International Journal of Statistics and Probability; Vol. 14, No. 1; March 2025
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Kumaraswamy Log Composite Distribution Based on the Median: Inference and Applications

Daniel Gálvez

Correspondence: Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile

Received: February 26, 2025 Accepted: March 28, 2025 Online Published: March 30, 2025

doi:10.5539/ijsp.v14n1p37 URL: https://doi.org/10.5539/ijsp.v14n1p37

Abstract

In this paper a new univariate distribution with bounded support on the Kumaraswamy distribution is defined. Some of its main properties are studied, and a Monte Carlo simulation is implemented to evaluate the behavior of the maximum likelihood estimators. Two applications to real data sets are presented. In particular goodness of fit and quantile regression are show in order to illustrate the potential and flexibility of this new distribution.

Keywords: Kumaraswamy distribution, Maximum likelihood estimates, Bounded support, Logarithmic distribution

1. Introduction

Distributions with bounded support are useful in many fields of real life, especially in the unit interval, which are essential for modeling percentages and rates. Due to this, new distributions in the unit interval have been developed in recent years to model various phenomena. Some of these works that have gained some popularity are the Unit logClog (Korkmaz and Korkmaz, 2023), Unit-improved second-degree Lindley (Altun and Cordeiro, 2020), Unit-Weibull (Mazucheli et al., 2018a), Unit-Chen (Korkmaz et al., 2022), Unit-inverse Gaussian (Ghitany et al., 2019), log-Bilal (Altun et al., 2021), UED (Bakouch et al., 2023), UHLG (Ramadan et al., 2022), among others. Many of these distributions arise from the necessity of consider more flexible distributions than the classics ones, such as the Beta, Generalized Beta (GB) (McDonald and Xu, 1995), Kumarawasmy (KM) (Kumaraswamy, 1980), Logistic-Normal (Aitchison and Shen, 1980), Johnson S_B (Johnson, 1949), unit-Gamma (Grassia, 1977), which generally do not have analytical forms for magnitudes of interest, as is the case of the Beta distribution whose cumulative distribution function (cdf) is a special function.

In the literature, there are many families of distributions, some of which include the T-X (Alzaatreh et al., 2013), Kw-G (Cordeiro and Castro, 2011), transmuted-G (Nofal et al., 2017), OBIII-G (Jamal et al., 2017), RL-G (Okutu et al., 2023), APT (Mahdavi and Kundu, 2017), ELG (Marinho et al., 2018), Tan-G (Souza et al., 2021), among others. In this article, we consider the complementary exponentiated G-logarithmic (CEGLn) class approach by Tahir and Cordeiro (2016), which has already been used by authors such as Hakamipour and Nadarajah (2022). In addition, new distributions have also been defined in this family, such as the EWL (Mahmoudi et al., 2014), GEL (Mahmoudi and Jafari, 2012) and CEL (Hakamipour and Nadarajah, 2022), although several of these authors use the EGLn or reparameterized versions of this distribution. Most of the distributions mentioned are defined in the positive interval; for this reason, a new distribution on the unit interval is proposed in this article, using recent methods proposed in the literature. Furthermore, it is shown that it presents a better fit than other distributions in the unit interval. A quantile regression model for the median is also implemented, based on the new distribution, which seems to be a good alternative to other models.

The rest of the paper proceeds as follows. Section 2 defines the proposed distribution and studies some of its mathematical properties. Section 3 performs a simulation study via Monte Carlo to evaluate the parameter estimates using maximum likelihood method. Section 4 presents some univariate applications of the new distribution. Section 5 presents the quantile regression model and an application. Finally, Section 6 presents a conclusion on everything shown in this paper.

2. The Proposed Distribution

The probability density function (pdf) of the Kumaraswamy distribution is given by

$$f_{KM}(x; a, b) = abx^{a-1}(1 - x^a)^{b-1}, \quad 0 < x < 1,$$

where $a > 0$ and $b > 0$ with cdf given by

$$F_{KM}(x; a, b) = 1 - (1 - x^a)^b, \quad 0 < x < 1. \tag{1}$$

Posteriorly, the cdf of the CEGLn class is defined by

$$F_{CEGLn}(x;\phi,\alpha,\boldsymbol{\tau}) = \frac{1}{-\log(\phi)}\sum_{n=1}^{\infty}\frac{\left[(1-\phi)G(x;\boldsymbol{\tau})^{\alpha}\right]^{n}}{n} = \frac{\log\left[1-(1-\phi)G(x;\boldsymbol{\tau})^{\alpha}\right]}{\log(\phi)}, \tag{2}$$

with $\phi \in (0,1)$, $\alpha > 0$ and $\boldsymbol{\tau}$ is the parameter vector. Recently, Attia (2024) defines a new distribution called Median Based Unit Rayleigh (MBUR) from the equation (3), by considering the cdf of the median of three iid random variables, which is given by

$$G(x;\boldsymbol{\tau}) = 3F(x;\boldsymbol{\tau})^2 - 2F(x;\boldsymbol{\tau})^3, \quad x \in \mathbb{R}. \tag{3}$$

Another approach to obtain this result involves considering the cdf of the class of Beta-G distributions, as defined by Eugene et al. (2002), which is given by

$$G(x;\boldsymbol{\gamma},\boldsymbol{\tau}) = \int_0^{F(x;\tau)}\frac{t^{\gamma_1-1}(1-t)^{\gamma_2-1}}{B(\gamma_1,\gamma_2)}\,dt,$$

where $B(\gamma_1,\gamma_2)$ is the Beta function, $\boldsymbol{\gamma} = (\gamma_1,\gamma_2)$, $\gamma_1 > 0$ and $\gamma_2 > 0$. If we set $\gamma_1 = \gamma_2 = 2$, this simplifies to:

$$G(x;\boldsymbol{\tau}) = \int_0^{F(x;\tau)}\frac{t(1-t)}{B(2,2)}\,dt = 3F(x;\boldsymbol{\tau})^2 - 2F(x;\boldsymbol{\tau})^3, \quad x \in \mathbb{R}.$$

If we considere this last equations as baseline function $G(x,\boldsymbol{\tau})$, and F as the cdf of the Kumaraswamy distribution, that is, replacing (1) in (3), we have (4).

$$G(x;a,b) = 3(1-(1-x^a)^b)^2 - 2(1-(1-x^a)^b)^3, \quad 0 < x < 1. \tag{4}$$

So if we replace the equation (4) in (2) we obtain the proposed distribution, with cdf

$$F_X(x) = \frac{\log\left[1-(1-\phi)\left\{3(1-(1-x^a)^b)^2 - 2(1-(1-x^a)^b)^3\right\}^{\alpha}\right]}{\log(\phi)}, \quad 0 < x < 1, \tag{5}$$

and pdf

$$f_X(x) = \frac{6(1-\phi)}{-ln(\phi)}\frac{\alpha[3(1-(1-x^a)^b)^2 - 2(1-(1-x^a)^b)^3]^{\alpha-1}(1-(1-x^a)^b)abx^{a-1}(1-x^a)^{2b-1}}{\left[1-(1-\phi)\{3(1-(1-x^a)^b)^2 - 2(1-(1-x^a)^b)^3\}^{\alpha}\right]}, \quad 0 < x < 1. \tag{6}$$

Considering the relationship that the proposed distribution has with the median, we call this distribution a complementary exponentiated median-based Kumaraswamy-G-logarithmic (CEMBKGLn) distribution. It should also be noted that the cdf defined in (4) is a particular case of the Extended Kumaraswamy distribution (EKw), as defined by Carrasco and Cordeiro (2017). Specifically, when $X \sim EKw(a,b,2,2,1)$, its cdf corresponds to (4). Some limit cases of the CEMBKGLn distribution as $\phi \uparrow 1$ are:

- If $\alpha = a = b = 1$ then $X \sim Beta(2,2)$
- If $\alpha = b = 1$ then $X \sim GB(a,1,0,2,2)$
- If $\alpha = 1$ then $X \sim EKw(a,b,2,2,1)$

Considering the four parameters of the proposed distribution, various forms of density are obtained. Figure 1 shows some cases for different values of the parameters.

2.1 Quantile Function

The quantile function $Q(p)$ of the CEMBKGLn distribution has close form which is given by

$$Q(p) = \left(1-\left(1-\left\{\frac{1}{2} - sin\left[\frac{sin^{-1}(1-2u)}{3}\right]\right\}\right)^{1/b}\right)^{1/a}, \tag{7}$$

where $u = \left(\frac{1-\phi^p}{1-\phi}\right)^{1/\alpha}$. The quartiles of the CEMBKGLn distribution are obtained by setting $p = 0.25, 0.5, 0.75$ in equation (7).

2.2 Density Expansion

Using the geometric series expansion, the new pdf (6) can be expanded as follows

$$
\begin{aligned}
f_X(x) &= \frac{6\alpha(1-\phi)}{-\log(\phi)}\sum_{i=0}^{\infty}(1-\phi)^i[3(1-(1-x^a)^b)^2-2(1-(1-x^a)^b)^3]^{\alpha(i+1)-1}(1-(1-x^a)^b)abx^{a-1}(1-x^a)^{2b-1}\\
&= \sum_{i=0}^{\infty}\frac{-(1-\phi)^{i+1}}{(i+1)\log(\phi)}6\alpha(i+1)[3(1-(1-x^a)^b)^2-2(1-(1-x^a)^b)^3]^{\alpha(i+1)-1}(1-(1-x^a)^b)abx^{a-1}(1-x^a)^{2b-1}\\
&= \sum_{i=0}^{\infty}\omega(i,\phi)f_{EG}(x;1,\alpha(i+1),\phi,a,b) \qquad (8)
\end{aligned}
$$

where

$$f_{EG}(x;1,\alpha(i+1),\phi,a,b) = 6\alpha(i+1)[3(1-(1-x^a)^b)^2-2(1-(1-x^a)^b)^3]^{\alpha(i+1)-1}(1-(1-x^a)^b)abx^{a-1}(1-x^a)^{2b-1}$$

is the generalized exponential pdf defined by Cordeiro et al. (2013), with baseline function (4), and $\omega(i,\phi) = \frac{-(1-\phi)^{i+1}}{(i+1)\log(\phi)}$ are the weights. In other words, (8) shows that the pdf of the distribution can be written as a linear combination of the generalized exponential distribution. The result obtained in (8) has also been proved, for a general case, by Oluyede et al. (2020).

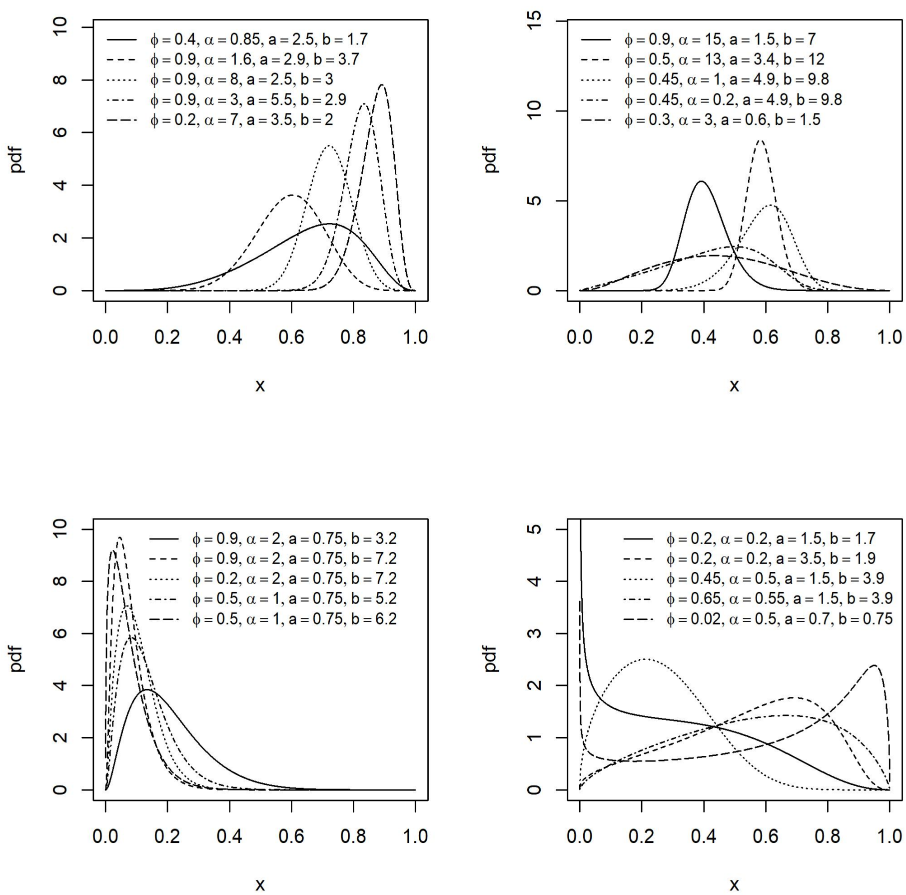

Figure 1. Plots of the CEMBKGLn density for some parameter values

2.3 Moments

Using (8) we can compute the $r-$th moment as

$$\begin{aligned}\mathbb{E}(X^r) &= \sum_{i=0}^{\infty}\omega(i,\phi)\int_0^1 x^r f_{EG}(x;1,\alpha(i+1),\phi,a,b)dx\\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\omega(i,\phi)\frac{(-1)^{j+k}\Gamma(\alpha(i+1))\alpha(i+1)}{k!\Gamma(\alpha(i+1)-j)\Gamma((j+1)-k)}\int_0^1 x^r G(x;a,b)^k g(x;a,b)dx\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function, $G(x;a,b)$ is the cdf defined in (4) and $g(x;a,b)$ is its respective pdf.

3. Parameter Estimation

In this section we introduce the maximum likelihood estimation method and conduct a simulation study to evaluate its performance and accuracy.

3.1 Maximum Likelihood Estimation

Let $X_1,...,X_n$ be a random sample of size n from the CEMBKGLn distribution with cdf and pdf given by (5) and (6), respectively, and the parameter vector $\boldsymbol{\theta}=(\phi,\alpha,a,b)$. The log-likelihood function of this sample is

$$\begin{aligned}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} &\log(6)+\log(1-\phi)-\log(-\log(\phi))-\log\left(1-(1-\phi)\{3(1-(1-x_i^a)^b)^2-2(1-(1-x_i^a)^b)^3\}^{\alpha}\right)\\ &+ln(\alpha)+(\alpha-1)\log\left(3(1-(1-x_i^a)^b)^2-2(1-(1-x_i^a)^b)^3\right)+\log(1-(1-x_i^a)^b)+\log(a)\\ &+\log(b)+(a-1)\log(x_i)+(2b-1)\log(1-x_i^a). \qquad (9)\end{aligned}$$

Obtaining the partial derivatives of (9), one may get

$$\begin{aligned}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\phi} &= \frac{n}{\phi-1}-\frac{n}{ln(\phi)\phi}-\sum_{i=1}^{n}\frac{G(x_i;a,b)^{\alpha}}{1-(1-\phi)G(x_i;a,b)^{\alpha}}\\ \frac{\partial\ell(\boldsymbol{\theta})}{\partial\alpha} &= \frac{n}{\alpha}+\sum_{i=1}^{n}\log(G(x_i;a,b))+\frac{(1-\phi)log(G(x_i;a,b))G(x_i;a,b)^{\alpha}}{1-(1-\phi)G(x_i;a,b)^{\alpha}}\\ \frac{\partial\ell(\boldsymbol{\theta})}{\partial a} &= \frac{n}{a}+\sum_{i=1}^{n}\frac{A(x_i,a,b)(\alpha-1)}{G(x_i;a,b)}+log(x_i)-\frac{x_i^a log(x_i)}{1-x_i^a}+\frac{bx_i log(x_i)(1-x_i^a)^{b-1}}{1-(1-x_i^a)^b}+\frac{(1-\phi)A(x_i,a,b)\alpha G(x_i;a,b)^{\alpha-1}}{1-(1-\phi)G(x_i;a,b)^{\alpha}}\\ \frac{\partial\ell(\boldsymbol{\theta})}{\partial b} &= \sum_{i=1}^{n}\frac{B(x_i,a,b)(\alpha-1)}{G(x_i;a,b)}-\frac{(1-x_i^a)^b\log(1-x_i^a)}{1-(1-x_i^a)^b}+\frac{1}{b}+2\log(1-x_i^a)+\frac{(1-\phi)\alpha G(x_i;a,b)^{\alpha-1}B(x_i,a,b)}{1-(1-\phi)G(x_i;a,b)^{\alpha}}\end{aligned}$$

where

$$\begin{aligned}A(x_i,a,b) &= 6bx_i^a\log(x_i)\,(1-x_i^a)^{2b-1}\left(1-(1-x_i^a)^b\right),\\ B(x_i,a,b) &= -6\,(1-x_i^a)^{2b}\left(1-(1-x_i^a)^b\right)\log\left(1-x_i^a\right).\end{aligned}$$

To determine the maximum likelihood estimators, one must solve estimation equations:

$$\left(\frac{\partial\ell(\boldsymbol{\theta})}{\partial\phi}\quad\frac{\partial\ell(\boldsymbol{\theta})}{\partial\alpha}\quad\frac{\partial\ell(\boldsymbol{\theta})}{\partial a}\quad\frac{\partial\ell(\boldsymbol{\theta})}{\partial b}\right)^{\top}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}=\mathbf{0}.$$

Since the system has no analytical solution, numerical methods such as Newton-Raphson, L-BFGS, must be used. These methods are already implemented in various software such as `R`, `Mathematica` and `Matlab`. For the purposes of this study, the `optim` function in `R` is used, employing the L-BFGS-B method due to the constraint $\phi\in(0,1)$. The Fisher information matrix can be estimated using

$$\boldsymbol{J}(\hat{\boldsymbol{\theta}})=-\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

since it is a consistent estimator. When n is large and the regularity conditions are met, we have that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\xrightarrow{d}\mathcal{N}(\mathbf{0},\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}),$$

so that it is possible to obtain the approximate variance of the estimators, using the diagonal of the matrix $\boldsymbol{J}(\hat{\boldsymbol{\theta}})^{-1}$.

3.2 Simulation

To evaluate the performance of the maximum likelihood estimators, the Monte Carlo method is employed. Simulations are conducted using the proposed distribution, generated via the inverse transform method, as the quantile function (7) is available in analytical form. Sample sizes of $n = 50, 100, 200$ and 500 are considered. For each sample, the maximum likelihood estimates (MLEs) are computed. This process is repeated $N = 10000$ times across three different parameter scenarios: $I : \phi = 0.5, \alpha = 1, a = 1, b = 1$, $II : \phi = 0.1, \alpha = 2, a = 1.5, b = 2$, $III : \phi = 0.8, \alpha = 1, a = 3, b = 3.5$. The following quantities are calculated based on the results of the MLEs $\hat{\nu} = \phi, \alpha, a, b$:

1. The Mean: $\overline{\nu} = \sum_{j=1}^{N} \frac{\hat{\nu}_j}{N}$
2. The Bias: $Bias(\hat{\nu}) = \overline{\nu} - \nu$
3. The root mean-square error: $RMSE(\hat{\nu}) = \sqrt{(1/N) \sum_{j=1}^{N} (\hat{\nu}_j - \nu)^2}$

Table A1 in the Appendix presents the results of the simulation study. The results show that as the sample size increases, the mean estimates of the parameters approach their true values, while the bias and RMSE decrease.

Based on these results, it can be concluded that the maximum likelihood method provides acceptable results. However, the parameter α shows the highest bias and RMSE. Even so, similar results can be observed in some simulation studies, for example, in the articles by Huang and Oluyede (2014), Makubate and Musekwa (2024), Y?ld?r?m et al. (2023) and Chaisee et al. (2024).

To reduce the bias of $\hat{\alpha}$, various methods have been proposed in the literature, such as the technique proposed by Cox and Snell (1968) and the parametric Bootstrap method introduced by Efron (1982). In both cases, if the maximum likelihood estimator is considered, a bias correction can be made to reduce the bias of the estimator. Different works such as those of Ahmed et al. (2024), Sangpoom and Klomwises (2021) and Sungboonchoo (2025) have adopted these methodologies, demonstrating a reduction in parameter bias as well as a decrease in RMSE, particularly in relatively small sample sizes. Alternatively, other estimation methods can also be considered, such as the least squares estimates, Anderson-Darling method, percentile estimates or Cramr-von Mises estimates, in order to obtain different perspectives of the behavior of the $\hat{\alpha}$ value according to the estimation method. All these approaches for bias reduction will be considered in future work.

4. Univariate Applications

In this section, we compare the fit of the proposed distribution CEMBKGLn with the fit of four competitors on two different real data sets. The competitors considered are:

1. Beta pdf:
$$f_{Beta}(x; \alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \quad \alpha, \beta > 0$$
2. KM pdf (Kumaraswamy, 1980):
$$f_{KM}(x; a, b) = abx^{a-1}(1 - x^a)^{b-1}, \quad a, b > 0$$
3. I-UDa pdf (Condino and Domma, 2023):
$$f_{I-UDa}(x; \beta, \lambda, \delta) = \frac{\beta\lambda\delta}{x\sqrt{1 - x^2}} \left[\log\left(\frac{1 + \sqrt{1 - x^2}}{x}\right)\right]^{-\delta-1} \left\{1 + \lambda \left[\log\left(\frac{1 + \sqrt{1 - x^2}}{x}\right)\right]^{-\delta}\right\}^{-\beta-1}, \quad \beta, \lambda, \delta > 0$$
4. UW2 pdf (Reyes et al., 2023):
$$f_{UW2}(x; \theta) = \frac{\theta\beta^{\theta} x^{\theta-1}(1 - x)^{\theta-1}}{\left[(\beta x)^{\theta} + (1 - x)^{\theta}\right]^2}, \quad \beta, \theta > 0$$

To compare these distributions, the Kolmogorov-Smirnov (K-S) statistic and the associated p-values are considered.

4.1 Data Set 1

The first data set was extracted from the SINCA website https://sinca.mma.gob.cl/index.php/, which corresponds to hourly measurements of relative humidity (in percentage) in the city of Talagante, Chile, between the periods 2022/11/01 -

2023/03/29. This period is considered because it is the hottest of the year. In particular, it is of interest to analyze the maximum relative humidity per day during this period. A scaling was carried out to the unitary interval of the variable of interest, that is, dividing the percentages by 100, having a sample size $n = 149$.

Table 1 shows the estimations of each distribution along with their standard errors (SEs), as well as the value of the Kolmogorov-Smirnov statistic and its respective p-value. The CEMBKGLn distribution achieves a better fit than the other distributions in terms of the K-S test, since the associated p-value is higher than the rest. Figure 2 shows the fits of the respective densities.

Table 1. MLEs and their SEs (on second line) of the fitted models and goodness-of-fit statistics for the first data set

Models	**Parameter estimations**	**K-S**	**p-value**
CEMBKGLn(ϕ, α, a, b)	0.0012 7.5598 1.7504 1.6307	**0.0729**	**0.4064**
	0.0027 11.7035 1.2938 0.3978		
I-UDa(β, λ, δ)	0.9117 0.0355 4.0521	0.0867	0.2118
	0.4348 0.0112 0.7709		
Beta(α, β)	14.2053 1.7009	0.0920	0.1598
	1.7228 0.1811		
Kumaraswamy(a, b)	12.2204 1.8169	0.0928	0.1531
	1.1495 0.2208		
UW2(β, θ)	0.0911 1.9251	0.0845	0.2372
	0.0068 0.1283		

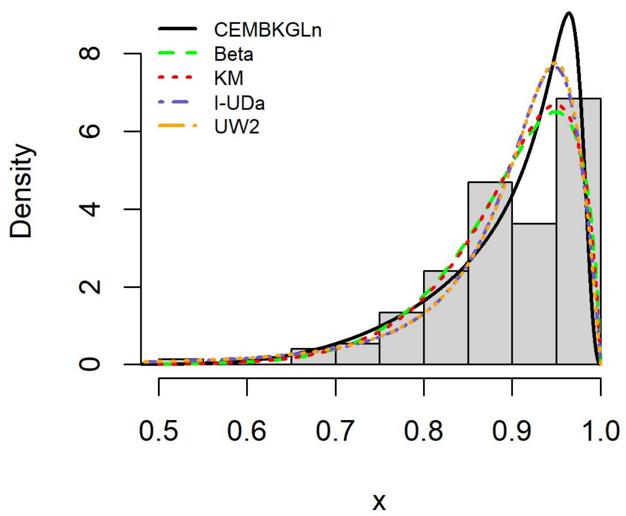

Figure 2. Fitted PDFs and the observed histogram for the first data set

4.2 Data Set 2

The second data set was extracted from Cordeiro and Bager (2012), and consists of 48 rock samples from an oil field. In particular it corresponds to twelve core samples taken from oil fields, which were sampled in four cross sections. Each core sample was measured in terms of permeability, and each cross section includes the following variables: total pore area, total pore perimeter and shape. The shape variable of the perimeter is analyzed as a function of the squared area.

Table 2 shows the estimations of each distribution along with their SEs, as well as the value of the Kolmogorov-Smirnov statistic and its respective p-value. The CEMBKGLn distribution achieves a better fit than the other distributions in terms of the K-S test, since the associated p-value is higher than the rest. Figure 3 shows the fits of the respective densities.

Table 2. MLEs and their SEs (on second line) of the fitted models and goodness-of-fit statis- tics for the second data set

Models	**Parameter estimations**	**K-S**	**p-value**
CEMBKGLn(ϕ, α, a, b)	0.9999 1053.5075 0.1770 3.0105	**0.0865**	**0.8651**
	1.9786 1665.0209 0.0896 0.8430		
I-UDa(β, λ, δ)	0.9359 3277.3125 9.9881	0.1503	0.2283
	0.1982 1875.0201 0.6584		
Beta(α, β)	5.9422 21.2074	0.1427	0.2820
	1.1814 4.3472		
Kumaraswamy(a, b)	2.7187 44.6603	0.1533	0.2092
	0.2935 17.5756		
UW2(β, θ)	3.9200 3.8275	0.09789	0.7471
	0.2574 0.4610		

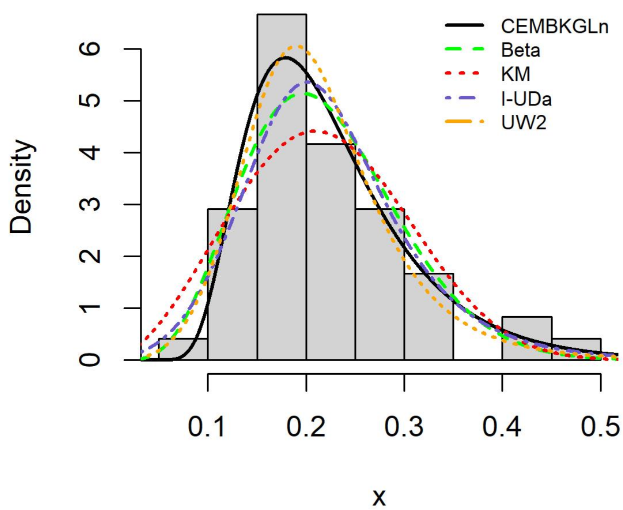

Figure 3. Fitted PDFs and the observed histogram for the second data set

5. Quantile Regression

Regression models for response variables bounded within the unit interval are important for modeling percentages and proportions. The most useful and popular is beta regression, defined by Ferrari and Cribari-Neto (2004), but when the response variable contains outliers, then it is better to consider a quantile regression model over classical regression models, since the mean is more sensitive to these values. As the CEMBKGLn distribution has no analytical form for the mean, but the quantile function is somewhat tractable, it is proposed to define a quantile regression model based on the median, since it is considered a competitor of the mean.

Considering the reparameterization suggested by Mitnik and Baek (2013) of the Kumaraswamy distribution, that is, solving for b in the equation $Q(0.5) = \mu$ we have

$$\Psi(\boldsymbol{\xi}) = b = \frac{\left(1 - \left\{\frac{1}{2} - sin\left[\frac{sin^{-1}(1-2u)}{3}\right]\right\}\right)}{\log(1-\mu^a)}, \tag{10}$$

where $u = \left(\frac{1-\phi^{1/2}}{1-\phi}\right)^{1/\alpha}$ and $\boldsymbol{\xi} = (\mu, \phi, \alpha, a)$. Reeplacing (10) in the density (6) the reparameterized density is given by

$$f_Y(y) = \frac{6(1-\phi)}{-ln(\phi)} \frac{\alpha[3(1-(1-y^a)^{\Psi(\boldsymbol{\xi})})^2 - 2(1-(1-y^a)^{\Psi(\boldsymbol{\xi})})^3]^{\alpha-1}(1-(1-y^a)^{\Psi(\boldsymbol{\xi})})a\Psi(\boldsymbol{\xi})y^{a-1}(1-y^a)^{2\Psi(\boldsymbol{\xi})-1}}{\left[1-(1-\phi)\left\{3(1-(1-y^a)^{\Psi(\boldsymbol{\xi})})^2 - 2(1-(1-y^a)^{\Psi(\boldsymbol{\xi})})^3\right\}^{\alpha}\right]}, \tag{11}$$

where $\mu \in (0, 1)$ represents the parameter of the quantile regression. Now, if we assume a random sample $Y_1, ..., Y_n$, with observed values $y_1, ..., y_n$ coming from the density (11), and we suppose that the median for each observation can be written as $\mu_i = g(\boldsymbol{x}_i^\top \boldsymbol{\theta})$, $i = 1, ..., n$, where $\boldsymbol{x}_i^\top = (x_{i1}, ..., x_{ip})^\top$ is the vector of covariates associated with the response y_i, $\boldsymbol{\theta} = (\theta_0, ..., \theta_p)$ is the vector of regression coefficients and $g(\cdot)$ is the link function.

For the purposes of this article, the logit link function is considered, so it is assumed that

$$g(\mu_i) = logit(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\theta}, \quad i = 1, 2, \ldots, n, \tag{12}$$

solving for μ_i we obtain the expression (13).

$$\mu_i = \frac{\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\theta}\right)}{1+\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\theta}\right)}. \tag{13}$$

Replacing the equation (13) in (10) the quantile regression model for the median is obtained. Considering the equation (11) and the parameter vector $\boldsymbol{\Omega} = (\phi, \alpha, a, \boldsymbol{\theta})$, the log-likelihood of the model is

$$\begin{aligned}\ell(\boldsymbol{\Omega}) = \sum_{i=1}^{n} & \log(6) + \log(1-\phi) - \log(-\log(\phi)) - \log\left(1-(1-\phi)\left\{3\left(1-(1-y_i^a)^{\Psi(\mu_i,\phi,\alpha,a)}\right)^2 - 2\left(1-(1-y_i^a)^{\Psi(\mu_i,\phi,\alpha,a))}\right)^3\right\}^{\alpha}\right) \\ & + ln(\alpha) + (\alpha-1)\log\left(3\left(1-(1-y_i^a)^{\Psi(\mu_i,\phi,\alpha,a))}\right)^2 - 2\left(1-(1-y_i^a)^{\Psi(\mu_i,\phi,\alpha,a))}\right)^3\right) + \log(1-(1-x_i^a)^{\Psi(\mu_i,\phi,\alpha,a)} + \log(a) \\ & + \log(\Psi(\mu_i,\phi,\alpha,a)) + (a-1)\log(y_i) + (2\Psi(\mu_i,\phi,\alpha,a)) - 1)\log(1-y_i^a)\end{aligned} \tag{14}$$

To obtain the maximum likelihood estimators, the equation (14) resulting from the quantile regression model must be maximized. For practical purposes, we will omit these expressions and proceed to use the `optim` function of `R`.

To illustrate the model, we consider the data used by Jodra and Jimenez-Gamero (2020), but we examine only the covariates that were significant in their study, that is, we use the response variable y = FIRMCOST/100, and the covariates SIZELOG (x_1) and INDCOST (x_2). Thus, (12) in this case is written as

$$logit(\mu_i) = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}, \quad i = 1, ..., 73.$$

Our model is compared with the UBS (Mazucheli et al., 2018b), Kumaraswamy (Kumaraswamy, 1980) and Johnson S_B (Jonhson, 1949) quantile regresion models. The coefficients are estimated using the `unitquantreg` library developed by Menezes and Mazucheli (2021). To evaluate the quality of fit of the models, the AIC (Akaike, 1974) and BIC (Schwarz, 1978) criteria are considered. To discriminate between the CEMBKLn with each competing models, the Vuong test (Vuong, 1989) is considered. In addition, to visually evaluate the fit of the model, the residuals are analyzed by randomized quantiles residuals defined by Dunn (1997), which are defined by

$$\hat{r}_i = \Phi^{-1}\left[F\left(y_i, \hat{\phi}, \hat{\alpha}, \hat{a}, \hat{\mu}_i\right)\right], \tag{15}$$

where $F\left(y_i, \hat{\phi}, \hat{\alpha}, \hat{a}, \hat{\mu}_i\right)$ is the cdf resulting from the parameterization obtained in (11) and Φ^{-1} is the quantile function of a standard normal distribution.

For numerical reasons, the parameter ϕ is reparameterized using a logit transformation. The R codes for the algorithm used in the simulation study in Section 3 and the quantile regression application are available at https://github.com/danielglvez/CEMBKGLn-distribution.

Table 3 shows the estimated values of the coefficients together with their associated variances and $p-$values, where the CEMBKGLn model obtains lowest AIC. Therefore, at least based on AIC criteria, it shows a better performance compared to the rest of the competitors considered. Regarding the BIC, the CEMBKGLn model was the second best, only surpassed by the Johnson S_B model, which has less parameters. On the other hand, as regards the estimation of the regression coefficients, the results are similar to those obtained with the Johnson S_B model, but quite different from those obtained with the KM and UBS models, which may be due to the distribution assumed for the response variable. Similar results can be seen, for example, in the articles by Korkmaz and Korkmaz (2023) and Fayomi et al. (2023).

Table 3. The results of fitted regression models with the considered model selection criteria

	CEMBKGLn			**Johnson S_B**			**KM**			**UBS**		
Parameters	**Estimate**	**SE**	**$p-$value**	**Estimate**	**SE**	**p-value**	**Estimate**	**SE**	**$p-$value**	**Estimate**	**SE**	**$p-$value**
θ_0	3.3874	1.0226	< 0.001	3.1910	1.2870	0.0131	0.5232	1.1600	0.6520	-3.6473	2.0956	0.0817
θ_1	-0.8165	0.1201	< 0.001	-0.8238	0.1533	< 0.001	-0.6718	0.1587	< 0.01	0.1178	0.2302	0.6084
θ_2	1.6987	0.4347	< 0.001	2.3424	0.7244	0.0012	6.4671	0.9126	< 0.01	2.7033	0.6978	< 0.001
ℓ	110.9130			107.3787			95.9179			38.4095		
AIC	-209.8260			-206.7574			-183.8359			-68.8190		
BIC	-196.0832			-197.5955			-174.6741			-59.6571		

Table 4 shows the values of the Voung statistic and their respective $p-$values, where it is obtained that the CEMBKGLn model is preferred with 95% confidence.

Figure 4 shows the residuals defined in (15). This figure shows that the UBS model has two values that are quite far from the confidence intervals, while the other models only have one value with these characteristics. On the other hand, the KM model seems to capture this extreme value better, but the residuals do not fit the straight line entirely, while in the CEMBKGLn model fits much better, thus showing that the fit provided by it is good.

Table 4. Observed values of Voung statistic and p-values

Comparisons	**Voung statistic**	**p-value**
CEMBKGLn vs UBS	16.6731	< 0.001
CEMBKGLn vs KM	2.4160	0.0156
CEMBKGLn vs Johnson S_B	2.4637	0.0137

6. Conclusion

In this article, a new unitary distribution with four parameters called CEMBKGLn was defined, which is useful for modeling asymmetric data. Some of its properties were studied, showing the relationship it has with the distribution proposed by Cordeiro et al. (2013). A simulation study was conducted to analyze the performance of maximum likelihood estimators. Two applications of the new distribution were developed using two real data sets, where it shows a better performance compared to classical distributions such as Beta and Kumaraswamy. A quantile regression model based on the median was defined, whose application to a real data set suggests greater flexibility compared to other distributions, obtaining good results in different criteria such as AIC and BIC, thus demonstrating that it is a useful model. Finally, the CEMBKGLn distribution turns out to be quite flexible to model data in the unit interval, and has a closed form for the quantile function, so it has potential applications when developing a quantile regression model.

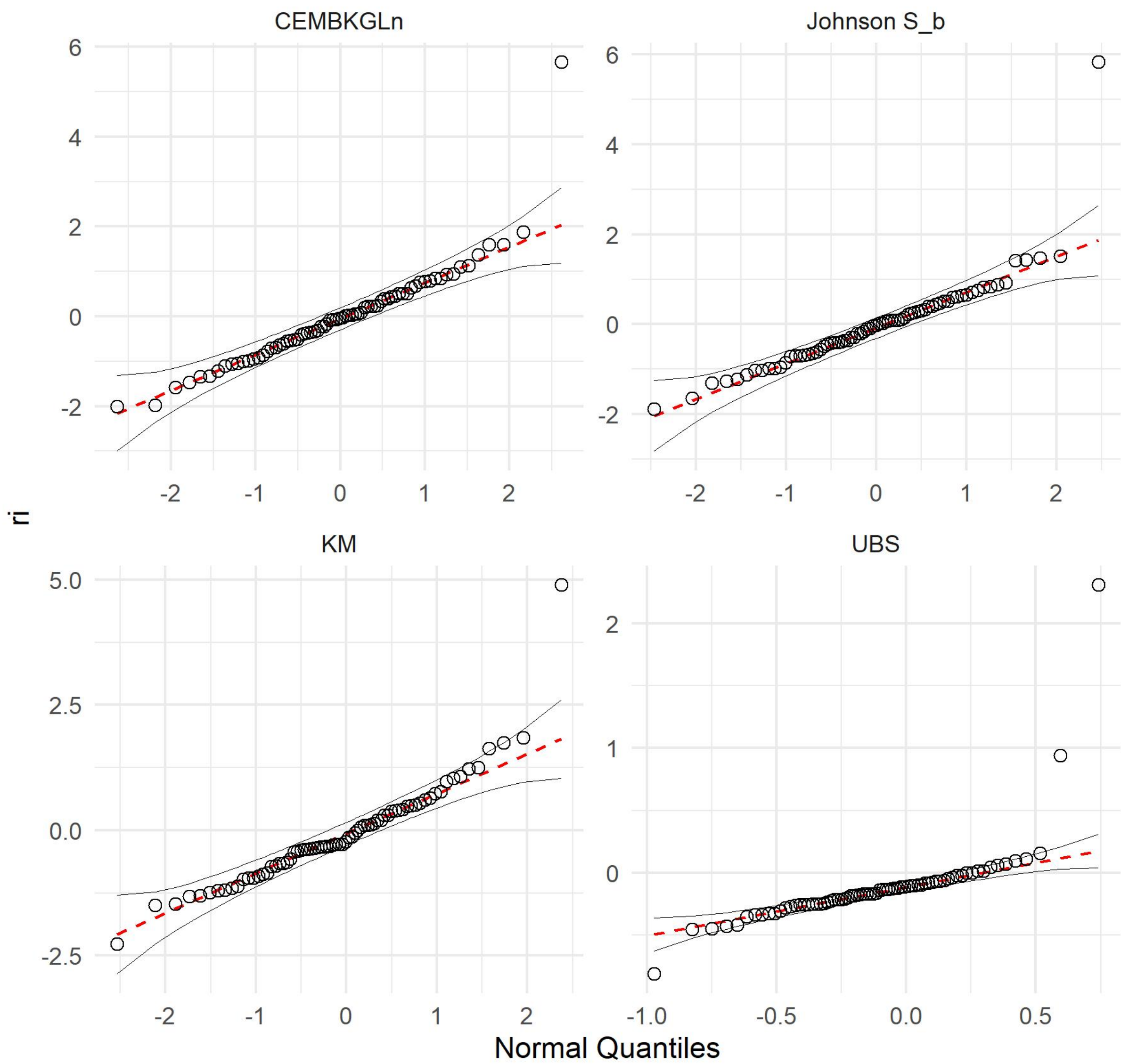

Figure 4. Randomized Quantile Residuals

Acknowledgments

Special thanks to Prof. Reinaldo Arellano-Valle for his advice and valuable feedback on this article.

References

Ahmed, A. A., Algamal, Z. Y., & Albalawi, O. (2024). Bias reduction of maximum likelihood estimation in exponentiated Teissier distribution. *Frontiers in Applied Mathematics and Statistics, 10*, 1351651. https://doi.org/10.3389/fams.2024.1351651

Aitchison, J., & Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika, 67*(2), 261-272. https://doi.org/10.2307/2335470

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716C723. https://doi.org/10.1109/TAC.1974.1100705

Altun, E., & Cordeiro, G. M. (2020). The unit-improved second-degree Lindley distribution: inference and regression modeling. *Computational Statistics, 35*, 259-279. https://doi.org/10.1007/s00180-019-00921-y

Altun, E., El-Morshedy, M., & Eliwa, M. S. (2021). A new regression model for bounded response variable: An alternative to the beta and unit-Lindley regression models. *Plos one, 16*(1), e0245627. https://doi.org/10.1371/journal.pone.0245627

Alzaatreh, A., Lee, C., & Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron, 71*(1), 63-79. https://doi.org/10.1007/s40300-013-0007-y

Attia, I. M. (2024). A Novel Unit Distribution Named As Median Based Unit Rayleigh (MBUR): Properties and Estimations. *arXiv preprint arXiv:2410.04132*.

Bakouch, H. S., Hussain, T., Tošió, M., Stojanović, V. S., & Qarmalah, N. (2023). Unit exponential probability distribution: Characterization and applications in environmental and engineering data modeling. *Mathematics, 11*(19), 4207. https://doi.org/10.3390/math11194207

Carrasco, J. M., & Cordeiro, G. M. (2017). An extension of the Kumaraswamy distribution. *International Journal of Statistics and Probability, 6*(3), 61. https://doi.org/10.5539/ijsp.v6n3p61

Chaisee, K., Khamkong, M., & Paksaranuwat, P. (2024). A New Extension of the Exponentiated WeibullCPoisson Family Using the Gamma-Exponentiated Weibull Distribution: Development and Applications. *Symmetry, 16*(7), 780. https://doi.org/10.3390/sym16070780

Condino, F., & Domma, F. (2023). Unit Distributions: A General Framework, Some Special Cases, and the Regression Unit-Dagum Models. *Mathematics, 11*(13), 2888. https://doi.org/10.3390/math11132888

Cordeiro, G. M., & R. B. dos Santos. (2012). The beta power distribution. *Brazilian Journal of Probabilityand Statistics, 26*(1), 88-112. https://doi.org/10.1214/10-BJPS124

Cordeiro, G. M., & De Castro, M. (2011). A new family of generalized distributions. *Journal of statistical computation and simulation, 81*(7), 883-898. https://doi.org/10.1080/00949650903530745

Cordeiro, G. M., Ortega, E. M., & da Cunha, D. C. (2013). The exponentiated generalized class of distributions. *Journal of data science, 11*(1), 1-27. https://doi.org/10.6339/JDS.2013.11(1).1086

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. https://doi.org/10.1111/j.2517-6161.1968.tb00724.x

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and graphical statistics, 5*(3), 236-244. https://doi.org/10.2307/1390802

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for industrial and applied mathematics.

Eugene, N., Lee, C., & Famoye, F. (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and methods, 31*(4), 497-512. https://doi.org/10.1081/STA-120003130

Fayomi, A., Hassan, A. S., & Almetwally, E. M. (2023). Inference and quantile regression for the unit-exponentiated Lomax distribution. *Plos one, 18*(7), e0288635. https://doi.org/10.1371/journal.pone.0288635

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics, 31*(7), 799-815. https://doi.org/10.1080/0266476042000214501

Ghitany, M. E., Mazucheli, J., Menezes, A. F. B., & Alqallaf, F. (2019). The unit-inverse Gaussian distribution: A new alternative to two-parameter distributions on the unit interval. *Communications in Statistics-Theory and methods, 48*(14), 3423-3438. https://doi.org/10.1080/03610926.2018.1476717

Grassia, A. (1977). On a family of distributions with argument between 0 and 1 obtained by transformation of the gamma and derived compound distributions. *Australian Journal of Statistics, 19*(2), 108-114. https://doi.org/10.1111/j.1467-842X.1977.tb01277.x

Hakamipour, N., Zhang, Y., & Nadarajah, S. (2022). A new family of compound exponentiated logarithmic distributions with applications to lifetime data. *Mathematica Slovaca, 72*(5), 1337-1354. https://doi.org/10.1515/ms-2022-0091

Huang, S., & Oluyede, B. O. (2014). Exponentiated Kumaraswamy-Dagum distribution with applications to income and lifetime data. *Journal of Statistical Distributions and Applications, 1*, 1-20. https://doi.org/10.1186/2195-5832-1-8

Jamal, F., Nasir, M. A., Tahir, M. H., & Montazeri, N. H. (2017). The odd Burr-III family of distributions. *Journal of Statistics Applications and Probability, 6*(1), 105-122. http://dx.doi.org/10.18576/jsap/060109

Jodra, P., & Jimenez-Gamero, M. D. (2020). A quantile regression model for bounded responses based on the exponential-geometric distribution. *REVSTAT-Statistical Journal, 18*(4), 415-436. https://doi.org/10.57805/revstat.v18i4.309

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika, 36*(1/2), 149-176. https://doi.org/10.2307/2332539

Korkmaz, M. C., Altun, E., Chesneau, C., & Yousof, H. M. (2022). On the unit-Chen distribution with associated quantile regression and applications. *Mathematica Slovaca, 72*(3), 765-786. https://doi.org/10.1515/ms-2022-0052

Korkmaz, M. Ç., & Korkmaz, Z. S. (2023). The unit logClog distribution: A new unit distribution with alternative quantile regression modeling and educational measurements applications. *Journal of Applied Statistics, 50*(4), 889-908. https://doi.org/10.1080/02664763.2021.2001442

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of hydrology, 46*(1-2), 79-88. https://doi.org/10.1016/0022-1694(80)90036-0

Mahdavi, A., & Kundu, D. (2017). A new method for generating distributions with an application to exponential distribution. *Communications in Statistics-Theory and Methods, 46*(13), 6543-6557. https://doi.org/10.1080/03610926.2015.1130839

Mahmoudi, E., & Jafari, A. A. (2012). Generalized exponentialCpower series distributions. *Computational Statistics & Data Analysis, 56*(12), 4047-4066. https://doi.org/10.1016/j.csda.2012.04.009

Mahmoudi, E., Sepahdar, A., & Lemonte, A. (2014). Exponentiated Weibull-logarithmic distribution: Model, properties and applications. *arXiv preprint arXiv:1402.5264*.

Makubate, B., & Musekwa, R. R. (2024). A novel technique for generating families of continuous distributions. *Statistics, Optimization & Information Computing, 12*(5), 1231-1248. https://doi.org/10.19139/soic-2310-5070-2068

Marinho, P. R. D., Cordeiro, G. M., Ramłrez, F. P., Alizadeh, M., & Bourguignon, M. (2018). The exponentiated logarithmic generated family of distributions and the evaluation of the confidence intervals by percentile bootstrap. *Brazilian Journal of Probability and Statistics, 32*(2), 281-308. https://doi.org/10.1214/16-BJPS343

Mazucheli, J., Menezes, A. F. B., & Ghitany, M. E. (2018a). The unit-Weibull distribution and associated inference. *J. Appl. Probab. Stat, 13*(2), 1-22.

Mazucheli, J., Menezes, A. F., & Dey, S. (2018b). The unit-Birnbaum-Saunders distribution with applications. *Chilean Journal of Statistics, 9*(1), 47-57.

McDonald, J. B., & Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics, 66*(1-2), 133-152. https://doi.org/10.1016/0304-4076(94)01612-4

Menezes, A. & Mazucheli, J. (2021). *unitquantreg: Parametric quantile regression models for bounded data.* R package version 0.0.6, https://andrmenezes.github.io/unitquantreg/.

Mitnik, P. A., & Baek, S. (2013). The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers, 54*, 177-192. https://doi.org/10.1007/s00362-011-0417-y

Nofal, Z. M., Afify, A. Z., Yousof, H. M., & Cordeiro, G. M. (2017). The generalized transmuted-G family of distributions. *Communications in Statistics-Theory and Methods, 46*(8), 4119-4136. https://doi.org/10.1080/03610926.2015.1078478

Okutu, J. K., Frempong, N. K., Appiah, S. K., & Adebanji, A. O. (2023). A New Generated Family of Distributions: Statistical Properties and Applications with Real-Life Data. *Computational and Mathematical Methods, 2023*(1), 9325679. https://doi.org/10.1155/2023/9325679

Oluyede, B. O., Mashabe, B., Fagbamigbe, A., Makubate, B., & Wanduku, D. (2020). The exponentiated generalized power series: Family of distributions: theory, properties and applications. *Heliyon, 6*(8). https://doi.org/10.1016/j.heliyon.2020.e04653

Ramadan, A. T., Tolba, A. H., & El-Desouky, B. S. (2022). A unit half-logistic geometric distribution and its application in insurance. *Axioms, 11*(12), 676. https://doi.org/10.3390/axioms11120676

Reyes, J., Rojas, M. A., Corts, P. L., & Arru, J. (2023). A new more flexible class of distributions on (0, 1): Properties and applications to univariate data and quantile regression. *Symmetry, 15*(2), 267. https://doi.org/10.3390/sym15020267

Sangpoom, S., & Klomwises, Y. (2021). Bias-corrected maximum likelihood estimation of the parameters of the modified power function distribution. *Trends in Sciences, 18*(19), 14-14. https://doi.org/10.48048/tis.2021.14

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461C464. https://doi.org/10.1214/aos/1176344136

Souza, L., O Jnior, W. R. D., Brito, C. C. R. D., Chesneau, C., Fernandes, R. L., & Ferreira, T. A. (2021). Tan-G class of trigonometric distributions and its applications. *Cubo (Temuco), 23*(1), 1-20. https://doi.org/10.4067/S0719-06462021000100001

Sungboonchoo, C. (2025). Improved Maximum Likelihood Estimator of the Extended Rama Distribution with Application to Lifetime Data. *Science and Technology Indonesia, 10*(1), 80C87. https://doi.org/10.26554/sti.2025.10.1.80-87

Tahir, M. H., & Cordeiro, G. M. (2016). Compounding of distributions: a survey and new generalized classes. *Journal of Statistical Distributions and Applications, 3*, 1-35. https://doi.org/10.1186/s40488-016-0052-1

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: journal of the Econometric Society, 57*(2) 307-333. https://doi.org/10.2307/1912557

Yıldırım, E., Iıkkan, E. S., Gemeay, A. M., Makumi, N., Bakr, M. E., & Balogun, O. S. (2023). Power unit Burr-XII distribution: Statistical inference with applications. *Aip Advances, 13*(10). https://doi.org/10.1063/5.0171077

Appendix

Table A1. Monte Carlo simulation results: Mean estimates, Biases and RMSEs

n	Parameter	I			II			III		
		Mean	Bias	RMSE	Mean	Bias	RMSE	Mean	Bias	RMSE
50	ϕ	0.4841	-0.0158	0.4247	0.4478	0.3478	0.5391	0.5650	-0.2349	0.4970
	α	4.8459	3.8459	9.4439	6.3562	4.3562	15.6794	6.7558	5.7558	15.9884
	a	0.9450	-0.0549	1.1630	3.5004	2.0004	4.7779	2.5626	-0.4373	2.1660
	b	1.3787	0.3787	1.0975	4.3617	2.3617	12.3096	5.2567	1.7567	1.3794
100	ϕ	0.5011	0.0011	0.3955	0.3933	0.2933	0.4827	0.5781	-0.2218	0.4715
	α	3.1762	2.1762	5.0115	3.6149	1.6149	6.6915	3.7172	2.7172	6.9747
	a	0.9228	-0.0771	0.9677	2.8941	1.3941	3.2697	2.5255	-0.4744	1.6461
	b	1.1271	0.1271	0.3837	2.8675	0.8675	4.3324	3.9473	0.4473	2.6578
200	ϕ	0.5009	0.0009	0.3558	0.3436	0.2436	0.4268	0.5672	-0.2327	0.4202
	α	2.2252	1.2252	2.7265	2.6449	0.6449	3.5678	2.3810	1.3810	3.5325
	a	0.9327	-0.0672	0.7977	2.5201	1.0201	2.2364	2.6179	-0.3820	1.3221
	b	1.0555	0.0555	0.2107	2.3935	0.3935	1.2135	3.6589	0.1589	1.4953
500	ϕ	0.4912	-0.0087	0.2980	0.2669	0.1669	0.3323	0.5743	-0.2256	0.3849
	α	1.6689	0.6689	1.4081	2.1624	0.1624	2.0044	1.7144	0.7144	1.6230
	a	0.9183	-0.0816	0.5613	2.1537	0.6536	1.4191	2.5962	-0.4037	0.9706
	b	1.0173	0.0173	0.1118	2.1757	0.1757	0.4245	3.4355	-0.0644	0.7253