Turkish Journal of Agriculture - Food Science and Technology, 6(8): 1030-1034, 2018

Turkish Journal of Agriculture - Food Science and Technology

Available online, ISSN: 2148-127X
www.agrifoodscience.com,
Turkish Science and Technology

Use of Non-Parametric Approaches on Normality of Hydrologic Variables

Kadri Yürekli*, Müberra Erdoğan, Mehmet Murat Cömert

Department of Biosystem Engineering, Faculty of Agriculture, Gaziosmanpasa University, 60240 Tokat, Turkey.

ARTICLE INFO

Research Article

Received 20 March 2018
Accepted 15 May 2018

Keywords:
Normality
Yesilirmak basin
Streamflow
Seasonal streamflow
Kolmogorov–Smirnov

*Corresponding Author:
E-mail: kadriyurekli@yahoo.com

ABSTRACT

Parametric approaches in statistical analysis assume that any given data are normally distributed. Therefore, the test of whether this conventional assumption is valid should be made in this context of the available data's normality before being passed to the application of statistical tests. The paper is focused on the normality methodologies commonly used in literature, named Kolmogorov-Smirnov, Jarque-Bera, D'agostino, Anderson Darling, Shapiro-Wilk and Ryan Joiner. In the study, the seasonal maximum data from eight streamflow gauging stations in Yesilirmak Basin was used as material. The normality in the 59% of the whole data sets were obtained as the highest result by the Kolmogorov –Smirnov approach, when compared to the other normality tests considered in the study.

DOI: https://doi.org/10.24925/turjaf.v6i8.1030-1034.1927

Introduction

In statistical analysis of hydro-meteorological variables, having knowledge about distribution characteristic of a variable to be analysed is very crucial to decide on selection of statistical approach. Özer (2007) reported that the applicability of parametric tests to hydrologic variables was associated with the normally distributed data, otherwise, non-parametric tests should be used. Okman (1994) stated that many hydro-meteorological data showed a right skewed distribution. Das and Imon (2016) imply, it is commonly believed a given data follows normal distribution, and in this sense, before applying any statistical test method to the sample data, the data should be checked whether its observations are departure from normality. Pearson and Please (1975) reported invalidity of some statistical tests such as the t and F test when the normality condition of data was not achieved. Beside, Bera and Jarque (1982) drawn attention that the results concerning with homoscedasticity and serial dependence tests come up with under the condition in which the observations are normally distributed could be led to misinterpretation in the non-normality condition. Under the light of the above, the assumption of normality is a very vital to deduce a reasonable and reliable judgment from statistical analysis of the data.

There are several procedures such as graphical and statistical tests being parametric or non-parametric for the normality assumption. But, graphical approaches give information only about shape of the distribution but, it does not provide a statistically significance result about whether or not the data comes from a normal distribution. Öztuna et al (2006) emphasized that the sample size had an effect on normality test and, in the small sample size circumstance, the null hypothesis related to normality is generally accepted. The basic objective in the study is implement the methods providing visual perspective and the non-parametric test procedure to the data sequences.

Material and Methods

Yesilirmak River basin area which was selected as study region, is approximately 5% of surface area of Turkey. The river basin is situated between 39º 30' and 41º 21' North latitude, 34º 40' and 39º 48' East longitude. Yesilirmak River is one of the major rivers of Turkey and its long is 519 kilometres. The river arises from Kosedag located in the northeast of Sivas province and, joins to Black Sea in district of Carsamba of Samsun province.

There are three main tributaries of the Yesilirmak River, named as Kelkit, Cekerek and Tersakan. Its water is mostly used for purposes as irrigation, drinking, fisheries and wildlife. But, the river has been exposed to pollution due to population growth and rapid industrialization. In terms of land use, presence of forest, cultivated land and pasture land in the basin are about 39%, 39% and 19%, respectively. Due to irregular streamflow regime of Yesilirmak river, flooding in river basin occurs in various times, especially during the period in April, May and June months (Munsuz ve Ünver 1983; Yürekli, 2017; Kurunç et al., 2005; Lekesiz et al., 2007).

In the study, data from eight streamflow gauging stations operated by The General Directorate of State Hydraulic Works (DSI) was used as a material. Figure 1 shows the location map of the streamflow gauging stations. Some characteristics belonging to eight stations were given in Table 1. In the study, streamflow data of the period in which there is the missing data were completed by using Grey System Theory (Wen, 2004). Monthly maximum streamflow value for each month of the relevant year was selected among the daily mean streamflow data for the study. But, the study was conducted on the data sequences in four seasons, names of which were season-I (S-I), season-II (S-II), season-III (S-III) and season-IV (S-IV), respectively. The maximum data of each season was formed by selecting among monthly maximum streamflow values in October, November and December for S-I, January, February and March for S-II, April, May and June for S-III and, July, August and September for S-IV.

The normality analysis of seasonal maximum data set from eight streamflow gauging stations was performed with non-parametric approaches, including Kolmogorov-Smirnov (KS), Jarque-Bera (JB), D'agostino (DA), Anderson Darling (AD), Shapiro-Wilk (SW) and Ryan Joiner (RJ). A detailed description of these methods was not intended for the purposes of reducing volume in the article. These approaches are disclosed in the literatures in detail (Özer, 2007; Jarque and Bera, 1980; D' Agostino et al., 1990; Anderson and Darling, 1952; Shapiro and Wilk, 1965; Yıldırım, 2013; Ryan and Joiner, 1973; Das and Imon, 2016).

Table 1 The streamflow stations used in the study

Station Code	Streamflow (Location)	Longitude (East)	Latitude (North)	Record Length
1401	Kelkit Stream (Fatlı)	36°59'56"	40°28'42"	74
1402	Yesilirmak (Kale)	36°30'45"	40°46'18"	75
1412	Çorum Çat River (Seyhoglu Bridge)	35°25'03"	40°27'06"	60
1413	Yesilirmak (Durucasu)	36°06'43"	40°44'40"	58
1414	Yesilirmak (Sütlüce)	36°07'05"	40°26'03"	59
1418	Yesilirmak (Gömelönü)	37°07'43"	40°18'42"	51
1424	Çekerek Stream (Cırdak Bridge)	36°08'47"	40°0'29"	45
1432	Tersakan Stream (Ahmetsaray)	35°53'15"	40°59'13"	14

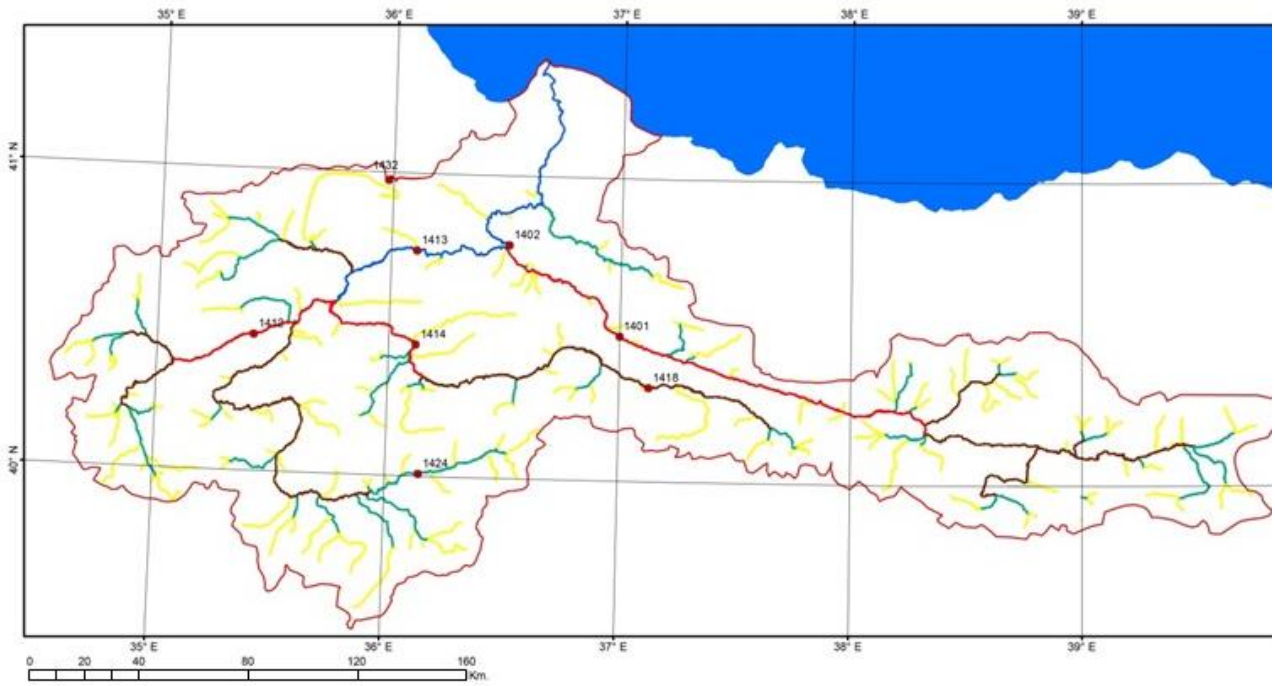

Figure 1 The location map of the streamflow gauging stations

Results

The results of non-parametric approaches applied to seasonal data sequences from each streamflow station are presented in the following tables (Table 2-9). The data normality is accepted in condition where is T_{KS} smaller than KS_{critic} from the table at 5% confidence level with respect to Kolmogorov –Smirnov (KS) test. As can be seen the tables, the data sets of two seasons in stations of 1401, 1413, 1418 and1424 were accepted as statistically normal. All seasons of 1402 and 1432 stations showed a characteristic normally distributed while the station 1414 had statistical normality in three seasons. Whereas none of the seasons in the station 1412 was statistically normal. As the Jarque-Bera (JB), there were no a statistically normally distributed data in any season of 1402, 1412, 1413 and 1418 stations. But, there were statistical normality in three seasons in the station 1432 when having normality in one seasonal maximum data in 1401, 1414 and 1424 stations.

Normality in the four seasons for 1401 station, three season for 1414 and 1432 stations, one season for 1413, 1418 and 1424 stations, and none of all seasons for 1402 and 1412 stations was found out by using D'agostino test (DA). The test results related to the AD, SW and RJ for the considered four periods of 1402 and 1412 stations was similar to that of the DA in terms of non-normality. The data belonging to one period of 1401, 1413, 1414, 1418, 1424 and 1432 stations showed a statistical normal distribution with the AD approach. The probability level symbolized as $P(T_{AD})$ in the tables and representing the test statistic value (T_{AD}) of the AD method implies non-normality when the probability value of $P(T_{AD})$ is smaller than the probability level of the 5% corresponding with the critical test value. The same results for the mentioned stations in the above AD normality test method were also obtained with the RJ test. This conclusion was from the result in which the probability, $P(T_{RJ})$, of the RJ test value(T_{RJ}) was greater than the 5% of significance level. In accordance with the SW methodology, the normality was detected in one season for 1401, 1414 and 1418 stations and, two seasons for 1432 station when the probability, $P(T_{SW})$, associated with the SW test statistic value (T_{SW}) was greater than the 5% probability level corresponding to the critical test value.

Table 2 Normality test results of the station 1401

Season	Normality Tests											
	KS		JB		DA		AD		SW		RJ	
	T_{KS}	KS_{critic}	T_{JB}	JB_{critic}	T_{DP}	DA_{critic}	T_{AD}	$P(T_{AD})$	T_{SW}	$P(T_{SW})$	T_{RJ}	$P(T_{RJ})$
S-I	0.144	0.158	15.16	5.99	0.2735	0.2729 - 0.2863	1.805	<0.005	0.905	0.000	0.950	<0.010
S-II	0.183		6.87		0.2765		2.326	<0.005	0.918	0.000	0.961	<0.010
S-III	0.066		3.06		0.2803		0.457	0.259	0.968	0.054	0.984	0.063
S-IV	0.194		8.75		0.2757		3.794	<0.005	0.873	0.000	0.938	<0.010

Table 3 Normality test results of the station 1402

Season	Normality Tests											
	KS		JB		DA		AD		SW		RJ	
	T_{KS}	KS_{critic}	T_{JB}	JB_{critic}	T_{DP}	DA_{critic}	T_{AD}	$P(T_{AD})$	T_{SW}	$P(T_{SW})$	T_{RJ}	$P(T_{RJ})$
S-I	0.151	0.157	25.39	5.99	0.2702	0.2729 - 0.2863	1.584	<0.005	0.912	0.000	0.954	<0.010
S-II	0.136		14.33		0.2701		1.053	0.009	0.949	0.005	0.973	<0.010
S-III	0.148		283.6		0.2429		2.618	<0.005	0.814	0.000	0.896	<0.010
S-IV	0.139		301.3		0.2366		3.844	<0.005	0.761	0.000	0.867	<0.010

Table 4 Normality test results of the station 1412

Season	Normality Tests											
	KS		JB		DA		AD		SW		RJ	
	T_{KS}	KS_{critic}	T_{JB}	JB_{critic}	T_{DP}	DA_{critic}	T_{AD}	$P(T_{AD})$	T_{SW}	$P(T_{SW})$	T_{RJ}	$P(T_{RJ})$
S-I	0.357	0.175	3047.33	5.99	0.1335	0.2717 - 0.2865	14.021	<0.005	0.326	0.000	0.554	<0.010
S-II	0.183		434.7		0.2291		3.686	<0.005	0.720	0.000	0.841	<0.010
S-III	0.189		15.24		0.2628		3.035	<0.005	0.849	0.000	0.924	<0.010
S-IV	0.339		981.15		0.1592		12.241	<0.005	0.422	0.000	0.638	<0.010

Table 5 Normality test results of the station 1413

Season	Normality Tests											
	KS		JB		DA		AD		SW		RJ	
	T_{KS}	KS_{critic}	T_{JB}	JB_{critic}	T_{DP}	DA_{critic}	T_{AD}	$P(T_{AD})$	T_{SW}	$P(T_{SW})$	T_{RJ}	$P(T_{RJ})$
S-I	0.222	0.178	308.6	5.99	0.2178	0.2714 - 0.2865	5.122	<0.005	0.664	0.000	0.808	<0.010
S-II	0.103		8.1		0.2762		0.569	0.134	0.952	0.023	0.974	0.027
S-III	0.132		20.11		0.2684		1.287	<0.005	0.910	0.000	0.953	<0.010
S-IV	0.221		2732.05		0.1752		6.947	<0.005	0.483	0.000	0.679	<0.010

Table 6 Normality test results of the station 1414

Season	Normality Tests											
	KS		JB		DA		AD		SW		RJ	
	T_{KS}	KS_{critic}	T_{JB}	JB_{critic}	T_{DP}	DA_{critic}	T_{AD}	$P(T_{AD})$	T_{SW}	$P(T_{SW})$	T_{RJ}	$P(T_{RJ})$
S-I	0.187	0.177	6.94	5.99	0.2747	0.2715 - 0.2865	2.025	<0.005	0.903	0.000	0.954	< 0.010
S-II	0.064		1.70		0.2835		0.288	0.606	0.979	0.418	0.991	> 0.100
S-III	0.115		7.26		0.2755		1.176	<0.005	0.925	0.001	0.964	< 0.010
S-IV	0.132		51.73		0.2630		0.977	0.013	0.904	0.000	0.947	< 0.010

Table 7 Normality test results of the station 1418

Season	Normality Tests											
	KS		JB		DA		AD		SW		RJ	
	T_{KS}	KS_{critic}	T_{JB}	JB_{critic}	T_{DP}	DA_{critic}	T_{AD}	$P(T_{AD})$	T_{SW}	$P(T_{SW})$	T_{RJ}	$P(T_{RJ})$
S-I	0.250	0.190	57.64	5.99	0.2295	0.2706 - 0.2865	5.607	<0.005	0.700	0.000	0.835	<0.010
S-II	0.125		244.75		0.2422		1.633	<0.005	0.796	0.000	0.884	<0.010
S-III	0.099		28732.1		0.2833		0.555	0.145	0.957	0.062	0.982	>0.100
S-IV	0.227		59.29		0.2404		3.576	<0.005	0.776	0.000	0.897	<0.010

Table 8 Normality test results of the station 1424

Season	Normality Tests											
	KS		JB		DA		AD		SW		RJ	
	T_{KS}	KS_{critic}	T_{JB}	JB_{critic}	T_{DP}	DA_{critic}	T_{AD}	$P(T_{AD})$	T_{SW}	$P(T_{SW})$	T_{RJ}	$P(T_{RJ})$
S-I	0.379	0.198	2737	5.99	0.1093	0.2695 - 0.2866	12.340	<0.005	0.237	0.000	0.464	<0.010
S-II	0.111		3.03		0.2811		0.639	0.090	0.949	0.049	0.978	0.087
S-III	0.133		15.26		0.2680		0.994	0.011	0.909	0.002	0.951	<0.010
S-IV	0.254		32.56		0.2382		4.299	<0.005	0.740	0.000	0.860	<0.010

Table 9 Normality test results of the station 1432

Season	Normality Tests											
	KS		JB		DA		AD		SW		RJ	
	T_{KS}	KS_{critic}	T_{JB}	JB_{critic}	T_{DP}	DA_{critic}	T_{AD}	$P(T_{AD})$	T_{SW}	$P(T_{SW})$	T_{RJ}	$P(T_{RJ})$
S-I	0.286	0.349	2.35	5.99	0.2579	0.2568 - 0.2857	1.318	<0.005	0.775	0.002	0.890	< 0.010
S-II	0.128		0.409		0.2742		0.278	0.5940	0.967	0.837	0.977	> 0.100
S-III	0.240		2.17		0.2657		11.58	<0.005	0.807	0.060	0.909	0.0140
S-IV	0.334		8.23		0.2184		22.07	<0.005	0.639	0.000	0.793	< 0.010

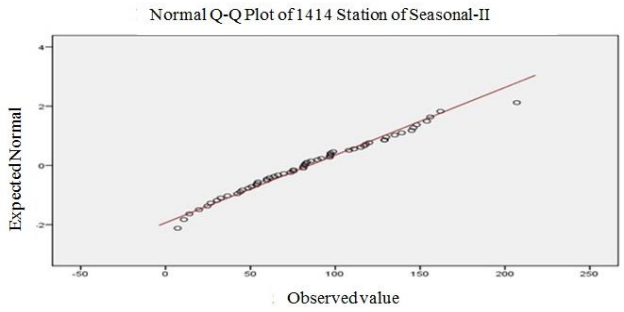

Figure 2 Normal quantile-quantile plot for the S-II series of the staion 1414

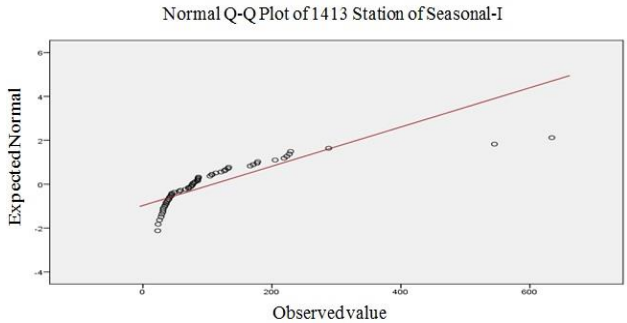

Figure 3 Normal quantile-quantile plot for the S-I series of the staion 1413

The normally distributed data sets among the 32 data sequences (the eight streamflow stations multiplied by four seasons) was presented in Table 10. As can be seen the table, the KS approach revealed the normality in 19 of 32 data series and, the DA method achieved the best other one result in the 13 of the total series after the KS. The six data sets by the JB and DA approaches and, the five data sets by the SW and RJ were found to have a statistically normal distribution

The Q-Q plot being a graphical tool is also widely used on judging if any given data set comprehensibly came from the theoretical normal distribution. The normality assumption of the existing data is realized if their points are approximately formed along the 1:1or 45 degree line when two quantile series are plotted against one another on Cartesian coordinate system. Two data sets, called as S-I for station 1413 and S-II for station 1414 were selected for a visual assessment. One (S-II) of these data sets had a statistically normal distribution and the other (S-I) was not normally distributed according to the above tests. The Q-Q plots of these data sets are presented in Figure 2 and 3.

Table 10 Normally distributed seasonal maximum data for all stations

Station	Normality Tests	Seasons			
		S-I	S-II	S-III	S-IV
1401	KS	✓		✓	
	JB			✓	
	DA	✓	✓	✓	✓
	AD			✓	
	SW			✓	
	RJ			✓	
1402	KS	✓	✓	✓	✓
	JB				
	DA				
	AD				
	SW				
	RJ				
1412	KS				
	JB				
	DA				
	AD				
	SW				
	RJ				
1413	KS		✓	✓	
	JB				
	DA		✓		
	AD		✓		
	SW				
	RJ				
1414	KS		✓	✓	✓
	JB		✓		
	DA	✓	✓	✓	
	AD		✓		
	SW		✓		
	RJ		✓		
1418	KS		✓	✓	
	JB				
	DA			✓	
	AD			✓	
	SW			✓	
	RJ			✓	
1424	KS		✓	✓	
	JB		✓		
	DA		✓		
	AD		✓		
	SW				
	RJ		✓		
1432	KS	✓	✓	✓	✓
	JB	✓	✓	✓	
	DA	✓	✓	✓	
	AD		✓		
	SW		✓	✓	
	RJ		✓		

Conclusion

Knowledge about the distribution of the origin from which the sample data was taken is very crucial to apply the parametric approaches to a given data set. In cases where the distribution pattern of the available data is unknown, the use of parametric tests could lead to inaccurate inference. Yıldırım (2013) recommends non-parametric approaches in such cases. In the study, the six non-parametric methodologies were taken into consideration for normality analysis of the 32 seasonal maximum data sequences from the eight streamflow gauging stations in Yesilirmak Basin. The highest number of data normality (in 19 of 32) was found by the Kolmogorov-Smirnov test. The second highest number of data normality (in 13 data sets) was obtained from the D'agostino test.

References

Anderson TW, Darling DA. 1952. Asymptotic Theory of Certain Goodness-of-fit Criteria Based on Stochastic Processes. The Annals of Mathematical Statistics 23(2): 193-212.

Bera AK, Jarque CM. 1982. Model specification tests: A simultaneous approach. Journal of Econometrics 20: 59-82.

Das KR, Imon AHMR. 2016. A Brief Review of Tests for Normality. American Journal of Theoretical and Applied Statistics, 5(1): 5-12.

D'Agostino RB, Belanger A, D'Agostino RB Jr. 1990. A Suggestion for using Powerful and Informative Tests of Normality, The American Statistician, 44: 316-321.

Jarque CM, Bera AK. 1980. Efficient Tests for Normality Homoscedasticity and Serial İndependence of Regression Residuals, Econometric Letters, 6, pp. 255–259.

Kurunc A, Yurekli K, Cevik O. 2005. Performance of Two Stochastic Approaches for Simulating river Water Quality and Streamflow. Environmental Modeling & Software, 20: 1995-2000.

Lekesiz MC, Mesci Y, Yorulmaz T. 2007. River Basin Management Applications Yesilirmak River Basin Development Project Model. International Congress River Basin Management, 22-24 March, Antalya.

Munsuz N, Ünver İ. 1983. Türkiye Suları, Ank. Üniv. Ziraat Fak. Yay. 392 s.

Okman C. 1994. Hidroloji. Ankara Üniversitesi Ziraat Fakültesi Yayınları No:1388, Ders Kitabı:402, Ankara.

Öztuna D, Elhan AH, Tüccar E. 2006. Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions. Turkish Journal of Medical Sciences, 36(3): 171-176.

Özer A. 2007. Comparison of Normality Tests (M.Sc. Thesis). Ankara University Graduate School of Natural and Applied Sciences. Department of Animal Science.

Pearson ES, Please NW. 1975. Relation Between the Shape of Population Distribution and the Robustness of Four Simple Statistical Tests. Biometrika 62: 223-241.

Ryan TA, Joiner BL. 1973. Minitab: A Statistical Computing System for Students and Researchers. The American Statistician, No. 27, pp. 222–225.

Shapiro SS, Wilk MB. 1965. An Analysis of Variance Test for Normality (complete samples). Biometrika 52(3/4): 591-611.

Wen KL. 2004. Grey Systems: Modeling and Prediction. Yang's Scientific Research Institute, 253 p.

Yıldırım N. 2013. Goodness of Fit Tests for Normal Distribution and a Simulation Study (M.Sc. Thesis). Gazı University Institute Science and Technology.

Yürekli K. 2017. Variability Analysis on Water Quality of Streamflow from Yesilirmak Basin in Turkey. Gaziosmanpaşa Üniversitesi Ziraat Fakültesi Dergisi, 34 (1): 33-37.