INTRODUCTION TO HYDROLOGICAL TIME SERIES ANALYSIS

Jorge A. GUZMAN
Ma. Librada CHU

GRUPO EN PREDICCIÓN Y MODELAMIENTO HIDROCLIMATICO – GPH
ESCUELA DE INGENIERÍA CIVIL
UNIVERSIDAD INDUSTRIAL DE SANTANDER

Revision: Bucaramanga, Julio 2004

TABLE OF CONTENTS

1. THE HYDROLOGICAL TIME SERIES

1.1 Introduction

The rapid increase of population places additional stress on water resources. Water is no longer a simple commodity. Water has become a social matter, a political issue even a religious concern. Water is a lot more than mere engineering. It is life.

Water resources studies have greatly benefited from the advances in science and technology to better understand and assess the complex world of hydrology. Specifically, collection and analysis of hydrological data have improved significantly during the last three decades. More gauging stations were established and sophisticated instruments were introduced to collect hydrological data. Along with the evolution of fast and efficient personal computers, softwares were developed and made available for data management and analysis.

Hydrological data are used for many purposes such as the design and operation of hydraulic structures, water resources management, environmental studies and other related themes. In most cases hydrological data are in the form of a *time series.* A time series is an arrangement of values in accordance with its time (which can be constant or variable) of occurrence. The purpose of time series analysis is two-fold. First, it is used to obtain an understanding of the basic principles of the time series – its variability, and characteristics of its periodic and irregular movements and second, to predict the behavior of the time series in the future.

The analysis of the time series consists of the description and measurement of the various changes or movements as they appear in the series during a period of time. These changes or movements may be classified as (Arkin and Colton, 1964):

1. **Secular trend**, or the long time growth or decline within the data. The period covered should include not less than ten years.

2. **Seasonal variation**, or the more or less regular movement within the twelve-month period. This movement occurs year after year and is caused by changing seasons.
3. **Cyclical movement**, or the swing from prosperity (peak) through recession, depression, recovery, and back again to prosperity. This movement varies in time, length and intensity.
4. **Residual, accidental or random variations**, including such unusual disturbances as wars, disasters, strikes, fads, or other non-recurring factors.

1.2 Components of hydrological time series

Hydrological time series are characterized by two main components: the deterministic components and the stochastic component. Deterministic components are exhibited by the presence of trend and periodicity (seasonal and cyclical variations). Trend is caused by inconsistencies and non-homogeneities in the time series. Non-homogeneity results from natural or man-made changes to the environment during the period of record. Climatic changes and changes in land use can cause non-homogeneity in time series. Inconsistencies on the other hand result from the change in the amount of systematic errors associated with the recording of the data during the period of observations (Hall, 2002). Inconsistencies can be attributed to reading and recording errors by the observer, failure of the instrument, and changes in gauging location and procedure.

Periodicity on the other hand is mainly caused by astronomic cycles. The principal astronomical cycles are the day (based on the rotation of the Earth on its axis), the year (based on the revolution of the Earth around the Sun), and the month (based on the revolution of the Moon around the Earth).

The stochastic component includes residual, or random variations in the hydrological cycle as it integrates with other complex systems (atmosphere, biosphere, lithosphere). Turbulence, large-scale vorticity, heat conversion, atmospheric opacity for incoming and outgoing radiation waves, random thermodynamics, and many other processes in the earth's environments are responsible for randomness. These sources of randomness produce the variations in time series referred to as the stochastic components (Salas, 1980).

Shown in Figures 1.1 to 1.4 are some evident components of time series. Figure 1.1 shows the mean monthly atmospheric carbon dioxide concentration measured in Mauna Loa Observatory in Hawaii. The plot shows very strong deterministic components as exhibited by an upward trend and regular cycles.

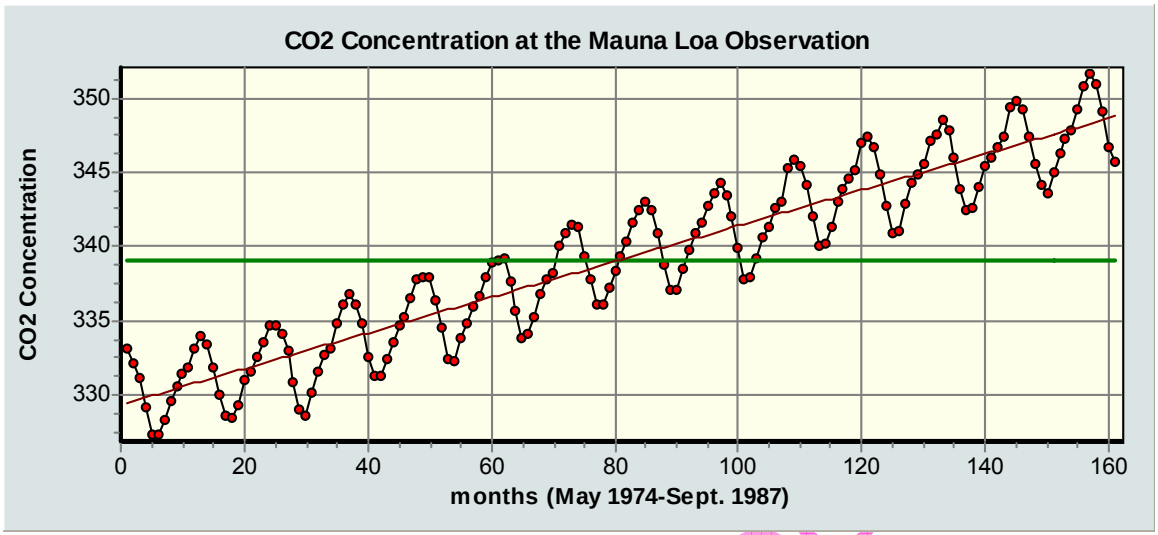


FIGURE 1.1 Mean monthly CO2 concentrations at the Mauna Loa Observatory in Hawaii, USA from May 1974 to September 1987 (http://www.itl.nist.gov/div898/handbook/ , 2004)

Figure 1.2 shows the mean monthly evaporation of the Geul catchment measured at Maastricht airport in The Netherlands. The plot shows a noticeable deterministic component exhibited by regular cycles occurring every 12 months and some stochastic fluctuations.

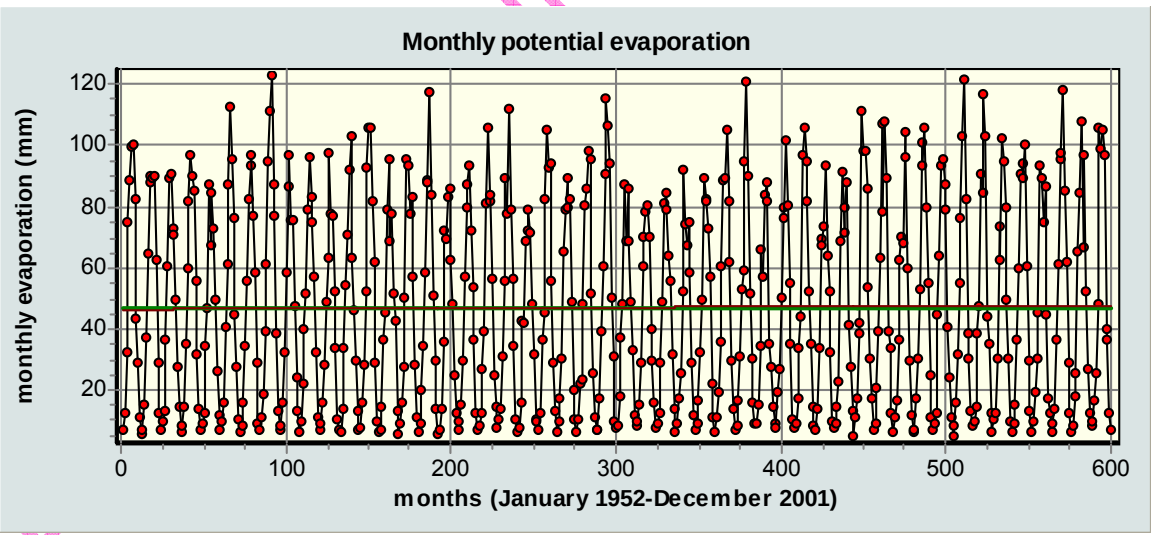


FIGURE 1.2 Monthly potential evaporation of the Geul Catchment, The Netherlands from January 1952 to December of 2001

Figure 1.3 shows the rainfall time series for the month of March in the Geul catchment in The Netherlands. A strong stochastic component is evident in the plot manifested by irregular fluctuations. Deterministic component in the form of an increasing trend is also present in the data.

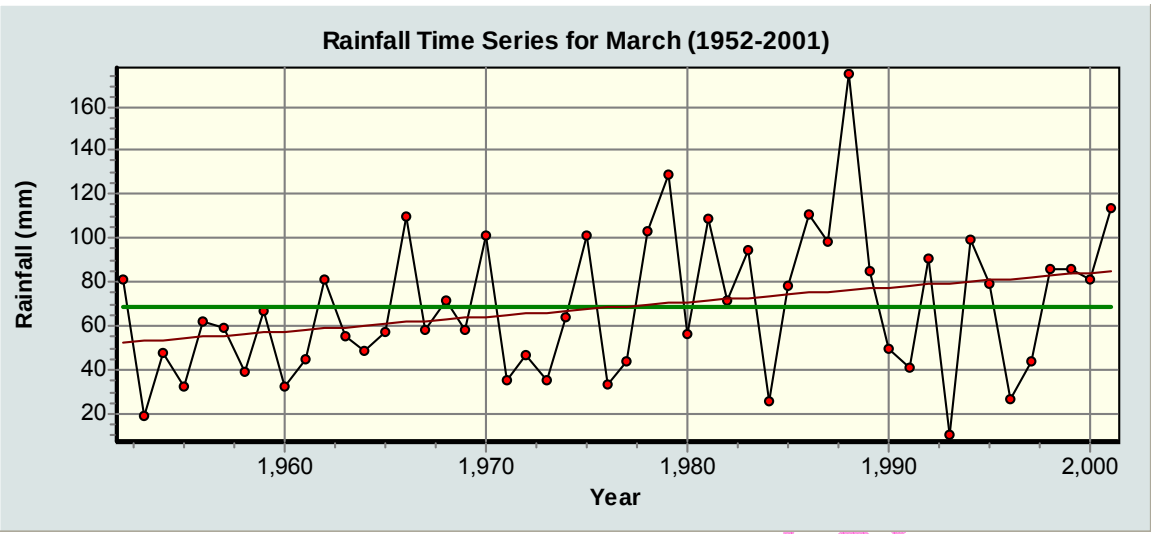


FIGURE 1.3 Rainfall time series for March of the Geul catchment, The Netherlands

Figure 1.4 is a plot of annual absolute maximum discharge of Zulia river measured in Puerto Leon in Colombia from 1973 to 1999. The series has a strong stochastic component with very little deterministic component exhibited by a slight trend.

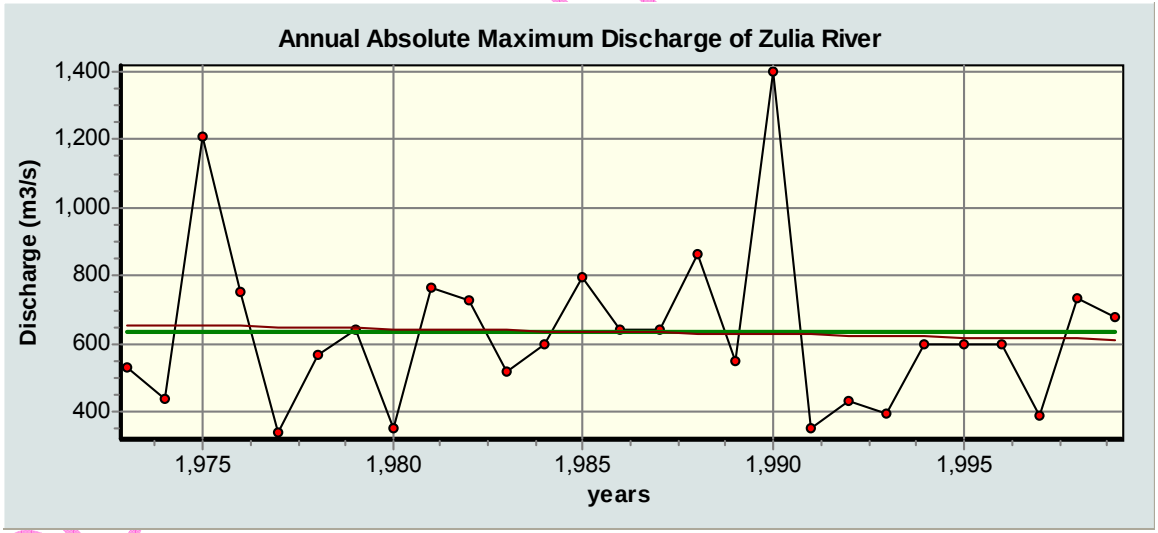


FIGURE 1.4 Annual Maximum Discharge of Zulia River, Colombia

The first step in time series analysis is to isolate the different components - trend, periodicity and the stochastic component. Once the components have been separated from each other, they can be described, and quantified by a process that can represent each one.

Different methods are used to represent these components ("fitting" is the common word used for describing and measuring the components). The most common way of measuring trend for example is by using moving average and

the “least square” method. Seasonal and cyclical movements are dealt with harmonic analysis, the most common of which are the autocorrelation analysis and spectrum analysis. There are several models, which can be used to describe the stochastic component. The most common examples of which are the *autoregressive integrated moving average (ARIMA)* models.

1.3 Time interval

The time resolution used to describe the hydrological variables plays an important role in the variability of the data. Hydrological time series are normally presented as an annual, monthly, daily, or hourly series. Figure 1.5 shows a comparison of the variability of the data at different time resolutions of discharge variables of Zulia river in Colombia. It can be seen that the variability decreases as data aggregation moved from daily values to monthly values to annual values. This change in variability can be attributed to the rapid response of catchment and the non-linear relation of the runoff processes.

The analysis of time series becomes simpler as the time resolution moves from short time interval (say minutes, hours, days) to a larger time interval. The simplest analysis is carried out for a time series of yearly interval. However, there are some trade offs in using annual time series instead of one with shorter interval. For example, daily (or even monthly) time series of evaporation will show strong periodicity while periodicity may not be evident with annual evaporation series. On the other hand, using a finer resolution series will show strong autocorrelation i.e. the value of the current variable is highly dependent on the values of the previous variables making the series non-random.

The analysis of time series is related in various ways to the selected time interval. Depending on the purpose of analysis, the time interval of the variables must be chosen with caution.

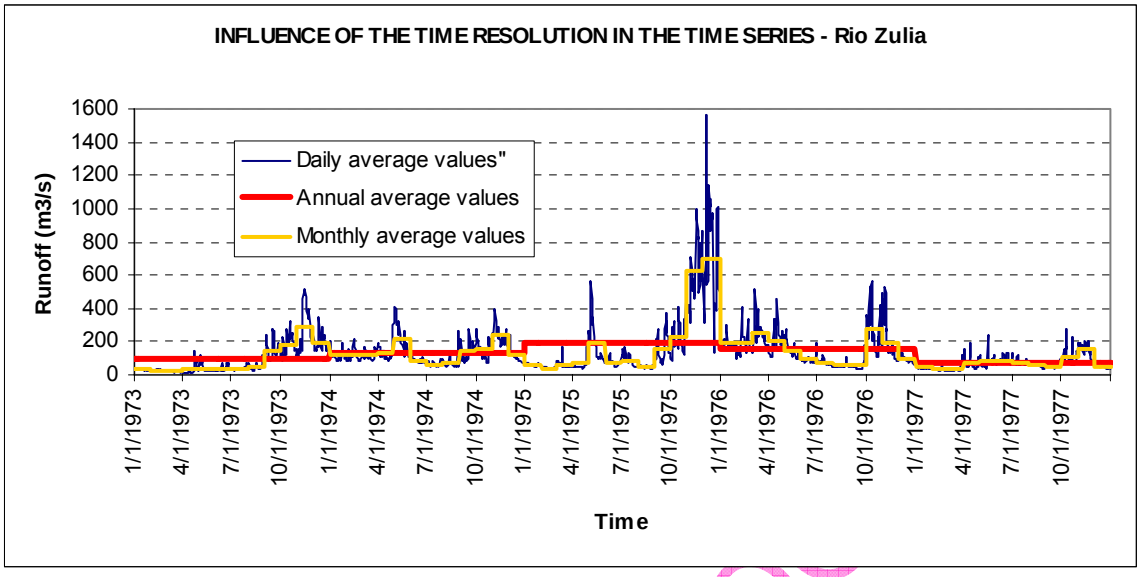


FIGURE 1.5 Runoff time series of Zulia river in Colombia illustrating the variability of the data at different time resolutions.

2. SCREENING OF HYDROLOGICAL DATA

There are four assumptions that typically underlie all measurement processes; namely, that the data from the process at hand "behave like"

1. random drawings;
2. from a fixed distribution;
3. with the distribution having fixed location; and
4. with the distribution having fixed variation.

Predictability is an all-important goal in science and engineering. If the four underlying assumptions hold, then we have achieved probabilistic predictability - the ability to make probability statements not only about the process in the past, but also about the process in the future. In short, such processes are said to be "in statistical control" (NIST/SEMATECH, 2004):

Advances in technology enable the use of sophisticated gauging procedures, which make data collection more easy and efficient. However, population explosion, urbanization, land use change, agricultural modification and the like can introduce non-stationary attributes in the data. Non-stationary behavior in a time series is manifested by the presence of periodicity and trend, which can be in the form of non-homogeneity and inconsistency.

A common assumption in time series analysis is that the data is stationary. Many statistical tests and hydrological models assume stationarity of the time series. A stationary process has the property that the estimates of the basic statistics of the data (mean, variance, and autocorrelation structure) do not change in time i.e., the basic characteristics of the data is unaffected by a shift in time origin. However, these characteristics may be affected whenever a trend and/or a positive or negative jump (slippage) are produced in hydrologic time series by non-homogeneity and/or inconsistency. In general, a stationary time series exhibits the following characteristics:

1. The data are random and independent. Being random means that each value of the variable has an equal chance of occurring. Independence means that the value of one variable is not dependent on (or affected by) the value of another variable.

2. It does not exhibit a significant trend (increasing or decreasing tendency)
3. The variance is stable (it is constant over time)
4. The mean is stable (no significant shift or jump in the data).
5. It does not exhibit regular cycle or seasonality

This chapter will present some basic screening methods which can identify non-stationary attributes in a hydrological time series. The tests are usually applied to time series of hydrological data that are summated over a year or a season. It is assumed that if the data are acceptable at this level of aggregation, they will be equally acceptable at lower levels that cover, say, a month or a day (Dahmen and Hall, 1989).

Each method is provided with a worked example, which can be carried out using a spreadsheet program or calculator. However, a better option is to use computer software for the computational analysis. For this purpose SPELL-stat will be introduced in each test. A complete description on how to use SPELL-Stat in screening hydrological data is presented in Section 2.6.

2.1 Preliminary screening of data

Before analyzing any collection of data, the first consideration must be the characteristics of the data themselves. These characteristics will determine the choice of appropriate data analysis procedure. The first step is to describe and summarize the data in forms, which convey their important characteristics. The basic statistical characteristics of the data (mean, variance, standard deviation, skewness, etc) can give a general overview about the time series.

After the basic statistics of data have been determined, a preliminary screening procedure can be carried out by plotting the time series. Plotting the data with finer resolution (minutes, hourly, daily) is very important in detecting cycles and errors in the data. Figure 2.1 shows a plot of the rainfall and runoff of the Geul catchment in The Netherlands. Also in the plot are the runoff measured at nearby stations Hommerich and Gulpen. It can be seen that the discharges measured at Meersseen from the middle of June to early July 1974 were extremely too high. The graph shows that by comparing these discharges with that of the nearby stations and of the precipitation, one can conclude the there are some errors in the data. The entries during these periods can be considered “suspicious” and further verification should be carried out.

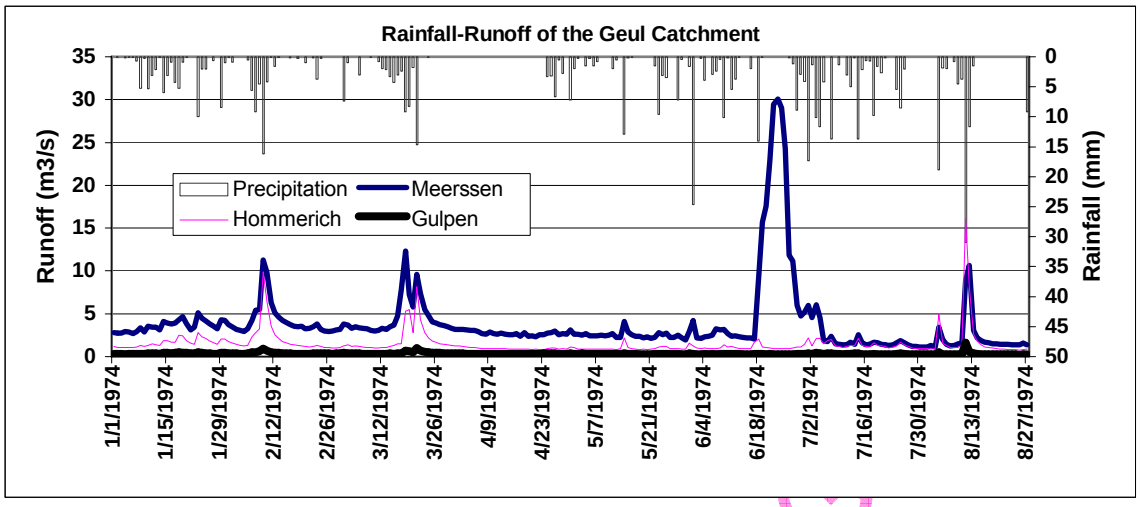


FIGURE 2.1 Daily time series of rainfall and runoff of the Geul catchment, The Netherlands.

After investigating the plot of daily (or hourly) values, it is also important to plot the annual time series of the data. Plotting the annual or seasonal time series can reveal evident non-stationary attributes of the data. Figure 2.2 shows non-stationary attributes of the annual maximum runoff data measured at Café Madrid of Rio de Oro, Bucaramanga in Colombia. The plot shows a decreasing trend indicated by the trendline (thinner line). A slip in the data is also evident in the early 1980's. The plot also shows some indications that the variance (spread of the data about the mean) is unstable.

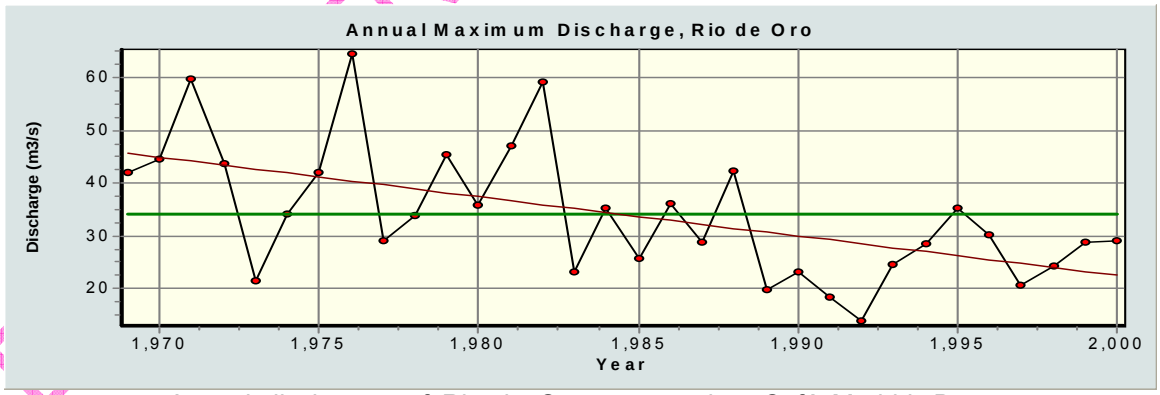


FIGURE 2.2 Annual discharges of Rio de Oro measured at Café Madrid, Bucaramanga, Colombia

Although the time series plot is a good tool in checking some non-stationary behavior of and errors in the data, a formal screening procedure is a requisite. Most time series do not exhibit obvious non-stationary behaviors and only an appropriate statistical test can detect these non-stationary behaviors.

2.2 Test for randomness

The randomness assumption is the most critical and the least tested of all the other assumptions that underlie measurement processes. One specific and common type of non-randomness is autocorrelation. Autocorrelation is the correlation between Y_t and Y_{t-k}, where k is an integer that defines the lag for the autocorrelation. That is, autocorrelation is a time dependent non-randomness. This means that the value of the current point is highly dependent on the previous point if k = 1 (or k points ago if k is not 1) (NIST/SEMATECH, 2004):

If the data are not random due to autocorrelation, then

1. Adjacent data values may be related.
2. There may not be *n* independent snapshots of the phenomenon under study.
3. There may be undetected "junk"-outliers.
4. There may be undetected "information-rich"-outliers

In general, if the assumption of randomness does not hold, all the usual statistical tests are invalid and the parameter estimates become suspect and unsupportive.

2.2.1 Test for serial correlation

Autocorrelation means correlation by itself and is measured by the *serial correlation coefficient*. Autocorrelation coefficients are ordinary linear correlation coefficients between a time series and the same time series an interval of time later. If the data are both random and independent, the (population) serial correlation coefficient is zero for all lags other than zero lag whose value is unity (meaning all values are correlated with themselves). Autocorrelation analysis is presented in a correlogram (see Figures 2.3 and 2.4). It is a plot of the serial coefficients against the lags. The serial correlation coefficient, r_L is computed for lags L = 1,2,…L_{max} where L_{max} should not exceed N/4 using the following expression (Box and Jenkins, 1970).

$$r_L = \frac{\sum_{i=1}^{N-L}(x_i - \bar{x})(x_{i+L} - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad (2.0)$$

Where N is the number of data, x_i is an observation and x_{i+1} is the following observation. The "Lag-one" serial correlation coefficient r_1 is usually considered an adequate parameter in hydrology to judge whether a sequence is random or not.

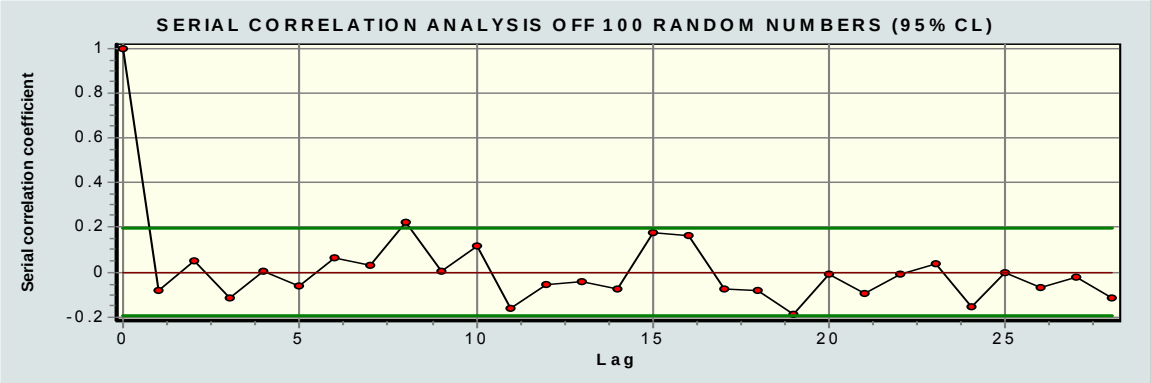


FIGURE 2.3 Correlogram of 100 random numbers

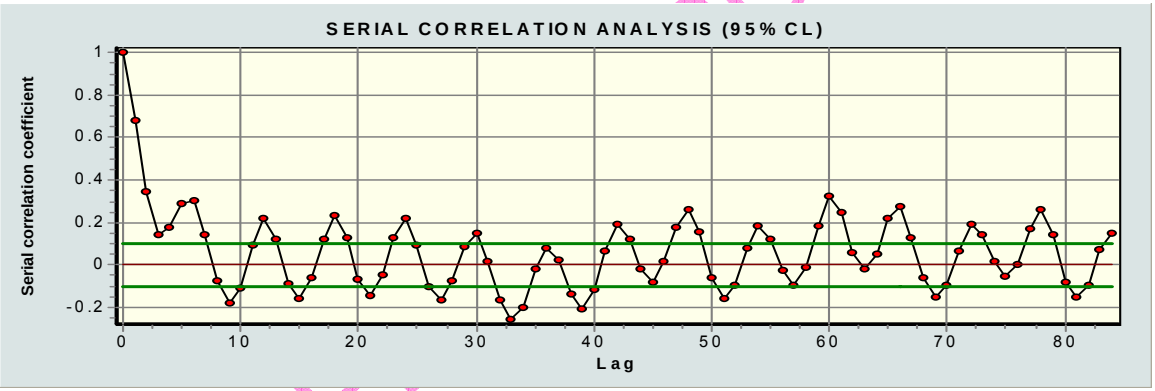


FIGURE 2.4 Correlogram of mean monthly discharge of Rio de Oro, Bucaramanga, Colombia

Null Hypothesis: the serial correlation coefficients are not significantly different from zero.
Alternative Hypothesis: the serial correlation coefficients are significantly different from zero.

The Anderson (1942) test is used to provide confidence limits to the lag k serial correlation coefficient, r_k at 95% level of confidence.

$$CL(r_k) = \frac{-1 \pm 1.96\sqrt{n-k-1}}{n-k} \qquad (2.1)$$

For the whole correlogram, the 95% level of confidence is given in Equation 2.2 (Hall, 2002). Figures 2.3 and 2.4 are examples of correlograms generated

by SPELL-Stat with confidence limits set at 95 % and are indicated by thick lines in the plot.

$$CL(r_k) = \frac{\pm 1.96}{\sqrt{N}} \qquad (2.2)$$

Example 2.1 Test for the absence of autocorrelation is performed to the mean annual flow of the Geul River in The Netherlands. The first serial coefficient is computed using Eq. 2.0. The solution is summarized in Table 2.1

Year	Q(m^3/s)	(X_i-X_m)	$(X_{i+L}-X_m)$	$(X_i-X_m)(X_{i+L}-X_m)$	$(X_i-X_m)^2$	Year	Q(m^3/s)	(X_i-X_m)	$(X_{i+L}-X_m)$	$(X_i-X_m)(X_{i+L}-X_m)$	$(X_i-X_m)^2$
1953	2.94	-0.640	-0.220	0.141	0.410	**1977**	3.09	-0.490	-0.700	0.343	0.241
1954	3.36	-0.220	-0.570	0.126	0.049	**1978**	2.88	-0.700	-0.040	0.028	0.491
1955	3.01	-0.570	0.820	-0.468	0.325	**1979**	3.54	-0.040	0.420	-0.017	0.002
1956	4.40	0.820	0.730	0.598	0.672	**1980**	4.00	0.420	0.530	0.222	0.176
1957	4.31	0.730	-0.890	-0.650	0.532	**1981**	4.11	0.530	0.800	0.423	0.280
1958	2.69	-0.890	-1.420	1.265	0.793	**1982**	4.38	0.800	0.370	0.296	0.639
1959	2.16	-1.420	0.560	-0.795	2.018	**1983**	3.95	0.370	0.300	0.111	0.137
1960	4.14	0.560	0.440	0.246	0.313	**1984**	3.88	0.300	0.140	0.042	0.090
1961	4.02	0.440	-0.450	-0.198	0.193	**1985**	3.72	0.140	0.840	0.117	0.019
1962	3.13	-0.450	-1.510	0.680	0.203	**1986**	4.42	0.840	0.700	0.587	0.705
1963	2.07	-1.510	0.940	-1.419	2.281	**1987**	4.28	0.700	0.080	0.056	0.489
1964	4.52	0.940	2.520	2.367	0.883	**1988**	3.66	0.080	-1.180	-0.094	0.006
1965	6.10	2.520	2.590	6.525	6.348	**1989**	2.40	-1.180	-1.180	1.393	1.393
1966	6.17	2.590	1.170	3.029	6.706	**1990**	2.40	-1.180	-1.070	1.264	1.393
1967	4.75	1.170	0.970	1.134	1.368	**1991**	2.51	-1.070	-0.670	0.718	1.146
1968	4.55	0.970	1.000	0.969	0.940	**1992**	2.91	-0.670	0.050	-0.033	0.449
1969	4.58	1.000	-0.080	-0.080	0.999	**1993**	3.63	0.050	0.360	0.018	0.002
1970	3.50	-0.080	-0.580	0.047	0.006	**1994**	3.94	0.360	-1.690	-0.608	0.129
1971	3.00	-0.580	-0.190	0.111	0.337	**1995**	1.89	-1.690	-0.880	1.488	2.858
1972	3.39	-0.190	-0.330	0.063	0.036	**1996**	2.70	-0.880	-0.870	0.766	0.775
1973	3.25	-0.330	0.440	-0.145	0.109	**1997**	2.71	-0.870	1.120	-0.975	0.758
1974	4.02	0.440	-1.540	-0.677	0.193	**1998**	4.70	1.120	0.570	0.638	1.253
1975	2.04	-1.540	-1.770	2.727	2.373	**1999**	4.15	0.570	0.520	0.296	0.324
1976	1.81	-1.770	-0.490	0.868	3.134	**2000**	4.10	0.520			0.270

TABLE 2.1 Test for the absence of serial correlation

The mean of the data, X_m = 3.58 m^3/s

From Eq. 2.0, the first serial correlation coefficient (for L=1) can be defined as,

$$r_1 = \frac{\sum_{i=1}^{N-1}(x_i - X_m)(x_{i+1} - X_m)}{\sum_{i=1}^{N}(x_i - X_m)^2} = \frac{23.542}{45.250} = 0.520$$

Also from Table 2.1, the value of $\Sigma(X_i-X_m)(X_{i+L}-X_m)$ is found to be 23.542 and that of $(X_i-X_m)^2$ is equal to 45.250 so that the value of the first serial correlation is equal to 0.520.

The null hypothesis H_o can be stated as, "the value of the first serial correlation coefficient is not significantly different from zero". To find out whether the null hypothesis is true, we have to assign confidence limits to the first serial correlation coefficient using Eq. 2.1.

The *upper confidence limit*, $UCL(r_1) = \dfrac{-1+1.96\sqrt{48-1-1}}{48-1} = 0.262$

The *lower confidence limit*, $LCL(r_1) = \dfrac{-1-1.96\sqrt{48-1-1}}{48-1} = -0.304$

Since the computed value of r_1, which is 0.520 lies outside the confidence limits set at 95% level of confidence, the null hypothesis H_o is rejected. There is a significant difference between r_1 and zero which indicates that the data are serially correlated.

2.2.2 Pre-whitening

Several approaches have been suggested for removing the serial correlation from a data set prior to analysis. One of the most common approaches is to pre-whiten the series. The pre-whitening approach involves calculating the serial correlation and removing the correlation if the calculated serial correlation is significant at the 5% level. Pre-whitening is accomplished through (Burn and Hag Elnur, 2002)

$$yp_{t+1} = y_{t+1} - ry_t \tag{2.3}$$

where yp_{t+1} is the pre-whitened series value for time interval t, y_t the original time series value for time interval t, and r is the estimated first serial correlation coefficient. Data should be in standardized form with a mean of 0 and a standard deviation of 1 before pre-whitening is carried out.

Example 2.2 The annual mean runoff of the Geul river was found to be serially correlated. To remove the serial correlation, pre-whitening is performed to the series. The series is standardized (with a mean of 0 and standard deviation of 1) prior to pre-whitening. The standardized series is then pre-whitened by applying Eq. 2.3. Finally, the pre-whitened series is transformed back to the "real-world" set up. The procedure is summarized in Table 2.2.

In Table 2.2, "Transformed Q" is the pre-whitened time series of the annual mean runoff of the Geul river. This series is tested for the absence of serial

correlation to check if the pre-whitening process indeed removed the correlation. A procedure similar to that performed in Example 2.1 is carried out to the new series. The first serial correlation r_1 of the new series is found to be equal to 0.137. The confidence limits are determined and found to be

The *upper confidence limit*, $UCL(r_1) = \dfrac{-1+1.96\sqrt{47-1-1}}{47-1} = 0.264$

The *lower confidence limit*, $LCL(r_1) = \dfrac{-1-1.96\sqrt{47-1-1}}{47-1} = -0.308$

Since the first serial correlation coefficient r_1 of the pre-whitened series lies within the confidence limits set at 95% level of confidence, the null hypothesis is accepted. The pre-whitened data are random and independent.

Year	Q(m3/s)	Standardized Q	Pre-whitened Q	Transformed Q	Year	Q(m3/s)	Standardized Q	Pre-whitened Q	Transformed Q
1953	2.94	-0.653			**1977**	3.09	-0.500	0.439	4.01
1954	3.36	-0.225	0.115	3.69	**1978**	2.88	-0.714	-0.454	3.14
1955	3.01	-0.581	-0.464	3.12	**1979**	3.54	-0.041	0.330	3.90
1956	4.40	0.835	1.138	4.70	**1980**	4.00	0.428	0.449	4.02
1957	4.31	0.744	0.309	3.88	**1981**	4.11	0.540	0.317	3.89
1958	2.69	-0.907	-1.294	2.31	**1982**	4.38	0.815	0.534	4.10
1959	2.16	-1.448	-0.975	2.62	**1983**	3.95	0.377	-0.047	3.53
1960	4.14	0.570	1.324	4.88	**1984**	3.88	0.305	0.109	3.69
1961	4.02	0.448	0.151	3.73	**1985**	3.72	0.142	-0.017	3.56
1962	3.13	-0.459	-0.692	2.90	**1986**	4.42	0.856	0.782	4.35
1963	2.07	-1.539	-1.301	2.30	**1987**	4.28	0.713	0.268	3.84
1964	4.52	0.958	1.759	5.31	**1988**	3.66	0.081	-0.290	3.30
1965	6.10	2.568	2.070	5.61	**1989**	2.40	-1.203	-1.245	2.36
1966	6.17	2.639	1.303	4.86	**1990**	2.40	-1.203	-0.577	3.01
1967	4.75	1.192	-0.181	3.40	**1991**	2.51	-1.091	-0.465	3.12
1968	4.55	0.988	0.368	3.94	**1992**	2.91	-0.683	-0.116	3.47
1969	4.58	1.019	0.505	4.08	**1993**	3.63	0.051	0.406	3.98
1970	3.50	-0.082	-0.612	2.98	**1994**	3.94	0.366	0.340	3.91
1971	3.00	-0.592	-0.549	3.04	**1995**	1.89	-1.723	-1.913	1.70
1972	3.39	-0.194	0.114	3.69	**1996**	2.70	-0.897	-0.001	3.58
1973	3.25	-0.337	-0.236	3.35	**1997**	2.71	-0.887	-0.420	3.17
1974	4.02	0.448	0.623	4.19	**1998**	4.70	1.141	1.603	5.15
1975	2.04	-1.570	-1.803	1.81	**1999**	4.15	0.580	-0.013	3.57
1976	1.81	-1.804	-0.988	2.61	**2000**	4.10	0.530	0.228	3.80

TABLE 2.2 Pre-whitening process of the serially correlated runoff series of the Geul river

Example 2.3 The test for the absence of serial correlation is carried out using SPELL-Stat. Figures 2.5 and 2.6 show the correlogram of the mean annual discharge of the Geul river in The Netherlands before and after the pre-whitening process, respectively. Notice that the first serial coefficient of the original series (Fig. 2.5) lies outside the 95% confidence limits set. The data are pre-whitened and the new series is tested for autocorrelation. Figure 2.6 shows the correlogram of the pre-whitened series. It can be seen that the first

serial correlation coefficient of the pre-whitened series is now within the confidence limits set at 95%.

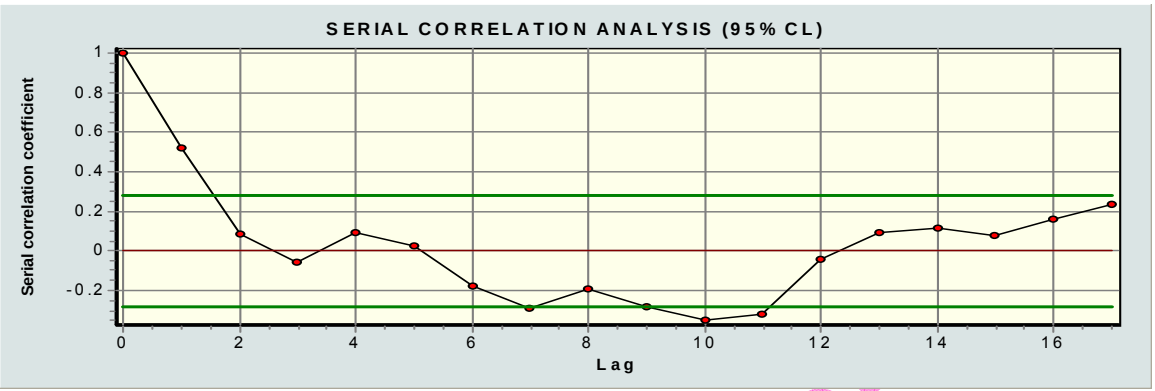


FIGURE 2.5 Correlogram of the mean annual flow of the Geul river, The Netherlands

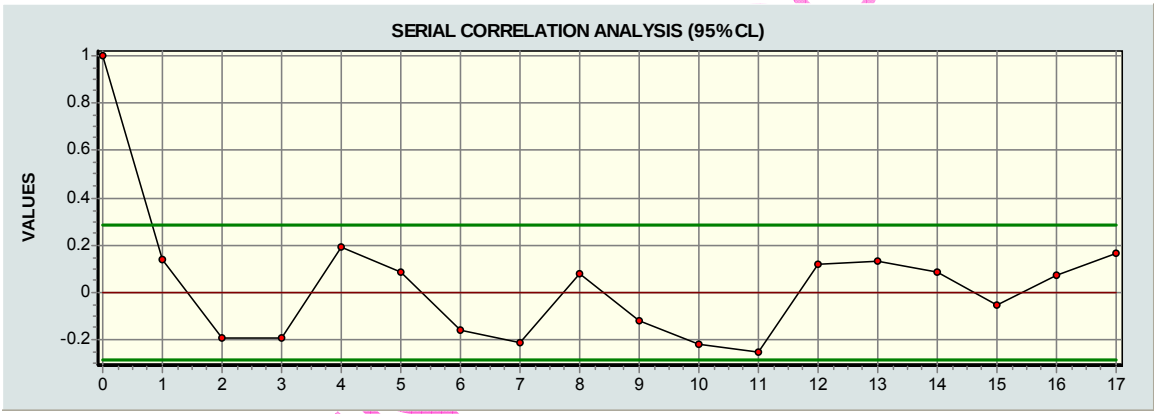


FIGURE 2.6 Correlogram of the mean annual flow of the Geul river after pre-whitening.

2.3 Parametric vs Nonparametric Tests

A statistical analysis requires suitable statistical methods to be chosen. The first consideration is whether to use a parametric or nonparametric test in the analysis. Hypothesis tests which assume that the data have a particular (usually normal) distribution are called *parametric tests.* This is because the information contained in the data is summarized by parameters, usually the mean and standard deviation, and the test statistic is computed using these parameters. On the other hand, hypothesis tests that do not require the data to follow certain distribution are called *nonparametric tests.*

Test procedures should be selected so that they have greater power for the types of data encountered. Comparison of the power of two test procedures, one parametric and one nonparametric, can be based on the tests' *asymptotic relative efficiencies (ARE),* a property of their behavior with large samples. A test with larger ARE will have generally greater power. For non-

normal data, the ARE of nonparametric tests can be many times those of parametric tests. Thus their power to reject the null hypothesis H_o when it is truly false is generally much higher in this case. When data are produced by a normal distribution, nonparametric tests have generally lower (5-15%) ARE than parametric tests (Helsel and Hirsch, 1992).

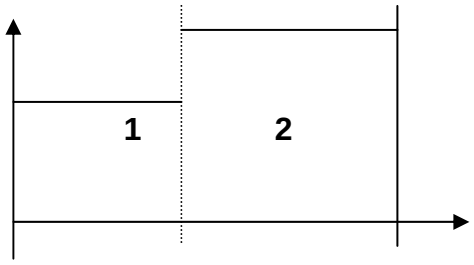
2.4 Parametric Test

Parametric test assumes that the data follow a given distribution usually a normal or Gaussian distribution. Prior to using a parametric test to evaluate stationarity of the data, a test for normality should be carried out to ensure that the data follow a normal distribution. Test for normality is performed by fitting the set of data to a normal distribution. In some cases the data may deviate slightly from a normal distribution. To make the data more normally distributed, they may be transformed by taking their logarithms. The transformed data have to be fitted to a normal distribution to check whether the transformed data are in fact normally distributed. This procedure (distribution fitting) is explained thoroughly in Chapter 3 Frequency Distribution.

The *Standard Normal Homogeneity Test (SNHT)* is a parametric test used to identify non-homogeneities in the time series by assuming the hypothesis of normality of the series. The test can detect sudden shifts of the mean level as well as gradual changes in the mean compared with surrounding sites. Sudden shifts are often related to changes in the gauging location and practices. Gradual changes maybe caused by changes in land use.

2.4.1 SNHT for single shift

The test uses a statistic T_o to compare the mean of the first *a* years of the record with that of the last *n-a* years. The test statistics, T_o will be small if the null hypothesis H_o is true, whereas large values of T_o make the H_A hypothesis more probable. A possible shift is located at the year A, when T_o reaches a maximum at the year a=A. The test statistic T_o was defined by Alexandersson and Moberg (1997).



a **n**

Test value at "a" (the time of break in the data)

$$T_a = a\bar{z}_1^{-2} + (n-a)\bar{z}_2^{-2} \quad (2.4)$$

$$\bar{z}_1 = \frac{1}{n}\sum_{i=1}^{a} z_i \quad 2.5)$$

$$\bar{z}_2 = \frac{1}{n-a}\sum_{i=a+1}^{n} z_i \quad (2.6)$$

$$z_i = \frac{Y_i - \bar{Y}}{s} \quad \text{(standardized data)} \quad (2.7)$$

Test Statistics:

$T_o = \max T_a$

where $\bar{Y}$ is the mean, s the standard deviation, and *n* the total number of years.

If T_o is larger than a certain critical level the series should be classified as non-homogeneous at a certain level (say with 95% level of confidence). This means that the break in the data at "a" year is significant so that the mean values before and after the break are significantly different. The year (or month and year) which is the most probable for a break (change point) can also be obtained from the test. More precisely it is the last year (or the last month and year) with the former mean level (Alexandersson, 1986).

Example 2.3 Standard normal homogeneity test for single shift is performed to the absolute maximum discharge of Zulia river in Colombia. Prior to performing the test, the series should be checked if normally distributed. In this example it is assumed that the series is normally distributed.

The test statistics T_a is computed for all possible values of *a,* the year of possible break in the data using Eq 2.4. The test statistics T_o corresponds to the maximum value of T_a. Table 2.3 shows that the maximum value of T_a is 2.453, which corresponds to *a* equal to 18. This indicates that there is a shift in the mean on the 18th year (or 1990) of the data. If we are going to evaluate the mean of the data before and after the break, we can see that the mean changes from 0.213 to –0.426, respectively indicating a downward shift at the mean level.

Table 2.4 shows the critical levels of the single shift test for the corresponding number of entries. For our data, *n* is equal to 27, which unfortunately is not fond in the table. If we consider 95% level of confidence the critical values for n equal to 20 and 30 are 6.95 and 7.65, respectively. This means that the critical value for n equal to 27 is between 6.95 and 7.65, which is much greater than the computed value of T_o equal to 2.453. Since T_o is statistically small, the null hypothesis H_o that there is no shift is true.

Figure 2.7 shows the plot of T_a as generated by SPELL-stat for the absolute maximum discharge of river Zulia.

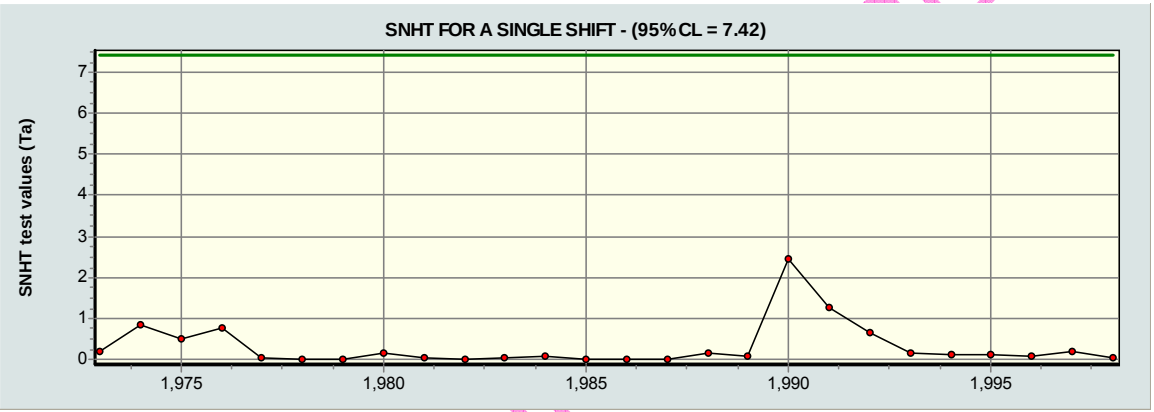


FIGURE 2.7 Test statistics for SNHT single shift generated by SPELL-Stat

Year	Q (m3/s)	Standardized	A	n-a	z1	z2	Ta
1973	529.4	-0.435	1	26	-0.435	0.017	0.196
1974	440.5	-0.801	2	25	-0.618	0.049	0.825
1975	1212	2.378	3	24	0.381	-0.048	0.489
1976	754	0.491	4	23	0.408	-0.071	0.782
1977	340	-1.215	5	22	0.083	-0.019	0.043
1978	570	-0.268	6	21	0.025	-0.007	0.005
1979	645	0.041	7	20	0.027	-0.010	0.007
1980	353	-1.162	8	19	-0.121	0.051	0.167
1981	768	0.548	9	18	-0.047	0.023	0.030
1982	725.6	0.374	10	17	-0.005	0.003	0.000
1983	521.6	-0.467	11	16	-0.047	0.032	0.041
1984	600.2	-0.143	12	15	-0.055	0.044	0.065
1985	797.4	0.669	13	14	0.001	-0.001	0.000
1986	643	0.033	14	13	0.003	-0.003	0.000
1987	643	0.033	15	12	0.005	-0.006	0.001
1988	866.4	0.954	16	11	0.064	-0.094	0.163
1989	550.8	-0.347	17	10	0.040	-0.068	0.074
1990	**1400**	**3.152**	**18**	**9**	**0.213**	**-0.426**	**2.453**
1991	353.7	-1.159	19	8	0.141	-0.335	1.273
1992	433.4	-0.830	20	7	0.092	-0.264	0.658
1993	395	-0.989	21	6	0.041	-0.143	0.158
1994	600	-0.144	22	5	0.032	-0.143	0.125
1995	600	-0.144	23	4	0.025	-0.143	0.096
1996	600	-0.144	24	3	0.018	-0.142	0.068
1997	390	-1.009	25	2	-0.023	0.291	0.183
1998	732	0.400	26	1	-0.007	0.182	0.035
1999	679.2	0.182	27	0			0.000

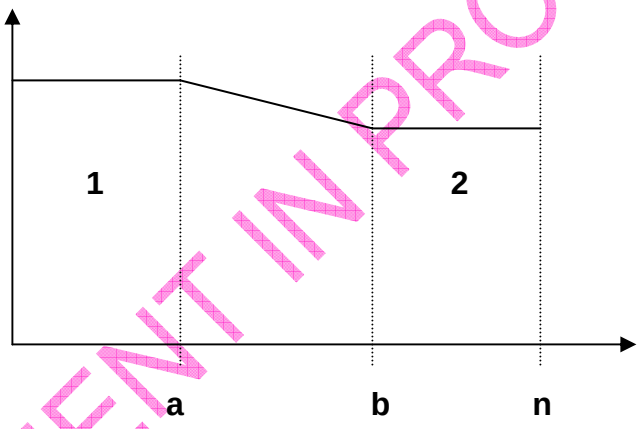
Table 2.3 Test for a single shift at mean level using SNHT

2.4.2 SNHT for linear trend

Alexandersson and Moberg (1997) developed a new test for the detection of linear trends of arbitrary length in normally distributed time series. Assuming that the mean level of a given series changes linearly from *a* to *b*, the test statistics T_{max} is defined in Eq. 2.8. It is further assumed that the minimum number of years in order to accept it as a real trend should be 5 years (b-a > 5 years).

$$T_{\max} = \max_{\substack{a,b \\ 1\le a<b\le n}} \left\{ -a\mu_1^2 + 2a\mu_1 \bar{z}_1 - \mu_1^2 SB - \mu_2^2 SA + 2\mu_1 SZB + 2\mu_2 SZA - 2\mu_1\mu_2 SAB - (n-b)\mu_2^2 + 2(n-b)\mu_2 \bar{z}_2 \right\} \tag{2.8}$$

where:



a	Last year before a possible shift or trend
b	Last year of a possible trend
$\bar{z}_1$ *and* $\bar{z}_2$	Estimated mean level of standardized differences before and after a possible shift or trend, respectively.
μ_1 *and* μ_2	Theoretical mean level of standardized differences before and after a possible shift or trend, respectively

The auxiliary symbols used in the calculation of the test statistics for the trend are given as follows:

$$SA = \sum_{i=a+1}^{b} (i-a)^2 / (b-a)^2 \tag{2.9}$$

$$SB = \sum_{i=a+1}^{b} (b-i)^2 / (b-a)^2 \tag{2.10}$$

$$SZA = \sum_{i=a+1}^{b} z_i (i-a)/(b-a) \tag{2.11}$$

$$SZB = \sum_{i=a+1}^{b} z_i (b-i)/(b-a) \tag{2.12}$$

$$SAB = \sum_{i=a+1}^{b} (b-i)(i-a)/(b-a)^2 \tag{2.13}$$

$$SK = \frac{(-SAB)}{SA+n-b} \tag{2.14}$$

$$SL = \frac{(n-b)\bar{z}_2 + SZA}{SA+n-b} \tag{2.15}$$

$$\mu_1 = \frac{a\bar{z}_1 + SZB - SL\,x\,SAB}{a+SB+SK\,x\,SAB} \tag{2.16}$$

$$\mu_2 = \mu_1 SK + SL \tag{2.17}$$

$$\bar{z}_1 = \frac{1}{a}\sum_{i=1}^{a} z_i \tag{2.18}$$

$$\bar{z}_2 = \frac{1}{n-b}\sum_{i=b}^{n} z_i \tag{2.19}$$

The critical levels for both the single shifts and the trend test are given in Table 2.4.

N	90% C.L.	95% C.L.
10	5.05	5.7
20	6.1	6.95
30	6.65	7.65
40	7	8.1
50	7.25	8.45
60	7.4	8.65
70	7.55	8.8
80	7.7	8.95
90	7.8	9.05
100	7.85	9.15
150	8.05	9.35
200	8.2	9.55
250	8.35	9.7
300	8.5	9.8

TABLE 2.4 Critical levels for single shifts and trend test of SNHT (Alexandersson, 1986)

Performing SNHT for trend using spreadsheet or calculator entails enormous amount of work, which is no longer practical. The use of computer software for this purpose is a must.

2.5 Nonparametric test

Nonparametric tests or distribution-free methods are alternative test procedures if there are no knowledge whatsoever about the distributions of the underlying populations, except perhaps that they are continuous.

Data analysts are using nonparametric or distribution-free procedures with increasing frequency. There are many applications in science and engineering where the data are reported not as values on a continuum but rather on an ordinal scale such that it is quite normal to assign ranks to the data (analysis of ranks). However, there are a number of disadvantages associated with nonparametric tests. Primarily, they do not utilize all the information provided by the sample, and thus a nonparametric test will be less efficient than the corresponding parametric test procedure when both are applicable. Consequently, to achieve the same power, a nonparametric test will require a larger sample size than will the corresponding parametric test (Myers and Walpole, 1993).

The following non-parametric tests are used to detect the presence of homogeneity or inconsistency in a hydrological times series:

- Test for change point
- Test for the absence of linear trend
- Test for stability of mean and variance

2.5.1 Change point test

The change point test determines the year in which there is a possible change (jump or slippage) in the data. This provides a guide as to where the time series will be divided when applying the test for the stability of means and variances (split record test). When the location of the possible break in the series is identified, the data can be divided at the change point by considering equal number of data on both sides of the change point. The least number of data will be used and the rest will be disregarded. This will define the *two equal non-overlapping sub-sets* which will be used for split record test.

Consider a sequence of random variables X_1, X_2, …., X_T. This series is said to have a change point at time τ if the X_t for t = 1, …, τ have a common distribution function $F_1(x)$ and X_t for t = τ+1, …, T have a common distribution function $F_2(x)$, where $F_1(x) \neq F_2(x)$. The Pettitt (1979) change point test involves computing the test statistic $U_{t,T}$ for all t, $1 \leq t \leq T$ from (Hall, 2002):

$$U_{t,T} = \sum_{i=1}^{t} \sum_{j=t+1}^{T} \text{sgn}(X_i - X_j) \tag{2.20}$$

where

sgn(x) = 1 if x > 0
sgn(x) = 0 if x = 0
sgn(x) = -1 if x < 0

The null hypothesis is tested using

$$K_T = \max \left| U_{t,T} \right| ; 1 \leq t \leq T \tag{2.21}$$

Null Hypothesis: there is no significant change in the data, τ = T
Alternative Hypothesis: there is a significant change in the data, $1 \leq \tau \leq T$

The probability associated with K_T may be approximated by Equation 2.22. A probability greater than 0.8 is considered a significant change point.

$$P_T = 1 - \exp\left(\frac{-6K_T^2}{T^3 + T^2} \right) \tag{2.22}$$

Example 2.4 Let us perform change point test to the mean annual flow of Rio de Oro in Bucaramanga, Colombia. The data is consists of 32 years of annual discharge measurements.

Year	Q(m3/s)	Year	Q(m3/s)	Year	Q(m3/s)	Year	Q(m3/s)
1969	21.379	1977	15.015	1985	12.744	1993	14.317
1970	19.985	1978	17.037	1986	19.266	1994	16.75
1971	31.265	1979	21.472	1987	15.528	1995	16.991
1972	22.497	1980	20.258	1988	23.493	1996	20.204
1973	10.306	1981	24.963	1989	13.008	1997	15.413
1974	19.974	1982	23.645	1990	15.283	1998	18.048
1975	17.222	1983	14.306	1991	10.44	1999	19.561
1976	33.591	1984	15.455	1992	8.581	2000	13.269

The statistic $U_{t,T}$ is determined for all values of t (t = 1, 2, …..32) using Eq. 2.20.

For t = 1; j = 1 + 1 = 2

$U_{1,32} = sgn(X_1 - X_2) + sgn(X_1 – X_3) + ……..+ sgn(X_1 – X_{32})$

For t = 2; j = 2 + 1 = 3

$U_{2,32} = sgn(X_1-X_3) + sgn(X_1 – X_4) + ……..+ sgn(X_1 – X_{32}) + sgn(X_2 - X_3) + sgn(X_2 – X_4) + ……..+ sgn(X_2 – X_{32})$

For t = 3; j = 3 + 1 = 4

$U_{3,32} = sgn(X_1 - X_4) + sgn(X_1 – X_5) + ……..+ sgn(X_1 – X_{32}) + sgn(X_2 – X_4) + sgn(X_2 – X_5) + ……..+ sgn(X_2 – X_{32}) + sgn(X_3 – X_4) + sgn(X_3 – X_5) + ……..+ sgn(X_3 – X_{32})$

The procedure is repeated for all values of "t" keeping in mind the following definitions:

If $(X_i – X_j) = 0$; $sgn(X_i – X_j) = 0$
If $(X_i – X_j) > 0$; $sgn(X_i – X_j) = 1$
If $(X_i – X_j) < 0$; $sgn(X_i – X_j) = -1$

The results are summarized in Table 2.5. For each $U_{t,T}$, the associated probability is computed using Eq. 2.22. The maximum probability is found to be 0.9894 corresponding to t = 14 which is the year 1982. Since the computed probability is greater than 0.8, the break in the series is considered significant. Further test should be carried out to reinforce this conclusion.

The tremendous amount of work involved in performing change point test can be reduced greatly by using SPELL-stat. An example of a change point test performed using SPELL-Stat is shown in Figure 2.7. The probabilities associated with candidate change point are plotted against time. It can be seen in the plot that the probabilities from 1980 to 1989 are greater than 0.8, which indicate significant change points. However, the highest probability (1982) corresponds to the location of the most probable break in the data. If historical records are available, this conclusion can be verified.

		t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
j	year	X	X1-Xj	X2-Xj	X3-Xj	X4-Xj	X5-Xj	X6-Xj	X7-Xj	X8-Xj	X9-Xj	X10-Xj	X11-Xj	X12-Xj	X13-Xj	X14-Xj	X15-Xj	X16-Xj	X17-Xj	X18-Xj	X19-Xj	X20-Xj	X21-Xj	X22-Xj	X23-Xj	X24-Xj	X25-Xj	X26-Xj	X27-Xj	X28-Xj	X29-Xj	X30-Xj	X31-Xj	X32-Xj
1	1969	21.379																																
2	1970	19.985	1																															
3	1971	31.265	-1	-1																														
4	1972	22.497	-1	-1	1																													
5	1973	10.306	1	1	1	1																												
6	1974	19.974	1	1	1	1	-1																											
7	1975	17.222	1	1	1	1	-1	1																										
8	1976	33.591	-1	-1	-1	-1	-1	-1	-1																									
9	1977	15.015	1	1	1	1	-1	1	1	1																								
10	1978	17.037	1	1	1	1	-1	1	1	1	-1																							
11	1979	21.472	-1	-1	1	1	-1	-1	-1	1	-1	-1																						
12	1980	20.258	1	-1	1	1	-1	-1	-1	1	-1	-1	1																					
13	1981	24.963	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	-1																				
14	1982	23.645	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	1																			
15	1983	14.306	1	1	1	1	-1	1	1	1	1	1	1	1	1	1																		
16	1984	15.455	1	1	1	1	-1	1	1	1	-1	1	1	1	1	1	-1																	
17	1985	12.744	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	1	1																
18	1986	19.266	1	1	1	1	-1	1	-1	1	-1	-1	1	1	1	1	-1	-1	-1															
19	1987	15.528	1	1	1	1	-1	1	1	1	-1	1	1	1	1	1	-1	-1	-1	1														
20	1988	23.493	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1													
21	1989	13.008	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	1	1	-1	1	1	1												
22	1990	15.283	1	1	1	1	-1	1	1	1	-1	1	1	1	1	1	-1	1	-1	1	1	1	-1											
23	1991	10.44	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1										
24	1992	8.581	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1									
25	1993	14.317	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	-1	1	-1	1	1	1	-1	1	-1	-1								
26	1994	16.75	1	1	1	1	-1	1	1	1	-1	1	1	1	1	1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1							
27	1995	16.991	1	1	1	1	-1	1	1	1	-1	1	1	1	1	1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1						
28	1996	20.204	1	-1	1	1	-1	-1	-1	1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1					
29	1997	15.413	1	1	1	1	-1	1	1	1	-1	1	1	1	1	1	-1	1	-1	1	1	1	-1	-1	-1	-1	-1	1	1	1				
30	1998	18.048	1	1	1	1	-1	1	-1	1	-1	-1	1	1	1	1	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	1	1	-1			
31	1999	19.561	1	1	1	1	-1	1	-1	1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	1	-1	-1		
32	2000	13.269	1	1	1	1	-1	1	1	1	1	1	1	1	1	1	1	1	-1	1	1	1	-1	1	-1	-1	1	1	1	1	1	1	1	0
	Ut,T		17	28	57	78	49	58	59	90	74	74	93	108	135	160	141	132	107	112	105	128	105	92	65	34	17	12	15	26	15	16	21	
	Probability		0.05	0.13	0.44	0.66	0.35	0.45	0.46	0.76	0.62	0.62	0.78	0.87	0.96	0.99	0.97	0.95	0.87	0.89	0.86	0.95	0.86	0.78	0.53	0.19	0.05	0.03	0.04	0.11	0.04	0.04	0.08	

TABLE 2.5 Summary of change point test performed to the runoff time series of Rio de Oro

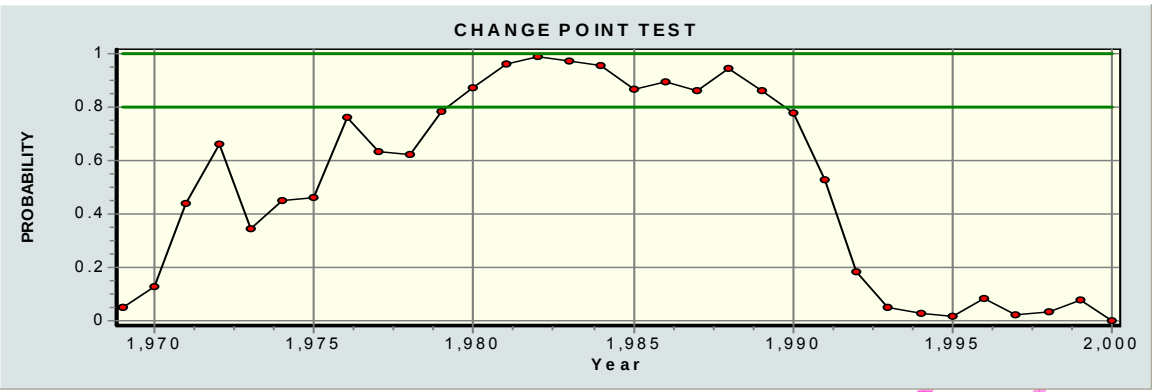


FIGURE 2.8 Change point test of the mean annual flow of Rio de Ore, Bucaramanga, Colombia

2.5.2 Test for the presence of a trend

The absence of trend is verified using Spearman's Rank-Correlation method. It involves computing the Spearman coefficient of rank correlation R_{sp}

$$R_{sp} = 1 - \frac{6\sum Di^2}{n(n^2-1)} \tag{2.23}$$

$$D_i = Kx_i - Ky_i \tag{2.24}$$

where Kx_i is the chronological order number of the observations. The series of observations, y_i, is transformed to its rank equivalent, Ky_i, by assigning the chronological order number of an observation in the original series to the corresponding order number in the ranked series, y. If two or more observations have the same value, the average rank Ky_i is calculated (Dahmen and Hall, 1990). A test-statistic t_t is used to test the null hypothesis against the alternative hypothesis. The test statistic is defined as

$$t_t = R_{sp}\sqrt{\frac{n-2}{1-R_{sp}^2}} \tag{2.25}$$

The test statistic t_t follows Student's t-distribution with v = n –2 degrees of freedom, where n is the number of elements in a sample.

Null Hypothesis: there is no correlation between the order in which the data are observed and the increase (or decrease) in the magnitude of the data, R_{sp} = 0 (there is no trend).

Alternative Hypothesis: there is a correlation between the order in which the data are observed and the increase (or decrease) in the magnitude of the data, $R_{sp} \neq 0$ (there is a trend).

The null hypothesis is accepted if the computed test statistic t_t lies within the acceptance region bounded by t{v,p} and t{v, 1-p} of a student t distribution with v = n –2 degrees of freedom. H_o is accepted if t{v, 2.5%} < t_t < t{v,97.5%} at 5% level of significance (two-tailed test).

Example 2.5 Spearman's rank correlation method is applied to the mean annual flow of Rio de Oro in Bucaramanga, Colombia

year	Q(m3/2)	Ranked Q	Chronological rank of Q, Kxi	Sorted rank of Q, Kyi	Di	Di^2
1969	21.379	8.581	1	25	-24	576
1970	19.985	10.306	2	22	-20	400
1971	31.265	10.44	3	31	-28	784
1972	22.497	12.744	4	27	-23	529
1973	10.306	13.008	5	2	3	9
1974	19.974	13.269	6	21	-15	225
1975	17.222	14.306	7	17	-10	100
1976	33.591	14.317	8	32	-24	576
1977	15.015	15.015	9	9	0	0
1978	17.037	15.283	10	16	-6	36
1979	21.472	15.413	11	26	-15	225
1980	20.258	15.455	12	24	-12	144
1981	24.963	15.528	13	30	-17	289
1982	23.645	16.75	14	29	-15	225
1983	14.306	16.991	15	7	8	64
1984	15.455	17.037	16	12	4	16
1985	12.744	17.222	17	4	13	169
1986	19.266	18.048	18	19	-1	1
1987	15.528	19.266	19	13	6	36
1988	23.493	19.561	20	28	-8	64
1989	13.008	19.974	21	5	16	256
1990	15.283	19.985	22	10	12	144
1991	10.44	20.204	23	3	20	400
1992	8.581	20.258	24	1	23	529
1993	14.317	21.379	25	8	17	289
1994	16.75	21.472	26	14	12	144
1995	16.991	22.497	27	15	12	144
1996	20.204	23.493	28	23	5	25
1997	15.413	23.645	29	11	18	324
1998	18.048	24.963	30	18	12	144
1999	19.561	31.265	31	20	11	121
2000	13.269	33.591	32	6	26	676
					sumDi^2	7664
					Rsp	-0.40469
					t	-2.42395

TABLE 2.6 Spearman's rank correlation analysis applied to annual mean discharge of Rio de Oro in Bucaramanga, Colombia

The original data are ranked in ascending order. The rank of the data as observed is assigned, K_{xi} while rank of the same data in ascending order is

assigned Kyi. Spearman's rank correlation R_{sp} is computed using Eq. 2.24. The solution is presented in Table 2.6.

In order to determine if the computed R_{sp} does not significantly differ from zero, the test statistics t_t is computed using Eq 2.25. Using a table for Student's t-distribution (see appendix), the critical values of the test statistics at 5% level of significance with v equal to 32-2 = 30 are found to be:

t{30, 2.5%} = -2.04, and t{30, 97.5} = 2.04

t{30, 2.5%} = -2.04 < -2.42 < t{30, 97.5} = 2.04 *FALSE!*

Since the computed value of t equal to –2.42 lies outside the acceptance region at 5% level of significant, the null hypothesis H_o is rejected. This indicates that the time series does have a trend (a decreasing trend).

2.5.3 Test for stability of the variance

The appropriate test statistics in testing the stability of the variance is the ratio of the variances of two non-overlapping subsets of the series (identified by the change point test). The F-distribution or Fisher distribution provides a useful indication for the stability of the variance (de Laat, 2001). The number of data *n* in the test series should be equal to or greater than 10. The test statistic is thus

$$F_t = \frac{\sigma_1^2}{\sigma_2^2} = \frac{s_1^2}{s_2^2} \tag{2.26}$$

Null Hypothesis: there is no significant difference between the variance of the two non-overlapping sub-sets, $S_1^2 = S_2^2$ (variance is stable).
Alternative Hypothesis: there is a significant difference between the variance of the two non-overlapping sub-sets, $S_1^2 \neq S_2^2$ (variance is not stable).

The null hypothesis is accepted if the test statistics F_t lies within the acceptance region bounded by $F\{v_1, v_2, p\}$ and $F\{v_1, v_2, 1\text{-}p\}$of the Fisher distribution with degrees of freedom $v_1 = n_1 – 1$ and $v_2 = n_2 - 1$. H_o is accepted if $F\{v_1, v_2, p\} < F_t < \{v1, v_2, 1\text{-}p)$.

The F-distribution is not symmetrical for v_1 and v_2 meaning the lower critical boundary is not equal to the upper boundary. Many tables of F-distribution present only the values of F greater than 1, i.e. only the values of higher probability. If the computed test statistic F_t is less than 1, it is possible to use

these tables by changing F_t to $1/F_t$. Subsequently, it is necessary to interchange the values of v_1 and v_2 in cases where v_1 is not equal to v_2.

2.5.4 Test for stability of the mean

The test for the stability of the variance has to be performed before this test, as statistically the variances of the sub-sets should not be different. The means of the same sub-sets can be compared to verify whether the mean is stable during the whole period of observation (de Laat, 2001). The student t-test will be used for this purpose.

The test statistic, t_t, has Student's t-distribution for a sample which is normally distributed. The test may also be applied for non-normal distributions and is best for approximately equal lengths of sub-sets.

$$t_t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad (2.27)$$

Null Hypothesis: there is no significant difference between the mean of the two non-overlapping sub-sets, $\overline{X_1} = \overline{X_2}$ (the mean is stable).
Alternative Hypothesis: there is a significant difference between the mean of the two non-overlapping sub-sets of the series (the mean is not stable).

The null hypothesis is accepted when the computed test statistics t_t lies within the acceptance region bounded by t{v,p} and t{v, 1-p} of a student t distribution with degrees of freedom $v = n_1 - 1 + n_2 - 1$. H_o is accepted if $t\{v, p\} < t_t < t\{v,1-p\}$.

Example 2.6 Let us apply F-test and t-test to the runoff data of Rio de Oro. The series was previously tested for change point and a significant change point is identified in 1982 (14^{th} year). We would like to find out if the break in the data has made the variance and mean unstable. The series is divided into two equal non-overlapping subsets as identified by the change point test. Sub-set 1 is from 1969 to 1982 (1-14 years) and sub-set 2 is from 1983 to 1996 (15-28 years). The solution is summarized in Table 2.7. Figure 2.8 shows the plot of the series indicating the mean of the subsets.

Sub-set1			Sub-set2		
i	**year**	**Q(m3/2)**	**i**	**year**	**Q(m3/2)**
1	1969	21.379	1	1983	14.306
2	1970	19.985	2	1984	15.455
3	1971	31.265	3	1985	12.744
4	1972	22.497	4	1986	19.266
5	1973	10.306	5	1987	15.528
6	1974	19.974	6	1988	23.493
7	1975	17.222	7	1989	13.008
8	1976	33.591	8	1990	15.283
9	1977	15.015	9	1991	10.44
10	1978	17.037	10	1992	8.581
11	1979	21.472	11	1993	14.317
12	1980	20.258	12	1994	16.75
13	1981	24.963	13	1995	16.991
14	1982	23.645	14	1996	20.204
mean		21.329			15.455
variance		36.191			14.932
				t_t	3.074
				F_t	2.424

TABLE 2.7 F- test and t-test performed for stability of variance and mean

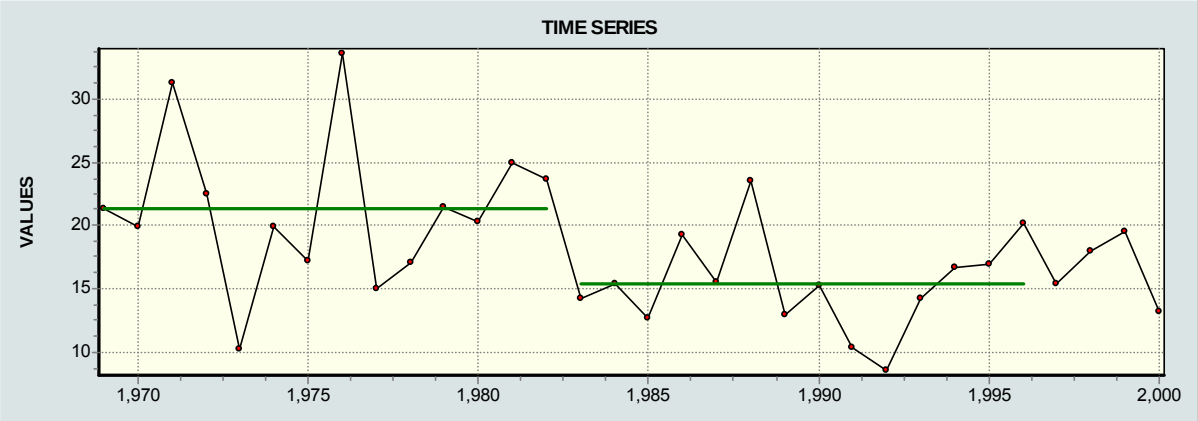


FIGURE 2.8 Time series plot generated by SPELL-Stat showing the slippage in the time series

The computed value of the test statistic $F_t = 2.424$ is obtained using Eq. 2.26. The critical values for F-distribution at 5% level of significance, $v_1 = 14\text{-}1 = 13$ and $v_2 = 14\text{-}1 = 13$ are:

F{13, 13, 2.5%} = 0.321, and F{13, 13, 97.5%} = 3.115

F{13, 13, 2.5%} = 0.321 < F_t = 2.424 < F{13, 13, 97.5%} = 3.115 *TRUE!*

Since the computed test statistics F_t equal to 2.424 is contained within the acceptance region at 5% level of significance, the null hypothesis H_o is accepted. The variance of the time series is stable.

Having established that the variance is stable, we can proceed with the test for the stability of the mean. The test statistics t_t is determined using Eq. 2.27. The critical values of the Student t-distribution at 5% level of significance at v = 28-2 = 26 are found to be:

t{26, 2.5%}= -2.04, and t{26, 97.5%} = 2.04

t{26, 2.5%}= -2.04 < t_t = 3.074 < t{26, 97.5%} = 2.04 *FALSE!*

Since the computed test statistics t_t equal to 3.074 lies outside the acceptance region at 5% level of significance, the null hypothesis is rejected. The mean of the time series is not stable.

The results of the non-parametric tests carried out for the annual mean discharge of Rio de Oro reveal the following conclusions:

- There is a significant break in the time series in 1982
- There is a significant downward trend in the time series
- The variance of the time series is stable at 5% level of significance but the mean is unstable.

The nonparametric tests show that the time series is not stationary.

2.6 Working with SPELL-STAT

All the tests discussed in this chapter can be carried out by SPELL-Stat easily. SPELL-Stat is a time series analysis program which can evaluate the homogeneity and consistency as well as the randomness of time series among its other functions. It is an easy-to-use software with minimal data input.

2.6.1 Data input

The data recognized by SPELL-Stat is in text-file format. The simplest way to prepare the data is using a spreadsheet program like EXCEL. The data should be presented in two columns. The first column contains the date (e.g. dd/mm/yyyy) and the second contains the variable. However SPELL can only recognize a real number date format which represents the number of days since 12 AM of December 30, 1899 (OLE 2.0 format). In order not to complicate things let us consider the following data prepared in EXCEL. Usually time series data are presented in a format similar to table *A*. In order to make this format acceptable to SPELL, the date should be converted to a

real date format. This is carried out by changing the format of the date from "*date*" to "*number*". Once the date is converted to a format similar to table *B*, the data can be exported to SPELL.

Date	*Variable*
1-Jan-76	*5.3*
2-Jan-76	*6.5*
3-Jan-76	*4.3*
4-Jan-76	*7.5*
5-Jan-76	*6.2*
6-Jan-76	*5.4*
7-Jan-76	*2.4*
8-Jan-76	*6.5*
9-Jan-76	*6.8*
10-Jan-76	*7.2*
11-Jan-76	*7.8*
12-Jan-76	*7.5*
13-Jan-76	*6.9*
14-Jan-76	*5.1*
15-Jan-76	*4.3*
16-Jan-76	*3.2*
17-Jan-76	*2.4*
18-Jan-76	*3.5*

(A)

Highlight the "Date" column, go to FORMAT, choose CELL, then click NUMBER. The date column well be changed to something like this ⇒

Numerical format	Variable
27760	5.3
27761	6.5
27762	4.3
27763	7.5
27764	6.2
27765	5.4
27766	2.4
27767	6.5
27768	6.8
27769	7.2
27770	7.8
27771	7.5
27772	6.9
27773	5.1
27774	4.3
27775	3.2
27776	2.4
27777	3.5

(B)

To export the data to SPELL, we simply have to copy the data and paste it in the ***TEXT EDITOR*** window which can be opened through the ***FILE*** menu. The data can be saved easily as text file.

What happens if we are dealing with annual or monthly data? Since SPELL can only recognize a date format which is consist of day, month, and year, it is necessary to input fictitious values for the missing time element. There is nothing to worry because this can be edited easily in SPELL.

2.6.2 The tests

SPELL-Stat can perform both parametric and non-parametric tests. The data has to be opened from the ***FILE*** option in the main menu. SPELL will automatically display the time series in the plot window and the corresponding values in the main window. To perform the desired tests, one has to go to the ***TEST*** option in the main menu.

SPELL-Stat has an intensive ***HELP*** section which can guide the users in performing the analysis.

3. FREQUENCY DISTRIBUTION

3.1 Probability concepts in hydrology

One of the most common problems faced by a hydraulic structural engineer is to estimate the design flood that the proposed structure is expected to withstand. Since all projects are planned for the future, the structural designer is uncertain as to the precise conditions to which the structures will be subjected to. He may know the intended loads on the structure or the magnitude of flood as in the case of designing a hydraulic structure but has no assurance that these loads will not be exceeded. Because of this uncertainty, reasonable assumptions are formulated allowing a generous factor of safety keeping in mind that construction involves cost. Since the exact sequence of flow (streamflow, runoff, discharge) for the future cannot be predicted with 100% accuracy, something must be said about the *probable variation* in flow so that the project can be completed on the basis of a calculated risk.

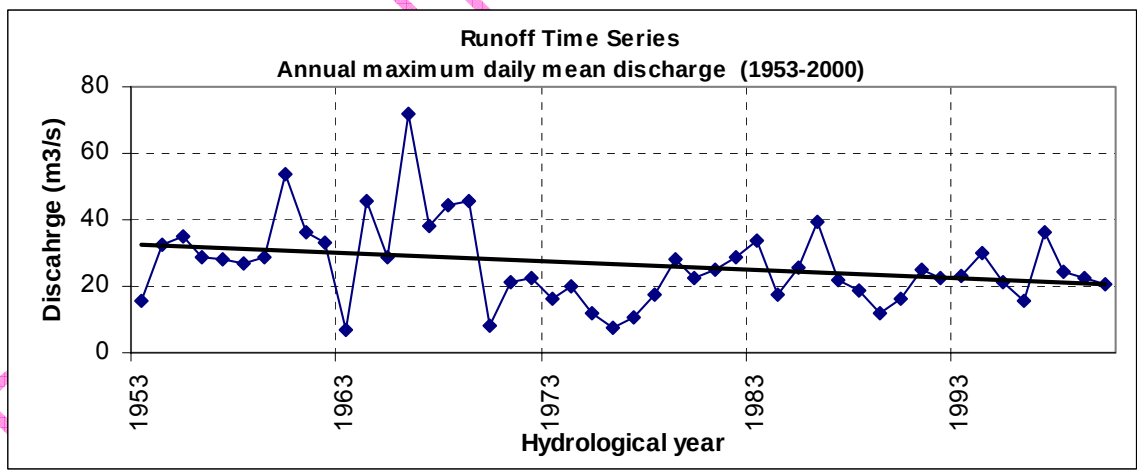


FIGURE 3.1 Annual maximum discharge of the river Geul, The Netherlands

Let us take the time series plot of annual maximum discharge of the river Geul in The Netherlands measured at Meerssen. The data set in Figure 3.1 is fitted with a linear trendline to show the tendency of the maximum discharges. It can be seen that for the 48-year record, the discharges have a decreasing tendency. Despite the decreasing tendency exhibited by the plot of the data, it

is not correct to say that the discharge of the future, say 10 years ahead will continue to decrease.

Now let us group the annual maximum discharges of the Geul according to magnitude. Table 3.1 contains the lists of annual maximum flow of the river Geul from 1953 to 2000 ranked in order of magnitude. The data are grouped in a *class interval* and the frequency of occurrence of each flood is tallied and shown in Table 3.2. The cumulative frequency is obtained by adding the frequency of each class interval starting form the highest magnitude. Both the frequency and cumulative frequency are plotted and shown in Figures 3.2 and 3.3

Rank	Year	Peak flow (m3/s)	Rank	Year	Peak flow (m3/s)
1	1966	71.96	25	1998	24.30
2	1960	53.60	26	1993	23.24
3	1964	45.43	27	1972	22.67
4	1969	45.37	28	1992	22.39
5	1968	44.45	29	1999	22.38
6	1986	39.07	30	1980	22.25
7	1967	38.28	31	1987	22.09
8	1997	36.31	32	1995	21.47
9	1961	36.14	33	1971	21.44
10	1955	34.76	34	2000	20.41
11	1983	33.64	35	1974	20.02
12	1962	33.06	36	1988	18.55
13	1954	32.72	37	1978	17.61
14	1994	29.83	38	1984	17.59
15	1956	28.84	39	1990	16.29
16	1959	28.53	40	1973	15.96
17	1965	28.53	41	1996	15.85
18	1982	28.48	42	1953	15.74
19	1979	28.40	43	1989	11.97
20	1957	27.91	44	1975	11.82
21	1958	26.99	45	1977	10.48
22	1985	25.83	46	1970	7.88
23	1981	25.05	47	1976	7.80
24	1991	24.73	48	1963	6.64

TABLE 3.1 Annual maximum flood data of the Geul river

Class Interval	Frequency	cumulative frequency
10	3	48
20	10	45
30	22	35
40	8	13
50	3	5
60	1	2
70	0	1
80	1	1

TABLE 3.2 Frequency table

The frequency of occurrence is plotted against the corresponding class interval. Figure 3.2 gives a clearer picture of the distribution of flood magnitudes. We can readily see the magnitude of flood that occurred the most and the least. For example, it can be seen right away in the figure that a

flood with a magnitude of 30 to 40 m^3/s occurred more than 20 times in the last 48 years while a magnitude of 70 to 80 m^3/sec occurred the least.

Depending on the purpose of the proposed structure, the design flood can be estimated roughly using a plot similar to Figure 3.2. However, Figure 3.2 does not give us an idea of the probability that the given flood will occur with such magnitude. For purposes of engineering design, the probability of occurrence is important so that the project can be designed based on calculated risk and sufficient factor of safety.

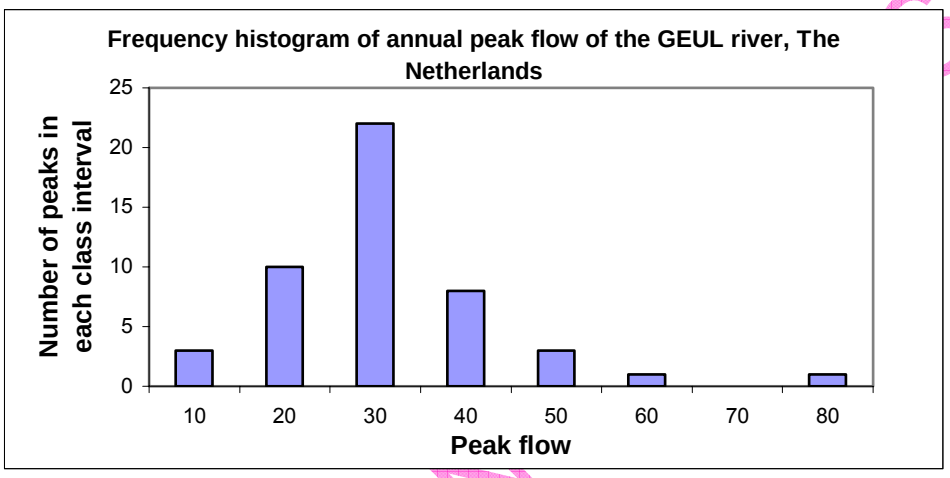


FIGURE 3.2 Frequency histogram of annual peak flow of the Geul river, The Netherlands.

To derive more information from the data set, a cumulative frequency graph (Figure 3.3) is prepared. This is a plot of the total number of floods above the lower limit of the class interval. For example, it can be seen in Figure 3.3 that 35 out of the 48 floods had magnitudes greater than 30 m^3/s. So we can say that the probability that the magnitude of flood will be greater than 30 m^3/s is 72.9% or 35 out of 48.

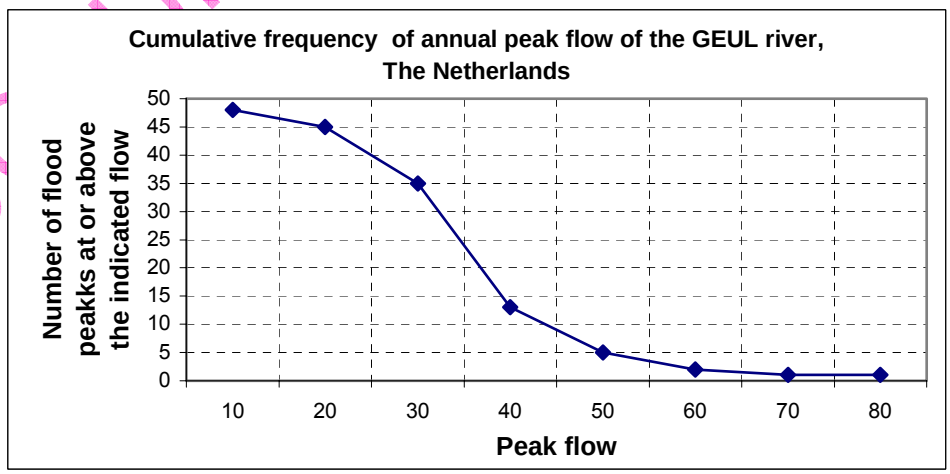


FIGURE 3.3 Cumulative frequency of annual peak flow of the Geul river, The Netherlands

In order to generate reliable information from a given data set, a long record (at least 30 to 40 years) of data should be available. A reliable analysis also requires that all data in a series be gathered under similar condition, i.e., the time series should be stationary. Stationarity in a time series means the statistics of the series (mean, variance, etc.) do not change in time.

3.2 Return Period

For the purpose of engineering design, much attention is paid to the event equaled or exceeded on average once every ***T*** years, also referred to as the *T-year event*. The *T-year event* represents the return period or recurrence interval of a given event. It is the average interval in years between events, which equal or exceed the considered magnitude of event.

If P is the probability that the event will be equaled or exceeded in a particular year, the return period T may be expressed as

$T = 1/P$

One should keep in mind that a return period of a certain event e.g. 10 years does not imply that the event occurs at 10 years interval. It means that the probability that a certain value (e.g. rainfall depth, discharge) is exceeded in a certain years is 10%.

P is the probability of the occurrence of an event.
1-P is the probability that the event will not occur.
(1-P)(1-P) is the probability that the event will occur in two successive years.
$(1-P)^3$ is the probability that the event will occur in three successive years.
$(1-P)^N$ is the probability that the event will occur in *N* successive years. N can be the design life of the structure.
Hence, $J = 1 - (1-P)^N$ is the probability that the event will occur during the span of *N* years.

3.3 Frequency distribution

When a long period of data is available, it is possible to use a smaller class interval and the histogram in Figure 3.2 might approach a smooth frequency distribution similar to that in Figure 3.4. The ordinates of Figure 3.4 are the probability density and the abscissa the magnitude of the floods. The ratio of the shaded area to the area under the entire curve is the probability that the given event will be equaled or exceeded. If a frequency distribution of a given data set is identified, the task of determining the magnitude of an event

corresponding to a given probability of exceedence becomes a simple task. However, only a very large number of samples (i.e., a long record) will permit accurate definition of a distribution. It is therefore necessary that a long record be made available to get reliable results from a frequency analysis.

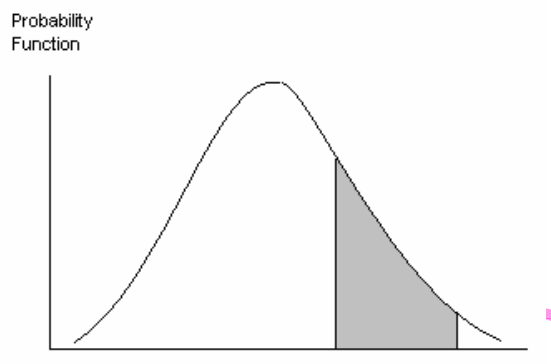


FIGURE 3.4 Frequency distribution

3.4 Probability Distribution

Probability distributions are a fundamental concept in statistics. They are used both on a theoretical level and a practical level. Some practical uses of probability distributions are (*NIST/SEMATEC*, 2004):

- To calculate confidence intervals for parameters and to calculate critical regions for hypothesis tests.
- For univariate data, it is often useful to determine a reasonable distributional model for the data.
- Statistical intervals and hypothesis tests are often based on specific distributional assumptions. Before computing an interval or test based on a distributional assumption, we need to verify that the assumption is justified for the given data set. In this case, the distribution does not need to be the best-fitting distribution for the data, but an adequate enough model so that the statistical technique yields valid conclusions.
- Simulation studies with random numbers generated from using a specific probability distribution are often needed

Probability distributions are typically defined in terms of the probability density function. For a continuous function, the probability density function (pdf) is the probability that the variate has the value x. Since for continuous distributions the probability at a single point is zero, this is often expressed in terms of an integral between two points.

$$\int_a^b f(x)dx = \Pr[a \leq X \leq b] \qquad 3.1$$

The cumulative distribution function (cdf) is the probability that the variable takes a value less than or equal to x. That is

$$F(x) = \Pr[X \leq x] = \alpha \qquad 3.2$$

For a continuous distribution, this can be expressed mathematically as

$$F(x) = \int_{-\infty}^{x} f(\mu) d\mu \qquad 3.3$$

Shown in Figure 3.5 are the plots of the probability distribution function and the cumulative distribution function of a normal distribution.

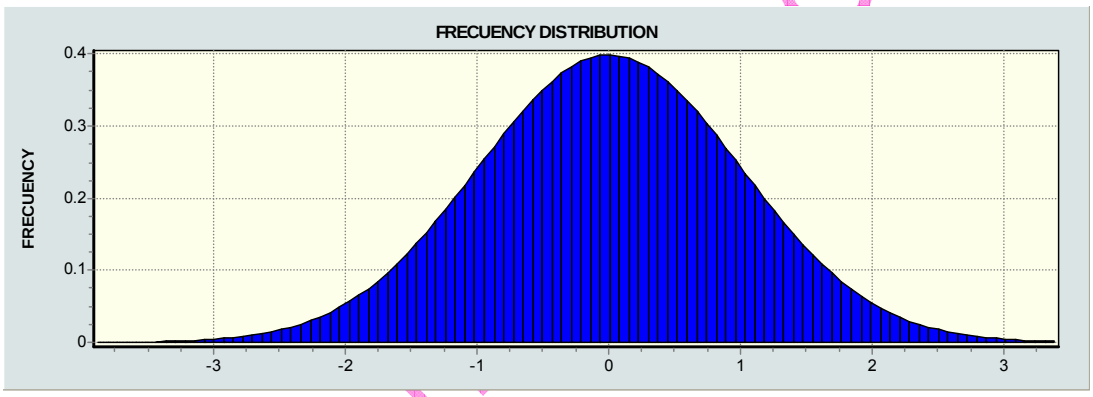


FIGURE 3.5a Probability distribution function of a normal distribution

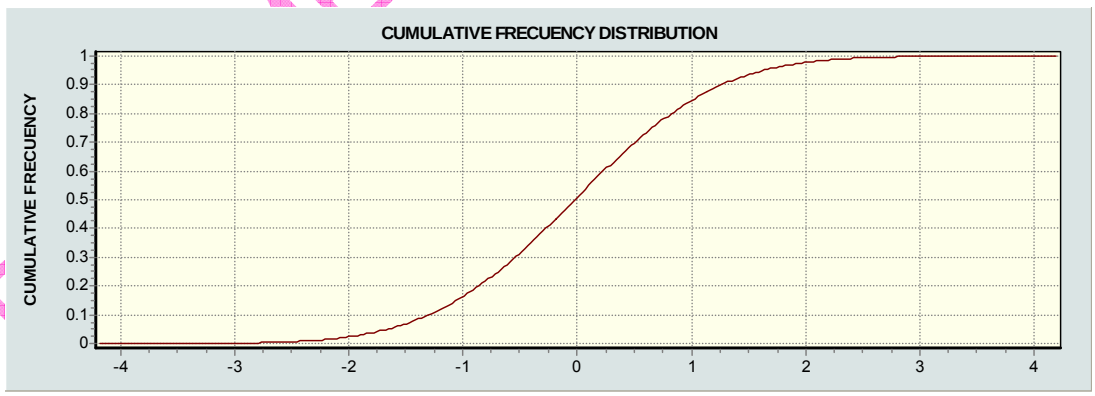


FIGURE 3.5b Cumulative distribution function of a normal distribution

3.5 Standard distributions

Many kinds of events conform to one of several standard frequency distributions that have been studied at length and the equation of distribution well established. The probability of such events can be determined quite easily.

One of the most common problems in designing hydraulic structures is estimating the magnitude of the flood corresponding to a given return period. For as long as the available data is sufficiently long, this procedure is reduced to simply fitting a frequency distribution to the annual floods. The estimate is then obtained by extrapolating the fitted frequency distribution to the desired return period. In this chapter, examples are carried out using a spreadsheet program and SPELL-Stat. The method used in fitting frequency distribution is the methods of moment. This is an analytical method, which depends upon the property of the distribution parameters, which in turn are uniquely defined by their moments. The first two (mean and variance), or maybe three (skewness) are generally sufficient. The number of moments required being equal to the number of parameters to be determined.

A probability distribution is characterized by location, scale and shape parameters. Location, scale, and shape parameters are typically used in modeling applications. The location parameter ***c*** depends upon the first order about the origin, the scale parameter ***a*** is related to the second central moment, and the shape parameter ***b*** depends upon the third central moment (Hall, 2002)

The general procedure in fitting a frequency distribution can be summarized as follows:

1. Choose the distribution to be fitted to the data;
2. Rank the data in ascending order of magnitude and compute the corresponding plotting position, F_i (probability of non-exceedence);
3. Convert the plotting positions to standardized variate, y_i;

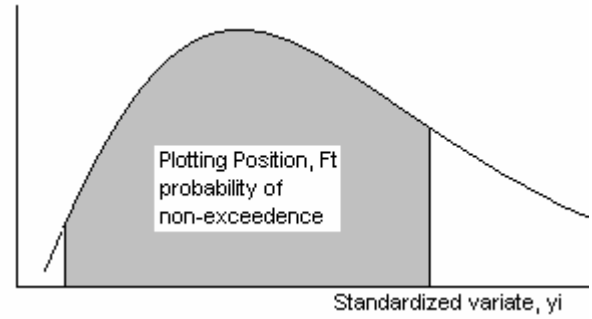


FIGURE 3.6 Plotting position and standardized variate

4. Compute the sample estimates of the mean, variance and (if necessary) skewness from the sample observations;
5. Estimate the location, *c* and scale, *a* parameters (and the shape, *b* parameter if required) from the moments;
6. Estimate the quantiles using the corresponding value of the standardized variate and the distribution parameters;
7. Determine the ninety-five percent confidence limits (two-tailed test) to the quantile estimate.

3.6 Some standard distributions

3.6.1 The Normal or Gaussian Distribution

The *probability distribution function* (PDF) of a normal distribution is given by

$$f(x) = \frac{1}{a\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left[\frac{x-c}{a} \right]^2 \right\} \qquad 3.4$$

where ***c*** is the location parameter and ***a*** is the scale parameter

The CDF of this distribution is

$$G(y) = \frac{1}{2\pi} \int_{-\infty}^{y} \exp\left(-\frac{y^2}{2} \right) dy \qquad 3.5$$

Normal distribution fitting follows the following steps:

1. Rank the data in ascending order

2. Compute the probability of non-exceedence, F_i

$$F_i = \frac{i-\alpha}{n+1-2\alpha} \qquad 3.6$$

where: i rank
α is equal to 0.375 (Blom's formula) for the normal distribution
n number of data

3. Convert the probability of non-exceedence derived from Eq. 3.5 to standardized variate, y_i using the table for standard normal distribution (or spreadsheet).

4. Compute the distribution parameters
   a = sample standard deviation
   c = sample mean

5. Estimates the quantiles using

$$X_{est} = ay + c \tag{3.7}$$

where y is the corresponding standardized variate

6. Determine the variability of the quantile estimates

$$\text{var } X_{est} = \frac{\sigma^2}{n}\left[1 + \frac{ny^2}{2(n-1)}\right] \tag{3.8}$$

where σ^2 is the sample variance

7. Compute the confidence limits based on 95%level of confidence (two-tailed test)

$$CL_{(x)} = X_{est} \pm t_{97.5,n-1}\sqrt{\text{var } X_{est}} \tag{3.9}$$

where $t_{97.5,n-1}$ is the value of Student's t-distribution for a 97.5 percent level of confidence (two-tailed test and n-1 degrees of freedom)

8. The magnitude of the T-year flood is then equal to

$$Q = ay + c \tag{3.10}$$

where ***a*** is the sample standard deviation, ***c*** is the sample mean, and ***y*** is the standardized variate corresponding to a given return period.

3.6.2 The Gumbel Distributrion (general extreme value)

The PDF of the extreme value type 1 (EV1) or Gumbel distribution is given as:

$$f(x_1) = \frac{1}{a}\exp\left[-\frac{x_1 - c}{a} - \exp\left(\frac{x_1 - c}{a}\right)\right] \quad 3.11$$

which may also be expressed in terms of its CDF

$$F(x_1) = \exp\left[-\exp\left(-\frac{x_1 - c}{a}\right)\right] \quad 3.12$$

where ***c*** is the location parameter, and ***a*** is the scale parameter

Gumbel distribution fitting follows the following steps:

1. Rank the data in ascending order
2. Compute the probability of non-exceedence, F_i

$$F_i = \frac{i - \alpha}{n + 1 - 2\alpha} \quad 3.13$$

where: i rank
α is equal to 0.44 (Gringorten's formula)
n number of data

3. Convert the probability of non-exceedence to standardized variate, y_i using

$$y_i = -\ln(-\ln F_i) \quad 3.14$$

4. Compute the distribution parameters

$$\sigma^2 = \frac{\pi^2 a^2}{6} \quad 3.15$$

$$\mu = c + 0.5772a \quad 3.16$$

where σ^2 is the variance, μ is the mean, ***c*** is the location parameter, and ***a*** is the scale parameter

5. Estimates the quantiles using

$$X_{est} = ay + c \qquad 3.17$$

where y is the corresponding standardized variate

6. Determine the variability of the quantile estimates

$$\text{var } X_{est} = \frac{a^2}{n}\left(1.17 + 0.196y + 1.099y^2\right) \qquad 3.18$$

7. Compute the confidence limits based on 95%level of confidence (two-tailed test)

$$CL_{(x)} = X_{est} \pm t_{97.5,n-1}\sqrt{\text{var } X_{est}} \qquad 3.19$$

where $t_{97.5,n-1}$ is the value of Student's t-distribution for a 97.5 percent level of confidence (two-tailed test and n-1 degrees of freedom)

8. The magnitude of the T-year flood is then equal to

$$Q = ay + c \qquad 3.20$$

where ***a*** is the sample standard deviation, ***c*** is the sample mean, and y is the standardized variate corresponding to a given return period.

Example 3.1 The following table contains the annual average daily discharges from a gauging station at Puerto Leon, in the Zulia River basin of Colombia. Choose an appropriate frequency distribution to represent the sample, and estimate the 100-year flood at this site. What is the 95% confidence interval for the flood of this return period?

Year	Q (m3/s)	Year	Q (m3/s)	Year	Q (m3/s)
1973	90.22	1982	143.27	1991	87.53
1974	132.51	1983	79.71	1992	79.70
1975	195.21	1984	96.94	1993	83.67
1976	152.08	1985	134.65	1994	108.41
1977	71.16	1986	109.67	1995	120.40
1978	99.38	1987	82.33	1996	130.70
1979	147.42	1988	145.01	1997	76.39
1980	76.25	1989	109.96	1998	116.54
1981	181.52	1990	173.77	1999	177.45

First let us fit a normal distribution to the data by following the steps listed in Section 3.6.1. The parameters of a normal distribution are ***a***, the scale parameter which is equal to the standard deviation and ***c***, the location

parameter which is represented by the sample mean. In this example ***a*** is found to be 36.40 m^3/s and c is equal to 118.59 m^3/s.

The 95% confidence limits are constructed by first ranking the data in ascending order of magnitude. The plotting positions are computed according to Blom's formula (Eq. 3.6) and then converted to standardized variate using the table for the areas under a normal curve. A spreadsheet program like EXCEL can also easily convert the plotting position to standardized variate.

Sorted Q	Rank	Fi	Yi	Xest	Var Xest	LCL	UCL
71.16	1	0.023	-1.997	45.91	150.660	20.87	70.95
76.25	2	0.060	-1.558	61.88	110.923	40.39	83.36
76.39	3	0.096	-1.303	71.16	92.326	51.56	90.77
79.70	4	0.133	-1.112	78.10	80.600	59.79	96.42
79.71	5	0.170	-0.955	83.81	72.331	66.46	101.16
82.33	6	0.206	-0.819	88.78	66.167	72.18	105.37
83.67	7	0.243	-0.696	93.24	61.433	77.25	109.23
87.53	8	0.280	-0.583	97.35	57.751	81.85	112.85
90.22	9	0.317	-0.477	101.21	54.888	86.09	116.32
96.94	10	0.353	-0.377	104.88	52.694	90.07	119.68
99.38	11	0.390	-0.280	108.41	51.070	93.83	122.99
108.41	12	0.427	-0.185	111.85	49.950	97.43	126.27
109.67	13	0.463	-0.092	115.23	49.294	100.91	129.56
109.96	14	0.500	0.000	118.59	49.078	104.30	132.88
116.54	15	0.537	0.092	121.94	49.294	107.62	136.26
120.40	16	0.573	0.185	125.32	49.950	110.90	139.74
130.70	17	0.610	0.280	128.76	51.070	114.18	143.34
132.51	18	0.647	0.377	132.30	52.694	117.49	147.11
134.65	19	0.683	0.477	135.97	54.888	120.85	151.08
143.27	20	0.720	0.583	139.82	57.751	124.32	155.33
145.01	21	0.757	0.696	143.93	61.433	127.94	159.92
147.42	22	0.794	0.819	148.40	66.167	131.80	164.99
152.08	23	0.830	0.955	153.36	72.331	136.01	170.71
173.77	24	0.867	1.112	159.07	80.600	140.76	177.39
177.45	25	0.904	1.303	166.01	92.326	146.41	185.61
181.52	26	0.940	1.558	175.30	110.923	153.81	196.78
195.21	27	0.977	1.997	191.27	150.660	166.23	216.31
a	36.40		T	2.04			
c	118.59						

TABLE 3.3 Normal distribution fitting for the annual average daily discharges of Zulia river

The quantile estimates or the estimated values of the data using normal distribution are determined using Eq. 3.7. The variances of the estimates are computed from Eq. 3.8. The confidence limits are obtained from Eq. 3.9 using a Student's t-value of 2.04 for a 97.5% level of confidence and 26 (n-1) degrees of freedom. The computations are summarized in Table 3.3 and the plot is shown in Figure 3.7

It can be seen in Figure 3.7 that the data can be described by a normal distribution. The magnitude of the 100-year return period flood can therefore be estimated using the derived parameters of a normal distribution. The probability corresponding to a return period of 100 years is 1/100 or 0.01. This represents the probability of occurrence or exceedence of the event. The probability of non-exceedence (plotting position, F_i) is then equal to 1-0.01 =

0.99. Using the table for areas under a normal curve, the standardized variate for a probability of non-exceedence equal to 0.99 is 2.326. The 100-year flood is estimated using Eq. 3.7 and found to be

Q_{100} = 36.4 (2.326) + 118.59 = 203.27 m^3/s

The confidence limits for the 100-year flood is computed as follows:

$$\text{var}\, Q_{100} = \frac{(36.4)^2}{27}\left[1 + \frac{27(2.326)^2}{2(27-1)}\right] = 186.988$$

$$CL_{(Q)} = 203.27 \pm 2.04\sqrt{186.988}$$

LCL = 175.37 m^3/s
UCL = 231.17 m^3/s

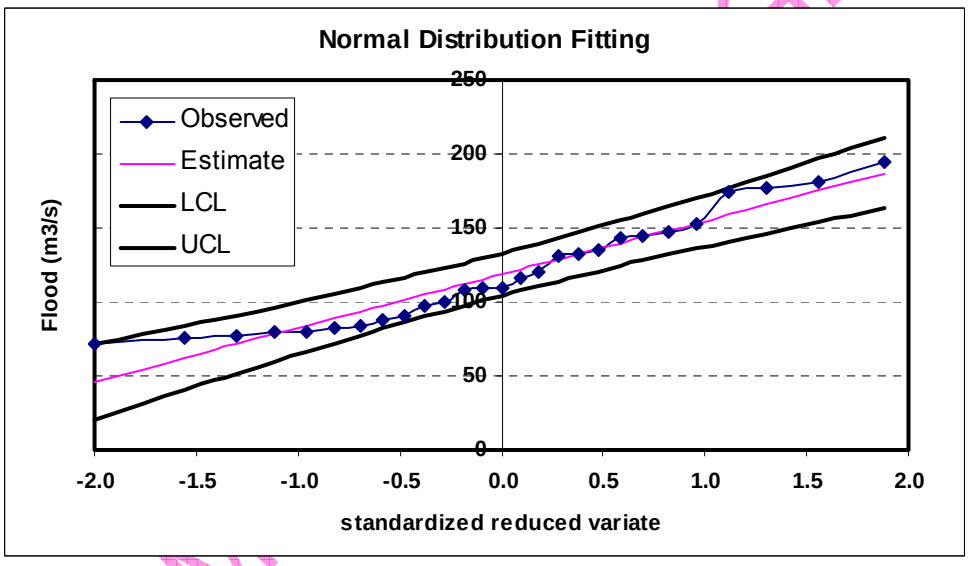


FIGURE 3.7 Plot of normal distribution fitting

It can be seen in Figure 3.7 that the data can be described by a normal distribution. With confidence limits set at 95%, it is expected that 95% of the data should lie within the confidence limits set and this is evident in Figure 3.7. The magnitude of the 100-year flood as estimated by the normal distribution is found to be 203.27 m^3/s. The estimate also lies between the confidence limits at 95% level of confidence.

The Gumbel distribution (EV1) is fitted to the annual average daily discharges of Zulia river by following the procedure described in Section 3.6.2. The parameters of the distribution ***a*** (scale) and ***c*** (location) are defined by Eqs. 3.15 and 3.16, respectively. The mean of the data is 118.5867 and the variance is 1325.108 so that the values of the parameters are

$$a = \sqrt{\frac{6(1325.108)}{\pi^2}} = 28.383$$

$$c = 118.5867 - 0.5772(28.383) = 102.20$$

In order to fit the 95% confidence limits to the data, the floods are ranked in ascending order of magnitude and the plotting positions computed using Gringorten formula (Eq. 3.13). The plotting positions are then converted to the corresponding standardized variate using Eq. 3.14.

Year	Q (m3/s)	Sorted Q	Rank	Fi	Yi	Xest	Var Xest	LCL	UCL
1973	90.22	71.16	1	0.021	-1.356	63.72	87.258	44.67	82.78
1974	132.51	76.25	2	0.058	-1.049	72.42	64.873	55.99	88.85
1975	195.21	76.39	3	0.094	-0.859	77.83	54.068	62.83	92.83
1976	152.08	79.70	4	0.131	-0.708	82.10	47.216	68.08	96.12
1977	71.16	79.71	5	0.168	-0.578	85.79	42.491	72.49	99.09
1978	99.38	82.33	6	0.205	-0.460	89.14	39.165	76.37	101.90
1979	147.42	83.67	7	0.242	-0.350	92.27	36.881	79.88	104.65
1980	76.25	87.53	8	0.279	-0.245	95.26	35.442	83.11	107.40
1981	181.52	90.22	9	0.316	-0.143	98.16	34.741	86.14	110.18
1982	143.27	96.94	10	0.353	-0.042	101.02	34.721	89.00	113.04
1983	79.71	99.38	11	0.389	0.058	103.86	35.362	91.73	116.00
1984	96.94	108.41	12	0.426	0.159	106.73	36.672	94.37	119.08
1985	134.65	109.67	13	0.463	0.262	109.63	38.684	96.94	122.32
1986	109.67	109.96	14	0.500	0.367	112.61	41.456	99.47	125.74
1987	82.33	116.54	15	0.537	0.475	115.68	45.077	101.98	129.38
1988	145.01	120.40	16	0.574	0.588	118.89	49.673	104.51	133.26
1989	109.96	130.70	17	0.611	0.707	122.26	55.415	107.08	137.45
1990	173.77	132.51	18	0.647	0.833	125.85	62.545	109.72	141.99
1991	87.53	134.65	19	0.684	0.970	129.72	71.399	112.48	146.96
1992	79.70	143.27	20	0.721	1.118	133.95	82.466	115.42	152.47
1993	83.67	145.01	21	0.758	1.284	138.65	96.477	118.61	158.69
1994	108.41	147.42	22	0.795	1.472	143.99	114.579	122.15	165.82
1995	120.40	152.08	23	0.832	1.692	150.24	138.712	126.21	174.26
1996	130.70	173.77	24	0.869	1.961	157.86	172.466	131.07	184.65
1997	76.39	177.45	25	0.906	2.311	167.80	223.558	137.30	198.30
1998	116.54	181.52	26	0.942	2.826	182.42	313.322	146.31	218.53
1999	177.45	195.21	27	0.979	3.870	212.04	548.542	164.26	259.81
mean	**118.59**	**a**	**28.38**	**t**	**2.040**				
variance	**1325.11**	**c**	**102.20**						

TABLE 3.4 Gumbel distribution fitting

The flood estimates are obtained using Eq. 3.17. For a 97.5% level of confidence and a degree of freedom equal to 26, the value of the Student's t distribution is found to be 2.04. The 95% confidence limits are computed using Eq 3.19. The solution is presented in Table 3.4 and the plot shown in Figure 3.8

Figure 3.8 shows the observed floods, the estimated floods using Gumbel distribution and the confidence limits at 95% level of confidence. It can be seen in the plot that the observed variables are contained within the confidence limits set. This shows that Gumbel distribution can be used to

describe the annual average daily discharges of Zulia river and can be used to estimate the magnitude of the 100-year flood.

To estimate the magnitude of the 100-year flood, the probability (plotting position) of 0.99 was converted to standardized reduced variate and was found to be 4.6. Using Eq. 3.17 the magnitude of the 100-yr flood was found to be:

Q_{100} = 28.383(4.6) + 102.2 = 232.77 m^3/s

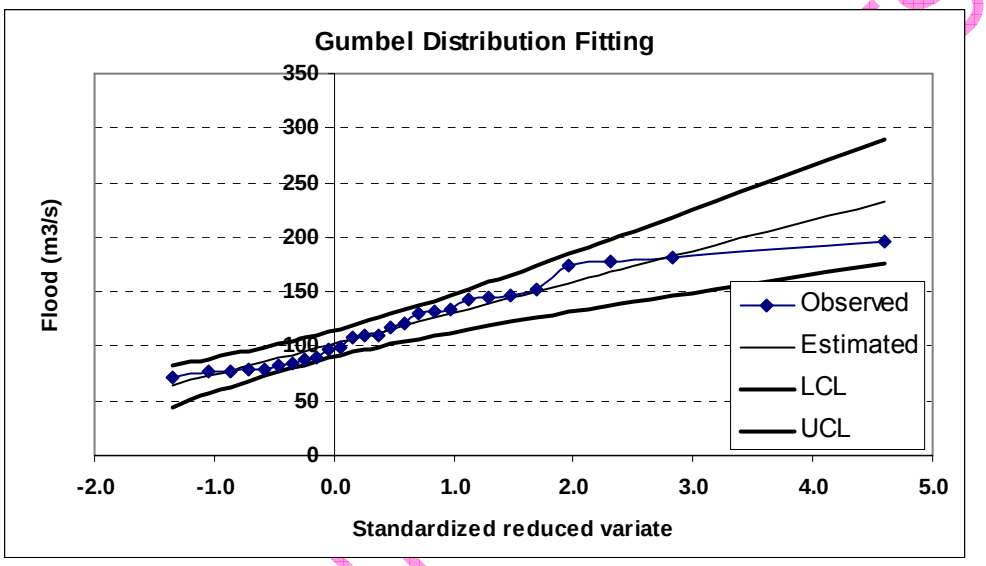


FIGURE 3.8 Plot of Gumbel distribution

The 95% confidence limits for the 100-year flood were set as follows:

$$\text{var}\, Q_{100} = \frac{(28.383)^2}{27}\left(1.17 + 0.196(4.6) + 1.099(4.6)^2\right) = 755.683$$

$$CL_{(Q)} = 232.77 \pm 2.04\sqrt{755.683}$$

LCL = 176.69 m^3/s
UCL = 288.85 m^3/s

3.6.3 Exponential Distribution (Partial Duration Series)

Another frequency distribution used in frequency analysis is the *Pearson type 3 distribution.* Its PDF may be expressed as

$$f(x) = \frac{1}{a^b \Gamma(b)} (x-c)^{b-1} \exp\left(-\frac{x-c}{a}\right) \qquad 3.21$$

where a, b, and c are the scale, shape and location parameters and Γ( ) is the complete Gamma function.

The Pearson type 3 distribution has 2 special cases when the location and shape parameters take on particular values. The first case is when c = 0 and the second special case is the exponential distribution. This distribution is especially used for partial duration series analysis and its PDF is given by

$$f(x) = \frac{1}{a} \exp\left(-\frac{x-c}{a}\right) \qquad 3.22$$

The distribution has affixed skewness of 2 and its CDF is reduced to

$$G(y) = 1 - \exp(-y) \qquad 3.23$$

Annual series analysis requires a continuous record of ideally 20 to 30 years. When the record is ten years or less, the annual series approach provides too few data to discriminate between the estimates of high return period events produced by fitting different frequency distributions. For this case, the partial duration series approach based upon the *peak-over-threshold (POT)* model, can offer substantial advantages (Hall, 2002). This model describes the distribution of ***m*** independent events within the ***n*** years of record, where ***m***>***n***, all of which exceed a threshold magnitude X_{th}.

Fitting an exponential distribution to partial duration series follows the following steps:

1. Rank the data in ascending order

2. Compute the probability of non-exceedence, F_i

$$F_i = \frac{i-\alpha}{n+1-2\alpha} \qquad 3.24$$

where: i rank
α is equal to 0.44 (Gringorten formula)
n number of data

3. Convert the probability of non-exceedence to standardized variate, y_i using

$$y_i = -\ln(1 - F_i) \qquad 3.25$$

4. The distribution parameters can be determined using the following methods:

   a) METHOD I: Choosing the threshold and treating the ***m*** events as a random sample from an exponential distribution with an unknown parameter ***a***;

      Using either moments or maximum likelihood

$$a = \overline{X} - X_{th} \qquad 3.26$$

      where $\overline{X}$ is the mean of the ***m*** events and X_{th} is the threshold

   b) METHOD II: Choosing ***m*** and estimating both the threshold and the parameter ***a*** from the sample

      The maximum likelihood estimates of ***a*** and X_{th} are

$$a = m\frac{\overline{X} - X_{\min}}{m-1} \qquad 3.27$$

$$X_{th} = X_{\min} - \frac{a}{m} \qquad 3.28$$

      where X_{min} is the smallest sample event

5. The quantile estimate is given by

$$X_i = X_{th} + ay_i \qquad 3.29$$

6. Variability of quantile estimate

   Method I:

$$Var\ X_{est} = \frac{a^2}{n}\left[1 + \frac{(y_i)^2}{K}\right] \qquad 3.30$$

   where K=m/n

Method II:

$$VarX_{est} = \frac{a}{m}\left[\frac{(1-y_i)^2}{m-1} + (y_i)^2\right] \qquad 3.31$$

7. Compute the confidence limits based on 95%level of confidence (two-tailed test)

$$CL_{(x)} = X_{est} \pm t_{97.5,n-1}\sqrt{\text{var } X_{est}} \qquad 3.32$$

where $t_{97.5,n-1}$ is the value of Student's t-distribution for a 97.5 percent level of confidence (two-tailed test and n-1 degrees of freedom)

8. The magnitude of the T-year flood is then equal to

$$X_t = X_{th} + a\ln(KT) \qquad 3.33$$

Example 3.2 The following table shows the flood peaks over a threshold of 49 m^3/s observed over a period of 50 months at station 45806, the River Creedy, a catchment of 262 km^2 in the Southwest of England (Hall, 2001). Choose the appropriate distribution and estimate the magnitude of a 50-year flood in this site.

Date	Flood m3/s	Date	Flood m3/s	Date	Flood m3/s
19-Jan-65	49.32	19-Jan-66	54.73	23-Dec-67	53.94
29-Jan-65	86.57	22-Oct-66	59.60	8-Jan-68	146.32
17-Nov-65	87.97	19-Dec-66	52.07	8-May-68	67.45
9-Dec-65	62.10	15-Feb-67	50.84	10-Dec-68	96.62
2-Jan-66	65.52	20-Feb-67	56.33	24-Dec-68	127.52
10-Jan-66	181.70	8-Mar-67	53.00	7-Jan-69	89.18
25-Jan-66	118.76	15-Oct-67	54.10	17-Jan-69	60.72
25-Feb-66	71.19	31-Oct-67	109.68	22-Feb-69	139.41
16-Apr-66	122.44	4-Nov-67	78.22	24-Feb-69	60.72

Since the threshold is given, the parameters of the distribution are derived from Method I. The distribution parameter ***a*** is obtained from Eq. 3.26, the value of which is 83.56 – 49 = 34.56. The data are ranked in ascending order and the plotting positions are obtained using Gringorten formula. The standardized variates corresponding to each plotting position follow Eq. 3.25. The flood estimates are obtained from 3.29 and the corresponding variances from Eq. 3.30. The 95% confidence limits are computed from Eq. 3.32 using a Student's t value of 2.04 for 26 degrees of freedom. The solution is summarized in Table 3.5 and the plot is shown in Figure 3.9.

Date	Q (m3/s)	Sorted Q	Rank	Fi	Yi	Xest	Var Xest	LCL	UCL
19-Jan-65	49.32	49.32	1	0.021	0.021	49.72	238.847	18.19	81.25
29-Jan-65	86.57	50.84	2	0.058	0.059	51.05	238.983	19.51	82.58

17-Nov-65	87.97	52.07	3	0.094	0.099	52.43	239.262	20.87	83.98
9-Dec-65	62.10	53.00	4	0.131	0.141	53.86	239.703	22.28	85.45
2-Jan-66	65.52	53.94	5	0.168	0.184	55.36	240.326	23.74	86.99
10-Jan-66	181.70	54.10	6	0.205	0.229	56.93	241.156	25.25	88.61
25-Jan-66	118.76	54.73	7	0.242	0.277	58.57	242.219	26.82	90.32
25-Feb-66	71.19	56.33	8	0.279	0.327	60.29	243.550	28.46	92.13
16-Apr-66	122.44	59.60	9	0.316	0.379	62.11	245.189	30.16	94.05
19-Apr-66	54.73	60.72	10	0.353	0.435	64.02	247.183	31.95	96.09
22-Oct-66	59.60	60.72	11	0.389	0.493	66.05	249.589	33.82	98.27
19-Dec-66	52.07	62.10	12	0.426	0.556	68.20	252.479	35.78	100.61
15-Feb-67	50.84	65.52	13	0.463	0.622	70.49	255.938	37.86	103.13
20-Feb-67	56.33	67.45	14	0.500	0.693	72.95	260.077	40.05	105.85
8-Mar-67	53.00	71.19	15	0.537	0.770	75.60	265.033	42.39	108.81
15-Oct-67	54.10	78.22	16	0.574	0.853	78.47	270.987	44.89	112.05
31-Oct-67	109.68	86.57	17	0.611	0.943	81.59	278.173	47.57	115.62
4-Nov-67	78.22	87.97	18	0.647	1.043	85.03	286.911	50.48	119.59
23-Dec-67	53.94	89.18	19	0.684	1.153	88.85	297.641	53.65	124.04
8-Jan-68	146.32	96.62	20	0.721	1.277	93.14	310.996	57.17	129.12
8-May-68	67.45	109.68	21	0.758	1.419	98.05	327.917	61.10	134.99
10-Dec-68	96.62	118.76	22	0.795	1.585	103.76	349.891	65.60	141.92
24-Dec-68	127.52	122.44	23	0.832	1.783	110.61	379.422	70.88	150.35
7-Jan-69	89.18	127.52	24	0.869	2.031	119.17	421.176	77.30	161.03
17-Jan-69	60.72	139.41	25	0.906	2.360	130.56	485.211	85.63	175.50
22-Feb-69	139.41	146.32	26	0.942	2.856	147.68	599.473	97.73	197.63
24-Feb-69	60.72	181.70	27	0.979	3.880	183.08	904.674	121.72	244.44
mean	83.56	a	34.6	T	2.04				
threshold, Xth	49	K	5.4						

TABLE 3.5 Partial duration series fitting – METHOD I

Figure 3.9 shows the fitted distribution to the data. It can be seen that the observed floods lie within the confidence limits set at 95% level of confidence. The parameters estimated from the distribution can be used to estimate the magnitude of the 50-year flood. The magnitude of the 50-year flood is estimated using Eq 3.33. There are ***m*** = 27 independent events in ***n*** = 5 years so that the event ratio ***k*** = 5.4. The magnitude of the 50-year flood is computed as

$Q_{50} = 49 + 34.56(5.4*50) = 242.48\ m^3/s$

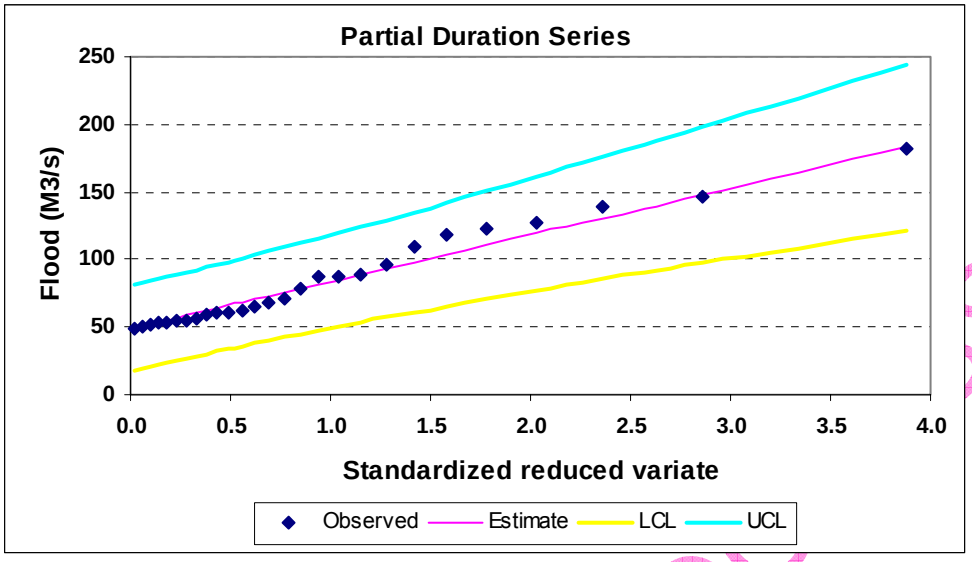


FIGURE 3.9 Plot of partial duration fitting – METHOD I

Example 3.3 The following table presents twelve flood peaks recorded over a three-year period at a gauging station in Europe. Estimate the ten-year flood at this site using the *peak-over-threshold* method and the second method of data extraction. Construct the 95% confidence limits to the fitted distribution (Hall 2001).

Date	Flood m3/s	Date	Flood m3/s	Date	Flood m3/s
24-Jan-89	11.78	22-Mar-90	13.12	1-Jan-91	10.55
2-Feb-89	15.57	18-May-90	20.64	3-Apr-91	49.45
23-Jul-89	17.25	29-Aug-90	18.17	14-Jul-91	22.44
30-Nov-89	29.34	3-Oct-90	26.96	24-Dec-91	34.55

The parameters of the distribution are obtained using Method II because the threshold was not defined. There are ***m*** = 12 independent events (floods) in ***n*** = 3 years, so that the event ratio, ***k*** = 4. The value of the distribution parameter ***a*** is computed using Eq. 3.27 where the average mean value $\bar{x}$ = 22.485 and the minimum values x_{min} = 10.55, so that ***a*** = 13.02. The threshold value, X_{th} = is obtained from Eq. 3.28 as 9.465 m^3/s.

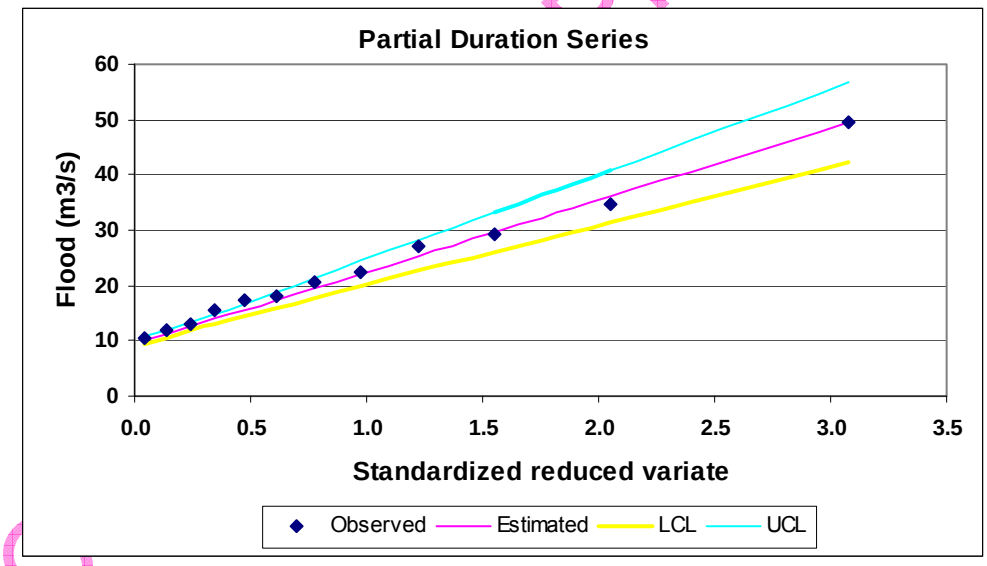
To construct the 95% confidence limits, the data must be ranked in ascending order of magnitude. The plotting positions are obtained using Gringorten formula and the corresponding standardized variate follow Eq. 3.25. The variance of the estimates is computed using Eq. 3.31. The confidence limits are fitted using a ***t*** value of 2.2 for degrees of freedom equal to 11. Table 3.6 contains the summary of the distribution fitting and Figure 3.10 shows the plot of the distribution fitting.

Date	Flood m3/s	Sorted Q	Rank	Fi	Yi	Xest	Var Xest	LCL	UCL
24-Jan-89	11.78	10.55	1	0.046	0.047	10.08	0.092	9.41	10.75
2-Feb-89	15.57	11.78	2	0.129	0.138	11.26	0.094	10.58	11.93
23-Jul-89	17.25	13.12	3	0.211	0.237	12.55	0.118	11.80	13.31
30-Nov-89	29.34	15.57	4	0.294	0.348	13.99	0.173	13.08	14.91
22-Mar-90	13.12	17.25	5	0.376	0.472	15.61	0.269	14.47	16.75
18-May-90	20.64	18.17	6	0.459	0.614	17.46	0.424	16.03	18.89
29-Aug-90	18.17	20.64	7	0.541	0.779	19.61	0.664	17.82	21.40
3-Oct-90	26.96	22.44	8	0.624	0.978	22.19	1.037	19.95	24.43
1-Jan-91	10.55	26.96	9	0.706	1.225	25.42	1.633	22.60	28.23
3-Apr-91	49.45	29.34	10	0.789	1.555	29.71	2.653	26.13	33.29
14-Jul-91	22.44	34.55	11	0.871	2.050	36.16	4.669	31.40	40.91
24-Dec-91	34.55	49.45	12	0.954	3.075	49.50	10.682	42.31	56.69
Mean	22.485	a	13.02	t	2.2				
Min	10.55	Xth	9.465						

TABLE 3.6 Partial duration fitting – METHOD II

The observed floods follow a peak-over-threshold distribution as shown in Figure 3.10. The parameters derived in this distribution can be used to estimate the magnitude of the 10-year flood. Using Eq. 3.33 the estimate of the 10-year flood is equal to

$Q_{10} = 9.465 + 13.02 \ln(4*10) = 57.494 \text{ m}^3/\text{s}$



3.6.4 Distribution fitting using SPELL-Stat

Distribution fitting can be carried out using SPELL-Stat. It is as easy as opening the file containing the data, clicking *VIEW* in the main menu and choosing the distribution to be fitted. The estimate of the T-year flood can be determined by clicking *Estimate* in the plot window. A dialogue window will show asking for the required return period. Shown in Figures 3.11 to 3.14 are the plots of data fitted to the distribution described earlier.

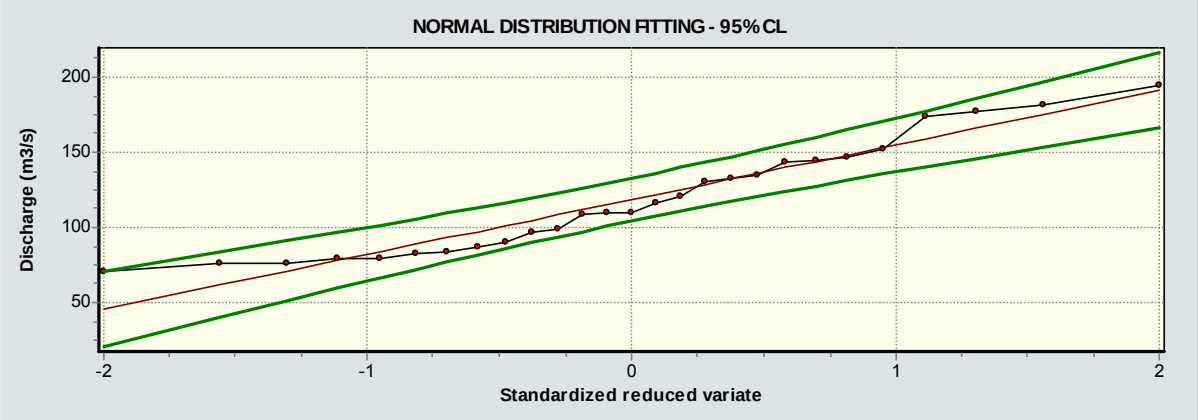


FIGURE 3.11 Normal distribution fitting by SPELL-Stat for the annual average daily discharges of Zulia river in Colombia.

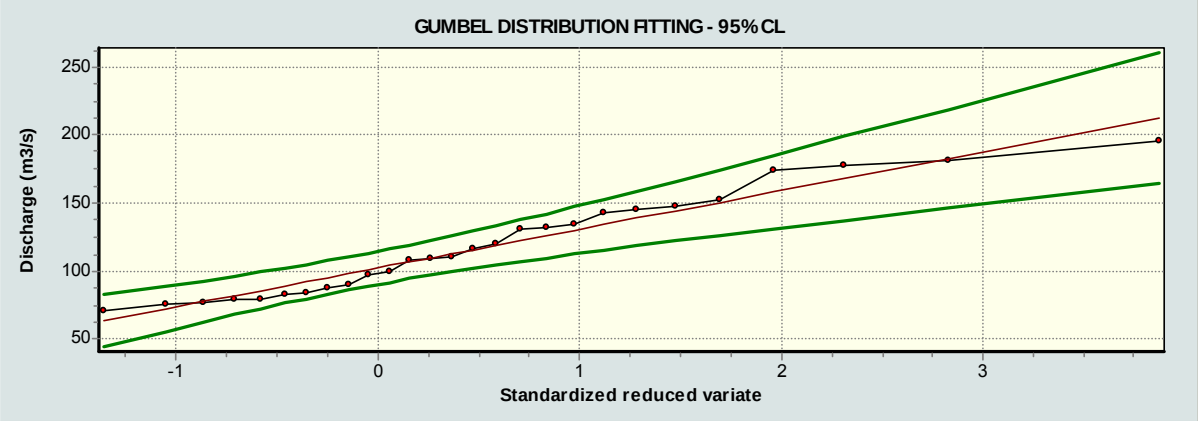


FIGURE 3.12 Gumbel distribution fitting by SPELL-Stat for the annual average daily discharges of Zulia river in Colombia.

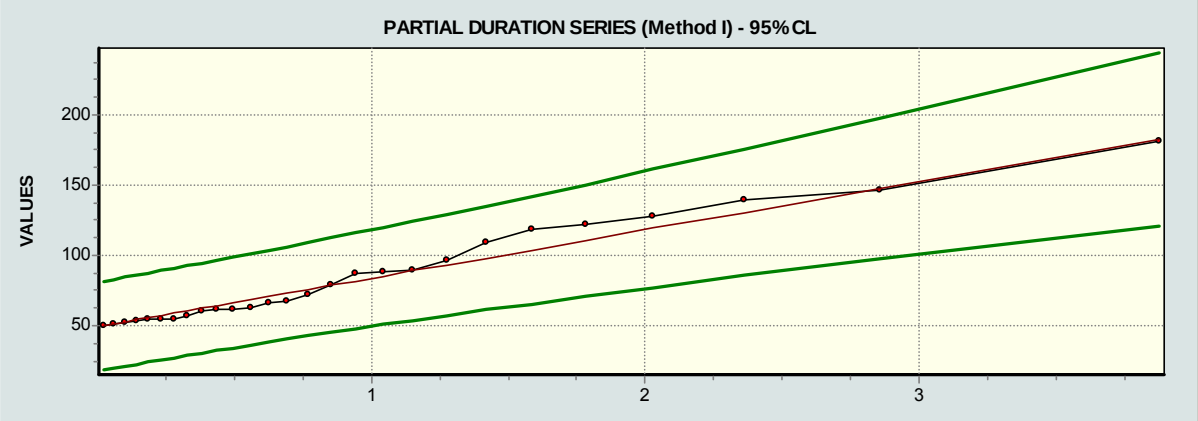


FIGURE 3.13 Peak-over-threshold model (Method I) by SPELL-Stat applied to flood measurements of River Creedy, England.

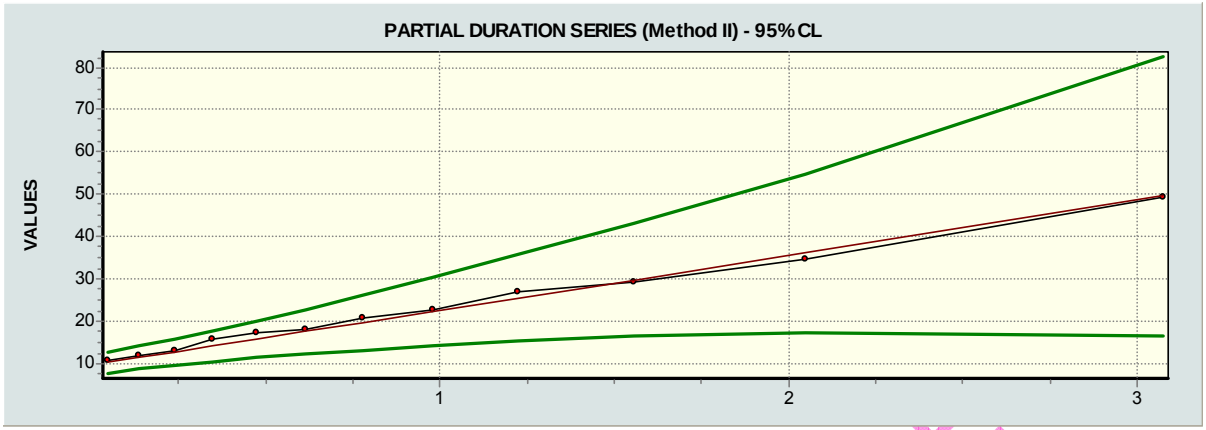


FIGURE 3.14 Peak-over-threshold model (Method II) by SPELL-Stat applied to flood measurements in a gauging station in England.

4. REFERENCES

Anderson, R.L., 1942. Distribution of the Serial Correlation Coefficients, Annals of Mathematics and Statistics, 13, 1-13

Alexandersson, H., 1986. A Homogeneity Test Applied to Precipitation Data, Journal of Climatology, Vol. 6: 661-675

Alexandersson, H. and Moberg, A., 1997. Homogeneity of Swedish Temperature Data-Part I, International Journal of Climatology, Vol. 17: 23-34

Arkins, H. and Colton, R.R. 1964. Statitical Methods, Barnes and Nobles, New York

Box, G.E.P. and Jenkins, G.M., 1970. Time Series Analysis, forecasting and Control, Holden-Day, san Francisco

Burn, D.H. and HagElnur, M.A., 2002. Detection of Hydrologic Trends and Variability, Journal of Hydrology 255: 107-122

Dahmen, E.R. and Hall, M.J., 1990. Screening of Hydrological Data: Test for Stationarity and Relative Consistency, Publication 49, ILRI, Wageningen, The Netherlands

Hall, M.J., 2002. Statistics and Stochastic Processes in Hydrology. Lecture notes HH296/02/1, IHE Delft, The Netherlands

Helsel, D.R. and Hirsch, R.M., 1992. Statistical Methods in Water Resources: New York, Elsevier Science Publishers

Laat, P.J.M. de, 2001. Workshop in Hydrology. Lecture notes HH275/01/1, IHE Delft, The Netherlands

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, 2004

Pettitt, A.N., 1979. A Non-Parametric Approach to the Change-Point Problem. Applied Statistics, 28 (2), 126-135

Salas, J.D., Delleur J.W., Yevjevich, V. and Lane, W.L., 1980. Applied Modelling of Hydrologic Time Series. Water Resources Publications, Colorado, U.S.A.

Walpole, R.E. and Myers, R.H. 1993. Probability and Statistics for Engineers and Scientists, 5th Edition. Mcmillan Inc., New York