






Incorporating Clustering Techniques into GAMLSS

Thiago G. Ramires ^{1,*}, Luiz R. Nakamura ^{2,†}, Ana J. Righetto ^{3,†}, Andréa C. Konrath ^{2,†}
and Carlos A. B. Pereira ^{4,†}

¹ Campus Apucarana, Universidade Tecnológica Federal do Paraná, Apucarana 86812-460, Brazil

² Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Florianópolis 88040-900, Brazil; luiz.rn@gmail.com (L.R.N.); andreak@gmail.com (A.C.K.)

³ Alvaz Agritech, Londrina 86050-268, Brazil; ajrighetto@gmail.com

⁴ Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo 05508-090, Brazil; cadebp@gmail.com

* Correspondence: thiagogentil@gmail.com

† These authors contributed equally to this work.

Abstract: A method for statistical analysis of multimodal and/or highly distorted data is presented. The new methodology combines different clustering methods with the GAMLSS (generalized additive models for location, scale, and shape) framework, and is therefore called c-GAMLSS, for “clustering GAMLSS.” In this new extended structure, a latent variable (cluster) is created to explain the response-variable (target). Any and all parameters of the distribution for the response variable can also be modeled by functions of the new covariate added to other available resources (features). The method of selecting resources to be used is carried out in stages, a step-based method. A simulation study considering multiple scenarios is presented to compare the c-GAMLSS method with existing Gaussian mixture models. We show by means of four different data applications that in cases where other authentic explanatory variables are or are not available, the c-GAMLSS structure outperforms mixture models, some recently developed complex distributions, cluster-weighted models, and a mixture-of-experts model. Even though we use simple distributions in our examples, other more sophisticated distributions can be used to explain the response variable.

Keywords: bimodal distributions; GAMLSS; mixture models; regression models; statistical learning



Citation: Ramires, T.G.; Nakamura, L.R.; Righetto, A.J.; Konrath, A.C.; Pereira, C.A.B. Incorporating Clustering Techniques into GAMLSS. *Stats* **2021**, *4*, 916–930. <https://doi.org/10.3390/stats4040053>

Academic Editors: Marta Nai Ruscone and Daniel Fernández

Received: 30 September 2021
Accepted: 11 November 2021
Published: 12 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Multiple regression models in which the response variables are generated by distributions that belong to a family of parametric probability distributions have been widely used. The goal is to explain the behavior of a response variable, denoted here as Y . The increasing ease with which data can be collected and stored allows for quick construction of databases of ever-increasing sizes. With large volumes of data available for study, patterns of greater complexity are being observed, forcing the search for more flexible probabilistic (regression) models to deal with such patterns. Examples of such complexities include asymmetry around central values, the presence of a high excess kurtosis, multimodality, and other aspects of separate subgroups with little variability. These anomalies, common in a large database, can be caused by the effect of latent variables or latent categories, which are variables or categories that are not observed or collected.

For example, consider the weight distribution of a certain animal species for which the sexes of the sampled units are not recorded. A possible difference between the average weights of males and females may lead to a bimodal distribution of frequencies. Other aspects not noted in the database may also be responsible for additional anomalies. Another common issue in establishing species profiles occurs when the variabilities of the two subgroups determine the corresponding profiles. If latent characteristics are not considered, marginal distributions of the response variables can produce anomalies such as those illustrated in Figure 1—(a) bimodality with little variability between subgroups; (b) high

asymmetry on the right with little difference between location parameters. This high asymmetry can be explained by the variability differences among subgroups.

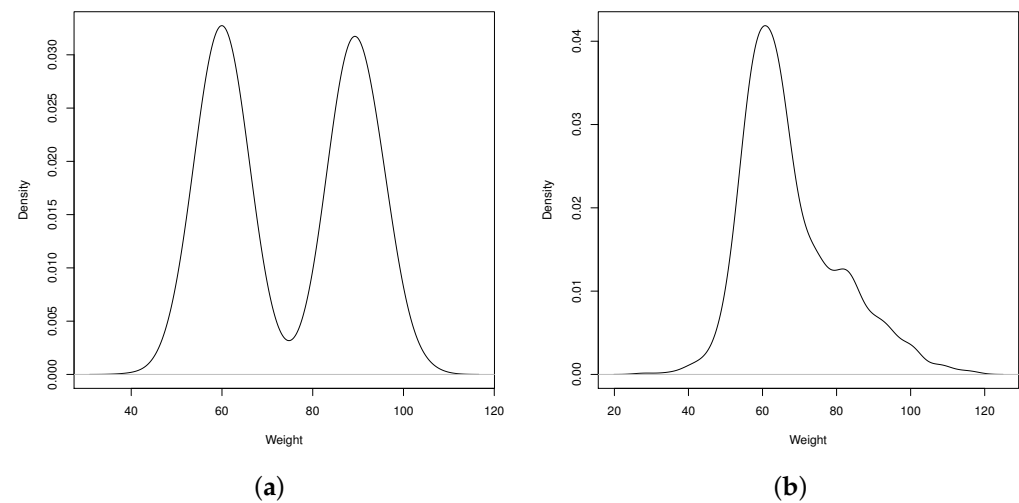


Figure 1. Densities generated from two different Gaussian distributions: (a) $N(60, 5)$ and $N(90, 5)$; and (b) $N(60, 5)$ and $N(75, 15)$.

Several methods of handling these kinds of behavior are described in the literature. It is common for analysts not to have access to features that may be causing these distortions in the target variable. Usually, it is assumed that the response variables have bimodal and/or skewed distributions. New overly complex probabilistic distributions are being developed to deal with these problems, such as in [1–3]. When one uses these new distributions to explain such distortions, one is only considering the marginal distribution of the target variable, and not the possibility that they might be due to one or more additional features.

Mixture distributions, such as in [4], are a promising alternative approach. The simplest way to use this kind of model in problems such as those in Figure 1 is to consider the observations of the response variable to have been generated by two distinct distributions that although they belong to the same family of distributions, have different values between their respective parameters. Mathematically, this can be written as

$$f(y; \xi) = p \omega_1(y, \xi_1) + (1 - p) \omega_2(y, \xi_2), \quad (1)$$

where p represents the proportion of observations generated by distribution ω_1 , $(1 - p)$ is from ω_2 , and ξ_1 and ξ_2 are parameter vectors. For instance, if both generators are considered to follow Gaussian distributions, then we have $\omega_1 \sim N(\mu_1, \sigma_1)$ and $\omega_2 \sim N(\mu_2, \sigma_2)$. Some other well-known options for dealing with such behavior available in the literature are the class of mixture-of-expert models (MoE) [5] and cluster-weighted models (cwm) [6].

The present work provides another alternative approach in which clustering techniques are used early in the modeling process. The use of these tools has the following objectives: (i) identify different clusters in the data set from a possible latent explanatory variable (e.g., in a bimodal data set, two different clusters would be created, producing a new dummy variable, while a trimodal data set would lead to three clusters, generating a factor with three levels, and so on); and (ii) include this new covariate in a regression model to explain the specific behavior observed in the response, as in [7]. In our approach, we use the GAMLSS framework, short for “generalized additive models for location, scale and shape” [8] with simple distributions. As stated in [9] the use of simple distributions with highly sophisticated regression-type models may return reliable and easily interpreted results.

The paper is organized as follows. In Section 2, we briefly describe some clustering methods that could be used in the proposed methodology. In Section 3, we incorporate clusters as latent variables into the GAMLSS framework. In Section 4, we perform some simulation studies to validate the proposed modeling process. In Section 5, four different applications are discussed. In the first two, no extra covariates are available, and we compare the proposed methodology to both mixture models and sophisticated distributions. In the others, authentic explanatory variables are also considered, and we compare our approach with cluster-weighted regression models and the mixture-of-experts model. Finally, Section 6 contains concluding remarks.

2. Clustering Methods

In this section, we present a brief summary of four well-established clustering methods that will be considered to identify different groups in both Sections 4 and 5. Nevertheless, it is worth noting that any clustering method may be applied in the proposed framework in this paper (e.g., the ones presented at <https://cran.r-project.org/web/views/Cluster.html>, accessed on 1 July 2021), as we show in the last two applications in Sections 5.3 and 5.4.

2.1. k-Means Clustering

The k-means clustering method is the most well-known and popular grouping method in the literature. The basic idea of the k-means method is to minimize intra-cluster variation. There are several algorithms available to minimize that variation, of which the best-known were pioneered by [10,11]. Here we focus only on the algorithm available in [11], which defines the total intra-cluster variation as the sum of squared Euclidean distances between items and the corresponding centroid $\sum_{y_i \in C_k} (y_i - \mu_k)^2$, where y_i is a point that belongs to cluster C_k and μ_k is the mean value of the points assigned to the cluster C_k . The main idea of this method is to minimize $J(C) = \sum_{k=1}^K \sum_{y_i \in C_k} (y_i - \mu_k)^2$, the sum of the squared distances between points and their respective clusters' centroids over all K clusters. This is known to be an NP-hard problem [12].

2.2. Ward's Hierarchical Clustering

Ward's hierarchical clustering method is the only one among the agglomerative methods that is based on a sum-of-squares method and produces groups that minimize intra-group dispersion at each binary fusion [13]. The function to be minimized is given by

$$D(C_1, C_2) = \frac{|C_1||C_2|}{|C_1| + |C_2|} \|C_1 - C_2\|^2.$$

As can be seen in [13], Ward's method shares the sum-of-squares minimization criterion with k-means partitioning. A full flowchart regarding the hierarchical grouping procedure is available in the original paper by [14].

2.3. Fuzzy Clustering

Fuzzy clustering allows a given observation (or an individual) to belong to more than one cluster, generalizing partitioning methods such as k-means [15]. Here, the objective function to be minimized is

$$\sum_{k=1}^K \frac{\sum_{i,j=1}^n u_{ik}^2 u_{jk}^2 d(i,j)}{2 \sum_{j=1}^n u_{jk}^2},$$

where u_{ik} represents membership of object i in cluster k , $d(i, j)$ are the dissimilarities and n is the number of observations. The considered objective function may vary in this method,

however, as mentioned above, any clustering method (and objective function) can be used in the proposed methodology.

2.4. Model-Based Clustering

Model-based clustering is based on finite Gaussian mixture modeling [16] and may be achieved using an EM (expectation-minimization) algorithm as described in [17]:

1. A maximum number K of clusters is defined and a mixture model is considered;
2. A hierarchical agglomeration to approximately maximize the classification likelihood for each model is carried out, followed by obtaining classifications for up to K groups;
3. The EM algorithm is performed for each model and each number of clusters;
4. A goodness-of-fit measure (such as Bayesian information criterion) is computed for the one-cluster case for each model and for the mixture model based on the estimates obtained in Step 3.

3. Incorporating Clustering into GAMLSS

As highlighted in Section 1, many different sophisticated statistical distributions have been proposed to deal with some specific patterns presented in the marginal distributions of the response Y . In this paper, we use simpler distributions, modeling not only the mean (location) parameter μ in terms of the latent variable (clusters) and any other known features (explanatory variables) through the GAMLSS (generalized additive models for location, scale, and shape) framework [8], a generalization of both well-established generalized linear models (GLM) [18] and generalized additive models (GAM) [19].

Generically, let $Y \sim \mathcal{D}(\theta)$, i.e., Y follows any distribution (that does not necessarily belong to the exponential family) with parameter vector θ . For an extensive list of distributions used in GAMLSS, see [20]. Clustering GAMLSS (c-GAMLSS), which considers different clusters obtained through one of the methods described in Section 2, can then be described as follows:

$$g_r(\theta_r) = X_r \beta_r + \sum_{j=1}^{J_r} s_{jr}(x_{jr}) + \sum_{k=2}^K C_k \psi_{kr}, \quad r = 1, \dots, R, \quad (2)$$

where $g_r(\cdot)$ denotes an appropriate link function for parameter θ_r , usually determined by the range of the parameter, m_r denotes the number of explanatory variables related to the r th parameter, X_r is a known $n \times (m_r + 1)$ model matrix, $\beta_r = (\beta_{0r}, \beta_{1r}, \dots, \beta_{m_r r})^\top$ is a parameter vector of length $(m_r + 1)$, $s_{jr}(\cdot)$ are smooth functions of x_{jr} (e.g., p-splines [21]), C_k denotes latent variables (clusters) and ψ_{kr} is a parameter vector of length $(K - 1)$. Note here that any and all of the parameters of the response distribution may be modeled as functions of a given set of explanatory and latent variables. As in the classical GAMLSS framework, it is possible to consider interactions between covariates and varying coefficient terms (as introduced in [22]), where the varying coefficient function fits separate smooth curves against X for each cluster [23]. Finally, model (2) can be seen as a flexible expert-network mixture of experts [24]. However, in this approach, the mixing proportion does not depend on the covariates.

Due to the high flexibility of GAMLSS, there are multiple techniques that may be used to select and fit a suitable model for the response variable considering the available covariates. They are extensively described in [23]. In the c-GAMLSS framework, we can use the following steps in the model-selection process:

- Step 1: Select a clustering method (e.g., among the ones presented in Section 2) to create different groups (estimating the number of clusters can be achieved using one of multiple available strategies. See, for instance, [25]).
- Step 2: Select subsets of authentic and latent variables for each of the parameters of the response-variable distribution, using any of the applied strategies for GAMLSS models [23]. The most common procedure is a stepwise procedure called "Strategy A" [23,26,27]).

Strategy A uses a goodness-of-fit measure to select the most suitable model to describe the data being studied. One may say that the Bayesian information criterion (BIC) [28] would perform better than the Akaike information criterion (AIC) [29] in the c-GAMLSS framework [30,31]. However, since we may consider smoothing functions within model (2), BIC can lead to underfitting (i.e., oversmoothing) [32]. A richer discussion regarding this topic can be seen in [27]. Furthermore, as in MoE, an important issue to be addressed in the future is whether a given covariate may be (or not) considered in either or both of the above-mentioned steps.

If only latent variables (clusters) are used to explain the behavior of a given response, then model (2) reduces to

$$g_r(\theta_r) = \psi_{1r} + \sum_{k=2}^K C_k \psi_{kr}, \quad r = 1, \dots, R, \quad (3)$$

which can be seen as a mixture model [24] with no covariates affecting the mixing proportion. It is worth mentioning that the new coefficient ψ_{1r} in (3) is the intercept associated with each of the regression structures related to the first cluster obtained from any of the methods described in Section 2. In model (2), the intercept is the first element of each vector of parameters β_r .

As a straightforward example, letting Y have a Gaussian distribution, with vector of parameters $\theta_r = (\theta_1, \theta_2)^\top = (\mu, \sigma)^\top$ in (3), results in the following:

$$\begin{aligned} g_1(\theta_1) &= \mu = \psi_{11} + \sum_{k=2}^K C_k \psi_{k1} \\ g_2(\theta_2) &= \log(\sigma) = \psi_{12} + \sum_{k=2}^K C_k \psi_{k2}. \end{aligned} \quad (4)$$

Note that we choose appropriate link functions for both parameters (identity and logarithm functions for μ and σ , respectively). See [23] for more information.

Regarding the inference processes, all model parameters can be estimated by the penalized maximum-likelihood method through the Rigby and Stasinopoulos (RS) and/or Cole and Green (CG) algorithms described in [8,33]. As stated in [23], both algorithms are stable and fast using simple starting values, such as constants.

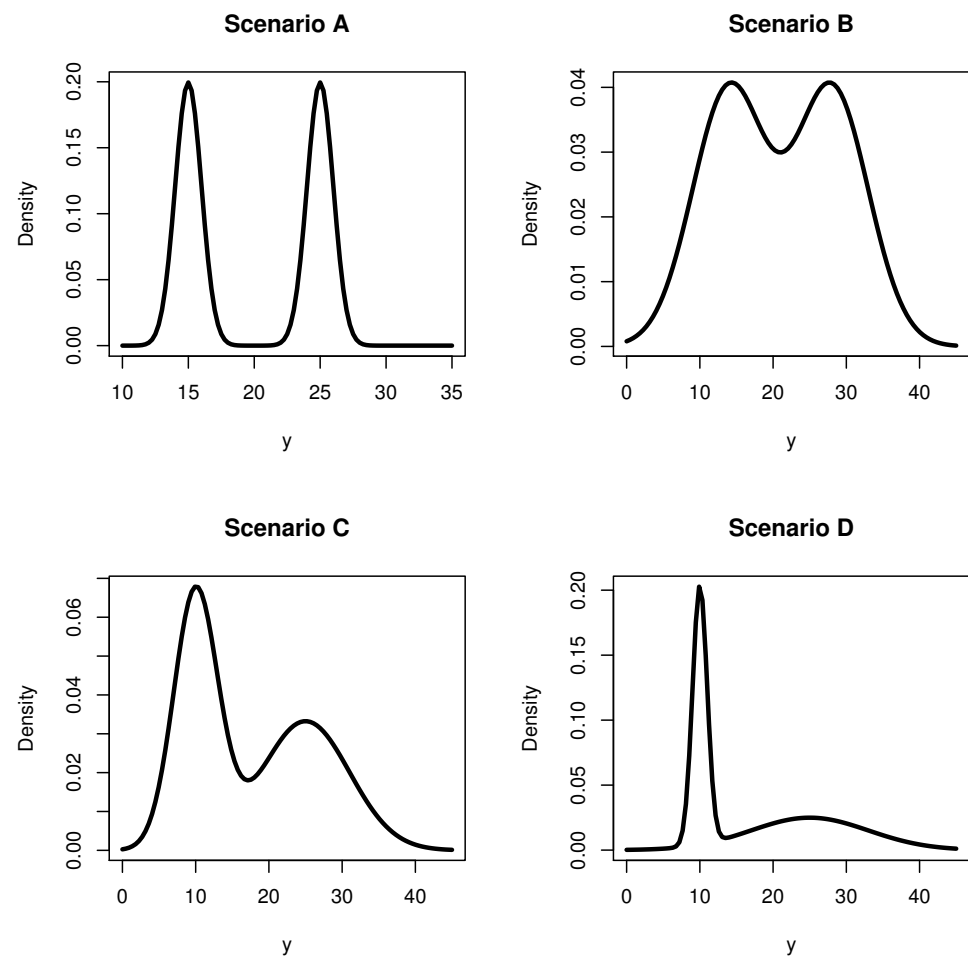
Implementation of fitting a c-GAMLSS model (including or not extra covariates and considering different response-variable distributions) is easily achieved using a new generic function called `gamlss.cluster()`, based on the `gamlss` package [33] for the R software environment [34] and available at <https://git.io/JtOZW> (accessed on 1 July 2021).

4. Simulation

In this section, we conduct some Monte Carlo simulation studies to understand the behavior of the c-GAMLSS framework based on the Gaussian distribution and compare it to the usual Gaussian mixture model approach, considering four different scenarios (marginal response-variable shapes) where each is composed of two different clusters. All observations were generated in the R software environment via the `rnorm()` function. The averages (μ) and standard deviations (σ) for each cluster and scenario are reported in Table 1 and the resulting empirical densities (considering both scenarios together) are displayed in Figure 2.

Table 1. Means and standard deviations of each cluster (from a Gaussian distribution) considered in four different scenarios.

Scenario	Cluster 1		Cluster 2	
	μ_1	σ_1	μ_2	σ_2
A	15	1	25	1
B	14	5	28	5
C	10	3	25	6
D	10	1	25	8

**Figure 2.** Shape of the empirical marginal density in each scenario.

For each scenario, we evaluate the maximum-likelihood estimates (MLEs) of the parameters for both approaches and then, after all replications, we compute the biases and mean squared errors (MSEs) based on the average estimates. In order to obtain the MLEs for the mixture models, we are using the `normalmixEM()` function available in the `mixtools` package for the R software environment, which returns the best fitted model after five attempts. All results here are obtained from 1000 Monte Carlo replications. Here we are considering two different sample sizes for each cluster, 50 and 300, totaling $n = 100$ and $n = 600$ observations for each data/replication. Results are displayed in Table 2.

Table 2. The biases and MSEs for the c-GAMLSS and mixture models based on 1000 simulations using different sample sizes.

Scenario	Parameter	<i>n</i> = 50 in Each Cluster				<i>n</i> = 300 in Each Cluster			
		c-GAMLSS		Mixture		c-GAMLSS		Mixture	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
A	μ_1	0.001	0.020	0.016	0.091	0.000	0.003	0.025	0.124
	σ_1	−0.038	0.009	−0.061	0.060	−0.010	0.001	−0.004	0.084
	μ_2	−0.009	0.019	−0.023	0.095	−0.003	0.003	−0.028	0.126
	σ_2	0.006	0.008	−0.061	0.060	0.006	0.001	−0.004	0.084
	<i>p</i>	0.000	0.000	0.001	0.000	0.000	0.000	0.002	0.001
B	μ_1	−0.401	1.650	0.330	5.931	−0.355	0.283	0.095	1.317
	σ_1	−0.747	0.858	−1.198	2.848	−0.580	0.357	−0.300	0.417
	μ_2	0.427	1.724	−0.193	5.638	0.377	0.291	−0.132	1.331
	σ_2	−0.529	0.510	−1.196	2.845	−0.556	0.329	−0.300	0.417
	<i>p</i>	0.046	0.005	0.127	0.031	0.015	0.000	0.046	0.005
C	μ_1	0.197	0.585	0.178	1.396	0.003	0.048	−0.005	0.107
	σ_1	0.022	0.398	−0.114	0.327	−0.120	0.037	−0.017	0.034
	μ_2	0.822	2.123	−0.016	2.074	0.525	0.459	−0.043	0.324
	σ_2	−0.860	1.508	−3.110	10.007	−0.583	0.415	−3.015	9.127
	<i>p</i>	0.044	0.004	0.048	0.007	0.019	0.000	0.016	0.001
D	μ_1	0.007	0.023	0.052	0.401	0.006	0.004	0.007	0.058
	σ_1	0.005	0.018	−0.005	0.022	0.009	0.003	−0.003	0.003
	μ_2	0.730	1.908	0.165	1.749	0.659	0.674	0.023	0.310
	σ_2	−0.622	1.129	−6.997	49.048	−0.491	0.378	−6.996	48.992
	<i>p</i>	0.024	0.001	0.021	0.002	0.021	0.001	0.007	0.000

Note: Based on model (4), in c-GAMLSS $\mu_1 = \psi_{11}$, $\mu_2 = \psi_{11} + \psi_{21}$, $\sigma_1 = \psi_{12}$ and $\sigma_2 = \psi_{12} + \psi_{22}$, where ψ_{1r} and ψ_{2r} , $r = 1, 2$, are coefficients associated with the first and second clusters, respectively.

Please note that we have five different parameters in Table 2: μ_1 , σ_1 , μ_2 , σ_2 and p . Regarding the mixture model, all parameters were already introduced in Equation (1) and refer to the mean and standard deviation of the first cluster, mean and standard deviation of the second cluster and the proportion of observations that will be modeled by the first Gaussian distribution, respectively. However, we note here that we temporarily reparameterize the c-GAMLSS model in this subsection to compare the performance of the two approaches, since they will have the same interpretation. This reparameterization, based on model (4), is given as follows: $\mu_1 = \psi_{11}$, $\mu_2 = \psi_{11} + \psi_{21}$, $\sigma_1 = \psi_{12}$, $\sigma_2 = \psi_{12} + \psi_{22}$ and p is the proportion of observations in the first cluster obtained through the best clustering method available in Table 1 for each of the scenarios (Scenario A: k-means; Scenario B: Fuzzy; Scenario C: model-based; and Scenario D: model-based).

In Scenario A we may note that the c-GAMLSS method clearly performs better than the Gaussian mixture models for both sample sizes considered (nevertheless, as expected, biases and MSEs decrease for both methods in the greater sample). For Scenario B ($n = 50$ in each cluster), we may note that the absolute biases for the standard deviation (σ_1 and σ_2) estimates are greater for the c-GAMLSS method; however, the MSEs are drastically smaller, indicating a better accuracy. Considering $n = 300$ in each cluster, all results are clearly favorable to the new proposed alternative. In Scenario C, the MSEs obtained for σ_2 by the mixture model are 6.67 and 21.99 times greater than the one returned by the c-GAMLSS method for $n = 50$ and $n = 300$ in each cluster, respectively. Finally, for Scenario D, the MSE returned by the Gaussian mixture when $n = 300$ in each cluster for the standard deviation σ_2 was almost 130 times greater than the one returned by c-GAMLSS. We can conclude that our proposed methodology clearly outperformed the already well-known Gaussian mixture model in all simulated scenarios.

5. Applications

In this section, we provide two applications comparing the adequacy of the c-GAMLSS framework, Gaussian mixture models, and some sophisticated bimodal distributions for

marginally modeling the response variable (i.e., not considering any further explanatory variables), and another two applications where extra authentic covariates exist, comparing c-GAMLSS to the mixture-of-experts models (MoE; Gaussian parsimonious clustering models with covariates) and cluster-weighted models (cwm). In order to compare these approaches, both AIC and BIC statistics are calculated.

The flexible recently proposed four-parameter distributions considered in the first two applications, where only the target variable is available, are: the odd log-logistic exponentiated Gumbel (OLLEGu) [1], extended generalized odd half-Cauchy log-logistic (EGOHC-LL) [2], and exponentiated log-sinh Cauchy (ELSC) [3] distributions. All parameter roles and their respective ranges from each of these distributions are described in Table 3.

Table 3. Parameter ranges and roles in the four-parameter distributions considered.

	OLLEGu		EGOHC		ELSC	
	Range	Role	Range	Role	Range	Role
μ	\mathbb{R}	Location	\mathbb{R}^+	Shape	\mathbb{R}	Location
σ	\mathbb{R}^+	Scale	\mathbb{R}^+	Scale	\mathbb{R}^+	Scale
ν	\mathbb{R}^+	Shape	\mathbb{R}^+	Shape	\mathbb{R}^+	Skewness
τ	\mathbb{R}^+	Shape	\mathbb{R}^+	Shape	\mathbb{R}^+	Shape

5.1. Actuarial Sciences Data

In the first application, we present a data set already studied in [1], where the authors compare their new model with some of its sub-models and other alternative distributions. The data consist of the lifespans (in years) of 280 retired women covered by the Mexican public health-insurance system who had temporary disabilities and died during 2004. The data are more fully reported in [35].

Table 4 contains the MLEs, their respective SEs, and the AIC and BIC statistics for all fitted models. The c-GAMLSS framework based on the Gaussian distribution clearly performs better than all other approaches, yielding the lowest AIC and BIC values (1788.35 and 1802.90, respectively). Please note that in this particular case, the latent variable (clusters) was considered only for the parameter μ , i.e., both clusters present the same dispersion (σ), and hence no estimates for ψ_{22} are displayed. This was achieved through the stepwise method (Strategy A), which also selected the k-means clustering method as the most appropriate to divide these data.

Table 4. MLEs of all model parameters on actuarial sciences data, their corresponding SEs (in parentheses), and the AIC and BIC values.

Model	Estimates				AIC	BIC
c-GAMLSS ($\psi_{11}, \psi_{21}, \psi_{12}$)	37.967 (0.526)	17.400 (0.701)	1.760 (0.042)		1788.35	1802.90
EGOHC-LL (μ, σ, ν, τ)	44.837 (1.454)	13.375 (1.210)	0.511 (0.095)	8.688 (2.935)	2091.69	2106.23
OLLEGu (μ, σ, ν, τ)	62.140 (4.002)	13.372 (0.213)	14.716 (0.304)	0.380 (0.328)	2106.3	2120.9
ELSC (μ, σ, ν, τ)	3.828 (0.037)	0.098 (0.084)	0.355 (0.184)	1.014 (0.187)	2113.80	2128.34
Gaussian mixture ($\mu_1, \sigma_1, \mu_2, \sigma_2, p$)	35.429 (1.688)	5.440 (0.856)	51.554 (1.601)	8.471 (0.936)	0.234 (0.094)	2116.10 2134.30

The Gaussian mixture model did not perform well when compared to the other approaches (Table 4), returning the highest AIC and BIC values among all models (2116.10 and 2134.30, respectively), presenting a roughly unimodal curve in Figure 3a. Moreover,

the c-GAMLSS method was able to precisely identify two different clusters (red and blue curves).

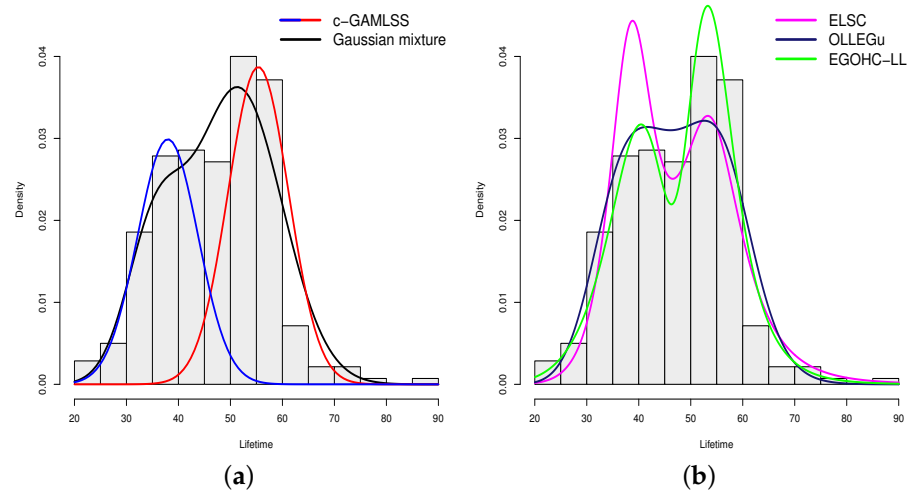


Figure 3. The actuarial sciences data and all fitted models: (a) c-GAMLSS framework and mixture model (b) ELSC, OLLEGu and EGOHC-LL distributions. Please note that the blue and red colors on the c-GAMLSS approach indicate the two different selected clusters.

Although the OLLEGu distribution was elected as the best one to describe these data according to [1] when compared to some other models, we can see that the EGOHC-LL distribution, proposed two years earlier by [2], returned better AIC and BIC statistics (2091.69 and 2106.23, respectively). Their fitted densities are displayed in Figure 3b.

5.2. Ozone Data

The second application corresponds to daily ozone-level measurements (in ppb = ppm × 1000) collected in New York during 1973 and is available in [36]. As with the previous application, we provide in Table 5 the MLEs, their corresponding SEs and AIC and BIC statistics. Once again, the proposed methodology is the best alternative to explain the behavior of the data since it returned the lowest AIC and BIC values (936.18 and 947.78, respectively) and the Gaussian mixture approach performed poorly. In this application, the c-GAMLSS framework based on the Gaussian distribution uses the model-based clustering selected via Strategy A, where the generated latent variable was included in both regression structures (μ and σ).

Table 5. MLEs of the model parameters for the ozone data, the corresponding SEs (given in parentheses), and the AIC and BIC statistics.

Model	Estimates				AIC	BIC
c-GAMLSS ($\psi_{11}, \psi_{21}, \psi_{12}, \psi_{22}$)	20.776 (1.221)	55.291 (4.261)	2.302 (0.086)	1.008 (0.136)	936.18	947.78
OLLEGu (μ, σ, ν, τ)	29.628 (0.278)	10.071 (3.598)	1.096 (5.259)	0.334 (0.472)	1054.41	1065.28
EGOHC-LL (μ, σ, ν, τ)	22.995 (5.259)	2.647 (0.472)	0.830 (0.278)	6.572 (3.598)	1062.50	1073.40
Gaussian mixture ($\mu_1, \sigma_1, \mu_2, \sigma_2, p$)	21.410 (2.297)	10.910 (1.743)	69.950 (7.895)	31.457 (3.818)	0.555 (0.090)	1063.93 1077.52
ELSC (μ, σ, ν, τ)	3.790 (0.412)	0.294 (0.111)	0.364 (0.216)	0.942 (0.461)	1095.32	1106.19

These data were also analyzed in [2], where the EGOHC distribution was the model selected when compared to some of its sub-models and other distributions. Nevertheless, we can see that the OLLEGu model returns better AIC and BIC values (1054.41 and 1065.28, respectively). Finally, all fitted densities are displayed in Figure 4.

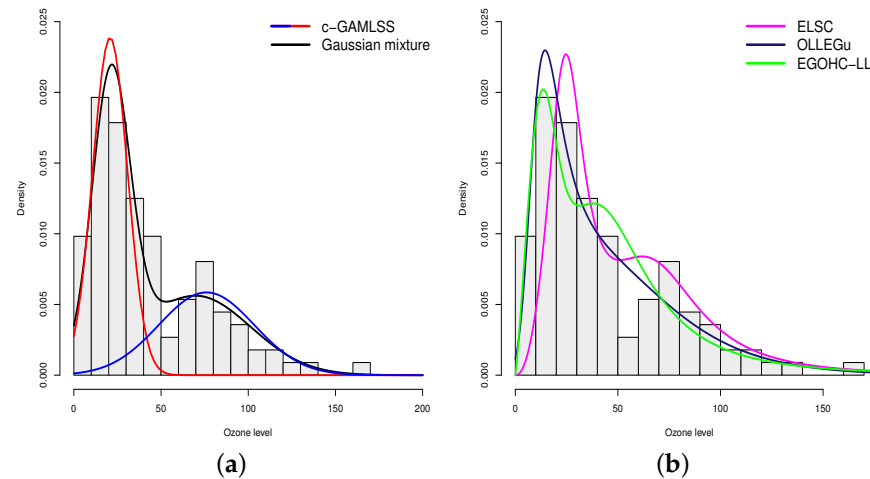


Figure 4. The ozone data and all fitted models: (a) c-GAMLSS framework and mixture model (b) ELSC, OLLEGu, and EGOHC-LL distributions. Note that the blue and red colors on the c-GAMLSS approach indicate the two different selected clusters.

5.3. Two-Moon

The two-moon is a well-known synthetic data set introduced by [37] which can be obtained in the `clusterSim` package in the R software environment, and is mainly used in studies regarding clustering methods. The generated data consist of two variables: the response Y and an authentic covariate X , both defined on the real numbers. The relationship between Y and X , as well as the two clusters (in red and black colors) are shown in Figure 5. We may note a clear nonlinear relationship between response and the explanatory variable, and hence clustering algorithms that consider linear separations may not be suitable for this example. Here we are considering the cluster classification already available in the data set [37]. Furthermore, this nonlinear relationship indicates that the use of smoothing functions in the regression structure would be appropriate, and the shape of such relationship may depend on each cluster, so considering an interaction between each cluster and the explanatory variable X may also be necessary.

Now we compare c-GAMLSS based on the Gaussian distribution, considering the presence of an authentic explanatory variable, against three different models: (i) fitting a regression structure only for the location parameter, with a constant variance (reducing to a c-GLM); (ii) incorporating smoothing functions in this regression structure (reducing to a c-GAM); and (iii) considering a mixture-of-experts model (Gaussian parsimonious clustering models with covariates, obtained using the `MoEClust` package for the R software environment). We shall highlight here that due to the behavior observed in Figure 5, in the c-GLM structure, we are considering a quadratic term and interactions between covariate X and the clusters, whereas in both c-GAM and c-GAMLSS, we fit a varying coefficient term (pvc).

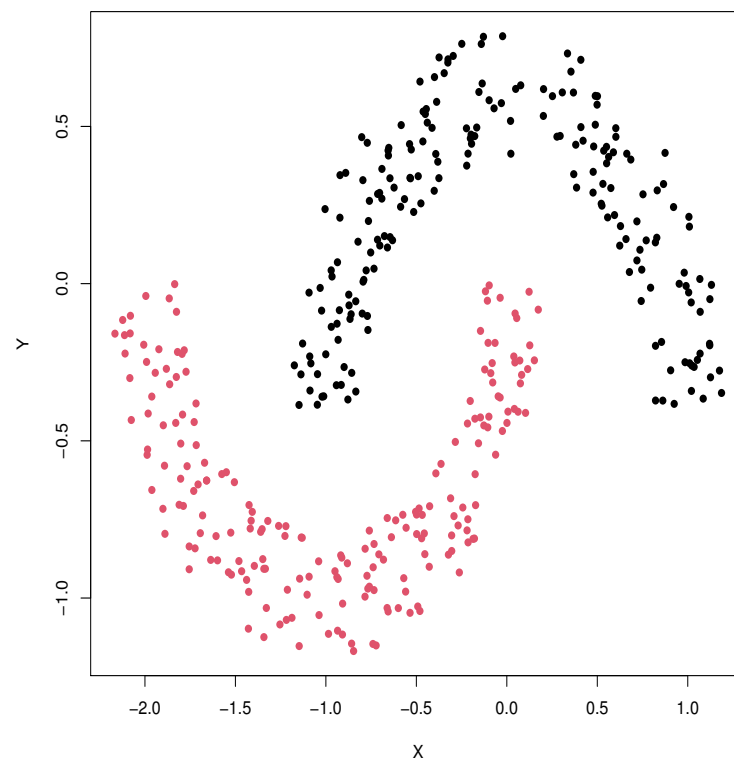


Figure 5. Scatter plot of Y against X for the moon data.

Table 6 displays all regression structures for each fitted model. c-GAMLSS presented the lowest AIC value (−279.4), whereas the c-GLM model returned the lowest BIC value (−202.7). We emphasize here that the BIC criterion highly penalizes models that consider smoothing functions to explain a given response-variable parameter. However, as can be seen in Figure 6, such functions are quite important to explain the nonlinear effect of X on each of the response-variable parameters (mean μ and standard deviation σ).

Table 6. Regression structures for all fitted models applied to the two-moon data and their corresponding AIC and BIC statistics.

Model	Regression Structure	AIC	BIC
c-GAMLSS	$\mu = \beta_{01} + pvc(x, by = C)$ $\sigma = \exp[\beta_{02} + pvc(x, by = C)]$	−279.4	−159.3
c-GAM	$\mu = \beta_{01} + pvc(x, by = C)$ $\sigma = \exp(\beta_{02})$	−248.1	−191.5
c-GLM	$\mu = \beta_{01} + \beta_{11}x + \beta_{21}x^2 + \psi_{21}C + \psi_{31}xC + \psi_{41}x^2C$ $\sigma = \exp(\beta_{02})$	−230.7	−202.7
MoE	$\mu_1 = \beta_{01} + \beta_{11}x + \beta_{21}x^2$ $\mu_2 = \beta_{02} + \beta_{12}x + \beta_{22}x^2$ $p = \text{logistic}(\beta_{03})$ $\sigma = \exp(\beta_{04})$	−186.3	−154.38

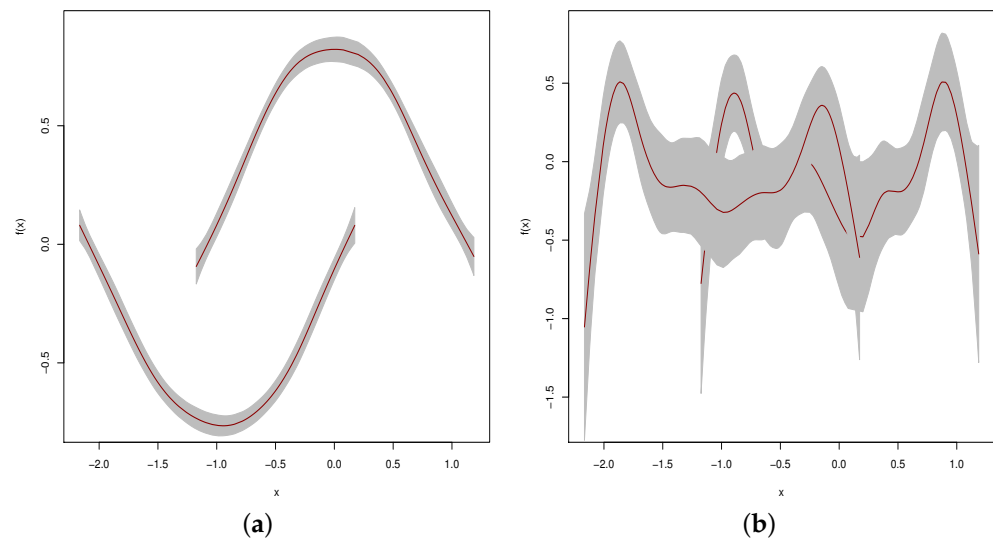


Figure 6. For “two-moon” data and c-GAMLSS, partial effects of the interaction between X and clusters on (a) μ and (b) σ .

5.4. Eruption Data

In this last application we revisit the well-known Old Faithful Geyser data from Yellowstone National Park in Wyoming, USA [38]. This data set contains 299 pairs of measurements regarding the times between the beginnings of successive eruptions, continuously collected over the first 15 days of August of 1985, and can be obtained in the MASS package for the R software environment. The explanatory variable X , which represents eruption duration, can also be used here to explain the response variable.

Here, we apply the full model displayed in Equation (2) to these data, disregarding the smoothing functions, i.e., the eruption duration X is treated as a linear predictor of the c-GAMLSS model. Furthermore, in order to show the great flexibility of this approach, we now consider a more complex distribution (apart from the celebrated Gaussian) to explain the target variable. In the classical GAMLSS context, when $Y > 0$, one of the most-used distributions is the Box-Cox Cole and Green (BCCG), a three parameter distribution where $\mu > 0$ is the median, $\sigma > 0$ is the approximate coefficient of variation and $-\infty < \nu < \infty$ is the skewness parameter. For further details regarding the BCCG, see [20].

We compare the results obtained by the c-GAMLSS framework based on both the Gaussian and BCCG distributions with a cluster-weighted model (cwm) [6], the estimates of which were obtained using the `cwm` function in the `flexCWM` package. All model structures and their respective AIC and BIC statistics are presented in Table 7, where we see that c-GAMLSS based on the BCCG distribution provides a better fit than the other two alternatives since its AIC and BIC values are the smallest (1840.6 and 1870.2, respectively). A visual comparison of all models is provided in Figure 7 (the graphical representation of c-GAMLSS based on both Gaussian and BCCG distributions overlap). Finally, we also present in Figure 8 the term plot for each effect fitted, i.e., the effects of each explanatory variable on the model parameters.

Table 7. Regression structures for c-GAMLSS (based on the BCCG and Gaussian distributions) and cwm models applied to the eruption data and their corresponding AIC and BIC values.

Model	Regression Structure	AIC	BIC
c-GAMLSS (BCCG)	$\mu = \exp(\beta_{01} + \beta_{11}X + \psi_{21}C + \psi_{31}XC)$ $\sigma = \exp(\beta_{02} + \beta_{12}X)$ $\nu = \psi_{13} + \psi_{23}C$	1840.6	1870.2
c-GAMLSS (Gaussian)	$\mu = \beta_{01} + \beta_{11}X + \psi_{21}C + \psi_{31}XC$ $\sigma = \exp(\psi_{12} + \psi_{22}C)$	1970.4	1992.6
cwm	$\mu_1 = \beta_{01} + \beta_{11}duration$ $p = \text{logistic}(\beta_{03})$ $\mu_2 = \beta_{02} + \beta_{12}duration$ $\sigma = \exp(\beta_{04})$	2203.6	2229.5

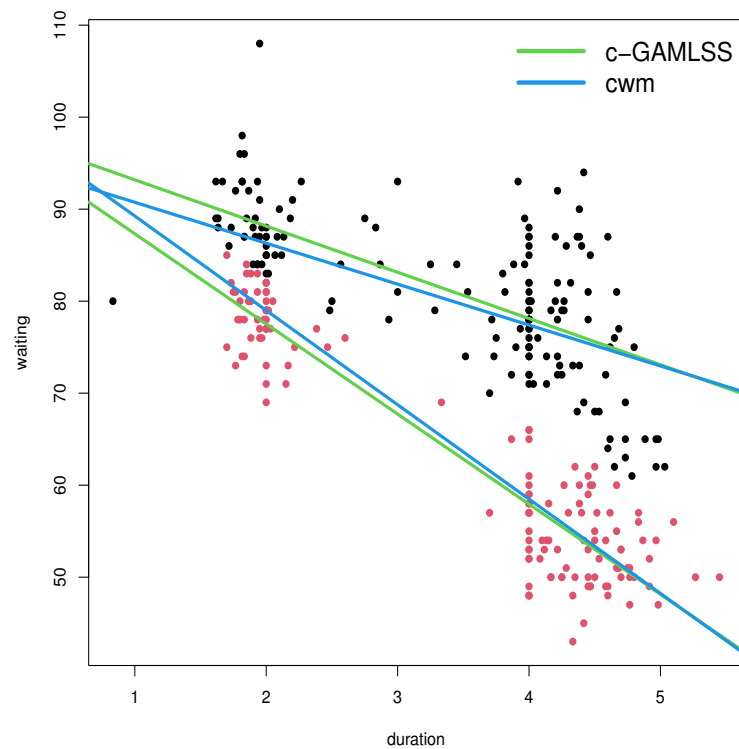


Figure 7. The eruption data and the fitted regression models based on the c-GAMLSS and cwm frameworks.

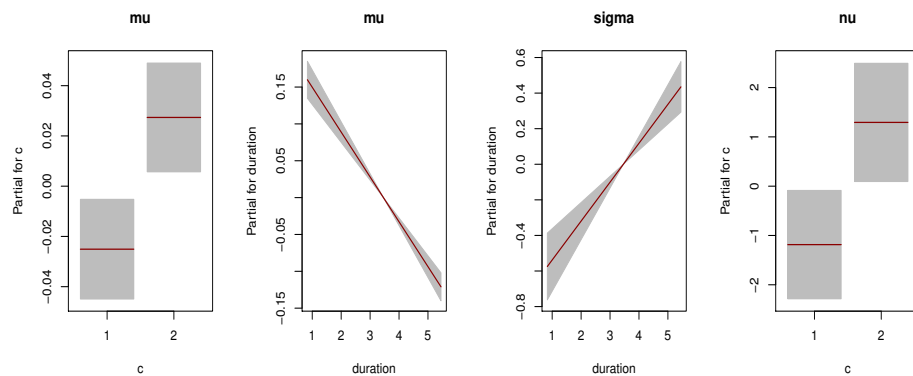


Figure 8. The fitted terms for the c-GAMLSS model based on the BCCG distribution on eruption data.

6. Conclusions

This article introduces c-GAMLSS—clustering generalized additive models for location, scale, and shape. It is an alternative model to deal with highly distorted data and even

though only bimodal responses were considered in this paper, this approach can also be applied in multimodal response problems. It is based on adding clustering techniques to the celebrated GAMLSS framework. It provides a new generic function implemented in R, a widely used statistical software. This new approach helps to make the fitting process more reliable, outperforming Gaussian mixture models, as illustrated in both simulation and real studies. Moreover, c-GAMLSS based on quite simple distributions returned better (smaller) AIC and BIC values than highly complex recently proposed distributions, cluster-weighted models and a mixture-of-experts model. Nevertheless, more sophisticated distributions may be used with c-GAMLSS for further applications.

Author Contributions: Conceptualization, T.G.R.; methodology, T.G.R., L.R.N. and A.J.R.; formal analysis, T.G.R., L.R.N., A.J.R., A.C.K. and C.A.B.P.; writing—original draft preparation, T.G.R. and L.R.N.; writing—review and editing, T.G.R., L.R.N., A.J.R., A.C.K. and C.A.B.P.; supervision, L.R.N.; project administration, T.G.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Actuarial sciences, ozone and moon data can be found in [35–37], respectively. The eruption data may be obtained in the MASS package for the R software environment and further information may be found in [38]. The new generic function for use in R is made available by the authors at <https://git.io/JtOZW> (accessed on 1 July 2021).

Acknowledgments: The first author would like to thank the multiuser lab (LAMAP-UTFPR). All authors gratefully acknowledge Mark Gannon for the extensive review.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alizadeh, M.; Ramires, T.G.; MirMostafae, S.M.T.K.; Samizadeh, M.; Ortega, E.M.M. A new useful four-parameter extension of the Gumbel distribution: Properties, regression model and applications using the GAMLSS framework. *Commun. Stat.-Simul. Comput.* **2019**, *48*, 1746–1767. [[CrossRef](#)]
2. Cordeiro, G.M.; Ramires, T.G.; Ortega, E.M.M.; Alizadeh, M. The new family of distributions and applications in heteroscedastic regression analysis. *J. Stat. Theory Appl.* **2017**, *16*, 401–418. [[CrossRef](#)]
3. Ramires, T.G.; Ortega, E.M.M.; Cordeiro, G.M.; Hens, N. A bimodal flexible distribution for lifetime data. *J. Stat. Comput. Simul.* **2016**, *86*, 2450–2470. [[CrossRef](#)]
4. McLachlan, G.; Peel, D. *Finite Mixture Models*; Wiley: Hoboken, NJ, USA, 2000.
5. Gormley, I.C.; Frühwirth-Schnatter, S. Mixture of experts models. In *Handbook of Mixture Analysis*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2019; pp. 271–307.
6. Subedi, S.; Punzo, A.; Ingrassia, S.; McNicholas, P.D. Cluster-weighted *t*-factor analyzers for robust model-based clustering and dimension reduction. *Stat. Methods Appl.* **2015**, *24*, 623–649. [[CrossRef](#)]
7. Candillier, L.; Tellier, I.; Torre, F.; Bousquet, O. Cascade evaluation of clustering algorithms. *Eur. Conf. Mach. Learn.* **2006**, *2006*, 574–581. [[CrossRef](#)]
8. Rigby, R.A.; Stasinopoulos, D.M. Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C Appl. Stat.* **2005**, *54*, 507–554. [[CrossRef](#)]
9. Ramires, T.G.; Nakamura, L.R.; Righetto, A.J.; Carvalho, R.J.; Vieira, L.A.; Pereira, C.A.B. Comparison between highly complex location models and GAMLSS. *Entropy* **2021**, *23*, 469. [[CrossRef](#)]
10. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1967; Volume 1, pp. 281–297. [[CrossRef](#)]
11. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28*, 100–108. [[CrossRef](#)]
12. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognition Lett.* **2010**, *31*, 651–666. [[CrossRef](#)]
13. Murtagh, F.; Legendre, P. Ward’s hierarchical agglomerative clustering method: Which algorithms implement Ward’s criterion? *J. Classif.* **2014**, *31*, 274–295. [[CrossRef](#)]
14. Ward, J.H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [[CrossRef](#)]
15. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2005.
16. Scrucca, L.; Fop, M.; Murphy, T.B.; Raftery, A.E. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **2016**, *8*, 289–317. [[CrossRef](#)] [[PubMed](#)]

17. Fraley, C.; Raftery, A.E. Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631. [[CrossRef](#)]
18. Nelder, J.A.; Wedderburn, R.W.M. Generalized linear models. *J. R. Stat. Soc. Ser. A (General)* **1972**, *135*, 370–384. [[CrossRef](#)]
19. Hastie, T.J.; Tibshirani, R.J. *Generalized Additive Models*; John Wiley & Sons: Hoboken, NJ, USA, 1990.
20. Rigby, R.A.; Stasinopoulos, D.M.; Heller, G.Z.; Bastiani, F.D. *Distributions for Modeling Location, Scale and Shape: Using GAMLSS in R*; CRC Press: Boca Raton, FL, USA, 2019.
21. Eilers, P.H.; Marx, B.D. Flexible smoothing with B-splines and penalties. *Stat. Sci.* **1996**, *11*, 89–102. [[CrossRef](#)]
22. Hastie, T.; Tibshirani, R. Varying-coefficient models. *J. R. Stat. Soc. Ser. B (Methodol.)* **1993**, *55*, 757–779. [[CrossRef](#)]
23. Stasinopoulos, D.M.; Rigby, R.A.; Heller, G.Z.; Voudouris, V.; Bastiani, F.D. *Flexible Regression and Smoothing: Using GAMLSS in R*; CRC Press: Boca Raton, FL, USA, 2017.
24. Murphy, K.; Murphy, T.B. Gaussian parsimonious clustering models with covariates and a noise component. *Adv. Data Anal. Classif.* **2020**, *14*, 293–325. [[CrossRef](#)]
25. Fu, W.; Perry, P.O. Estimating the number of clusters using cross-validation. *J. Comput. Graph. Stat.* **2020**, *29*, 162–173. [[CrossRef](#)]
26. Nakamura, L.R.; Rigby, R.A.; Stasinopoulos, D.M.; Leandro, R.A.; Villegas, C.; Pescim, R.R. Modelling location, scale and shape parameters of the Birnbaum-Saunders generalized *t* distribution. *J. Data Sci.* **2017**, *15*, 221–237. [[CrossRef](#)]
27. Ramires, T.G.; Nakamura, L.R.; Righetto, A.J.; Pescim, R.R.; Mazucheli, J.; Rigby, R.A.; Stasinopoulos, D.M. Validation of Stepwise-Based Procedure in GAMLSS. *J. Data Sci.* **2021**, *19*, 96–110. [[CrossRef](#)]
28. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [[CrossRef](#)]
29. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [[CrossRef](#)]
30. Keribin, C. Consistent estimation of the order of mixture models. *Sankhyā: Indian J. Stat. Ser. A* **2000**, *62*, 49–66.
31. Leroux, B.G. Consistent estimation of a mixing distribution. *Ann. Stat.* **1992**, *20*, 1350–1360. [[CrossRef](#)]
32. Hossain, A.; Rigby, R.; Stasinopoulos, M.; Enea, M. Centile estimation for a proportion response variable. *Stat. Med.* **2016**, *35*, 895–904. [[CrossRef](#)]
33. Stasinopoulos, D.M.; Rigby, R.A. Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* **2007**, *23*, 1–46. [[CrossRef](#)]
34. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
35. Balakrishnan, N.; Leiva, V.; Sanhueza, A.; Cabrera, E. Mixture inverse Gaussian distributions and its transformations, moments and applications. *Statistics* **2009**, *43*, 91–104. [[CrossRef](#)]
36. Nadarajah, S. A truncated inverted beta distribution with application to air pollution data. *Stoch. Environ. Res. Risk Assess.* **2008**, *22*, 285–289. [[CrossRef](#)]
37. Zhou, D.; Bousquet, O.; Lal, T.N.; Weston, J.; Schölkopf, B. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*; MIT Press: Vancouver and Whistler, BC, Canada 2003; pp. 321–328.
38. Azzalini, A.; Bowman, A.W. A look at some data on the old faithful geyser. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1990**, *39*, 357–365. [[CrossRef](#)]