Estimation of shape # parameter in Kumaraswamy distribution using Maximum Likelihood and Bayes method

Cite as: AIP Conference Proceedings **1862**, 030160 (2017); https://doi.org/10.1063/1.4991264
Published Online: 10 July 2017

H. G. Simbolon, I. Fithriani, and S. Nurrohmah

Estimation of Shape $\boldsymbol{\beta}$ Parameter in Kumaraswamy Distribution Using Maximum Likelihood and Bayes Method

H. G. Simbolon, I. Fithriani, and S. Nurrohmah[a)]

Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

[a)]Corresponding authors: snurrohmah@sci.ui.ac.id

Abstract. This paper discusses the Maximum Likelihood (ML) and Bayes method for estimating the shape β parameter in Kumaraswamy distribution. Both methods will be compared according to Mean Square Error (MSE) obtained from each estimator. In the Bayes method, two Loss functions will be used, i.e., the Square Error Loss Function (SELF) and Precautionary Loss Function (PLF). Then, the Posterior Risk obtained from both loss functions will be compared. The comparison will be applied to hydrological data as a recommendation for the best method of representing the data. Hydrological data used in this study is a water storage in Shasta Reservoir, obtained from the California Data Exchange Center. By using the Mathematica Software and the formulas from both methods one obtains a statistic which can nicely describe the data and also predict the next observation of a reservoir in a certain time.

INTRODUCTION

One of the greatest achievements in hydrology was made by an Indian Hydrologist, Poondi Kumaraswamy (1930-1988). Started from his experiment with a conclusion that any probability distribution function like Normal, Log-Normal, Gamma, Beta, and other distributions cannot fit hydrological data, Kumaraswamy developed a new distribution that faithfully fits hydrological data which is later known as Kumaraswamy Distribution [1].

This paper will focus on the Kumaraswamy distribution especially for estimating the β *shape* parameter by using the Maximum *Likelihood* (ML) and Bayes method. In the Bayes method, two loss functions will be used, the Square Error Loss Function (SELF) and the Precautionary Loss Function (PLF). In addition to the Bayes method, it is necessary to know the prior distribution based on the recommendation from [2].

Next, there will be a simulation data generated by Mathematica software, then using the formulas from both methods we will get a recommendation for any hydrological case. The recommendation will be used for Shasta Reservoir Capacity Data so that we know the characteristics from data or the distribution that can describe it.

THEORETICAL ANALYSIS

Kumaraswamy Distribution

Beginning with the observation of various hydrological data, Kumaraswamy [1] tried to establish a distribution function which can represent the empirical distribution of the data. Generally, there are four forms of the empirical distribution graph of hydrological data as shown in Fig. 1. Suppose the random variable Z at any hydrological data with the minimum point z_{min} and the maximum point z_{max}. Note Fig. 1 we can see that the distribution function is a combination of the two functions, namely F_0 and $(1 - F_0)P(z)$ so that it can be

International Symposium on Current Progress in Mathematics and Sciences 2016 (ISCPMS 2016)
AIP Conf. Proc. 1862, 030160-1–030160-8; doi: 10.1063/1.4991264
Published by AIP Publishing. 978-0-7354-1536-2/$30.00

written $F(z) = F_0 + (1 - F_0)P(z)$, where $P(z)$ is a general function will be found from the four graphs of empirical distribution.

First Empirical Distribution Graph. Note that the form of the first function is a function rises linearly starting from (z_{min}, F_0) to $(z_{max}, 1)$. By utilizing the formation of a linear function, we can find the function of first empirical distribution, that is

$$F(z) = F_0 + (1 - F_0)x, \quad P(z) = x = \frac{z - z_{min}}{z_{max} - z_{min}}, \quad 0 \le x \le 1 \tag{1}$$

Second, Third, and Fourth Empirical Distribution Graph. By keeping into consideration the conclusions $P(z)$ in the first form function as in Equation 1, there are three possibilities for common function $P(z)$ as follows.

- The first possibility: $P(z) = x^\alpha = \left(1 - (1 - x^\alpha)\right)$
- The second possibility: $P(z) = (1 - (1 - x)^\alpha)$
- The third possibility: $P(z) = \left(1 - (1 - x^\alpha)^\beta\right)$

When the first or second possibility is used then it can only describe the first, second, and third form of empirical distribution graph. While the third possibility is already encapsulate the first and second possibilities and can describe the shape of the fourth graph empirical distribution. It can be concluded that the third possibility is the right choice to describe the four forms of empirical distribution graph of hydrological data.
Hence, the distribution function is,

$$F(z) = F_0 + (1 - F_0)\left[1 - (1 - x^\alpha)^\beta\right], \quad 0 \le x \le 1,\ \alpha > 0,\ \beta > 0 \tag{2}$$

while the probability density function is

$$f(z) = \begin{cases} F_0, \ x = 0 \\ (1 - F_0)\left[\alpha\beta x^{\alpha-1}(1 - x^\alpha)^{\beta-1}\right], \ 0 < x < 1 \end{cases} \tag{3}$$

which became known as Kumaraswamy CDF and PDF. F_0 is a probability for the smallest observation (z_{min}) occurs. α and β are shape parameters which cause PDF or CDF variates.

Reservoir

The reservoir is an artificial or natural area used to store water [3]. The reservoir has a capacity of water with the lower limit is zero and the upper limit is the maximum capacity of the reservoir. Generally, the volume of stored water at a certain time never reached the lower limit, or it can be said the reservoir is never empty. Therefore, the discussion of the data capacity of the reservoir is always assumed $F_0 = 0$. This assumption will be used next.

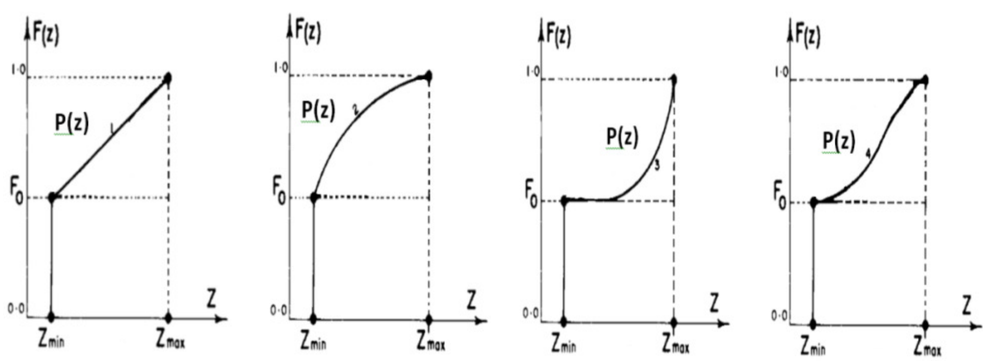

FIGURE 1. Empirical Distribution Graphs of Hydrological Data

ML Estimation

Suppose a random sample of size n drawn from a population that follows the Kumaraswamy distribution, namely $X_1, X_2, \dots, X_n$. P.d.f for X_i is $f(x_i|\alpha,\beta) = \Pr(X_i = x_i|\alpha,\beta) = \alpha\beta {x_i}^{\alpha-1}(1-{x_i}^{\alpha})^{\beta-1},\ 0 < x_i < 1.$ The likelihood function obtained $L(\alpha,\beta) = \alpha^n\beta^n \prod_{i=1}^{n} {x_i}^{\alpha} \prod_{i=1}^{n}(1-{x_i}^{\alpha})^{\beta-1}$.

Because of the shape of the likelihood function ln provide maximum value equal to the likelihood function, then to facilitate the calculation, use the form of the likelihood function ln is [4],

$$\ln[L(\alpha,\beta)] = n\ln\alpha + n\ln\beta + \alpha\sum_{i=1}^{n}\ln x_i + (\beta-1)\sum_{i=1}^{n}\ln(1-{x_i}^{\alpha})$$

To obtain β estimation, we will look for the value of β that maximizes the likelihood function by lowering $\ln[L(\alpha,\beta)]$ against β and equate the derivative equation with zero, obtained

$$\hat{\beta}_{ML} = -\frac{n}{\sum_{i=1}^{n}\ln(1-{x_i}^{\alpha})} \tag{4}$$

which $\hat{\beta}_{ML}$ is the estimated point β from Maximum Likelihood method (ML). In addition, the α estimation can be obtained by solving non-linear equation $\frac{n}{\alpha} + \sum_{i=1}^{n}\ln x_i - (\beta-1)\sum_{i=1}^{n}\frac{{x_i}^{\alpha}\ln x_i}{1-{x_i}^{\alpha}} = 0$ numerically using Mathematica software.

Bayes Estimation

The initial stage needs to be done is to determine the prior distribution [4]. Prior Gamma distribution is used as a conjugate prior distribution for β [2]. Suppose β random variable with following prior density function ;

$$P(\beta) = \frac{\beta^{\delta-1}}{\Gamma(\delta)b^{\delta}}e^{-\frac{\beta}{b}}, \quad b > 0, \delta > 0 \text{ or it can be written } P(\beta) \propto \beta^{\delta-1}e^{-\frac{\beta}{b}}.$$

Note that β is Gamma distribution with parameter δ and b. By using the method of moments will be obtained parameter values $\delta = \frac{\bar{X}^2}{\sum_{i=1}^{n}\frac{{x_i}^2}{n}-\bar{X}^2}$ and $b = \frac{\sum_{i=1}^{n}\frac{{x_i}^2}{n}-\bar{X}^2}{\bar{X}}$.

Posterior Distribution

Gamma prior distribution and the likelihood function that have been acquired are combined to generate a posterior distribution using Bayes theorem and obtained

$$\mathrm{P}(\beta|x) = \frac{\beta^{\delta+n-1}\exp\left(-\beta\left[\frac{1}{b}-\sum_{i=1}^{n}\ln(1-{x_i}^{\alpha})\right]\right)}{\frac{\Gamma(\delta+n)}{\left(\frac{1}{b}-\sum_{i=1}^{n}\ln(1-{x_i}^{\alpha})\right)^{\delta+n}}} \tag{5}$$

Point Estimation

Would be used the following definition;
$\hat{\beta}_{SELF}$: β point estimation of Bayes method using SELF,
$\hat{\beta}_{PLF}$: β point estimation of Bayes method using PLF.

Point Estimation by SELF. Based on the definition of Square Error Loss Function (SELF) obtained [5],

$$L(\beta,\hat{\beta}_{SELF}) = (\beta-\hat{\beta}_{SELF})^2$$

Based on the definition of posterior risk obtained

$$RP(\hat{\beta}_{SELF}) = E[L(\beta,\hat{\beta}_{SELF})] = E[(\beta-\hat{\beta}_{SELF})^2]$$

$$RP(\hat{\beta}_{SELF}) = \int_0^{\infty} \beta^2 P(\beta|x)\, d\beta - 2\hat{\beta}_{SELF} \int_0^{\infty} \beta\, P(\beta|x) d\beta + \hat{\beta}_{SELF}{}^2$$

Point estimation of loss function is a value estimation that minimizes the posterior risk [6]. Therefore, the value of $\hat{\beta}_{SELF}$ will be found to minimize $RP(\hat{\beta}_{SELF})$ by using the value of the first derivation $RP(\hat{\beta}_{SELF})$ equal to zero and obtained

$$\hat{\beta}_{SELF} = \int_0^{\infty} \beta\, P(\beta|x) d\beta = E[\beta|x] = \frac{\delta+n}{\frac{1}{b} - \sum_{i=1}^{n} \ln(1-x_i{}^{\alpha})} \tag{6}$$

Point Estimation by PLF. Based on the definition of the Precautionary Loss Function (PLF) obtained [5]

$$L(\beta, \hat{\beta}_{PLF}) = \frac{(\beta-\hat{\beta}_{PLF})^2}{\hat{\beta}_{PLF}}$$

Based on the definition of posterior risk obtained

$$RP(\hat{\beta}_{PLF}) = E[L(\beta, \hat{\beta}_{PLF})] = E\left[\frac{(\beta-\hat{\beta}_{PLF})^2}{\hat{\beta}_{PLF}}\right]$$

$$RP(\hat{\beta}_{PLF}) = \frac{1}{\hat{\beta}_{PLF}} \int_0^{\infty} \beta^2 P(\beta|x) d\beta - 2\int_0^{\infty} \beta\, P(\beta|x) d\beta + \hat{\beta}_{PLF}$$

Point estimation of loss function is a value estimation that minimizes the posterior risk [6]. Therefore, the value of $\hat{\beta}_{PLF}$ will be found to minimize $RP(\hat{\beta}_{PLF})$ by using the value of the first derivation $RP(\hat{\beta}_{PLF})$ equal to zero and obtained

$$\hat{\beta}_{PLF} = \left[\int_0^{\infty} \beta^2 P(\beta|x) d\beta\right]^{1/2} = \left[E[\beta^2|x]\right]^{1/2} = \frac{\sqrt{(\delta+n+1)(\delta+n)}}{\frac{1}{b} - \sum_{i=1}^{n} \ln(1-x_i{}^{\alpha})} \tag{7}$$

Posterior Risk

The previous estimation has been obtained by Bayes method using Square Error Loss Function (SELF) and Precautionary Loss Function (PLF). Next, we will find posterior risk from both of point estimations.

Posterior Risk of SELF Point Estimation. We have obtained SELF posterior risk, that is

$$RP(\hat{\beta}_{SELF}) = \int_0^{\infty} \beta^2 P(\beta|x)\, d\beta - 2\hat{\beta}_{SELF} \int_0^{\infty} \beta\, P(\beta|x) d\beta + \hat{\beta}_{SELF}{}^2$$

By substitute SELF point estimation, we will get

$$RP(\hat{\beta}_{SELF}) = \frac{\delta+n}{\left(\frac{1}{b} - \sum_{i=1}^{n} \ln(1-x_i{}^{\alpha})\right)^2} \tag{8}$$

Posterior Risk of PLF Point Estimation. We have obtained PLF posterior risk, that is

$$RP(\hat{\beta}_{PLF}) = \frac{1}{\hat{\beta}_{PLF}} \int_0^{\infty} \beta^2 P(\beta|x) d\beta - 2\int_0^{\infty} \beta\, P(\beta|x) d\beta + \hat{\beta}_{PLF}$$

By substitute SELF point estimation, we will get

$$RP(\hat{\beta}_{PLF}) = 2\left[\frac{[(\delta+n+1)(\delta+n)]^{1/2} - (\delta+n)}{\frac{1}{b} - \sum_{i=1}^{n} \ln(1-x_i{}^{\alpha})}\right] \tag{9}$$

Data Simulation

First, we will compare β shape parameter estimation by Maximum Likelihood (ML) method and Bayes method based on Mean Square Error (MSE). The smaller MSE value, the better estimation will be [7]. Then, we will compare β shape parameter estimation obtained by Bayes method using Square Error Loss Function (SELF) and Precautionary Loss Function (PLF) based on Posterior Risk. The smaller Posterior Risk value, the better estimation will be [6]. In this research, we use $\alpha = 0.25, 0.5, 1, 4$ and $\beta = 0.2, 0.4, 1, 3$ for sample size $n = 25, 50, 100, 200$. For MSE simulation, it will be resurrected 1000 samples using Mathematica software and the conclusion is obtained as in Table 1. As a conclusion to Posterior Risk simulation can be seen in Table 2.

TABLE 1. Conclusion for Mean Square Error (MSE) Simulation

α	β	Best Method
0.25	0.2	ML
	0.4	SELF
	0.6	SELF
	1	PLF
	3	ML
0.5	0.2	ML
	0.4	ML & SELF
	0.6	SELF
	1	PLF
	3	ML
0.75	0.2	ML
	0.4	SELF
	0.6	SELF
	1	PLF
	3	ML
1	0.2	ML
	0.4	ML
	0.6	SELF
	1	PLF
	3	ML
4	0.2	ML
	0.4	ML
	0.6	ML
	1	PLF
	3	ML

TABLE 2. Conclusion for Posterior Risk Simulation

α	β	Best Method
0.25	0.2	SELF
	0.4	SELF
	0.6	SELF
	1	SELF & PLF
	3	PLF
0.5	0.2	SELF
	0.4	SELF
	0.6	SELF
	1	SELF & PLF
	3	PLF
0.75	0.2	SELF
	0.4	SELF
	0.6	SELF
	1	SELF & PLF
	3	PLF
1	0.2	SELF
	0.4	SELF
	0.6	SELF
	1	SELF & PLF
	3	PLF
4	0.2	SELF
	0.4	SELF
	0.6	SELF
	1	SELF & PLF
	3	PLF

APPLICATION

We will find β shape parameter estimation of Kumaraswamy distribution using Maximum Likelihood (ML) and Bayesian methods based on Shasta Reservoir capacity data. The reservoir is located in California, United States $(40^o 43' 07'' \mathrm{U}, 122^o 25' 08'' \mathrm{B})$. The reservoir has a height of 602 ft (183 m), a length of 3460 ft (1050 m), and a total capacity of 4.552 million acre-ft (5.615 million dam3). Shasta Reservoir capacity data can be obtained in [8]. The 20 observations of reservoir capacity before and after transformation is presented in Table 3.

Before finding the estimation, we need to set the value of α parameter. This can be obtained by observing the data histogram and Kumaraswamy PDF variation in Fig. 2. It can be seen that the shape of the histogram is a type I which can be described by Kumaraswamy distribution with parameter $\alpha > 1$ and $\beta > 1$. Therefore, we will continue the estimation research by using,

$$\alpha = 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 10.5$$

Based on the results obtained in Table 4, it can be seen that β estimation by Bayes method reaches a value greater than one starting when α is set greater than 3.5. Meanwhile, when α determined by ML method obtained, $\hat{\alpha}_{ML} = 6.35 \approx 6.5$.

Due to only comparing β parameter, α needs to be established without any relation with value defined by other methods. Therefore in determining α value, authors chose criteria $\alpha = \frac{6.5-3.5}{2} = 5$ and obtained three models as follows.

TABLE 3. Shasta Reservoir Capacity Data Each February from 1991 to 2010

Year	Capacity	Transformed Capacity	Year	Capacity	Transformed Capacity
1991	1542838	0.338936	**2001**	3495969	0.768007
1992	1966077	0.431915	**2002**	3839544	0.843485
1993	3459209	0.759932	**2003**	3584283	0.787408
1994	3298496	0.724626	**2004**	3868600	0.849868
1995	3448519	0.757583	**2005**	3168056	0.69597
1996	3694201	0.811556	**2006**	3834224	0.842316
1997	3574861	0.785339	**2007**	3772193	0.828689
1998	3567220	0.78366	**2008**	2641041	0.580194
1999	3712733	0.815627	**2009**	1960458	0.430681
2000	3857423	0.847413	**2010**	3380147	0.742563

TABLE 4. Estimation and Posterior Risk of β Shape Parameter of Kumaraswamy Distribution Based on Shasta Reservoir Capacity Data Each February from 1991 to 2010

α	$\hat{\beta}_{ML}$	$\hat{\beta}_{SELF}$	$\hat{\beta}_{PLF}$	RP($\hat{\beta}_{SELF}$)	RP($\hat{\beta}_{PLF}$)
1.5	0.9415	0.80827	0.817514	0.015029	0.009244
2	1.1769	0.877605	0.887642	0.017718	0.010037
2.5	1.43215	0.934765	0.945456	0.020101	0.010691
3	1.71143	0.982934	0.994175	0.022226	0.011242
3.5	2.01846	1.0241	1.03581	0.024127	0.011713
4	2.35693	1.05962	1.07174	0.02583	0.012119
4.5	2.73068	1.09049	1.10296	0.027357	0.012472
5	**3.14383**	**1.11748**	**1.13026**	**0.028727**	**0.012781**
5.5	3.60082	1.14116	1.15421	0.029958	0.013052
6	4.10647	1.16202	1.17531	0.031064	0.01329
6.5	4.66602	1.18045	1.19396	0.032057	0.013501
7	5.28521	1.19678	1.21046	0.032949	0.013688
7.5	5.97029	1.21126	1.22511	0.033751	0.013853
8	6.7281	1.22413	1.23813	0.034472	0.014
8.5	7.56612	1.23558	1.24971	0.035121	0.014131
9	8.49251	1.24579	1.26004	0.035704	0.014248
9.5	9.51621	1.2549	1.26926	0.036228	0.014352
10	10.647	1.26304	1.27749	0.036699	0.014446
10.5	11.8956	1.27032	1.28485	0.037123	0.014529

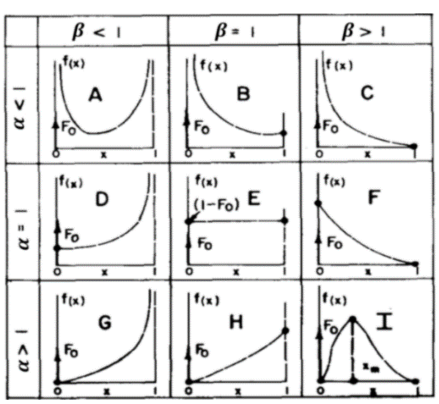

FIGURE 2. Variation of Kumaraswamy PDF

1st Model Kumaraswamy Distribution with $\alpha = 5$ *and* $\hat{\beta}_{ML} = 3.14383$

PDF obtained $f(x) = 15.71915x^4(1-x^5)^{2.14383}, \quad 0 < x < 1$

CDF obtained $F(x) = 1-(1-x^5)^{3.14383}, \quad 0 < x < 1$

The representative graphic for PDF and CDF can be seen in Fig. 3 and Fig. 4.

2nd Model Kumaraswamy Distribution with $\alpha = 5$ *and* $\hat{\beta}_{SELF} = 1.11748$

PDF obtained $f(x) = 5.5874x^4(1-x^5)^{1.11748}, \quad 0 < x < 1$

CDF obtained $F(x) = 1-(1-x^5)^{1.11748}, \quad 0 < x < 1$

The representative graphic for PDF and CDF can be seen in Fig. 5 and Fig. 6.

3rd Model Kumaraswamy Distribution with $\alpha = 5$ *and* $\hat{\beta}_{PLF} = 1.13026$

PDF obtained $f(x) = 5.6513x^4(1-x^5)^{1.13026}, \quad 0 < x < 1$

CDF obtained $F(x) = 1-(1-x^5)^{1.13026}, \quad 0 < x < 1$

The representative graphic for PDF and CDF can be seen in Fig. 7 and Fig. 8.

From the graph, it can be seen that from the three models, 1st Model has a distribution function which can adjust the function of data empirical distribution. Furthermore using Kolmogorov-Smirnov test, we also get the same conclusion that the best model represents Shasta Reservoir capacity data is 1st Model.

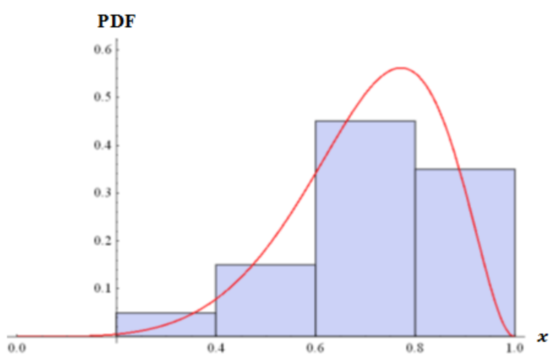

FIGURE 3. Kumaraswamy PDF 1st Model

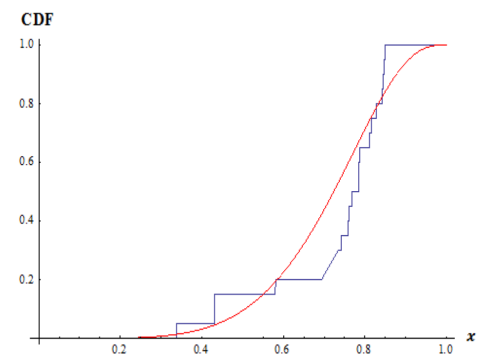

FIGURE 4. Kumaraswamy CDF 1st Model

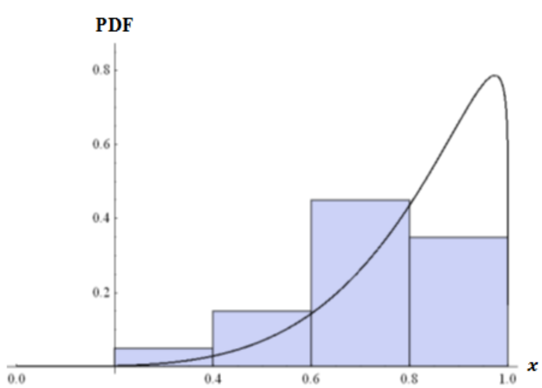

FIGURE 5. Kumaraswamy PDF 2nd Model

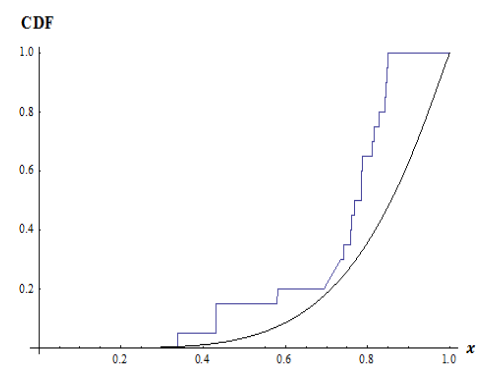

FIGURE 6. Kumaraswamy CDF 2nd Model

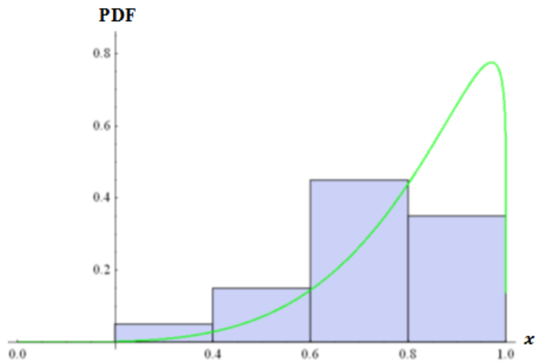

FIGURE 7. Kumaraswamy PDF 3rd Model

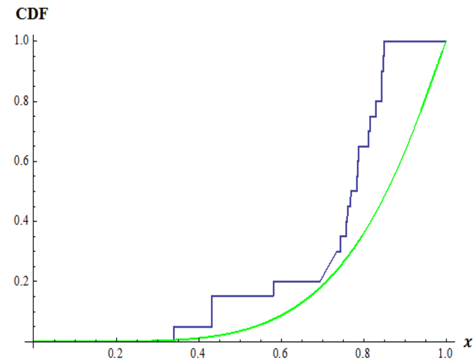

FIGURE 8. Kumaraswamy CDF 3rd Model

CONCLUSIONS

β shape parameter estimated by the Kumaraswamy distribution and obtained by using the Maximum Likelihood (ML) method only relies on information from the data,whereas the β shape parameter estimation of Kumaraswamy distribution obtained by using the Bayes method depends on a combination of data and β shape parameter information. To determine the best estimation method we use the Mean Square Error (MSE) on each of the estimation obtained from both methods. The estimation is called good if MSE is the smallest one. For comparing the estimation obtained from Bayes method one uses the Posterior Risk based on Square Error Loss Function (SELF) and Precautionary Loss Function (PLF). Estimation is called good when the Posterior Risk becomes the smallest.

REFERENCES

1. P. Kumaraswamy, J. Hydrol. **46**, 79-88 (1980).
2. E. K. Al-Hussaini and Z. F. Jaheen, Commun Statist-Theory Meth **24**, 1829-1842 (1995).
3. World Commission on Dams, *Dams and Development: A New Framework for Decision-Making* (2000), available at http://www.rivernet.org/general/wcd/wcd_overview_english.pdf
4. R. V. Hogg, J. W. McKean, and A. T. Craig, *Introduction to Mathematical Statistics, 7th ed* (Prentice Hall, New Jersey, 2012).
5. S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models: From Data to Decisions 4th ed* (John Willey & Sons, Hoboken, New Jersey, 2012).
6. N. Balakrishnan, E. Castillo, and J. M. Sarabia, *Advances in Distribution Theory, Order Statistics, and Inference* (Birkhäuser, New York, 2006).
7. A. S. Cohen and M. T. Kane, and S. H. Kim, Appl. Psychol. Meas. **25**, 136-145 (2010).
8. California Data Exchange Center (2016), available at http://cdec.water.ca.gov/