J. Korean Soc. Math. Edu. Ser. B: Theoretical Math. Pedagogical Math.
https://doi.org/10.7468/jksmeb.2025.32.1.1 ISSN(Print) 3059-0604
Volume 32, Number 1 (February 2025), Pages 1–11 ISSN(Online) 3059-1309

MODELLING MONTHLY RAINFALL IN THE CASE OF EXCESSIVE ZEROS

Hamid Ghorbani

Abstract. In the frequency analysis of rainfall amounts for arid and semi-arid regions, it is common to observe zero values for dry days. Since continuous distributions cannot describe such data, a discrete-continuous mixture version of standard distributions is defined to model the data. In this paper, the zero-adjusted versions of several distributions such as gamma, inverse Gaussian, log-normal, Weibull, log-logistic, generalized gamma, and generalized inverse Gaussian were fitted to monthly rainfall. The models were evaluated using a combination of the K-S test and AIC criteria, which indicated that the zero-adjusted gamma distribution provided the most satisfactory fit to the data.

1. Introduction

The frequency analysis of hydrological variables, generally defined as the quantification of the expected number of occurrences of an event of a certain magnitude, is perhaps the first and most common application of probability and statistics in the field of water resources engineering. In summary, the purpose of frequency analysis methods performed on a sample of observed data is to estimate the probability that a random variable is equal to or greater than a given quantity. Note that there are two types of approaches in the analysis of hydrological data, namely the analysis based on the data of each (local) station and the regional analysis. If rainfall data are only available for one station, the frequency analysis, the so-called at-station (or local) analysis method, is applied. Otherwise, if rainfall observations are recorded at different stations in a region and these data are analyzed together, the so-called regional frequency analysis is applied [18]. Regarding the type of data considered in this paper, the station method has been used to find the best probability distribution for the monthly rainfall in one of the hydrometric stations of the Zayanderud

Received by the editors August 29, 2024. Revised October 12, 2024. Accepted Nov. 17, 2024.
2020 *Mathematics Subject Classification.* 62M10, 60E05, 62F10.
Key words and phrases. goodness-of-fit statistic, maximum likelihood estimation, probability distributions, Zayanderud catchment basin, zero-adjusted distributions.

catchment over a period of 49 years from 1970 to 2018. Statistical modelling of such historical and measured hydrological data allows us to simplify the probabilistic prediction of future rainfall with the desired confidence. In fact, the inherent correlation in time series data will help us to predict the future based on historical data. In particular, the primary goal of frequency analysis is to relate the probability of occurrence of extreme events to the frequency of their occurrence by using the statistical distributions [12].

Modelling rainfall data can be distinguished into two parts: rainfall occurrence [10] and rainfall amount [7], [14]. A model of rainfall occurrence is a model that provides a sequence of dry and wet time, while a model of rainfall amounts simulates the amount of rainfall occurred in a desired time interval. Many studies have been carried out to model rainfall using standard distributions such as Gamma, generalized Pareto, Gumbel, Log-Normal, Pearson Type III, Log-Pearson Type III, Weibull, etc. to model the amount of rainfall in humid areas where the data have a right-skewed positive behavior or considering a large time scale for which the true zeros rainfall has no chance, [16], [23]. However, considering only such situations, does not follow the nature of rainfall where there are time intervals that do not rain at all. However, in many biometric, ecological or hydrological situations, it is common to observe semi-continuous variables with true zeros and positive continuous values. For example, it is common to find a large proportion of zeros in rainfall data sets in arid and semi-arid regions (e.g. on daily, monthly or seasonal scales). For this type of data, it is mathematically impossible to fit standard distributions with positive support by maximum likelihood methos, since taking the logarithmic transformation of the vector containing zero values is not allowed. On the other hand, ignoring zeros is also a bad idea, as it makes it impossible to predict the probability of zero and leads to poor inference of the other parameters. To overcome this problem, the assumed probability distributions are updated by adding a zero-adjusted parameter to the distribution to account for the extra zeros in the data.

In this paper, after introducing some zero-adjusted probability distributions, they are used to model the rainfall data in the case of excessive zeros. To estimate the parameters by the maximum likelihood method, the `gamlss.inf` package in the `R` statistical software was implemented, [5].

2. Mixed Random Variables

Mixed random variable Y is a mixed variable (continuous-discrete) whose distribution is a mixture of continuous and discrete distribution with supports equal to R_1 and R_2 respectively. The density function of Y has the following form [15]:

$$\int_{R_1} g_Y(y) + \sum_{R_2} g_Y(y) = 1.$$

The mixed distributions that we encounter in `gamlss.inf` package are those which are either zero-inflated or zero-adjusted ones. The zero-adjusted version of the density function, $f_X(.)$, having a positive support, is:

$$g_Y(y) = \begin{cases} \omega, & y = 0 \\ (1-\omega) f_X(y), & y > 0, \end{cases} \tag{1}$$

in which $f_X(.)$ has a positive skew [13] and $0 < \omega < 1$. Note that the following relations hold for the mean and variance of the g_Y distribution, [22]:

$$E_g(Y) = (1-\omega)E_f(Y), \quad Var_g(Y) = (1-\omega)Var_f(Y) + \omega(1-\omega)E_f(Y).$$

For example, suppose the random variable X has a gamma density function with parameters $\alpha = \frac{1}{\sigma^2}$ and $\lambda = \mu\sigma^2$ and the following density function:

$$f_X(x) = \frac{y^{\frac{1}{\sigma^2}-1} e^{\frac{-y}{\sigma^2\mu}}}{\Gamma(\frac{1}{\sigma^2})(\sigma^2\mu)^{\frac{1}{\sigma^2}}}, \quad x > 0; \mu, \sigma > 0.$$

The density function of the random variable Y, which has a zero-adjusted gamma distribution with parameters μ, σ and ω, shown as $Y \sim \texttt{ZAGA}(\mu, \sigma, \omega)$, is:

$$g_Y(y) = \begin{cases} \omega & y = 0 \\ (1-\omega)\dfrac{y^{\frac{1}{\sigma^2}-1} e^{\frac{-y}{\sigma^2\mu}}}{\Gamma(\frac{1}{\sigma^2})(\sigma^2\mu)^{\frac{1}{\sigma^2}}} & y > 0 \end{cases}$$

Similarly, the other families of zero adjusted distributions can be generated, for example the zero adjusted inverse Gaussian (`ZAIG`), log-normal (`ZALN`), Weibull (`ZAWEI`), log-logistic (`ZALLG`), generalized gamma (`ZAGG`), generalized inverse Gaussian (`ZAGIG`) distributions, all of which were fitted to the rainfall data in this study.

3. Statistical Distributions in R Software

3.1. Common distributions Fitting distributions to data is one of the most common tasks in statistics and involves choosing the appropriate probability distribution

to model the random variable and estimating the parameter for that distribution. The statistical package `fitdistrplus` contains commands for fitting the distributions defined in the core of the software `R` to a set of data using the maximum likelihood method, [4]. These distributions cover a wide range of all common probability distributions in statistics. In special cases, if we want to fit another distribution that is only defined in a specific statistical package, we can use the programming features in order to fit it using the capabilities of the `fitdistrplus` package. These capabilities, in addition to estimating the distribution parameters and calculating the standard error of the estimates, include the Kolmogorov-Smirnov goodness-of-fit test and useful goodness-of-fit diagrams.

3.2. Mixed distributions The `R` package `gamlss` [24], at the moment, supports two distributions namely the zero-adjusted gamma and the zero-adjusted inverse normal distributions. Using the capabilities of the package `gamlss.inf`, we can create the zero-adjusted version of the existing distributions with positive support $(0, \infty)$, like log-normal. On the other hand, for example the zero-adjusted log-normal has a total of three parameters, two of which are related to the log-normal distribution and the other one the probability of taking zero value. In practice, for complex data sets, for the part of the data that lies on a positive real line, we may need a distribution with more than two or three parameters to properly capture the variation in the data. For situations like this, we may create a four parameter distribution using the existing distributions with a support on $(-\infty, \infty)$ in two steps, first by taking an exponential transformation or left truncating at zero, a positive value random variable is created. Then using the facilities of the package `gamlss.inf`, a parameter ω, accounts the excessive zero-value observations, is added to the resulted density function in the first step, which has now a positive support, to get finally a new zero-adjusted distribution which by default was not included in `gamlss.inf` package. In this way, depending on the behavior of the data at hand, any common real valued density function, available in any `R` package, could be converted by the researcher to a desired zero-adjusted one, then statistical inference for it, would be done using the facilities of the `fitdistrplus` and `gamlss.inf` packages.

4. Fitting Distribution to Rainfall Data

The common practice for choosing the appropriate distribution for fitting to rainfall data is first, selecting a number of parametric families of distributions as

candidate ones then fitted them to data one after another. Afterward, based on appropriate goodness-of-fit test, any single fitted distributions are assigned into two categories, either it fits data successfully or fails to fit. Finally, base on a goodness-of-fit criterion, the best fitted candidate, among those fitted successfully, is determined. It is worth mentioning that, due to nature of the data, there is no guarantee that the best selected distribution would adequately model the upper tail of the data. Because this part of the data controls both the magnitude and frequency of extreme events. On the other hand, all fitting methods are biased against the tail since only a very small fraction of the experimental data belongs to the tail (unless a very large sample is available) and the values obtained as estimates of the model parameters reflects, more or less, the ability of model to describe the majority of data belong to the body of the distribution, not the minor part belong to tail of distribution. In other words, this model best describes a large part of the data, not the tail of the data. In application, this type inefficiency of the model to describe extreme values causes inappropriate consequences in hydrological designs [19].

The Normal, log-normal, Gumbel, gamma, Pearson type III, log-Pearson type III, Weibull distributions are among the most important and common probability distributions used in hydrology, [6], [11]. However, the process of choosing an appropriate density function, to best describe the data, among the numerous available models then fitting it has always been a challenge. In the following, the April rainfall records for Ghale Shahrokh station, which belongs to upstream Zayanderud basin, over a period of 49 years from 1970 to 2018 has been considered. When our aim is to model the monthly rainfall, our data include just the rainfall data correspond to a certain month of different years. The reason is that, the random mechanism governing rainfall, for example, in March is not the same as in July. The five point summary statistics for the data (in mm) are min=0, 1st Qu.=30, median=44, 3rd Qu.=80 and max=166.50, where the maximum rainfall was recorded on 2017.

It is worth mentioning that for mixed distribution in (1), given a random sample $X_1, X_2, \cdots, X_n$ that containing $n-m$ zeroes (dry days), the likelihood of the random sample with parameter w and θ is:

$$L(\omega, \theta|x) = \omega^{n-m}(1-\omega)^m \prod_{i=1}^{m} f_{x_i}(\theta).$$

For obtaining the maximum likelihood estimates (MLEs) of the parameters the logarithm of $L(\omega, \theta|x)$ which is a nonlinear equation of the parameters need to be

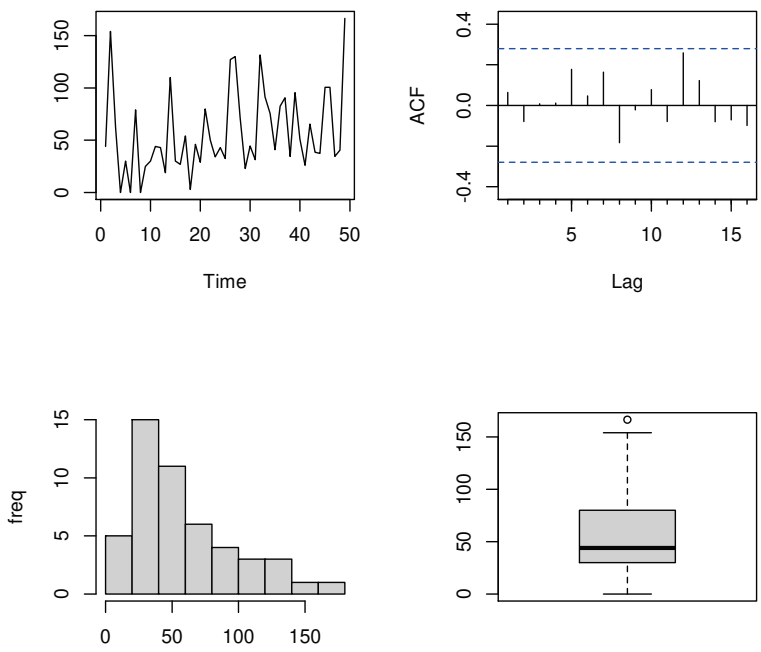

Figure 1. The plots are describing different aspects of the April rainfall records for Ghale Shahrokh station over a period of 49 years from 1970 to 2018.

maximized using the numerical iterative box-constrained optimization routines, [20]. The MLE of ω does not depend on the form of $f(.)$ and is always given by $\hat{\omega} = \frac{n-m}{n}$, i.e. the proportion of zeros observed in the data.

4.1. Test of randomness Autocorrelation functions are usually used to visually check the randomness of a time series. If the time series is random, we expect that all autocorrelations are assumed to be zero for each time lag. The Ljung-Box test is used for testing the null hypothesis that the autocorrelations of a stochastic process for different time delays are all zero. This test is widely used in econometrics and other applications of time series analysis [3]. In addition the package `randtests` provides several non-parametric tests for the randomness of a sequence of observations, among which we can mention the Run test and the Bartels rank test [1]. For the observed data, the randomness assumption is not rejected based on the Bartels rank test at the 0.05 level (statistic=-0.24, n=49, p-value=0.81). Figure 1 shows the time series plot, the sample autocorrelation function (ACF), the histogram and the box plot for the data. The ACF plot exhibits no significant autocorrelation in assumed lags.

4.2. The results of fitting the zero-adjusted distributions Table 1 shows the result of fitting the zero-adjusted gamma (`ZAGA`), inverse Gaussian (`ZAIG`), log-normal (`ZALN`), Weibull (`ZAWEI`), log-logistic (`ZALLG`), generalized gamma (`ZAGG`) and generalized inverse Gaussian (`ZAGIG`) distributions, respectively. The table displays the maximum likelihood estimates of the parameters and, the Kolmogorov-Smirnov (K-S) goodness of fit test statistic and its corresponding p-value, along with the Akaike information criterion (AIC) for each fitted distribution. For the fitted models the corresponding standard errors of the estimated parameters have been displayed. Reporting the standard errors is necessary for measuring the accuracy of the estimates of the parameters and testing their significance. In this study, a combination of the K-S test and AIC has been utilized to evaluate the degree to which different models fit the observed data. The null hypothesis for the K-S test at the significance level of 0.05 determines how likely is each parametric model to describe the rainfall data for the selected station. The KS test statistic quantifies the distance between the empirical cumulative distribution function and the reference cumulative distribution functions, which are subjected for fitting to the data under the null hypothesis. The null hypothesis is rejected if the corresponding p-value of the K-S test is less than the pre-assumed significance level. Regarding the reported p-values in Table 1, all distributions fit the data well. The relative goodness of fit of different models accepted by the K-S test are further compared with the AIC, an estimator of prediction error. The equation for AIC is calculated as $\text{AIC} = -2\log\left(L(\hat{\omega}, \hat{\theta}|x)\right) + 2k$, where $\hat{L}(.|x)$ is the likelihood function of the fitted model for the observed sample and k is the number of estimated parameters in the model. It balances the trade-off between maximum likelihood estimate of the selected parametric cdf and complexity based on the number of parameters. Thus, allowing us to compare and choose the most appropriate parametric model for the data. Thus, AIC provides a means for model selection [21]. Given a collection of models for the data, estimates the quality of each model, relative to each of the other models. The preferred model is the one with the minimum AIC value.

In Table 1, we find the best fitting model is the zero-adjusted gamma distribution (ZAGA) with the lowest AIC value of 480.12. One of the main advantages of AIC is that it is easy to calculate and apply, as it only requires the likelihood function and the number of parameters of the model. AIC is asymptotically effective and unbiased since the test is based on the maximum likelihood function and if the sample size is sufficiently larger than 30, the test will yield fairly accurate result [17]. The

Table 1. The estimated parameters, the corresponding standard errors and the goodness-of-fit statistic for the distributions fitted to the April rainfall records for Ghale Shahrokh station. For all zero-adjusted distributions $\hat{\omega} = 0.06$.

Model	Estim. (s.e.)	K-S (p-value)	AIC
`ZAGA`	$\hat{\mu}$=60.87 (1.00)	0.12 (0.42)	480.12
	$\hat{\sigma}$=0.62 (1.00)		
`ZAIG`	$\hat{\mu}$=60.86 (1.07)	0.15 (0.22)	495.39
	$\hat{\sigma}$=0.11 (1.05)		
`ZALN`	$\hat{\mu}$=3.90 (0.10)	0.097 (0.78)	484.94
	$\hat{\sigma}$=0.69 (1.05)		
`ZAWEI`	$\hat{\mu}$=68.5 (1.00)	0.14 (0.29)	481.30
	$\hat{\sigma}$=1.70 (1.00)		
`ZALLG`	$\hat{\mu}$=3.92 (0.01)	0.096 (0.72)	480.87
	$\hat{\sigma}$=0.37 (1.01)		
`ZAGG`	$\hat{\mu}$=60.89 (1.00)	0.12 (0.42)	482.12
	$\hat{\sigma}$=0.62(1.00)		
	$\hat{\nu}$=1 (0.57)		
`ZAGIG`	$\hat{\mu}$=60.87 (1.00)	0.12 (0.42)	482.12
	$\hat{\sigma}$=22.94(1.00)		
	$\hat{\nu}$=2.60 (0.57)		

sample size of this study is greater than 30, hence AIC can be applied to determine the best model. In small to moderate sample size applications where the candidate collection includes models of high dimension, AIC may severely underestimated and the criterion may favor the overfitted models [2]. It is also noteworthy that the distribution of the K-S test statistic is not dependent on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test. Despite these advantages, the K-S test has several important limitations. The K-S test is more sensitive (i.e., more power) near the center of the distribution than at the tails which might lead to incorrect inferences, see [9] for more details. It is noteworthy that the zero-adjusted version of the distributions of the Lindley family, as referenced in [8], were also subjected to fitting to the data in order to ascertain the potential for a superior fit. However, none of the distributions exhibited a superior fit to the data when compared to the zero-adjusted gamma distribution.

5. Conclusions

This paper addresses a key issue in rainfall frequency analysis, particularly in arid and semi-arid regions where data sets often comprise zero rainfall values. In

such instances, the standard distributions typically employed in existing literature are unsuitable for modelling such data sets, as they are defined on a positive range of values, which makes them incompatible with the statistical characteristics of the data. Moreover, the results may vary notably when the analysis is conducted with and without zero values. It is thus necessary to develop an innovative frequency analysis method that is suitable for rainfall data including zeros. In this study, the zero-adjusted version of the standard distributions commonly used in hydrology for modelling rainfall amount in wet regions have been introduced as a special case of a mixed distribution, which is a combination of the discrete and continuous components. The objective was to identify the most suitable model for monthly rainfall records for a typical hydrometric station in a semi-arid region in the middle of Iran over a period of 49 years from 1970 to 2018. This was achieved by fitting a series of distributions, including zero-adjusted version of gamma, inverse Gaussian, log-normal, Weibull, log logistic, generalized gamma and generalized inverse Gaussian distributions. The maximum likelihood method, the most widely used parameter estimation method, was employed to fit these distributions. The models were evaluated using a combination of the K-S test and AIC criteria. The results indicated that the zero-adjusted gamma distribution was the most satisfactory fit for the data. In conclusion, the findings of this study can be applied to the modelling of other hydrological variables, such as streamflow, drought magnitude, wind speed, and so forth.

References

1. F. Caeiro & A. Mateus: randtests: Testing randomness in R. R package version 1.0.2, 2024.
2. J.E. Cavanaugh & A.A. Neath: The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. WIREs Computational Statistics Journal **11** (2019), 1460. https://doi.org/10.1002/wics.1460
3. J.D. Cryer & K-S. Chan: Time Series Analysis: With Applications in R. Springer, 2ed., New York, 2008. http://dx.doi.org/10.1007/978-0-387-75959-3
4. M.L. Delignette-Muller & C. Dutang: fitdistrplus: An R package for fitting distributions. Journal of Statistical Software **64** (2015), 1-34. https://doi.org/10.18637/jss.v064.i04
5. M. Enea, M. Stasinopoulos, B. Rigby & A. Hossain: gamlss.inf: Fitting Mixed (Inflated and Adjusted) Distributions. R package version 1.0-1, 2019.

6. T.A. Gado, A.M. Salama & B.A. Zeidan: Selection of the best probability models for daily annual maximum rainfalls in Egypt. Theoretical and Applied Climatology **144** (2021), 1267-1284. https://doi.org/10.1007/s00704-021-03594-0
7. H. Ghorbani: Determination of best-fit probability distribution of annual maximum daily precipitation (Case study- Isfahan and Kashan stations). Mathematics and Society **7** (2022), 9-18. https://doi.org/10.22108/msci.2022.134147.1512
8. H. Ghorbani & M. Irshad: Comments on Irshad et al. (2021): "The Zografos-Balakrishnan Lindley distribution properties and applications". Statistica **82** (2022), 45-64. https://doi.org/10.6092/issn.1973-2201/13650
9. M. Goldman & D.M. Kaplan: Comparing distributions by multiple testing across quantiles or CDF values. Journal of Econometrics **206** (2018), 143-166. https://doi.org/10.1016/j.jeconom.2018.04.003
10. J.D.S. Jale, J. Xavier, F.A. Sívio, É.F.M. Xavier, T. Stošić, B. Stošić & T.A.E. Ferreira: Application of Markov chain on daily rainfall data in Paraba-Brazil from 1995-2015. Acta Scientiarum. Technology **41** (2019), 37186. https://doi.org/10.4025/+actascitechnol.v41i1.37186
11. M. M. Khudri & F. Sadia: Determination of the best fit probability distribution for annual extreme precipitation in Bangladesh. European Journal of Scientific Research **103** (2013), 391-404.
12. N.T. Kottegoda: Stochastic Water Resources Technology. Macmillan Press, New York, 1980. https://doi.org/10.1007/978-1-349-03467-3
13. K. Krishnamoorthy: Handbook of Statistical Distributions with Applications. CRC Press, 2ed., New York, 2016. https://doi.org/10.1201/b19191
14. M. Lee, H. An, S. Jeon, S. Kim, K. Jung & D. Park: Development of an analytical probabilistic model to estimate runoff event volumes in South Korea. Journal of Hydrology **612** (2022), 128129. https://doi.org/10.1016/j.jhydrol.2022.128129
15. R. Maity: Statistical Methods in Hydrology and Hydroclimatology. Springer, 2ed., Singapore, 2022. https://doi.org/10.1007/978-981-16-5517-3
16. R. Mudashiru, I. Abustan, N. Sabtu, H. Mukhtar & W. Balogun: Choosing the best fit probability distribution in rainfall design analysis for Pulau Pinang, Malaysia. Modeling Earth Systems and Environment **9** (2023), 3217-3227. https://doi.org/10.1007/s40808-022-01668-0
17. F.M. Mutua: The use of the Akaike Information Criterion in the identification of an optimum flood frequency model. Hydrological Sciences Journal **39** (1994), 235-244. https://doi.org/10.1080/02626669409492740
18. M. Naghettini: Fundamentals of Statistical Hydrology. Springer, Switzerland, 2017. https://doi.org/10.1007/978-3-319-43561-9

19. S.M. Papalexiou, D. Koutsoyiannis & C. Makropoulos: How extreme is extreme? An assessment of daily rainfall distribution tails. Hydrology and Earth System Sciences **17** (2013), 851-862. https://doi.org/10.5194/hess-17-851-2013
20. Y. Pawitan: In All Likelihood: Statistical Modelling and Inference using Likelihood. Oxford Science Publications, Clarendon Press, Oxford, 2013. https://doi.org/10.1093/oso/9780198507659.001.0001
21. E. Popat & A. Hartmann: Exploring cumulative probability functions for streamflow drought magnitude: Global scale analysis and parametric vs. non-parametric comparisons. Journal of Hydrology **637** (2024), 131426. https://doi.org/10.1016/j.jhydrol.2024.131426
22. R.A. Rigby, M. Stasinopoulos, G.Z. Heller & F.De. Bastiani: Distributions for Modeling Location, Scale, and Shape-Using GAMLSS in R. Chapman and Hall/CRC, New York, 2019. https://doi.org/10.1201/9780429298547
23. M. Rizwan, L. Anjum, Q. Mehmood, J.N. Chauhdary, M. Yamin, M. Awais , M.A. Muneer & M. Irfan: Daily maximum rainfall estimation by best-fit probability distribution in the source region of Indus River. Theoretical and Applied Climatology **151** (2023), 1171-1183. https://doi.org/10.1007/s00704-022-04334-8
24. M. Stasinopoulos, R.A. Rigby, Z.H. Gillian & F.De. Bastiani: Flexible regression smoothing: using GAMLSS in R. Chapman and Hall/CRC, New York, 2019. https://doi.org/10.1201/b21973

Professor: Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran
Email address: hamidghorbani@kashanu.ac.ir