PROCEEDINGS OF THE LATVIAN ACADEMY OF SCIENCES. Section B, Vol. 72 (2018), No. 3 (714), pp. 184–192.
DOI: 10.2478/prolas-2018-0005

DE GRUYTER OPEN

Research Methods

MODERN TECHNIQUES IN DATA ANALYSIS, WITH APPLICATION TO THE WATER POLLUTION

Haroon M. Barakat[1], Osama Mohareb Khaled[2,#], and N. Khalil Rakha[3]

[1] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt

[2] Department of Mathematics, Faculty of Science, Port Said University, Port Said, Egypt

[3] Department of Physics and Engineering Mathematics, Faculty of Engineering, Port Said University, Port Said, Egypt

[#] Corresponding author, osam87@yahoo.com

Communicated by Aleksandrs Šostaks

This paper presents a comparison of most capable families of distributions for modelling asymmetry. Kum-normal, stable-symmetric normal family and two of the full families were chosen, where the quality of the fit, the flexibility and the amount of asymmetry parameters were factors used for comparison. The objective of this study was to generate data with increasing levels of asymmetry and to choose the best fit. The distributions were also compared in modelling two data sets of pollution of the drinking water in the El-Sharkia governorate in Egypt. Much of this paper is concerned with the distribution theory, exploring the properties of some new recent families of distributions and, where appropriate, extolling their virtues. Relatively, much of this paper is devoted to practical application.

AMS 2010 Subject Classification: *62-07; 62E10; 62F99.*

Key words: *mixture distribution, full family of distributions, data modelling, skewness, kurtosis, drinking-water quality.*

INTRODUCTION

Statistical distributions are commonly applied to describe real world phenomena. In statistics, the normal distribution is the most popular model in applications to real data. When the number of observations is large, it can serve as an approximate distribution for other models. Nevertheless, if the data is asymmetric, normal distribution will not be a good choice. Therefore, the interest in developing more flexible statistical distributions remains strong in the statistics profession. Many generalised families of distributions have been developed and applied to describe various phenomena. A common feature of these generalised distributions is that they have more shape parameters. In various situations it is useful to deal with a generalised family which allows a "continuous" variation from normality to non-normality. Among the generalised families of distributions are Azzalini's skew normal distribution (due to Azzalini, 1985), kum-normal distribution (due to Cordeiro and Castro, 2011), stable-symmetric normal distribution (due to Barakat, 2015) and the full families (defined by Barakat and Khaled, 2017). It is notable that the last mentioned families of distributions own the property of "strict inclusion" of the normal distribution and are considered to be mathematical tractable at the same time.

For the data modelling purpose, we note that the kurtosis and the skewness uniquely determine the type of many distribution functions (df's). Therefore, practically, any df should belong to one and only one of the following nine types: (1) symmetric and mesokurtic, denoted by 00, (2) symmetric and leptokurtic (positive excess kurtosis or the df has a more acute peak around the mean and fatter tails than normal df), denoted by 0+, (3) symmetric and platykurtic (negative excess kurtosis or the df has a lower, wider peak around the mean and thinner tails), denoted by 0–, (4) positive symmetric and mesokurtic, denoted by +0, (5) positive symmetric and leptokurtic, denoted by ++, (6) positive symmetric and platykurtic, denoted by +–, (7) negative symmetric and mesokurtic, denoted by –0, (8) negative symmetric and leptokurtic, denoted by –+, and (9) negative symmetric and platykurtic, denoted by —. Consequently, the number of possible types of df' is the crucial factor of the efficiency of any family in describing many different real data. In other words, we expect that the capability of any family for describing several real data of different statistical types increases as the number of possible types of the df's increases. Barakat and Khaled (2017) called the family which contains the nine possible types of df's a full family. On the other hand, since the normal df is the only known df, which is of 00 type, then it is naturally to take it as the base. Therefore,

the following three properties are essential for any expanded family to be used in the modelling data:

1. The resulting family is mathematically tractable, i.e. we can derive its distributional characteristics, e.g., the moments, median, mode, skewness and kurtosis. Moreover, via the added parameters, we can control some of these characteristics.

2. Strict inclusion of the normal df has to be used.

3. The family should include the largest possible number of the types 00, 0+, ..., ++, +– of distributions.

4. The wide range of the indices of skewness and kurtosis has to be used (the Pearson coefficients of skewness, γ_1, and kurtosis, γ_2 are $-\infty \le \gamma_1 \le +\infty$ and $\gamma_2 \ge 0$).

In the light of the above four features, the Azzalini's skew-normal distribution contains at most {00, 0+, +0, ++, –0, –+} types of df's. The two parameter Kum distribution was created by Kumaraswamy (1980), in applications in hydrology. The two parameter Kum-normal distribution (denoted by KumN) is defined by

$$\text{KumN}(x;\, a,\, b,\, \mu,\, \sigma) = 1-\left(1-\Phi^{a}\left(\frac{x-\mu}{\sigma}\right)\right)^{b},\ a,\, b>0, \quad (1.1)$$

where $a > 0$ and $b > 0$ are the shape parameters, while $\mu \in \Re$ (the symbol $\Re$ for the real number line) is the location parameter (mean), $\sigma > 0$ is the scale parameter (standard deviation) and $\Phi(.)$ is the standard normal distribution. Michelle *et al*. (2012) presented a comparison of the Azzalini's skew normal distribution and the KumN distribution. The quality of the fit, the flexibility and the amount of asymmetry parameters were factors used for this comparison. The comparison revealed that the kumN distribution proved to be effective in adjusting asymmetric data as much as the Azzalini's skew normal distribution. As the level of asymmetry increases, the kumN distribution shows better fitting than the Azzalini's skew normal distribution. For the real data, both families had the same fitting quality. Moreover, this study showed that although the kumN distribution presents a better fit, there are limitations on the shape of both families.

Barakat (2015) defined the stable symmetric family of df's as a family that contains the reverse of every df that belongs to it. The stable-symmetric normal distribution (SSN) is defined by $\text{SSN}(x;\, \alpha,\, c) = \alpha\Phi(x+c)+\overline{\alpha}\Phi(x-c)$ where $0 \le \overline{\alpha} = 1-\alpha \le 1$, $c \in \Re$. Barakat (2015) showed that this family possesses a remarkable wide range of the indices of skewness and kurtosis. Besides, it contains all the possible types of df's, except 0+. Moreover, the SSN family is more tractable than the Azzalini's skew normal and the KumN distributions, since the SSN family is no more than being a linear combination of the normal df's. Therefore, this family has the same degree of complexity of the normal df itself. Barakat (2015) fitted the location-scale SS-normal family $\text{SSN}(x;\, \alpha,\, c;\, \mu,\, \sigma) = \text{SSN}\left(\frac{x-\mu}{\sigma}, \alpha, c\right)$, $\sigma > 0$ or equivalenty

$$\text{SSN}(x,\, \alpha,\, c;\, \mu,\, \sigma) = \alpha\Phi\left(\frac{x-\mu_1}{\sigma}\right)+\overline{\alpha}\Phi\left(\frac{x-\mu_2}{\sigma}\right), \quad (1.2)$$

where $\mu_1 = \mu - c\sigma$ and $\mu_2 = \mu + c\sigma$, to a real data set (Example 3.1). Barakat (2015) compared the fitness with many other distributions. The obtained result indicated that the SSN fitted the best among all the distributions, which have been considered.

Barakat and Khaled (2017) suggested a method for constructing tractable full families via the mixture of the SSN family and any tractable symmetric leptokurtic df, which is called complementary df. Actually, given three parameters $0 \le \alpha, \beta \le 1$ and $c \in \Re$ the full family of df's, denoted by $\text{FN}(x:\alpha,\beta,c)$ is given by $\mathcal{L}$ L

$$\text{FN}(x:\alpha,\beta,c,||\,\mathcal{L}) = \beta[\alpha\Phi(x+c)+\overline{\alpha}\Phi(x-c)]+\overline{\beta}\,\mathcal{L}(x)$$

where $\mathcal{L}(x)$ is a complementary df. Barakat and Khaled (2017) suggested that both the logistic and the Laplace df's are two complementary df's, which are defined respectively by

$$\mathcal{L}_1(x) = \frac{1}{1+e^{-x}},\ x \in (-\infty, \infty)$$

and

$$\mathcal{L}_2(x) = \begin{cases} \dfrac{1}{2}e^{x}, & \text{if} \quad x < 0, \\ 1-\dfrac{1}{2}e^{-x}, & \text{if} \quad x \ge 0. \end{cases}$$

The coefficients of excess kurtosis of the logistic and Laplace df's are 1.5 and 3, respectively. Clearly, each of the full families $\text{FNi}(x:\alpha,\beta,c|\,\mathcal{L}_i)$, $i = 1,2$, is more tractable than the Azzalini's skew normal and the KumN distributions. Moreover, each of them possesses a very wide range of skewness and kurtosis indices. Actually, this range of indices is at least not less than the range of the indices of the SSN family. This is because the SSN family is contained in both, $\text{FN1}(x:\alpha,\beta,c|\,\mathcal{L}_1)$ and $\text{FN2}(x:\alpha,\beta,c|\,\mathcal{L}_2)$. Finally, each of the suggested full families contains all the possible types of df's. Therefore, the full family of df's provides a general and flexible mechanism for fitting a wide spectrum of real world dataset. The location-scale FNi family: $\text{FNi}(x:\alpha,\beta,c;\mu,\sigma|\,\mathcal{L}_i)$, $\sigma > 0$, $i = 1, 2$, may be written in the form:

$$\text{FNi}(x:\alpha,\beta,c;\mu,\sigma|\,\mathcal{L}_i) = \alpha_1\Phi\left(\frac{x-\mu_1}{\sigma}\right)+\alpha_2\Phi\left(\frac{x-\mu_2}{\sigma}\right)+$$

$$+\alpha_3\mathcal{L}_i\left(\frac{x-\mu_3}{\sigma}\right),\ i = 1, 2, \quad (1.3)$$

where $\alpha_1 = \beta\alpha.$; $\alpha_2 = \beta\overline{\alpha}$; $\alpha_3 = \overline{\beta}$; $\mu_1 = \mu - \mu\sigma$; $\mu_2 = \mu - \mu\sigma$ and $\mu_3 = \mu$. Barakat and Khaled (2017) confirmed the outperforming of the proposed full-families, by fitting the fam-

ily FNi($x:\alpha,\beta,c;\mu,\sigma,\mathcal{L}_1$), $\sigma > 0$, to a huge real data set, where the fitness was compared with the location-scale Azzalini family and the location-scale SSN family.

Clearly, each of the full families FN1 and FN2 has no theoretical advantage than the other (at least based on the aforesaid four essential properties). Thus, it seems that the only way to compare them is through carrying out a simulation study by generating data with increasing levels of asymmetry and choosing the best fit. Although both FN1 and FN2 outperform, the SSN family theoretically to a little extent, the full families have a greater number of shape parameters that earn the SSN some potential practical advantage. Actually, this advantage is based on the fact that any extra unknown parameter in the family will need to be estimated and this estimation is always accompanied with some random error. Therefore, we expect that as the number of the unknown parameters increases, the performance of the family becomes poor. Consequently, the main aim of this work was to verify the performance of SSN, FN1, FN2 and KumN families in modelling asymmetry, using criteria information on the quality of fit and likelihood ratio test (LRT).

For comparison, we used simulated data, generated from the distribution created by Tukey, called *g* and *h*, with increasing levels of asymmetry. All analyses were conducted using R. In this study we pursued the study implemented by Michelle *et al*. (2012) to compare the Azzalini's skew normal distribution and the KumN distribution. Since the study of Michelle *et al*. (2012) revealed that the performance of KumN family slightly surpasses the performance of the Azzalini's skew normal, we chose only the KumN family to be included in our study. Then, we present data from *g* and *h* distribution, tests for normality, information criteria and the LRT used to compare the four families. Finally, the four chosen families are compared in modelling two data sets of pollution of the drinking water in the El-Sharkia governorate in Egypt.

ASYMMETRY TESTS AND COMPARISON CRITERIA

Firstly, we generate several random samples with different levels of asymmetry from the *g* and *h* distribution, which has some facility in generating asymmetric data values. The *g* and *h* distribution was suggested by Tukey (1977) and discussed by Hoaglin and Peters (1979) and Hoaglin (1983). This distribution is defined by transforming the standard normal variable, Z to $T_{g,h}(Z) = \frac{\exp(gZ)-1}{g}\exp\left(\frac{h}{2}Z^2\right)$, where *g* is a real constant and *h* is a nonnegative real constant (where $h \in \Re$, see Martinez and Iglewicz, 1984). It can be shown that $T_{0,0}(Z) = Z$ (i.e. the distribution is symmetric when $g = 0$) and the distribution of $T_{0,h}(Z)$ is symmetric with increasingly heavy tails as h increases. That is, the distribution of $T_{g,h}(Z)$ has heavier tails than the normal for $h > 0$. Moreover, $T_{g,h}(Z) = \frac{\exp(gZ)-1}{g}$, which coincides with the location-scale log-normal distribution. The sign of g controls the direction of skewness but not its amount. Positive values of *g* skew the distribution to right tail while negative values of *g* skew the distribution to the left tail. In summary, $T_{h,g}(Z)$ introduces skewness through the factor involving g and elongation through the factor involving h. Finally, since the transformation $T_{h,g}(.)$ is one to one, the df of Tukey (g, h) random variable X can be written as $F_X(x) = \Phi(T_{h,g}^{-1}(x))$. Clearly, this distribution is not easily mathematically tractable, because it does not possess a simple expression for density.

We simulate eleven random samples from this distribution, each of size 100. Moreover, the generated values from the *g* and *h* distribution, which were split by the degree of asymmetry starting from the level –9.171929 (the corresponding sample called Scenario 1, and it is denoted by *r*1(–1,1), where this sample is generated for $g = -1$, $h = 1$) and ending by the level 9.147666 (also, the corresponding sample called Scenario 11 is denoted by *r*11(1,1) where this sample is generated for $g = h = 1$). To generate such data, we used the function rgh (random) for generating vectors containing 100 random values. Table 1 presents these scenarios, with the corresponding levels of asymmetry.

Secondly, a comparison was made between distributions through performing tests to verify the normality and symmetry to ensure that the simulated data from the *g* and *h* distribution was statistically asymmetric. For such verification and to confirm the data does not follow a normal distribution, we carried out the Shapiro-Wilk test (see, Ferreira, 2009). Also, to verify that the asymmetry parameter does not equal zero, we used the D'Agostino test (D'Agostino, 1978). Moreover, both the Shapiro-Wilk and the D'Agostino tests were implemented by using the R-package. Here the Shapiro-Wilk test (shapiro.test()) was a part of the basic package and the D'Agostinotest (agostino.test()) was a part of the package moments. To compare the fittings we used the LRT and to select the best fit we used Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criteria (BIC) (Schwarz, 1978). These criteria are based on the likelihood value of the model, the number of

Table 1

RESULTS OF THE NORMALITY TESTS AND ESTIMATES OF ASYMMETRY OF THE DATA SIMULATED FROM THE *g* AND *h* DISTRIBUTION

Scenario	Shapiro-Wilk *p*-value	D'Agostino *p*-value	Asymmetry level
r1(-1,1)	2.2e-16	2.2e-16	-9.171929
r2(-0.5,0.9)	2.2e-16	2.2e-16	-8.712417
r3(-1,0)	1.755e-07	1.473e-05	-1.192421
r4(0,0)	0.9824	0.9125	-0.02490355
r5(0,0.5)	0.6692	0.6236	0.2845326
r6(0.2,0.1)	0.005998	0.003184	0.7327836
r7(0.3,0)	1.98e-10	2.56e-7	1.554027
r8(0.4,0.8)	1.208e-10	2.811e-08	1.744873
r9(0.7,0.7)	2.2e-16	2.2e-16	4.020554
r10(1,0)	4.179e-15	2.2e-16	5.693755
r11(1,1)	2.2e-16	2.2e-16	9.147666

observations and the number of parameters thereof. Furthermore, the LRT (see, Casella and Berger, 2010) was performed to compare what distribution best fitted the simulated data, according to the degree of asymmetry. Table 1 shows that the *p*-values in Scenarios 4 and 5 for the Shapiro-Wilk and D'Agostino tests, are greater than 5%, i.e. the simulated data follow a normal distribution and therefore are symmetrical. For the other scenarios, the *p*-values are less than the level of significance for both tests, characterising asymmetric distributions. It is worth noting that the degree of asymmetry increases as the value of the parameter *g* increases. Moreover, the sign of *g* controls the direction of skewness.

Finally, Tables 2 and 3 present the results of information criteria AIC, BIC and the LRT for each scenario, for the families kumN, SSN, FN1 and FN2. The LRT in Tables 2 and 3, for all scenarios, confirms the equality of the fitting for the SSN and KumN. However, there is a favour for SSN with small AIC and BIC. In addition, we have the same result for the FN1 and FN2 families, as well as FN1 and KumN, with a favour for FN1. This is because FN1 has a smaller AIC and BIC. In all scenarios there are significant differences in fitting FN1 and SSN in favour of FN1. The same result holds for FN2 compared with the families SSN and kumN, except that in r2 , the LRT confirms the equality of the fitting for the FN2 and SSN with is a favouring of the SSN with small AIC and BIC. In summary, the performance of the FN1 family remarkably outperforms all other families in all considered scenarios, and in performance this family is followed by FN2, except in r2 where the family SSN surpasses the family FN2. The third family is SSN, while the last family is KumN.

Table 2

RESULTS OF THE AIC, BIC AND LRT FOR KumN, SSN, FN1 ANDFN2

Scenario	Family	AIC	BIC	LRT(p-value)		
				FN2	SSN	KumN
r1(-1,1)	FN1	525.75	543.99	1	2.63e-31	1
	FN2	938.70	943.12	-	2.93e-25	1
	SSN	828.87	844.50	-	-	1
	KumN	6914.03	6924.46	-	-	-
r2(-0.5,0.9)	FN1	290.01	308.24	1	7.34e-21	1
	FN2	926.75	931.18	-	1	2.03e-80
	SSN	200.23	215.87	-	-	1
	KumN	548.29	558.71	-	-	-
r3(-1,0)	FN1	329.51	347.75	1	6.90e-30	1
	FN2	357.24	361.66	-	5.96e-36	1
	SSN	198.54	214.18	-	-	1
	KumN	22618.32	22628.74	-	-	-
r4(0,0)	FN1	325.60	343.84	1	3.95e-29	1.33e-7
	FN2	958.48	962.90	-	6.03e-167	1
	SSN	198.10	213.73	-	-	1
	KumN	2897.78	2908.20	-	-	-
r5(0,0.5)	FN1	327.70	345.94	1	1.53e-29	1
	FN2	1024.89	1029.31	-	2.45e-181	5.44e-161
	SSN	198.32	213.95	-	-	1
	KumN	274.68	285.10	-	-	-

REAL DATA MODELLING

Water is indispensable to keep life. The availability of safe drinking water is a pivotal issue for all humans. If our bodies are not continuously supplied with water, they will become dehydrated, and the vital organs will deteriorate until death. Moreover, water acts as a purifier in our bodies. If enough water is not consumed, one would not be able to properly flush out the kidneys and/or the livers. In addition the colon would not be able to expel bowels properly and completely. This results in keeping unhealthy toxins in the body. However, according to medical experts, an individual needs to consume at least 2 liters of water daily for basic survival. Therefore, we should be able to find the best path of labour for protecting people against water-borne disease.

For all the aforementioned reasons, safe and clean drinking water is a human right. Therefore, the drinking water should be clear and free from bad tastes and smells. Above all it should be free from pathogens such as bacteria, viruses, protozoan parasites, and chemical pollutants to meet the biological and chemical standards. There is a large percentage of the World population in developing countries, especially

Table 3

RESULTS OF THE AIC, BIC AND LRT FOR KumN, SSN, FN1 AND FN2

Scenario	Family	AIC	BIC	LRT(p-value)		
				FN2	SSN	KumN
r6(0.2,0.1)	FN1	325.08	343.32	1	5.03e-29	1
	FN2	1046.29	1050.71	-	4.79e-186	1
	SSN	198.06	213.69	-	-	1
	KumN	1293.20	1303.61	-	-	-
r7(0.3,0)	FN1	326.66	344.89	1	2.50e-29	1
	FN2	1032.66	1037.08	-	4.83e-183	0.67
	SSN	198.24	213.88	-	-	1
	KumN	1019.51	1029.93	-	-	-
r8(0.4,0.8)	FN1	349.39	367.62	1	1.46e-32	1
	FN2	1122.23	1126.65	-	8.75e-201	1
	SSN	206.20	221.83	-	-	1
	KumN	664.66	675.08	-	-	-
r9(0.7,0.7)	FN1	329.86	348.10	1	7.31e-30	1
	FN2	1036.19	1040.61	-	1.21e-183	1
	SSN	199.01	214.64	-	-	1
	KumN	33656.04	33666.46	-	-	-
r10(1,0)	FN1	323.45	341.69	1	1.11e-28	1
	FN2	969.82	974.24	-	1.97e-169	1
	SSN	198.00	213.63	-	-	1
	KumN	4805.15	4815.57	-	-	-
r11(1,1)	FN1	332.52	350.76	1	3.16e-30	1
	FN2	1042.13	1046.54	-	1.02e-184	1
	SSN	200.01	215.64	-	-	1
	KumN	3069.40	3079.82	-	-	-

in villages not having access to safe drinking water and suffering from water-borne and related diseases such as diarrheal diseases.

In Egypt, the geographic location and the population development of the El-Sharkia governorate make it vulnerable to the problems caused by water pollutants. Actually, about 77.4% of people of El-Sharkia live in rural areas and about 18.3% of the inhabitants still use untreated groundwater drawn from shallow aquifers, by means of hand pumps, for drinking and other household purposes. These shallow aquifers easily become polluted with hazardous constituents and toxic chemicals.

We focus in this study on two pollutants: chlorides (Figs. 3–5) and sulphates (Figs. 6–9). A set of selected pollutants were measured to assess the quality of drinking water in the El-Sharkia governorate. The two data sets (each consisted of 202 measurements) for chlorides and sulphates concentration (mg/L) were collected by the Egyptian Water and Waste Water Regulatory Agency (EWRA) during two years (2009–2010).

The descriptive diagram and the summary statistics for these data sets for chlorides and sulphates concentration (mg/L) are given in Figures 1, 2 and Table 4, respectively. The estimates of the unknown parameters of the families

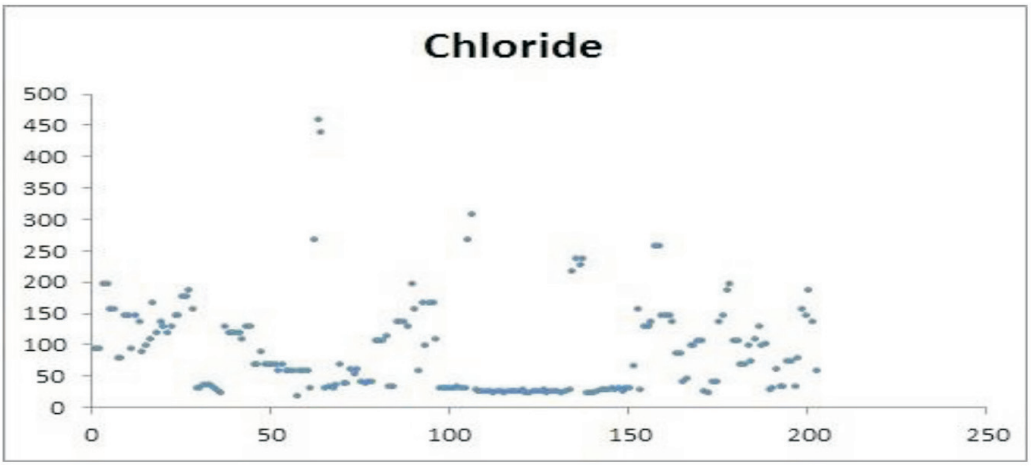

Fig. 1. Chemical chlorides in El-Sharkia drinking water.

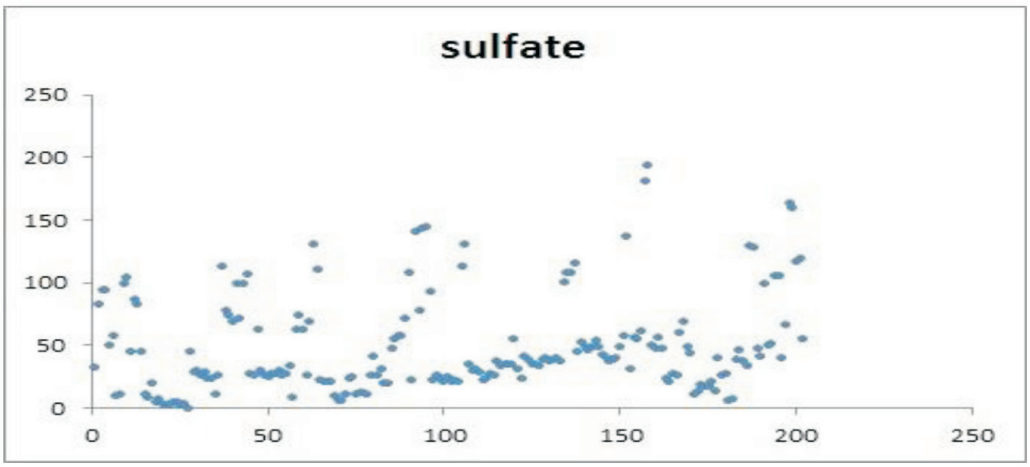

Fig. 2. Chemical sulphates in El-Sharkia drinking water.

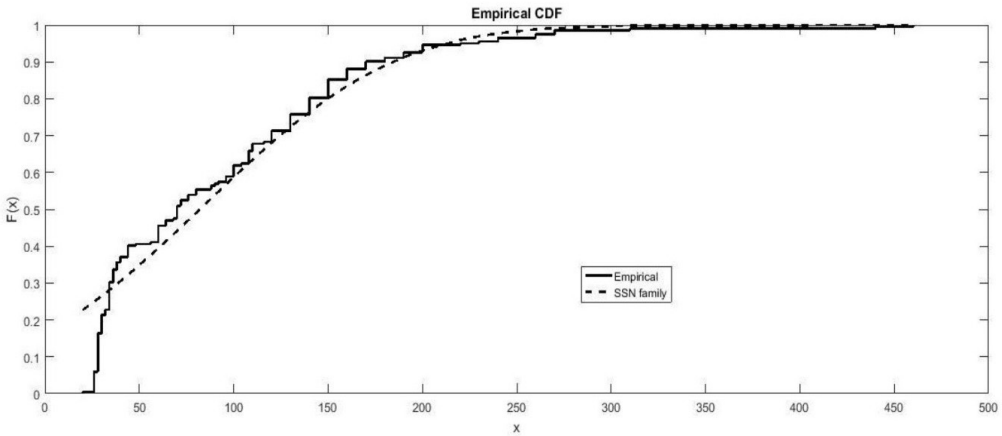

Fig. 3. SSN family for chlorides concentration (mg/L).

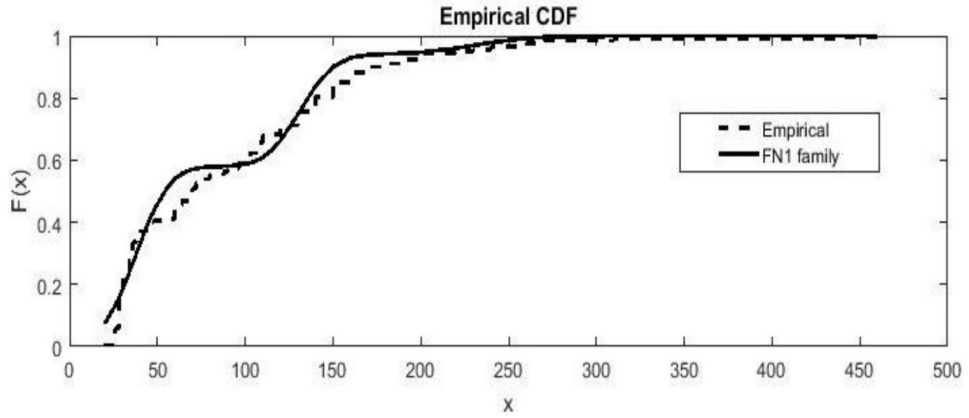

Fig. 4. FN1 family for chlorides concentration (mg/L).

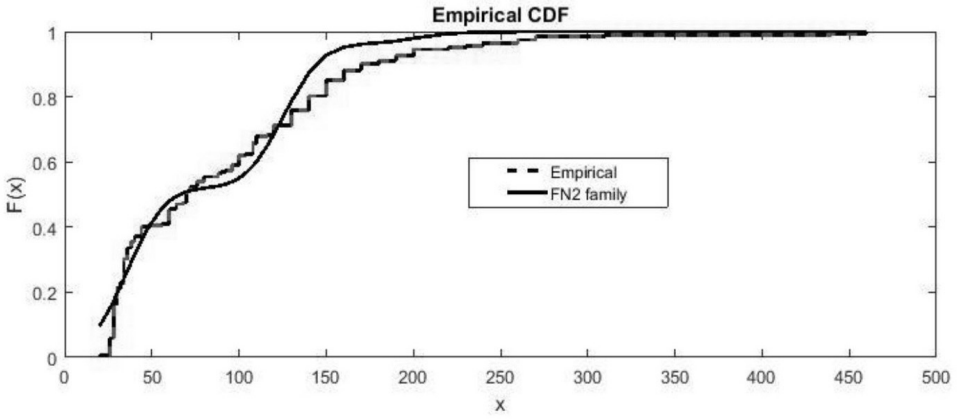

Fig. 5. FN2 family for chlorides concentration (mg/L).

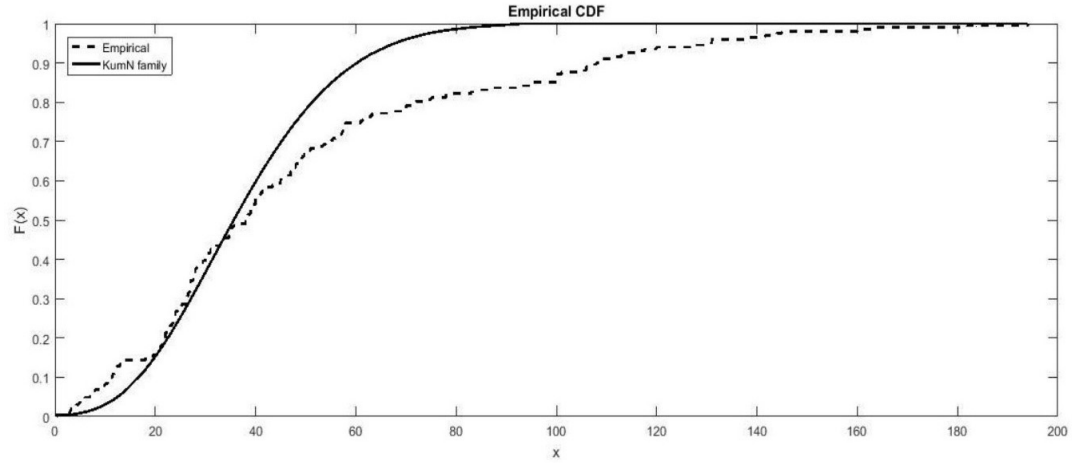

Fig. 6. KumN family for sulphates concentration (mg/L).

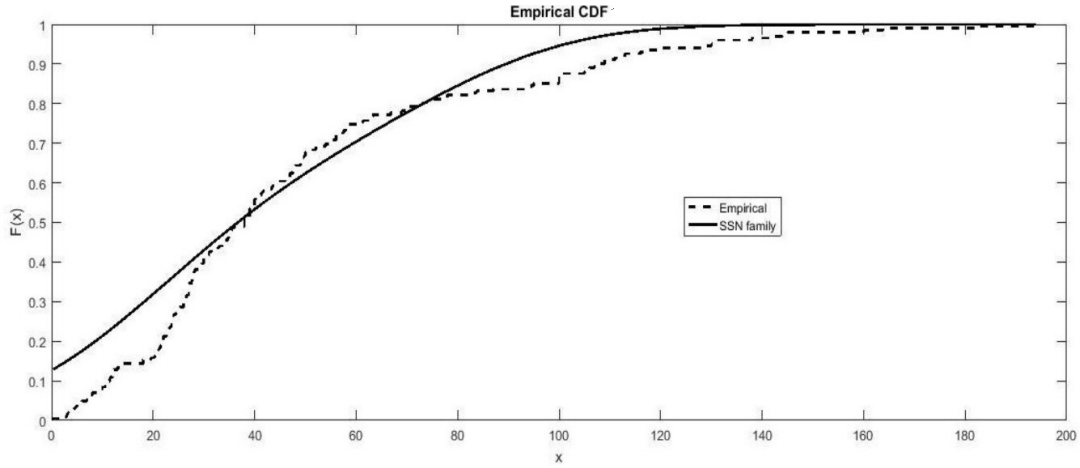

Fig. 7. SSN family for sulphates concentration (mg/L).

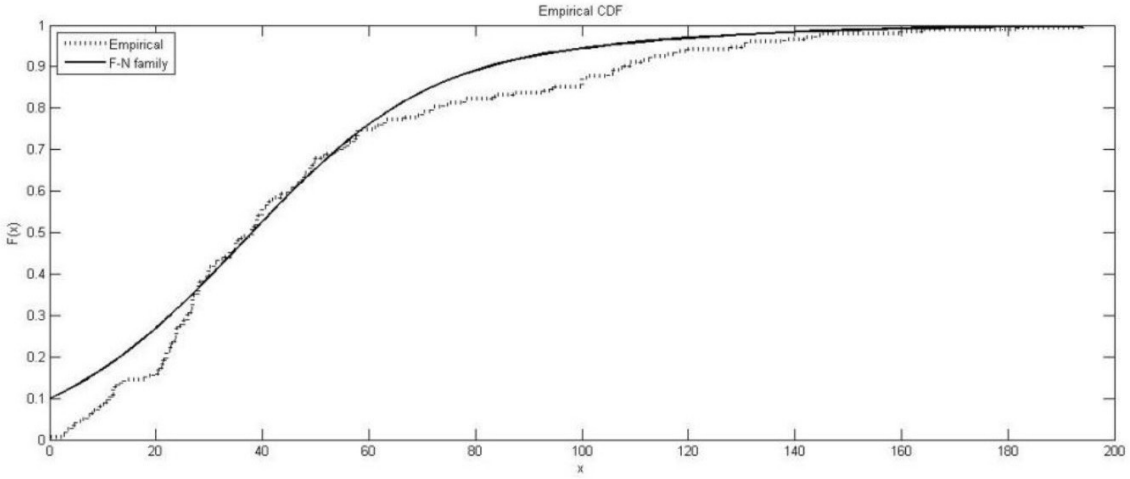

Fig. 8. FN1 family for sulphates concentration (mg/L).

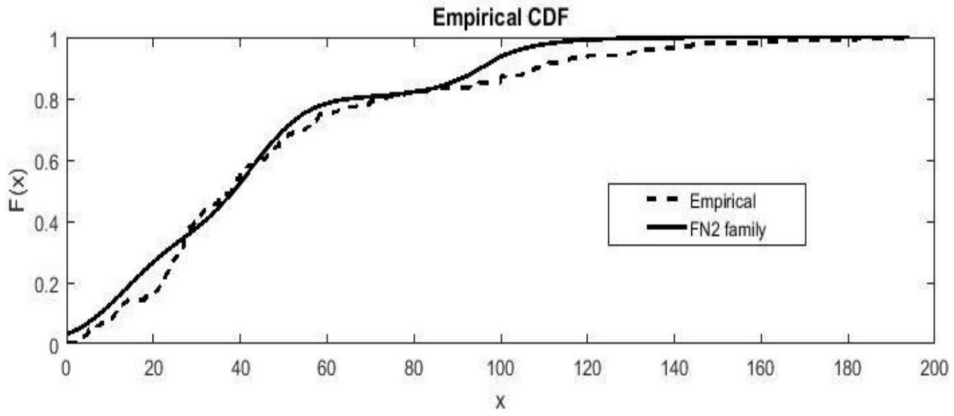

Fig. 9. FN2 family for sulphates concentration (mg/L).

Table 4

SUMMARY STATISTICS

Descriptive statistics for water pollutant								
Pollution	n	minimum	maximum	median	mean	SD	skewness	kurtosis
Chloride	202	20	460	70	91.82	71.68	1.80	5.23
Sulphate	202	0.3	194	38	48.77	38.096	1.39	1.64

(1.1)–(1.3) were calculated by using maxLik package, where these estimates are summarised in Tables 5 and 6. We compared the performances of the four families defined in (1.1)–(1.3) in fitting the data sets of the two pollutants by using AIC and BIC criteria and LRT. The results of these comparisons are presented in Tables 7 and 8 for the chlorides and sulphates concentration (mg/L), respectively. Based on LRT and then on the AIC and BIC criteria, Table 7 shows the performance of the family FN1 as the best, followed by FN2, and then followed by SSN and KumN, respectively. The same result is obtained in Table 8, for the sulphates concentration. Finally, we checked the fitting of these families with the Kolmogorov-Smirnov (K-S) test, where we have four functions [*H, P, KSSTAT, CV*]. Namely, *H* is equal to 0 or 1, *P* is the *p*–value, *KSSTAT* is the maximum difference between the data and the fitting curve and *CV* is a critical value. Thus:

- We accept H_0, if $H = 0$, $KSSTAT \leq CV$ and $P >$ level of significant,

Table 5

PARAMETER ESTIMATION FOR CHLORIDES CONCENTRATION (mg/L) IN EL-SHARKIA DRINKING WATER

ML parameters estimation							
KumN	a	b		μ			σ
	0.032	0.516		191.05			29.68
SSN	α	$\bar{\alpha}$		μ_1	μ_2		σ
	0.28	0.72		24.5	102.6		73.48
FN1	α_1	α_2	α_3	μ_1	μ_2	μ_3	σ
	0.58	0.36	0.06	37.4	131.3	230.45	15.25
FN2	α_1	α_2	α_3	μ_1	μ_2	μ_3	σ
	0.52	0.44	0.04	35.5	125.5	200.5	17.21

- We reject H_0, if $H = 1$, $KSSTAT > CV$ and $P \leq$ level of significant.

The results of the K-S test are summarised in Tables 9 and 10 for chlorides and sulphates concentration (mg/L), respectively. Table 9 shows that the family KumN failed to fit the

Table 6

PARAMETER ESTIMATION FOR SULPHATES CONCENTRATION (mg/L) IN EL-SHARKIA DRINKING WATER

ML parameters estimation							
KumN	a		b		μ		σ
	2.153		0.151		11.02		9.45
SSN	α		$\overline{\alpha}$		μ_1	μ_2	σ
	0.64		0.36		20.47	75	24
FN1	α_1	α_2	α_3	μ_1	μ_2	μ_3	σ
	0.1	0.79	0.11	3.76	37.78	95.45	25.24
FN2	a_1	α_2	α_3	μ_1	μ_2	μ_3	σ
	0.35	0.45	0.2	13.4	142.7	95.42	9.85

Table 7

COMPARING BETWEEN THE FAMILIES (1.1–-(1.3) FOR CHLORIDES CONCENTRATION (mg/L)

Family	AIC	BIC	LRT (p-value)		
			FN2	SSN	KumN
FN1	572.01	1153.20	1	3.75e-12	1
FN2	580.55	1170.30	-	7.33e-16	1
SSN	541.70	1089.90	-	-	1
KumN	677.74	1358.40	-	-	-

Table 8

COMPARING BETWEEN THE FAMILIES (1.1)–(1.3) FOR SULPHATES CONCENTRATION (mg/L)

Family	AIC	BIC	LRT (p-value)		
			FN2	SSN	KumN
FN1	1014.70	2038.50	1	1	1
FN2	1268.90	2546.90	-	5.04e-96	7.44e-100
SSN	1045.40	2097.40	-	-	1.37e-6
KumN	1031.80	2068.10	-	-	-

Table 9

K-S TEST FOR MODELING CHLORIDES CONCENTRATION (mg/L)

df	*H*	*P*	*KSSTAT*	*CV*	Decision
KumN	1	0.0035	0.1173	0.0853	reject the null hypothesis
SSN	0	0.0707	0.0785	0.0853	accept the null hypothesis
FN1	0	0.14	0.0689	0.0853	accept the null hypothesis
FN2	0	0.0854	0.0772	0.0853	accept the null hypothesis

Table 10

K-S TEST FOR SULPHATES CONCENTRATION (mg/L)

df	*H*	*P*	*KSSTAT*	*CV*	Decision
KumN	0	0.07	0.0803	0.0853	accept the null hypothesis
SSN	0	2.349	0.059	0.0853	accept the null hypothesis
FN1	0	174.0286	0.0333	0.0853	accept the null hypothesis
FN2	0	0.5425	0.0381	0.0853	accept the null hypothesis

data set while the other families fit the data. Table 10 shows that all families succeeded to fit the given data. However, relaying on the estimated values of KSSTAT, we can easily see that the performance of the family FN1 is the best, followed by FN2, and then SSN and KumN, respectively, for both pollutants. The estimated mean of the chlorides and sulphates concentration (mg/L) calculated in the best fitted model FN1 are 40.72 and 82.78, respectively. Moreover, as an important application of the selected model (FN1) of the considered pollutants, given in Tables 9 and 10, we calculate the probabilities of exceeding the allowed upper limits for those pollutants in the light of the "Guidelines for Drinking Water, published by the World Health Organisation (Anonymous, 1995). These allowed upper limits for the chlorides and sulphates concentration (mg/L) are 200 and 250, respectively. Therefore, the probabilities of exceeding the allowed upper limits for those pollutants are $P_1 = 0.052$ and P_1= 2.4e-4. Evidently, on the one hand, the probability of exceeding of sulphates concentration (mg/L) is very small. Therefore, this pollutant does not represent any real danger to the public health. On the other hand, although the probability of the exceeding of the chlorides concentration (mg/L) is small, but if we bear in mind that for every 100 liters of drinking water there are 5 liters for which the level of the chlorides exceeds the allowed upper limits, we see that this pollutant represents a concrete danger to the public health.

ACKNOWLEDGMENTS

The authors of this research would like to thank the referee for the valuable suggestions and comments, which improved the presentation substantially. Moreover, they would like to thank Dr. Khaled Z. Elwakeel, Department of Environmental Monitoring and Environmental Management, Faculty of Science, Port Said University, for providing the data.

REFERENCES

Akaike, H. (1973). Information theory and the maximum likelihood principle. In: Petrov, B. N., Csaki, F. (eds.). *Second International Symposium on Information Theory, AkademiaiKiado, Budapest*, pp. 267–281.

Anonymous (1996). *Guidelines for drinking water quality*. 2nd ed. Vol. I. World Health Organization (WHO). 990 pp.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist*., **12** (2), 171–178.

Barakat, H. M. (2015). A new method for adding two parameters to a family of distributions with application to the normal and exponential families. *Statist. Meth. Applic*., **24** (3), 359–372.

Barakat, H. M., Khaled, O. M. (2017). Toward the establishment of a family of distributions that may fit any dataset. *Commun. Statist. Simul. Comput*., **48** (8), 6129–6143.

Casella, G., Berger, R. L. (2010). *Inferência Estatística* 2ª ed. Saraiva. 588 pp.

Cordeiro, G. M., Castro, M. (2011). A new family of generalized distributions. Commun. *Statist. Simul. Comput*., **81** (7), 883–898.

D'Agostino, R. B. (1978). Transformation to normality of the null. *Biometrika*, **57** (3), 679–681.

Ferreira, D. F. (2009). *Estatística Basica´*. UFLA, Lavras. 664 pp.

Hoaglin, D. C., Peter, S. C. (1979). Software for exploring distributional shapes. In:: *Proceedings of Computer Science and Statistics: 12th Annual Symposium on the Interface, Ontario. Canada.* Iniverty of Waterloo, pp. 418–443.

Hoaglin, D. C. (1983). Summarizing shape numerically: The g-and-h distributions. In: *Exploring Data, Tables, Trends and Shapes.* Wiley, New York, pp. 461–513.

Kumaraswamy, P. (1980). Generalized probability density-function for double bounded random-process. *J. Hydrol.*, **462**, 79–88.

Martinez, J., Iglewicz, B. (1984). Some properties of the Tukey g and h family of distributions. *Commun. Statist. Theory Meth.*, **13** (3), 353–369.

Michelle, A. C., Denismar, A. N., Eric, B. F. (2012). Kumaraswamy normal and Azzalini's skew normal modeling asymmetry. *Sigmae, Alfenas*, **1** (1), 65–83.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6** (2), 461–464.

Tukey, J. W. (1977). Modern techniques in data analysis. Proceeding of NSF Sponsored Regional Research Conference at Southeastern Massachusetts University, North Dartmouth, MA.

Received 12 October 2017
Accepted in the final form 9 January 2018

DATU ANALĪZES JAUNĀS METODIKAS AR PIELIETOJUMU ŪDENS PIESĀRŅOJUMA MODELĒŠANAI

Rakstā ir salīdzinātas sadalījumu funkciju saimes, kuras tiek izmantotas asimetriju modelēšanai. Pētījumā pamatā ir datu ģenerēšana ar augošiem asimetriju līmeņiem ar mērķi izvēlēties atbilstošākos no tiem asimetriju modelēšanai. Modelēšanas rezultāti tika salīdzināti ar divām Ēģiptes *El-Sharkia* reģiona dzeramo ūdeni raksturojošām oficiālām datu kopām.