference, yellow and red cards, position, classification to UEFA league, relegation, shots, shots on goal, clean sheets, offsides, fouls, fouled (received fouls), tackles, interception, possession, dribble, shot concede and pass accuracy. This data set is available for consulting at http://www.whoscored.com.

For all the studied leagues, the game structure remains the same, i.e. in all leagues the football teams play against each other twice (home and away game). However in the Bundesliga, there are only 18 teams, while the other leagues are composed by 20, for this reason the total number of games are different, 34 and 38, respectively, and therefore the resulting number of observations is $n = 234$. Another difference among these leagues is the way that the teams who will be playing at the UEFA Leagues (Champions and Europe) are selected. The same occurs with the number of relegations. To avoid any problems with the different amount of games for each league, we used all information (explanatory variables) per game. Also we standardised the variables yellow and red cards to avoid problems related to scale.

### 3.2. Marginal analysis

As described in Section 2.3, the first step to fit a GAMLSS model is to select a suitable distribution for the response variable. Here, in addition to the logitSHASHo and logitSST distributions, we consider some interesting distributions with support on the unit interval such as the logit-normal (logitNO) and the generalised beta distributions. Their density functions are given respectively by

$$f_Y(y|\mu,\sigma) = \frac{1}{y\,(1-y)\,\sigma\,\sqrt{2\pi}}\,\exp\left\{-\frac{1}{2\sigma^2}\left[\log\left(\frac{y}{1-y}\right)-\log\left(\frac{\mu}{1-\mu}\right)\right]^2\right\},$$

where $0<y<1$, $0<\mu<1$ and $\sigma>0$ and

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{\tau\,\nu^{\beta}\,y^{\tau\,\alpha-1}(1-y^{\tau})^{\beta-1}}{B(\alpha,\beta)\,\left[\nu+(1-\nu)\,y^{\tau}\right]^{(\alpha+\beta)}},$$

where $0<y<1$, $\alpha=\mu\,(1-\sigma^2)/\sigma^2$ and $\beta=(1-\mu)\,(1-\sigma^2)/\sigma^2$, and $\alpha>0$, $\beta>0$. Figure 3(a) shows a box plot of the points rate and Figure 3(b) displays the histogram of the points rate and the density functions of the fitted logitSHASHo, logitSST, logitNO,

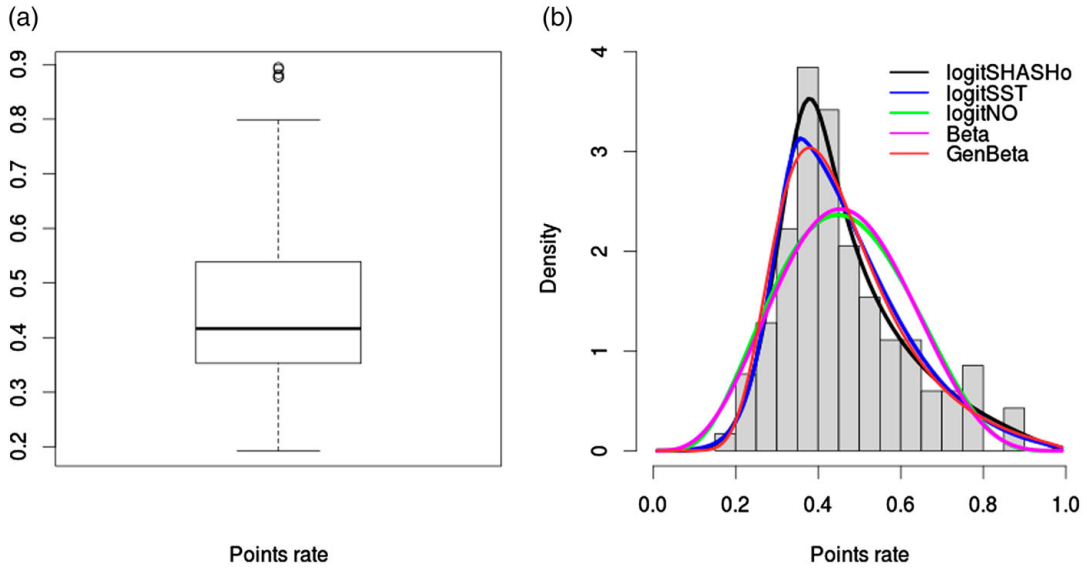


**Figure 3.** (a) Box plot and (b) distribution of the response variable points rate.

**Table 2.** Descriptive statistics for the response variable points rate.

| Mean | Median | SD | Skewness | Kurtosis | Minimum | Maximum |
|---|---|---|---|---|---|---|
| 0.4587 | 0.4165 | 0.1523 | 0.8230 | 3.2200 | 0.1930 | 0.8950 |

beta and generalised beta distributions that are suitable options since we can see that the response variable is slightly positively skewed on support $(0, 1)$. Table 2 confirms this statement (skewness $= 0.8230$) and also displays values of the mean, median, standard deviation (SD), kurtosis, minimum and maximum values. We shall highlight here that there is a slightly difference between the beta GAMLSS model and the beta regression model proposed by [10], i.e. for the GAMLSS parameterisation $Var(Y) = \sigma^2\mu(1-\mu)$ and in [10]'s parameterisation $Var(Y) = (1+\phi)^{-1}\mu(1-\mu)$ where $\phi > 0$ is a precision parameter.

### 3.3. Modelling of distribution parameters

Figure 4 displays scatter plots of the response variable against some of the available explanatory variables, which are used to select additive terms that will compose the regression model to explain the points rate. We can clearly see, in the first column of Figure 4, that the explanatory variables shots on goal per game (ShG.Pg), clean sheet (ClSh) and shots per game (Sh.Pg) have a positive linear relationship with points rate (correlations equal to 0.82, 0.70 and 0.71, respectively). On the other hand, shots conceded per game (Sh.con.Pg) has a negative linear relationship with the response variable (correlation equals to $-0.68$). Since these explanatory variables present high correlations, we can conclude that, probably, they will be considered in the best final models, modelling some of the parameters (mainly $\mu$) as linear functions. Finally, it seems that the number of yellow cards (YC) does not have a strong linear relationship with total points (correlation equals to $-0.26$ ). However, this explanatory variable presents a great variability in points rate, indicating that we may need to consider it in order to model parameters related to the scale and/or skewness and kurtosis of the response variable. Please note that, in order to keep a clear scatter plot,

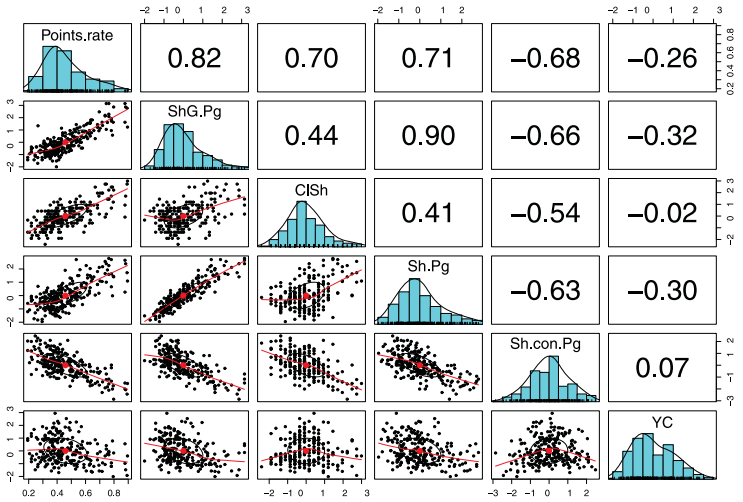


**Figure 4.** Scatter plots and correlation coefficient between response and some explanatory variables.

**Table 3.** Statistics from the fitted models.

| Model | Parameters | df | GD | AIC | SBC |
|---|---|---|---|---|---|
| logitSHASHo | 4 | 10 | **−656.28** | **−636.27** | **−601.72** |
| Beta | 2 | 7 | −645.19 | −631.59 | −600.50 |
| logitSST | 4 | 8 | −646.19 | −628.19 | −597.09 |
| logitNO | 2 | 9 | −644.72 | −626.72 | −595.62 |
| Generalised beta | 4 | 9 | −642.48 | −624.48 | −593.38 |

the variables displayed here are the ones which were selected to compose the best fitted model presented further on this section (please check the supplementary file to see the other relationships).

Here, we applied the steps described in Section 2.4 to select the explanatory variables using each of the five distributions named in Section 3.2. Table 3 displays the number of distribution parameters, the total effective degrees of freedom used in the model, the values of the global deviance (GD), AIC and SBC for each model, which were used to compare the fitted models.

Table 3 indicates that the best fitted model according to both the AIC and SBC criteria is the one based on the logitSHASHo distribution (−636.27 and −601.72, respectively). As we highlighted in Section 1, the beta regression model is not very flexible when we need to model skewness and kurtosis, thus despite the close values from especially the SBC criterion when we compare it to the model based on the logitSHASHo distribution, the residual analysis of this model indicates some problems that will be discussed further in Section 3.4.

The best model from the logitSHASHo distribution under the GAMLSS parametric framework (4) and its estimates and standard errors (in in parentheses) are listed in Table 4.

### 3.4. Model checking

Figure 5 displays the normalised quantile residuals from the chosen logitSHASHo model. Panels (a) and (b) show the residuals against fitted values of $\mu$ against an index, respectively. The kernel density estimate of the residuals can be seen in Panel (c), whereas Panel

**Table 4.** Estimates of the coefficients associated with each covariate considered in the response variable distribution parameters and their respectively standard errors (in parentheses).

| Variable | $\mu$ | $\log \sigma$ | $\nu$ | $\log \tau$ |
|---|---|---|---|---|
| Intercept | −0.19 (0.02) | −1.96 (0.20) | 0.07 (0.06) | −0.39 (0.11) |
| ShG.Pg | 0.49 (0.04) | 0.13 (0.06) | – | – |
| ClSh | 0.29 (0.02) | – | – | – |
| Sh.Pg | −0.13 (0.04) | – | – | – |
| Sh.con.Pg | −0.05 (0.02) | – | – | – |
| YC | – | – | −0.12 (0.04) | – |

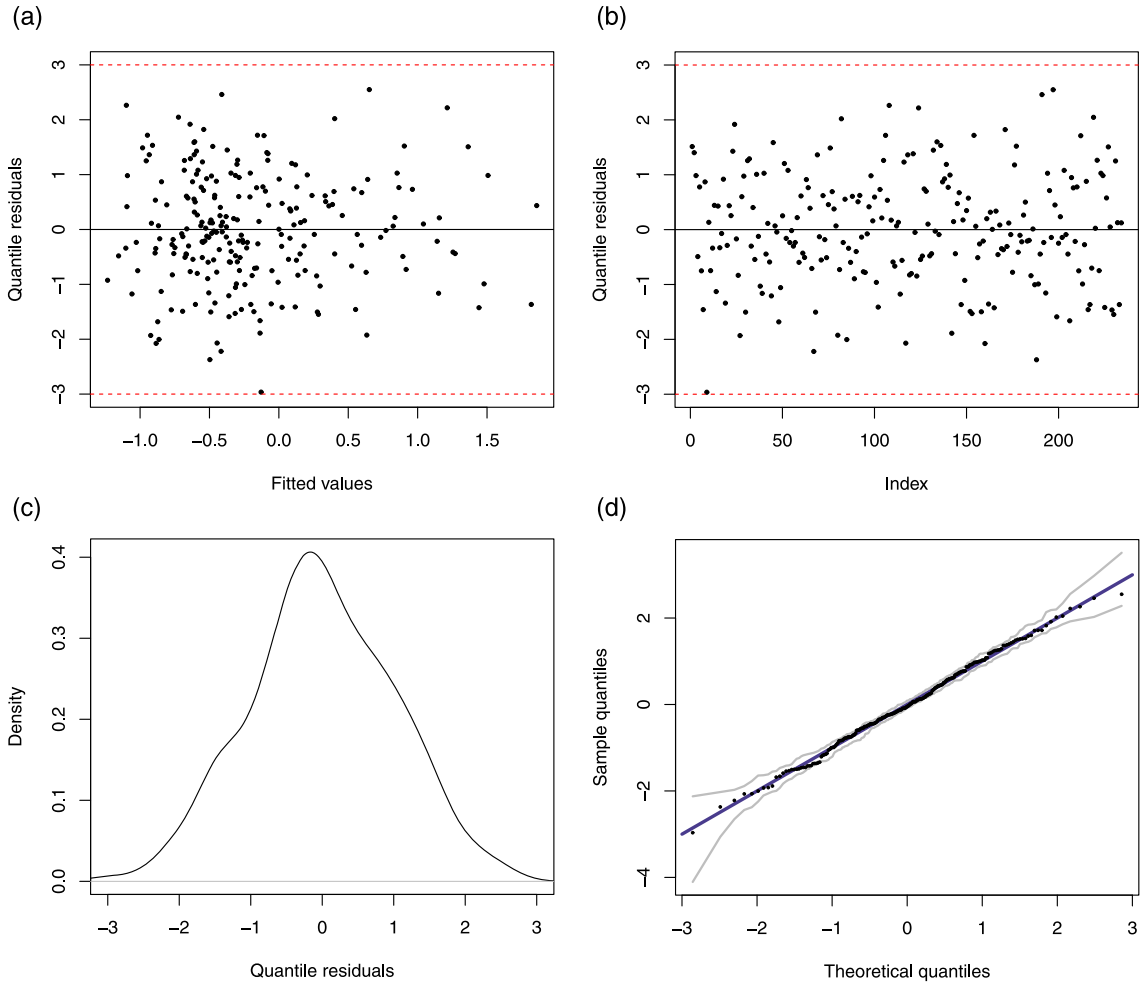


**Figure 5.** The residuals from the logitSHASHo model: (a) against fitted values; (b) against index; (c) kernel density estimate; and (d) simulated envelope.

(d) presents a simulated envelope [2], where all dotted points are within the greyish bands. We may observe in these plots that the normalised quantile residuals seem to follow approximately a normal distribution indicating a suitable fitted model.

In addition, we display in Figure 6 worm plots obtained from the best fitted models based on every distribution considered in this paper. We can see that the models based on the logitSHASHo and logitSST distributions do not present any trend (vertical shift, slope, quadratic or cubic shape), thus it fitted really well the skewness and kurtosis present in the response variable. As for the other models based on the beta, logitNO and generalised

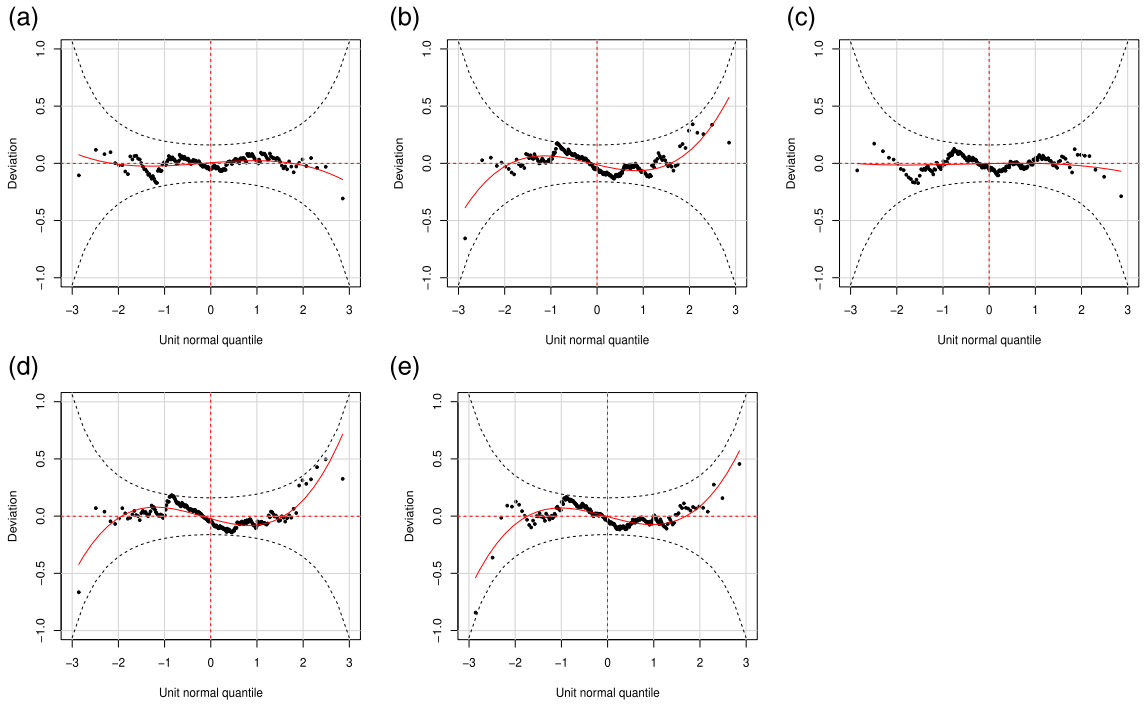


**Figure 6.** Worm plots of the (a) logitSHASHo; (b) beta; (c) logitSST; (d) logitNO and (e) generalised beta GAMLSS models.

beta distributions, we can clearly see that the normalised quantile residuals present a cubic shape, indicating a problem with their kurtosis. Hence, based on Figure 6 and on AIC and SBC measures presented in Table 3, we can conclude that the logitSHASHo model is the most suitable model among the five used in this paper to explain the current data.

## 4. Discussion

The results obtained through the logitSHASHo fitted model given in Table 4 show that shots on goal (ShG.Pg) and clean sheet (ClSh) are positively associated with the location parameter $\mu$ in the points rate distribution. This relationship can be explained using simple game facts: a team that avoids conceding goals (clean sheet) and creates real chances of scoring (shots on goal) has more chance to be successful. The explanatory variable shots per game (Sh.Pg), surprisingly, has a negative effect on the location parameter $\mu$, even though it presents a high positive correlation with points rate (Figure 4). Statistically, this can be explained by using the value of the partial correlation between shots per game and points rate given shots on goal, which is negative ($-0.11$). In practice, we may say that just arbitrarily shooting has a negative effect, since in many cases a football team trying to score from anywhere in the pitch, without presenting any real risk for the opposite team is less likely to score. The last covariate used to model $\mu$ is shots conceded (Sh.con.Pg) which has a negative effect on it, since it increases the chances to concede goals and consequently to lose the match.

A fact that should be highlighted is the absence of some explanatory variables to explain the location parameter, specifically, possession (please check the supplementary file in order to see this relationship in the provided scatter plot), leagues and season. For season, we note that the football teams for those leagues have a points rate average that does not

tend to change from season to another. However, possession and leagues are quite interesting and may be more difficult to be accepted. Generally, football experts and commentators claim that some leagues are more competitive than others, but the results obtained in this paper may indicate that this may be too subjective and not be completely right, since the effect of league is not significant in the model. Regarding possession, we can try to explain its absence with different strategies during the game, i.e. only keeping the ball may not be sufficient for having a good points rate, once in some situations a team can play very defensively and try to score in counter attacks (this may indicate that both strategies can be equally efficient).

The value of the scale parameter $\sigma$ increases according to the number of shots on goal per game (ShG.Pg) by the football teams and yellow cards (YC) has a negative linear effect on the skewness parameter $\nu$. Finally, the kurtosis parameter $\tau$ is a constant smaller than one ($\approx 0.68$), i.e. the final model from the logitSHASHo distribution presents heavy tails. These characteristics indicate why the normalised residuals obtained from the beta regression model (Figure 6b) did not behave well, since it is not capable to model different degrees of skewness and kurtosis.

## 5. Concluding remarks

In this paper, we proposed the logit sinh–arcsinh distribution (logitSHASHo) defined on the unit interval $(0, 1)$ which admits high degrees of skewness and kurtosis. Also, it is very versatile and it can be used to analyse different types of data sets. Based on the logitSHASHo distribution, we proposed a logitSHASHo regression model using the flexibility of the GAMLSS framework. The new regression model can be used as an alternative to the beta regression model to fit heavy-tailed response variables on the unit interval $(0, 1)$. The distribution parameters are modelled as parametric linear and/or additive non-parametric smoothing functions of explanatory variables. Furthermore, we performed some simulation studies for the new regression model under different sample sizes and we tested the performance of the explanatory variable selection method through the `StepGAICAll.A` function available in the `gamlss` package. We also discussed model checking analysis using the normalised quantile residuals in the new regression model fitted to a real data. An application related to the points rate of the football teams data set demonstrated that it can be used quite effectively to provide better fits than other regression models in the literature.

## References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automat. Control 19 (1974), pp. 716–723.

[2] A.C. Atkinson, *Plots, Transformations, and Regression: an Introduction to Graphical Methods of Diagnostic Regression Analysis*, Clarendon Press, Oxford, 1985.
[3] D.R. Brillinger, *Modeling game outcomes of the Brazilian 2006 series a championship as ordinal-valued.*, Braz. J. Probab. Stat. 22 (2008), pp. 89–104.
[4] K. Bury, *Statistical Distributions in Engineering*, Cambridge University Press, New York, 1999.
[5] P.H.R. Cerqueira, L.R. Nakamura, R.R. Pescim and R.A. Leandro, *Investigating the underlying causal network on European football teams*, J. Data Sci. 15 (2017), pp. 293–312.
[6] T.J. Cole and P.J. Green, *Smoothing reference centile curves: the LMS method and penalized likelihood*, Stat. Med. 11 (1992), pp. 1305–1319.
[7] F. De Bastiani, R.A. Rigby, D.M. Stasinopoulos, A.H.M.A. Cysneiros and M.A. Uribe-Opazo, *Gaussian Markov random field spatial models in GAMLSS*, J. Appl. Stat. 45 (2018), pp. 168–186.
[8] P.K. Dunn and G.K. Smyth, *Randomized quantile residuals*, J. Comput. Graph. Statist. 5 (1996), pp. 236–245.
[9] C. Fernández and M.F. Steel, *On Bayesian modeling of fat tails and skewness*, J. Amer. Statist. Assoc. 93 (1998), pp. 359–371.
[10] S.L.P. Ferrari and F. Cribari-Neto, *Beta regression for modelling rates and proportions*, J. Appl. Stat. 31 (2004), pp. 799–815.
[11] A. Hossain, R. Rigby, M. Stasinopouos and M. Enea, *Centile estimation for a proportion response variable*, Stat. Med. 35 (2016), pp. 895–904.
[12] N.L. Johnson, S. Kotz and N. Balakrishnan, *Continuous Univariate Distributions*, Vol. 2. Wiley, New York, 1995.
[13] M.C. Jones and A. Pewsey, *Sinh–arcsinh distribution*, Biometrika 96 (2009), pp. 761–780.
[14] D. Karlis and I. Ntzoufras, *Bayesian modeling of football outcomes: using the Skellam's distribution for the goal difference*, IMA J. Manag. Math. 20 (2009), pp. 133–145.
[15] A. Lee, *Modeling scores in the Premier League: Is Manchester United really the best?*, Chance 10 (1997), pp. 15–19.
[16] F. Louzada, A.K. Suzuki and L.E.B. Salasar, *Predicting match outcomes in the English Premier League: which will be the final rank?* J. Data Sci. 12 (2014), pp. 235–254.
[17] S. Nadarajah and S. Kotz, *Multitude of beta distributions with applications*, Statistics 41 (2007), pp. 153–179.
[18] L.R. Nakamura, P.H.R. Cerqueira, T.G. Ramires, R.R. Pescim, R.A. Rigby and D.M. Stasinopoulos, *Beyond regression: what does really affect a football team on championships*, Proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), pp. 91–94 (2017).
[19] L.R. Nakamura, R.A. Rigby, D.M. Stasinopoulos, R.A. Leandro, C. Villegas and R.R. Pescim, *Modelling location, scale and shape parameters of the Birnbaum–Saunders generalized t distribution*, J. Data Sci. 15 (2017), pp. 221–238.
[20] R Core Team, A language and environment for statistical computing. Software available at http://cran.r-project.org/.
[21] T.G. Ramires, L.R. Nakamura, A.J. Righetto, E.M.M. Ortega and G.M. Cordeiro, *Predicting survival function and identifying associated factors in patients with renal insufficiency for the metropolitan area of Maringá, Brazil*, Cadernos de Saúde Pública 34 (2018), pp. 1–13.
[22] R.A. Rigby and D.M. Stasinopoulos, *Generalized additive models for location, scale and shape, (with discussion)*, Appl. Stat. 54 (2005), pp. 507–554.
[23] G. Scandroglio, A. Gori, E. Vaccaro and V. Voudouris, *Estimating VaR and ES of the spot price of oil using futures-varying centiles*, Int. J. Financ. Eng. Risk Manage. 1 (2013), pp. 6–19.
[24] G. Schaugerger, A. Groll and G. Tutz, *Analysis of the importance of on-field covariates in the German Bundesliga*, J. Appl. Stat. 9 (2018), pp. 1561–1578.
[25] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. 6 (1978), pp. 461–464.
[26] D.M. Stasinopoulos and R.A. Rigby, *Generalized Additive Models for Location Scale and Shape (GAMLSS) in R*, J. Stat. Software 23 (2007), pp. 1–10.
[27] D.M. Stasinopoulos, R.A. Rigby, G.Z. Heller, V. Voudouris and F. De Bastiani, *Flexible Regression and Smoothing: Using GAMLSS in R*, Chapman and Hall/CRC, London, 2017.
[28] S. van Buuren and M. Fredriks, *Worm plot: a simple diagnostic device for modelling growth reference curves*, Stat. Med. 20 (2001), pp. 1259–1277.

[29] V. Voudouris, R. Gilchrist, R. Rigby, J. Sedwick and D. Stasinopoulos, *Modelling skewness and kurtosis with the BCPE density in GAMLSS*, J. Appl. Stat. 39 (2012), pp. 1279–1293.
[30] D.D. Zhang, D.H. Yan, Y.C. Wang, F. Lu and S.H. Liu, *GAMLSS-based nonstationary modeling of extreme precipitation in Beijing–Tianjin–Hebei region of China*, Natur. Hazards 77 (2015), pp. 1037–1053.