Zekâi Şen

Innovative Trend Methodologies in Science and Engineering

Springer

Innovative Trend Methodologies in Science and Engineering

Zekâi Şen

Innovative Trend Methodologies in Science and Engineering

Springer

Zekâi Şen
Turkish Water Foundation
Istanbul
Turkey

and

Faculty of Meteorology and Arid Lands,
Excellency Center for Climate Change
Research, Faculty of Earth Sciences
King Abdulaziz University
Jeddah
Saudi Arabia

ISBN 978-3-319-52337-8 ISBN 978-3-319-52338-5 (eBook)
DOI 10.1007/978-3-319-52338-5

Library of Congress Control Number: 2016963783

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

RAHMAN VE RAHİM OLAN ALLAH'IN ADI İLE
IN THE NAME OF ALLAH, THE MOST MERCIFUL, THE MOST COMPASSIONATE

I hope that each individual will try to perform his/her intellectual ability and moral behavior along an increasing trend for the common share and prosperity of humanity in homogeneous and isotropic manner

Preface

Scientific and technological developments in any discipline have become heavily dependent on the digital data treatment for better and refined deductions from available time series records that reflect the behavior of natural phenomena or artificial events as for their performances. These phenomena and events are rather complex, uncertain at times, vague, and even incomplete in their past records, which are also embedded with some deterministic components such as linear or nonlinear trends, sudden jumps, and seasonalities, each of which provide useful information for prediction of the future behavior so as to be able to control the natural events to a certain extent. Especially, trend component is the most sought one, because it shows the direction of general tendency within the partially uncertain events and especially since four decades their search has become a very significant task concerning climate change effects on environmental, social, and health aspects; economic growth indices; business affairs; and industrial production quality controls.

Currently, there is a trend in automation and data exchange in manufacturing technologies that leads to a new paradigm shift in industry that is now referred to as the Industry 4.0, which will be empowered only with innovative methodological procedures. Trend identification, determination, future extension, and de-trending procedures will gain refined and progressive advancement that will pave way towards better management and control of the phenomenon concerned in scientific, technological, engineering, environmental, social, economic, business, and health aspects. The success of industrial machines to predict failures and trigger maintenance processes autonomously or self-organized logistics, which react to unexpected sudden changes (jumps) or gradual and monotonic expected changes in the behavior of the phenomenon or production concerned. In order to arrive at meaningfully guiding information, it is necessary to provide useful insight into the prediction and management procedures through data processing by means of innovatively advanced analytical approaches and algorithms. The information generation algorithms should be able to detect and address visible and hidden issues in environmental changes such as the climate change impacts on different disciplines, machine degradations, depreciations, or improvements in the final industrial production.

In order to achieve effective prediction, management, and information generation, one of the most important data processing issues is the identification and determination of trend component in a given record especially in the form of time series. This is the main purpose of this book where after an effective literature review in the first three chapters different types of innovative trend analysis methodologies are presented in the science philosophical, logical, rational, and linguistic foundation leading to the probabilistic, statistical, and stochastic aspects for better and refined trend identification. The innovative trend template provides first of all visual inspection for verbal information deductions not only for holistical purposes, but also for providing better views in terms of at least three categories as "low," "medium," and "high" record values within the data. Spatial and partial trend component identification methodologies are also provided with simple but illuminating examples. Innovative trend simulation studies and trend test statistical procedures are explained along with actual example applications from different parts of the world. Apart from the classical trend analysis on the average, possible trend behavior in terms of standard deviation is also presented with innovative approaches under the title of variability. Last but not least, after a brief explanation of fuzzy logic modeling principles, fuzzy trend analysis fundamentals are explained. In the final chapter, several examples are presented concerning the climate change impact in terms of trend analyses.

The content of this book is an outcome from a series of lectures by the author at the Technical University of Istanbul and also at the King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia. Furthermore, many aspects of the trend analysis have been discussed with international students from different countries personally and through the electronic communication systems. I appreciate all of these precious discussions, which accumulated and led to the production of this book. It will give me pleasure and self-satisfaction if the content of this book serves to those who are interested in the trend analysis.

In writing an international book one has to be very patient and confine his/her attention for many hours, days, months, and even years without care for many other things and therefore needs the support of many others. I appreciate and thank those who have encouraged me to write this book and at the top of the list is my wife Mr. Fatma Şen for her endurance, patience, and encouragement.

Çubuklu, Istanbul, Turkey Zekâi Şen
2016

Contents

Introduction

1

Abstract

Trend analysis has an interdisciplinary context that is shared by many researchers all over the world. The preliminary recommendation in this chapter is about visual trend examination and identification in a given time series to feel what are the possibilities of trend existence either holistically or partially. In this manner the researcher will be able to decide which type of the probabilistic, statistical, and mathematical approach for its objective determination. A brief discussion about trend analysis usage is presented on the basis of a set of disciplines. Additionally, pros and cons about trend analysis approaches are presented briefly and finally future trend research directions are mentioned with the purpose of this book.

Keywords

Concept · Definition · Disciplines · Purpose · Trend · Visualization · Mathematics · Statistics

1.1 General

Modern lifestyle at every aspect can be improved further through the measurements, mathematical models, control and prediction for future time periods at short-, medium- and long terms. With the computational facilities and treatment of data many social, economic, health, earth, environment and engineering systems can be modeled for prediction purposes. In practice, natural or artificial time series records at regular time intervals are available, but unfortunately they are evaluated for certain purposes without a complete description of deterministic and stochastic parts. Each time series is full of various qualitative and quantitative features, which are ready for

Z. Şen, *Innovative Trend Methodologies in Science and Engineering*,
DOI 10.1007/978-3-319-52338-5_1

logical, rational, analytical, probabilistic, statistical, stochastic, and fuzzy scientific assessments for deduction of useful information in practical applications. Time series component identifications are the most important issues that are backbones of fruitful developments in any discipline.

Among the most significant components of a time series is the trend evolution lines, which indicate continuous increase, decrease, or stability (balance, neutrality) along the time axis. Trends are desirable in many human activities depending on the final goal. For instance, the Olympic Games record breakings are indicators of increasing trend, because each game is sought to perform better than the previous ones. In general, any societal development is measured through various indices, which indicate either a positive or negative change or neural state. Any development, concerning a system, can be felt first intuitively and subjectively and later on objective decision can be achieved through the necessary measurements, which provide databases, and subsequently, their treatments, by convenient scientific methodologies leading to useful information.

Time series trend analysis research and application studies have increased during the last 25 years as a result of interest in the global warming and climate change impacts on natural events in addition to more refined economic and business prediction purposes. There are trend identification methodologies and statistical trend significance tests, but each one with different set of assumptions, which may not be simultaneously valid within the measurement data. In general, the following points are among the trend study purposes in various disciplines.

(1) Performance of any system toward better conditions, which implies an increasing trend embed within the time series,
(2) Measurement of system performance quantitatively, whether there is an improvement (positive trend) or depreciation (negative trend) by time,
(3) Assessment of any system as to its balance and steadiness about temporal evolution, which is reflected by a neutral (no trend) case,
(4) Random or stochastic behavior identification of a system after systematic variations (trend, shift, seasonality–periodicity) elimination from a given time series,
(5) Quality control of a manufacturing system such as factories and depreciation (decreasing trend) of the machine performances,
(6) Economic performance measure (increasing or decreasing trend) of any societal activity (business, economics, population, etc.) variation with time.

Trends are also indicators of significant correlation between successive event occurrences and time or among a set of events at different measurement locations in space. They are gradually introduced into the records, because of natural or man-made (artificial) effects. The shifts (jumps) are also due to the similar effects, but as rather sudden changes at times step by step. Most often, the gradual changes in environmental phenomena are results of global warming, climate change, population growth, and assessment of available resources. It is also possible to check gradual urbanization impacts on some changes due to the environmental activities

around a measurement site. Replacement of measurement instrument by a recent one or even with similar instrument and compulsory change of location may cause jump (shift), i.e., sudden step changes in the measurements or records. For this purpose, it is necessary that the records at any measurement site must not be numerical only, but also linguistically available causative and consequent information are needed for better identifications, interpretations, and predictions (see Chaps. 4 and 6).

Trend analyses are significant not only in the earth systems researches, but they have a larger domain of applications in different disciplines including quality control, economics, pattern recognition, digital signal processing and, in general, in data mining works. Among the various disciplines, the interest in trend analysis can be summarized along the following points.

(1) Any researcher that works with time series records would like to identify the system response in terms of systematic variations (trends, seasonality–periodicity and jumps), nonsystematic and uncertain residuals,
(2) Detection of trends indicate the general tendency toward increasing or decreasing directions or stability in the system response,
(3) Graphical representation of time series is the first step in data processing prior to quantitative theoretical technique applications and with naked eye one may search linguistically (verbally) for different variation patterns leading to preliminary qualitative and fuzzy information deductions (Şen 2010).

1.2 Trend Definition and Analysis

Any systematic and continuous increase or decrease along time axis is referred to as temporal trend, which may be in the linear or nonlinear forms. Trends are almost everywhere, but one begins to think or feels about their existence unless someone talks or when s/he is asked to provide evidence about them. For instance, since birth human beings are in increasing trend as for the tallness is concerned, but it is not linear throughout the life. Anyone feels in comfort, if his/her income increases with time. In general, trends are systematic changes in natural, social and artificial events over relatively longer time periods preferably with at least 30 or more sample.

There is a variety of trend definition depending on the purpose. In general, it is a tendency in which some event develops as increasing (upward) or decreasing (downward) changes. Each trend has a general direction, which may also be expressed in terms of drift, shift, swing, course, current, leaning, tendency, and inclination and synonymously as bias and bend. The term trend may also have social context as modern model, fashion, mode, type, style, vogue, and rage. Some examples are increasing as warming trend, fashion trend, upwards economic trend, downward trade trend, quality trend, stock market trend, business trend, etc.

Trend identification and detection procedure reviews are available in the literature (Esterby 1996; Hess et al. 2001). They lack a comprehensive text that covers potential applications in global warming, climate change, hydrology, environment, health, engineering and climatology disciplines, which are in increasing need for objective trend identification and prediction. Trend analysis reviews are focused on a single and monotonic trend search in a given time series with an emphasis on some classically favorable techniques only. In this book, after extensive literature review and criticisms, innovative trend identification and detection procedures are presented with rational and logical bases. No need to say that at the dawn of twenty first century, there is a need to highlight the importance of time series analysis in many disciplines including water resources planning, management and new issues of sustainable management, where innovative trend analysis techniques are ready to pave objective ways for logical interpretation and quantitative calculations.

1.2.1 Conceptual and Visual Trends

Mental and logical visualization change inspections with time are very helpful to generate illuminating ideas about the process concerned prior to any quantitative applications and theoretical developments. There are two ways to establish preliminary ideas about the temporal evolvement of any event performance. These are conceptualization of the event through mental experiments with a set of possible logical rules without any data availability and visualization of the event by means of graphical representations provided that there are measurements in the form of a time series.

In any scientific work, provided that the numerical data are available, the preliminary work is to try, visualize and explore the data behavior in graphical forms, which trigger the mind and creative thinking through the geometrical shapes. This is already reflected in saying that one picture is worth of thousand words. Especially, in the time series analysis, the temporal evolution of the phenomenon concerned can be grasped through the relevant graphs so as to see the random and systematic (trend, seasonality, sudden jumps) behaviors. The graphical representation and its visual interpretation provide valuable qualitative (verbal, linguistically) information, which are the basic ingredients of original scientific developments prior to any quantitative evaluation. Qualitative information deduction from the temporal behavior of a time series depends on the grasp and intuitive ability of a person, and although a set of subjective information are derived, among them there are also objective supportive ones. For instance, in order to be successful in a business, one may think about the basic principles and rules that are necessary to provide steadily increasing income economy, and accordingly, systematic implementations of the conceptualized systematic and rivalry rules into application. This is a simple way of increasing trend conceptualization.

In general, trend reflects the relationship between two variables; one may think and reach to a conclusion that there is a direct and increasing (or decreasing) relationship between two variables. Humans can conceptualize such two-variable

relationships, because in almost all cases, everybody is capable to appreciate logically whether there is a direct or indirect relationship between any two variables of his/her concern. Without any specialization, if someone is asked, say, about the possible relationship between the rainfall and its consequent runoff event, then s/he responds that there is a direct relationship, which means that increase in the former variable implies increase in the other in the form of increasing trend. After the decision on the direct or inverse relationship, the next question is whether this relationship is in the linear or nonlinear form? Another alternative to these questions is that there might not be any relationship between the two variables. As a result of these two questions, there are six possible and simple alternatives, each one of which is the answer for dependent and independent variable, Y and time, t, and evolution in the form of two-variable relationship in any discipline with mathematical certainty as in Fig. 1.1.

After the aforementioned conceptualizations and explanations, one can conclude that mentally, there are two questions; what are the proportionality relationships between two variables and what the shape (geometry) of the relationship is.

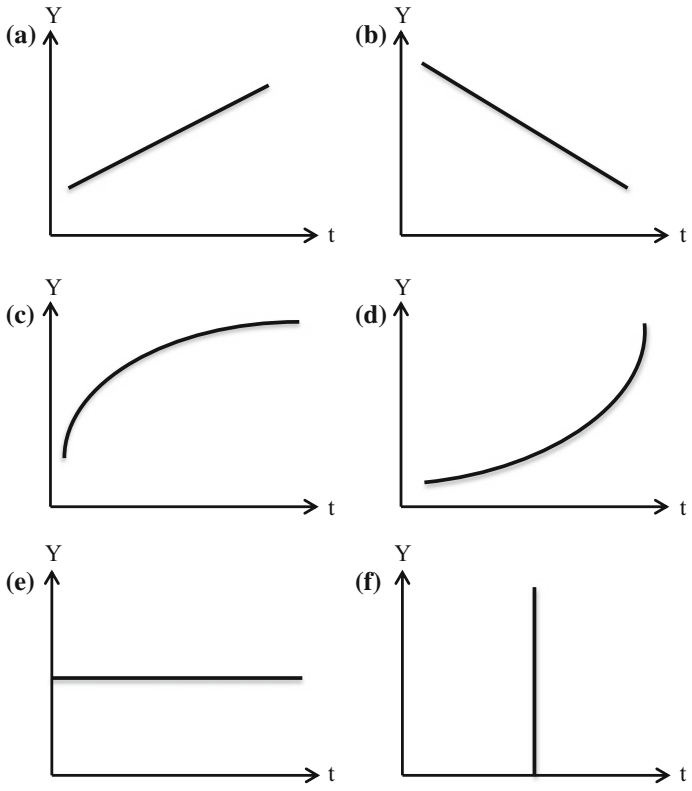


Fig. 1.1 Proportionality and geometry relationships, **a** direct-linear, **b** inverse-linear, **c** direct-nonlinear, **d** inverse-nonlinear, **e** no relation, **f** no relation

On the other hand, one can visualize temporal evolution of an event, provided that there are measurements, which help to fix the position, if the event performs on the event variable-time coordinate system. If there are no random errors in the measurements and the system is performing deterministically without any random component then the plot of the measurement data appears as a systematic scatter of points along one of the graphs in Fig. 1.1. However, in case of random variabilities in addition to the systematic variations, the resulting scatter of points appears in one of the six alternatives as in Fig. 1.2.

The trends in each one of these graphs are obvious and naked eye transforms visual information into mind and then appropriate qualitative interpretations can be deduced accordingly and they pave way toward probabilistic, statistical, stochastic or mathematical assessment. As for the measurements, whatever is the sensitivity, there are always measurement errors or inherent structural randomness during the evolution of the event. Trends in these graphs are representatives of systematic variability and deviations from each trend are the random or stochastic component of the variable.

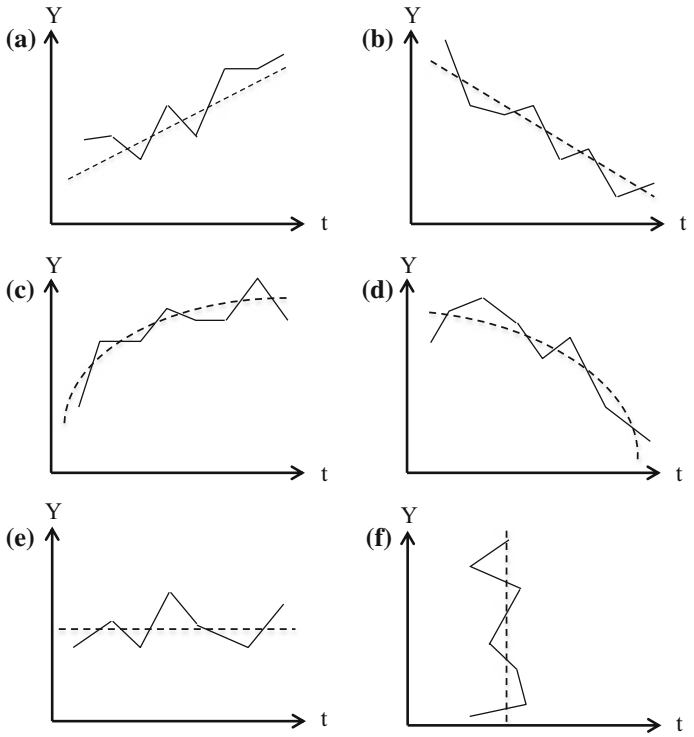


Fig. 1.2 Different time series and trends, **a** linearly increasing, **b** linearly decreasing, **c** nonlinearly increasing, **d** nonlinearly decreasing, **e** no trend (independence), and **f** no trend (independence)

Conceptual and visual trend evaluations provide linguistic (verbal) and partially fuzzy knowledge and information (Chaps. 7–8), which are qualitative, but they are the fundamentals of subsequent mathematical and statistical trend deduction, identification, determination and quantitative assessments as well as interpretations that are the main topics in the next chapters of this book. It should be emphasized at this junction that expertize about an event can be gained through such basic human intuitional and visual conceptions prior to any mathematical and statistical data treatment.

1.2.2 Mathematical Trend

Conceptual and visual trend alternatives provide the geometrical (functional) relationships of different forms without symbolic (mathematical) expressions, which provide preliminary objective definition, identification and description of a trend. If digital data are available in the form of time series then their treatment through scientific methodologies require first the establishment of mathematical foundations. For this purpose, simple mathematical functions must be kept in the library to study a time series for trend analysis. In practical applications, most often trend implies linear forms as increasing or decreasing tendencies. Hence, frequent trend searches are confined to Fig. 1.1a, b mathematically and Fig. 1.2a, b statistically. These have the simplest mathematical forms with two parameters a and b. For the linear trends in Fig. 1.1, the trend components are completely deterministic without random deviations, and therefore, the mathematical form is given as,

$$Y = a \pm bt, \tag{1.1}$$

where positive (negative) sign is for increasing (decreasing) trend. Equation (1.1) is the mathematical expression of Fig. 1.1a, b. The parameter values represent the intercept on the vertical axis and the slope of the line, respectively (see Fig. 1.3).

Equation (1.1) takes an uncertainty form by addition of an uncertain (random) element, u, with zero arithmetic average into the deterministic trend component as follows:

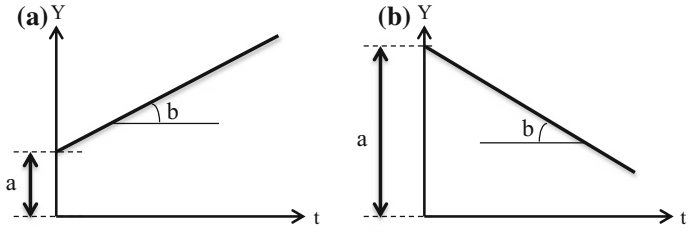


Fig. 1.3 Linear trend parameters

$$Y = a \pm bt + u. \tag{1.2}$$

This expression represents Fig. 1.2a, b, because of its linear structure. In conceptual trend works, sometimes it is possible to judge the parameter values, although there is no measurement. For instance, if there is no currency in the credit card one cannot buy goods, but depending on the amount of credit the amount of shopping increases, and therefore, the parameter a is equal to zero. Another example is the relationship between the rainfall, R, and the surface water flow, F, over a land piece, where there is no surface flow prior to the rainfall. Logically, one can conceptualize that the surface flow cannot be more than the rainfall, and hence, the slope parameter, b, value must be less than 1. In this example, the linear line also passes through the origin, and consequently, the trend line between the rainfall and surface water flow passes through the origin ($a = 0$) with slope $b < 1$. However, precise determination of the slope value necessitates simultaneous rainfall and surface flow measurements.

As for the nonlinearity trends, mathematical functions may be in different forms including polynomial, exponential, power, logarithmic and other functions, but they are not frequently used in practice. The most widely used nonlinear trend description is in the form of second order polynomial function as,

$$Y = a \pm bt \pm ct^2, \tag{1.3}$$

where c is an additional parameter that indicates the curvature of the nonlinear trend. It is the mathematical formulation of Fig. 1.1c, d. In case of uncertainty ingredient component existence, it can be rewritten as,

$$Y = a \pm bt \pm ct^2 + u. \tag{1.4}$$

This is the valid mathematical counterpart of Fig. 1.2c, d.

It is also possible to describe the trend components mathematically by differential equations. The first order differential expression represents linear trend and depending on its sign, it may be increasing (positive sign) or decreasing (negative sign) trend. However, the second order differential equation represents the nonlinear form again depending on the sign, where positive (negative) sign implies concave upward (downward) curvature. These alternatives are presented in Fig. 1.4.

The first order differential term of Eq. (1.1) leads to the following expression that is in accordance with Fig. 1.4a, b

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = \pm b \tag{1.5}$$

On the other hand, the first and second order differentials of Eq. (1.2) yield the following two differential equations.

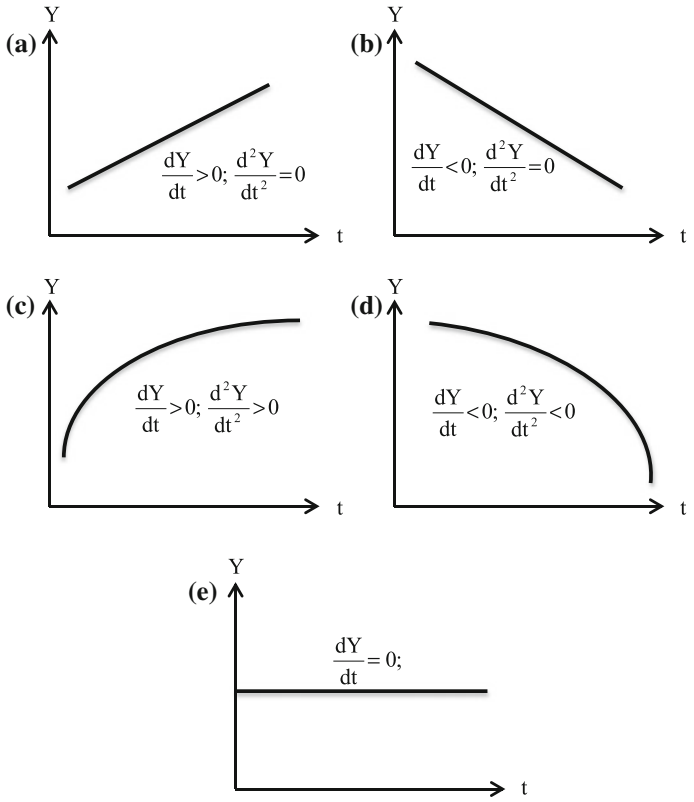


Fig. 1.4 Partial differentials of trends

$$\frac{dY}{dt} = \pm b \pm 2ct \tag{1.6}$$

and

$$\frac{d^2Y}{dt^2} = \pm 2c \tag{1.7}$$

Deterministic mathematical expressions without any random component are not used much in the data mining studies.

1.2.3 Statistical Trend

Variety of statistical tools is employed for trend analysis as will be explained in the following chapters and they are accessible to anyone who is interested in such works.

After the visual inspection of time series possible trend, sudden changes, outliers and random pattern around the trend component can be interpreted linguistically. Subsequently, data values can be converted to moving average value, which clarifies the background patters well because of smoothing (Chap. 2, Sect. 2.8.1). Finally, a regression model can be fitted to the final pattern (Chap. 3, Sect. 3.4.1).

In most contexts, trends are formed and interpreted from sets of data through probability, statistics and stochastic methodologies, which imply that there are random element embeds in the systematic deterministic components (trend, seasonality, step and shift-jump) in a time series. Natural and artificial time series measurements are not free of errors or inherent random components. In industrial machines, there are measurement errors but in natural, social and economic events additionally there are uncontrollable inherent random ingredients. As mentioned before, in Sect. 1.2.1 time series in Fig. 1.2 have random components, and therefore, deterministic equations cannot describe such time series completely. In order to represent them with trend component, an extra uncertainty component, u_i, is added to the mathematical expressions. The symbolic representation of a time series is $Y_1, Y_2, \ldots, X_n$ or Y_i ($i = 1, 2, \ldots, n$), where n is the number of samples. For statistical expression of time series with a linear trend component, the mathematical formulation can be written in the most explicit form as,

$$Y_i = a \pm bt_i + u_i. \tag{1.8}$$

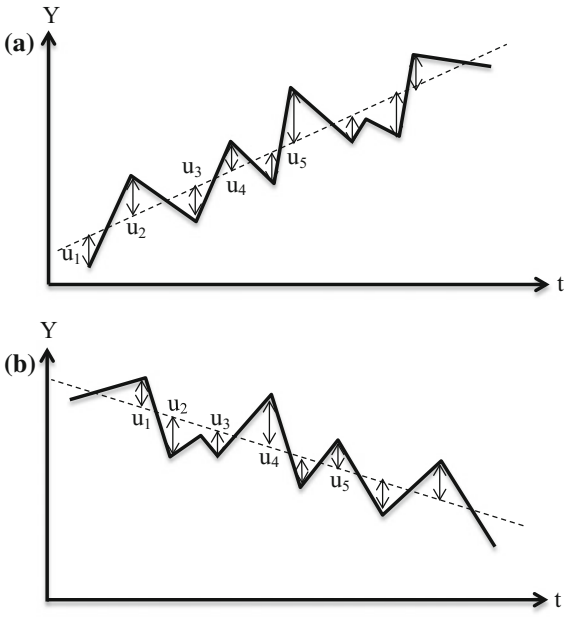


Fig. 1.5 Uncertainty components **a** increasing trend, **b** decreasing trend

Figure 1.5 indicates the graphical representation of the time series with uncertainty terms that are represented by vertical deviations from the trend line.

This figure indicates that the uncertainty components with respect to the trend line has + and – values. It brings to the mind logically as the first condition, for the best trend representation, the summation of the uncertainty terms must be equal to zero.

$$\sum_{i=1}^{n} u_i = 0 \tag{1.9}$$

This is a necessary condition but not sufficient, because the + and − deviations may be far away from the trend line, but still their summation may appear as zero. In the case of complete determinism, satisfaction of this condition is possible only when each one of the uncertainty amount is equal to zero, which is never the case in natural or artificial time series. Completely deterministic case can be represented by taking the absolute value of each uncertainty term and see whether their summation is equal to zero.

$$\sum_{i=1}^{n} |u_i| = 0 \tag{1.10}$$

However, in practice there is never a completely deterministic case, but there are uncertainties, and therefore, the logic deducts that the summation of absolute errors must be as small as possible (minimum). In practical calculations, instead of the absolute value, the square of each term is adapted for calculation convenience and the second and most significant condition for trend identification is the following expression.

$$\sum_{i=1}^{n} u_i^2 = \text{minimum} \tag{1.11}$$

Equations (1.9) and (1.11) are the basic requirements in the classical statistical regression analysis, which is one of the trend identification techniques in the literature (see Chap. 3).

1.3 Trend in Some Disciplines

In various aspects of life, trends are everywhere vivid, but they need close care for their identification. Trends may be beneficial or harmful depending on the disciplines, circumstances, and the purpose of the study or project.

1.3.1 Atmospheric Sciences

Each phenomenon in atmospheric sciences has uncertainty component and during long time durations trends are also observable in time series records. Among the major atmospheric scientists are meteorologists and climatologists who try and study characteristics of atmospheric physics, air mass movements and related processes in order to quantify and make reliable predictions about the atmospheric environmental activities. Weather forecasting is one of the major subjects, which involves uncertainties, but also local or temporal monotonic or partial trend components are also necessary in the prediction studies (see Chaps. 3 and 8). During the prediction studies, it is essential to identify and interpret climate trends for better understanding of the weather conditions and variabilities. Atmospheric studies are also very significant in air pollution control, agriculture, forestry, air and sea transportation, defense, and the study of possible trends in the Earth's climate, such as global warming, droughts, floods and ozone depletion. Currently, among the most important trend identification studies are concerned with the global warming, greenhouse gases concentrations and climate change impacts (Chaps. 7–8).

1.3.2 Environmental Sciences

Environmental changes can be detected and estimated through the classical statistical methodologies, where the trend analysis plays very significant role. In recent decades, tremendous amount of environmental data have accumulated in digital medium and their treatments, especially in the forms of time series, reveal decisive and conclusive results not only for the management but also for the quality and quantity trend variations and their controls.

It is necessary to understand, identify and quantify the possible temporal and spatial changes in different aspects of environmental sciences such as air, soil, and water quality and quantity. Especially, description of past trends and variations are important for understanding the basic generation mechanism of the phenomenon concerned and then to make future projections for the purpose of monitoring and combating the intervention of any undesired effect. Detection and estimation of possible trends in time series related to environmental sciences can be obtained through a set of familiar classical procedures. This is especially significant due to the explosion of environmental information records that provide a common basis for data treatment so as to reach meaningful and applicable results for better functioning of environmental systems.

1.3.3 Earth Sciences

In earth sciences, most often not only temporal but especially spatial trend searches are important in order to appreciate, understand and then explore the mineralogical, water, soil, oil, and different industrial raw material existences in a region. In earth

sciences especially trend surface analyses occupy a significant role in description of surface and subsurface geological tendencies (Chap. 6). Trend surface analysis helps to separate the spatially available data at irregularly scattered points in a study area into three components, namely regional trend effect, significant localized features and random component that cannot be expressed mathematically except by probabilistic, statistical, geostatistical and stochastic methodologies (Davis 1986).

1.3.4 Engineering

In many engineering aspects, trend analysis plays dominant role especially those events that have relationships with natural phenomena. In the past, many water resources planning, management and operation studies assumed implicitly that the time series (temperature, streamflow, precipitation records) are stationary (Maas et al. 1962). However, the stationarity assumption is no longer valid due to human disturbances in the atmospheric and hydrologic environments (IPCC 2007, 2013, 2014). By now, numerous studies have demonstrated that the stationarity principle is dead, because of substantial impacts due to climate change in the atmospheric events (Milly et al. 2008).

Changes in the means of hydrometeorological time series and in their extreme values may imply trend existence, which must be identified and separated from the main series so as to render it into a stationary state. Since almost three decades environmental, atmospheric, hydrologic, climatologic and agricultural degradations have been searched through trend analysis, especially by employing some classical methodologies such as the Mann–Kendall (MK) analysis (Mann 1945; Kendall 1970). Additionally, trend slope determination by median slope calculation has also been used coupled with the trend detection as suggested by Sen (1968).

Efficient, effective and optimum management of water resources requires the identification of trends not only monotonically over the whole time period, but also whether the "low", "medium", and "high" values have separate trends (Chaps. 3 and 8). These help also to identify drought and flood occurrences in their increasing or decreasing frequencies. In general, a monotonic trend is a gradual change over the whole record period and it is expected to continue in the future. However, for "low", "medium", and "high" values trend searches, the periods are comparatively shorter. As Zhang et al. (2010) rightly suggests, the hydrological literature has so far devoted very limited attention to the characterization of trend pattern. They sought abrupt or gradual trend patterns in a given time series.

1.3.5 Global Warming

Compared to the past, especially in the twentieth and the current centuries, human population increase, extravagant style of life, land use, economic ambitions, wastages, and fossil energy use to support these activities, initially environment and currently atmosphere had to absorb all the remnants in the forms of particulate

matter and greenhouse gasses. Consequently, the chemical composition of the lower atmosphere (troposphere) had started to change such that the extraterrestrial irradiation could not escape back to the atmosphere in the form of short waves, and therefore, accumulation of especially carbon dioxide gas led to warming in the troposphere. The emission releases into the atmosphere cause to global warming as explained in detail by Intergovernmental Panel on Climate Change (IPCC) reports in 2007 and 2013. The most likely changes in physical climate variables or climate forcing agents are identified based on current knowledge, following the IPCC AR5 uncertainty guidance (Mastrandrea et al. 2010). Global warming and climate change terms are used interchangeably for average temperature rise in the Earth's climate system in the form of increasing trend. Global warming causes changes in the climate variables such as the hydrometeorological records, which affect consequently the water resources and the food production that are of utmost significance to human beings. More detailed information is presented in Chaps. 7–8.

1.3.6 Climate Change

Trend analysis evaluation is also needed for long-term infrastructure design and risk analysis in hydro-meteoro-climatic and social origin time series. As stated by Fatichi et al. (2013), due to climate change assessment, trend identification, detection and evaluation are important issues in different disciplines.

Climate does not play role only on present day human activities, but more significantly and scientifically its future predictions are among the most desirable elements so as to mitigate and to provide adaptive decisions, projects, plans, necessary preventive structures and their right as well as adjusted operations in order to reduce expected climate change impacts up to a maximum safety level. Water and food security plans, human health improvements, environmental protections, social and economic affairs are all related to climate change in the long run, but on the present day weather affects in the short-run. All these effects are identifiable through effective trend analysis as presented in this book. IPCC (2007) report expresses the importance of future climate change expectations on different regions, sectors and define the role of anthropogenic atmospheric pollution due to increase of greenhouse gases in an unprecedented rate, which must be offset for prosperous and sustainable future expectancies in various walks of life.

Especially, after 1970 the realization of global warming and the climate change impacts, it has been understood that in a variety of situations some hydroclimatological and economy variables change over time, and this gives rise to a linear or nonlinear trends in the related time series.

1.3.7 Social Sciences

Social and cultural values and practices in a society are changing with time and if asked to most of young people, they may be happy, in general, for the changes, but

the elder generation by remembering old good days may complain about the deterioration of the societal and cultural virtues. The former (latter) generation looks for an increasing (decreasing) trend, but whatever the circumstances, there are steady changes due to economic prosperity, population growth, land use, forestry, global warming, climate change, terrorism, etc. Social trends cannot be for long-term and continuous variations and even one can observe such trend changes over 5-year or shorter duration. Even though the social changes may be initiated by a small group of people, but their effect may spread to cover large portions from the society.

1.3.7.1 Economy

Economic trends are with all of us every day vividly. For instance, shopping implies adding to the consumer spending trend, but to the business gaining trend. The interest rate is also an economic trend and the longer is the time without payment, the more will be the interest amount along an increasing trend. Unemployment rate is another economic trend. In the economy domain, the foreign exchange rate in unstable societies is in an increasing trend direction. If the exchange rate of each month is plotted against time then an economic time series emerge with a trend component.

Measure of any economic development is through the observation, identification and comparison of present levels with the previous cases, and accordingly, either an increasing or decreasing trend shows the degree of the development. In the study of economic trends, the main focus is on the development trends in the recent decades as a result of increasing globalization of knowledge, technology and economy. Industrial processes, information, telecommunication, investments and unprecedented digital data records have led to further researches on a number of trends in economics.

In the economy discipline, a trend can be defined as the overall direction in which a nation's economy changes by time. Economists and especially, financial departments in the governmental or private sectors must be aware of the prevailing direction of the economic trends. Provided that they are able to detect the current economic trends, they are then empowered with more reliable, accurate and effective plans for their establishments.

1.3.7.2 Business

In the strategic business development, it is necessary to observe trends related to the business sector. Careful concern about the business trends helps to improve the market possibilities. Early identification of these trends adds to the future value of the business, because accordingly, the best and successful strategic decision can be taken. With a good background of the past and present economic trends, one is capable to preserve the status of the business affairs and avoid unwanted possible occurrences in the future transactions. Especially, with the availability of computers and fast computation facilities early interpretation of such trends helps to augment the business capacity leading to business growth strategical planning. Changes in the business trends may occur due to the increasing or decreasing product or service

usages, children to stay at home during longer durations, pricing such as the increasing use of online purchasing, changes in the interest rates and in the global factors such as the world economy, housing demand, etc.

Anyone who is working in the business affair should try and identify the most important trends for his/her market and the ones that are not important for the same market can be overlooked. Global watch on the business and economy trends may also help the investor to expect similar trend effects in his/her country or location. The side effect trends on the business must also be watched for a successful strategic business planning. In the business sector, it is not necessary that one should look in finer detail for trend identification and most often conceptual and visual trend assessments and interpretations in linguistic (verbal) statements may be more effective. The growths of business always vary, but mostly accord with an increasing trend.

Future business projections can be achieved through verifiable methods including trend analysis. Trends may also be early warning tools for impending failure outcomes. If accurate and reliable numerical and verbal information are available then trend analysis provides a precise medium for future expectations. Trend analysis is used to forecast market trends, sales growth, inventory levels and interest rates.

1.3.7.3 Health

Trend analyses are also important in any society for health care, because all efforts are toward the improvement of human health. Trends in disease, death and behaviors such as smoking, alcohol drinking are among the public health domains and they show the healthcare directions, assessments in addition to better service planning and policy developments. It is possible to make future predictions and occurrence frequencies and rates based on numerical data examination by time in search for temporal trends.

Trend analysis in health sector indicates the performance of a service usage whether it is beneficial (increasing trend) or not (decreasing trend). In case of benefit, the slope of the trend indicates the rate of change as "quick" or "slow" in linguistic terms. It is also possible to compare one time interval with other as will be explained in the following chapters of this book, to appreciate the effectiveness of the program and whether there is a steady increase. By means of temporal trend analysis in the health domain, it is also possible to compare the situations among geographical locations and populations. Healthcare service improvements may be aided by trend analysis on the basis of estimating possible future likely occurrences, frequencies and rates.

In trend analysis of time series, the first step is to plot available data and then try and observe through examination the change rate. Hence, one is able to grasp overview of the general trend shape, outliers as extreme values, and hence, the researcher gains career experience even though it may be conceptual and visual in fuzzy linguistic (verbal) terminologies (Ross 1995; Şen 2010). In health sector most

often employed trend analyses have linguistic, verbal, probabilistic, statistical and fuzzy diagnoses and interpretations in addition to objective identifications by quantitative methods provided that numerical data are available.

1.4 Pros and Cons of Trend Analysis

With the widespread availability of data virtually in every field and the computer's capability to process them applications for trend analysis seem almost limitless. Since, a trend analysis is based on verifiable data it can be subjected to thorough scrutiny for validation. The use of numbers makes the analysis more exacting. A trend analysis can be replicated, checked, updated and refined when necessary.

Historical data may not give a true picture of an underlying trend. Extreme events like severe floods, droughts and earthquakes distort a normal trend line, while others are more subtle. A major problem in forecasting trends involves turning point identifications. With hindsight, turning points are clearly visible, but it can be difficult to tell in the moment whether they are mere aberrations or the beginning of new trends (Chap. 7). Long-term projections need more data to support them and that may not always be available particularly for a new business or product line. In any case, the further out one forecast the greater is the error possibility, because the passage of time introduces inevitably the effects of new variables.

In the application of any trend methodology prior to the application one should care for underlying assumptions and hypotheses so as to reach at reliable conclusions. Otherwise, even the possible trend component in the data may be weakened.

1.5 Future Research Directions

Time series analysis has a tremendous research and especially application possibilities not only in the natural domains but also in many data mining studies concerning various disciplines as mentioned in Sect. 1.3. In the last two to three decades, the trend analysis has become one of the boiling research, development and application aspects in different disciplines. Although the classical Mann–Kendall trend test is overwhelmingly used in trend identification and assessment in any time series, it has serious drawbacks as for the basic assumptions of sample size, normal (Gaussian) probability distribution functions (pdf) and serial correlation structure. This test is considered as the classical approach in the literature. However, there is still room and need for more efficient and powerful trend detection tests in future time series analyses, which are within the bulk goal of this book. Furthermore, research should be directed toward the consideration of contemporary multiple comparison tests in addition to the commonly used tests for checking homogeneity in the natural and artificial time series records. In many studies, a time series is considered as stationary random variable based on trend

tests only. Although there are less common methods such as t-test and nonparametric Mann–Whitney test, they have not been employed in full scale (Chap. 3).

Periodicity and dependent structure (persistence) features are ignored in many time series analysis studies. Each time series component (trend, periodicity, jump and randomness) is very important in many planning, operation, management and maintenance projects. Parallel to ever growing global warming and climate change events, the trend search in recorded time series occupies the top priority in the time series analysis. As a result, trend analysis gained acceleration over the last 30 years. It is expected that the well-known "greenhouse effect" will alter the timing and magnitude of many hydrological, environmental, social and health events, leading to the possibility of environmental and socioeconomic dislocations that can be pressed partially by trend component. For instance, trends have important implications for the planning and management of water resources in the future (Gleick 1989; IPCC 2007). Additionally, variability features are also searched through the trend analysis as will be explained later in this book.

Hirsh et al. (1982) stated that for a "next generation" of trend analysis techniques in response to the observations need recent and longer monitoring data sets, new questions about the effectiveness of control efforts and the availability of new statistical tools. They identified seven critical attributes for the next generation of trend analysis. It has been stated that the current trend analyses should,

(1) focus on revealing the nature and magnitude of change rather than strict hypothesis testing,
(2) not assume that the flow-concentration relationship is constant over time,
(3) make no assumptions that seasonal patterns repeat exactly over the period of record, but allow the shape of seasonality to evolve over time,
(4) allow the shape of an estimated trend to be driven by the data and not constrained to follow a specific form such as linear or quadratic; trend patterns should be allowed to differ for different seasons or flow conditions,
(5) provide consistent results describing trends in both concentration and load,
(6) provide not only estimates of trends in concentration and flux, but also trend estimates where the variation in, say, water quality due to variation in streamflow has been statistically removed,
(7) include diagnostic tools to assist in understanding the nature of the changes that have taken place over time, e.g., to identify particular time periods of year or conditions during which quality changes are most pronounced.

1.6 Purpose of This Book

During the last 50 years' time series analyses methodologies have been applied in a variety of fields including hydrology, meteorology, climatology, geology, oceanography, seismology, oceanography, economics, health, space research, earth,

marine and agricultural sciences, etc. This book presents not only a review of trend analysis applications in different domains through a set of classical methodologies, but also provides innovative methodologies for the most effective ways of trend identification, determination, assessment and interpretation. The comprehensive review indicates the convenience of the available trend analysis methodologies and their adaptations based on rational, logical and a set of scientific assumptions for each approach. On the other hand, many social, health, environmental and engineering aspects attract worldwide attention for time series analysis techniques application. Although there are numerous applications of the well-known time series and trend analyses in different domains, unfortunately less number of studies is developed on innovative approaches/methodologies or even modification of existing approaches for trend analysis. Hence, the main goal of this book, after the introduction of the classical time series and trend analyses methodologies, is to present up to date modernly developed innovative trend analyses of different types, which provide simple, effective, rational and logical linguistically interpretations and quantitative theoretical developments with almost no assumption. In the past, most researchers have employed the applications of the classical trend analyses without significant improvements except few modifications that could not avoid the drawbacks and assumptions. In the applications, such trend analyses have been applied so frequently and consequently that other time series features (stationarity, homogeneity, periodicity and persistence) have been overlooked by depending on unchecked assumptions. It is hoped that in the future studies in addition to trend methodological developments, the basic features and assumptions are also taken into consideration. It is recommended in this book that in the future more effective innovative methodologies should be discovered or at least the existing ones are modified such that the applications comply by the basic theoretical assumptions.

Future researches are expected to deal with the applications of time series analysis techniques in different disciplines toward more robust and widely acceptable interpretation and implementation possibilities. The most important and preliminary requirement is to have reliable measurement records for providing better and useful methodologies.

References

Davis, J. C. (1986). *Statistics and data analysis in geology*. New York: Wiley.

Esterby, S. R. (1996). Review of methods for the detection and estimation of trends with emphasis on water quality applications. *Hydrological Processes, 10*(2), 127–149.

Fatichi, S., Ivanov, V. Y., Caporali, E. (2013). Assessment of a stochastic downscaling methodology in generating an ensemble of hourly future climate time series. *Climate Dynamics, 40*, 1841–1861.

Hess, A., Iyer, H., & Malm, W. (2001). Linear trend analysis: A comparison of methods. *Atmospheric Environment, 35*(30), 5211–5222.

Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of trend analysis for monthly water-quality data. *Water Resources Research, 18,* 107–121.

IPCC. (2007). Climate change 2007: Impacts, adaptation, and vulnerability. Contribution of Working Group II to the fourth assessment report of the intergovernmental panel on climate change. Cambridge, UK: Cambridge University Press.

IPCC. (2013). Climate change 2013: The physical science basis. Contribution of Working Group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge, UK: Cambridge University Press.

IPCC. (2014). Climate change 2014: Impacts, adaptation, and vulnerability. Contribution of Working Group II to the fifth assessment report of the intergovernmental panel on climate change. Cambridge, UK: Cambridge University Press.

Kendall, M. G. (1970). *Rank correlation methods* (4th ed.). London: Griffin.

Maas, A., Hufschmidt, M., Dorfman, R., Thomas, H. A., Jr., Marglin, S., & Fair, G. M. (1962). *Design of water resource systems*. Cambridge, MA: Harvard University Press.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13*(3), 245–259.

Mastrandrea, M. D., Heller, N. E., Root, T. L., & Schneider, S. H. (2010). Bridging the gap: Linking climate-impact research with adaptation planning and management. *Climate Change, 100,* 87–101.

Milly, P. C. D., Julio, B., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, P., et al. (2008). Stationarity is dead: Whither water management? *Science, 319,* 573–574.

Ross, J. T. (1995). Fuzzy logic with engineering applications. McGraw-Hill, Inc., 600 p.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association, 63*(324), 1379–1389.

Şen, Z. (2010). *Fuzzy logic and hydrological modeling* (p. 340). New York: Taylor and Francis Group, CRC Press.

Zhang, X., Harvey, K. D., Hogg, W. D., & Yuzyk, T. R. (2010). Trends in Canadian streamflow. *Water Resources Research, 37,* 987–998.

Uncertainty and Time Series

2

Abstract

Trends are one of the deterministic parts of a given time series apart from the natural or artificial seasonality and uncertain components. Trend analysis is a search for deterministic trend in an uncertain environment, therefore, the basic concepts of uncertainty are explained as stochastic and completely random variables and their importance in trend identification studies. Since, probability and statistics are main subjects for such a search various probabilistic and statistical concepts are presented in an effective manner so that prior to a proper trend analysis the reader can appreciate the fundamental elementary concepts, which are in later chapters are employed for the main goal. In classical trend analyses, the most restrictive assumption requirement is the serial independence of given time series, various correlation measurement suggestions are reflected from the literature. In the meantime for classical trend analysis, the characteristics of a time series are explained for proper application of the methodologies.

Keywords

Correlation · Frequency · Histogram · Homogeneity · Seasonality · Stationarity · Uncertainty

2.1 General

Uncertainty has many connotations to common people and experts grasp it in rather different ways; some considers it as entirely unknown and unpredictable information and to some others, it is partial information and knowledge. The uncertainty is everywhere and one cannot get rid of it completely. Initial knowledge and information are concepts that depend on personal observations and experience.

Z. Şen, *Innovative Trend Methodologies in Science and Engineering*,
DOI 10.1007/978-3-319-52338-5_2

Uncertainty can be avoided by a set of simplifying assumptions about the phenomenon concerned. For instance, Newtonian classical physics is entirely deterministic. Today almost all branches of science (environmental, atmospheric, earth, engineering, economics, health and social) are confronted with uncertainty ingredients and many scientific deterministic foundations take uncertainty forms in terms of random, probability, statistics, chaos, fractal, stochastic, quantum, and fuzzy implications. In many scientific and technological institutions determinism dominates the education systems. Famous philosophers and scientists spell out the uncertainty and fuzzy ingredients that are essential bases of scientific progress. For instance, Russell (1948) stated that

> All traditional logic habitually assumes that precise symbols are being employed. It is, therefore, not applicable to this terrestrial life but only to an imagined celestial existence.

On the other hand, as for the verbal and linguistic fuzzy conceptions Zadeh (1965) said that

> As the complexity of a system increases, our ability to make precise and yet significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance (or relevance) become almost mutually exclusive characteristics.

During human thinking evolution, the premises include uncertainty elements such as vagueness, ambiguousness, possibility, probability, random, and fuzziness. Implication of mathematical structure from the mental thinking process might seem exact, but even today it is understood as a result of scientific development that at every stages of modeling, physical, or mechanical, there are uncertainty pieces, if not in the macroscale, at least at the microscale. It is clear today that mathematical conceptualization and idealization leading to satisfactory mathematical structure of any physical actuality is often an approximation, because as Popper (1954) states that scientific facts are falsifiable.

The word uncertainty reminds also the probability of occurrences with attachment of a certain percentage. It is not uncommon that everyone is confronted with probability statements, especially in natural, social, and financial conservations. For instance, what is the probability of weather status for tomorrow? what will be the income depending on the number of clients in the next month? what is the probability of investing on stock? These queries involve uncertainty and their daily answers are quantitatively through subjective percentage numbers, which are, the probability statements. Probability in the statistics context helps to investigate past records of uncertain events so as to make future predictions and dependable decisions.

Prior to the explanations of systematic components such as trends in any time series, it is preferable to equip the reader with uncertainty concepts. Uncertainty or randomness and stochasticity are counter concepts to determinism. For instance, astronomic events were thought as rather uncertain phenomena by early human beings, but today they are well known and even one is capable to calculate through the scientific methodologies the position of any planet at any future time.

Additionally, any gadget, instrument, automation or machine can be described by deterministic formulations, equations, and logical rules.

Many natural events in atmospheric, environmental, earth, oceanography, meteorology, hydrogeology, earthquake and social, economic, health, business, and similar events do not provide completely well-established behaviors, because they include some deterministic parts that can be identified by mathematical statistical methods, but the residuals from the deterministic components are random in character. Any natural phenomenon takes place under the combined effects of physical, chemical, mechanical, and thermodynamic conditions and evolves temporally and spatially according to a set of certain laws. These effects cannot be identified accurately whatever the monitoring instrumentation and scientific methodologies are. For instance, hydrological events such as rainfall, runoff, infiltration, and evaporation cannot be controlled precisely over large areas and future times. The uncertainties in these events affect the economic, environmental, social, and even physiological conditions of human societies. For instance, human civilization has long been deeply affected by impacts of droughts on economic, environmental and social sectors (Wilhite 1993).

On the other hand, the scientific models that are suitable for description of an event might not be reliable with high confidence. At this stage, it is very convenient to remember the statement by Einstein that

> so far as the laws of mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality.

Natural event uncertainty is associated with not knowing if and/or when, say for instance, a rainfall event will cause to the exceedence of a given design discharge. Additionally, model uncertainty is the inaccuracy of the model used to estimate the design discharge. In addition to these two uncertainty types the third one is the measurement error.

The uncertainty in the earth and atmospheric systems arises from the conviction that generalizations are immensely complicated instantiations of abstract and often universal physical laws. Such generalizations always contain assumptions of boundary and initial conditions. The researchers cannot control these conditions with certainty.

Earth systems sciences deal with spatial and temporal structures (trends, periodicities, jumps) in natural phenomena at every scale for the purpose of predicting the future replicas of the similar phenomenon, which help to make significant decisions in planning, management, operation, and maintenance of natural events that are related to social, environmental, and engineering activities. These phenomena are sampled by measurements with uncertainty ingredient; their analysis, control, and prediction need to use uncertainty techniques for reliable predictions. Natural phenomena cannot be monitored at a set of desired instances and locations, and therefore, such restrictive time and location conditions bring additional irregularity into the measurements. For instance, floods, earthquakes, car accidents, illnesses, and rock fracture occurrences are among the irregularly distributed temporal and spatial events. Uncertainty and irregularity are the common properties of

natural phenomena measurements in many researches, but the analytical solutions through numerical approximations require mostly regularly available initial and boundary conditions that cannot be obtained by lying regular measurement sites or time intervals. In an uncertain environment any cause is associated with different effects each with different level of possibility. Herein, possibility means some preference index for the occurrence of each effect. The greater the possibility index, the more frequent the event's occurrence.

2.2 Random and Randomness

Random and randomness are the two terms that are used in statistical sense to describe any phenomenon, which is unpredictable with any degree of certainty. An illuminating definition of randomness is provided by famous statistician Parzen (1960) as,

> A random (or chance) phenomenon is an empirical phenomenon characterized by the property that its observation under a given set of circumstances does not always lead to the same observed outcome (so that there is no deterministic regularity) but rather to different outcomes in such a way that there is a statistical regularity.

The statistical regularity implies group and subgroup behaviors of a large number of observations so that the predictions can be made for each group more accurately than individual ones. For instance, provided that a long sequence of temperature observations are available at a location, it is then possible to say quite confidently that the weather will be warm, cool, cold, or hot tomorrow than specifying exactly by degree of centigrade prediction. The statistical regularities are as a result of some astronomical, natural, environmental, and social effects.

Deterministic phenomena are those in which outcomes of individual events are predictable with complete certainty under a given set of circumstances, provided that the initial and boundary conditions are known. It is necessary to check the validity of the assumption sets and initial conditions. In a way, with idealization concepts, assumptions, and simplifications deterministic scientific researches yield conclusions in the forms of algorithms, procedures, or mathematical formulations, which should be used with caution. The very essence of determinism is the idealization and assumptions so that uncertain phenomenon becomes graspable and conceivable to work with the available physical concepts and mathematical procedures. In a way, idealization and assumption sets render random phenomenon into conceptually certain case by trashing out the uncertainty components. A significant question that may be asked at this point is that, is there not any benefit from the deterministic approaches in natural studies, where the events are uncertain? The answer to this question is affirmative, because in spite of the simplifying assumptions and idealizations, the skeleton of the uncertain phenomenon can be captured by deterministic methods. For instance, determination of a trend component in a time series is a good example.

Even after the separation of a time series from its systematic components such as trend, the residuals should be checked for various feature properties so as to be able to apply probabilistic, statistical, and stochastic methodologies, which have a set of assumptions such as stationarity or weakly stationarity, homogeneity, independence or dependence, persistence or fuzziness.

2.3 Empirical Frequency and Distribution Function

An empirical work is possible provided that there is a set of measurements about the event concerned. In the content of this book measurements are considered as a sequence of records at equal time intervals, which are referred to as the time series. In mathematical notation Y_1, Y_2, …, Y_n is a time series with n samples. It can be shown succinctly as Y_i (i = 1, 2, …, n). Such a time series is shown for annual precipitation records in Fig. 2.1

Visual inspection of this figure indicates that obviously there is not a systematic follow-up between the successive years, and therefore, it is a random time series. This visualization is obtained by looking at the figure directly. It is possible to search visually for any trend, seasonality, and shift components. However, in this figure, it is not possible to identify any one of these deterministic components.

It is possible to obtain the number of frequencies provided that the variation domain is divided into equal length classes, which is shown in Fig. 2.2 with five classes. The frequency means the number of data values that fall within the class considered. For instance, in Fig. 2.3 the numbers of frequencies from left to right classes are 3, 30, 43, 27, and 13, respectively. The summation of these frequencies is equal to 116, which is the number of data in Fig. 2.1.

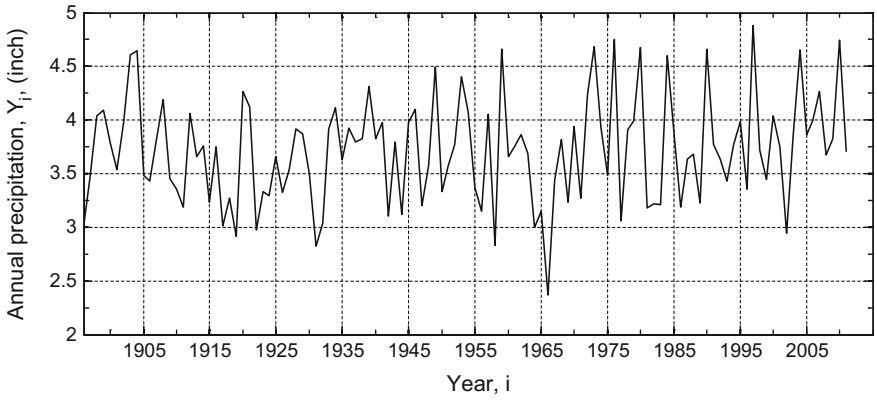


Fig. 2.1 Annual precipitation time series

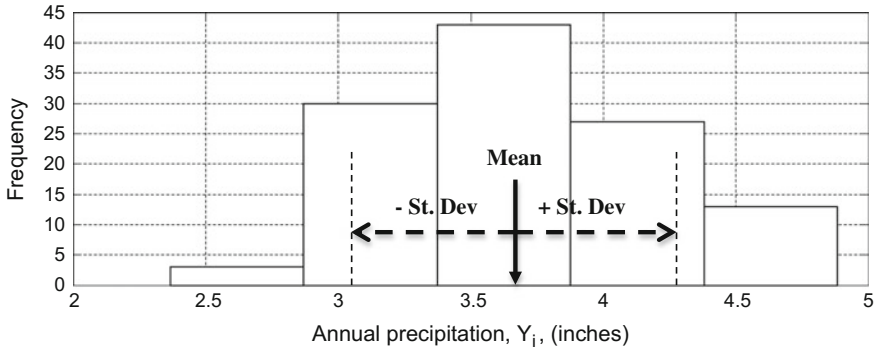


Fig. 2.2 Empirical frequency distributions

In order to appreciate the frequency concept, Fig. 2.2 is shown on the left-hand side in Fig. 2.3 vertically, and hence, one can understand, which data values fall into which class.

The frequency diagram in Fig. 2.2 seems almost symmetrical, which means that the number of data values more (less) than the arithmetic average is almost 50%. Such a symmetrical frequency distribution function has the arithmetic average (mean) value at or very close to the symmetry axis as shown in the same figure. As a statistical rule, in symmetrical distributions, the mean value is almost equal to the mode (the most frequently occurring value) and median (the value that divides the frequency distribution function into two halves) values.

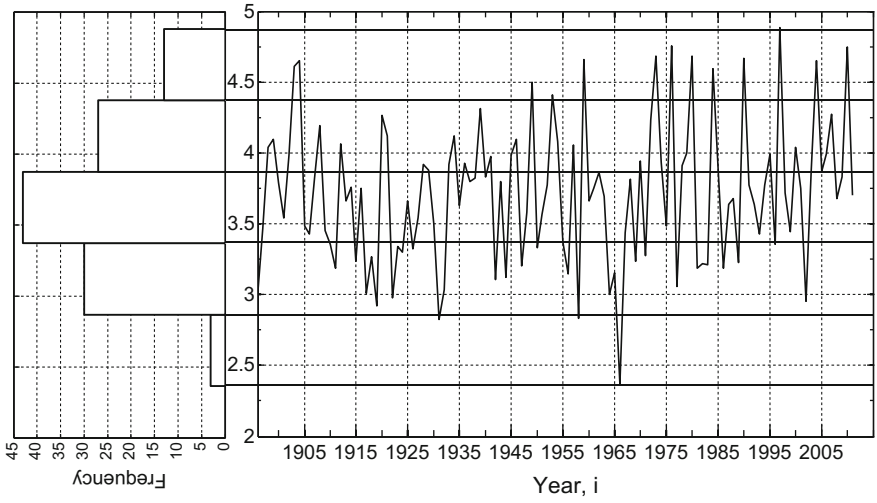


Fig. 2.3 Time-series and frequency combination

After all what have been explained above, it is possible to write down an equation among the class frequencies in a time series. If there are m classes each with frequencies f_i (i = 1, 2, …, m) then,

$$f_1 + f_2 + \cdots + f_m = n, \tag{2.1}$$

where n is the number of data in the given time series.

It is also possible to appreciate the standard deviation value, which indicates arithmetic average deviations around the mean value. With this concept in mind, positive and negative standard deviation values are shown on the right and left of the arithmetic average value in Fig. 2.2. Apart from the symmetric frequency distribution function, various skewed (nonsymmetric) types are shown in Fig. 2.4.

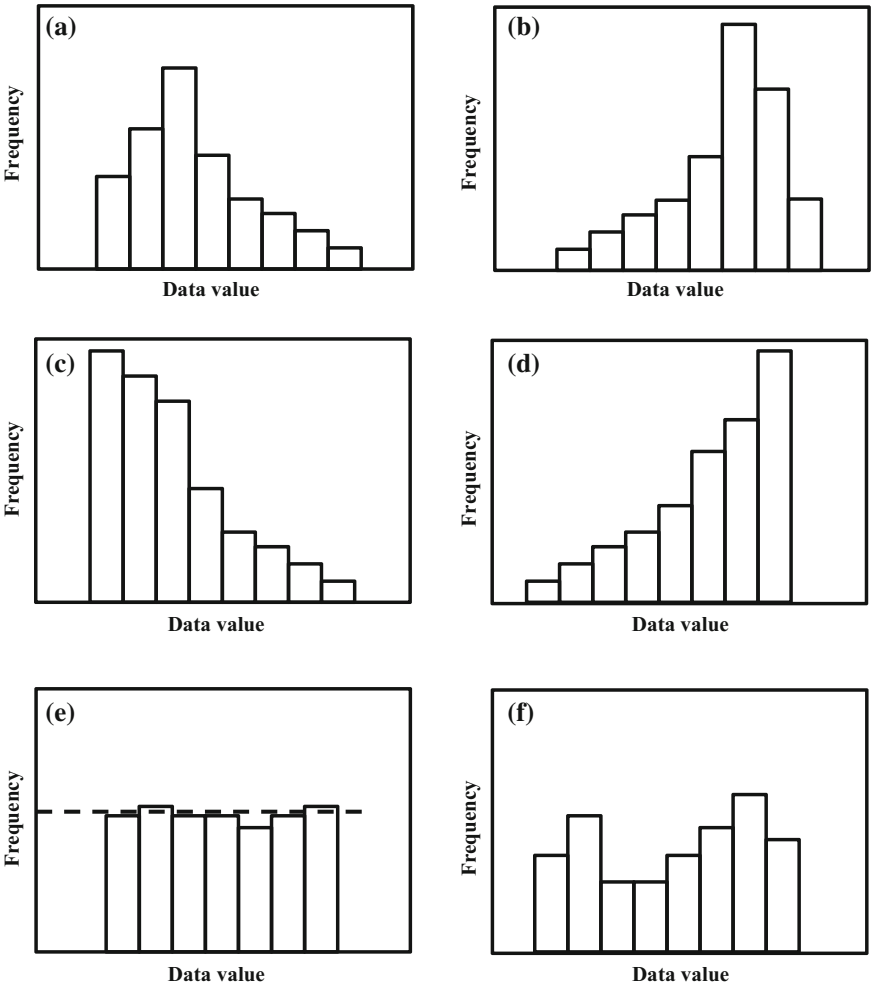


Fig. 2.4 **a** Negatively skewed, **b** positively skewed, **c** exponential, **d** J-shaped, **e** uniform, **f** bimodal empirical distributions

In practical studies, any one of these empirical frequency distribution functions emerges from a given time series data. It is possible to make various visual interpretations from each empirical frequency distribution function and this point is left to the reader so that s/he can increase personal expert views.

A very significant question at this stage; is it possible to identify any systematic component from the empirical frequency distribution functions? The answer is that it is not possible to deduce any trend component from the empirical frequency distributions.

2.3.1 Empirical Frequency and Trend

The unique way to be able to identify trend component through the employment of empirical frequency analysis is possible if the given time series is divided into two halves and for each half the empirical frequency distribution functions are obtained and compared with each other. For visual inspection, Fig. 2.5 presents a time series of 100 data values, where one can see visually that there is an increasing trend.

In order to see objectively whether there is a trend in the given time series or not, the time series is divided into two halves 50 and 50 data values and the resulting two empirical frequency distributions are given in Fig. 2.6.

Comparison of the two-half empirical frequency distributions indicates that there is significant shift toward the high values as obvious from Fig. 2.6b.

Figure 2.7 indicates decreasing trend component in the time series and it is identifiable even by naked eye visually without any quantitative methodology applications.

This time series is also divided into two halves and their empirical frequency distribution functions' comparison provides information about the decreasing trend as the frequency distributions in Fig. 2.8.

Comparison of the second half empirical frequency distribution with the first one indicates that there is a decreasing trend.

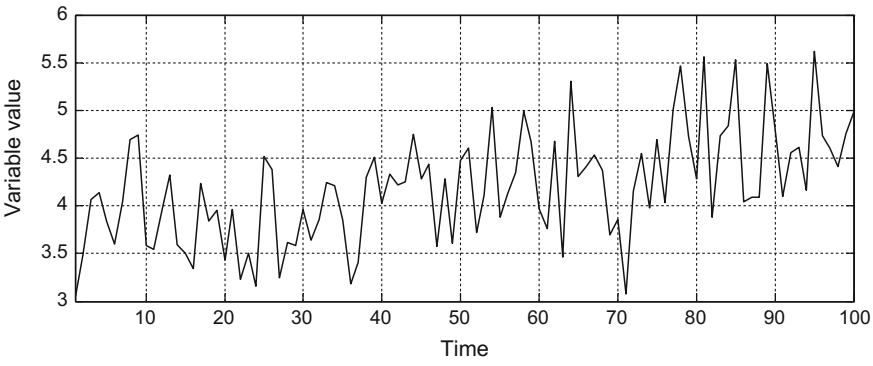


Fig. 2.5 Time series with visual increasing trend

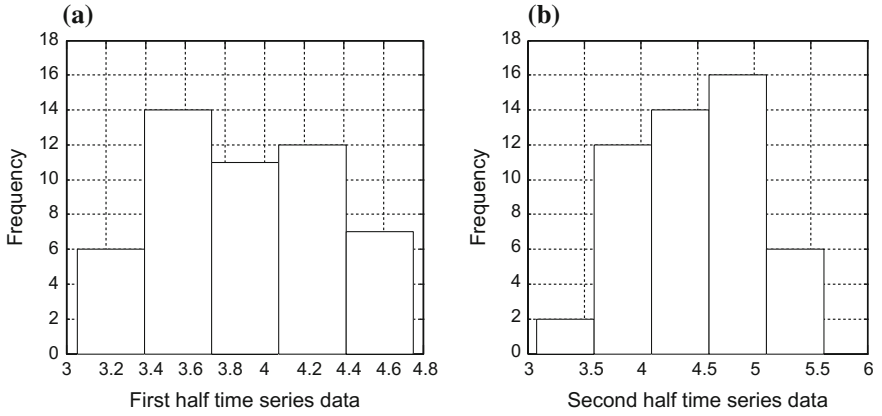


Fig. 2.6 Two halves empirical frequency distributions

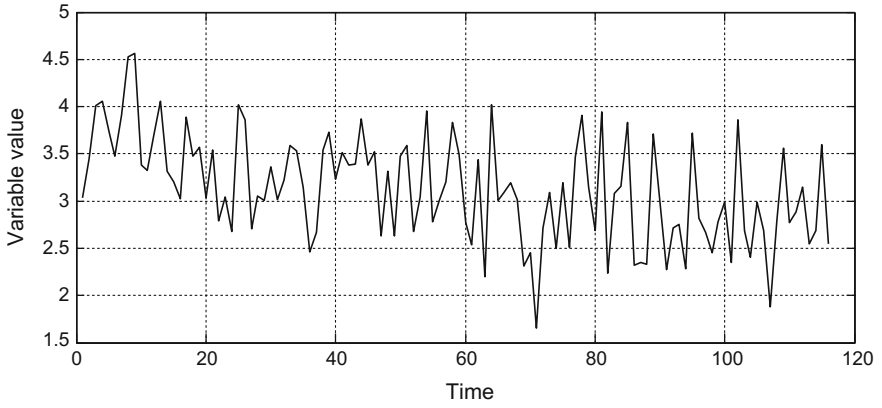


Fig. 2.7 Time series with visual decreasing trend

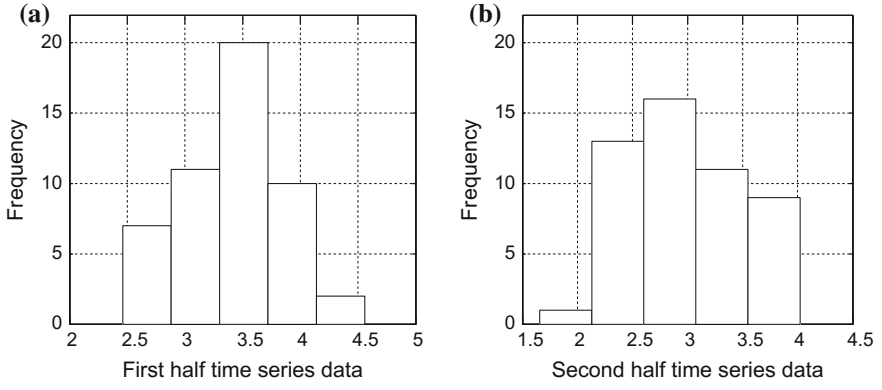


Fig. 2.8 Two halves empirical frequency distributions

Objective decision about the trend component existence within a time series by two-half empirical frequency distributions is possible through the chi square test.

Şen (2012, 2014) has proposed an innovative trend analysis methodology by using the concept of dividing a given time series into two equal halves. However, it is also possible to divide the time series into more than two equal size pieces and then in each piece, one can search for possible trend through the same innovative methodology. Various forms of the innovative trend methodology are explained in different chapters of this book (Chaps. 5–8).

2.4 Theoretical Probability Distribution Function (Pdf)

The probability has been expressed in daily life as percentages, but in the statistical context it may assume any value between 0 and 1, inclusive. When its value is equal to zero (one) then the event is absolutely impossible (possible). If the question is what is the probability in a given class then one can divide both sides of Eq. (2.1) by n, and hence, the probability of ith class is defined objectively as f_i/n. Finally, Eq. (2.1) can be written as,

$$\frac{f_1}{n} + \frac{f_2}{n} + \cdots + \frac{f_n}{n} = 1 \tag{2.2}$$

or with probability, p_i (i = 1, 2, …, m), notations,

$$p_1 + p_2 + \cdots + p_m = 1 \tag{2.3}$$

After these definitions, it becomes clear that in order to obtain the pdf one needs to divide each class frequency diagram in the empirical frequency distribution by the number of data and the resulting graph is referred to as histogram. Since by definition the area under any theoretical pdf is equal to 1, the histograms must be prepared preferably in such a way that the empirical area under it must also be equal to 1.

As mentioned in the previous subsection for the empirical frequency distribution, it is possible to search for possible trend component by comparison of two-half pdfs of a given time series. This is used descriptively for global warming discussions as in Fig. 2.9, where pdfs are shown theoretically as continuous curves.

In this figure, the pdf A can be regarded as the first half of a time series and B and C are the second halves. Comparison of the first half pdf (A) with the second (B) indicates that there is a shift toward the high values. If this shift is not sudden but gradual then it evolves with time along a smooth trend component. This last statement implies that there is an increasing trend in the given time series. If the amount of increase is asked then one can say that the arithmetic average (the peak) value, μ, of the first half has shifted toward the right by amount of $\Delta\mu$, and since this amount took place during the half time, $n/2$, evolution of a given time series of

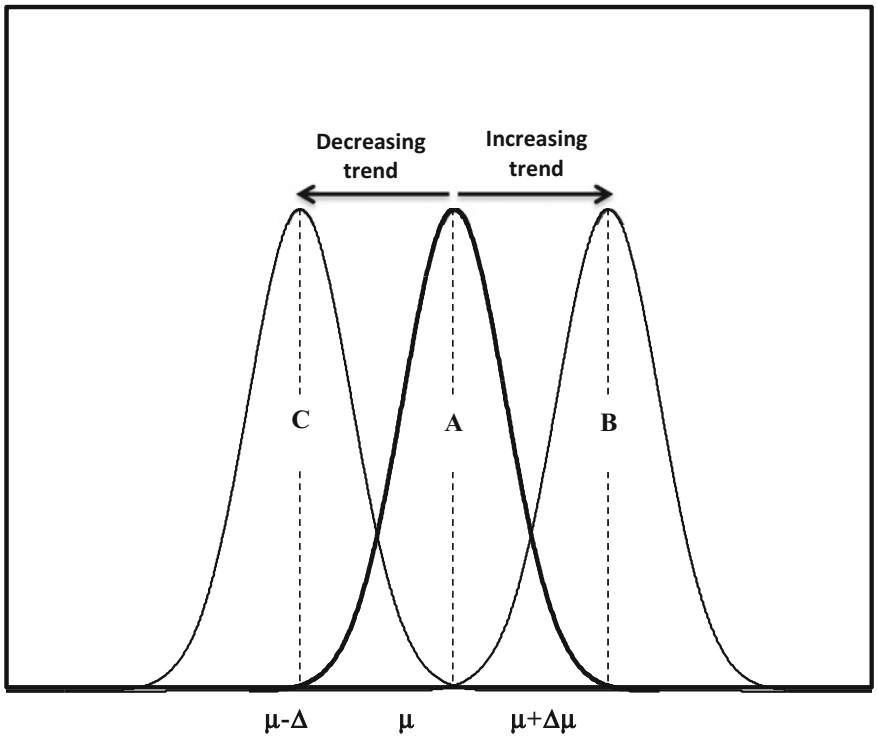


Fig. 2.9 Probability distribution function trend implications

duration (sample number), n, then the slope, S, of the trend component can be calculated as,

$$S = \frac{(\mu + \Delta\mu) - (\mu)}{(n/2)} = + \frac{\Delta\mu}{(n/2)} \tag{2.4}$$

This statement says that in order to find the slope of the trend, take the difference between the arithmetic averages of the two halves and divide it by the half number of data. This statement is one of the fundamental points about the innovative trend analysis in Chap. 5. In cases of nonsymmetrical pdfs, instead of arithmetic average, if possible, mode or preferably median value can be adapted.

On the other hand, if pdf C is considered as the second half, then there is a decreasing trend component in the time series and the slope should be calculated similar to Eq. (2.4) as follows.

$$S = \frac{\mu - (\mu + \Delta\mu)}{(n/2)} = - \frac{\Delta\mu}{(n/2)} \tag{2.5}$$

These interpretations indicate that visual inspections can lead to quantitative trend slope calculations in the simplest manner. These expressions will be used in the subsequent chapters, and especially during the explanation of the innovative trend analyses methodologies (Chap. 5).

2.5 Statistical Modeling

Complex interactions among the natural event characteristics give rise to spatial and temporal evolution of the phenomenon concerned, which must be controlled in a scientific manner so as to render its consequences to beneficial forms for human activities. For instance, prior to computer age the runoff event analysis dates back to the original work of Ripple (1881), who presented a deterministic graphical method for determination of the necessary reservoir capacity from an available sequence of recorded runoffs. This capacity is regarded in its simplest form as a prediction for future runoff regulations. However, such an approach has several drawbacks as follows.

(1) The historical sequences will not reappear in the same order in future,
(2) The statistical correlation structure will not have the same pattern,
(3) The location of the extreme values along the time axis will not be in the same order as in the historic records.

The use of computers in natural event modeling led researchers to an explosion of simulation models for prediction purposes. Subsequently, a host of physical, conceptual, or black box type models are developed continuously and introduced into the literature. However, initially most of these models aimed at preserving some low order statistics, but later more specific real-time prediction processes are presented with rather simple recurrence model types, which extract necessary information from the available historical data, and later, their future predictions are achieved. Among the most important statistical parameters are the mean, standard deviation, coefficient of skewness, and the autocorrelation coefficients.

Especially, in long periods of time, the Hurst coefficient is also suggested for modeling purposes. Mandelbrot and Wallis (1969a, b, c) works led them to set horizons of the fractal geometry, which plays significant role in the investigation of chaotic behaviors of dynamic systems. In order to construct a dynamic model for the simulation of any natural phenomena, it is necessary to have a finite record of past observations. Given a historical record, the estimation process consists of computing an estimate of the variable concerned at time lead k, the position of which relative to observation period leads to three types of estimations problems.

(1) The estimation of state at any time instant during observation period is referred to in statistics as “smoothing” operation or in mathematics as “interpolation”,

(2) Estimation of the state at the final observation time instant is called as "filtering",
(3) State variable estimation at a time instant after the final observation, which is referred to in uncertainty domain as "prediction" or in certainty, i.e., in mathematics domain as "extrapolation".

In addition to these stages, after the model adaptation and determination of its parameters there is "verification" stage where the suitability of chosen model to the historical observation sequence is sought. This stage includes search for suitable model theoretically to make parameter estimates for the model and to check the suitability of the model. However, in any study, the most important stages are the identification, verification, and subsequent prediction phases, and they follow each other.

The basic estimation work has been performed by Gauss in early 1800s who tried to fit the most suitable curve through the scatter of points by having the least squares technique as a criterion, which constitutes without exception the basis of any uncertainty event assessment in statistics and stochastic process modeling studies. The successful application of the least squares technique for almost two centuries is due to the following factors.

(1) The minimization of sum of squared errors leads to a system of linear equations, which are easy to solve and do not require an extensive theory. This approach is used frequently in trend identification calculations,
(2) The sum of the squares corresponds in many different contexts to various interpretations such as in physics, the energy is expressed as the sum of squares; in mechanics it represents moment of inertia, in statistics it provides the variance about the fitted curve, and consequently, it can be used as a measure of the goodness-of-fit test,
(3) An assumption of a definite explicit analytical form to represent that the observed data constitutes the principal application of the classical least squares technique,
(4) Without proposing an explicit analytical expression, it is possible to apply the least squares technique to filtering problems. For instance, a known differential equation may represent the phenomenon concerned. Likewise, the storage and continuity equations are explicit expressions for certain phenomenon in physics,
(5) Wiener (1949) has founded a different application version of the least squares technique by assuming certain statistical properties for the useful signal and noise constituents of observation sequences. The significant difference of Wiener's approach lies in the fact that the useful and noise parts are characterized not by analytical forms, but by their statistical properties, such as the mean values are supposed to be zero or rendered to zero and also serial and cross-autocorrelations,

(6) After 1960 in order to reduce the computation burden, Kalman suggested an elegant procedure for the adaptive prediction in the form of recursive filtering. This technique is generally considered as igniting the widespread interest in the subject of estimation.

2.5.1 Deterministic-Uncertain Model

In various modeling studies in different disciplines, there are input and output random structured variables. In any system design, the input and output variable measurements show randomness in the sense that any data value cannot be predicted from the previous data values with certainty, therefore, they must be treated by probabilistic, statistical, and stochastic models. On the basis a convenient prediction model, future replica of the output variables can be obtained within certain limits of errors such as ±5 or 10%. This was the main problem in front of system planners and managers and the question is which value to adopt in the model procedural design? In the beginning of the twentieth century, Hazen (1914) has suggested the use of the following procedure without the availability of any computer and even calculator. His method was random drawing of paper pieces that are mixed in a sack. Each paper piece had a certain number written on it and then folded and put into the sack. If the past observations of any phenomenon are denoted as a sequence, $Y_1, Y_2, \ldots, Y_n$, it is possible to calculate its various statistical parameters (arithmetic average, standard deviation, serial correlation coefficient, etc.). This sequence is the naturally ordered record of measurements, and the statistical parameters are dependent on the whole data values and they are valid for the record duration only. These historical data series can be used for the simple prediction of the future values according to the following steps.

(1) Historical data record: There are daily, monthly, or annual records of past variable measurements. Let the number of records be n,
(2) Each one of these measurements is written separately on equal size paper pieces. They are folded and then put into a sack,
(3) The pieces are drawn one after the other, and hence, a new time series is constructed of the same duration or even longer with the same data values but at different times,
(4) After each drawing, the paper pieces are either returned to the sack or not. In the former case, it is possible to generate sequences as long as desired. However, in the latter, the maximum length of the synthetic data can be equal to the length of the original data. After the generation is complete, whole pieces can be retuned into the sack and again another random sequence (replica) can be generated. In this manner one is able to obtain an ensemble of synthetic sequences (time series).

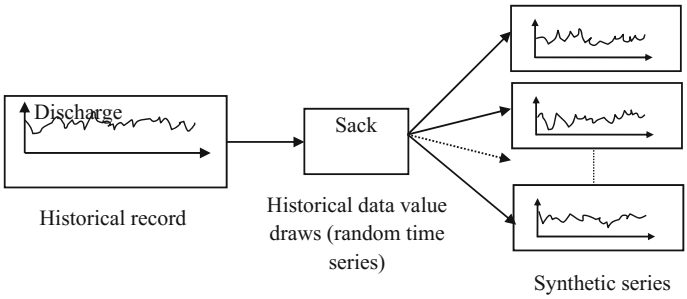


Fig. 2.10 Deterministic-uncertain method sequences

After the completion of such a generation procedure, the input sequences are treated for the assessment of decision or determination of design quantity. For instance, if the sequences are supply levels of some quantity then by knowing the demand level, it is possible to decide about the supply sufficiency time durations. In the case of insufficiency, additional supply is withdrawn from available storages.

The sequences obtained by this aforementioned deterministic-uncertain method can be referred to as synthetic sequences (replicas), which have the following points in common.

(1) Each synthetic sequence has the same arithmetic average as the original sequence,
(2) Each synthetic sequence has the same variance and the standard deviation as the original sequence,
(3) Other statistical parameters (mode, median, skewness, kurtosis, etc.) are also the same in addition to the relative frequencies, hence also the relative frequency distribution,
(4) The major assumption in such a draw system is that each one of the generated sequence is regarded as independent from others, but this is not valid practically. Each one of the generated sequence has its own serial correlation coefficient that may be significantly different from each other. The general procedural function of this deterministic-uncertain methodology is shown in Fig. 2.10.

2.5.2 Probabilistic-Statistical Model

This procedure is more developed than the previous one and instead of using the same data values in synthetic sequence generation, the relative frequency distribution of the measured data is adopted as the root of the generation procedure and it is fitted with the most convenient theoretical pdf through the chi square test. This time the data are not drawn from the sack with the repetition of the historical data,

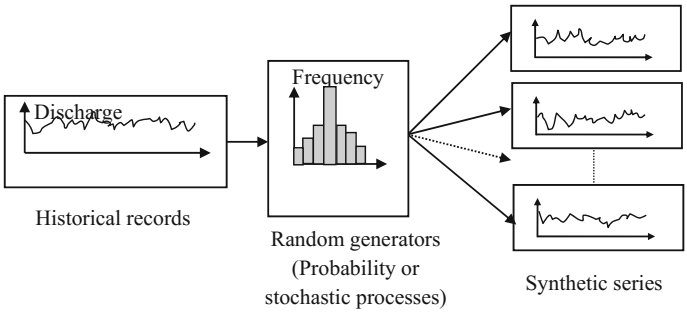


Fig. 2.11 Statistical procedure stages

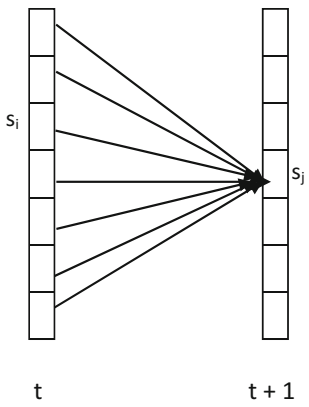
but the synthetic sequence values are drawn from the theoretical pdf automatically in computers from random number generators. This approach yields synthetic sequences that have in the long run almost the same statistical parameters, but it also provides extreme value contribution into the generation procedure, which is never possible with the deterministic-uncertainty method. Figure 2.11 shows the basic stages in the statistical procedure.

In this generation process, although the original time series statistical structure is preserved, but the serial correlation coefficient is not taken into consideration.

2.5.3 Transitional Probability Model

Although the probabilistic-statistical methods are based on the statistical parameter preservation by a theoretical pdf adaptation as explained above, they depend on the class interval relative frequencies obtained from a given measurement series. In the transitional probability approach, the sequence of class intervals are considered to remain the same at each time instant as shown in Fig. 2.1, but transition probabilities or transition frequencies are considered from one class interval at t instant to the next one at $t + 1$ instant as in Fig. 2.12.

Fig. 2.12 State and transition probabilities



If there are m class intervals, there will be m class interval relative frequencies, which are referred to herein as the state probabilities. Furthermore, there are $m \times m$ interclass interval transition probabilities, which are relative joint frequencies. Hence, instead of the statistical parameters, the state and transition probabilities are used in the modeling of a given time series. These are known in the literature as the Markov chain models (Feller 1968; Box and Jenkins 1970). Their application requires the following steps:

(1) Construction of the histogram from a given time series,
(2) Calculation of the class interval frequencies (state probabilities) from the histogram,
(3) Calculation of transition relative frequencies (transition probabilities) between two successive time instances.

The transition probabilities can be considered in the form of a matrix, where rows are for time instant t, and columns for $t - 1$. Such a matrix is called as the transition probability matrix, P_T.

$$P_T = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{1n} \\ p_{21} & p_{22} & p_{23} & p_{24} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{2n} \\ p_{31} & p_{32} & p_{33} & p_{34} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{3n} \\ p_{41} & p_{42} & p_{33} & p_{44} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{4n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ p_{n1} & p_{n2} & p_{n3} & p_{n4} & \cdot & \cdot & \cdot & \cdot & \cdot & p_{nn} \end{bmatrix} \tag{2.6}$$

Since the transition from state, say, s_i at time t, to state, s_j at time $t - s$ is the same as the transition from state s_j at time $t - 1$ to s_i at time t, the transition matrix will have a diagonal symmetric form, i.e., $p_{ij} = p_{ji}$. Hence, $m(m - 1)/2$ transition probabilities are necessary for the definition of the transition probability matrix. The transition probabilities along the major diagonal are all equal to 1 ($p_{ii} = 1$), because they represent the transition from a state to itself. The state and the transition matrix provide the basis of future phenomenon prediction.

2.6 Stochastic Models

These are the most advanced alternatives for synthetic time series generation with inclusion of all the properties in the previous models and additionally they have sound probabilistic, statistical, and transitional foundations collectively. Any time series Y_i ($i = 1, 2, \ldots, n$), has in general four distinctive components as the periodic

fluctuations, P_i, trend, T_i, sudden jump (step) J_i and stochastic, S_i components. Hence, it is possible to write a given time series mathematically as,

$$Y_i = P_i + T_i + J_i + S_i \tag{2.7}$$

After the identification of sudden jump and its separation, this expression becomes with remaining components of jump (sudden change) free Y_i time series as,

$$X_i = P_i + T_i + S_i \tag{2.8}$$

After the trend separation trendless time series, Z_i, expression takes the following form.

$$Z_i = X_i - T_i = P_i + S_i \tag{2.9}$$

It is now time to try and separate the periodic component, which can be done through the harmonic analysis (Sect. 2.6.3).

Following the separation of the periodic component, the remaining stochastic part, S_i, has inherit random variability that can be treated by probabilistic, statistical, and stochastic evaluations methodologies (Box and Jenkins 1970). Most often, these methodologies are applied automatically to available records, and consequently, there are numerous papers published in different disciplinary journals that do not provide any new approach, but the application of well-known methods to specific data sets leads to desired information. It is recommended that original methodology development foundations are first based on the qualitative information deductions, which help to establish theoretical backgrounds with a set of fundamental assumptions among which the homogeneity and stationarity are the most important ones.

2.6.1 Homogeneity (Consistency)

This assumption is valid only in the case when the record series originate from the same population. This implies that the record series has a constant time invariant arithmetic average, which also means that the record temporal variation is free of trend or jump (shift) components. Otherwise, the records have heterogeneous structure, which need comparatively rather complex mathematical, probabilistic, statistical, and stochastic treatments.

Since, in homogeneous series the arithmetic average is time variant, the simplest method to check for homogeneity is to compare the arithmetic averages after the division of the original time series into two or more portions of the same length. Buishand (1982, 1984), Jayawardena and Lau (1990) have summarized the application of three statistical homogeneity tests.

Under a null hypothesis, H_o, a given time series, Y_i (i = 1, 2, …, n) has the same mean value throughout the effective time period. The alternative hypothesis, H_a, is generally vague, since often no reliable prior information is available about possible changes in the mean. Of course, Y_i's have some empirical pdf and in the application of the homogeneity test, a theoretical joint pdf is assumed for the Y_i's.

In general, tests require serially independent structure for time series. If the tests are performed on seasonal or annual time series then this point cannot be a significant restriction. The test statistic pdfs are derived for stochastically independent and identically distributed time series. If there are slight departures from the normality, the test can still be applied confidently. In practical homogeneity tests, generally the pdf of test statistics is overlooked. The properties of test statistics are illustrated for the case that the Y_i's are normally distributed with mean (Buishand 1982).

$$E(Y_i) = \begin{cases} \mu & i = 1, 2, \ldots, m \\ \mu + \Delta & i = m+1, m+2, \ldots, n \end{cases} \tag{2.10}$$

and the variance of the time series is simply,

$$\mathrm{Var}(Y_i) = \sigma_Y^2 \tag{2.11}$$

According to this model, there is a shift (sudden jump) in the time series of magnitude Δ after m observations, and therefore, it is not a homogeneous time series. Homogeneity implies that the data in the series belong to one population, and hence, have a time invariant mean. Heterogeneity may arise due to changes in the method of data collection and/or the environment in which it is done (Fernando and Jayawardena 1994).

2.6.2 Stationarity

Different samples from the same population have practically the same statistical parameters within the range of sampling error (variability). Any time series with all the statistical parameters without significant change is referred to as the strictly stationary process. This is an impossible property in natural records. However, in practical applications, weakly (second order) stationary records are suitable for the application of the classical statistical methodologies including the stochastic processes. This type of stationarity implies that the time series has the first-order (arithmetic average) and second-order (variance) moments depending on the time differences (Box and Jenkins 1970). Independence of the variance from time is referred to as the homoscedascity in the statistics literature.

In order to check the stationarity property, at least two non-overlapping parts are considered from the original time series. If these two subseries look similar then visually one can say that the original series is stationary. This implies that stationary time series cannot include trends, jumps, or periodicities. Stationarity can be

checked either by parametric or nonparametric tests. Parametric tests are employed usually in the analysis of economic time series based on a certain number of data (Aigner et al. 1977; Bauer 1990).

The researchers who care for the frequency properties of time series prefer to work with nonparametric stationarity tests. Among these researchers are electronic engineers and a certain branch of statisticians and stochastic process experts. They consider the whole system as a "black box", where only input and output signals are important and the system identification may be achieved through some simple procedures such as the regression technique and spectral analyses. Depending on the work type, researcher uses parametric and nonparametric approaches. The significance of nonparametric approaches is that they are not based on the assumption of normal pdf. This point makes the nonparametric approaches to be used more frequently in practical applications even though they are less powerful than the parametric alternatives. As suggested by Bethea and Rhinehart (1991) in order to reach almost to the same conclusions, the nonparametric tests need 5–35% more data than parametric tests.

2.6.3 Periodicity (Seasonality)

It is well known that the periodic fluctuations are embedded into a natural time series as a result of mainly astronomic events such as Earth's rotation around the sun annually with implication of seasonality; diurnal variations due to day–night variations. In the social and economic time series, seasonality is the main factor for the periodic component existence. In general, such variations in any time series records become graspable and quantifiable at time scales less than a year (daily, weekly, monthly, three-monthly, and six-monthly). Figure 2.13 presents different periodicities in the given time series.

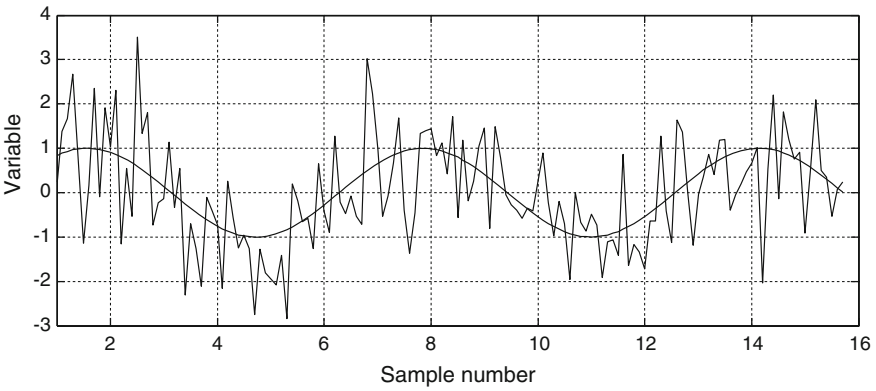


Fig. 2.13 Periodicity (seasonality) components

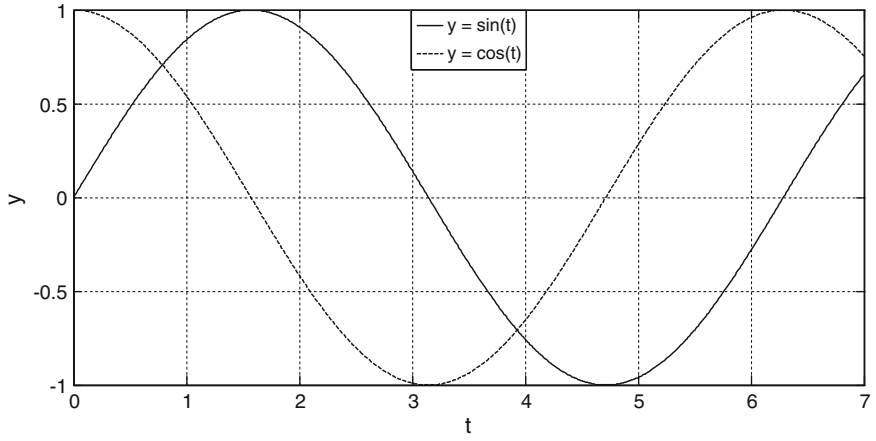


Fig. 2.14 Sine and cosine waves

In order to quantify and detect the periodicity, the commonly used methodology is the Fourier series (Maidment and Parzen 1984; Kite 1989; Jayawardena and Lai 1989; Pugacheva et al. 2003). Some researchers like Jayawardena and Lai (1989) have used the autocorrelation technique for testing the periodicity in time series.

Periodic component is the part of time series which reflects the seasonal effects. The astronomical effects in any time series can be observed provided that record durations are less than one year such as day, week, month, and season. Periodic fluctuations can be expressed as regular sine and cosine waves as in Fig. 2.14.

These waves have their amplitudes, a, basic wave period, T, and phase angle, Θ. These three quantities define a sine wave as,

$$Y_t = a \sin\left(2\pi \frac{t}{T} + \Theta\right) \tag{2.12}$$

Cosine waves are also defined in a similar way. In order to identify the periodic component in a time series, a series of waves first a sine (and cosine) wave is considered with basic wave length equal to the whole record length, then equal to the half of the total length, then one-third, etc. The summation of regular waves leads to a rather random (irregular) looking wave as in Fig. 2.15.

In this manner, it is possible to approach an irregular wave, like a given time series, by the summation of various regular waves. A regular sine wave can be expressed as,

$$Y_{i1} = a_1 \sin\left(2\pi \frac{1}{n} i + \Theta_1\right) \tag{2.13}$$

Each wave is called as a harmonic. In general, the frequency of jth harmonic has a frequency as j/n and its regular wave expression is,

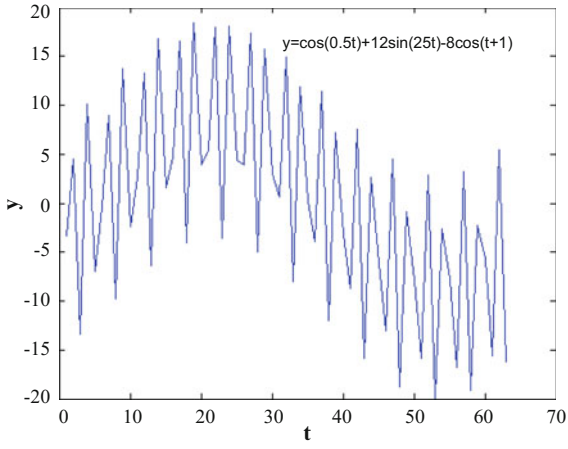


Fig. 2.15 Regular wave summations

$$Y_{ij} = a_j \sin\left(2\pi \frac{1}{n} i + \Theta_j\right) \tag{2.14}$$

The expansion of this sine wave gives,

$$Y_{ij} = a_j \sin\left(2\pi \frac{j}{n} i\right) \cos\Theta + a_j \cos\left(2\pi \frac{j}{n} i\right) \sin\Theta \tag{2.15}$$

Since $a_j \cos\Theta_j$ and $a_j \sin\Theta_j$ are constant, they are represented by A_j and B_j, and hence, the previous expression takes the following form,

$$Y_{ij} = A_j \sin\left(2\pi \frac{j}{n} i\right) + B_j \cos\left(2\pi \frac{j}{n} i\right) \tag{2.16}$$

This is the contribution of jth harmonic to ith data value. If m harmonics are considered then the time series will have the approximation as,

$$Y_i = \sum_{i=1}^{m} \left[A_j \sin\left(2\pi \frac{j}{n} i\right) + B_j \cos\left(2\pi \frac{j}{n} i\right)\right] \tag{2.17}$$

This has a zero arithmetic average value, and hence, it alone cannot represent the arithmetic average of the time series. Therefore, it is necessary to add the average term, $\overline{Y}$, which leads to,

$$Y_i = \overline{Y} + \sum_{i=1}^{m} \left[A_i \sin\left(2\pi \frac{j}{n} i\right) + B_i \cos\right] \left(2\pi \frac{j}{n} i\right) \tag{2.18}$$

In practice, since the number of harmonics, namely m, is a finite value not more than 7, this last expression approaches a given time series with error term, h_i, and finally, the periodicity equation takes the following form.

$$Y_i = \overline{Y} + \sum_{i=1}^{m} \left[A_i \sin\left(2\pi \frac{j}{n} i\right) + B_i \cos \right] \left(2\pi \frac{j}{n} i\right) + h_i \tag{2.19}$$

In this expression, A_j's and B_j ($j = 1, 2, \ldots, m$), there are $2m$ unknowns. They can be obtained from a given time series value by the minimization of sum of error squares, min (Σh_i^2) condition, leading to the following expressions.

$$A_j = \frac{2}{n} \sum_{i=1}^{n-1} Y_i \cos\left(2\pi \frac{j}{n} i\right) \tag{2.20}$$

and

$$B_j = \frac{2}{n} \sum_{i=1}^{n-1} Y_i \sin\left(2\pi \frac{j}{n} i\right) \tag{2.21}$$

The summation of the squares of these terms is equivalent to the variance of the given time series as,

$$\sigma_j^2 = A_j^2 + B_j^2 \tag{2.22}$$

and the phase angle is defined as,

$$\Theta_j = \tan^{-1}\left(\frac{A_j}{B_j}\right) \tag{2.23}$$

The major defect of this approach is that the frequencies must be whole number divisions as $1/n$, $2/n$, …, m/n.

2.6.3.1 Known Period Case

If the basic period in a time series is known, then the periodicity component can be eliminated by using simple statistical parameters without any consideration of trigonometric functions. For instance, if hourly data are available, then it is known that the periodicities are confined within the 24-h period, and therefore, a table similar to Table 2.1 can be presented for the exposition of available data and there are $N = 24n$ hourly values, where n is the number of days. If such a time series is shown as Y_0, Y_1, Y_2, …, Y_{N-1}, their exposition is given in Table 2.1.

In the last two rows, the arithmetic averages, ($\overline{Y}_i$, $i = 0, 1, 2, \ldots, 23$), and the standard deviations, (σ_i, $i = 0, 1, 2, \ldots, 23$), of hourly data are calculated. If the arithmetic average of each hour is subtracted from the corresponding hourly data,

Table 2.1 Hourly data

Y_0	y_1	y_2	.	.	.	y_{24}
Y_{24}	y_{25}	y_{26}	.	.	.	y_{48}
Y_{48}	y_{49}	y_{50}	.	.	.	y_{60}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
$Y_{24(n-1)}$	$y_{24(n-1)+1}$	$y_{24(n-1)+2}$	.	.	.	y_{24n-1}
$\overline{Y}_0$	$\overline{Y}_1$	$\overline{Y}_2$	.	.	.	$\overline{Y}_{23}$
σ_0	σ_1	σ_2	.	.	.	σ_{23}

Table 2.2 Periodicity free hourly data

$Y_0 - \overline{Y}_0$	$Y_1 - \overline{Y}_1$	.	.	.	$Y_{23} - \overline{Y}_{23}$
$Y_{24} - \overline{Y}_0$	$Y_{25} - \overline{Y}_1$	.	.	.	$Y_{47} - \overline{Y}_{23}$
$Y_{48} - \overline{Y}_0$	$Y_{49} - \overline{Y}_1$	.	.	.	$Y_{60} - \overline{Y}_{23}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$Y_{24(n-1)} - \overline{Y}_0$	$Y_{25(n-1)+1} - \overline{Y}_1$	.	.	.	$Y_{24n-1} - \overline{Y}_{23}$
0	0	.	.	.	0
σ_0	σ_1	σ_2	.	.	σ_{23}

then the remaining term does not have any more periodic fluctuation on the arithmetic average level. This subtraction procedure is shown in Table 2.2.

One can notice that in this table, the arithmetic average of each column is zero, but the standard deviations remain without any change. In order to eliminate the periodicity effect in the standard deviation, each column in the previous table must be divided by the standard deviation of the column leading to Table 2.3. The time series in this table is a standardized data, because it has zero arithmetic average and unit variance.

Table 2.3 Standardized data

$(Y_0 - \overline{Y}_0)/S_0$	$(Y_1 - \overline{Y}_1)/S_1$	.	.	.	$(Y_{23} - \overline{Y}_{23})/S_{23}$
$(Y_{24} - \overline{Y}_0)/S_0$	$(Y_{25} - \overline{Y}_1)/S_1$	.	.	.	$(Y_{47} - \overline{Y}_{23})/S_{23}$
$(Y_{48} - \overline{Y}_0)/S_0$	$(Y_{49} - \overline{Y}_1)/S_1$	.	.	.	$(Y_{60} - \overline{Y}_{23})/S_{23}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$(Y_{24(n-1)} - \overline{Y}_0)/S_0$	$(Y_{25(n-1)+1} - \overline{Y}_1)/S_1$	.	.	.	$(Y_{24n-1} - \overline{Y}_{23})/S_{23}$
0	0	.	.	.	0
1	1	.	.	.	1

2.7 Time Series Truncation

It is possible to explore the internal structure of any time series by truncating it at a certain truncation level, Y_0 (Şen 2015). Such a truncation gives rise to two-valued verbal variables such as deficit/surplus, dry/wet, cloudy/non-cloudy, flood/drought, hot/cold, rainy/non-rainy, gain/loss, etc. These two-valued variables help decision maker to base his/her final plans toward a certain goal. In some system design studies, the variables must be categorized into two classes on the basis of a certain truncation level. Let us consider that for practical applications, the time series given in Fig. 2.16 is truncated at Y_0 level.

After the truncation, the time series is converted into two mutually exclusive events along the time axis as surplus, S_i, and deficit, D_i. In mathematical sense, surpluses have positive and deficits have negative values. In general, when a time series, Y_i ($i = 1, 2, \ldots, n$) is truncated at Y_0 constant level then at the ith location, there is either $S_i = Y_i - Y_0 > 0$ or deficit $D_i = Y_0 - Y_i < 0$. The following properties are observable from such a truncation.

(1) Along the time series, there are appearances of S_i and D_i in a randomly alternate manner. The first important point is that at two successive time instances there are four possible events as deficit–surplus (DS), deficit–deficit (DD), surplus–deficit (SD), or surplus–surplus (SS),
(2) If there are n elements in a time series with n_d deficits then the number of surpluses is,

$$n_S = n - n_d$$

or

$$n_d + n_S = n \tag{2.24}$$

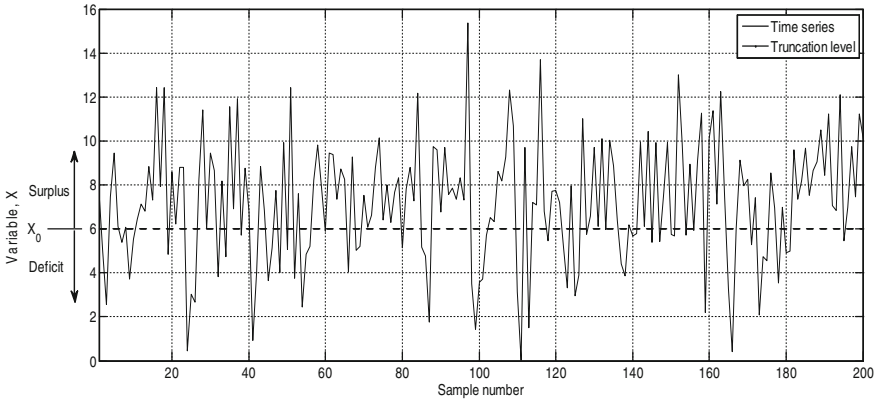


Fig. 2.16 Time series truncation

Dividing both sides by the total number of elements, n, yields,

$$p_S + p_d = 1 \tag{2.25}$$

Herein, $p_d = n_d/n$ is the probability (percentage) of deficits and likewise $p_s = n_s/n$ is for surplus probability. Depending on the level of truncation, Fig. 2.17 indicates the relationship between the truncation level and these probabilities,

It is noted that the truncation level changes between the maximum, $Y_{\max}$ and minimum, $Y_{\min}$ data values. In the case of symmetric relative frequency distribution like the normal (Gaussian) pdf, the truncation level that is equal to the arithmetic average is also equal to the median and model levels, which implies that $p_s = p_d = 0.5$. In this case approximately, $Y_0 = (Y_{\max} + Y_{\min})/2$.

(3) In case of uninterrupted sequence of two or more deficit (surplus) events, a deficit (surplus) period is valid. These periods follow each other along the time axis alternatively. In a given time series, the difference between the number of surplus and deficit periods is either 0 or 1.
(4) The maximum deficit duration corresponds to the critical deficit period within the given time series,
(5) The transition from a deficit period to surplus has DS bivariate event, whereas SD bivariate event is valid in the case of surplus followed by deficit period. The first bivariate event is referred to as the upcrossing and the second one as downcrossing event. The more these bivariate variables are in a time series the less is the dependence.
(6) The summation of deficits (surpluses) along a deficit (surplus) duration is referred to as the deficit (surplus) magnitude.
(7) The division of magnitude to duration is the deficit (surplus) intensity.

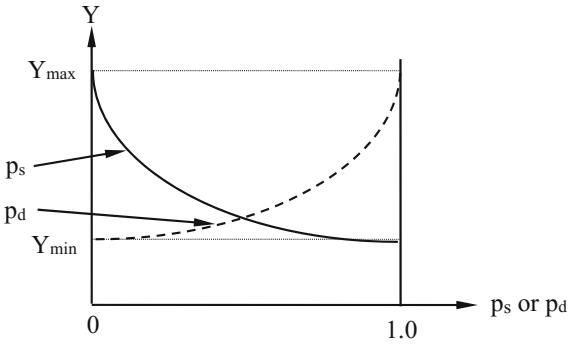


Fig. 2.17 Surplus and deficit percentages

2.7.1 Statistical Truncations

In the statistical studies, the deviations from the arithmetic average for a given time series data $(Y_i - \overline{Y})$ are very significant. Such deviations constitute the basic definitions of variance, covariance, correlation coefficients, and the coefficient of determination in regression analysis. It is possible to categorize the overall time series on the basis of standard deviation distances above and below of the arithmetic average as in Fig. 2.18.

In this figure, σ indicates the standard deviation of the whole time series. With 1, 2, and 3 standard deviation limits, a given time series may be viewed in seven categories as in Table 2.4 with different specifications.

In practice, most of the time series values fall within the normal limits with extreme values outside above and below normal extreme limits. In order to standardize all the time series to a common dimensionless base, the standard values, y_i, can be obtained according to the following formulation.

$$y_i = \frac{Y_i - \overline{Y}}{\sigma_Y} \quad (i = 1, 2, \ldots, n) \tag{2.26}$$

In the case of normal (Gaussian) pdf consideration of 1, 2, 3, and 4 standard deviation values around the arithmetic mean leads to the following numerical percentages.

In interval,

$-1 < y_i < +1$ 68.269%, i.e., with probability 0.68269
$-2 < y_i < +2$ 95.450%, i.e., with probability 0.95450
$-3 < y_i < +3$ 99.730%, i.e., with probability 0.99730
$-4 < y_i < +4$ 99.994%, i.e., with probability 0.99994

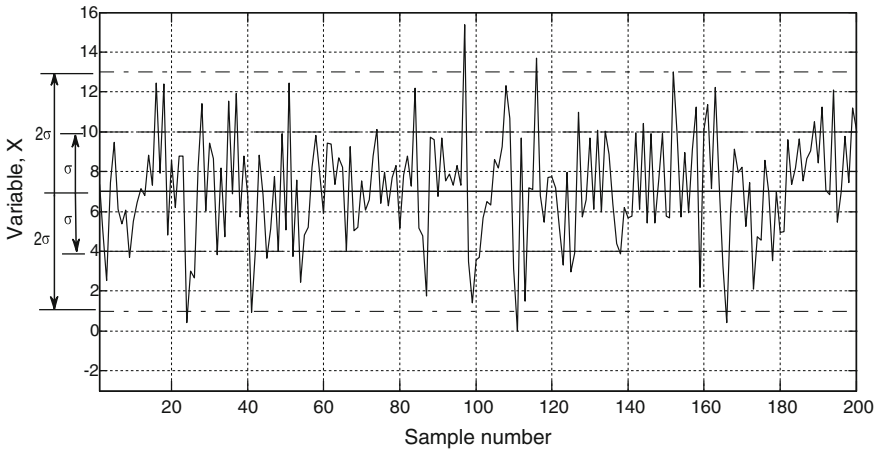


Fig. 2.18 Standard deviation truncation

Table 2.4 Truncation levels and specifications

Truncation	Specification
$\overline{Y}+3\sigma<Y_i$	Above normal extreme
$\overline{Y}+3\sigma<Y_i<\overline{Y}+2\sigma$	Rather super-normal extreme
$\overline{Y}+2\sigma<Y_i<\overline{Y}+1\sigma$	Above normal
$\overline{Y}+1\sigma<Y_i<\overline{Y}-1\sigma$	Normal
$\overline{Y}-1\sigma<Y_i<\overline{Y}-2\sigma$	Below normal
$\overline{Y}-2\sigma<Y_i<\overline{Y}-3\sigma$	Rather subnormal extreme
$Y_i<\overline{Y}-3\sigma$	Below normal extreme

In Fig. 2.19, a standard normal (Gaussian) pdf is shown with arithmetic mean $\bar{y}=0$ and variance 1 in addition to the categorical division according to the standard deviation at three levels.

Different from the statistical truncation, there are others that are useful for various human activities. In such truncations, the comfort and benefit of humans are taken into consideration. These are referred to herein as the engineering truncations. For instance, for the comfort of humans, the temperature must not be under 15 °C, and for the plant life below 7 °C. The daily water demand of Istanbul City, Turkey, is 1.5×10^6 m^3, which can be considered as the truncation level for water supply to the city.

In the previous explanations, the values below and above of any truncation level are given in terms of numbers, percentages, or probabilities. However, as shown in Table 2.5, it is also possible to specify different phenomena with different words verbally.

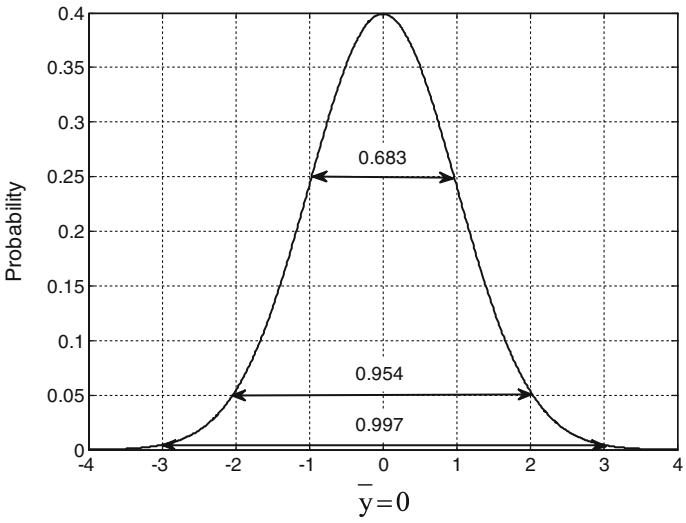


Fig. 2.19 Standart normal distributions

Table 2.5 Time series truncation and specifications

	Temperature	Rainfall	Runoff	Humidity	Cloud	General
If $Y_i < Y_0$	Cold	Rainy	Dry	Humid	Open	Deficit
If $Y_i > Y_0$	Hot	Non-rainy	Wet	Non-humid	Close	Surplus

These bivariate specifications play important role in many diverse human social, environmental, economic, health, and engineering activities.

2.8 Data Smoothing

In the records of time series, there are local haphazard and very random fluctuations that may sometimes hide the general variation trend. In order to get rid of these disturbances, it is necessary to smooth the time series through some procedures. In general, the equation for a time series can be written as composed of deterministic, D_i, and stochastic, S_i, parts similar to Eq. (2.7). The time series components are already explained in Sect. 2.6. The summation of the random component, and hence, its arithmetic average is equal to zero, and therefore, the arithmetic average of the process is equal to the arithmetic average of the deterministic part, i.e., $\overline{Y} = \overline{D}$. This shows that random component can be eliminated through some average procedure. The remaining deterministic part is the smoothened part of the time series which is shown in Fig. 2.20.

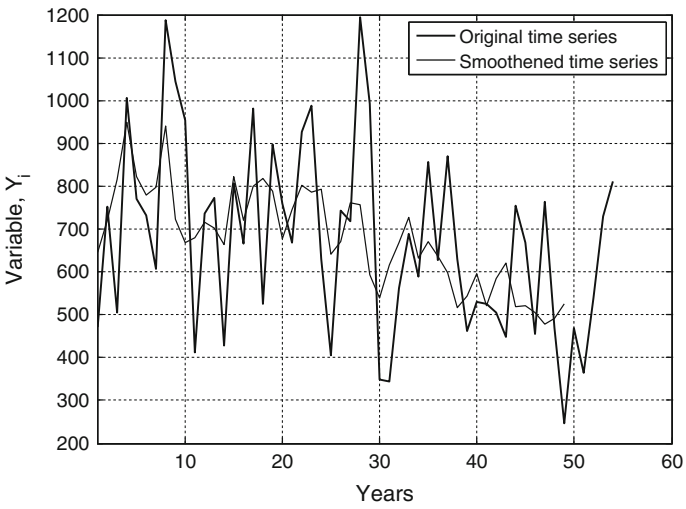


Fig. 2.20 Smoothened time series

2.8.1 Moving Averages

Informal regression methods based on moving averages at certain window widths are used as the smoothing techniques to disclose possible hidden trend components. Moving average methodology helps to identify and highlight possible long-term nonlinear trends with smoothing of short-term fluctuations. Moving average procedure is commonly used in many economy sectors.

The most frequently used procedure for smoothing is the moving average approach, where a certain length of series is replaced in an overlapping manner by the arithmetic average. For instance, in Fig. 2.20, 5-year window width is used for successive arithmetic moving average smoothening. In most applications, first, third-order moving average procedure is recommended and subsequently the order can be increased up to seventh order, if necessary. If a time series is, Y_i (i = 1, 2, …, n), its third-order moving average smoothing X_i (i = 1, 2, …, n − 2) can be achieved as follows.

$$X_1 = \frac{Y_1 + Y_2 + Y_3}{3},\ X_2 = \frac{Y_2 + Y_3 + Y_4}{3},\ \ldots,\ X_{n-2} = \frac{Y_{n-2} + Y_{n-1} + Y_n}{3}$$

In a third-order moving average, there are n − 2 terms from a time series of length n. In the case of m-order moving average procedure, there are $n - m + 1$ terms, X_i (i = 1, 2, …, $n - m + 1$).

The above-mentioned moving average gives equal weights to each part of smoothened time series. However, in some cases, it is necessary to give different weights to each smoothening term. In practice, the most frequently used versions are as follows.

$$X_i = \frac{Y_i + 2Y_{i+1} + Y_{i+2}}{4} \tag{2.27}$$

or

$$X_i = \frac{Y_i + 4Y_{i+1} + 6Y_{i+2} + 4Y_{i+3} + Y_{i+4}}{16} \tag{2.28}$$

2.8.2 Difference Smoothing

A very simple procedure is the successive difference method, which for a given series, Y_i, with lag-one difference operation, the terms in a new time series, X_i, become with n – 1 terms as,

$$X_1^{(1)} = Y_2 - Y_1,\ X_2^{(1)} = Y_3 - Y_2,\ \ldots,\ X_{n-1}^{(1)} = Y_n - Y_{n-1}$$

If the difference is taken at lag-m apart from each other, then the new series X_i has $n - m$ terms as,

$$X_1^{(m)} = Y_m - Y_1,\ X_2^{(m)} = Y_{m+1} - Y_2,\ \ldots,X_m^{(m)} = Y_n - Y_m$$

Example 2.1 In Table 2.6, 23 terms are given as a time series in the first column and it is smoothened according to difference procedure at lags 1, 2, 3, and 10 in the same table. It is obvious that lag-10 differences have more fluctuations, because the successive terms become more independent from each other. In general, the further away the two values are from each other the less is the dependence between them.

Table 2.6 Application of difference procedure

Data (X_i)	Differences			
9.05	1	2	3	10
	−0.96			
8.08		0.79		
	−0.17		0.96	
7.97		1.75		
	1.58		−1.00	
9.50		0.75		
	2.33		−4.67	
11.83		−3.92		
	−1.59		7.05	
10.84		3.13		242.2
	1.54		−1.08	
11.78		2.06		67.2
	3.59		−4.95	
15.37		−2.90		−366.1
	0.69		3.38	
16.06		0.48		544.5
	1.17		−4.83	
17.23		−4.35		−622.2
	−3.18		6.67	
14.05		2.32		661.8
	−0.86		−3.13	
13.19		−0.81		−714.2
	−1.67		5.79	
11.52		4.98		759.1
	3.31		−8.88	
14.83		3.90		−708.8
	−0.59		5.02	
14.23		1.12		533.7

(continued)

Table 2.6 (continued)

Data (X_i)	Differences			
	0.53		−3.13	
14.77		−2.01		−317.6
			1.03	
13.29		−0.98		162.2
	−2.46		0.78	
10.83		−0.20		−90.9
	−2.66		2.15	
8.17		1.95		
	−0.71		0.04	
7.46		1.99		
	1.28		−1.22	
8.74		0.77		
	2.05		−3.75	
10.79		−2.98		
9.86				
Sample number, n				
23	22	21	20	13
Average				
11.68	0.0372	0.001	−0.189	11.60

2.9 Jump (Shift)

This implies a sudden change (downward or upward) in a time series, which may also have a linear downward or upward trend. Such changes are common in financial time series and also in a surface flow discharge record series after the construction of a dam or a diversion channel (Fig. 2.21).

For a jump component, there is almost a sudden change in the effective environmental conditions.

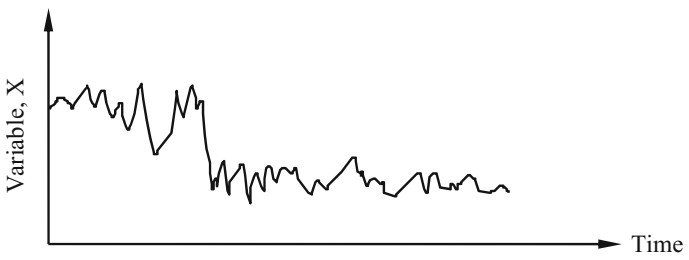


Fig. 2.21 Sudden jump components in a time series

(1) If the location of the measurement station is changed, then the change in the environmental conditions may lead to sudden effects in the time series measurements. Such a jump may be very distinctive as shown in Fig. 2.21,
(2) After natural hazards, there may appear sudden jumps in the measurements. For instance, eruptions may load the lower atmosphere with dust, and accordingly, this may cause sudden jumps in some meteorological records,
(3) Due to miss-maintenance, local defects in the instruments may lead to sudden jumps. For instance, if there appears a small hole in the rain gauge then the measurements will be lower than before.

2.10 Correlation Coefficients

In any time series, apart from the visible components in a graph, there are also non-visible features that need identification and quantitative evaluation. These features are concerned with the internal structure of a given time series. The most significant one is the serial dependence, which is concerned with the question, whether there is an effect of any event occurrence in a time to the next time step occurrence? Such a feature, which is referred to as the serial autocorrelation is an indispensable component relevant to any time series whether natural or artificial. It is also known as a memory effect in two types as the short-memory and long-memory effects, where the latter type is the persistence. Scientific terminology for the short memory effect is the autocorrelation coefficient. Linguistically and qualitatively, the short-term memory effect can be expressed as "time series low values follow low values and high values follow high values". This expression provides an ability to visualize a time series and then to deduce whether there is a short-memory effect or not.

In general, processes can be viewed under two very broad categories as dependent and independent according to time scale considerations. Usually, the smaller the time interval between two successive events the greater is the dependence, and hence, there is persistence, but large-time apart natural events imply independence. This classification is also in accord with the rarity or frequency of the event. For instance, flood and earthquake occurrences are among rare natural events that occur along time axis, and therefore, they are considered as independent from each other. Similar arguments are valid also for low natural events such as droughts. In such problems, the serial (internal) correlation coefficient is ignored and the probabilistic treatment of the successive event occurrences becomes very easy according to the probabilistic modeling.

Correlation coefficients are useful in determination of the relationship strength between two variables. Two different time series can be related to each other proportionally, inversely or there may not be any correlation between them such a relationship is calculated by the cross-correlation coefficient. For the quantification

of correlation, there are different procedures as parametric and nonparametric alternatives.

On the other hand, within the same time series the successive data values might affect each other, which is expressed by the serial correlation coefficient. For instance, rainfall of today might be affected partially from yesterday's rainfall occurrence. In general, rainy periods follow rainy periods and dry periods follow dry periods. Furthermore, high rainfall amounts follow high amounts and low values follow low values. These two statements indicate that so far as the rainfall occurrences and their amounts are concerned, there are serial (internal) relationships to a certain extent.

In mathematics, when two variables are related to each other their variation or plots on a Cartesian coordinate system does not appear as a horizontal or vertical line (see Fig. 1.2e, f), but rather a straight line with a slope or a curve with many tangential slopes (see Fig. 1.2a–d). The simplest form of dependence has linearity, which is always used in the statistics or stochastic modeling works. If there are two different time series, they can be plotted as one versus the other. By visual inspection of the scatter points, one can appreciate whether the dependence is high or low and directly or reversely proportional. In the case of scatters around a straight line (trend line) there is dependence.

For serial correlation structure, time series Y_1, Y_2, …, Y_n, is shifted by a certain lag (for instance lag-one) so as to obtain another parallel time series as Y_2, Y_3, Y_4, …, Y_{n-1}. They have $n - 1$ common point. The scatter diagram of these two time series gives rise to $n - 1$ scatter points on the Cartesian coordinate system (Fig. 2.22).

If straight-line trend appears through the scatter points then it is possible to conclude that there is dependence between the two variables, otherwise they are independent. The most suitable straight line through these scatter points gives the dependence measurement as its slope. The more the deviation of the slope from ±45° (1:1 and −1:−1) line is, the smaller is the dependence. In Fig. 2.23 an independent scatter diagram is shown.

2.10.1 Pearson Correlation Coefficient

There are two types such as serial correlation and cross-correlation. The serial correlation coefficient, ρ_k, is expressed for a given time series, Y_i, (i = 1, 2, …, n) and lag-k as follows.

$$\rho_{\mathrm{sk}} = \frac{\sum_{i=1}^{n-k} \left(Y_i - \overline{Y}\right)\left(Y_{i-k} - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n-k} \left(Y_i - \overline{Y}\right)^2} \sqrt{\sum_{i=n-k}^{n} \left(Y_i - \overline{Y}\right)^2}} \tag{2.29}$$

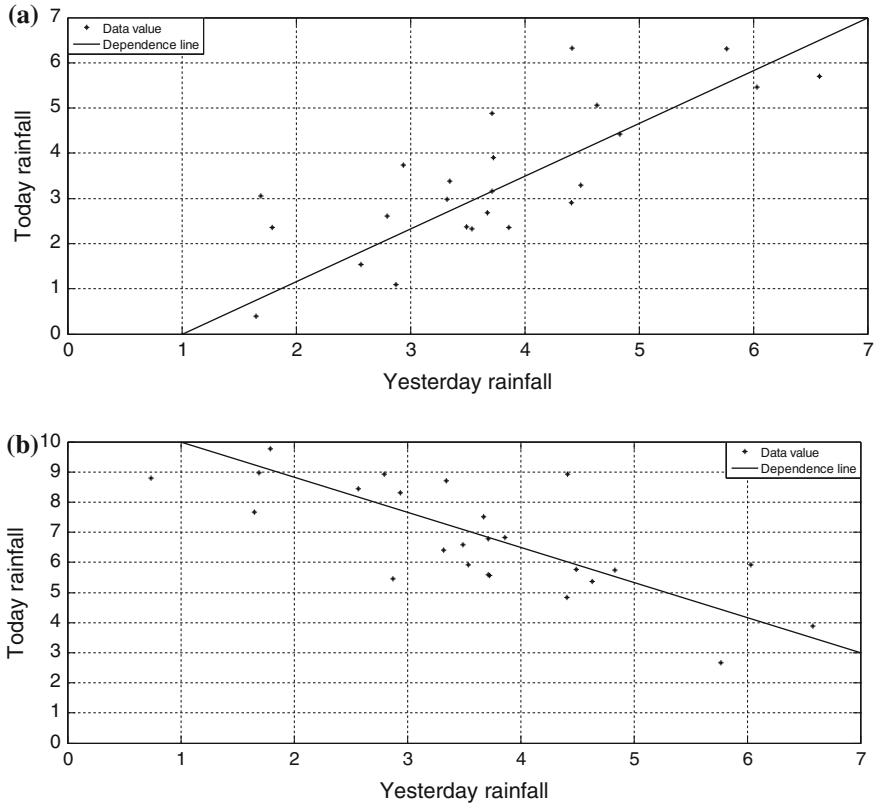


Fig. 2.22 Dependent scatter diagrams, **a** positive dependence, **b** negative dependence

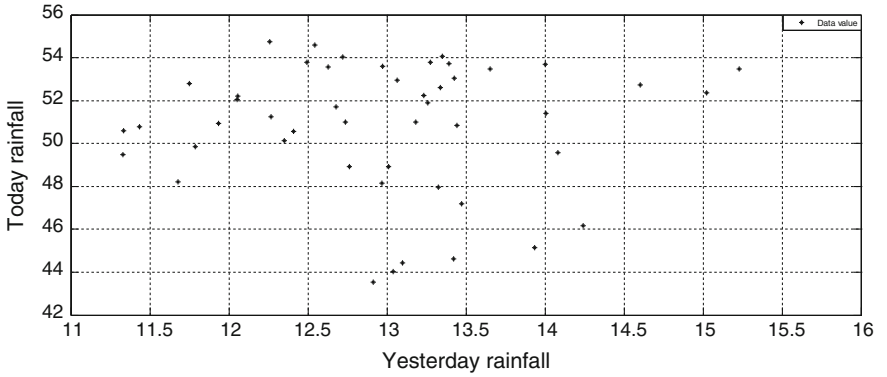


Fig. 2.23 Independent scatter diagrams

On the other hand, similarly the Pearson cross-correlation, ρ_c, between two time series Y_i and X_i is defined as,

$$\rho_c = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{(Y_i - \overline{Y})^2}}, \tag{2.30}$$

where $\overline{X}$ and $\overline{Y}$ are the mean of respective time series. It is also possible to search for lag-one or more cross-correlations.

The correlation coefficient takes values between −1 and +1. The closer the cross-correlation coefficient values are to zero the more random, i.e., independent are the two time series, otherwise, values close to +1 (−1) imply positively (negatively) strong cross- or serial correlations. Positive correlation means direct proportionality (see Fig. 2.22a) and negative value shows inverse proportionality (see Fig. 2.22b). In the case of positive dependence, high (low) values follow high (low) values, whereas in the case of negative dependence high (low) values follow low (high) values. The dependence that is calculated through Eq. (2.29) is the serial correlation or autocorrelation coefficient. Similarly to lag-one, lag-two or more lag correlation coefficients can be calculated simply from a given time series. In general, there is an upper practical limit for lag-k as $k \leq n/3$. The theoretical distribution parameters as the average and variance for lag-one in Eq. (2.29) are given as (Anderson 1942),

$$\overline{\rho_{\text{sk}}} = -\frac{1}{(n-1)} \tag{2.31}$$

and,

$$\sigma_{\text{sk}}^2 = \frac{1}{(n-1)}, \tag{2.32}$$

respectively. The pdf of ρ_c was shown to be asymptotically normal with the mean $E(\rho_P) = 0$ and variance as in Eq. (2.32). In the test of serial correlation, a single-tail normal pdf is used.

As mentioned earlier, the Pearson correlation coefficients assume any value between −1 and +1, inclusive with specifications in Table 2.7. One should not memorize this table, because they are more or less the subjective opinion of the author. Other authors may deviate slightly from these specifications owing to their experiences, but such deviations are not significant in practical works. It must be kept in mind that correlation coefficients are the measure of linear dependence between two variables or within the same time series. If the correlation is not linear, then these definitions are invalid.

Table 2.7 Correlation coefficient classes

Numerical value intervals	Linguistic interpretations
$\rho_P = -1.0$	Completely negative dependence
$-1.0 < \rho_P < -0.9$	Strong negative dependent
$-0.9 < \rho_P < -0.7$	Quite negative dependence
$-0.7 < \rho_P < -0.5$	Weak negative dependence
$-0.5 < \rho_P < -0.3$	Very weak negative dependence
$-0.3 < \rho_P < -0.1$	Insignificant negative dependence
$\rho_P = 0.0$	Complete independence
$0.1 < \rho_P < 0.3$	Insignificant positive dependence
$0.3 < \rho_P < 0.5$	Very weak positive dependence
$0.5 < \rho_P < 0.7$	Weak positive dependence
$0.7 < \rho_P < 0.9$	Quite positive dependence
$0.9 < \rho_P < 1.0$	Strong positive dependence
$\rho_P = 1.0$	Complete positive dependence

The following points are the deficiencies of the Pearson correlation coefficient concept in practical applications.

(1) Even if the correlation is not linear, the correlation value will appear between −1 and +1. This may not have logical and physical meaning, because the Pearson correlation coefficient definition is valid for linear relationships,
(2) If there are one or more extreme values in a time series, then these values affect Eq. (2.30) in such a manner that the correlation coefficient appears biased and/or unrepresentative,
(3) The data must abide with a normal pdf, otherwise, the correlation coefficient is not meaningful,
(4) For meaningful and reliable correlation coefficient calculations, the standard deviation of the data must be constant, i.e., homoscedasticity property must be valid,
(5) Correlation coefficient definition in Eq. (2.30) cannot be used for verbal and linguistic data,
(6) If data is transformed by any means to have normal pdf then the correlation coefficient of the transformed data is not the same with the original data (Şen 1977). Even the reverse transformation does not guarantee that the correlation coefficient is equal to the observed correlation value.

After what has been said above, it is obvious that the domain of the Pearson correlation coefficient is rather restrictive, and prior to its use all necessary assumptions must be cared for their validity.

2.10.2 Kendall Correlation Coefficient

In order to alleviate the defects in the Pearson correlation coefficient, other procedures are suggested for the same purpose. One of these techniques is the consideration of data ranks instead of data values in natural sequence. This is the requirement of the Kendall correlation coefficient, ρ_K, which gets rid of the extreme value effects. This coefficient can be used even though the data may have a skewed pdf. It is applicable even in the cases of some missing data or incomplete measurements. In general, for the same data series Kendall correlation coefficient is smaller than the Pearson coefficient. For this reason, although strong correlation is observed through the use of the Pearson correlation coefficient as 0.9 or more, the same is valid as 0.7 in the case of Kendall correlation coefficient. Kendall coefficient can be calculated without any calculator even by hand. It is also capable to measure nonlinear correlations. The superiority of this coefficient over the Pearson coefficient is due to the following points.

(1) It measures even the nonlinear relationships,
(2) It is not affected by extreme values,
(3) Even after the transformations Kendall coefficient remains the same.

For instance, if $\log(Y_i)$ and $\log(X_i)$ are used instead of the original data (Y_i and X_i), the Kendall correlation coefficient will remain the same. In the calculation of this correlation coefficient the following steps are necessary,

(1) Rank one of the time series into ascending order and replace the data in the next time series with the ranks of this time series. Hence, one of the time series is ordered and the other had replacement of values according to the ranks of the first one. In the case of correlation, there will appear simultaneous increase in both series. If there is increase in the ranked time series, and decrease in the other time series implies then a negative correlation is valid. Otherwise, there is no correlation between them,
(2) Any data value in the ranked series, say Y_i, is compared with all the data after its location, Y_j ($j = i + 1, i + 2, \ldots, n$), and if $Y_i < Y_j$ then +, otherwise for $Y_i > Y_j$ a − sign is attached. For the data value at place i, there are $(n - i)$ number of alternative + and − signs. If the same procedure is repeated for all the data values without any equal data value to each other, then there are $n(n - 1)/2$ signs. If half of these have + sign then the data sequence is considered as independent. If + (−) signs are more than − (+) signs then there is positive (negative) dependence,
(3) If the total numbers of + and − signs are denoted by P and N then the Kendal correlation coefficient is defined as,

$$\rho_K = \frac{2(P - N)}{n(n - 1)} \tag{2.33}$$

By definition this has values between −1 and +1,

(4) For the test of independence the necessary statistical quantity, K, is defined as,

$$K = P - N \tag{2.34}$$

which is the difference between the numbers of + and − signs. The K value may be positive or negative with zero expectation. Theoretical studies indicate that its standard deviation can be expressed as,

$$\sigma_K = \sqrt{\frac{n(n-1)(2n+5)}{18}} \tag{2.35}$$

On the other hand, for more than 10 data values, the distribution of the Kendall correlation coefficient approaches the Gaussian pdf. Whether ρ_K is different from zero can be tested by the standard normal pdf. If the necessary standard value is S, then the test statistics can be calculated as,

$$z_s = \begin{cases} \frac{S-1}{\sigma_s} & \text{if} \quad S \geq 0 \\ 0 & \text{if} \quad S = 0 \\ \frac{S+1}{\sigma_s} & \text{if} \quad S \leq 0 \end{cases} \tag{2.36}$$

If this standard value is less than the critical value found on the basis of a certain significance level then the data is considered as independent.

2.10.3 Spearman Correlation Coefficient

In the nonparametric statistics domain, the analogous to the Pearson correlation is named as the Spearman's rank correlation coefficient. Pearson correlation coefficient requires that both variables should comply by the normal pdf, which is not the case in many disciplines. For calculating this nonparametric correlation coefficient, both data sets are ordered separately from each other. Hence, there are two sequences of ranks, one for Y time series, $R(Y_i)$, and other for X time series, $R(X_i)$. If for each i the ranks are of Y_i equal to ranks of X_i, then the Spearman's rank correlation is regarded as perfect. The rank correlation is defined as the sum of the difference between the corresponding ranks of Y_i and X_i. Analogous to the parametric version of the coefficients the correlation values are scaled between −1 (perfect negative) and +1 (perfect positive) correlation. In between the value is equal to zero indicating no correlation. Spearman's rank correlation calculation steps are as follows.

(1) As the null hypothesis, H_0, the correlation between Y_i and X_i is assumed as equal to zero. This is referred to as the hypothetical correlation value, $\rho_s = 0$,
(2) Alternative hypothesis, H_a is that this correlation coefficient is different than zero,

(3) The test statistic, ρ_S, which is referred to as the Spearman's rank correlation coefficient is defined in terms of each data set ranks and the number, n, of data in each set as,

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n}[R(Y_i) - R(X_i)]}{n(n^2-1)} \tag{2.37}$$

As with the other nonparametric methods, values of X_i and Y_i can vary extensively without affecting the final result. It is necessary to keep in mind that ρ_S does not imply good linear relationship, rather than linearity. It is quite possible to obtain low Spearman's rank correlation coefficient for high Pearson's parametric correlation coefficient. However, in many applications, it is unusual for Pearson's coefficient to provide a statistical test result markedly superior to Spearman's rank correlation approach even with normally distributed data. Another version of the previously defined Spearman correlation coefficient can be found as follows,

$$\rho_S = \frac{\sum_{i=1}^{n} R(Y_i)R(X_i) - n\left(\frac{n+1}{2}\right)^2}{\frac{n(n^2-1)}{12}} \tag{2.38}$$

In the case of positive correlation, high values of Y_i ranks follow high X_i ranks; otherwise there is a negative correlation. Theoretical studies indicate that in the case of trend nonexistence for big data values, this coefficient appears according to a Gaussian pdf with the following average and variance expressions as,

$$\bar{\rho}_S = 0 \tag{2.39}$$

and

$$V_{\rho_S} = \frac{1}{(n-1)}, \tag{2.40}$$

respectively. The test must be carried out with two-tailed pdf by the assumption as the null hypothesis that there is no trend component in the time series. The test value at any significance level results as $\rho_{\text{sig}} > \rho_S$, otherwise the time series is not homogeneous. In the case of significant level $\rho_S > 0$ implies the existence of an increasing trend.

2.11 Persistence/Nonrandomness

Persistence is one of the most important properties in many system designs concerning the storage capacity of reservoirs, average return periods, failure risks, hidden periodicities, trends, and drought properties. Its consideration in analytical derivations of design criteria presents difficulties and for this reason most often the analytical expressions are obtained on the basis of nonpersistent (independent or short-memory) processes. Although the conventional autocorrelation coefficients and functions are used in many design problems, but the very definition of the autocorrelation function requires that the underlying process generating mechanism abide with normal (Gaussian) pdf. It is therefore, necessary to convert non-Gaussian pdf into normal pdf in order to make benefit of the available analytical expressions. During the transformation process, the very persistence genuine property of the basic variables is not preserved although the statistical parameters such as the average, standard deviation, skewness coefficient, and kurtosis are maintained in the transformed normal pdf.

Persistence and randomness are two distinctive properties of a time series. Randomness is another term for nonpersistence and it is defined as the independence among time series values, whereas persistence (correlation) occurs provided that the successive time series data affect each other. Persistence (correlation) is a tendency of the successive time series values to "remember" their antecedent values' influence.

2.11.1 Short-Memory (Correlation) Components

Simple successive dependence models are representations of a linear line on a Cartesian coordinate system between the value from the time series and the following value. Hence, given a time series of Y_i, Y_2, …, Y_n with n observations for the simplest successive dependence at lag-one apart, this sequence yields $n - 1$ points on a scatter diagram as shown in Fig. 2.22a, b. It is the trend slope of the straight line that is a representative of the simple dependence, i.e., short-term correlation. All the serial correlations are obtained in this manner. In Fig. 2.22a, b, the horizontal axis represents the previous value, say, Y_{i-1}, whereas the vertical axis is for current value, Y_i. It is possible to infer the simplest model (lag-one Markovian) mathematically as the straight line with deviations, u_i, from this line as,

$$Y_i = a + bY_{i-1} + u_i, \tag{2.41}$$

where, a and b are the model parameters. Such a simple model does not have any assumption concerning the pdf of the time series, but the model is based on the linearity assumption. This is one of the most significant conclusions about the serial dependence that the classical correlation coefficient measures the linear dependence, and therefore, prior to its application, it is necessary to look at the scatter diagram of successive values so as to infer whether this assumption is valid.

Otherwise, in the case of nonlinearity one can still obtain classical correlation coefficient but without knowing the underlying facts of nonlinearity. Unfortunately, most often in practical applications, this point is overlooked and moreover the correlation coefficient is calculated and applied rather blindly. The model parameters, a and b, can be obtained in any manner without formal procedures. For instance, Eq. (2.41) can be considered without error and in this case any two data values give rise to two equations, which can be solved for a and b parameters. If this procedure is applied to all possible pairs from the sequence, then a set of a and b parameters can be calculated and their averages are adopted as a and b. However, this is very naive way of parameters estimation (Chap. 4).

If the necessary tests are not performed and the data are not checked for the basic assumptions then all what have been explained above leave suspicions in the coefficient estimations. In practical studies, researchers most often do not care or even think about these restrictive assumptions, and consequently, the coefficient estimations might remain biased. Even the amount of the global bias is not known, and therefore, bias correction procedures cannot be defined and applied (Şen 1974). Hence, the parameter estimates of Eq. (2.41) remain under suspicion. In order to avoid all these restrictive assumptions rather than the application of procedural regression analysis to data with a set of restrictive assumptions, it may be preferable to try and preserve only the arithmetic averages and variances of the sequence. After all the arithmetic averages and variances are the most significant statistical parameters in any design work.

Equation (2.41) can also be interpreted as a first-order Markov process. In such a case, since always a physical value is assumed to exist, i.e., there is no zero value, it is possible to consider that $a = 0$ in this equation. On the other hand, with theoretical restrictive assumptions similar to the regression approach especially the normal (Gaussian) pdf of the physical variable, it can be shown that, Eq. (2.41) can be brought into a stochastic process form as follows.

$$(X_i - \mu) = \rho(X_{i-1} - \mu) + \sigma\left(1 - \rho^2\right)^{1/2} u_i, \tag{2.42}$$

where μ, σ, ρ, and u_i are the arithmetic average, standard deviation, first-order correlation coefficient, and uncertain residual error term, respectively. In such a model, u_i is normally distributed random variable with zero mean and unit standard deviation. The stochastic model in Eq. (2.42) generates normally (Gaussian) distributed variables. It is important to notice at this stage that the correlation coefficient is defined for linear dependence and for normal pdf stochastic variates only.

2.11.2 Long-Memory (Persistence) Component

Persistence is commonly referred to as long-term dependence between successive observation values in a time series. Natural variables (landslide, earthquake, flood, and tsunami) are uncertain in character but there are embedded features that give

rise to better quantitative description of these series. Among these features are long-term averages and standard deviations and better frequency distribution behaviors according to a theoretical pdf such as normal, logarithmic normal, Gamma, Gumbel, Pearson, etc. However, none of these features are capable to give the measure of successive dependence, except the correlation coefficient or persistence measures such as rescaled ranges (Hurst 1951; Şen 1974). Dependence measures including classical correlation and persistence strength decreases as the basic time interval of the time series increases. For instance, daily records have more dependence characteristic than annual series. Even in high (floods) or low (droughts) natural events there are persistence, but unfortunately for the sake of brevity and simplicity these are ignored in many practical applications. Assumption of serial independence makes calculations simple within the probability theory only but the conclusions always appear as over-estimations. For instance, the ideal expected size of a storage reservoir, $E(R)$, is given simply for a first-order Markovian process as (Şen 1974),

$$E(R) = \sqrt{\frac{2(1-\rho)}{\pi(1+\rho)}}\sigma, \tag{2.43}$$

where ρ is the lag-one serial correlation coefficient and σ is the standard deviation of time series. The ratio, $(1 - \rho)/(1 + \rho)$ is smaller than 1, and consequently, the expected size for dependent process is smaller than the independent case, which appears for $\rho = 0$ as,

$$E(R) = \sqrt{\frac{2}{\pi}}\sigma \tag{2.44}$$

It is, therefore, very significant to consider the short- or long-term dependences within any natural phenomena, if the design is expected to perform in the best possible manner economically. On the other hand, Douglas et al. (2000) have shown that even in the low values of seven-day basic time interval, there are significant serial correlations reaching to 25% at the runoff stations throughout the USA.

2.11.2.1 Rescaled Range and Hurst Phenomenon

Hurst (1951, 1956), Şen (1974) studied long-term fluctuations within a large number of geophysical records and found that

$$\frac{R_n}{S_n} \approx n^h, \tag{2.45}$$

where R_n is the range of cumulative departures from the sample mean and S_n is the standard deviation estimation. The range is based on the cumulative sums of departures from the time series arithmetic average as in Fig. 2.24.

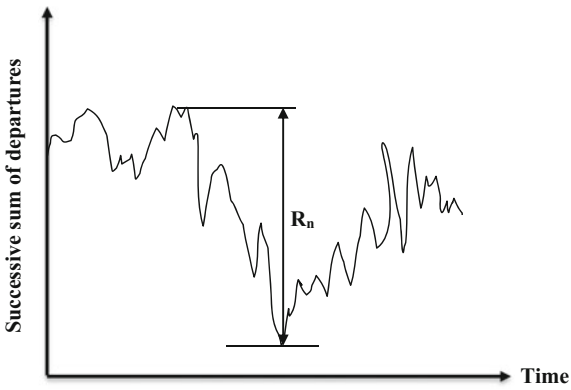


Fig. 2.24 Range of a time series

In this empirical formulation, h is referred to as the Hurst coefficient, which assumes values theoretically between 0 and 1 but for independent processes its value is equal to 0.5. In many geophysical phenomena h does not appear as 0.5, and hence, its deviation from 0.5 is called as the "Hurst phenomenon" implying long-term dependence, i.e., persistence. Such a discrepancy has been accounted on the basis of three factors.

(1) The non-normality of the pdf of the underlying variables,
(2) Effect of small samples, i.e., bias effect in the statistical sense,
(3) The autocorrelation structure.

Especially, the last factor has caused introduction of different theoretical stochastic models among which the "fractional Brownian processes" (Mandelbrot and Wallis 1968) are the major ones in addition to the white Markov (Şen 1974) or AutoRegressive Integrated Moving Average (ARIMA) processes (Box and Jenkins 1974). Division of the range, R, which is the storage volume, by the standard deviation, S, theoretically leads for serially independent processes to the expectation value as (Feller 1968),

$$E\left(\frac{R_n}{S_n}\right) = 2\sqrt{\frac{2}{\pi}n\sigma}, \tag{2.46}$$

where n is the sample length. However, for serially dependent processes, the expectation of the rescaled range, R/S, is derived by Şen (1974) as follows.

$$E\left(\frac{R}{S}\right) = \frac{2\sqrt{n}}{\sqrt{\pi(n-1)}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sum_{k=1}^{n} k^{-\frac{1}{2}}\left[\frac{1+\rho}{1-\rho} - \frac{2\rho\left(1-\rho^k\right)}{k(1-\rho)^2}\right]^{\frac{1}{2}}. \tag{2.51}$$

References

Aigner, D., Knox Lovell, C. A., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics, 6,* 21–37.

Anderson, T. W. (1942). Distribution of the serial correlation coefficients. *The Annals of Mathematical Statistics, 13*(1), 1–13.

Bauer, P. W. (1990). Recent developments in the econometric estimation of frontiers. *Journal of Econometrics, 46,* 39–56.

Bethea, R., & Rhinehart, R. (1991). *Applied engineering statistics* (p. 312). Marcel Dekker Inc.

Box, G. E. P., and Jenkins, G. (1974). Time series analysis, forecasting and control, Holden-Day, San Francisco.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Buishand, T. A. (1984). Tests for detecting a shift in the mean of hydrological time series. *Journal of Hydrology, 73,* 51–69.

Buishand, T. A. (1982). Some methods for testing the homogeneity of rainfall records. *Journal of Hydrology, 58,* 11–27.

Douglas, E. M., Vogel, R. M., & Kroll, C. N. (2000). Trends in flood and low flows in the United States: Impact of spatial correlation. *Journal of Hydrology, 1–2,* 90–105.

Fernando, D. A. K., & Jayawerdena, A. W. (1994). Generation of forecasting of mansoon rainfall data. In: *Proceedings of the 20th WEDC Conference on Affordable Water Supply and Sanitation, Columbo, Sri Lanka* (pp. 310–313).

Feller, W. (1968). *An introduction to probability theory and its applications* (Vol. I, 3rd ed.). John Wiley and Sons. Co.

Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Trans. ASCE, 77,* 1308.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers, 116,* 70–808.

Hurst, H.E. (1956). Methods of using long-term storage in reservoirs. In: *Proceedings of the Institution of Civil Engineers, Part I*, (pp. 519–542).

Jayawardena, A. W., & Lai, F. (1989). Time series analysis of water quality data in Pearl river, China. *Journal of Environmental Engineering, 115*(3), 590–607. ASCE.

Jayawardena, A. W. & Lau, W. H. (1990). Homogeneity tests for rainfall data. *Journal of the Hong Kong Institution of Engineers*, 22–25.

Kite, G. (1989). Use of time series analyses to detect climatic change. *Journal of Hydrology, 111*, 259–279.

Maidment, D. R., & Parzen, E. (1984). Time patterns of water use in six Texas cities. *Journal of Water Resources Planning and Management, 110*(1), 90–106. ASCE.

Mandelbrot, B. B., and Wallis, J. R. (1968). Noah, joseph and operational hydrology. *Water Resources Research, 4*(5), 909–918.

Mandelbrot, B. B., & Wallis, J. R. (1969a). Computer experiments with fractional Gaussian Noises. Part 1—Averages and variances. *Water Resources Research, 5*(1), 228–241.

Mandelbrot, B. B., & Wallis, J. R. (1969b). Some long run properties of geophysical records. *Water Resources Research, 5*(2), 321–340.

Mandelbrot, B. B., & Wallis, J. R. (1969c). Robustness of the rescaled range R/S in the measurement of non-cyclic long-run statistical dependence. Water Resources Research.

Parzen, E. (1960). *Modern Probability and its applications*. New York: Wiley and Sons.

Popper, K. (1954). *The logic of scientific discovery*. Routledge is an imprint of the taylor & francis group. 494 p. ISBN 0-203-99462-0

Pugacheva, G., Gusev, A., Martin, I., Schuch, N., & Pankov, V. (2003). 22-year periodicity in rainfalls in littoral Brazil. Geophysical Research Abstracts, EGS - AGU -EUG Joint Assembly, Abstracts from the meeting held in Nice, France, April 6–11, 2003, 6797.

Ripple (1881) Diagram for storage capacity determination.

Russel, B. (1948). *Human knowledge: Its scope and limits*. London: George Allen and Unwin.
Şen Z. (1974). Small sample properties of stationary stochastic processes and hurst phenomenon in hydrology. Unpublished Ph. D. Thesis, University of London, 256 pp.
Şen, Z. (1977). Run-sums of annual flow series. *Journal of Hydrology, 35*, 311–324.
Şen, Z. (2012). Innovative trend analysis methodology. *Journal of Hydrologic Engineering, 17*(9), 1042–1046.
Şen, Z. (2014). Trend identification simulation and application. *Journal of Hydrologic Engineering, 19*(3), 635–642.
Şen, Z. (2015). *Drought modeling, prediction and mitigation* (p. 396). : Elsevier.
Weiner, N. (1949). *Extrapolation, interpolation and smoothing of stationary time series*. New York: Wiley.
Wilhite, D. A. (1993). The enigma of drought. In D. A. Wilhite (Ed.), *Drought assessment, management and planning: Theory and case studies*. Boston: Kluwer Academic Publishers.
Zadeh, L. A. (1965). Fuzzy Sets. *Information Control, 8,* 338–353.

Statistical Trend Tests

3

Abstract

There are various classically established trend identification tests in the literature and their preliminary explanations are useful for further and innovative trend proposal understandings. In general these methodologies are divided into two groups as parametric and non-parametric approaches. Each group is explained with its proper assumptions, restrictions and mathematical formulations so as to give the reader appreciation of the fundamental concepts, which are useful in the assessment of any trend identification procedure. The regression analysis, which is the first main methodology for the description of the mathematical expression of any trend, is presented with a set of restrictive assumption exposition that are not taken into consideration in many publications throughout the world. It is recommended that in the application of any methodology the researcher should be aware of the assumptions, restrictions and difficulties that may be confronted in the trend identification application researches.

Keywords

Assumptions · Non-parametric · Parametric · Regression · Restrictions · Statistics · Tests · Trend

3.1 General

Statistical methodologies are effective tools for digital data evaluation, deduction of central, deviational, skewness, kurtosis, and many representative parameters from a set of measurements. They are very effective in uncertainty cases that cannot be evaluated and interpreted by any other method. Statistics help to collect numerical data for organization, classification, representation, and population characteristic

Z. Şen, *Innovative Trend Methodologies in Science and Engineering*,
DOI 10.1007/978-3-319-52338-5_3

descriptions from a given sample of finite length. It deals with facts of a state in time, space, and spatiotemporal domains. In any discipline, quantification and summarization of available data in meaningful manner are possible through the application of the statistical methods. One can compare data groups' distribution properties with others and deduction of internal parametric quantities, hypothesis testing by considering suitable probability distribution functions (pdf), relevant interpretations in accordance with the probability statements and mathematical calculations to arrive at the necessary statistical inferences.

Statistics can be defined also as the science dealing with collection, classification, and interpretation of numerical data. It can arrive at quantitative results through the use of mathematical probability theories, which explore order and regularity on aggregates of more or less disparate elements. In this context, statistics is a branch of mathematics to deal with uncertainties and their long-term or large sample behaviors for meaningful deduction to make estimations and predictions of the uncertain phenomenon future behaviors. It also provides opportunities for preliminary scientific deductions from small samples.

Statistics provide a set of brief summary parameters for a given data in descriptive manner and these parameters can then be regarded as the population characteristics of the phenomenon concerned. Among such parameters are the arithmetic average, mode, and median that provide information about the general level of the data; the standard deviation, which represents the average deviation range around the arithmetic average and the skewness coefficient that shows possible imbalance between the big and small deviations around the mean value. All these statistics are very useful in the empirical and analytical studies for historical behavior and future predictions of the underlying generation mechanism of the phenomenon concerned.

Apart from the descriptive parameters, statistics provide opportunity to search for various components that build up a time series. The most important systematic components are the seasonality (periodicity), trend, and shift (step) embedded within the uncertainty (nonsystematic) component of stochasticity (Chap. 2). In any scientific and technological study, where uncertainty ingredient exists even at small extends, the statistical techniques are sought to make firm and valid decisions. For this purpose, the statistics literature gives great importance to confidence limit constructions, significance tests and regression approaches.

In this chapter, each one of aforementioned components will be explained through the classical statistics methodologies, but in the rest of this book trend identification, assessment, interpretation, and applications will be mentioned by means of various scientific and technological disciplines.

3.2 Nonparametric Tests

Normally these lack the power of parametric tests, because they do not consider the population distribution governing the process concerned (Chap. 2). They are based on the order statistics, i.e., ordering of the data from the smallest to the largest or

vice versa and the calculations are based only on the ranks. There is always a loss of information in the transformation from real-data values to ranks. The loss of information means that the result is conservative, which implies that rejection of a null hypothesis is less likely, so parametric tests should be preferred for better results. In general, the nonparametric tests are preferred because of their robustness. These order values are then attached with exceedance probabilities. In the nonparametric methods, estimates of rejection probabilities for null hypotheses are not attempted or if so they are highly uncertain. Nonparametric tests must be used in the following two cases.

(1) If the data are available in the ordinal scale then the use of nonparametric tests is a must, because the mean and standard deviation cannot be calculated. This is due to the fact that only the relative position of the data is meaningful and arithmetic operations are not applicable,
(2) If the frequency distribution of random event shows marked departures from the normal pdf, especially for small sample sizes. However, for large sample sizes, the central limit theorem allows the use of parametric methods on quite markedly non-normal pdf's.

On the other hand, nonparametric tests are valid regardless of sample size and type of pdf. Parametric statistics should be used with more than 30 sample sizes provided that tests such as normal scores or Kolmogorov–Smirnov (KS) lead to acceptance of normal pdf as null hypothesis. It is important to remember that all means of transformations should be applied to the data at hand in an attempt to normalize the data before parametric methods are abandoned.

Most often the nonparametric tests are based on the median value, since the calculation of the mean value does not have any relevance in nonparametric calculations. The median value is found without involved calculations, since it represents the mid-point (50–50%) position within the data sequence. It is possible to find nonparametric alternatives for the standard parametric tests. Nonparametric tests often have a null hypothesis as population medians equality.

3.2.1 Data Ordering (Ranks)

It is a simple procedure to obtain the position of a data value in the ordered sequence from highest (lowest) to lowest (highest) value. In Table 3.1, there are 10 data values, Y_i ($i = 1, 2, \ldots, n$), where n is the sample length. The corresponding ranks, r_i ($i = 1, 2, \ldots, n$) constitute another sequence ready for use in the nonparametric methods.

In this ordering, the lowest value has the smallest rank as 1. Nonparametric test results do not depend on the ordering from the smallest (biggest) to greatest (smallest) value. For any value greater than the greatest data value, the rank of X_8 should remain as $r_8 = 10$. These indicate the irrelevance of the absolute scale and the loss of information, if there are interval or ratio data types.

Table 3.1 Data and ranks

Data		Rank, r_i
Sequence, i	Value, Y_i	
1	2.3	3
2	0.5	1
3	5.0	7
4	6.2	8
5	3.1	4
6	4.9	6
7	8.2	9
8	9.7	10
9	2.1	2
10	3.4	5

3.3 Statistical Tests

There are various methods to explore the internal structure of a given time series by parametric and nonparametric methodologies. In this section, nonparametric tests are explained.

3.3.1 Wald–Wolfowitz

In this test, the run lengths are not taken into consideration and its application requires extra care. This test is neither powerful nor efficient, but can be used to determine whether observations of a random variable are independent, and consequently, in such a time series there is no trend component. The sum of squared lengths test is a more powerful procedure (Himmelblau 1969).

The adjacency test requires that the observations are identically and independently distributed under similar conditions (Kanji 2001). In case of large sample sizes, the difference sign tests are applicable under the similar conditions as in the adjacency test. The run and successive differences tests can be applied when the observations in time series appear under similar conditions. Another distribution free nonparametric test is the Mann–Whitney test (Sect. 3.3.5), which is applicable only when the observations are random and independent. The most popular trend test is the Kendall's rank correlation test, which is employed together with Sen (1968) slope and Mann (1945) sign methodologies.

3.3.2 Sign Test

This is the simplest nonparametric test analogue of the conventional t-test in the parametric statistics. The objective is to decide whether the data are drawn from a population with a specified median. The median, m_0, is defined as the point exactly

at the half-way location through the rank order; if the sample size is odd, it is the value at the midpoint position, i.e., at position $(n + 1)/2$. However, if the sample size is even then the median is taken as the arithmetic average of values at ranks $n/2$ and $(n + 1)/2$. The theoretical basis of nonparametric tests is the probability theory. In general, if a sample is drawn from a population with specified hypothetical median then each item in the data has a priori probability equal to 0.50 as being greater or smaller than the median. Knowing the hypothetical median, m_0 each data value can be attributed with a sign, + for $Y_i > m_0$ or − for $Y_i < m_0$, respectively. Logically, the far away the sample median is from the hypothetical median, the greater is the imbalance between the numbers of +'s and −'s, and consequently, less likely that the sample comes from that population. Sign test is performed according to the following steps for a given sequence of data.

(1) Suppose as zero hypothesis, H_0, that the population median is equal to the hypothetical median,
(2) Alternative hypothesis, H_a is that the population median is not equal to the hypothetical median,
(3) For independent and identically distributed observations, the hypothesis H_0 states that the median m_0 can be tested with the statistic, T, which is the number of observations greater than m_0. Each observation has probability 0.5 of being greater than m_0. If H_0 is true then T follows a Binomial pdf law with $p = 0.5$, otherwise, p will have some other value. If T is too different from $n/2$, as measured by the Binomial probabilities with $p = 0.5$, then H_0 is rejected, (Connover 1971).

It is also possible to perform this test by considering pairs of observation sequences. A paired-comparison of two-sample test can be investigated also by the sign test. Let pairs of samples be denoted by $X_1, X_2, X_3, \ldots, X_n$ and $Y_1, Y_2, Y_3, \ldots, Y_n$. The differences, $Z_i = X_i - Y_i$, are formed and if X_i and Y_i are interchangeable then the median of the differenced series is expected to be equal to zero. Interchangeability means that the two series originate from the same population. In order to confirm this point, a sign test can be performed on the set of differences with the null hypothesis, H_0.

3.3.3 Sign Difference Test

In a given time series, $Y_1, Y_2, \ldots, Y_n$, the number of, say, positives can be found after considering the difference of any term from its subsequent terms starting from the first term in the series. If the sign is indicated by S_i, then the new series will have $(n - 1)$ number of 1 or 0 terms in sequence defined as,

$$S_i = \begin{cases} \text{If} & X_{i+1} - X_i > 1 \quad (i = 1, 2, \ldots, n) \\ \text{Otherwise} & 0 \end{cases}$$

The summation of + signs, S_+ can be obtained as,

$$S_+ = \sum_{i=1}^{n-1} S_i \tag{3.1}$$

If there is not a trend in the time series then the number of +'s is equal to the number of –'s, otherwise, there is a trend. The theoretical average and variance values are given by,

$$\overline{T}_+ = \frac{(n-1)}{2} \tag{3.2}$$

and

$$V_{T_+} = \frac{n+1}{12}, \tag{3.3}$$

respectively. This variable is distributed according to the normal (Gaussian) pdf with these parameters, and therefore, the trend test can be achieved with normal pdf test.

3.3.4 Run Test

Similar to the previous one, it is possible to test and classify the data by considering the standard time series elements as +1 for $Y_i > 0$ and −1 when $Y_i < 0$, which leads to a two-valued sequence. Run test is based on the number of uninterrupted sequence of +1's and −1's. If the number of +1 runs is shown by P, its theoretical average value is defined as,

$$\overline{P} = \frac{2n_1 n_2}{n} + 1 \tag{3.4}$$

and the standard deviation as,

$$S_p = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - 1)}{n^2 (n-1)}} \tag{3.5}$$

With these two parameters the distribution of P complies by a Gaussian pdf. If n_1 and n_2 are the numbers of +1 and −1, respectively, in the whole two-valued sequence then $n_1 + n_2 = n$, where n is the number of data. Sneyer (1992) has recommended the use of Wald–Wolfowitz serial dependence test, which together

with the following Spearman order serial dependence procedure during the trend test (Chap. 2). The reason for this is the usual existence of a jump in some topics concerning the change in the data sequence. The trend analysis in a sequence can be achieved by the following two nonparametric tests.

3.3.5 Mann–Whitney (MW) Test

This is an analogous test to the two-sample *t*-test in the parametric statistics literature. It tests the null hypothesis of population medians equality from which two samples might be drawn. This test depends on the ranking of two data set mixtures. The basic idea is that if the two samples are from the same pdf with the same median value then the summation of ranks for each data set is expected to be equal. The following steps are necessary for successful application of this nonparametric test to two data sets.

(1) Consider as the null hypothesis H_0 that the median of data population X is equal to the median of population Y,
(2) As the alternative hypothesis H_a should signify that the median of data population X is not equal to the median of population Y,
(3) Calculate the test statistic, T, by considering ranks, r_i and the number of samples, n, in each data set as,

$$T = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2} \tag{3.6}$$

The first term on the right-hand side corresponds to the summation of ranks attributed to one data set of sample size n. Theoretically, the second term is the summation of ranks. The closer these two terms to each other, the median of X population is equal to the median of population Y. In case of two homogeneous time series and $n_1 = n_2$, then summation of ranks in each time series is similar to each other. However, when the data numbers are different, the homogeneity is measured by the closeness of r_1/n_1 and r_2/n_2. There are $n!/n_1!n_2!$ different time series. For $n_1 = n_2 = 10$, there are 184,756 different possibilities of two time series. The necessary test quantity is given as,

$$U_1 = M_1 - \frac{n_1}{2}(n_1 + 1) \tag{3.7}$$

or

$$U_2 = M_2 - \frac{n_2}{2}(n_2 + 1) \tag{3.8}$$

For each time series there is different test quantity. Theoretic distribution of this quantity shows that the mean and the standard deviation are,

$$\overline{U} = \frac{n_1 n_2}{2} \tag{3.9}$$

and

$$S_u = \left[\frac{n_1 n_2 (n_1 + n_2 - 1)}{12}\right], \tag{3.10}$$

respectively. The test quantity is distributed according to a normal pdf with these parameters, and hence, the MW test is similar to a normal distribution test.

(4) The sample sizes n_1 and n_2 of data sets are required for comparison of calculated T value with critical values from Table 3.2.

One of the main questions is, if the two time series are serially independent, is it possible that there may still be theoretically some difference between these time series? In order to answer to this question, the two time series are assumed to come from the same population, and hence, there is not significant difference between them. Such an assumption provides opportunity to exchange values between them. They may be mixed and the total sample length, n, is equal to the summation of the first, n_1, and the second time series sample length, n_2.

Example 3.1 In the following table the number of lightings is given during a rainfall seeding and non-seeding periods. It is thought that during the seeding there is expectancy of lighting reduction. For this purpose, in a random manner, the same cloud group is seeded and the number of lightings is measured. During $n_1 = 12$ seeding operation on the average 19.25, and during $n_2 = 11$ normal case on the average there were 69.45 lightings. If Table 3.3 is examined, it is observed that during the non-seeding period the lighting numbers do not appear as normally (Gaussian) distributed. At least existence of a high number as 358 indicates this fact.

Solution 3.1 In the application of MW test, the necessary calculations for the rank quantity are presented in Table 3.4. From Eq. (3.7) MW test quantity is $U_1 = 108.5 - 6(12 - 1) = 30.5$. Arithmetic average and standard deviation values are calculated from Eqs. (3.9) and (3.10) as $\overline{U} = 12 \times 11/2 = 66$ and $S_u = [12 \times 11 (12 + 11 - 1)/2]^{1/2} = 16.2$. Hence, the standardized U_1 value correspondence in a standard Gaussian pdf is $z = (30.5 - 60)/16.2 = -2.19$. On the other hand, in a theoretical standard normal pdf, the critical standard value at 5% significant level is $z_{cr} = 1.96$. Since, z value is greater than the critical level, it is understood that there is a significant reduction in the number of lighting after the seeding. Hence, heterogeneity is concluded.

Table 3.2 Mann–Whitney table

		n_2																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Alpha = 0.5 (two-twiled)																				
n_1	2							0	0	0	1	1	1	1	1	1	2	2	2	2
	3				0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	4			0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	5		0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	6		1	2	3	5	6	7	10	11	13	14	16	17	19	21	22	24	25	27
	7		1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	8	0	2	4	6	7	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	9	0	2	4	7	10	12	15	17	20	23	26	28	31″	34	37	39	42	45	49
	10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
	14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
	15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
	16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
	17	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
	18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
	19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
	20	2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

(continued)

Table 3.2 (continued)

		n_2																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Alpha = 0.10 (two-tailed)																				
n_1	2																			
	3		0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
	4		0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
	5		1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
	6		2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
	7		2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
	8		3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
	9		4	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
	10		4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
	11		5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
	12		5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
	13		6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
	14		7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
	15		7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
	16		8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
	17		9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
	18		9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
	19		10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
	20		11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138

(continued)

Table 3.2 (continued)

		n_2																		
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Alpha = 0.20 (two-tailed)																				
n_1	2		0	0	1	1	1	2	2	3	3	4	4	5	5	5	6	6	7	7
	3	0	1	1	2	3	4	5	5	6	7	8	9	10	10	11	12	13	14	15
	4	0	1	3	4	5	6	7	9	10	11	12	13	15	17	18	20	21	22	23
	5	1	2	4	5	7	8	10	12	13	15	17	18	20	22	23	25	27	28	30
	6	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31	34	36	38
	7	1	4	6	8	11	13	16	18	21	23	26	28	31	33	36	41	43	46	48
	8	2	5	7	10	13	16	19	22	24	27	30	33	36	39	42	45	48	51	54
	9	2	5	9	12	15	18	22	25	28	31	35	38	41	45	48	52	55	58	62
	10	3	6	10	13	17	21	24	28	32	36	39	43	47	51	54	58	62	66	70
	11	3	7	11	15	19	23	27	31	36	40	44	48	52	57	61	65	69	73	78
	12	4	8	12	17	21	26	30	35	39	44	49	53	58	63	67	72	77	81	86
	13	4	9	13	18	23	28	33	38	43	48	53	58	63	68	74	79	84	89	94
	14	5	10	15	20	25	31	36	41	47	52	58	63	69	74	80	85	91	97	102
	15	5	10	17	22	27	33	39	45	51	57	63	68	74	80	86	92	98	104	110
	16	5	11	18	23	29	36	42	48	54	61	67	74	80	86	93	0	106	112	119
	17	6	12	20	25	31	41	45	52	58	65	72	79	85	92	99	106	113	120	127
	18	6	13	21	27	34	43	48	55	62	69	77	84	91	98	106	113	120	128	135
	19	7	14	22	28	36	46	51	58	66	73	81	89	97	104	112	120	128	135	143
	20	7	15	23	30	38	48	54	62	70	78	86	94	102	110	119	127	135	143	151

Table 3.3 Lighting numbers

Seeding numbers	No-seeding data
49	61
4	33
18	62
26	45
29	0
9	30
16	82
12	10
2	20
22	358
10	63
34	

Table 3.4 Joint data

Lighting number	Is there seeding?	$R(SDS_1) + R(NS)$	Separate ranks	
			(SD)	(NS)
0	N	1		1
2	Y	1	2	
4	Y	2	2	
9	Y	3	3	
10	N	4	4	
10	Y	5.5		5.5
12	Y	5.5	5.5	
16	Y	7	7	
18	Y	8	8	
20	N	9	9	
22	Y	10		10
26	Y	11	11	
29	Y	12	12	
30	N	13	13	
33	N	14		14
34	Y	15		15
45	N	16	16	
49	Y	17		17
61	N	18	18	
62	N	19		19
63	N	20		20
82	N	21		21
358	N	22		22
Total rank			108.5	167.5

N No-seeding; *Y* Seeding

3.3.6 Kruskal–Wallis (KW) Test

This is simply an extension of the previous test and conventional *t*-test to the cases where there are multiple sample situations, say k data sets. It is therefore analogous to one-way analysis of variance. The basic hypothesis is that all the sample sets are drawn from populations with the same median value. The test is successful if k median values are significantly close to each other. Otherwise, the failure of this test is confirmed even if any one of the k median is different than the others. The following steps are necessary for the performance of KW test.

(1) The null hypothesis, H_0, is that the median of all source populations are equal,
(2) The alternative hypothesis H_a says that the median of at least one source population is different,
(3) The test statistic, T, is calculated from given k data sets as,

$$T = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{n_j} r_{ij}\right)^2}{n_j} - 3(N+1), \tag{3.11}$$

where there are k samples each with sizes n_1, n_2, n_3, …, n_k, making a total sample size of N. In Eq. (3.11) r_{ij} is the rank of the ith data point in the jth sample,

(4) T is distributed very similarly to chi-square pdf with $k - 1$ degrees of freedom. The sample value T should be compared with critical values from Table 3.5.

It is assumed that k time series originate from the same population. It is possible that in each time series the number of data is different from each other ($n_1 \neq n_2 \neq$, …, $\neq n_k$). In the application of this method the following steps are necessary.

(1) By mixing all the time series, a single time series is obtained that is composed of $n = n_1 + n_2 + \cdots + n_k$ data,
(2) These data are ordered from the smallest to the greatest,
(3) The ordered data are attached with ranks starting from 1, r_j (j = 1, 2, …, n). If there are equal data values in the sequence they are given arithmetic average of the ranks (Table 3.6).
(4) Find the ranks that correspond to each time series data. Hence, corresponding to n_k time series total ranks r_j (j = 1, 2, …, k) are calculated.

$$M_j = \sum_{i=1}^{n_k} M_{ij} \tag{3.12}$$

Table 3.5 Critical T values

df	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	–	–	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.02	0.051	0.103	0.211	4.605	5.991	7.378	9.21	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.86
5	0.412	0.554	0.831	1.145	1.61	9.236	11.07	12.833	15.086	16.75
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.69	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.18	2.733	3.49	13.362	15.507	17.535	20.09	21.955
9	1.735	2.088	2.7	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.94	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.92	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.3
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.66	5.629	6.571	7.79	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.39	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.26	9.591	10.851	12.443	28.412	31.41	34.17	37.566	39.997
21	8.034	8.897	10.283	11.591	13.24	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.26	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.98	45.559

(continued)

Table 3.5 (continued)

df	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
25	10.52	11.524	13.12	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.16	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.29
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.42	76.154	79.49
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.54	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Table 3.6 Wilcoxon signed rank method application

Time series data		Differences		Signed ranks	
X	Y	d_i	Order $\lvert d_i \rvert$	$d_i < 0$	$d_i > 0$
53	70	−17	20		20
54	66	−12	17.5		17.5
48	82	−34	21		21
46	58	−12	17.7		17.5
67	78	−11	16		16
75	78	−3	4.5		4.5
66	76	−10	14.5		14.5
76	70	6	9	9	
63	73	−10	14.5		14.5
67	59	8	11.5	11.5	
75	77	−2	2		2
62	65	−3	4.5		4.5
92	86	6	9	9	
78	81	−3	4.5		4.5
92	96	−4	7		7
74	73	1	1	1	
91	97	−6	9		9
88	75	13	19	19	
100	92	8	11.5	11.5	
99	96	3	4.5	4.5	
107	98	9	13	13	
Rank summations				78.5	152.5

Basic test hypothesis is that all k time series have the same behavior, i.e., they are homogeneous. The alternative hypothesis is that at least one of the k time series has the average difference behavior. Theoretic studies led to KW test quantity as,

$$K_W = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{[M_k - n_k(n+1)/2]^{1/2}}{n_k} \tag{3.13}$$

or simply as,

$$K_W = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{M_K^2}{n_k} - 3(n+1) \tag{3.14}$$

K_W has approximately $(k - 1)$ degrees of freedom and it complies by a chi-square pdf and chi-square test is applied for the final decision.

3.3.7 Nonparametric Correlation Coefficient

In parametric statistics the correlation coefficient is named as the Pearson correlation and it is defined as the product moment (Chap. 2). In the nonparametric statistics domain the analogous to the Pearson correlation is named as the Spearman's rank correlation coefficient. Pearson correlation coefficient requires that both variables should comply by the normal pdf, which is not the case in many natural, social, environmental, economic sciences. In calculation of this nonparametric correlation coefficient both data sets are ordered separately from each other. Hence, there are two rank sequences, one for X_i ($i = 1, 2, \ldots, n$) variable $R(X_i)$ and other for Y_i variable as $R(Y_i)$. If for each i the ranks of X equal to ranks of Y, then the Spearman's rank correlation is regarded as perfect. The rank correlation is defined as the sum of the difference between the corresponding ranks of X and Y. Analogously to the parametric version values of the coefficient are scaled between -1 (perfect negative correlation) and $+1$ (perfect positive correlation). In between there is a value, which is equal to zero indicating no correlation. Spearman's rank correlation calculation steps are as follows.

(1) As the null hypothesis, H_0, the correlation between X and Y time series is assumed equal to zero. This is referred to as the hypothetical correlation value, $r_s = 0$,
(2) Alternative hypothesis, H_a, is that this correlation coefficient is different than zero,
(3) The test statistic, r_S, which is referred to as the Spearman's rank correlation coefficient is defined in terms of each data set ranks and the sample number, n, of data in each set as,

$$r_S = 1 - \frac{6\sum_{i=1}^{n}[R(X_i) - R(Y_i)]}{n(n^2 - 1)} \tag{3.15}$$

(4) The calculated r_S is compared with critical values from Table 3.7.

For example, if there are 20 pairs of data with a value of 0.53 then there would be a probability between 0.01 and 0.005 that it had occurred by chance. In other words, one might expect to get this result by chance once every 100–200 times, and therefore, it indicates a very significant correlation between the two sets of data.

As with the other nonparametric methods, values of X and Y can vary extensively without affecting the final result. It is necessary to keep in mind that r_S do not imply good linear relationship. It is quite possible to obtain low Spearman's rank correlation coefficient for high Pearson's parametric correlation coefficient. However, in many applications, it is unusual for Pearson's coefficient to provide a statistical test result markedly superior to Spearman's rank correlation approach even with normally distributed data.

Table 3.7 Critical values for Spearman's rank correlation coefficient

n Number of pairs	Probability of record occurrence				
	0.1	0.05	0.025	0.01	0.005
4	1	1	1	1	1
5	0.7	0.9	0.9	1	1
6	0.6571	0.7711	0.8286	0.9429	0.9429
7	0.5714	0.6786	0.7857	0.8571	0.8929
8	0.5476	0.6529	0.7381	0.8095	0.8571
9	0.4833	0.6	0.6833	0.7667	0.8167
10	0.4424	0.5636	0.6485	0.7333	0.7818
11	0.4182	0.5273	0.6091	0.7	0.7545
12	0.3986	0.5035	0.5874	0.6713	0.2773
13	0.3791	0.478	0.5604	0.6484	0.6978
14	0.367	0.4593	0.5385	0.622	0.6747
15	0.35	0.4429	0.5179	0.6	0.6536
16	0.3382	0.4265	0.5029	0.5824	0.6324
17	0.3271	0.4124	0.4821	0.5577	0.6055
18	0.317	0.4	0.4683	0.5425	0.5897
19	0.3077	0.3887	0.4555	0.5285	0.5751
20	0.2992	0.3783	0.4438	0.5155	0.5614
21	0.2914	0.3687	0.4329	0.5034	0.5487
22	0.2841	0.3598	0.4227	0.4921	0.5368
23	0.2774	0.3515	0.4132	0.4815	0.5256
24	0.2711	0.3438	0.4044	0.4716	0.5151
25	0.2653	0.3365	0.3961	0.4622	0.5052
26	0.2698	0.3297	0.3882	0.4534	0.4958
27	0.2546	0.3233	0.3809	0.4451	0.4869
28	0.2497	0.3172	0.3739	0.4372	0.4785
29	0.2452	0.3115	0.3673	0.4297	0.4705
30	0.2407	0.3061	0.361	0.4226	0.4629

3.3.8 Spearman's Rho Test of Trend

The Spearman's rho trend test is based on the Spearman's rho statistic, which is the standard Pearson correlation coefficient between the rank of the annual summary statistics and the year (Kolaz and Swinford 1988, 1989; Sweitzer and Kolaz 1984; Lettenmaier 1976; EPA 1974). No trend and independent structure cases imply that all ranks are equally likely, which is used to test the statistical significance of the Spearman's rho statistic. If the value is significantly different from zero then it implies a significant trend. When ties in the annual summary statistics are present, then the significance level has to be adjusted in order to account for the number of ties. The linear regression power calculations are based on formulae that are incorrect for small samples but approximately correct for large ones (Lettenmaier 1976).

Consider Y and X as the ranks of the time (e.g., year) and data series (say, rainfall), respectively; the Spearman (1940) rho statistic, r_{sc}, for data with no tied ranks is given as,

$$r_{sc} = 1 - \frac{6\sum_{i=1}^{n}(Y_i - X_i)^2}{n(n^2 - 1)} \tag{3.16}$$

In some studies (Khaliq et al. 2009; Sonali and Kumar 2013), it was mentioned that this expression can be used even when there are ties in the data except that the convention is to take X as the average rank. If the ties are of significant extent then more suitable approach can be used for quantifying the extent of the statistical dependence in the data.

3.3.9 Turning Point Test

One of the easiest methods to apply is the turning point test, which can be visualized from the graphical representation of a time series, because it is possible to see the systematic and random variations. As stated by Shahin et al. (1993) Kendal phase test is more valid in case of points that tend to bunch together. Here, the difficulty is that a comparison of observed and theoretical numbers of phases by the usual chi-square test is invalidated due to the fact that the lengths of phases are not independent. Also, the distribution of phase lengths does not tend to be normal for large sample sizes, but the number of phases follows a normal pdf (Kendall 1973).

The convenience of trend tests to a time series structure is concerned mainly with the adaptability of a chosen test. The turning points and number of phase tests are practically outdated due to the availability of much more powerful tests (Shahin et al. 1993).

In case of a time series with trend component there will be more turning points (peaks and valleys) expectations than trend free time series. A turning point is defined at least by three subsequent time series terms. If a time series has n observations as Y_1, Y_2, Y_3, …, Y_n, then the consideration of this sequence as three successive terms from its beginning until the end indicates turning points as,

$$T_i = \begin{cases} \text{If} & Y_i < Y_{i+1} > Y_{i+2} \quad 1 \\ \text{Otherwise} & \qquad\qquad\qquad\quad 0 \end{cases} \tag{3.17}$$

The result has $(n - 2)$ terms with 1's and 0's in a randomly alternation manner. The total number of turning points, N_T, is given as,

$$N_T = \sum_{i=1}^{n-2} T_i \tag{3.18}$$

Theoretical studies by some researchers indicate that the turning point number has a normal (Gaussian) pdf with average and variance as, (Kendall and Stuart 1973),

$$\overline{N}_T = \frac{2}{3}(n-2) \tag{3.19}$$

and

$$V_{N_T} = \frac{16n-29}{90}, \tag{3.20}$$

respectively. Finally, the existence of a trend may be checked by a normal pdf test.

3.3.10 Mann–Kendall (MK) Test

Irrespective of the linearity or curvature of the trend, its identification is possible by nonparametric MK test. However, the serial correlation structure in a dependent time series is bound to affect the ability of the MK test (Yue et al. 2004). The assumptions of the classical parametric tests viz., normality, linearity, and independence are usually not met by many natural and artificial time series. Additionally, the questions of missing values, censored data, flow relatedness, and seasonality hinder the normal (Gaussian) pdf and MK test analysis depends on the sign difference between all combinations of earlier and later data measurements. This provides the possibility of $n(n-1)/2$ different types of differences each with sign 1, 0 or −1. In this sign procedure, there is no need for data pdf, since it is independent of any specific pdf. This test assumes that a value can always be declared less than, greater than, or equal to another value; that data are independent; and that the pdf of data remains constant either in the original or transformed units (Helsel and Hirsch 1992). The test statistics are invariant to transformations such as logs (i.e., the test statistics will be the same value for both raw and log-transformed data), and hence, the Mann–Kendall test is applicable in many situations.

The performance of a MK test requires computation of the difference between the successive measurements $(j-i)$ apart, where $j > i$, and according to the sign of the difference an integer value is attached to positive differences as 1, no differences as 0, and negative differences as −1. The test statistic, S, is then computed as the sum of the integers.

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sign}\left(Y_j - Y_i\right), \tag{3.21}$$

where Y_j and Y_i are the sequential data values in a sample of size n, and

$$\operatorname{sign}(Y_j - Y_i) = \begin{cases} 1 & \text{if} \quad (Y_j - Y_i) > 0 \\ 0 & \text{if} \quad (Y_j - Y_i) = 0 \\ -1 & \text{if} \quad (Y_j - Y_i) < 0 \end{cases} \tag{3.22}$$

For $n \geq 10$, Mann (1945) and Kendall (1975) have documented that S has approximately normal pdf with the mean $E(S) = 0$ and variance $V(S)$ given as,

$$V(S) = \frac{1}{18} n(n-1)(2n+5) \tag{3.23}$$

In case of tied ranks, $V(S)$ becomes,

$$V(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{k=1}^{m} t_k(t_k-1)(2t_k+5) \right], \tag{3.24}$$

where m is the number of tied groups, and t_k is the number of observations in the kth group. The standardized MK test statistic Z_{MK}, which follows the standard normal pdf with mean zero and unit variance is defined as,

$$Z_{\mathrm{MK}} = \begin{cases} \dfrac{S-1}{\sqrt{V(S)}} & \text{for} \quad S > 0 \\ 0 & \text{for} \quad S = 0 \\ \dfrac{S+1}{\sqrt{V(S)}} & \text{for} \quad S < 0 \end{cases} \tag{3.25}$$

A positive (negative) value of S indicates an upward (downward) trend. The trend is considered insignificant if Z_{MK} is less than the standard normal variate $Z_{\alpha/2}$, at $\alpha\%$ significance level. The trend is significant if $Z_{\mathrm{MK}} \geq Z_{\alpha/2}$.

To deal with the influence of autocorrelation on $V(S)$, Yue and Wang (2004) suggested a modification at the variance based on the effective sample size, n^*, for variance, $V^*(S)$ as,

$$V^*(S) = V(S)\frac{n}{n^*} \tag{3.26}$$

and

$$n^* = \frac{n}{1 + 2\sum_{i=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k}, \tag{3.27}$$

where ρ_k is the lag-k serial correlation coefficient of a given series X_i, which can be computed as,

$$\rho_k = \frac{\frac{1}{(n-k)}\sum_{i=1}^{n-k}\left[Y_i - E(Y_i)\right]\left[Y_{i+k} - E(Y_i)\right]}{\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - E(Y_i)\right]^2} \tag{3.28}$$

If S is a large positive (negative) number, later-measured values tend to be larger (smaller) than earlier values indicating an upward (downward) trend. Small absolute S values imply no-trend. The test statistic, τ, is,

$$\tau = \frac{2S}{n(n-1)} \tag{3.29}$$

This varies within the range from −1 to +1 and it is analogous to the correlation coefficient in the regression analysis. The null hypothesis, H_o, of no trend is rejected in cases when S and τ are significantly different from zero.

The slope of a significant trend can be calculated according to the formulation given by Sen (1968).

$$\beta = \text{median}\left(\frac{Y_j - Y_i}{j - i}\right) \tag{3.30}$$

For all $i < j$ where $1 < i\ n - 1$ and $2 < j < n$. One of the widely used non-parametric tests for detecting trends in the time series is the MK test (Mann 1945; Kendall 1955). This trend test is derived from a rank correlation test for two groups of observations. In the procedure, the correlation is considered between the rank order of the observed values and their orders. The null hypothesis, H_0, for this test is that the data are independent and randomly ordered, i.e., there is no trend or serial correlation structure among the observations.

The MK test statistic, S, for two sets of observations X_1, X_2, X_3, …, X_n and Y_1, Y_2, Y_3, …, Y_n is formulated as,

$$S = \sum_{i \le j}^{n} a_{ij} b_{ij}, \tag{3.31}$$

where

$$a_{ij} = \text{sgn}\left(X_j - X_i\right) = \begin{cases} 1 & X_i \le X_j \\ 0 & X_i = X_j \\ -1 & X_i \ge X_j \end{cases} \tag{3.32}$$

and b_{ij} is similarly defined for the observations in Y. Under the hypothesis that X and Y are independent and randomly ordered, the statistic S tends to normal pdf for large n, with mean and variance given by,

$$E(S) = 0 \tag{3.33}$$

and

$$\mathrm{Var}(S) = \frac{n(n-1)\ (2n+5)}{18}, \tag{3.34}$$

respectively. If the values are replaced with the time order of the time series X_i, (i = 1, 2, …, n) then it can be used as a trend test (Mann 1945). In this case, the statistic S reduces to,

$$S = \sum_{i \le j} a_{ij} = \sum_{i \le j} \mathrm{sgn}\left(X_j - X_i\right) \tag{3.35}$$

with the same mean and variance as before. Kendall (1955) gave a proof of the asymptotic normal pdf of the statistic, S. The significance of trends is tested by comparing the standardized test statistic, Z, as,

$$Z = \frac{S}{[\mathrm{Var}(S)]^{0.5}} \tag{3.36}$$

with the standard normal pdf at the desired significance level.

3.3.10.1 Mann–Kendall Trend Search

This is one of the most frequently used methodologies in trend search (Mann 1945; Kendall 1955). This trend test is derived from a rank correlation test. For two groups of observations trend is derived from a rank correlation test and the correlation between the rank order of the observed values and their order in time series. The null hypothesis, H_0, is that the data are independent and have random pdf, which implies no trend and no serial correlation. In the application of MK method, the following steps are valid.

(1) Starting from the left side of a time series, each value is compared with the ones on the left and the number of values greater than the considered value is counted and written. If these numbers are shown by n_i then the new series will have these values,
(2) The successive summation of these numbers, t_i, is considered for the statistical test as,

$$t_i = \sum n_i \tag{3.37}$$

(3) Theoretical studies indicated that this statistic has also normal pdf with the following arithmetic average and standard deviation values, (Kendall and Stuart 1952)

$$\bar{t} = \frac{n(n-1)}{4} \tag{3.38}$$

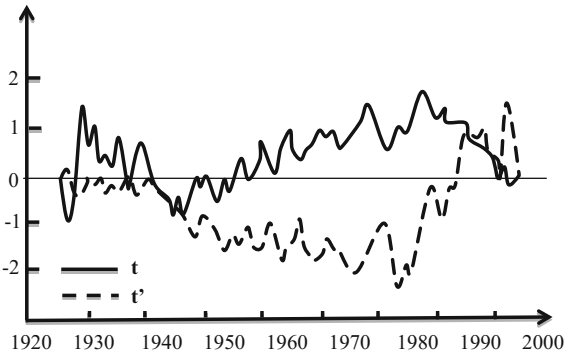


Fig. 3.1 Mann–Kendall trend search

and

$$\sigma_t = \left[\frac{n(n+1)\,(2n+5)}{72}\right]^{1/2} \tag{3.39}$$

(4) Previously found t_i values from Eq. (3.37) are standardized with these theoretical parameters,

$$u_i = \frac{t - \overline{t}}{S_t} \tag{3.40}$$

(5) The same procedure is applied to this time starting from the right side, i.e., beginning of the given time series, which leads to a similar series of $t_i^{'}$,

(6) These two series, t_i and $t_i^{'}$ are shown as time series in Fig. 3.1. At this stage two tailed test is applied. According to MK test, in the case of no trend, these two series intercross each other several times. In the case of trend the intersection of these two time series indicates the trend beginning.

3.3.10.2 Sen Slope Estimator

Sen (1968) suggested a nonparametric alternative for slope estimation, which is based on the calculation of slopes for all the pairs of time series data and then taking the median of these slopes as an estimate of the overall trend slope. This method is insensitive to outliers and can handle a moderate number of values below the detection limit and missing values. If time series sample length is n (or n periods of time), and Y_i is the data value for the ith time instant, there will be $n(n-1)/2$ possible pairs of time points (i, j) in which $i > j$. Hence, a pairwise slope, b_{ij}, is computed as,

$$b_i = \frac{Y_j - Y_k}{j - k} \quad (\text{for } j > k;\ i = 1, 2, \ldots, n), \tag{3.41}$$

where n is the sample size. In case of no trend, a given Y_i has the same chance to be above or below of another Y_j value, and therefore, there is approximately equal number of positive and negative slopes leading to almost zero median value.

In any given time series, there are $N = n(n - 1)/2$ slopes. The Sen slope is equal to the median of all these slopes. A two-tailed significance test can be obtained concerning this slope value by the nonparametric technique based on the normal (Gaussian) pdf. It is possible to calculate the confidence limits as,

$$\mathrm{CL}(\alpha) = \pm Z_{1-\alpha/2}\sqrt{\mathrm{VAR}(S)} \tag{3.42}$$

In this equation $Z_{1-\alpha/2}$ is taken from a standard normal (Gaussian) pdf.

$$b_{ij} = \frac{Y_i - Y_j}{i - j} \tag{3.43}$$

3.3.10.3 Spearman's Tau

In parametric statistics, the Pearson correlation is used and defined as the product moment with the basic assumption that the time series must abide by the normal (Gaussian) pdf. In the nonparametric statistics domain, the Spearman's rank correlation coefficient is used and it does not require normal (Gaussian) pdf. The data sets are ordered for calculating this nonparametric correlation coefficient. Hence, there are two sequences of ranks, one for X variable, $R(X_i)$, and other for Y variable, $R(Y_i)$. If the ranks of X are equal to ranks of Y, then the Spearman's rank correlation is perfect. The rank correlation is defined as the sum of the difference between the corresponding ranks of X and Y. Analogous to the parametric version of the coefficients, it is scaled between −1 (perfect negative correlation) and +1 (perfect positive correlation). In between the value that is equal to zero indicates no correlation. The Spearman test statistic, ρ_S, is defined in terms of each data set ranks and the number of sample, n, as,

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n}[R(X_i) - R(Y_i)]}{n(n^2 - 1)} \tag{3.44}$$

As with the other nonparametric methods, values of X and Y can vary extensively without affecting the final result. It is necessary to keep in mind that ρ_S does not imply good linear relationship rather than linearity.

3.3.10.4 Regression Trend

It is a parametric method and equivalent to the classical regression approach, where the independent variable is the time sequence (t = 1, 2, 3, …, n) with dependent variable of any physical, social, economic, etc. series, Y_i (i = 1, 2, 3, …, n). In general, the linear regression expression has the following form,

$$Y = a + S_r t, \tag{3.45}$$

where a is the intercept at $t = 0$ and S_r is the trend slope. The basic equations for parameter estimations are,

$$\widehat{S}_r = \frac{n\sum_{i=1}^{n} t_i Y_i - \sum_{i=1}^{n} t_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2} \tag{3.46}$$

and

$$\widehat{a} = \frac{\sum_{i=1}^{n} t_i \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} t_i \sum_{i=1}^{n} t_i Y_i}{n\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2} = \overline{y} - \overline{t}\,\widehat{S}_r \tag{3.47}$$

The trend slope estimation is $\widehat{S}_r$ and the trend line passes through the centroid point, which is defined by coordinates ($t/2$, $\overline{X}$). This centroid point is also used in the trend determination by Sen's slope and other methods in the following sequel.

3.3.11 Two-Sample Wilcoxon Test

Consider two independent random samples as $X_1, X_2, X_3, \ldots, X_n$ and $Y_1, Y_2, Y_3, \ldots, Y_n$. Let sample X be drawn from a population with distribution F_X and Y sample be taken from another population with distribution function F_Y.

$$F_X(X + \Delta) = F_Y(X) \tag{3.48}$$

This means that the two populations differ only by a shift, Δ. The null hypothesis is, H_0: $\Delta = 0$ and it is false if Δ has some value different from zero then one sample will tend to have larger values than the other. This can be measured by ranking the combined samples in the order of increasing size and summing the ranks of each sample within the combination. Let the sum of X ranks be r_X and similarly, r_Y for Y's. If r_X and r_Y are too different from each other then H_0 is rejected.

Another set of statistics, which measures the differences in ranking, is the number of inversions, which is also known as the Mann–Whitney statistics that is defined as I_X. It is the number of times X is greater than corresponding Y; and likewise I_Y is the number of times Y is greater than X. These have the theoretical formulations as,

$$I_X = r_X - \frac{m(m+1)}{2} \tag{3.49}$$

and

$$I_Y = r_Y - n\frac{n(n+1)}{2} \tag{3.50}$$

with the common expectation (mean) and variance as,

$$E(I_X) = E(I_Y) = \frac{mn}{2} \tag{3.51}$$

and

$$V(I_X) = V(I_Y) = \frac{mn(m+n+1)}{12}, \tag{3.52}$$

respectively. Finally, consideration of these statistics leads to the rejection of the null hypothesis, H_0, if the minimum of I_X and I_Y is too much smaller than $mn/2$.

3.3.11.1 Signed-Wilcoxon Test

For a given random sequence, Y_1, Y_2, Y_3, …, Y_n, it is necessary to test whether the population is symmetric about m. If so, the random sample of differences $(Z_i = Y_i - m;\ 1 \leq i \leq n)$ will be symmetric about zero. Let the null hypothesis, H_0, state that Y symmetry about m, is true. Negative values of Z_i form one sample of values, and positive Z_i values as a set is another sample. The Wilcoxon rank-sum statistic is computed from these artificially created two samples. If the sum of ranks for the positive values is too different from the sum of ranks of the negative values, H_0 is rejected.

3.3.11.2 Wilcoxon Signed Rank Test

X_i and Y_i (i = 1, 2, …, n) are two time series of sample length, n. The cross pair data difference series is $d_i = X_i - Y_i$. On the assumption that both time series originate from the same population, rationally and logically the number of +'s is almost the same with the −'s, additionally, the magnitudes of differences are close to each other. For this test, first of all the absolute values of these differences are considered and then ordered with the definition of a new D_i sequence as,

$$D_i = \text{mertebe}|F_i| = \text{mertebe}|x_i - y_i| \tag{3.53}$$

Based on the sign of this quantity, the ranks are grouped into + and − groups with the rank summation of each group statistic as,

$$T^+ = \sum_{F_i \geq 0} D_i \tag{3.54}$$

and

$$T^- = \sum_{F_i \leq 0} D_i \tag{3.55}$$

When $D_i = 0$ then the rank is distributed equally between T^+ and T^-; the summation of these two quantities is expected to be equal to the summation of the ranks.

$$T^+ + T^- = 1 + 2 + \cdots + n = \frac{n(n+1)}{2} \tag{3.56}$$

Enumeration possibility indicates that 2^n different time series can be obtained from two time series random mixture and then pairwise random drawing leads to new X_i and Y_i time series. Again application the same procedure to these two time series leads to 2^n values, which have theoretically a normal (Gaussian) pdf with the following average and standard deviation,

$$\overline{T} = \frac{n(n+1)}{4} \tag{3.57}$$

and

$$S_T = \left[\frac{n(n+1)\,(2n+1)}{24}\right]^{1/2}, \tag{3.58}$$

respectively.

Example 3.2 Wilcoxon sign rank method is applied to two time series given in Table 3.5. These data are the observed storm numbers at two different locations during 21 years. Due to the closeness of the two regions, it is expected that these time series are similar to each other.

Solution 3.2 It is possible to obtain $2^{21} = 2097152$ different time series from these data. After the calculations, it is possible to find that $T^+ = 78.5$, $\overline{T} = (21)(22)/4 = 115.5$ and $S_T = [21 \times 22(42 + 1)/24]^{1/2} = 28.77$. The standard value is obtained as $z = (78.5 - 115.5)/28.77 = -1.29$. According to a standard normal pdf test with 5% significance level, the critical value is 1.96, and hence, one can decide that time series come from the same population, and therefore, they are similar to each other.

3.3.12 von Neuman Test

This test is applied by the Neumann ratio, which is defined according to the following expression.

$$N = \frac{\sum_{i=1}^{n-1} (Y_i - Y_{i+1})^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}, \tag{3.59}$$

where Y_i's (i = 1, 2, …, n) represent time series and $\overline{Y}$ stands for the arithmetic average of the same series. If the mean value is constant, under the null hypothesis, the expectation of this ratio is equal to two, $E(N) = 2$. Qwen (1962) has given a table of percentage points of N for normal and independent pdf's. It has been stated that this ratio is closely related to the first-order serial correlation coefficient (WMO 1966). Yevjevich and Jeng (1969) presented a comprehensive study of the effect of changes in the mean on the correlation function.

3.3.13 Cumulative Departures Test

Some insights into the general features of a time series including possible changes and homogeneity can be gained through the graphical methods. Very classical methodology is the double-mass curve which is obtained by plotting the cumulative amounts of the station under consideration against the cumulative amounts of a set of records at neighboring stations (Searcy and Hardison 1960). In case of homogeneity the scatter of points falls along a straight-line (Chap. 6). It is also valid for the cumulative deviations from some average value, which have the advantage that changes in the mean amount are recognized easily (Craddock 1979).

Although such graphs are useful to detect the shifts in the mean value, but it is not obvious usually how real changes can be distinguished from purely random fluctuations. In order to clarify this point objectively, it is necessary to have a significance test. Although there are commonly employed statistical techniques in various disciplines such as in climatology and hydrology as stated by the World Meteorological Organization (WMO 1966), they are not based on some characteristic of the cumulative sums in the graphical analysis.

3.3.13.1 Cumulative Deviations

Homogeneity tests can be based on the cumulative sums of departures, S_k^*, from the record mean value, $\overline{Y}$.

$$S_k^* = \sum_{i=1}^{k} \left(Y_i - \overline{Y}\right), \quad k = 1, 2, \ldots, n \tag{3.60}$$

Herein, $S_n^* = 0$ implies that there is no trend in the time series. In a homogeneous record, the fluctuations of S_k^* appear around zero, since there is no systematic pattern in the deviations of Y_i's from the average value, $\overline{Y}$. On the other hand, most values of S_k^* are positive, if the Y_i's tend to be larger (smaller) than $\overline{Y}$ for $i \leq m$ (for $i > m$). In order to further explain this point an illustrative example is presented in Fig. 3.2.

On the other hand, rescaled adjusted partial sums are obtained by dividing the S_k^*'s by the sample standard deviation, D_Y, as,

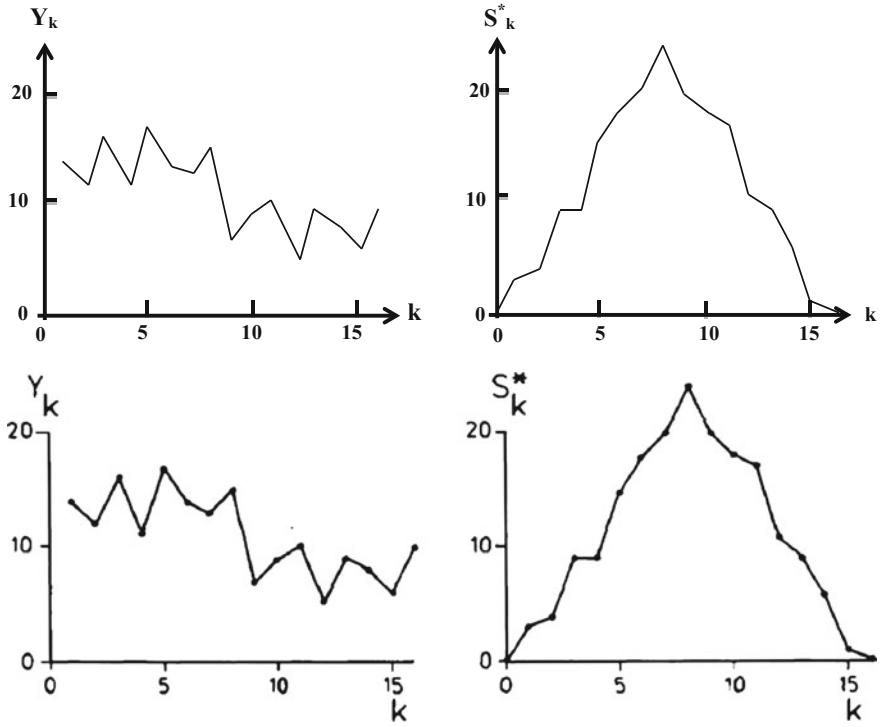


Fig. 3.2 Nonhomogeneous time series and adjusted partial sums

$$S_k^{**} = \frac{S_k^*}{D_Y} \quad (k = 0, 1\ldots, n) \tag{3.61}$$

Since S_k^{**}'s are not influenced by any linear transformation of the data, homogeneity tests are based on S_k^{**} calculations. A statistic that is sensitive to departures from homogeneity is given as,

$$\begin{aligned} Q &= \max\left|S_k^{**}\right| \\ &0 \le k \le n \end{aligned} \tag{3.62}$$

High Q values indicate a change in the level. The critical values for the test-statistic are given in Table 3.6. The percentage points in this table are based on 19,999 synthetic sequences of Gaussian pdf random numbers. For $n \to \infty$ the critical values of Q can be obtained from Table 3.8 of the Kolmogorow–Smirnov goodness-of-fit statistic.

Table 3.8 Percentage points of $Q/\sqrt{n}$ and $R/\sqrt{n}$

n	$Q/\sqrt{n}$			$R/\sqrt{n}$		
	90%	95%	99%	90%	95%	99%
10	1.05	1.14	1.29	1.21	1.28	1.38
20	1.10	1.22	1.42	1.34	1.43	1.60
30	1.12	1.24	1.46	1.40	1.50	1.70
40	1.13	1.26	1.50	1.42	1.53	1.74
50	1.14	1.27	1.52	1.44	1.55	1.78
100	1.17	1.29	1.55	1.50	1.62	1.86
∞	1.22	1.36	1.63	1.62	1.75	12.00

Another statistic for testing homogeneity is the range, R, given as follows.

$$R = \max_{0 \le k \le n}\left(S_k^{**}\right) - \min_{0 \le k \le n}\left(S_k^{**}\right) \tag{3.63}$$

Such ranges are very significant statistical parameters in reservoir storage capacity determination studies. Many works have been done by different researchers on the ranges and rescaled ranges (Hurst 1951; Şen 1974; Gomide 1978). Higher range values imply shifts (jumps) in the mean value of the time series. A figure with percentage points of R distribution under the null hypothesis is given by Wallis and O'Connell (1973).

3.3.14 Bayesian Test

Chernoff and Zacks (1964) and Gardner (1969) suggested the Bayesian procedures for the detection of changes in the mean value. In this test, the standard deviation, σ_Y, of the time series must be known. However, if the population standard deviation is not known then the sample standard deviation can be employed in the test. Gardner (1969) statistic for a two-sided test at an unknown point can be written as,

$$G = \sum_{i=1}^{n-1} p_k \left(\frac{S_k^*}{\sigma_Y}\right)^2 \tag{3.64}$$

Herein, p_k is for the prior probability that the shift occurs just after the kth observation.

The uniform prior pdf, P_k, has been given as,

$$U = \frac{1}{n(n-1)} \sum_{k(1)}^{n-1} \left(S_k^{**}\right)^2 \tag{3.65}$$

or the proportional pdf to $1/[n(n-1)]$, is,

Table 3.9 Percentage U and A points

n	U			A		
	90%	95%	99%	90%	95%	99%
10	0.336	0.414	0.575	1.90	2.31	3.14
20	0.343	0.447	0.662	1.93	2.44	3.50
30	0.344	0.444	0.691	1.92	2.42	3.70
40	0.341	0.448	0.693	1.91	2.44	3.66
50	0.342	0.452	0.718	1.92	2.48	3.78
100	0.341	0.457	0.712	1.92	2.48	3.82
∞	0.347	0.461	0.743	1.93	2.49	3.86

$$A = \sum_{k=1}^{n-1} \left(Z_k^{**}\right)^2 \tag{3.66}$$

Departures from homogeneity are evident provided that test-statistics have large values. Critical values for U and A are given in Table 3.9 with the percentage points that are based on 19,999 synthetic sequences of Gaussian pdf random numbers. The limiting distributions of U and A are those of certain test-statistics of the Cramer–von Mises type. The statistic U/n corresponds asymptotically with Smirnov's ω^2 and the statistic A with the Anderson–Darling statistic.

3.3.15 Relative Error Test

The relative error, α, between two values is equal to absolute difference between two variables $(X_j - X_i)$ divided by the bigger one and the multiplied by 100.

$$\alpha = \frac{\left|X_i - X_j\right|}{X_j} \tag{3.67}$$

In practice, if the relative error is less than 5% or in some other applications less than 10% then the difference between the two values is regarded as insignificant.

The relative error calculation does not require any specific pdf for the data, and hence, it is free of frequency diagram or pdf It is useful for the preliminary check on whether the data sequence is homogeneous or not.

Example 3.3 Check the homogeneity of 12 rainfall (cm) values given in the following at 5% level, 12.3, 14.2, 8.3, 11.2, 9.2, 13.4, 10.3, 12.9, 9.9, 11.5, 10.7, 8.9.

Solution 3.3 In the homogeneity test, the arithmetic average of the data can be compared with the arithmetic average of at least two complementary subdivisions. For this purpose, herein, the series is divided into two halves and the arithmetic averages of the whole series with the two halves are $\overline{X} = 11.06$ cm, $\overline{X}_1 = 11.43$ cm and $\overline{X}_2 = 10.70$ cm, respectively. Substitution of halve averages together with the

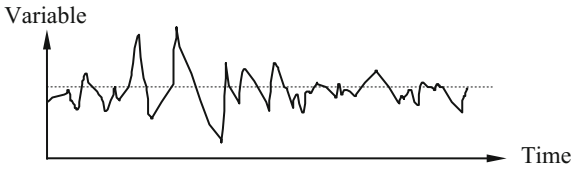


Fig. 3.3 Standard deviation heterogeneity

overall average into Eq. (3.67) leads to relative averages as $\alpha_1 = 0.03237$ and $\alpha_2 = 0.06387$. Since both of these are less than 10% significance level, the rainfall sequence is considered as homogeneous.

In some cases it is possible to obtain homogeneity on the average level, but the sequence may not be homogeneous with respect to other parameters. In order to check the homogeneity with respect to any parameter, the relative error procedure is similar as the average parameter. For instance, in Fig. 3.3, although the given series is homogeneous with respect to average parameter, but heterogeneity exists so far as the deviations are concerned, and hence, it needs for standard deviation homogeneity checking.

Example Check whether the rainfall sequence in the previous example is homogeneous or not?

Solution The standard deviations of the whole and two-half series are $\sigma_X = 1.78$ cm; $\sigma_1 = 2.13$ cm; and $\sigma_2 = 1.26$ cm. Substitution of these values into Eq. (3.67) leads to relative error percentages as $\alpha_1 = 0.1643$ and $\alpha_2 = 0.4084$. Since both of these errors are more than 5 or 10%, the series is not homogeneous with respect to standard deviation.

It is advised here for better and reasonable results, the series must be divided at least to three parts and the homogeneity test must be applied on each one through the use of relative error concept.

3.3.16 *t* Test

This test is useful in searching for whether the two time series come from the same population pdf's. For instance, in order to check whether two time series arithmetic averages come from the same population, Student's *t*-test is employed. If from a normal distribution pdf different time series of the same length, *n*, are drawn, their arithmetic averages are rather different from each other. The frequency distribution of these arithmetic averages comply by *t*-distribution for small samples, but as the sample number increase, it converges to a normal (Gaussian) pdf. Figure 3.4 shows the shapes of different *t*-distributions, $f(t)$.

The relative error concept explained in the previous section is attractive for those who do not care about the pdf and in the case of very small data numbers, but if

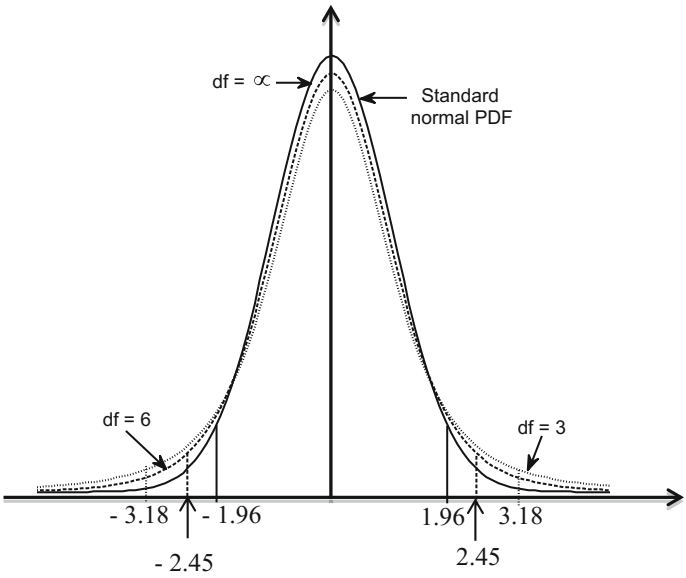


Fig. 3.4 Different *t*-distributions

sample length is more than 30, the statistical tests must be preferred. If two series with n_1 and n_2 data numbers have different arithmetic averages, $\overline{X}_1$ and $\overline{X}_2$, then the use of *t*-test comes into view. The probability values of a *t*-distribution are given in Table 3.10 for different degree of freedom.

The *t*-test can also be applied to two time series according to the standard deviation values equality or not as follows.

(1) The same standard deviation case: The test is performed only for the arithmetic averages. The difference between the two arithmetic averages, $\overline{X}_1 - \overline{X}_2$, is not sufficient for identification of whether they are significantly different from each other. The standard deviations must also play a role in the test. In order to obtain a dimensionless test statistic, the difference in the arithmetic averages is divided by the weighted average of the standard deviation by taking into consideration the data numbers, which leads to,

$$t = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sigma\sqrt{\left(\frac{n_1+n_2}{n_1 n_2}\right)}}, \tag{3.68}$$

where σ is the common standard deviation, which in terms of sample standard deviations (S_1 and S_2) given as,

Table 3.10 *t*-distribution values

df	α = 0.1	α = 0.05	df	α = 0.1	α = 0.05
1	6.314	12.706	21	1.721	2.08
2	2.92	4.303	22	1.717	2.074
3	2.353	3.182	23	1.714	2.069
4	2.132	2.776	24	1.711	2.064
5	2.015	2.571	25	1.708	2.06
6	1.943	2.447	26	1.706	2.056
7	1.895	2.365	27	1.703	2.052
8	1.86	2.306	28	1.701	2.048
9	1.833	2.262	29	1.699	2.045
10	1.812	2.228	30	1.697	2.042
11	1.796	2.201	35	1.69	2.03
12	1.782	2.179	40	1.684	2.021
13	1.771	2.16	45	1.68	2.014
14	1.761	2.145	50	1.676	2.009
15	1.753	2.131	60	1.671	2
16	1.746	2.12	70	1.667	1.994
17	1.74	2.11	80	1.664	1.99
18	1.734	2.101	90	1.662	1.987
19	1.729	2.093	100	1.66	1.984
20	1.725	2.086	120	1.658	1.98
			$\propto$	1.645	1.96

Note When the degrees of freedom, df, exceeds 30 critical values are not tabulated for every case

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 1} \tag{3.69}$$

This will have the following degree of freedom,

$$v = n_1 + n_2 - 2 \tag{3.70}$$

With these two last values, if one enters the standard *t*-distribution table and obtains the critical *t* value, t_{cr} at 5 or 10% level. Hence, if $t \leq t_{cr}$ then on the basis of arithmetic averages the two time series are indifferent.

(2) Different standard deviations case: The *t*-test statistics is defined for different σ_1 and σ_2 standard deviations of two time series as,

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}, \tag{3.71}$$

where n and m are the sample lengths of the time series. The corresponding degree of freedom is given as,

$$v = \left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right) \Big/ \left[\frac{\left(S_1^2/n\right)^2}{n-1} + \frac{\left(S_2^2/m\right)}{m-1}\right] \tag{3.72}$$

The use of the similar procedure with Table 3.10 leads to desired answer. The distribution and its shape depend on the degrees of freedom (df). The t-distribution has thicker tails than the normal distribution but as the df increases the t-distribution approximates the normal distribution. The area under the pdf to the right of t is equal to $\alpha/2$ (see Fig. 3.4).

3.3.17 Cramer Test

This test is for the comparison of long-term arithmetic average of a time series with shorter duration arithmetic averages from the same time series. The test statistic is given as,

$$t_k = \tau_k \left[n(N-2)/\left(N - n\tau_k^2\right)\right]^{1/2} \tag{3.73}$$

and

$$\tau_k = \frac{\overline{Y}_k - \overline{Y}}{\sigma}, \tag{3.74}$$

where

$$\overline{Y}_k = \frac{1}{n}\sum_{i=k+1}^{k+n} Y_i \tag{3.75}$$

and

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{3.76}$$

Herein, $\overline{Y}$ and σ are the arithmetic average and the standard deviation of whole data. However, $\overline{Y}_k$ is the arithmetic average of a sub-length with k data values within the whole n data number. Theoretical studies indicate that the test statistic in Eq. (3.73) has a t-distribution with $(n - 2)$ degrees of freedom. In order to depict any significant difference between the whole and partial arithmetic averages, it is necessary to make two tailed test.

3.3.18 F Test

The same question of significance on the basis of standard deviations σ_1 and σ_2 of two time series instead of average values can be tested by Fisher F-test instead of *t*-test. In order to understand whether the standard deviations of two time series are significantly different from each other the variances ratio is defined as,

$$F = \frac{\sigma_1^2}{\sigma_2^2} \tag{3.77}$$

This ratio is calculated such that it is greater than one, i.e., $\sigma_1^2 \geq \sigma_2^2$. Theoretically, the F ratio abides by Fisher (F) pdf, the properties of which are presented in Fig. 3.5.

(1) There are different F-distributions according to F ratio value and numerator and denominator degrees of freedom,
(2) Increase in the degrees of freedom of the numerator and denominator leads to more symmetric F pdf,
(3) For rather big n_2 values, the arithmetic average of F-distribution approaches to one and it is equal exactly to $(n_2 - 1)/(n_2 - 3)$.

According to two degrees of freedom defined as $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$, the critical F value, F_{cr} is read from the F-distribution table (Table 3.11) with a significance level, which is taken in most practical studies as 5% ($\alpha = 0.05$). If the F-ratio from Eq. (3.77) is less than the critical F value, then the two time series do not have significantly different standard deviations. Otherwise, the standard deviation difference is significant.

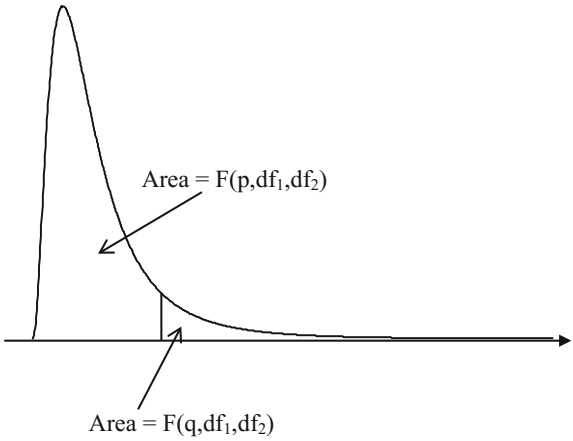


Fig. 3.5 F-distributions

Table 3.11 F distribution values

		v_2									
		1	2	3	4	5	6	7	8	9	10
v_1	1	39.86346	49.5	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498
	2	8.52632	9	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157
	3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24	5.23041
	4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988
	5	4.06042	3.77972	3.61948	3.5202	3.45298	3.40451	3.3679	3.33928	3.31628	3.2974
	6	3.77595	3.4633	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693
	7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251
	8	3.45792	3.11312	2.9238	2.80643	2.72645	2.66833	2.62413	2.58935	2.56124	2.53804
	9	3.3603	3.00645	2.81286	2.69268	2.61061	2.55086	2.50531	2.46941	2.44034	2.41632
	10	3.28502	2.92447	2.72767	2.60534	2.52164	2.46058	2.41397	2.37715	2.34731	2.3226
	11	3.2252	2.85951	2.66023	2.53619	2.45118	2.38907	2.34157	2.304	2.2735	2.24823
	12	3.17655	2.8068	2.60552	2.4801	2.39402	2.33102	2.28278	2.24457	2.21352	2.18776
	13	3.13621	2.76317	2.56027	2.43371	2.34672	2.28298	2.2341	2.19535	2.16382	2.13763
	14	3.10221	2.72647	2.52222	2.39469	2.30694	2.24256	2.19313	2.1539	2.12195	2.0954
	15	3.07319	2.69517	2.48979	2.36143	2.27302	2.20808	2.15818	2.11853	2.08621	2.05932
	16	3.04811	2.66817	2.46181	2.33274	2.24376	2.17833	2.128	2.08798	2.05533	2.02815
	17	3.02623	2.64464	2.43743	2.30775	2.21825	2.15239	2.10169	2.06134	2.02839	2.00094
	18	3.00698	2.62395	2.41601	2.28577	2.19583	2.12958	2.07854	2.03789	2.00467	1.97698
	19	2.9899	2.60561	2.39702	2.2663	2.17596	2.10936	2.05802	2.0171	1.98364	1.95573
	20	2.97465	2.58925	2.38009	2.24893	2.15823	2.09132	2.0397	1.99853	1.96485	1.93674
	21	2.96096	2.57457	2.36489	2.23334	2.14231	2.07512	2.02325	1.98186	1.94797	1.91967
	22	2.94858	2.56131	2.35117	2.21927	2.12794	2.0605	2.0084	1.9668	1.93273	1.90425

(continued)

Table 3.11 (continued)

		v_2									
		1	2	3	4	5	6	7	8	9	10
v_1	23	2.93736	2.54929	2.33873	2.20651	2.11491	2.04723	1.99492	1.95312	1.91888	1.89025
	24	2.92712	2.53833	2.32739	2.19488	2.10303	2.03513	1.98263	1.94066	1.90625	1.87748
	25	2.91774	2.52831	2.31702	2.18424	2.09216	2.02406	1.97138	1.92925	1.89469	1.86578
	26	2.90913	2.5191	2.30749	2.17447	2.08218	2.01389	1.96104	1.91876	1.88407	1.85503
	27	2.90119	2.51061	2.29871	2.16546	2.07298	2.00452	1.95151	1.90909	1.87427	1.84511
	28	2.89385	2.50276	2.2906	2.15714	2.06447	1.99585	1.9427	1.90014	1.8652	1.83593
	29	2.88703	2.49548	2.28307	2.14941	2.05658	1.98781	1.93452	1.89184	1.85679	1.82741
	30	2.88069	2.48872	2.27607	2.14223	2.04925	1.98033	1.92692	1.88412	1.84896	1.81949
	40	2.83535	2.44037	2.22609	2.09095	1.99682	1.92688	1.87252	1.82886	1.7929	1.76269
	60	2.79107	2.39325	2.17741	2.04099	1.94571	1.87472	1.81939	1.77483	1.73802	1.70701
	120	2.74781	2.34734	2.12999	1.9923	1.89587	1.82381	1.76748	1.72196	1.68425	1.65238
	inf	2.70554	2.30259	2.0838	1.94486	1.84727	1.77411	1.71672	1.6702	1.63152	1.59872

3.3.19 Truncation Test

A given time series, Y_i ($i = 1, 2, \ldots, n$), are truncated by a constant level, Y_0, and hence, multiple data values are reduced to two types, those greater than ($Y_i > Y_0$) or less than ($Y_i > Y_0$) the truncation level. The truncation level may be taken statistically as the arithmetic average, mode value, or the median value. For instance, if a particular value has high frequency of occurrence then the mode value is taken as truncation. There are also some engineering truncation levels for instance; water demand can be considered as a truncation level. In the data treatment procedures unless there is some valid arguments, always the statistical parameters are adapted as the truncation value. The truncation of a time series in Fig. 3.6 at a given level of Y_0 leads to two linguistic variables as surplus and deficit. This linguistic words may be plus/minus, yes/no, black/white, hot/cold, dry/wet, empty/full, etc. depending on the problem at hand.

If the surpluses (deficits) continue successively for a period, this is referred to as surplus (deficit) spell. The summation of surpluses and deficits is equal to data number. The number of surplus periods is equal to deficit periods or the difference between the two is equal to 1. The longer the duration of surplus (deficit) spell the more is the correlation of successive values within the time series. If the time series is homogeneous, then the number of surpluses is equal to deficits. In a homogeneous time series, the durations are rather small, random and close to each other. Under the light of all these arguments, if the two time series as in Fig. 3.7 are compared then the following points can be observed.

If the total number of deficit and surplus period is n_{ds} this value should remain between the following upper and lower limits for the time series to be homogeneous (Koçak and Şen 1998).

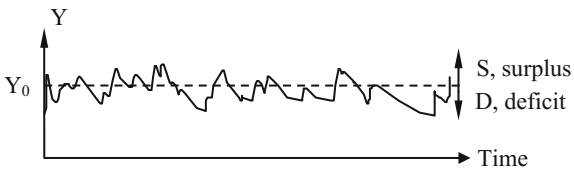


Fig. 3.6 SF truncation

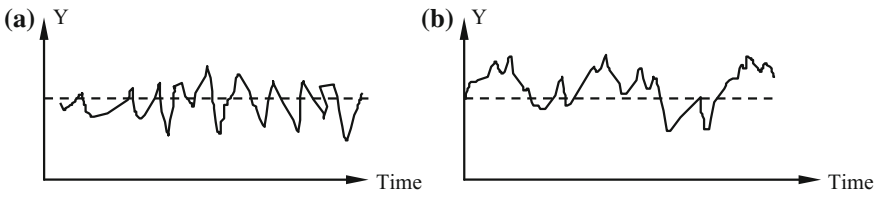


Fig. 3.7 Comparative homogeneity analysis

$$\begin{aligned} n_{\text{upper}} &= 1.104\frac{n}{2} + 2.384, \\ n_{\text{lower}} &= 0.931\frac{n}{2} - 1.829, \end{aligned} \tag{3.78}$$

where n is the data number.

3.3.20 Deviations Test

This is also a simple test for deciding whether the time series is homogeneous or not depending on its serial properties. For a given series $Y_1, Y_2, Y_3, \ldots, Y_n$ the deviations are calculated from the arithmetic average as $Y_1 - \overline{Y}, Y_2 - \overline{Y}, Y_3 - \overline{Y}, \ldots, Y_n - \overline{Y}$. This is equivalent to saying that the series is truncated at the arithmetic average $\overline{Y}$ level. It is possible to calculate the quantities of $A = \Sigma(Y_i)^2$ and $B = \Sigma(Y_i - Y_{i-1})$, for $i = 1, 2, 3, \ldots, n$ such than $\overline{Y} = N_n$, and finally, the ratio, $2A/B$. If the following expression is valid then the time series is homogeneous.

$$\left(1 - \sqrt{1/n}\right) \leq 2A/B \leq \left(1 + \sqrt{1/n}\right) \tag{3.79}$$

3.3.21 Subtraction Test

The corresponding values at two time series are subtracted from each other. Hence, two time series is reduced to a single one, which can be tested by some of the aforementioned tests. The subtraction test assesses the serial appearance of the differences. For this purpose, the arithmetic average of the difference time series is calculated and then subtracted from each one of the difference value. If the result is positive, it is labeled by S otherwise by D. This yields to a time series including succession of two symbols as SSSDDDSDSD… DSDSSSSDSDDDDS. This sequence is then partitioned into two-successive and nonoverlapping pairs as SS, SD, DD, SD, SD… DS, DS, SS, SD, SD, DD, DS. If the pairs have the same symbols they are regarded as the same, otherwise they are different. If the number of the same pairs is A and the different pairs is B, then the difference $A - B$ provides a quantity, which helps for the decision of the homogeneity of the two time series, otherwise, when the difference is outside of these two limits, the time series have heterogeneity.

The test is given as

$$-\sqrt{n-1} \leq A - B \leq +\sqrt{n-1} \tag{3.80}$$

3.3.22 Şen Autorun Test

It is generally accepted that if a time series is dependent the high values tend to follow high values and the low values tend to follow low values. This statement implies a dependent series in which periods of wet and dry spells tend to be greater than that in the case of an independent series. Such a property is termed "persistence." There are various attempts to measure "persistence" mainly by three procedures, namely, autocorrelation function, spectral analysis, and rescaled-range analysis (Chap. 2). Wet and dry spells are directly related to run properties in statistics as explained by various researchers (Parzen 1960; Feller 1968) and in hydrology (Yevjevich 1967; Saldarriga and Yevjevich 1979; Şen 1976).

Autocorrelation analysis is a means of measuring linear dependence between any two series. As stated by Feller (1968), it is by no means a general measure of dependence because it involves all the assumptions stated in Chap. 2. The rescaled-range analysis which was introduced into hydrology by Hurst (1951, 1956), has the advantage of being comparatively more robust than any other technique, and it is not very sensitive to the marginal pdf (Chap. 2).

The autorun coefficient has already been presented by Şen (1978) based on the median value and then it is generalized to other exceedance probability levels ($0 < p < 1$).

Joint and conditional probabilities are also measures of dependence. A joint probability is equal to the multiplication of a conditional probability by a marginal probability (Feller 1968). In particular, if Y_i and Y_{i-k} are two dependent events with joint probability, $P(Y_i, Y_{i-k})$, their conditional probability is denoted by $P(Y_i/Y_{i-k})$. Herein, k is referred to as the lag and indicates the time difference between the two events provided that $P(Y_{i-k})$ is the marginal probability of event Y_{i-k}. The probability statement between these two probabilities is,

$$P(Y_i, Y_{i-k}) = P(Y_i/Y_{i-k})\, P(Y_{i-k}) \tag{3.81}$$

Furthermore, the conditional probability is defined by Şen (1976) as the autorun coefficient, $r = P(Y_i/ Y_{i-k})$, and hence, the Eq. (3.81) yields,

$$r_k = \frac{P(Y_{i-k}, Y_i)}{P(Y_i)} \tag{3.82}$$

If a time series is truncated at an arbitrary constant level, Y_0, and then Eq. (3.82) can be defined in terms of probabilities as,

$$r_k = \frac{P(Y_{i-k} > Y_0\,,\, Y_i > Y_0)}{P(Y > Y_0)} \tag{3.83}$$

A special case is defined by Şen (1978) for the median, m, truncation level for which $P(Y_i > Y_0) = 0.5$, which is the exceedance probability. The non-exceedance probability is $q = 1 - p$. Generally, r_k, definition becomes,

$$r_k = \frac{P(Y_{i-k} > Y_0, Y_i > Y_0)}{p} \tag{3.84}$$

An estimate, $\hat{r}_k$, of r_k can simply be proposed by considering Eq. (3.84) together with the classical definition of probability in textbooks. The probability, $P(Y_i)$, is found by counting the total number, n_x, of occurrences, exceedances in the autorun case, and consequently, one can define simply,

$$P(Y_i > Y_0) = \frac{n_x}{n} \tag{3.85}$$

In the case of a joint event ($Y_i > Y_0$, $Y_{i-k} > Y_0$), in a sequence of n observations, $n - k$ possible alternatives exist for two observations at lag-k apart. If the number of joint events in a given sequence of length n is n_k, then from Eq. (3.85) one can obtain,

$$P(Y_i > Y_0, Y_{i-k} > Y_0) = \frac{n_k}{n-k} \tag{3.86}$$

The substitution of which into Eq. (3.81) leads to the small sample estimate of r_k as,

$$\hat{r}_k = \frac{n_k}{p(n-k)} \tag{3.87}$$

The numerator is an integer random variable, whereas the denominator is a fixed value for given n, k, and p. The random characteristics of $\hat{r}_k$ can be obtained directly from the characteristics of random variable, n_k. For instance, if the expected value $E(n_k)$ and the variance $V(n_k)$, of n_k are known, then the expectation and variance of $\hat{r}_k$ could be evaluated, respectively, as,

$$E(\hat{r}_k) = \frac{E(n_k)}{p(n-k)} \tag{3.88}$$

and

$$V(\hat{r}_k) = \frac{V(n_k)}{pq(n-k)^2} \tag{3.89}$$

On the basis of the frequency interpretation of the probability, the estimate $\hat{r}_k$ can be calculated by successive execution of the following steps:

(a) The exceedance probability, p, and its corresponding truncation level Y_0 is calculated from a given time series Y_1, Y_2, …, Y_n,
(b) The series is truncated at the level of x_0 giving rise to sequences of surpluses ($Y_i > Y_0$) and deficits ($Y_i \leq Y_0$),

(c) The number, n_k, of overlapping successive surplus pairs (observations lag-k apart) are counted,
(d) The estimate of r_k is then calculated from Eq. (3.89).

These four steps are distribution-free. If the length of time series is very large then $\hat{r}_k$ becomes,

$$r_k = \lim_{n\to\infty} \frac{2n_k}{n-k} = 2\,P(Y_i > m, Y_{i-k} > m) \tag{3.90}$$

Contrary to the autocorrelation analysis, the autorun analysis does not distort the dependence structure of the sequence considered. The autorun coefficients are easier to calculate than the autocorrelation coefficients. Furthermore, autorun analysis is directly related to run properties, which play an effective role in various engineering problems such as droughts, floods, reservoir operation, etc. Extensive simulation studies using a first-order Markov process are carried out and the autorun function is calculated according to Eq. (3.87). The simulation graphs are presented in Fig. 3.8 for $\rho = 0.9$. It is possible to obtain many autorun functions each for different truncation level, Y_0, but only one autocorrelation function exists.

Each one of the autorun functions starts from $r_0 = 1$ at lag zero and become asymptotic (large lags) to the exceedance probabilities. The decrease in the autorun function by lag appears according to an exponential function, and hence, it is possible to relate the autorun function to the exceedance probability and lag-k as,

$$r_k = 1 - (1-p)e^{-pk} \tag{3.91}$$

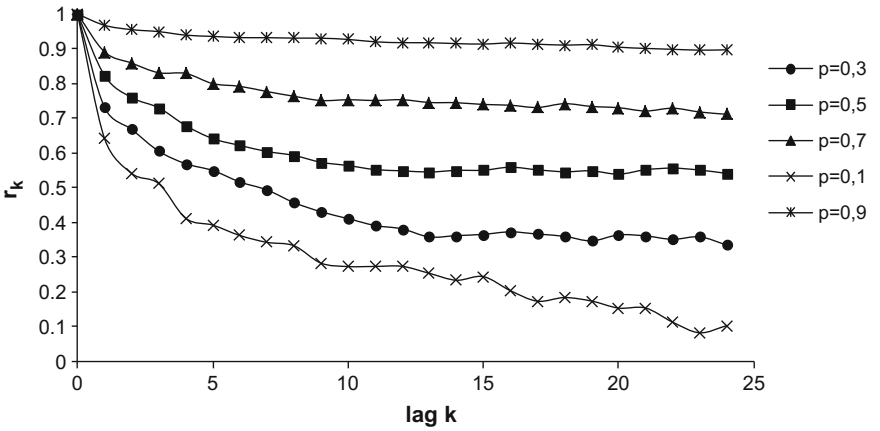


Fig. 3.8 Autorun coefficient variations with lag

3.3.23 Seasonal Kendall Test

Most of the natural phenomena show strong seasonal patterns. For instance, there are seasonal variations in temperature, evaporation, and precipitation records. In time series records seasonal component occupy significant portion. Seasonal rise and fall of temperature imply cooling and heating seasonality in energy demand.

The seasonality can be accounted by two hybrid procedures as the seasonal Kendall test for residuals from a simple linear regression analysis of Y versus X and the given data can be deseasonalized by subtracting seasonal means or medians from all the data within the same season (Chap. 2). The deseasonalized data is then regressed against time (Montgomery and Reckhow 1984).

First, Mann–Kendall procedure is applied to each season and then each season results are combined together. The seasonal Kendall test is weakly powerful but robust. The statistic, S_k, of this procedure is equal to the summation of the S_i value obtained from ith season.

$$S_k = \sum_{i=1}^{m} S_i, \tag{3.92}$$

where m is the number of seasons. If the product of the number of seasons, m, and number of years, n, $m \times n$ is greater than 30, the pdf of S_k has approximate normal pdf and it is negative for a declining trend, (Gilbert 1987). An estimate of the trend slope for Y over time can be computed as the median of all slopes between data pairs within the same season using a generalized version of the Sen's slope estimator (Sect. 3.3.8).

The seasonal Kendall test is a nonparametric method, which depicts seasonality on the basis of the MK test (Sect. 3.3.8). In practical applications most often the seasonality is concerned in the monthly or three-monthly (January–February–March, April–May–June, July–August–September, and October–November–December) epochs. Kendall's S statistic S_i for each season i is summed over all seasons to form the overall test statistic S_k.

The slope of trend in each season is calculated by the Sen's slope, which is the median annual slope of all possible pairs of values in each season. For instance, if for January the slope is sought and two different years' January parameter (median) n-year apart is considered. The slope is equal to the difference between the two values divided by n. A meaningful and statistically significant Kendall test has more than 1% of the median value.

In many natural time series, there is embedded seasonal effect either due to natural or man-made impacts. Hence, there is a strong seasonal variation in the time series. It is necessary to depend on the definition of seasonality prior to any trend procedure application. In practice frequently monthly or 3-monthly (quarterly block) seasonality is employed.

The cyclic (seasonal) variations can be described by combination of sine and cosine functions in the form of Fourier series as a form of parametric multiple

regression models with a trend component, which has the simplest mathematical form as,

$$Y_t = \alpha + \beta \sin\left(\frac{2\pi t}{n}\right) + \gamma \cos\left(\frac{2\pi t}{n}\right)\delta t + \varepsilon, \tag{3.93}$$

where n is the number of data per year (52 for weekly data, 12 for monthly data, 4 for seasonal data, 2 for 6-monthly data). The trend can be identified, if the parameter δ is significantly different from zero.

3.4 Unit Root Model Trend Determination

Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests are helpful as null hypothesis whether a given time series is stationary around a linear trend, which implies that the time series is "trend-stationary" and the alternative hypothesis unit root. In the KPSS test, the absence of a unit root is not a proof of stationarity but of trend-stationary. It is possible for a time series to be nonstationary without unit root yet be trend-stationary. In both unit root and trend-stationary processes, the mean can grow or decrease over time. However, in the presence of a random component, trend-stationary processes are mean-reverting, while unit-root processes have a permanent impact on the mean (i.e., no convergence over time).

KPSS-type tests are intended to complement unit root tests, such as the Dickey–Fuller (DF) tests. By testing both the unit root hypothesis and the stationarity hypothesis, one can distinguish series that appear to be stationary, series that have a unit root, and series for which the data (or the tests) are not sufficiently informative to be sure whether they are stationary or integrated.

Many economic and financial time series exhibit trending behavior or nonstationarity in the mean (see Chap. 8). Leading examples are asset prices, exchange rates and the levels of macro-economic aggregates like real GDP. An important econometric task is determination of the most appropriate trend form in the time series. For example, in autoregressive moving average (ARMA) modeling, the time series must be transformed to a stationary form prior to analysis. If the time series are trending, then some form of trend removal is required.

Two common trend removal or de-trending procedures are first differencing and time-trend regression. First differencing is appropriate for $I(1)$ time series and time-trend regression is appropriate for trend-stationary $I(0)$ time series. Unit root tests can be used to determine, if trending time series should be first differenced or regressed on deterministic functions of time to render the time series stationary. Moreover, economic and finance theories often suggest the existence of long-run equilibrium relationships among nonstationary time series variables. If these variables are $I(1)$, then co-integration techniques can be used to model these long-run relations. A common trading strategy in finance involves exploiting mean-reverting

behavior among the prices of assets pairs. Unit root tests can be used to determine, which pairs of assets appear to exhibit mean-reverting behavior.

The core of the following explanations is available at the Warsaw School of Economics (Syczewska 2010). Kwiatkowski–Phillips—Schmidt–Shin (KPSS) test that is introduced in 1992 is the test for time series stationarity versus alternative of unit root. Unit root tests started with classic DF test, and later on several refinements such as Perron-type tests have as a null hypothesis presence of unit root in the series with the alternative hypothesis of stationarity. The KPSS test differs from the majority of tests used for checking integration in that its null hypothesis of stationarity is a simple hypothesis.

3.4.1 Integration and Dickey–Fuller (DF) Test

Let a given time series record has a generation process as first-order AutoRegressive, AR(1) structure, which is also referred to as the first-order Markov process, and it is widely used in many branches of different disciplines.

$$Y_t = \Phi Y_{t-1} + \varepsilon_t, \tag{3.94}$$

where is a stationary disturbance term or autocorrelation coefficient and ε_t is independent stationary random variable. If $\Phi = 1$ characteristic equation of the process in Eq. (3.94) has a unit root then the process is nonstationarity. As long as ε_t is stationarity first differences of Y_t are stationary. The series Y_t is integration of the first order, $I(1)$. Provided that $\Phi < 1$, Y_t time series is stationary in the sense that it is integrated of order zero, $I(0)$. Integration order of Y_t determines its properties (Mills 1993).

In case of $I(0)$ the time series, $_t$, has the following statistical properties, which are common reflections of stationarity.

(1) The standard deviation of the time series is time independent and constant,
(2) The residual (random variable), ε_t affects the series in a sequence of random shocks,
(3) Expected time duration between zero crossings has a finite value,
(4) The autocorrelation function diminishes with increasing lag and the summation of the correlation coefficients attains to a constant value.

On the other hand, if the time series is integrated of order one, $I(1)$, then it will reflect the following features.

(1) The standard deviation of the time series tends to infinity with time,
(2) The residual, i.e., random shock independent variable, ε_t, effects on Y_t at each time instance, as the summation of all the previous random variables (shocks),
(3) Expected time duration between consecutive crossing points on $Y = 0$ line is an infinite number,
(4) The autocorrelation coefficient tends to infinity with increasing lags.

The aforementioned features have impacts on macro-economic time series records. These features should be taken into consideration in the construction of any econometric model.

The DF test (Dickey and Fuller 1979, 1981) is the null hypothesis that in a model, which is represented by Eq. (3.94) model for $\delta = \Phi - 1$ as follows,

$$\Delta y_t = \delta Y_{t-1} + \varepsilon_t \tag{3.95}$$

In case of $\delta = 0$ the time series, Y_t, is Brownian process with accumulation of random shocks. Alternative hypothesis as $\delta < 0$ implies that the variable is stationary.

The test statistics is computed as, $t = \hat{\delta}/\hat{\sigma_\delta}$ similar to the Student's t-ratio but with different pdf. In order to make a decision, a critical value is necessary at a chosen significance level. If the test statistic exceeds a critical value then the null hypothesis cannot be rejected about presence of unit root in a series. However, if test statistics is smaller than the critical value, then null hypothesis is rejected in favor of time series stationarity. In general, disturbance terms in Eq. (3.95) are correlated and in the augmented DF test, this correlation is taken into consideration by including lagged values of Y_t differences on the right-hand side of Eq. (3.95). One can also include a constant as follows.

$$Y_t = \alpha_0 + \Phi\, Y_{t-1} + \varepsilon_t \tag{3.96}$$

On the other hand, in case of a stationary time series with a linear trend around the mean, the formulation takes the following form.

$$y_t = \alpha_0 + \zeta\, t + \Phi\, y_{t-1} + \varepsilon_t \tag{3.97}$$

If $|\Phi| < 1$ then the time series is stationary around the linear trend, otherwise for $\alpha = 1$ there is a unit root ($\alpha = 1$) in the time series, which is nonstationary.

3.4.2 The Kwiatkowski, Phillips, Schmidt, and Shin Test

Kwiatkowski et al. (1992) suggested a test, shortly KPSS test, with a null hypothesis of stationarity time series around either mean or a linear trend against the alternative hypothesis with the assumption that a time series is nonstationary due to presence of a unit root. The difference of this test from the previous one is that the null hypothesis assumes presence of a unit root.

In this test, time series record is represented as a sum of three distinctive components, namely, deterministic trend, ζ_t, a random walk, r_t, and a stationary error term, ε_t.

$$\begin{aligned} Y_t &= \zeta t + r_t + \varepsilon_t \\ r_t &= r_{t-1} + u_t \end{aligned} \tag{3.98}$$

where ζ is the slope of trend, u_t is another error (random) term with zero mean and $\widehat{\sigma}_u^2$ similar to ε_t. In Eq. (3.98), r_0 is a constant value corresponding to the intercept of time series.

In case that the variance, $\widehat{\sigma}_u^2$, is greater than zero, then time series has nonstationary behavior (as sum of a trend and random walk), as a result of a unit root. Subtracting from both sides of the first line in Eq. (3.5) one can obtain,

$$\Delta y_t = \zeta + u_t + \Delta\varepsilon_t = \beta + w_t, \tag{3.99}$$

where w_t is generated by AR(1) process as $w_t = v_t + \Phi v_{t-1}$ (Kwiatkowski et al. 1992). Finally, the KPSS test may be expressed as,

$$\begin{aligned} Y_t &= \zeta + \beta Y_{t-1} + w_t \\ w_t &= v_t + \theta v_{t-1}, \quad \beta = 1 \end{aligned} \tag{3.100}$$

This equation provides an interesting relationship between KPSS test and DF test. The DF test checks for $\beta = 1$ on the assumption that $\theta = 0$; where θ is a nuisance parameter. However, Kwiatkowski et al. (1992) assume that β is a nuisance parameter and test whether $\theta = -1$, assuming that $\beta = 0$. They also introduce one-side Lagrange Multiplier test of null hypothesis $\sigma_u^2 = 0$ with assumption of a normal pdf and ε_t as identically distributed independent random variables with zero expected value and a constant variance, σ_ε^2. After all what have been explained above the KPSS test application steps are given along the following points.

(1) *A null hypothesis test of stationarity around a linear trend versus alternative hypothesis of a unit root presence.*

Let ε_t, denote estimated errors from a regression on a constant. Let estimate of variance be equal to a sum of error squares divided by number of observations, *n*. The partial sums of errors should be calculated as,

$$S_t = \sum_{i=1}^{t} e_i \quad (t = 1, \ldots, n) \tag{3.101}$$

This leads to the definition of the LM test statistic as,

$$\mathrm{LM} = \frac{\sum_{t=1}^{n} S_t^2}{\sigma_\varepsilon^2} \tag{3.102}$$

(2) *A null hypothesis test of stationarity around mean, versus alternative hypothesis of a unit root presence.*

The estimated errors, ε_t, are computed as residuals of regression of Y_t on a constant ($\varepsilon_t = Y_t - \overline{Y}$) the rest of definitions are unchanged.

Asymptotic properties of the statistic is based on assumption that ε_t have certain regularity properties defined by Phillips and Perron (1988, p. 336). The long-run variance is,

$$\sigma^2 = \lim\left[\frac{E\left(S_T^2\right)}{n}\right] \tag{3.103}$$

The consistent estimate of the long-run variance is the following formula as (Kwiatkowski et al. 1992),

$$s^2(k) = \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 + \frac{2}{n}\sum_{j=1}^{k} w\,(j,k)\sum_{t=s+1}^{n}\varepsilon_t\,\varepsilon_{t-1}, \tag{3.104}$$

where $w(j, k)$ denote weights depending on a choice of spectral window. The authors use the Bartlett window, i.e., $w(j,k) = 1 - j/(k+1)$, which ensures non-negativity. They argue that for quarterly data lag $k = 8$ is the best choice (if $k < 8$, test is distorted, if $k > 8$, power decreases) (Kwiatkowski et al. 1992). The KPSS test statistic is computed as a ratio of sum of squared partial sums, and estimate of long-term variance,

$$\widehat{\eta} = \frac{1}{n^2}\sum_{t=1}^{n}\frac{S_t^2}{s^2\,(k)} \tag{3.105}$$

Symbols $\widehat{\eta}_\mu$ and $\widehat{\eta}_\tau$ denote respectively the KPSS statistic for testing stationarity around mean and around a trend.

Asymptotic distribution of the KPSS test statistic is nonstandard; it converges to aBrownian bridges of higher order (Kwiatkowski et al. 1992, p. 161). The $\widehat{\eta}_\mu$ statistic for testing stationarity around mean converges to,

$$\widehat{\eta}_\mu \rightarrow \int_0^1 V^2(r)\,\mathrm{d}r, \tag{3.106}$$

where $V(r) = W(r) - rW(r)$ denotes a standard Brownian bridge defined for a standard Wiener process $W(r)$, and goes to weak convergence of probability measures.

The KPSS test statistic $\widehat{\eta}_\tau$ for stationarity around trend, i.e., for $\xi \neq 0$, weakly converges to a second order Brownian bridge, $V_2(r)$, is defined as (Kwiatkowski et al. 1992),

$$V_2(r)_2 = W(r) + (2r - 3r^2)\, W(1) + \left(-6r + 6r^2\right) \int_0^1 W(s)\mathrm{d}s \tag{3.107}$$

The statistic weakly converges to a limit,

$$\widehat{\eta}_\tau \rightarrow \int_0^1 V_2^2(r)\, \mathrm{d}r \tag{3.108}$$

The KPSS test is performed in a following way. One can test null hypothesis about stationarity around mean or around trend, against alternative hypothesis of nonstationarity of a time series due to a unit root presence. It is possible to compute value of a test statistic, $\widehat{\eta}_\mu$ or $\widehat{\eta}_\tau$, respectively. If computed value is greater than critical value, the null hypothesis of stationarity is rejected at given level of significance.

3.4.3 Critical Values of the KPSS Test

In the original Kwiatkowski et al. (1992) paper the results of Monte Carlo simulation concerning size and power of the KPSS test and asymptotic properties of the test statistics are obtained with use of Eqs. (3.107) and (3.108). Hence, there is a need of computing critical values for finite sample size.

Monte Carlo simulation experiments are carried out for computation of KPSS test critical values based on the definition Eq. (3.107) for Gauss PDF independent process generation.

Data generating process used for simulation corresponds to the models in Eqs. (3.103) and (3.104). Number of lags is equal to 8. The model has the following form.

$$Y_t = \zeta\, t + r_0 + \varepsilon_t$$

and two versions for $\xi = 0$ model has a constant only, and for $\xi \neq 0$ constant and a linear trend. The test statistic is computed for $k = 8$ from Eq. (3.102) as,

$$\mathrm{LM} = \frac{\sum_{t=1}^{n} S_t^2}{s^2(8)},$$

where, similar to Eq. (3.104) is given as

$$s^2(8) = \frac{1}{T}\sum_{t=1}^{n} \varepsilon_t^2 + \frac{2}{T}\sum_{j=1}^{8} w(j,8) \sum_{t=s+1}^{n} \varepsilon_t \, \varepsilon_{t-j}$$

Sample size is set at 15, 20, 25, 30, 40, 50 60, 70, 80, 90, and 100. Number of replication equals 50,000 and the result of the computed KPSS test statistic are given in Table 3.12.

3.4.4 Empirical Power of the KPSS

A set of simulation experiments is carried out for checking the power of the KPSS test. Sample sizes are set at n = 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, and 100, with 10,000 replications. A random walk with nonzero variance is generated for the alternative of the KPSS test, i.e., nonstationarity of a time series due to a unit root presence. As explained earlier the error term variance equal to zero corresponds to a stationarity null hypothesis. Earlier experiments have shown that particular value of variance, as long as it is nonzero, have little effect on the results. It is assumed that variance takes three values: 0 (as a benchmark), 0.5, 1.0, and 1.5. Hence, time series generation process has the following form as,

$$Y_t = \zeta\, t + r_t + \varepsilon_t,$$
$$r_t = r_{t-1} + u_t,$$

where ε_t are disturbances terms, and u_t is independent identically distributed random variable according to a normal pdf. These two random variables are also mutually independent.

The experiments are performed for two versions, namely, with and without linear trend. In former case ξ = 0.1, but in the latter case ξ = 0. Test statistics values are compared with the critical values and the results are shown in Table 3.13.

Table 3.14 shows the results of checking whether the value of $\hat{\sigma}_u^2$ chosen in simulation has an effect on the empirical power of the KPSS test. The regression is run of a rejection percentage on two variables with $\hat{\sigma}_u^2 = \{0.0, 0.1, 0.2, \ldots, 1.4\}$ and $\alpha \in \{0.95, 0.90, 0.50, 0.1\}$. The choice of $\hat{\sigma}_u^2$ does not influence the empirical power of the test for a model with a linear trend.

Table 3.15 presents computation results of the empirical KPSS test power for 25, 30, 40, 50, …, 90, 100 observations. In the DGP, the variance takes the values, 0 (as a benchmark; this corresponds to a null of stationarity); 0.1, 0.2, … 1.4.

Table 3.12 Critical values of the KPSS test statistics for 50,000 replications

α	Without trend	Linear trend
Sample size = 15		
0.990	0.48313288	0.41433477
0.975	0.45183890	0.38740080
0.950	0.42608752	0.36435597
0.900	0.39875209	0.34151076
0.500	0.31307493	0.27000041
0.100	0.24775830	0.22441514
0.050	0.23429514	0.21764786
0.025	0.22542106	0.21314635
0.010	0.21814408	0.20935949
Sample size = 20		
0.990	0.42612535	0.32710900
0.975	0.40672348	0.30130862
0.950	0.38874144	0.27736290
0.900	0.36425871	0.25185147
0.500	0.25352270	0.18687704
0.100	0.17868235	0.16048566
0.050	0.16906971	0.15720065
0.025	0.13643788	0.15498356
0.010	0.15862807	0.15314154
Sample size = 25		
0.990	0.42646756	0.25070640
0.975	0.40531466	0.22643507
0.950	0.38197871	0.20925595
0.900	0.35089080	0.19228778
0.500	0.21829452	0.15145480
0.100	0.14634008	0.12768145
0.050	0.13698973	0.12327038
0.025	0.13060574	0.12031411
0.010	0.12464071	0.11733054
Sample size = 30		
0.990	0.44132930	0.200256350
0.975	0.41341759	0.182619190
0.950	0.38597981	0.170824210
0.900	0.34684355	0.159814950
0.500	0.19372352	0.130842290
0.100	0.12651386	0.107294630
0.050	0.11659583	0.102870090
0.025	0.10982029	0.099837474
0.010	0.10324302	0096860084

(continued)

Table 3.12 (continued)

α	Without trend	Linear trend
Sample size = 40		
0.990	0.475156690	0.160712240
0.975	0.433931310	0.153045430
0.950	0.395416130	0.145912950
0.900	0.344645180	0.137426680
0.500	0.169521970	0.105376760
0.100	0.102161600	0.084344769
0.050	0.093148798	0.079845616
0.025	0.086938805	0.076609429
0.010	0.081393647	0.073462161
Sample size = 60		
0.990	0.528078950	0.162823760
0.975	0.468351880	0.150524540
0.950	0.412710640	0.139364470
0.900	0.345339490	0.125373750
0.500	0.150605630	0.083937329
0.100	0.080174225	0.062058256
0.050	0.071325513	0.058300785
0.025	0.065485408	0.055667655
0.010	0.060465037	0.052867329
Sample size = 70		
0.990	0.549813740	0.165354940
0.975	0.480306630	0.151748190
0.950	0.417638110	0.138643110
0.900	0.344672600	0.123379830
0.500	0.144216790	0.078741199
0.100	0.074349269	0.056075523
0.050	0.065356429	0.052371944
0.025	0.059376144	0.049654444
0.010	0.054097437	0.046988381
Sample size = 80		
0.990	0.569931730	0.171982270
0.975	0.493300620	0.154278010
0.950	0.424566150	0.139022010
0.900	0.346647950	0.121706290
0.500	0.141357440	0.075160709
0.100	0.069921053	0.051766687
0.050	0.060900214	0.047949210
0.025	0.054953008	0.045218561
0.010	0.049560220	0.042535417

(continued)

Table 3.12 (continued)

α	Without trend	Linear trend
Sample size = 90		
0.990	0.587101070	0.175304920
0.975	0.505673320	0.156321150
0.950	0.429490220	0.139885170
0.900	0.344908300	0.121399040
0.500	0.139052120	0.072447229
0.100	0.067168859	0.048548798
0.050	0.057816010	0.044612097
0.025	0.051877068	0.041908084
0.010	0.046780441	0.039386823
Sample size = 100		
0.990	0.594603380	0.177754650
0.975	0.510372830	0.157183470
0.950	0.431164860	0.139652320
0.900	0.343732070	0.120403750
0.500	0.135927460	0.070300302
0.100	0.064217752	0.045987879
0.050	0.055225906	0.042096564
0.025	0.049190208	0.039409235
0.010	0.043797820	0.036908227

3.4.5 Example: Comparison of the DF and KPSS Tests for Several Macro-Economic Time Series

Rao (1995) explains that Dickey et al. (1991) show results concerning integration and co-integration of several macro-economic variables. The data set consists of quarterly observations, starting in first quarter of 1953 and ending in the last quarter of 1988, hence, it covers 36 years with 144 measurements. As usual, testing of integration is an introductory step leading to/co-integration relationship estimation. It was performed with use of the DF test with three augmentations.

The testing is repeated for integration using DF test, and applied the KPSS test to the same data set, with use of Gaussian PDF.

The results for the DF test are given in Table 3.16. They are in perfect agreement with original results of Dickey (1991), where the null hypothesis of a unit root estimation cannot be rejected.

In Table 3.16 the symbol # means that computed value of the KPSS test statistic is greater than critical value for 100 observations.

3.4.5.1 Test of Stationarity Around Mean

For all variables computed KPSS test statistic is greater than the critical value. Hence, the null hypothesis of stationarity around mean is rejected.

Table 3.13 The empirical power of the KPSS test

Variance	Significance level					
	0.99	0.95	0.90	0.10	0.05	0.01
A. Results for model without trend						
The tested null hypothesis is of level stationarity						
Sample size = 80						
0.0	0.00850	0.04900	0.09710	0.89570	0.94740	0.99050
0.1	0.01090	0.05100	0.09890	0.90010	0.94900	0.98890
0.2	0.01060	0.05440	0.10530	0.90410	0.95200	0.99030
0.3	0.00980	0.04710	0.09900	0.89800	0.95000	0.99040
0.4	0.01030	0.04760	0.09980	0.90220	0.95170	0.99000
0.5	0.01090	0.05050	0.10010	0.89960	0.95000	0.99000
0.6	0.01050	0.04970	0.09770	0.89900	0.95060	0.99230
0.7	0.01050	0.04930	0.09810	0.89900	0.94560	0.98940
0.8	0.01200	0.04970	0.09630	0.90280	0.95020	0.99050
0.9	0.00930	0.04780	0.09750	0.90510	0.95080	0.99030
1.0	0.08700	0.04820	0.09800	0.89690	0.94880	0.99140
1.1	0.00780	0.04850	0.09910	0.90380	0.95290	0.99100
1.2	0.00890	0.04690	0.09680	0.90190	0.95220	0.99000
1.3	0.00960	0.04610	0.09500	0.89630	0.94850	0.99210
1.4	0.01130	0.04860	0.09580	0.89610	0.94700	0.98940
Sample size = 90						
0.0	0.00890	0.05060	0.10100	0.89820	0.94910	0.98850
0.1	0.00910	0.04920	0.10110	0.89980	0.94960	0.99070
0.2	0.01200	0.04730	0.09770	0.90270	0.95180	0.99130
0.3	0.01100	0.05180	0.09800	0.90460	0.95370	0.99020
0.4	0.01090	0.05110	0.10230	0.89470	0.94580	0.98940
0.5	0.00960	0.04760	0.09890	0.89760	0.95130	0.99080
0.6	0.00960	0.04740	0.09800	0.90050	0.95370	0.99160
0.7	0.01020	0.04770	0.10340	0.89730	0.94860	0.98930
0.8	0.00990	0.04930	0.09940	0.90080	0.94760	0.98880
0.9	0.00970	0.04650	0.09830	0.89780	0.94720	0.98840
1.0	0.00850	0.04580	0.09850	0.89650	0.94760	0.99100
1.1	0.00910	0.04660	0.09860	0.89680	0.94840	0.98910
1.2	0.01100	0.04830	0.10180	0.90040	0.95150	0.99090
1.3	0.01090	0.04760	0.09450	0.89830	0.94660	0.98830
1.4	0.00730	0.04690	0.09910	0.89640	0.94940	0.98900
Sample size = 100						
0.0	0.00960	0.04880	0.10480	0.90180	0.95040	0.99160
0.1	0.01050	0.04890	0.09820	0.90420	0.95170	0.99020
0.2	0.00950	0.05190	0.10550	0.89960	0.94860	0.99040
0.3	0.01260	0.05310	0.10560	0.90580	0.95360	0.99060

(continued)

Table 3.13 (continued)

Variance	Significance level					
	0.99	0.95	0.90	0.10	0.05	0.01
0.4	0.01060	0.05310	0.10170	0.90810	0.95380	0.99250
0.5	0.00940	0.05000	0.09970	0.89960	0.94960	0.99100
0.6	0.01030	0.04760	0.10010	0.90620	0.95080	0.99120
0.7	0.00850	0.08600	0.09990	0.899930	0.94840	0.99020
0.8	0.01020	0.04780	0.09950	089340	0.94730	0.99040
0.9	0.01170	0.04980	0.10120	090150	0.95240	0.98930
1.0	0.00900	0.04530	0.09410	0.90330	0.95260	0.99120
1.1	0.01010	0.05000	0.10160	0.89680	0.94880	0.98970
1.2	0.00990	0.04860	0.09520	0.90170	0.94810	0.98870
1.3	0.00950	0.04970	0.09970	0.89060	0.94560	0.99040
1.4	0.01130	0.05220	0.10060	0.90070	0.94830	0.99000
B. Results for model with linear trend						
The tested null hypothesis is of stationarity around linear trend						
Sample size = 80						
0.0	0.01020	0.04790	0.09780	0.89440	0.94720	0.99040
0.1	0.00890	0.05220	0.10610	0.89910	0.94910	0.99080
0.2	0.01000	0.04930	0.09960	0.90100	0.95150	0.99030
0.3	0.0960	0.04800	0.09710	0.90220	0.95310	0.98990
0.4	0.01130	0.05260	0.09940	0.90330	0.94910	0.98860
0.5	0.00990	0.5050	0.10160	0.90340	0.95260	0.99050
0.6	0.00930	0.05100	0.10310	0.89920	0.94920	0.99030
0.7	0.01120	0.05240	0.10010	0.89430	0.94680	0.98810
0.8	0.00850	0.04800	0.09970	0.89930	0.94980	0.99180
0.9	0.01090	0.05190	0.10340	0.89910	0.94700	0.99120
1.0	0.00940	0.04940	0.09720	0.90050	0.95170	0.99020
1.1	0.00960	0.05350	0.10340	0.90190	0.95070	0.98920
1.2	0.00960	0.05070	0.10390	0.89810	0.94970	0.98970
1.3	0.00830	0.04810	0.09490	0.90000	0.95260	0.99040
1.4	0.00860	0.04860	0.09730	0.89830	0.94790	0.99070
Sample size = 90						
0.0	0.00910	0.04930	0.09630	0.90010	0.95180	0.99080
0.1	0.00940	0.05090	0.09800	0.89990	0.95210	0.99020
0.2	0.00850	0.05020	0.09980	0.89390	0.94740	0.98890
0.3	0.00870	0.04720	0.09700	0.89960	0.94960	0.98960
0.4	0.00860	0.04740	0.09980	0.90020	0.95240	0.99030
0.5	0.01020	0.05020	0.10280	0.89870	0.94940	0.98880
0.6	0.00940	0.04640	0.09770	0.89910	0.95180	0.99040
0.7	0.00920	0.05030	0.10120	0.90380	0.95090	0.99000
0.8	0.00920	0.05170	0.10620	0.89830	0.95030	0.99120

(continued)

Table 3.13 (continued)

Variance	Significance level					
	0.99	0.95	0.90	0.10	0.05	0.01
0.9	0.01060	0.04790	0.09860	0.89720	0.94430	0.98850
1.0	0.00950	0.05000	0.10070	0.89820	0.95080	0.98950
1.1	0.01060	0.04930	0.10050	0.89940	0.94890	0.98980
1.2	0.01020	0.05210	010230	0.90090	0.94820	0.98920
1.3	0.01010	0.04470	0.09410	0.89860	0.94850	0.98860
1.4	0.00980	0.04860	0.10040	0.90110	0.95190	0.99090
Sample size = 100						
0.0	0.00920	0.04970	0.09870	0.89550	0.94960	0.98960
0.1	0.00880	0.04500	0.09830	0.89470	0.94770	0.98770
0.2	0.00690	0.04890	0.09940	0.90070	0.94990	0.98880
0.3	0.00920	0.05200	0.09970	0.90410	095130	0.99030
0.4	0.00850	0.05070	0.10360	0.89930	0.94790	0.98730
0.5	0.01000	0.04530	0.09580	0.90140	0.94910	0.98940
0.6	0.00940	0.05180	010170	0.90060	0.94910	0.98780
0.7	0.01150	0.05170	0.09980	089440	0.94560	0.98910
0.8	0.00920	0.05250	0.10340	0.90580	0.95070	0.98980
0.9	0.01060	0.04950	0.09730	0.89820	0.94920	0.98860
1.0	0.01250	0.05540	010640	0.90600	0.95450	0.98960
1.1	0.00870	0.04780	0.09920	0.90200	0.95060	0.98890
1.2	0.00980	0.05130	0.09950	0.89640	0.94650	0.98770
1.3	0.01100	0.05300	0.10000	0.89970	0.95010	0.99040
1.4	0.00980	0.05190	0.10170	090190	0.95090	0.98960

3.4.5.2 Test of Stationarity Around a Linear Trend

Only for real money category M1/P and rates of return from 10 Year Government bonds the null of stationarity around a trend cannot be rejected. For all other variables this hypothesis is rejected.

Rao (1995) and Dickey et al. (1991) conclude that both the DF test and the KPSS test give similar results. They stated that all variables can be modeled with use of AR model with trend, for money and rate of return from bonds coefficient of autoregression was smaller than 1, and all other variables have a unit root.

3.5 Parametric Tests

There are several parametric tests in the literature and the most famous one is the classical linear and multiple regression analyses.

Table 3.14 Effect of choice of $\hat{\sigma}_u^2$ value on empirical power of test $\hat{\sigma}_u^2$

Model without trend	Sample size 80	Sample size 90	Sample size 100
α	The value of $\hat{\sigma}_u^2$ in this regression is		
0.99	Significant	Significant	Insignificant
0.95	Significant	Significant	Insignificant
0.90	Significant	Insignificant	Significant
0.10	Insignificant	Significant	Significant
0.05	Insignificant	Insignificant	Significant
0.01	Insignificant	Insignificant	Significant
Model with a linear trend	80 observations	90 observations	100 observations
α	The value of $\hat{\sigma}_u^2$ in this regression is		
0.99	Insignificant	Significant	Significant
0.95	Insignificant	Insignificant	Significant
0.90	Insignificant	Insignificant	Insignificant
0.10	Insignificant	Insignificant	Insignificant
0.05	Insignificant	Significant	Insignificant
0.01	Insignificant	Insignificant	Insignificant

Source Own computations

1. For a fixed significance level α of the KPSS test compute its empirical power for different sample sizes and values of $\hat{\sigma}_u^2$
2. Run a regression of empirical power on sample size and value of $\hat{\sigma}_u^2$
3. Check significance of in this regression $\hat{\sigma}_u^2$

Table 3.15 The results of the Dickey–Fuller test for macroeconomic variables

Ml/P	Real money Ml
M2/P	Real money, M2
MB/P	Real monetary base
NM1M2/P	Part of M2 category outside Ml, real teams
K	Proportion of cash to checkable deposits
KSA	Proportion of cash to checkable deposits, seasonally adjusted
R3M	Nominal percentage rate for 3-month treasury Bills
R10Y	Nominal returns from 10-year government securities
RGNP	Real GNP

Variable	Model with a constant	Model with a constant and a linear trend	Variable	Model with a constant
K	−0.5490	−2.332	Δ_K	−4.223
M2/P	−0.8040	−4.737	$\Delta_{M2/P}$	−10.15
Ml/P	−0.8001	−1.542	$\Delta_{M1/P}$	−3.639
MP/P	0.4109	−2.624	$\Delta_{MB/P}$	−3.048
RGNP	−0.4672	−2.412	Δ_{RGNP}	−6.233
R3M	−2.324	−3.743	Δ_{R3M}	−6.346
R10Y	−1.874	−2.447	Δ_{R10Y}	−5.590
NM1M2/P	−2.156	−1.447	$\Delta_{NM1M2/P}$	−4.029

Table 3.16 The KPSS test statistics for the same variables

Variable	Test with a constant	Test with a trend
K	1.385#	0.2334#
M2 P	1.677#	0.2139#
Ml P	0.5210#	0.1382
RGNP	1.686#	0.2623#
R3M	1.380#	0.1717#
R10Y	1.543#	0.1338
NM1M2/P	1.651#	0.3835#

3.5.1 Regression Analysis

This statistical methodology provides the most reliable relationship between independent, Y, and dependent, X, or even among a single dependent and a set of independent variables, which is then referred to as a multiple regression method. A bivariate relationship has already been presented in Chap.1 by Eq. (1.2) with uncertainty component. Chapter 4 presents detailed procedure for trend identification in a given time series by bivariate regression analysis, where independent variable is time, t, and dependent variable is any variable in the form of time series.

A trend in a given time series can be distinguished mathematically and separated from the whole time series. The mathematical forms of the trends are either a straight-line or low-order polynomial. The graphical representation of a time series with an increasing trend component is shown in Fig. 3.9.

The mathematical expression of such a trend is a function of time according to the following equation,

$$Y = a + bt \tag{3.109}$$

The classical regression methodology end has to pass through the centroid, i.e., arithmetic average values of the two variables $(\overline{Y}, \overline{t})$ which leads to,

$$\overline{Y} = a + b\,\overline{t} \tag{3.110}$$

The second expression can be found after multiplying both sides of Eq. (3.109) by the independent t variable and then taking the arithmetic average of both sides will give,

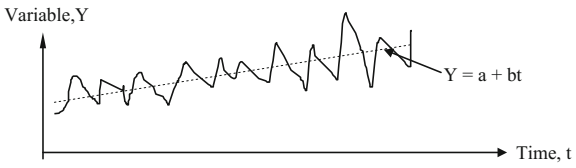


Fig. 3.9 Linear trend

Table 3.17 Trend calculation table

Time (t_i)	Data (Y_i)	t_i^2	$Y_i t_i$
t_1	X_1	t_1^2	$X_1 t_1$
t_2	X_2	t_2^2	$X_2 t_2$
t_3	X_3	t_1^2	$X_3 t_3$
.	.	.	.
.	.	.	.
.	.	.	.
t_n	Y_n	t_n^2	$Y_n t_n$
$\bar{t}$	$\bar{Y}$	$\overline{t^2}$	$\overline{Yt}$

$$\overline{Yt} = a\,\bar{t} + b\,\overline{t^2} \tag{3.111}$$

It is possible to find a and b parameters by a common solution of these two equation. Table 3.17 shows all the necessary calculation steps in column forms. In the first two columns are the time series values of Y and corresponding times, t. Other columns are opened according the requirements of Eq. (3.111). The last row in the two columns includes the arithmetic averages of them.

It is possible to prepare similar table for any nonlinear regression models in Chap. 2 similar to Eq. (3.111).

3.5.2 Regression Line Assumptions

Formally, it is possible to apply the classical least squares estimation of a and b, which lead to regression estimations, where the regression coefficient is equivalent with the first-order serial correlation coefficient. The application of regression approach has six restrictive assumptions as for the parameter estimations.

(1) **Linearity**: Regression technique fits a straight-line trend through a scatter of data points, and correlation analysis test for the "goodness-of-fit" of this line. Clearly, if the trend cannot be represented by a straight-line, regression analysis will not portray it accurately. The unrestricted model below does not require such a restriction, since it is concerned with the variances and arithmetic averages only. In the case of a definite trend, cross-correlation is necessary and it brings the linearity restriction by definition.

(2) **Normality**: It is widely assumed that use of the linear regression model requires that the variables have normal pdf. The requirement is not that the raw data be normally distributed, but the conditional distribution of the residuals should have normal pdf. It is necessary to test if the data have normal pdf. The annual time series have normal pdf, whereas the monthly or seasonal variables have skewed pdf's, and hence, as abide with logarithmic normal, gamma, Weibull, etc. pdf's.

(3) **Means of conditional distributions**: For every value of time series the mean of the differences between the measurement and prediction values obtained by Eq. (3.110) must be zero. If it is not, the coefficients of the regression equation (a and b) are biased estimates. The implication of major departure from this assumption is that there is a nonlinear trend in the scatter diagram.

(4) **Homoscedascity**: This means equal variances in the conditional distributions and it is an important assumption. If it is not satisfied then the regression equation coefficients (a and b) may be severely biased. In order to test for homoscedascity, the data must be subdivided into three or more nonoverlapping parts and the variance of each group is calculated. If there is a significant difference between any of these variances then the data does not have homoscedascity.

(5) **Autocorrelation**: The crux of this assumption is that the value of each observation on the independent time series is independent of all others within the series, so that one cannot predict the value of Y_i at time i, if one knows Y_{i-1} value at time, $i - 1$. There are two interpretations as to the importance of this assumption, one is substantively logical and the other is statistically logical. The statistical interpretation of autocorrelation relates to the linearity assumption.

(6) **Lack of measurement error**: This assumption requires the time series measurements are without error. If this is not the case, and the magnitude of the error is not known, then the coefficients of the regression equation may be biased to an extent that cannot be estimated.

3.5.3 Goodness of Fit (R^2) for Regression

After the regression parameter estimations the goodness of the regression should depend simply on the square of deviations of the data values from the corresponding theoretical regression line. Figure 3.10 indicates regression data scatter with nonlinear regression line and deviations.

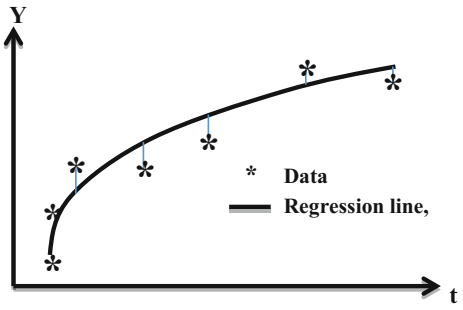


Fig. 3.10 Regression line components

If the data points lie on the regression line without any deviation then the arithmetic mean and the standard deviation of the original data and the regression data sets corresponding to independent variable, t, are exactly the same. Figure 3.10 indicates that there are deviations from the data set, and therefore, although the arithmetic average of the original and regression data sets may be the same, but variances are different. The variance, V_{data}, of original data set is always greater than the regression data set variance, V_{regr}, i.e., $V_{\text{data}} > V_{\text{regr}}$. V_{regr} is referred to also the explained variance, and hence, unexplained variance is $V_{\text{uexp}} = V_{\text{data}} - V_{\text{regr}}$. In case of complete explanation $V_{\text{data}} = V_{\text{regr}}$, which means that $V_{\text{regr}}/V_{\text{data}} = 1$. This is never possible in natural time series, and therefore, the goodness-of-fit quantity, R^2, is defined as follows.

$$R^2 = 1.0 - \frac{V_{\text{regr}}}{V_{\text{data}}} \tag{3.112}$$

R^2 is referred to as the coefficient of determination in the statistics literature. Its value falls in between 0.0 and 1.0, without any unit. The more the R^2 value, the better is the regression curve fitting.

3.5.4 Cumulative Sum (CUSUM) Method

This is a graphical approach for step change identifications in a given time series. It does not require any prior hypothesis or assumption about the occurrence times of the changes, but includes simple plotting with advantages. For instance, any change in slope is identified by the CUSUM graph, and this is comparatively very simple than conventional approaches. Simply the CUSUM variable, C_j, for the jth data value is the summation of deviations from the long term arithmetic average, μ, or the one calculated from the given time series, $\overline{Y}$ as,

$$C_j = \sum_{i=1}^{j} \left(Y_i - \overline{Y}\right) \tag{3.113}$$

This can be rendered into a dimensionless form after dividing it by the standard error of estimation, σ_e, as,

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - Y_{i-1}\right)^2}{2(n-1)}}, \tag{3.114}$$

where n is the sample length. Hence, the standardized CUSUM, C_{jS}, is,

$$C_{jS} = \frac{C_j}{\sigma_e} \tag{3.115}$$

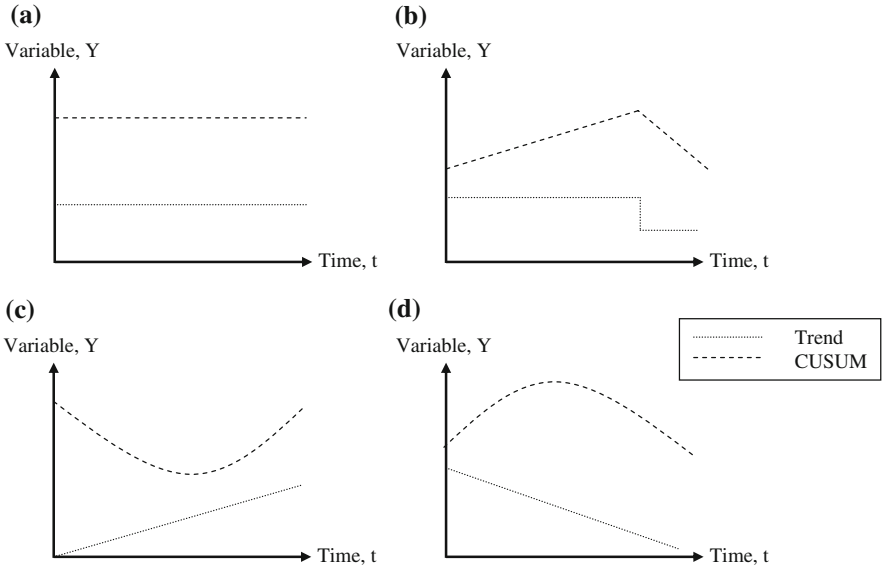


Fig. 3.11 CUSUM trend identification templates

The plot of C_{jS} versus j leads to the CUSUM graph and the vertical difference between the maximum and the minimum values in this graph yields the range, R, and in cases of $R > 10$ a significant statistical change comes into view (see Fig. 2.25). The shape of the CUSUM graph indicates whether there is a trend in the data.

(1) A straight-line indicates no trend (Fig. 3.11a),
(2) An abrupt change in slope indicates a step change in the variable (Fig. 3.11b),
(3) A downwards parabola indicates an increasing linear trend (Fig. 3.11c),
(4) An upwards parabola indicates a decreasing linear trend (Fig. 3.11d).

References

Chernoff, H., & Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Annals of Mathematical Statistics, 35,* 999–1018.
Connover, W. J. (1971). *Practical non-parametric statistics*. New York: John Wiley and Sons.
Craddock, J. M. (1979). Methods of comparing annual rainfall records for climatic purposes. *Weather, 34,* 332–346.
Dickey D. A. (1991). A primer on co-integration with an application to money and income, Federal Reserve Bank of St. Louis, 58–78, Reprinted in Rao (1995).
Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74,* 427–431.
Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica, 49,* 1057–1072.

Dickey, D. A., Dennis, W., Daniel, J., & Thornton, L. (1991). A primer on cointegration with an application to money and income, Federal Reserve Bank of St. Louis, 58–78, reprinted in Rao (1995).

EPA. (1974). Guideline for the evaluation of air quality trends. Office of air quality planning and standards, U.S. Environmental Protection Agency.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed., Vol. I), John Wiley and Sons Co.

Gardner, L. A., Jr. (1969). On detecting changes in the mean of normal variates. *Annals of Mathematical Statistics, 40,* 116–126.

Gilbert, R. O. (1987). *Statistical methods for environmental pollution monitoring*. New York: Van Nostrand Reinhold.

Gomide, F. L. S. (1978). Markovian inputs and the Hurst phenomenon. *Journal of Hydrology, 37,* 23–45.

Helsel, D. R., & Hirsch, R. M. (1992). *Statistical methods in water resources. Studies in environmental science 49*. New York: Elsevier. (Available on-line as a pdf file at: http://water.usgs.gov/pubs/twri/twri4a3/) (Retrieved September 15, 2011).

Himmelblau, D. M. (1969). *Process analysis by statistical methods*. New York: John Wiley and Sons.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers, 116,* 70–808.

Hurst, H. E. (1956). Methods of using long-term storage in reservoirs. *Proceedings of the Institution of Civil Engineers, Part I*, 519–542.

Kanji, G. K. (2001). *100 statistical tests*. New Delhi: Sage Publication, 111"pp.

Kendall, M. G. (1955). Rank correlation methods, Charles Griffin, London.

Kendall, M. G. (1973). *Time series*. London: Griffin.

Kendall, M. G. (1975). Rank correlation methods, 4th edition. Charles Griffin, London, U.K.

Kendall, M. G., & Stuart, A. (1952). The advanced theory of statistics. Vol. 1: Distribution theory. Griffin, London.

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistics* (Vol. II). New York: Hafner.

Khaliq, M. N., Ouarda, T. B. M. J., Gachon, P., Sushama, L., & StHilaire, A. (2009). Identification of hydrological trends in the presence of serial and cross correlations: A review of selected methods and their application to annual flow regimes of Canadian rivers. *Journal of Hydrology, 368*, 117–130.

Koçak, K., & Şen, Z. (1998). Applied examination of dry and wet day occurrences via markov chain approach. *Turkish Journal of Engineering and Environmental Science, 22,* 479–487.

Kolaz, D. J., & Swinford, R. L. (1988). *Ozone air quality: How does chicago rate? 81st annual meeting of the air pollution control association*. Texas, USA: Dallas.

Kolaz, D. J., & Swinford, R. L. (1989). Ozone trends in the greater chicago area. *Ozone Conference on Federal Controls for Ozone Around Lake Michigan, Lake Michigan States' Section and Wisconsin Chapter of the Air and Waste Management Association*.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54, 159–178 (North-Holland).

Lettenmaier, D. P. (1976). Detection of trends in water quality data from records with dependent observations. *Water Resources Research, 12*(5), 1037–1046.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13*(3), 245–259.

Mills, T. C. (1993). *The econometric modelling of financial time series*. Cambridge: Cambridge University Press.

Montgomery, R. H., & Reckhow, K. H. (1984). Techniques for detecting trends in lake water quality. *Water Resources Bulletin, 20*(1), 43–52.

Parzen, E. (1960). Modern probability theory and its applications. John Wiley & Sons, New York.

Phillips, P. C. B., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika, 75*, 335–346.

Qwen, D. B. (1962). *Handbook of srtatistical tables*. Reading, Massachusetts, USA: Addison-Wesley.

Rao, B. B. (1995). *Cointegration for the applied economist*. London: Macmillan.

Saldarriga, J., & Yevjevich, V. (1979). Application of run-lengths to hydrologic time series. Hydrology Paper 40. Colorado State University, Fort Collins, USA.

Searcy, J. K., & Hardison, H. H. (1960). Manual of hydrology: Part 1, general surface-water techniques double-mass curves. U.S. Geological Survey, Water Supply Paper 1541-B.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association, 63*(324), 1379–1389.

Şen, Z. (1974). Small sample properties of stationary stochastic processes and hurst phenomenon in hydrology. *Unpublished Ph. D. Thesis*, *University of London*, 256 pp.

Şen, Z. (1976). Wet and dry periods of annual flow series. *Journal of the Hydraulics Division*, 102, 1503–1514 (ASCE Proceedings of the Paper, 12497).

Şen, Z. (1978). Autorun analysis of hydrologic time series. *Journal of Hydrology, 36,* 75–85.

Shahin, et al. (1993). In D. Machiwal & M. K. Jha. *In hydrologic time series analysis: Theory and practice* (p. 301). Springer Publisher.

Sneyer, R. (1992). On the use of statistical analysis for the objective determination of climate change. *Meteorologishe Zeitshrifft, N.F.*, 247–256.

Sonali, P. & Kumar, D. N. (2013). Review of trend detection methods and their application to detect temperature changes in India. *Journal of Hydrology, 476*, 212–227.

Spearman, C. (1940). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101. doi:10.2307/1412159.

Sweitzer, T. A., & Kolaz, D. J. (1984). An assessment of the influence of meteorology on the trend of ozone concentrations in the chicago area. In *Air Pollution Control Association Specialty Conference on "Quality Assurance in Air Pollution Measurements," Boulder, Colorado, USA*.

Syczewska, E. M., (2010). *Empirical power of the Kwiatkowski-Phillips-Schmidt-Shin test*. Working Paper No. 3–10, Department of Applied Econometrics website at: http://www.sgh.waw.pl/instytuty/zes/wp/.

Wallis, J. R., & O'Connell, P. E. (1973). Firm reservoir yield—How reliable are historic hydrological records? *Hydrological Sciences Bulletin, 18,* 347–365.

WMO (World Meteorological Organization). (1966). Climate change. Technical Note No. 79, WMO No. 195, TP 200, I-20, WM, Geneva, Switzerland.

Yevjevich, V. (1967). *An objective approach to definition and investigation of continental hydrologic droughts*. Hydrology Paper 23. Colorado State University, Fort Collins, USA.

Yevjevich, V., & Jeng, R.J. (1969). *Properties of non-homogeneous hydrologic series*. Hydrolology Paper 32, Colo. State Univ., Fort Collins, USA.

Yue, S., & Wang, C. (2004). The Mann–Kendall test modified by effective sample size to detect trend in serially correlated hydrological series. *Water Resource Manage,* 18, 201–218.

Yue, S., Pilon, P., & Phinney, B. (2004). Canadian streamflow trend detection: impacts of serial and cross-correlation. *Hydrogical Sciences Journal, 48*(1), 51-64.

Temporal Trend Analysis

4

Abstract

There are various regression analyses in the literature concerning temporal trend analysis in a classical manner, which have their application domains in various contexts. Among these are the statistical conventional regression methodologies as explained to a certain extent in the previous chapter, unrestricted regression, and partial regression, and cluster regression methodologies. Detailed explanatory information is presented for each one of these methodologies explaining the basic requirements for their applications. In the meantime, there are some new concepts such as the trend polygons that are applicable for distinction between different time intervals such as months and associated trend components during transition from one time step to another.

Keywords

Cluster · Partial polygon · Regression · Temporal · Trend · Unrestricted

4.1 General

Trend identification in time series data is one of the major tasks for long-term changes and their impacts on various human activities including natural, social, economic, agricultural, climate, and engineering management systems. Practically, simple trend identifications are achieved by moving average procedure, which smooths fluctuations around a trend or periodicity component even without any mathematical expression for the trend. Theoretically, Mann–Kendall (MK) trend test is applied to a given time series for possible trend existence with subsequent application of Sen's slope mode calculation and then, through the classical regression analysis trend identification is achieved (Chap. 3). These methodologies

Z. Şen, *Innovative Trend Methodologies in Science and Engineering*,
DOI 10.1007/978-3-319-52338-5_4

exploit the whole time series and do not make any categorical distinction, say, among "low", "medium" and "high" values in search of partial trends, but they are more for monotonic trend determination. This section suggests temporal trend analysis where trends are sought with respect to some baseline either within the same time series or between two or more records. The former is referred to as serial-trend and the latter is cross-trend identifications.

Natural phenomena evolve by time and the main question is whether they are free of trend or there are time intervals during which increasing (decreasing) trends take place. It is serially possible, to compare the present-day conditions with respect to some previous nonoverlapping intervals (Şen 2010). For instance, one may compare a set of recent years' serial summer (annual, decadal) climatic situation with another previous years' serial summer (annual, decadal) climate appearance each after in ascending sorting. The last sentence implies that "low" ("medium", "high") values are comparable with "low" ("medium", "high") values. A given time series can be considered in terms of subseries and they can be compared with each other so as to appreciate whether an increasing (or decreasing) trend takes place in recent time period compared to previous periods. In such comparisons, human does not think about the absolute sequential appearances within each subseries but with their orders. In the statistical sense, order irrelevance leads one to think about nonparametric methodology, which constitutes the basis of the serial- and cross-trend identification procedures in this chapter. Such a procedure is not concerned with absolute time but its reasoning is based on relativistic bases, where one subseries is compared with respect to another subperiod of the same time series for the trend assessments. Consideration of subseries leads to better and finer interpretations, suggestions, and conclusions about the trends of various durations.

Nonparametric Mann–Kendall (Mann 1945; Kendall 1975) statistical test has been used for the last several decades in search for trends in past time series records to understand the environmental changes such as water pollution, climate change, and global warming (Chap. 3). Man–Kendall (MK) test assumes that the time series has independent serial structure, since dependence (positive serial correlation structure) may lead to trend detections in the absence of trends. Although a pre-whitening procedure application is proposed prior to MK test application for rendering the original series into independent series (von Storch 1995), it has been shown by Douglas et al. (2000) that such a pre-whitening may lead to trend detection that is less than the original series. A detailed account of pre-whitening procedure prior to trend detection has been presented by Yue and Wang (2002). They concluded that pre-whitening is not suitable for estimating the effects of serial correlation on the MK test when trend exists within a time series. Various simulation studies have been proposed for pre-whitening, but still one cannot be certain about the trend identification precision (Bayazit and Önöz 2007). Block bootstrap MK is a modified form of the MK test devised for serially correlated samples. Its Type I and Type II errors are investigated by a simulation study. It is found that the rejection rate of the hypothesis of no trend approaches the nominal significance level if the block length of bootstrap samples is chosen properly depending upon the sample size and lag-one autocorrelation coefficient (Önöz and Bayazit 2012).

Although the classical temporal trend tests have been useful and can be applied in the future, they can be criticized because of the following deficiencies:

(1) They search for temporal monotonic trend component mostly in holistic and monotonic manners without any distinction between "low", "medium" and high values. Additionally, one is not able to determine the subsection duration to compare it with the same length desirable duration at some other part of the given time series,
(2) The search is in the absolute time domain, which takes into account real-time sequence of the series, and hence, the serial correlation coefficient becomes important,
(3) The correlation structure of given time series is either assumed independent so that the classical trend tests can be applicable or rendered into an independent serial structure through pre-whitening procedure. However, the pre-whitening procedure disturbs the original structure of the time series.

The study by Yue et al. (2002) searched for the power of MK and Spearman rho tests through Monte Carlo simulations and found that the power of these tests depends on the preassigned significance level, trend magnitude, sample size, and the serial correlation of a time series. The bigger the absolute magnitude of trend, the more powerful is these tests at large sample sizes provided that the serial autocorrelation is negligibly small. Fatichi et al. (2009) mentioned about any increase in uncertainty when pronounced stochastic behaviors are present in the data.

The main purpose of this chapter is to present classical methodologies that are used for identification of temporal trends with applications and interpretations. Subsequently, pre-whitening and over-whitening procedures are explained for the satisfaction of the basic assumption of independence in a given time series so as to be able to apply the MK test.

4.2 Visual Inspection

It is advised in this section that prior to any theoretical methodology application for trend detection in a time series, one should plot the time series and then try to explore with necked eyes whether there may be trend embedded into the data structure. For this purpose, several replicates are presented in Fig. 4.1. The reader may look at them carefully and try to identify any trend component visually as much as possible. This will give him/her confidence experience and insight gain qualitatively for trend identification.

Trend analyst should evaluate his/her data qualitatively by different visual inspections as for the graphical presentations of the time series at hand, its internal (serial) structure as correlation, periodicities, visibility of possible trends and according to all the inspection then she/he can decide on a single or the two most

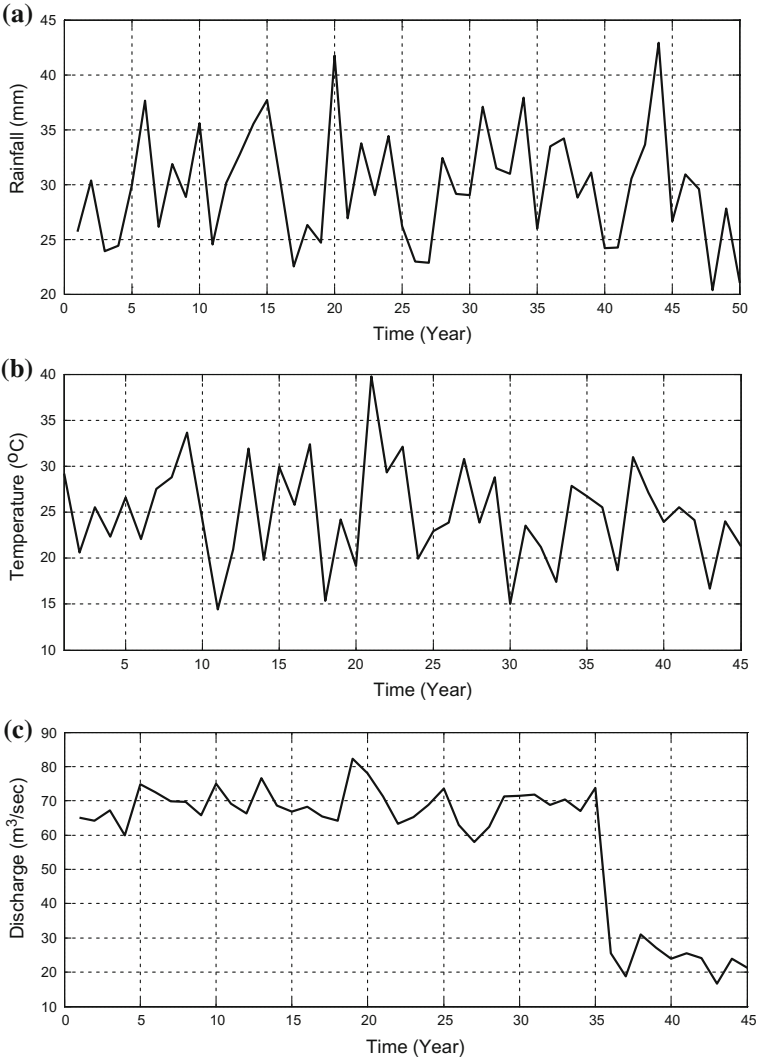


Fig. 4.1 Trend possibility time series

convenient statistical trend analyses methodologies that are suitable for the qualitative characteristics of the records. In an efficient time series, the sample size must be long enough, statistically at least 30 data records and preferably as long as possible, without gaps or with few gaps, consistent data measurements (without depreciation or change of the instrument except with calibration). In cases of monthly data, it is advised to have 10-year of records.

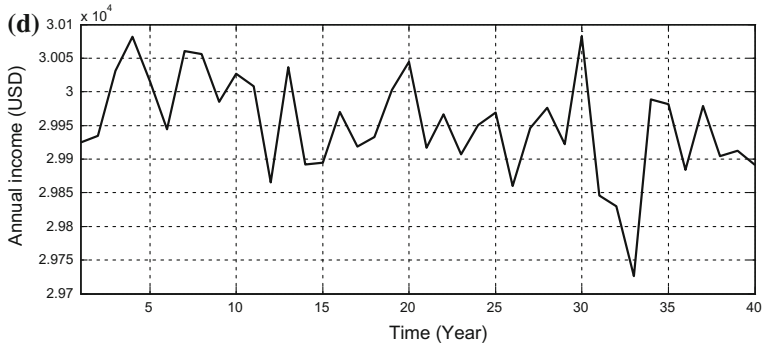


Fig. 4.1 (continued)

Most often trend analysis implies a single, unidirectional, and gradual change of monotonic component within given time series irrespective of partial or step changes. Depending on the type of project, environment impacts, and treatment one can sense possible trend expectation. It is advised that trend analyst should also consult other experts and even the local people who are concerned or affected by the trend causal effects. These qualitative knowledge and information, prior to trend methodology application, are very precious, especially, for the identification of step and partial trends. For instance, when abrupt changes, as in Fig. 4.1c, are within the time series then either partial trend identification methodology or step trend methods can be employed (Chap. 8).

It is not recommended that one should indulge in trend analysis without preliminary visual and exploratory inspection of the circumstances at the time series record locations, because such searches pave way to preliminary trend indications and provide strategy for proper and convenient trend analyses. During this inspection, the necessary conditions of assumption satisfactions must be searched, and especially, the probability distribution function (pdf) type as normality (Gaussian), variance constancy (homoscedasticity), and serial independence cases must be evaluated in an effective manner. If necessary, convenient transformations (pre-whitening, over-whitening, logarithmic, square root, or cubic root) should be applied.

4.3 Monotonic Trend Analysis

There are well-established methodologies for identification of monotonic trends with reliability provided that the basic assumptions are satisfied completely or to a significant extent. It is known that the statistical trend analysis is equivalent to hypothesis testing by null and alternative hypotheses. The null hypothesis is valid in case of no trend existence and it can be checked through different parametric tests concerning each method. If the null hypothesis is rejected, then one understands

that there is trend component in the given time series, but it does not yield the type of trend in increasing or decreasing direction, which needs additional method for its mathematical determination.

Monotonic trend analysis methodologies are either nonparametric, parametric, or mixture of the two. Although parametric tests are the most powerful alternatives, they need satisfaction of the basic assumptions that are rather difficult to encounter in natural time series. Especially, small sample sizes are the major problems in the methodological applications. The first gateway for the application of the parametric methods is that the time series should have normal (Gaussian) pdf. Parametric and nonparametric tests require that the time series has independent serial structure and homoscedasticity (constant variance). When the normality, constancy of variance, and serial independence assumptions are valid then the parametric method of regression line fitting becomes preferable application for trend identification. On the other hand, nonparametric methods are preferable in cases that the pdf of time series is nonnormal, existence of data gaps, and robustness against outliers.

Among the parametric methodologies in addition to the regression analysis are multiple linear regression, periodic functions, and others. Nonparametric methods are Mann–Kendall (MK), seasonal Kendall, and others. Among these techniques classically the most widely applied ones are the linear regression, MK, and seasonal Kendal procedures (Chap. 3).

Multiple linear regression method accounts for the effects of different variables on a responsive variable. It includes especially covariates in trend analysis in a single monotonic unidirectional manner. The simplest form of such a model has the following mathematical structure:

$$Y = \alpha + \beta t + \gamma X + \varepsilon, \tag{4.1}$$

where Y is the dependent variable, X is the independent variable, t is time, u is the random variable, and finally α, β, and γ are linear trend model parameters. In this expression, the null hypothesis, H_0, for trend analysis is that $\beta = 0$ and the t-statistics test can be used. On the other hand, if γ is not significantly different from zero, then the simple regression technique becomes applicable, because there is no effect of independent variable, X, on the dependent variable, Y. The best trend line implies that the residuals (error terms) are serially independent, has constant variance, and they have normal (Gaussian) pdf.

4.4 Scatter Diagrams and Regression Model

If the question is the search for the type of relationship between two variables, the practical answer can be given by plotting the corresponding values of two time series against each other on a Cartesian coordinate system. Consequently, the two time series data values give rise to scatter points as presented in Fig. 4.2, which provides visual inspection, preliminary feeling, and consideration of the

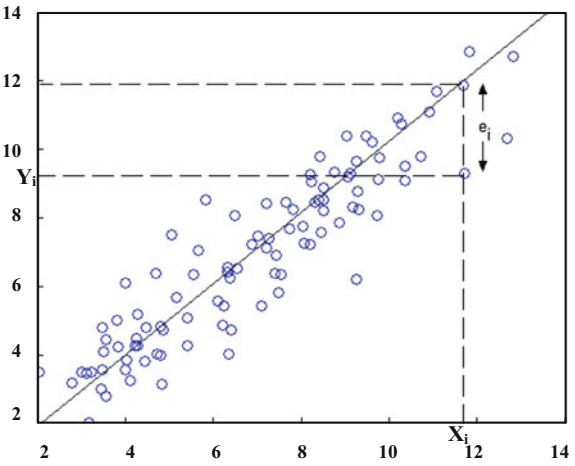


Fig. 4.2 Scatter diagram

relationship type between two variables. Such coordinate systems with data points are referred to as the scatter diagram in the statistics terminology.

Comparison of the scatter diagram with the functional relationships either in Fig. 4.3 or in any mathematics textbook visually provides the first opinion about the type of deterministic relation form (mathematical function), which shows the general trend between the two time series values.

The simplest of the possible relationships is the straight-line form, and it is frequently used in different disciplines.

$$Y = a + bX \tag{4.2}$$

This model is referred to as the simple regression, since X is regressed on Y. In any actual prediction model, there is more than one predictor, but the ideas for simple linear regression can be generalized easily to multiple linear regression. The representation of Eq. (4.2) on a Cartesian coordinate system yields to a straight line in the mathematical sense, but scatter of points in the statistical sense, where each one of these points is associated with the data pairs (X_i, Y_i) for n data pairs ($i = 1, 2, \ldots, n$). The relative position of the straight line must be determined in such a way that the summation of squared-deviation of each point from this straight line is the most possible smallest. Herein, the deviation is synonymously used as the error. As shown in Fig. 4.2, these deviations from the straight line may be decided as the horizontal, vertical, or perpendicular distances. However, in practical studies most often sum of vertical deviation squares are minimized to fix the regression line through the scatter diagram. The choice of the sum of the squared-error criteria is convenient not only that it is necessarily for the best model, but also it is mathematically tractable for the differentiations. The scatter diagrams are very significant for the identification of functional relationship between the variables. In any functional relationship such as in Eq. (4.2), there are parameters such as a and b. Herein, the Y variable (predictand)

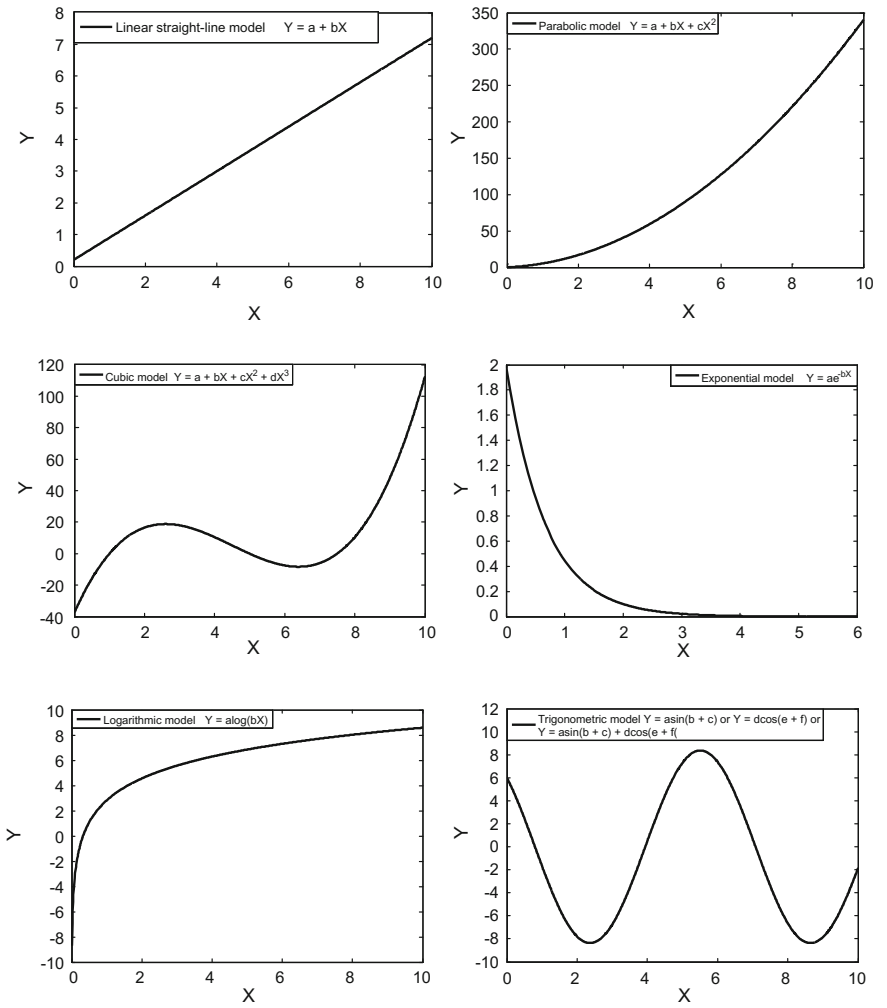


Fig. 4.3 Mathematical functions

is predicted from X predictor variable. After the decision of visual best relationship form, it is important to determine or estimate the model parameters from the available time series data values. The principle in the estimation is that the error (between any predicted, $\hat{Y}_i$ and measured, Y_i) sum squares value is the minimum. The parameters are dependent on the time series data, which can be expressed implicitly as

$$a = f_1(X, Y) \tag{4.3}$$

and

$$b = f_2(X, Y) \tag{4.4}$$

Finally, the whole question is now how to obtain the explicit formulation of the model parameters. For this purpose, the regression procedure is used which will be explained in the following sequel.

4.5 Linear Regression Model

In this section, only straight-line model is considered. The regression method is about the search procedure for the explicit expressions of Eqs. (4.3) and (4.4). In order to grasp the question conceptually, let Eq. (4.2) be written for the i-th data pairs as,

$$Y_i = a + bX_i \tag{4.5}$$

In Fig. 4.4, different straight-line forms are shown for different sets of model parameters.

In Fig. 4.4a there is quite steep slope. This implies that b parameter in Eq. (4.4) has a big positive value. In this manner, increase in X causes increase in Y. However, in Fig. 4.4b the situation is just the opposite and increase in X value gives rise to decrease in Y. Hence, b parameter expresses the slope of the straight line, and it is represented geometrically in Fig. 4.4d. For some β distance along the X axis, the corresponding vertical distance on the Y axis is α, and hence, the slope can be expressed as

$$b = \frac{\alpha}{\beta} \tag{4.6}$$

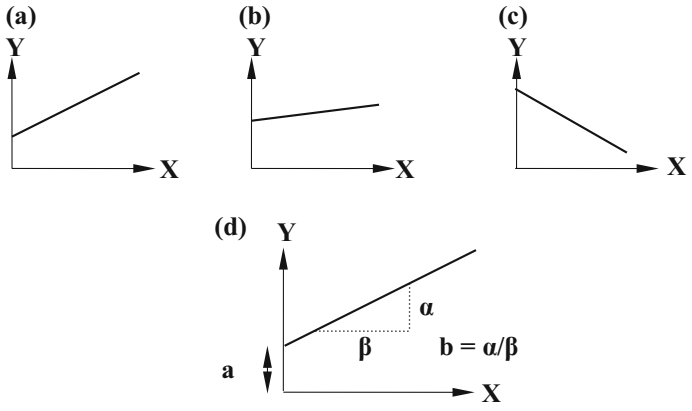


Fig. 4.4 Straight lines for different parameter sets

If $\beta = 1$, then the α value on the Y axis gives directly the slope of the straight line which is equal to b. This parameter is called as the regression coefficient. By definition, it is the change in dependent Y variable corresponding to each unit increment in X independent variable. In Fig. 4.4c, a corresponds to the ordinate of intercept point on the Y axis. Up to now, the regression parameters (a and b) are explained in a mathematical manner completely independent of time series data. It is already stated that these parameters should be determined such that the sum of deviations, i.e., error squares is minimum (Chap. 3).

4.5.1 Statistical Procedure

The straight-line parameters are estimated from the best model fitting through the scatter diagram shown in Fig. 4.5a. This means that the fitted straight line must be "close as much as possible" to overall scatter points. The statement "as much as possible" implies that the variance of the points from the straight line must have its minimum value. In general, in any classical straight-line model fitting, the deviations are adopted as the vertical errors that are parallel to Y axis as shown in Fig. 4.5a. Hence, the minimization of the total sum of error squares based on n data points can be expressed mathematically as,

$$\text{Min} \sum_{i=1}^{n} \left(\hat{Y}_i - Y\right)^2, \tag{4.7}$$

where $\hat{Y}_i$ shows the predicted Y value on the straight line corresponding to i-th independent data, X_i. The expression in Eq. (4.7) is known as the least square

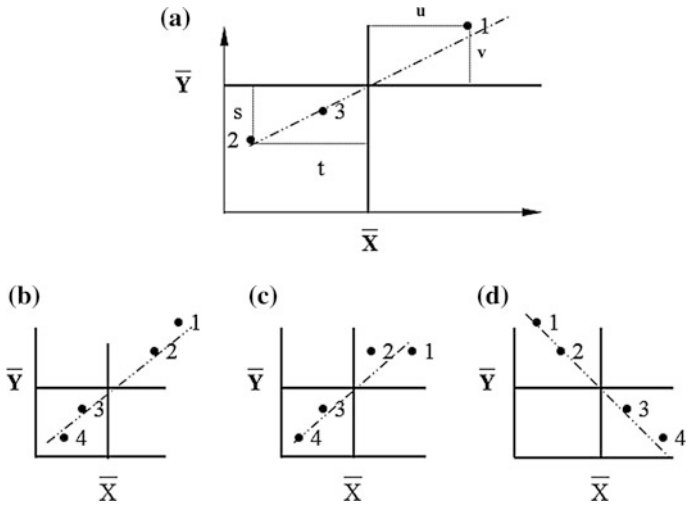


Fig. 4.5 Various deviations

procedure in the regression methodology. The main subject in any regression procedure is the relationship between the variances of dependent and independent variables (Y_i and X_i). With this information, let us concentrate on various Y_i and X_i points in Fig. 4.5. For better understanding, after the arithmetic averages, $\overline{X}$ and $\overline{Y}$ of the two variables, the contributions, u, v, s, and t deviations of points 1 and 2 to the overall variances, namely, S_X^2 and S_Y^2 are shown in Fig. 4.5a.

In Fig. 4.5a, the contribution from point 1 to S_X^2 and S_Y^2 are u^2 and v^2, respectively. In this manner, points 1 and 2 contributes significantly to variances S_X^2 and S_Y^2 because of their comparatively far away locations from the arithmetic averages $\overline{X}$ and $\overline{Y}$ whereas point 3 has very little contribution. In order to commonly account for these contributions collectively, it is necessary to develop the concept of covariance. In general, covariance is defined as the average value of products of deviations from the averages. Hence,

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \tag{4.8}$$

For instance, in Fig. 4.5a, the covariance contributions of points 1 and 2 are,

$$\left(X_1 - \overline{X}\right)\left(Y_1 - \overline{Y}\right) = uv$$
$$\left(X_2 - \overline{X}\right)\left(Y_{21} - \overline{Y}\right) = ts$$

respectively. For these two points from Eq. (4.8), the covariance can be expressed as,

$$\mathrm{Cov}(X,Y) = 0.5(uv + ts)$$

In Fig. 4.5b, c and d, the regression line slope appears as the ratio between this covariance and the variance, S_X^2, of independent variable as,

$$b = \frac{\mathrm{Cov}(X,Y)}{S_X^2} \tag{4.9}$$

It is obvious from Fig. 4.5 that the overall regression line crosses through the weight point $\left(\overline{X},\overline{Y}\right)$ of all X_i and Y_i data. With this information at hand, the intercept point ordinate of the straight line on the Y axis, the parameter a_{YX} can be easily calculated leading to,

$$\overline{Y} = a + b\overline{X} \tag{4.10}$$

and

$$a = \overline{Y} - b\overline{X} \tag{4.11}$$

The regression parameters must be found on the basis of "the least sum of error squares". For this purpose, the basic straight line can be rewritten for the i-th data point by taking into consideration the error term h_i as,

$$Y_i = a + bX_i + e_i \tag{4.12}$$

Hence, the error for i-th data value becomes as,

$$e_i = Y_i - (a + bX_i) \tag{4.13}$$

Furthermore, the error square is,

$$e_i^2 = [Y_i - (a + b\,X_i)]^2 \tag{4.14}$$

Finally, the sum of error squares over all the available data becomes,

$$H_T = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2 \tag{4.15}$$

In order to minimize this expression, mathematically, it is necessary to take the partial derivatives with respect to unknowns (herein the unknowns are a and b) and then equated to zero as follows:

$$\frac{\partial H_T}{\partial a_{YX}} = \sum_{i=1}^{n} 2[Y_i - (a + bX_i)](-1) = 0 \tag{4.16}$$

and

$$\frac{\partial H_T}{\partial b_{YX}} = \sum_{i=1}^{n} 2[Y_i - (a + bX_i)](-X_i) = 0$$

and after the simplification,

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i$$

and

$$\sum_{i=1}^{n} Y_i X_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2$$

Division of these equations by the number of data, n, leads to expressions that can be written in terms of arithmetic averages as,

$$\overline{Y} = a + b\overline{X} \tag{4.17}$$

and

$$\overline{YX} = a\overline{X} + b\overline{X^2} \tag{4.18}$$

It is obvious that Eq. (4.17) is equivalent to previously obtained Eq. (4.10) and on the other hand, substitution of a from Eq. (4.17) into Eq. (4.18) leads after the necessary algebraic manipulations to,

$$b = \frac{\overline{YX} - \overline{X}\,\overline{X}}{\overline{X^2} - \overline{X}^2} \tag{4.19}$$

which should have the similar interpretation with Eq. (4.8).

4.6 Unrestricted Regression Model

It has been suggested by Şen (2001) that since there are two parameters in a linear regression model, two conditions are sufficient for their estimations from a given set of data. Without any procedural restrictive assumptions first the average and then the variance of both sides in Eq. (4.10) lead to

$$\overline{Y} = a' + b'\overline{X} \tag{4.20}$$

and

$$\mathrm{Var}(\overline{Y}) = b'^2 \mathrm{Var}(\overline{X}), \tag{4.21}$$

where a' and b' are defined as the intercept and slope parameters of a restricted regression equation, respectively. Herein, for distinction unrestrictive model parameters are shown as a' and b', respectively. These two equations are the basis for conservation of the arithmetic mean and variances of dependent and independent data. The basic equation remains unchanged whether restrictive or unrestrictive model is used. Equation (4.20) implies that in both models, the centroid, i.e., averages are equally preserved. Furthermore, another implication from this statement is that both models yield close estimations around the centroid. The deviations between the two model estimations appear at independent and dependent time series values away from the arithmetic averages. The simultaneous solution of Eqs. (4.20) and (4.21) yields parameter estimates as

$$b' = \sqrt{\frac{\mathrm{Var}(\overline{Y})}{\mathrm{Var}(\overline{X})}} \tag{4.22}$$

and

$$a' = \overline{Y} - \sqrt{\frac{\mathrm{Var}(\overline{Y})}{\mathrm{Var}(\overline{X})}}\overline{(X)} \tag{4.23}$$

Physically, variations in the dependent variable data are always smaller than the independent variable data, and consequently, $\mathrm{Var}(\overline{X}) \geq \mathrm{Var}(\overline{Y})$. Under the light of Eq. (4.22) always $0 \leq b' \leq 1$. Furthermore, Eq. (4.22) is a special case of Eq. (4.21) when $r_{hs} = 1$. The same is valid between Eqs. (4.22) and (4.23). In fact, from these explanations, it is clear that all the restrictive assumptions bias effects are represented globally in r_{hs} which does not appear in the unrestrictive model parameter estimations.

Mathematically, the second term in Eq. (4.23) is always smaller than the first one and hence a' is always positive. The following relationships are valid between the restrictive and unrestrictive model parameters:

$$b' = \frac{b}{r_{hs}} \tag{4.24}$$

and

$$a' = \frac{a}{r_{hs}} - \left(1 - \frac{1}{r_{hs}}\right)(\overline{Y}) \tag{4.25}$$

These theoretical relationships between the parameters of the two models imply the following points. Since b and b' are the slopes of the straight lines, the restricted equation slope is smaller than the unrestricted approach ($b < b'$) according to Eq. (4.24) since always $r_{hs} > 0$ for the dependent and independent data scatter on a Cartesian coordinate system (see Fig. 7.1). It has already been said above that the two methods coincide practically around the centroid. This further indicates under the light of the previous statement that unrestricted model yields overestimates compared to the restricted estimations for dependent data greater than the average value but underestimation of dependent (4.25) shows that $a' > a$. Furthermore, the summation of model parameters is

$$a' + b' = \frac{a+b}{r_{hs}} - \left(1 - \frac{1}{r_{hs}}\right)\overline{Y} \tag{4.26}$$

These last expressions indicate that the two approaches are completely equivalent to each other only for $r_{hs} = 1$. Otherwise, unrestricted model parameter estimations are greater than the corresponding restricted regression coefficients.

4.6.1 Application

The determination of the parameters in Eq. (4.20) by means of unrestricted regression approach parameter estimations are represented for three global radiation estimations for Istanbul and Ankara, Turkey, stations. In the classical regression methodology as explained in the previous section, the data must abide by normal pdf, which is not the case for global solar irradiation records, and therefore, unrestricted regression approach is used for parameter estimations. The frequency distribution functions are overwhelmingly positively skewed. However, normality in the frequency distribution function implies validity of the regression coefficient in the classical least squares technique. In normal or nearly normally distributed date cases, there are minor differences between the estimates of the classical and unrestricted regression approaches. The smaller is the scatter around a straight line, the smaller the difference between the two methods parameter estimations. In natural events, the smaller the averaging time period (smaller than year as hours, days, weeks, months, seasons) the more the deviation from normality. Equation (4.20) parameter estimations by the classical least squares technique remain in bias. Parameter estimations according to restricted and unrestricted models are given in Table 4.1.

In Fig. 4.6 the two Angström straight lines obtained separately from the unrestricted and classical regression approaches are presented for Ankara station. It is noticed that both straight lines pass through the centroid $(\overline{X}, \overline{Y})$ of the scatter diagram.

Table 4.1 Regression and unrestricted method parameter estimations

Station name	Restricted		Unrestricted	
	a	b	a'	b'
Ankara	0.311	0.323	0.376	0.355
Istanbul	0.295	0.354	0.273	0.393

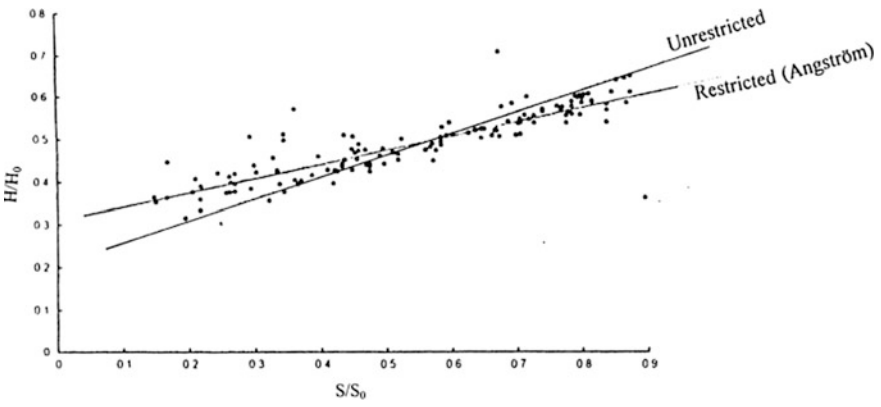


Fig. 4.6 Straight lines by the classical (restricted) and unrestricted regression methods

4.7 Partial Regression Method (PRM)

This model is a refinement of classical regression model as has been used by many researchers. Equation (4.20) is a dimensionless expression where parameters a and b are the regression line intercept with the Y axis and the slope of the straight line. Provided that simultaneous data on Y and X are available, a and b model parameters can be determined by use of the statistical regression approach. However, such an approach has the following some drawbacks:

(1) Since the whole data are processed the overall Y and X data averages are used in the parameter estimations.
(2) For both variables global variances are used without considering the variance variation (homoscedasticity) in the variation domain of Y or X.
(3) Once parameters a and b are estimated from the data, their substituted into Eq. (4.20) is for prediction of Y value without considering the variation range of this estimate. At best that can be done is to attach the upper and lower confidence limits to the prediction but it will again be dependent on the global variance.

The PRM method overcomes these drawbacks with more flexible and realistic dependent variable prediction. Although the same amount and type of data are necessary for the application of PRM approach similar to other regression methods, the former is more dynamic by taking into account the variation ranges that are likely to occur in the arithmetic average and variance values. In such a manner, the constancy of the variance in the classical regression method is avoided and better predictions with different confidence limits can be obtained according to the value of predictant. The PRM provides dependent variable predictions on the basis of partial averages and variances, and in this manner the variabilities in the data parameters are taken into consideration.

For instance, Fig. 4.7 shows the scatter diagram of an independent variable, X, versus Y. It is obvious that there are some outliers especially in the lower part of the scatter diagram, and they are eliminated from further consideration. The scatter diagram does not have uniform variance (homoscedasticity).

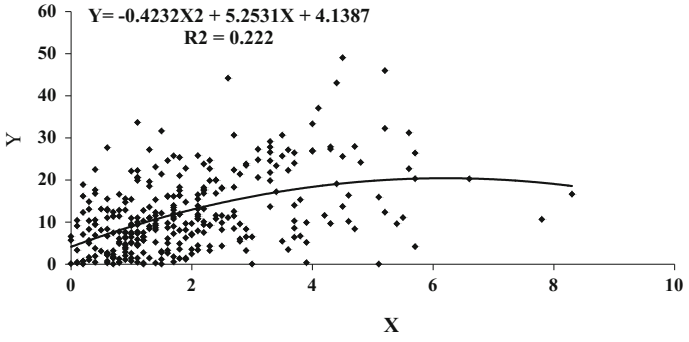


Fig. 4.7 Scatter diagram

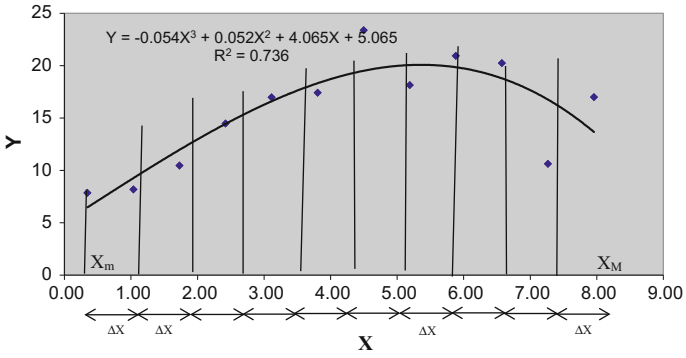


Fig. 4.8 PRM procedure

Hence, the classical regression technique with its variance constancy assumption cannot be valid for the modeling of relationship between dependent and independent variables. It is, therefore, necessary to search for a new model and this justifies the suggestion of PRM.

For better explanation of this model, it is necessary to subdivide the variation domain of the independent variable into nonoverlapping adjacent partial divisions as shown in Fig. 4.8.

Selection of small number of subinterval, ΔX, will average the data into coarse intervals with loss of information, and on the other hand, big interval number will get the procedure closer to the classical regression case and one should be then forced to assume that the variance is constant. Provided that the maximum and minimum data values are X_M and X_m, respectively, then with n number of subdivisions the subclass length, ΔX can be expressed as,

$$\Delta X = (X_M - X_m)/n \tag{4.27}$$

Accordingly the limits of each subdivisions are from the lowest subdivision, X_m to $X_m + \Delta X$; the second one is from $X_m + \Delta X$ and $X_m + 2\Delta X$; and, finally the last subdivision limits is from $X_m + (n - 1)\Delta X$ to X_M. It is possible to arrange this in a table form as in Table 4.2.

In this table, the summation of points in each subdivision is equal to the number of available data. It is also necessary that the mean of the subdivision means is equal to the overall mean value of the variable concerned. The representative value of subdivision is taken as the midpoint of each subdivision. For instance, the third subinterval has m_3 as the average of the subinterval lower and upper limits.

$$\begin{aligned} m_3 &= (X_m + 2\Delta X + X_m + 3\Delta X)/2 \\ &= X_m + 2.5\Delta X \end{aligned} \tag{4.28}$$

Hence, in general the representative of r-th subinterval is

Table 4.2 Subdivision characteristics

Number	Subdivision					
	Lower limit	Upper limit	New dependent variable	Number of data	Mean	Variance
1	X_m	$X_m + X$	Z_1	n_1	m_1	V_1
2	$X_m + \Delta X$	$X_m + 2\Delta X$	Z_2	n_2	m_2	V_2
3	$X_m + 2\Delta X$	$X_m + 3\Delta X$	Z_3	n_3	m_3	V_3
…	…	…	…	…	…	…
…	…	…	…	…	…	…
$n - 1$	$X_m + (n - 2)\Delta X$	$X_m + (n - 1)\Delta X$	Z_{n-1}	m_{n-1}	m_{n-1}	V_{n-1}
n	$X_m + (n - 1)\Delta X$	$X_m X$	Z_n	m_n	m_n	V_n

$$m_r = X_m + (r/2)\Delta X \tag{4.29}$$

The corresponding dependent variable average for this subinterval Z_r (r = 1, …, n) will be the average value of all the data points that fall within this subdivision as,

$$Z_r = (1/n_r)\sum Y_i, \tag{4.30}$$

where n_r is the number of data within the r-th subdivision. The calculations for Figs. 4.7 and 4.8 data are given in Table 4.3.

Table 4.3 PRM data presentation

No.	X		Y			Relative error (%)
	Class intervals	Interval mid-points	Average	Std. dev.	Prediction	
1	0–0.692	0.35	8	6.53	6.48	17.70
2	0.692–1.38	1.04	8	6.72	9.28	11.83
3	1.384–2.08	1.73	10	7.26	11.98	12.62
4	2.08–2.77	2.42	14	7.69	14.45	0.19
5	2.77–3.46	3.11	17	9.21	16.61	2.27
6	3.46–4.15	3.81	17	11.24	18.33	4.95
7	4.15–4.84	4.50	23	12.35	19.52	16.57
8	4.84–5.54	5.19	18	15.63	20.06	9.47
9	5.54–6.23	5.88	21	10.25	19.85	5.26
10	6.23–6.92	6.57	20	–	18.79	7.23
11	6.92–7.61	7.27	11	–	16.76	36.52
12	7.61–8.30	7.96	17	–	13.65	19.68
					Average	12.02

4.8 Cluster Regression and Markov Chain

In all the previous regression approaches, the dependence between successive data values are not taken into consideration serially, but an overall trend function is fitted to the available data at hand. Also the trend line is searched among two variables such as the time and the variable evolution values along time. Sometimes, there are trends and/or sudden jumps within the successive values of the variable concerned. As explained in Chap. 2, the lag-one or multi-lag apart from scatter diagram of the same variable appears as a straight line, which is already referred to as the correlation coefficient of that lag, but it does not give the internal trend component within the time series. It is well-known that either serial or cross-correlation coefficient is valid for stationary time series intact of trend component.

In order to explain the cluster regression approach, monthly water-level fluctuation records in the Van Lake in the eastern province of Turkey are taken into consideration as shown in Fig. 4.9. Lake Van has been subject to a net water level rise of about 2 m and consequently the low-lying, inundated areas along the shore are now giving problems to local administrators, governmental officials, and irrigation activities and to people's property (Kadıoğlu et al. 1997).

Whatever the causes might be, there has been a systematic increase in the water level of Lake Van. In the following subsection level changes will be modeled by means of a simple approach based on the combination of regression line and transition probability methods, which are the two basic components of the cluster regression methodology. The regression analysis is adopted because it furnishes the basis of the short-term persistence through an autocorrelation coefficient and probability, due to its suitability for clustering of points as a result of possible abrupt shifts.

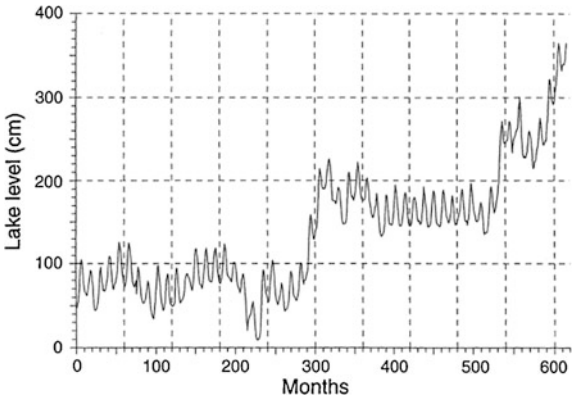


Fig. 4.9 Lake Van water level fluctuations

4.8.1 Cluster Regression Model

Classical regression analysis has several assumptions about the normality and independence of the residuals. Furthermore, an implied assumption that skips from the considerations in most regression line applications is that the scatter diagram should have the points distributed uniformly (homoscedascity) around a line. Unfortunately, this assumption is often overlooked, especially if the scatter diagram is not plotted. Uniform scatter of the points along the line is possible if the original records are homogeneous and stationary with no shifts, trends, or seasonality (Chaps. 2 and 3). If level shifts exist through time, then the scatter diagram will include clusters of points along the regression line. Confirmation of such clusters is obvious in Fig. 4.10, which shows the lake-level lag-one scatter diagram for monthly records from Lake Van.

The following conclusions are possible from interpretation of the scatter diagram in this figure

(1) The lag-one scatter diagram indicates an overall straight-line relationship between the successive lake-level occurrences. Existence of such a straight line corresponds to the first-order autocorrelation coefficient in the monthly lake-level time series. Hence, lake-level persistence is preserved by this straight line,
(2) The scatter of points around the straight line is confined within a narrow band, which implies that the prediction of immediate future levels cannot be very different from the current level provided that there are no shifts in the data,
(3) There are different cluster regions along the straight line. Such clusters are not expected in the classical regression approach but the existence of these clusters

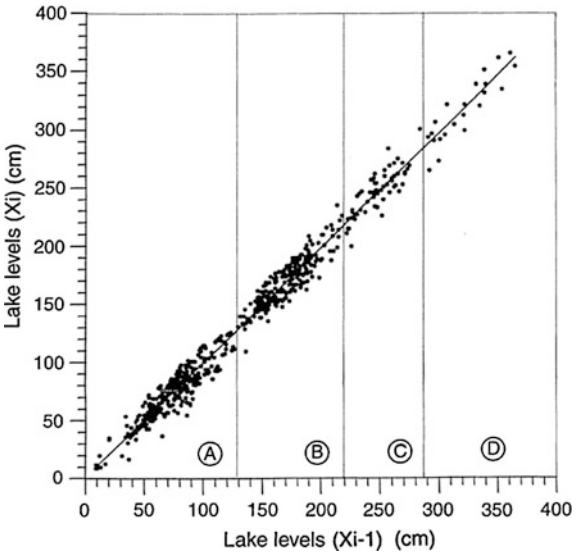


Fig. 4.10 Lag-one water level fluctuations and cluster boundaries

renders the classical regression analysis into a cluster regression analysis, the basis of which will be presented later in this section. Separate clusters correspond to periods of shifted lake level,

(4) Classical regression analysis provides a basis for predicting water levels relative to current, but in the cluster regression line approach, reliable predictions are only possible provided that the probability of cluster occurrences are taken into consideration. Herein, the questions arise as to which cluster is to be taken for future predictions? Should the future prediction remain within the same cluster? Any transition from one cluster to another means a shift in the water level. We need, therefore, to know the transitional probabilities among various clusters. The cluster regression depicts not only the autocorrelation coefficient but also the influence domain of each cluster as shown in Fig. 4.10 along the horizontal axis as *A*, *B*, *C*, and *D*. The influence domains help to calculate the transitional probabilities between the clusters from the original water level records,

(5) For any current cluster, it is possible to estimate future normal lake levels using the regression line equation.

For reliable estimations through cluster regression, one should follow these steps in sequence:

(1) In order to decide initially, which domain of influence (*A*, *B*, *C*, or *D*) should be taken into consideration, a uniform distribution function is considered that assumes random values between 0 and 400 cm,

(2) Generate a uniformly distributed random number and, accordingly, decide about the next cluster by considering influence domains. For instance, if the uniformly distributed random number is 272 then from Fig. 4.10 influence domain, *C* will be the current cluster,

(3) Generate another uniformly distributed random number and if the level remains within the same cluster then use the regression equation for estimation. Otherwise, take the average water level value in the new cluster.

The new level will be adopted as the midpoint of the cluster domains in Fig. 4.10. A better estimation might be based on the random variable generation again from a uniform distribution confined within the variation domain of each cluster. Furthermore, the value found in this manner will be added to a random residual value. This will then give the basis of the future water level estimations within the same cluster domain.

4.8.2 Application and Discussion

The cluster regression approach has been applied herein to the recorded water level fluctuations of the Lake Van. For this purpose, various lag scatter points of the successive levels are first plotted in Figs. 4.11 and 4.12.

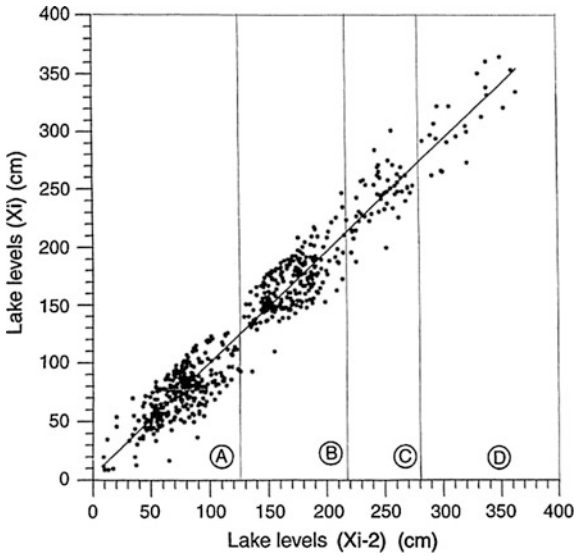


Fig. 4.11 Lag-two water level fluctuations and cluster boundaries

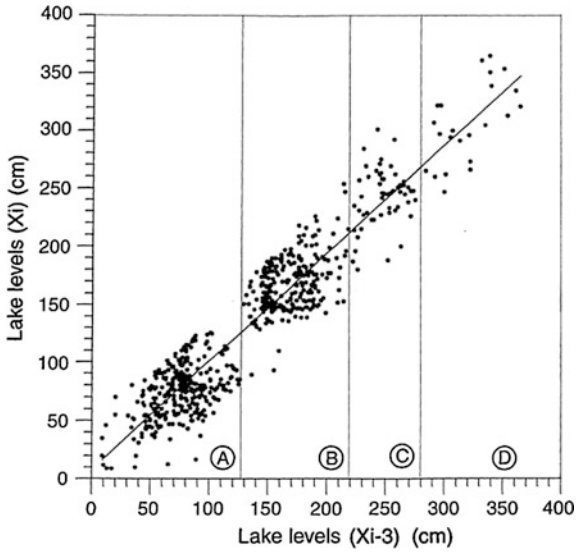


Fig. 4.12 Lag-three water level fluctuations and cluster boundaries

The general appearance of these figures implies the applicability of the cluster regression equation steps as mentioned in the previous section. In all the figures, there are straight lines and the transition boundaries between clusters of *A*, *B*, *C*, and *D* are given in Table 4.4 in addition to the boundaries of each cluster at different lags up to 9.

Table 4.4 Cluster regression boundaries and coefficients

Lag	Transboundary values			Regression coefficients	
	A–B	*B–C*	*C–D*	*a*	*b*
1	130	220	>285	0.985	1.459
2	125	218	>280	0.960	4.564
3	129	215	>280	0.930	8.239
4	130	219	>281	0.901	11.725
5	128	222	>280	0.878	14.486
6	128	212	>287	0.862	16.292
7	130	221	>296	0.853	17.114
8	132	225	>302	0.852	16.835
9	131	222	>308	0.858	15.532
Average	129	219		0.898	11.805

Here, *A* is considered as a low lake-level cluster, where only transitions from low level to low level are allowed. *B* and *C* refer to lower medium and upper medium level clusters and, finally, *D* is the cluster that includes highest levels only. It is obvious from Table 4.4 that the transition limits between *A* and *B*, and *B* and *C* are practically constant on average for all lags and equal to 129 and 219, respectively. However, the upper limit transition between *C–D* increases with the increase in the lag value. The difference between the first and ninth lags has a relative error percentage of (100 × (308)285)/308 = 7.4, which may be regarded as small for practical purposes.

The scatter diagrams in Figs. 4.11 and 4.12 yield the following specific interpretations for Lake Van level fluctuations:

(1) The scatter diagrams have four clusters with the densest point concentration in cluster *A* that represents low water level following low water levels. Extreme values of water level fluctuations have the least frequency of occurrences in cluster *D*.
(2) Irrespective of the lag value, points in the scatter diagram deviate from the regression line within a narrow band. This indicates that once the water level is within a certain cluster it will remain within this cluster with comparatively very high probability as will be argued later in this work. Furthermore, the transitions between the clusters are expected to take place rather rarely and in fact between the adjacent clusters only.
(3) In none of the scatter diagrams is transition of water level possible from one cluster to another nonadjacent one. This may be confirmed from the calculated transition matrix elements because there are no elements except along the main and the two off-diagonals.

Herein, only lag-one regression line will be considered to model the lake levels by considering transitional probabilities between adjacent clusters. The monthly level time series data for Lake Van yield lag-one transition probability matrix, [*M*] as follows:

$$[M] = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 291 & 2 & 0 & 0 \\ 1 & 233 & 4 & 0 \\ 0 & 3 & 54 & 2 \\ 0 & 0 & 1 & 20 \end{bmatrix} \end{array} \tag{4.31}$$

The diagonal values in this matrix are the numbers of transitions within each cluster. For instance, there are 291 transitions from low levels to low levels within cluster *A*. In the same matrix, intercluster transitions occur rather rarely along the off-diagonals, such as four transitions from cluster *B* to *C*. In classical stochastic processes, the calculation of transition matrix elements is based on the fundamental assumption that the process is time reversible. This is equivalent to saying that transitions as $A \rightarrow B$ is the same as $B \rightarrow A$. Consequently, the resulting matrix must be symmetrical. However, in the proposed method of cluster regression technique, only one-way transitions along the time axis toward future is allowed. This means that the transition along the time axis is irreversible. As a result of this fact the transition matrix is not symmetrical. Accordingly, the matrix in Eq. (4.31) is not symmetric; the transition from *C* to *B* is not equal to 4 but 3. Zero values next to the off-diagonals indicate that the water levels can move only to adjacent clusters. Hence, the possible transitions are *ABCD* only. For instance, transition to cluster *C* is possible 4 times from *B*, 54 times from previous *C* and only once from *D* with no transition from *A*, (hence a total of 59 transitions). Columnar values show transition to the cluster considered from other clusters and the transition probabilities can be calculated after dividing each value in the column by the column total. Hence, the transition probability matrix [*P*] becomes from Eq. (4.31) as

$$[P] = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0.9932 & 0.0068 & 0 & 0 \\ 0.0042 & 0.9790 & 0.0168 & 0 \\ 0 & 0.0508 & 0.9152 & 0.0339 \\ 0 & 0 & 0.0476 & 0.9524 \end{bmatrix} \end{array} \tag{4.32}$$

The linear regression line that relates two successive water levels, namely, W_i and W_{i-1}, can be obtained from the cluster scatter diagram in Fig. 4.11 as,

$$W_i = 0.98598 W_{i-1} + 1.45918 + \varepsilon_i \tag{4.33}$$

in which ε_i signifies the vertical random deviations from the regression line. Theoretically, these random deviations should have a Gaussian distribution function for the validity of the regression line and Fig. 4.13 indicates that they are normally distributed. In order to adopt Eq. (4.33) estimations with the cluster scatters, it is essential to take into account the following steps:

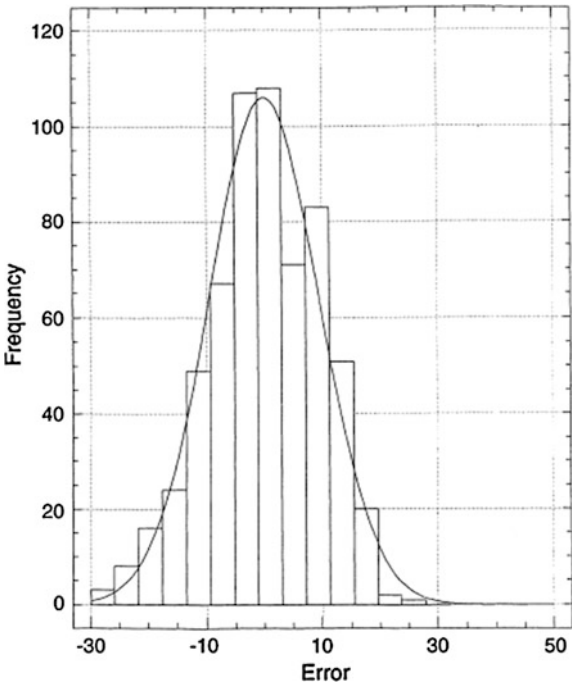


Fig. 4.13 Regression error distribution error

a. Because the most frequently occurring water levels are confined in cluster A, the initial state W_0 is selected randomly from the actual water levels in this cluster,
b. Decision whether there is transition to the next cluster is achieved through the transition probabilities given in matrix $[P]$ in Eq. (4.32). The transitions occur according to the following rules,

1. Transition to cluster A is possible only from cluster B or the level remains within the same cluster. From the transition matrix these have probabilities as 0.9932 and 0.0042 and their summation is equal to 1.0. In order to decide which one of these two clusters will be effective in the next time step, it is necessary to generate a uniform random number, ζi, which varies between zero and one. If $\zeta i < 0.9932$, then the water level will remain within cluster A, otherwise for $0.9932 < \zeta i < 1.0$ a transition occurs from cluster A to B. In the former case, after generating a normally distributed random number, ε_i, the new water level value is generated by the use of the clusteral regression model in Eq. (4.33). However, in the latter case, water level will be selected randomly from the range of water levels for cluster, B.
2. At any instant, transition to cluster B may take place from two adjacent clusters (A or C). The transitional probabilities from A and C are 0.0068 and 0.0508, respectively, with complementary probability of 0.9790 remaining within cluster B. Now the decision of transition to B will have three

independent regions of the uniform distribution, namely, if $0 < \zeta i < 0.0068$ then a transition occurs from A to B or when $0.0068 < \zeta i < 0.9858$ water level remains within cluster B and finally, for $0.9858 < \zeta i < 1.0$ a transition occurs from C to B. If the water level remains within cluster B, a normal variate is generated as ei and the regression expression in Eq. (4.33) is used to predict the next water level. In the transition cases, water level is depicted randomly from the available levels.

c. Transitions to cluster C show a similar mechanism to cluster B with different transition probabilities but the same generating mechanism,
d. Finally, transition to cluster D is possible only from cluster C, in addition to remaining in the same cluster. The application of all these procedures and steps to Lake Van monthly water level variations result in the development of the synthetic transition matrix $[M_s]$

$$[M_s] = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 287 & 2 & 0 & 0 \\ 1 & 230 & 3 & 0 \\ 0 & 4 & 56 & 2 \\ 0 & 0 & 1 & 19 \end{bmatrix} \end{array} . \tag{4.34}$$

Comparison of corresponding elements between the two last matrices shows that they differ by less than 5% relative error. This indicates that the preservation of transition numbers as probabilities in the predicted lake levels are indistinguishable

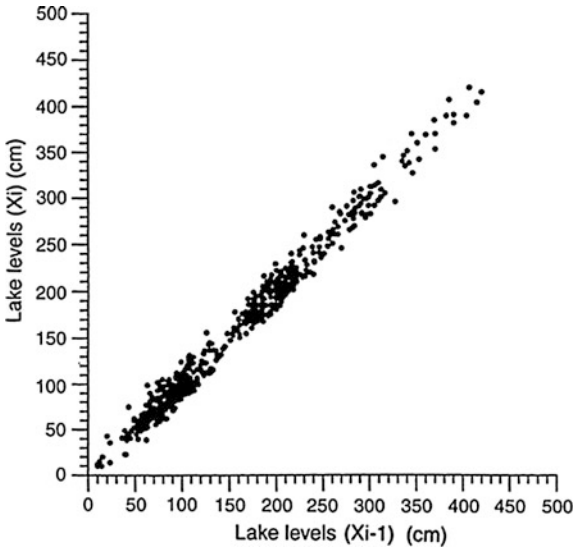


Fig. 4.14 Synthetic lag-one water level fluctuations

from the actual water level data. The synthetic cluster scatter diagram obtained from the use of Eqs. (4.32) and (4.33) is shown in Fig. 4.14 where the regression line has the form as

$$W_i = 0.978W_{i-1} + 1.47 + \varepsilon_i \tag{4.35}$$

Again comparison of this expression with Eq. (4.33) shows that the corresponding coefficients vary by less than 5% relative error. In other words, the autocorrelation coefficient in the prediction of water levels is preserved in spite of shifts in the original data.

The bases of a new regression equation with clusters are presented with an application to the water level fluctuations of Lake Van, eastern Turkey. The clusteral regression method provides the best regression line in addition to the cluster occurrences and transition probabilities along this line. Its difference from the classical regression approach lies in the appearance of nonoverlapping clusters. The cluster regression approach preserves all the statistical parameters in addition to the autocorrelation coefficient, which is a measure of short-term persistence in lake-level records. Any shifts in the data do not lead to spurious and unrealistic autocorrelation.

4.9 Trend Over-whitening Procedures

Most trend-detection studies using the MK test (Chap. 3) have assumed that sample data are serially independent, even though certain natural, economic, environmental, social, etc., series have statistically significant serial correlation. Furthermore, von Storch (1995) documented that the existence of positive serial correlation increases the probability that the MK test detects trend when no trend exists. In order to convert serially dependent time series into independent structure series Yue et al. (2002) suggested the pre-whitening procedure. Pre-whitening is a procedure for reduction of serial correlation within a given time series by adding white noise (serially independent) series to the original series. It is demonstrated that removal of positive serial correlation by pre-whitening removes a portion of actual trend.

None of the classical trend tests such as the MK test takes into account classical parametric and most commonly used serial correlation, and hence, they require independence structure in the applications. In general, independence test can be carried out mainly by examining the autocorrelation coefficients of the time series. If the absolute values of the autocorrelation coefficients for a time series consisting of n observations are not larger than the typical critical value, i.e., $1.96/\sqrt{n}$ corresponding to the 5% significance level (Douglas et al. 2000), then the observations in this time series can be accepted as being independent from each other. The significance of the trend is determined using Kendall's test because it does not assume an underlying pdf of the data series. There is, however, a problem associated with the Kendall test in that the result is affected by serial correlation of the series.

Specifically, a positive autocorrelation, that is likely the case for most natural time series records in the residual time series, will result in more false detection of a significant trend than specified by the significance level (von Storch 1995; Zhang and Zwiers 2004). This would make the trend detection unreliable. Some authors have tried to alleviate this important drawback by pre-whitening the given time series data. For instance, von Storch (1995) indicated that the existence of positive serial correlation increases the probability that the MK test detects trend when the given time series is trend free. He then proposed removal of the serial correlation through pre-whitening procedure prior to the application of MK test. However, Douglas et al. (2000) further explained the reduction in the serial correlation after pre-whitening but with some loss in trend information. Furthermore, Yue et al. (2002) explored the influence of pre-whitening and they found that removal of positive serial correlation by pre-whitening removes a portion of trend.

There is, however, a problem associated with the MK test in that the result is affected by serial correlation of the series. Specifically, a positive autocorrelation, that is likely the case for most climatological data in the residual time series, will result in more false detection of a significant trend than specified by the significance level (von Storch 1995; Zhang and Zwiers 2004). This would make the trend detection unreliable. Some authors have tried to alleviate this important drawback by pre-whitening the given hydro-climatic series. For instance, von Storch (1995) indicated that the existence of positive serial correlation increases the probability that the MK test detects trend when the given time series is trend free. He then proposed removal of the serial correlation through pre-whitening procedure prior to the application of MK test. However, Douglas et al. (2000) further explained the reduction in the serial correlation after pre-whitening but with some loss in trend information. Furthermore, Yue et al. (2002) explored the influence of pre-whitening and they found that removal of positive serial correlation by pre-whitening remove a portion of trend.

Different statistical methodologies are employed to identify possible trend component in any time series. Pre-whitening (PW) procedure has been suggested to reduce the serial correlation effect on Mann–Kendall (MK) trend analysis. In this section, instead of PW, over-whitening (OW) procedure is suggested (Şen 2016), which generates serially independent series with the same trend slope value. Analytically necessary formulations for OW are presented with a nonparametric but simple innovative trend assessment procedure, which are supported by extensive simulation studies.

4.9.1 Over-whitening (OW) Process

The purpose of over-whitening procedure is to reduce the original time series serial dependence function down to almost independent serial structure without any harm on the trend component. Let Y_t represent the original time series with monotonic trend component and addition of an independent random (white noise) component, ε_t, with zero mean gives rise to over-whitened time series, Z_t, as,

$$Z_t = Y_t + \varepsilon_t \tag{4.36}$$

It is possible to rewrite explicitly by considering a monotonic linear trend component with slope, β, and OW independent time series, η_t, with zero mean and unit standard deviation as,

$$Z_t = X_t + \beta t + \gamma\eta_t, \tag{4.37}$$

where X_t is a stationary process and γ is the OW standard deviation. In order to facilitate the analytical derivations, it is assumed that X_t is in the form of the standardized AR(1) process zero mean and unit variance. Time conditional expectation of both sides in Eq. (4.37) leads to,

$$E(Z_t|t) = E(X_t|t) + E(\beta t|t) + E(\gamma\eta_t|t)$$

Since $E(X_t) = E(\eta_t) = 0$, then

$$E(Z_t|t) = \beta t \tag{4.38}$$

which is the trend component only.

For the variance calculation, square of both sides in Eq. (4.37) and then expectations result in,

$$E(Z_t^2|t) = E(X_t^2|t) + E(\beta^2 t^2|t) + E(\gamma^2\eta_t^2|t) + 2E(X_t\beta t|t) + 2E(X_t\gamma\eta_t|t) + 2E(\beta t\gamma\eta_t|t)$$

Herein, by definition, $E\left(X_t^2|t\right) = 1$; $E\left(\gamma^2\eta_t^2|t\right) = \gamma^2$; $E(X_t\beta t) = 0$; $E(X_t\gamma\eta_t|t) = 0$ (since η_t is independent from X_t); $E(\beta t\eta_t|t) = 0$. Consideration of the variance definition as $\mathrm{Var}(Z_t|t) = E\left(Z_t^2|t\right) - E^2(Z_t|t)$ after the necessary algebraic calculations gives,

$$\mathrm{Var}(Z_t|t) = 1 + \gamma^2 \tag{4.39}$$

On the other hand, the k-order conditional covariance of the whole time series can be obtained as follows:

$$\begin{aligned}\mathrm{Cov}(Z_t, Z_{t-k}|t) &= E\{[X_t + \beta t + \gamma\eta_t][X_{t-k} + \beta(t-k) + \gamma\eta_{t-k}]\} \\ &= E(X_tX_{t-k}) + E[X_t\beta(t-k)] + E(X_t\gamma\eta_{t-k}) + E(\beta_tX_{t-k}) + E\left[\beta^2 t(t-k)\right] + E(\beta t\gamma\eta_{t-k}) \\ &= E(\gamma\eta_tX_{t-k}) + E[\gamma\eta_t\beta(t-k)] + E(\gamma\eta_t\gamma\eta_{t-k})\end{aligned}$$

By taking into consideration the AR(1) process autocorrelation coefficient structure and independence of X_t and η_t processes, this expression can be simplified as,

$$\mathrm{Cov}(Z_t, Z_{t-k}|t) = \rho^k + \beta^2 t(t-k)$$

Here, ρ is the first-order autocorrelation coefficient of the AR(1) process. The k-th order dependency coefficient ρ_k of Z_t process can be calculated after dividing this last expression by Eq. (4.39),

$$\rho_k = \frac{\rho^k + \beta^2 t(t-k)}{1+\gamma^2} \tag{4.40}$$

At this stage, it is very convenient to remember that Yue et al. (2002, 2003) stated a modified methodology, where the slope of trend is first estimated and then the record is de-trended. Subsequently, the lag-k (example lag-one) serial correlation coefficient of the de-trended series is estimated, and then the series is pre-whitened. They also argued that the removal of the trend as a first step may allow for more accurate estimate of the population's lag-one autocorrelation coefficient, and subsequently better estimation of trend. In order to apply these arguments in an analytical manner, let us assume that $\beta = 0$, and hence, the trend is removed from Eq. (4.37), which leads to the autocorrelation coefficient from Eq. (4.40) as,

$$\rho_k = \frac{\rho^k}{1+\gamma^2} = \alpha\rho^k, \tag{4.41}$$

where $0 < \alpha < 1$, and herein, it is referred to as the dependence reduction factor. Hence, the OW k-th order autocorrelation function is a function of the lag-k autocorrelation coefficient of the original time series, and the OW standard deviation, γ. Equation (4.41) implies that the autocorrelation structure of any given time series with lag-one autocorrelation can be reduced to the first-order over-whitened serial correlation coefficient, ρ_o, as,

$$\rho_o = \alpha\rho_1 \tag{4.42}$$

The relationship between ρ_o and ρ through α can be seen from Fig. 4.15. Depending on the first-order serial correlation of the original series, ρ_1, with trend component, one can reduce it as small as possible desired, by selecting a convenient α dependence reduction factor. However, it is not possible to obtain absolutely independent process unless the time series itself originally has completely independent structure, i.e., $\rho_1 = 0$. This point agrees with the statement by Yue and Wang (2002) and Bayazıt and Önöz (2007) that by PW, the dependence structure can be reduced such that the serial correlation coefficients becomes close to zero. The same statement is valid also for OW procedure.

This figure together with Eq. (4.42) indicate that it is not possible to make the autocorrelation structure of the time series purely independent even after over-whitening, because for such a case α should be equal to zero, which is not possible practically. Yue and Wang (2002) stated that "…pre-whitening is not suitable for eliminating the effect of serial correlation coefficient on the MK test when trend exists in a time series", because "….pre-whitening will remove a

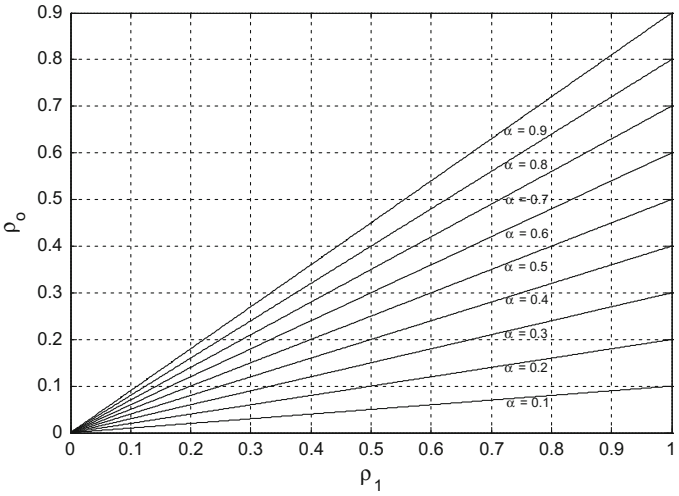


Fig. 4.15 OW chart

portion (equal to the lag-one autocorrelation coefficient) of trend, and hence, reduces the probability of rejecting the null hypothesis when it is false". The OW procedure will not be affected from such shortcomings.

Additionally, Yue and Wang (2004a, b) stated that "Pre-whitening a time series using spurious or contaminated serial correlation coefficient is fundamentally wrong", because "…the existence of a trend in a time series will produce a spurious serial correlation when there is no serial correlation, and the presence of trend will increase the estimate of positive serial correlation coefficient when the serial correlation exists…". This last part of the statement is obvious from Eq. (4.40), where there is an additional term in the numerator and whatever the trend slope (positive or negative) due to β^2 term there will always be an increase in the serial correlation coefficient. The confirmation of this last sentence has been given by Bayazıt and Önöz (2007) by saying that "…the trend (upward or downward) always has a positive contribution to serial correlation". In order to avoid this effect Yue and Wang (2004b) proposed that the existing trend component should be removed from a time series first, and then the lag-one serial correlation coefficient may be computed from the residuals, so that it is no longer affected by the trend". In the OW procedure, the trend remains as it is without removal opposite to the case of PW.

The standard deviation of over-whitening component can be calculated from the last two parts of Eq. (4.41) as,

$$\gamma = \sqrt{\frac{1}{\alpha} - 1} \tag{4.43}$$

Since α is always positive γ will be positive. In practical applications convenient α value should be chosen from the chart in Fig. 4.15 or calculated from Eq. (4.42) as,

$$\alpha = \frac{\rho_o}{\rho_1}, \tag{4.44}$$

where ρ_o is desired OW first-order correlation coefficient that should be chosen so as to make the over-whitened time series to have almost independent correlation structure and ρ_1 is the first-order serial correlation coefficient of the original time series, i.e., available hydro-meteorological record. In order to confirm these equations, in the following subsection simulation results are treated with innovative trend template approach.

4.9.2 Simulation

Two different simulation studies are presented in this section. The first one indicates that whatever the first-order serial correlation coefficient and the trend slope are the innovative trend procedure yields the same template without OW. The second set of simulation is for over-whitened time series with different OW standard deviations. These simulation studies indicate that whether OW is applied or not the innovative methodology yields the same trend pattern.

For this purpose, a series of Monte Carlo computer simulations are performed with a set of statistical parameters by the use of AR(1) process. The mean and variance of the stochastic process are assumed as zero and one, respectively. The simulation procedure takes into consideration the set of autocorrelation coefficients ($\rho = \pm 0.1, \pm 0.3, \pm 0.5, \pm 0.7$ and ± 0.9) and the trend slopes set ($s = \pm 0.001, \pm 0.003, \pm 0.005, \pm 0.007$ and ± 0.009). The length of the synthetically generated series is adapted as 1,000. After alternative combinations between a serial correlation coefficient and the slope set, all innovative trend templates have appeared in the same pattern without any distinction among the simulation results. Herein only

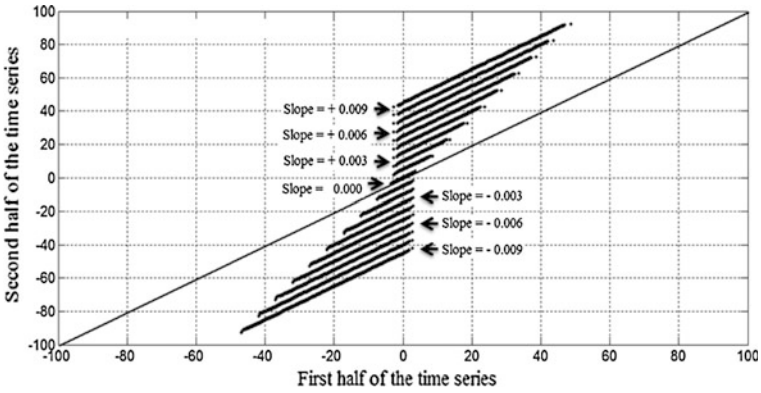


Fig. 4.16 General innovative template for trend slopes irrespective of serial correlation coefficient

one innovative trend template is presented in Fig. 4.16 that covers all possible combination cases.

Another set of simulation is carried out for a high serial correlation coefficient equal to 0.9 with trend component of 0.0003 and the simulation is carried out along the following steps.

(1) 1,000 Gaussian independent random variables are generated with zero mean and unit standard deviation,
(2) These random variables are converted to AR(1) process with serial correlation coefficient equal to 0.9,
(3) The sequence is embedded with a trend component of slope equal to 0.003. The resulting time series is shown in Fig. 4.17a,
(4) Innovative trend template is obtained shown in Fig. 4.17b,
(5) In order to over-whiten the sequence according to Eq. (4.37), one needs, first, to decide about the serial dependence reduction for OW. If is chosen as $\alpha = 0.01$ then Eq. (4.43) yields the standard deviation of the over-whitening sequence as $\gamma = 9.95$. This factor plays the role of scaling, but does not affect the embedded trend component,
(6) Addition of the component over the same sequence and then the application of the innovative trend procedure yields the trend template as in Fig. 4.17c, which is scaled down version of Fig. 4.17b,
(7) Figure 4.17b, c are overlapped and the final result, in Fig. 4.17d, shows that whether over-whitened or not the same trend is preserved even in the over-whitened series.

4.9.3 Application

The applications of aforementioned innovative trend methodology and other procedures to factual data are presented for three different cases all with annual temperature records. These are New Jersey State, USA, and global annual temperature anomalies annual temperature records with more than 100 years (116 and 134 years). Florya meteorology station in Istanbul, Turkey, has shorter duration (56 years). The Office of the New Jersey State Climatologist has gathered and quality checked New Jersey state-wide annual temperature records going back to 1895, and has made these data available on-line (http://epa.gov/climatechange/index.html; http://climate.rutgers.edu/stateclim). These data, as summarized and charted by the Department, show a statistically significant rise in average state-wide precipitation and temperature over the last 116 years.

Global annual temperature anomalies (^{o}C) are computed using data meteorological stations from 1981 to 2014. The anomalies are relative to the 1951–1980 base period means of 30-year data. The Internet site for this data set is available at http://cdiac.ornl.gov/ftp/trends/temp/hansen/gl_land.txt.

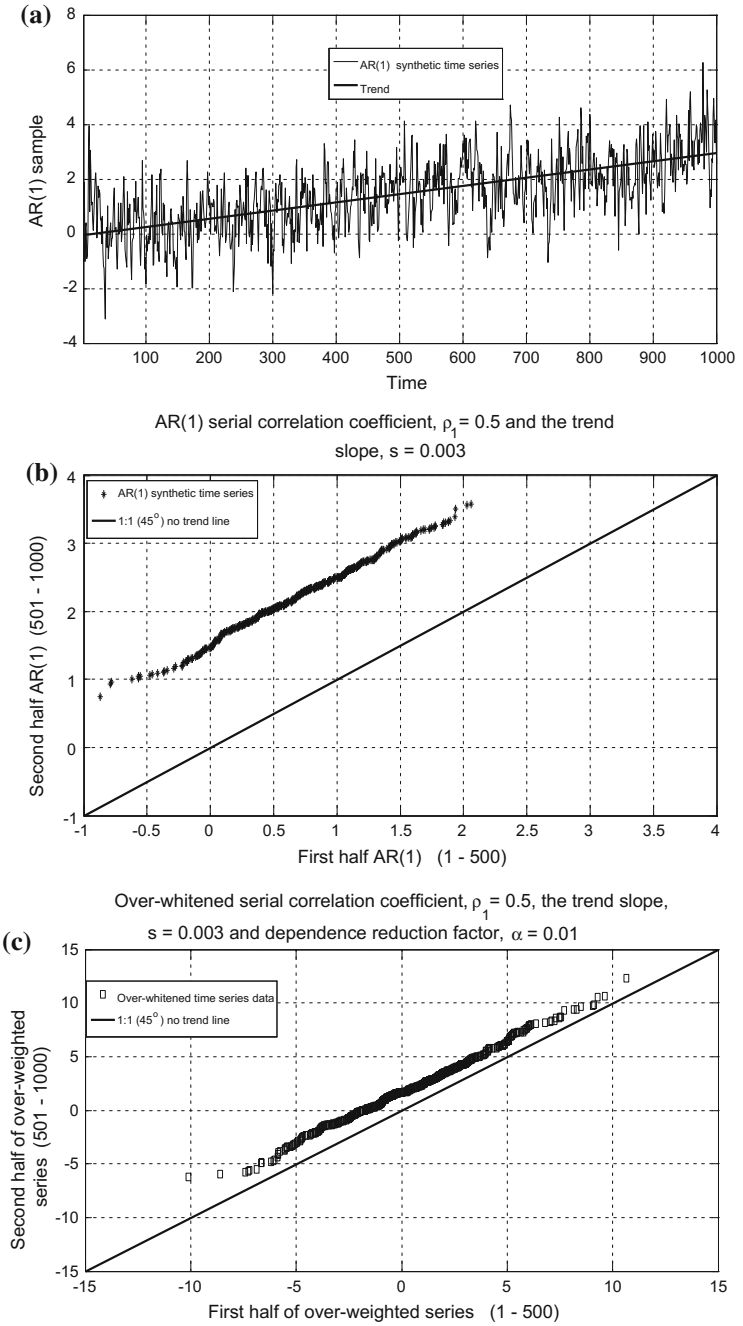


Fig. 4.17 **a** Trend embedded time series, **b** trend prior to OW, **c** trend posterior to OW, **d** prior to and posterior to OW

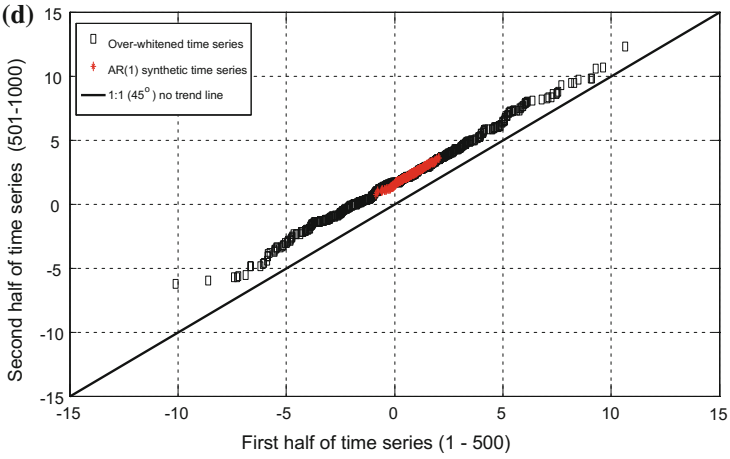


Fig. 4.17 (continued)

The Florya data are from the Turkish Meteorology Service and it is from 1936 to 2006, inclusive. The basic statistical features of these records are given in Table 4.5. The dependence reduction factor, α, is chosen as the lowest appearing value in Fig. 5.22 as 0.1 for each record.

In order to carry out the OW procedure software is written in Matlab programing language and its logical steps are as follows:

(1) Standardize each record by subtracting the mean value and then dividing this difference by the standard deviation. As mentioned earlier, the standardization procedure does not affect the existing trend component in the original time series and the same is valid for the autocorrelation coefficient,
(2) The standardized series is over-whitened by a white noise (completely independent random series) time series with zero mean and OW standard deviation (the last column in Table 4.5). The probability distribution function (pdf) of OW process is always a normal (Gaussian) pdf,
(3) OW process renders the dependent serial structure of the standardized time series into independent case,
(4) The software generates a set of Gaussian independent time series of the same length with the original time series and records their trend slopes individually at the memory. Hence, an ensemble of independent time series are generated,
(5) The trend slopes in these independent time series are averaged so as to find the trend slope after OW procedure.
(6) Generate innovative template for the standardized and over-whitened time series and show the first half, m_1, and second half, m_2, arithmetic averages of the over-whitened time series on the same graph,
(7) Calculate the trend slope, S, value from the over-whitened time series by the well-known Sen (1968) procedure that is invariably employed in MK trend

Table 4.5 Statistical and over-whitening parameters

Location	Record duration (year)	Statistical parameters			OW parameters		
		Arithmetic average (μ)	Standard deviation (σ)	Correlation coefficient (ρ_1)	Dependence reduction factor (α)	Correlation coefficient (ρ_o) (Eq. 5.20)	Standard deviation (γ) Eq. (5.19)
New Jersey	1895–2010	53.037	2.302	0.386	0.1	0.038	3.162
Istanbul	1936–2006	14.000	0.603	0.301	0.1	0.031	3.162
Global	1881–2014	0.0108	0.382	0.909	0.1	0.090	3.162

analysis. Additionally, in this work the innovative trend slope, S_I, is calculated by the subtracting the second half's arithmetic average from the first half arithmetic average and then dividing this difference by the half-length, $n/2$, of the over-whitened time series as,

$$S_I = \frac{2(m_2 - m_1)}{n} \tag{4.45}$$

It has been observed during this study that the difference between this equation result and the classical within the limits of ±5% relative error, which is well acceptable for practical purposes,

(8) The software produces the autocorrelation function of the original record with the over-whitened one as a graph,

(9) Finally, standardized and the over-whitened time series are shown with the trend on the same graph.

The application of all these steps to New Jersey State annual temperature time series produces three graphs, which are given in Fig. 4.18.

As for the innovative trend graph in Fig. 4.18a, OW innovative and OW MK slopes are given as 0.018629 and 0.018463, respectively and they are practically equal to each other. The scatters of standardized and over-whitened time series are very close to each other. In this and other similar graphs for other two stations, if the over-whitened time series is regenerated by the same software then there may be sampling errors, but the results will always remain within less than ±5% relative error and even very less than this value. The OW process renders the original first-order serial correlation coefficient down to 0.028, which is the required result even from the PW procedure as in the literature (Fig. 4.18b). Both time series (standardized and OW) are presented in Fig. 4.18c with two trend straight lines (OW innovative and OW MK) and they fall on each other, which indicate the validity of the OW approach with the use of Eq. (4.45).

The same arguments are valid for the Florya meteorology station graphs in Fig. 4.19, where the negligible difference between the two trends is visible. However, the difference is well less than practically acceptable 5% (Fig. 4.19c).

Finally, Fig. 4.20 is for the global annual temperature anomalies time series, where one can appreciate similar conclusions as for the two previous examples. The significantly distinctive point in this time series is that its first-order serial correlation coefficient is very high as 0.909, but OW procedure reduces its effect down to the very small value equal to 0.096, which is an evident for the effectiveness of the OW procedure Fig. 4.20.

It is well-documented by now that due to climate change and landscape alterations the hydrological cycle is affected with some increases (decreases) in terms of trends depending on the location on the world. The tendencies are embedded in different hydrological records including temperature, precipitation, runoff, soil moisture, evaporation, etc. It is, therefore, necessary to detect these trends in an

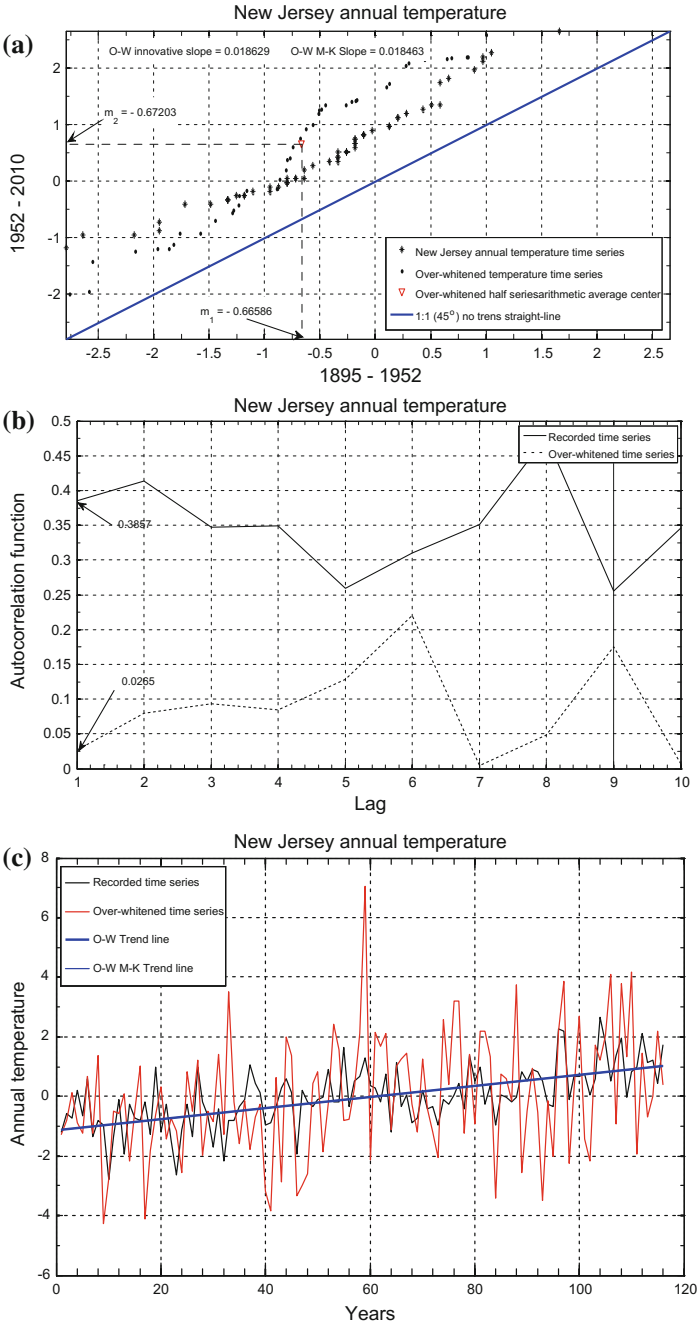


Fig. 4.18 New Jersey annual temperature O-W procedure applications, **a** standardized and OW time series innovative template, **b** autocorrelation graph, **c** time series and trend graph

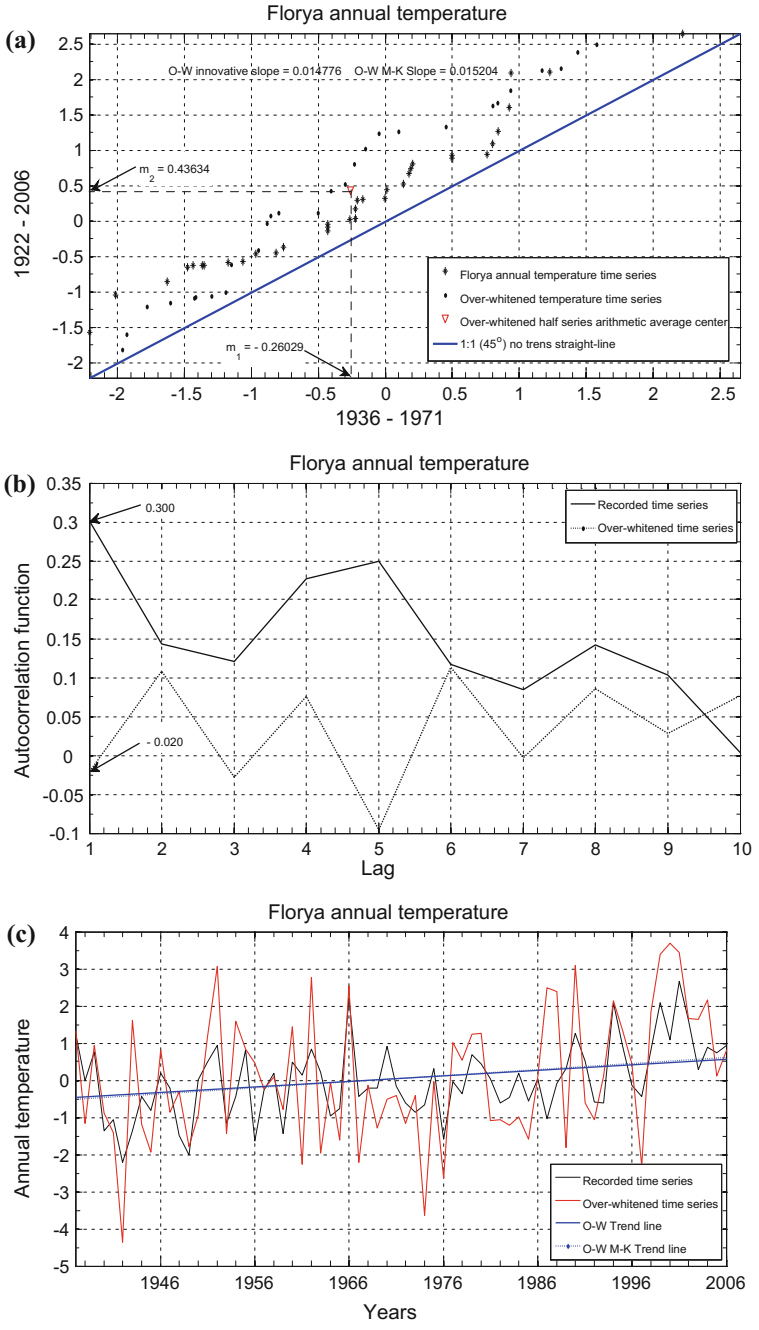


Fig. 4.19 Florya annual temperature OW procedure applications, **a** standardized and OW time series innovative template, **b** autocorrelation graph, **c** time series and trend graph

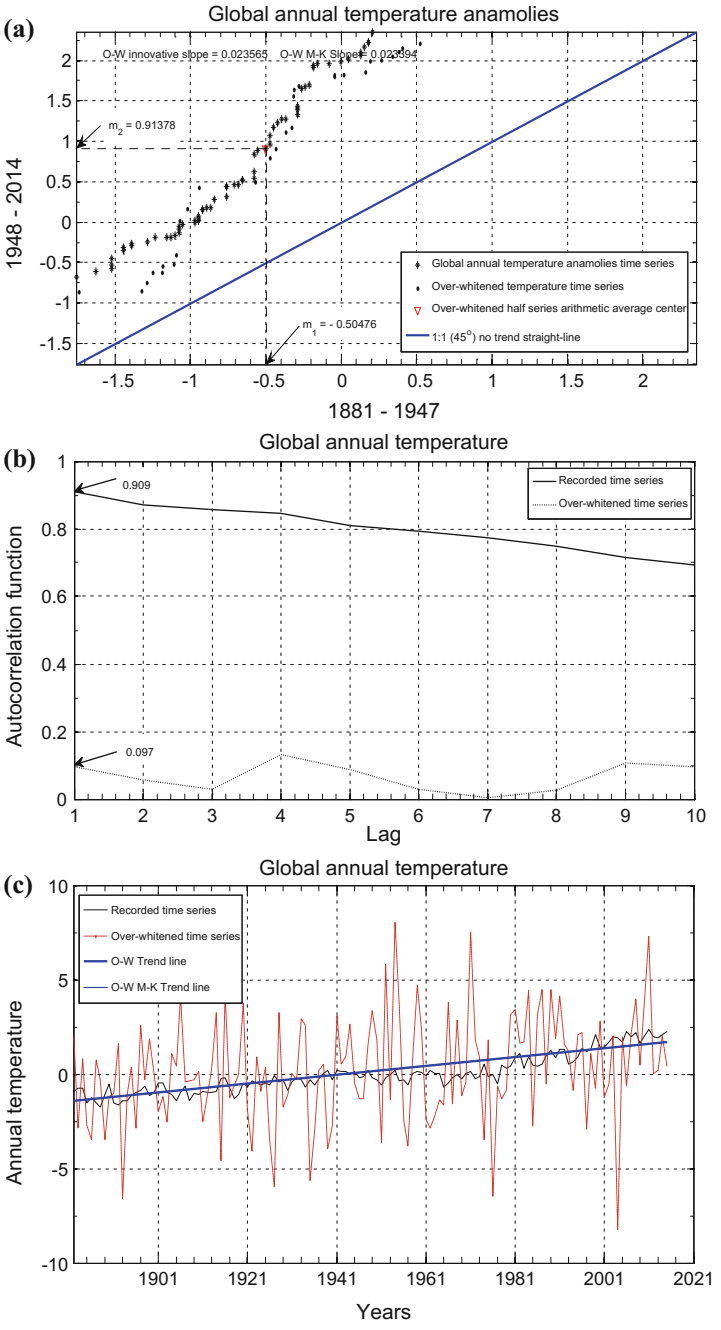


Fig. 4.20 World annual temperature OW process applications, **a** standardized and OW time series innovative template, **b** autocorrelation graph, **c** time series and trend graph

objective manner. In the literature the most frequently used procedure for this purpose is the Mann–Kendal (MK) trend test statistics, where the most important element is the trend slope determination. This test statistics is affected by the serial correlation structure and it is more valid for independent processes. For this purpose, various authors tried to overcome this restriction by trying to pre-whiten (PW) the given time series. In this paper, another new procedure is suggested as over-whitening (OW), where the given time series is superimposed by an independent Gaussian time series with zero mean and standard deviation according to derived analytical derived expressions in the text. Additionally, an innovative trend template concept is used for showing that the dependence or independence serial structure of any time series does behave in the same manner on the template. All the necessary stochastic formulation derivations are presented for the application of the OW procedure. Accordingly, extensive simulation studies have been carried out to validate formulations and procedures. The innovative template and OW methodologies are applied to three annual temperature time series, which are New Jersey State, USA, Florya meteorology station at Istanbul, Turkey and the global annual temperature anomalies from the Internet. The application of OW procedure application to these records yielded reliable results and this new way of trend application can be used in the future in many applications.

References

Bayazit, M., & Önöz, B. (2007). To prewhiten or not to prewhiten in trend analysis? *Hydrological Sciences Journal, 52,* 611–624.

Douglas, E. M., Vogel, R. M., & Kroll, C. N. (2000). Trends in floods and low flows in the United States: Impact of spatial correlation. *Journal of Hydrology, 240,* 90–105.

Fatichi, S., Barbosa, S. M., Caporali, E., & Silva, M. E. (2009). Deterministic versus stochastic trends: Detection and challenges. *Journal of Geophysical Research*, *114*.

Kadıoğlu, M., Şen, Z., & Batur, E. (1997). The greatest soda-water in the world and how influenced by climatic change. *Geophysicae, 15*, 1489–1497.

Kendall, M. G. (1975). *Rank correlation methods*. New York: Oxford University Press.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13,* 245–259.

Önöz, B., & Bayazit, M. (2012). Block bootstrap for Mann-Kendall trend test of serially dependent data. *Hydrological Processing, 26,* 3552–3560.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 63*, 1379–1389.

Şen, Z. (2001). Angström equation parameter estimation by unrestricted method. *Solar Energy, 71* (2), 95–107.

Şen, Z. (2010). *Fuzzy logic and hydrological modeling*. New York: Taylor and Francis Group, CRC Press. 340 pp.

Şen, Z. (2016). Hydrological trend analysis with innovative and over-whitening procedures. *Hydrological Science Journal* (in print).

von Storch, H. (1995). Misuses of statistical analysis in climate research. In H. V. Storch & A. Navarra (Eds.), *Analysis of climate variability: Applications of statistical techniques* (pp. 11–26). New York: Springer.

Yue, S., & Wang, C. Y. (2002). Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test. *Water Resource and Research, 38*. doi:10.1029/2001WR000861.

Yue, S., & Wang, C. Y. (2004a). Reply to comment by M. Bayazit & B. Önöz on Applicability of pre-whitening to eliminate the influence of serial correlation on the Mann-Kendall test. *Water Resources Research, 40,* W08802.

Yue, S., & Wang, C. Y. (2004b). Reply to comment by X. Zhang and F. W. Zwiers on Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test. *Water Resources Research, 40,* W03806.

Yue, S., Pilon, P., & Cavadias, G. (2002). Power of Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology, 259,* 254–271. (*Annales Geophysicae*, *5*, 1489–1497).

Yue, S., Pilon, P., & Phinney, B. (2003). Canadian streamflow trend detection: impacts of serial and crosscorrelation. *Hydrological Sciences Journal, 48*(1), 51–63.

Zhang, X., & Zwiers, F. W. (2004). Comment on "Applicability of prewhitening to eliminate the influence of serial correlation on the MannKendall test" by Sheng Yue and Chun Yuan Wang. *Water Resources Research, 40*, W03805. doi:10.1029/2003WR002073.

Innovative Trend Analyses

5

Abstract

Innovative trend analysis is the most modern, simple, easy to interpret, and effective trend analysis procedure that incorporates first visual inspection for identification of the trend type whether increasing, decreasing, or no trend cases and then provide numerical calculation for the trend slope again by a very simple formulation. All the classical trend determination methodologies try to find holistic monotonic trend either over the whole record period or on pieces of subperiods. However, the innovative trend method compares last parts of any desired duration record length with earlier perions within the time series itself, hence, one can appreciate the trend variation within the record itself. Another innovative trend method is based on the number of crossings along the trend line, which should have the maximum number of crossing. This procedure helps to identify also the surplus and deficit parts of a given time series with respect to the trend line.

Keywords

Crossing · Innovative · Over-whitening · Simulation · Slope · Intercept

5.1 General

There are commonly used trend identification techniques such as Mann–Kendall (MK) and Spearman's Rho (SR) tests as explained in Chap. 3, but their validity is possible under a set of restrictive assumptions such as independent structure of the time series, normality of the distribution and length of data. It is also not possible to calculate trend magnitude (slope) except through regression approach, which brings additional assumptions for the theoretical validation in practical applications.

Z. Şen, *Innovative Trend Methodologies in Science and Engineering*,
DOI 10.1007/978-3-319-52338-5_5

Recent hydrologic regime changes due to potential climate variability impacts brought into focus the search for effective trend identification analysis. Numerous works in different parts of the world showed quasi-periodic natural behavior and systematic trends of key climate variables due to climate change and/or climate variability (Chap. 6). It is well known that changing climate is expected to have notable impacts on the rainfall–runoff processes due to increasing or decreasing trends in hydro-meteorological time series (floods, droughts, heat waves, etc.). These impacts can no longer be assumed to be stationary, which means that future replicates are no more statistically indistinguishable from the historical counterparts. If climate change is not taken into account then such changes or variability can lead to underestimation/overestimation of parameters for the design and operation of water infrastructures, water shortages, water stresses, and agricultural failures. Although some test procedures are presented for trend identification, there are restrictive assumptions with respect to serial structure (ignorance of correlation coefficient), normal probability distribution function (PDF) of the variables and rather lengthy datasets.

Two commonly used trend tests are Mann–Kendall (Mann 1945; Kendall 1975) test and Spearman's Rho test to the data set (Sen 1978). In many studies, these two nonparametric rank-based statistical tests are used for detecting monotonic trends in a given time series. The power of these tests has not been well documented but the simulation results by Yue et al. (2002a, b, c) indicate that the power depends on the pre-assigned significance level, magnitude of trend, sample size, and the amount of variation within a time series. That is, the bigger the absolute magnitude of trend, the more powerful are the tests; as the sample size increases, the tests become more powerful; and as the amount of variation increases within a time series, the power of the tests decrease. When a trend is present, the power is also dependent on the PDF type and the skewness coefficient. The simulation results also demonstrate that these two tests have similar power in detecting a monotonic trend, to the point of being indistinguishable in practice.

In the past, time series were often assumed as stationary or weakly stationary stochastic processes for simulation purposes. Due to anthropogenic (human disturbance) effects on climate, environment, drainage basin and atmosphere, such an assumption is not valid anymore. However, this is almost the case with economic time series. This implies that future predictions cannot be regarded as statistically indistinguishable from the past records. Current anthropogenic impacts substantially affect natural, environmental and economic variables. For instance, events as droughts, floods, and streamflow discharges are also influenced by climate impacts. Monotonic and steadily increasing trends in past records lead to the alteration of planning, operation and management practices of atmospheric researchers, meteorologists, climatologists, economics, and hydrologists alike. Therefore, prior to any future predictions, it is necessary to try and identify possible monotonic trend components in any given time series. Trend identification analyses have been extensively employed in natural works (Kalra et al. 2008; Miller and Piechota 2008; McCabe and Wolock 2002; Lins and Slack 1999; Douglas et al. 2000; Lettenmaier et al. 1994; Groisman et al. 2001).

This section presents preliminary results and applications of two effective and potential innovative trend identification methodologies that do not require many of the restrictive assumptions. The first one is concerned with the plot of a set of subseries from the original time series on a Cartesian coordinate system, where 45° straight-line implies no trend but any plot appearance above (below) this line implies increasing (decreasing) trends. The same methodology is capable to provide trend magnitude (slope) calculation. The other one, crossing trend analysis, depends on the crossing number of a given time series at the arithmetic average truncation level.

5.2 Probability Distribution-Statistical Parameter Trend Implications

The most important trend or shift component indicator in any time series is the PDF provided that there is a long series of available data that can be divided into at least two nonoverlapping equal parts. The frequency distribution function (or histogram) for each part is than fitted to a theoretical PDF and the comparison of these two PDFs provide first a visual inspection about the possibility of trend or sudden shift (jump) component. For the sake of explanation herein the theoretical PDFs are assumed as a normal PDF. Comparison of the relative position of these two PDFs to each other leads to seven different cases The first case is shown in Fig. 5.1 where

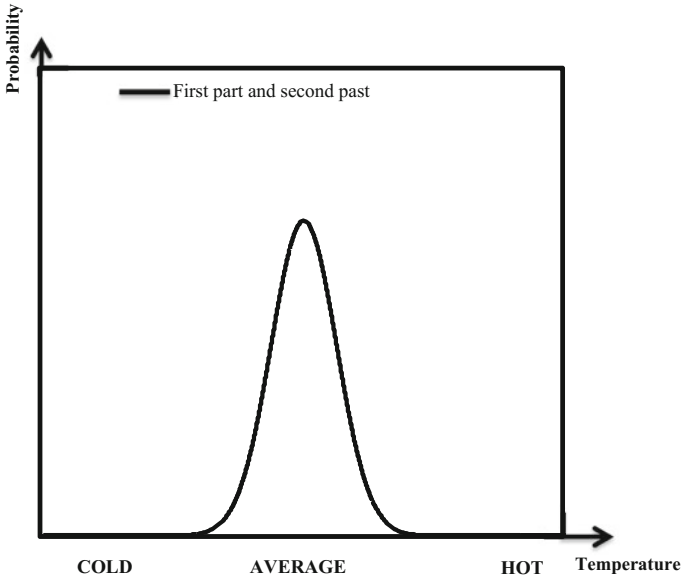


Fig. 5.1 No trend PDF

the two PDFs of the first and next half of the available time series record fall on each other then there is no trend component within the time series (Fig. 5.1). The property implies also that the time series is strictly stationary, because all the statistical parameters are constant along the time series. If one interprets this graph under the light of recent climate change she/he can state that there is no climate change and "hot," "mild," and "cold" climate states remain almost the same by time. Notice that for climate change interpretation the horizontal axis is taken as representative of temperature records.

As in Fig. 5.2 if there is a shift of the first part (past records) PDF toward higher values then there is the possibility of either an increasing trend or a jump that maybe sudden or over a very short period of time. One can decide qualitatively by visual inspection of the time series graph whether it is a trend or a jump. If the increase in the time series values toward recent values seems as gradual then one can conclude that there is an increasing trend, otherwise it is a jump. An important point at this point is that the time difference between the two PDFs in Fig. 5.2 is equal to the half duration time of the time series record duration. This last statement implies that in case of a trend there is a gradual increase from the statistical parameters of the first half toward the second half. This is a very important scrap of information, which enables one to calculate any parameters change slope by taking the difference between the two parameters and its division by the half duration.

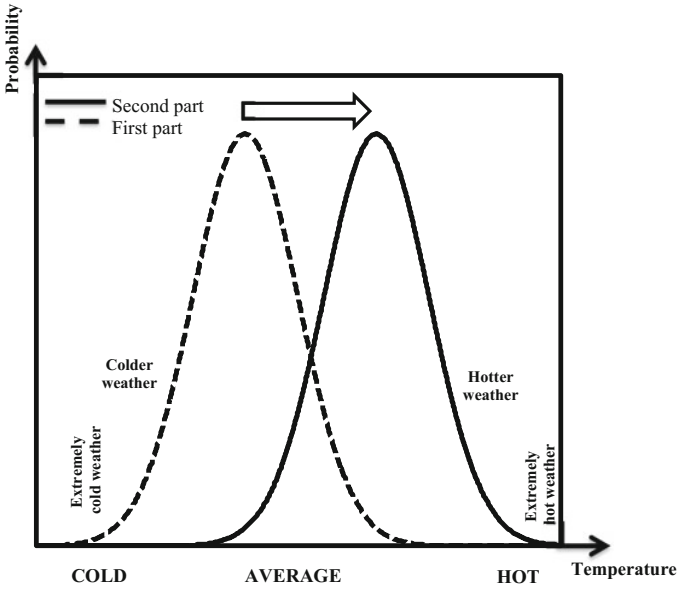


Fig. 5.2 Increase trend PDFs

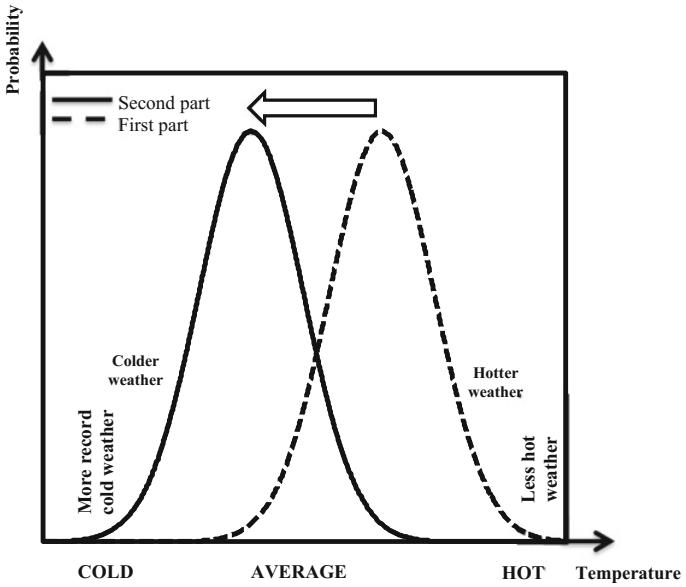


Fig. 5.3 Decreasing trend

In Fig. 5.2 by taking into consideration that the horizontal axis is for temperature records then one can make climate change implication interpretations as "colder," "extremely cold," "hotter," and "extremely hot" weather conditions. The reader may have his/her interpretations.

Decreasing trend or downward jump possibilities are shown in Fig. 5.3 on the basis of two-halves PDFs. The shift in the first part PDF is toward lower data values. Similar to the previous case the statistical parameters are decreased and the slope values can be calculated for each statistical parameter.

The previous graphs collectively imply that although there are changes in the arithmetic average values, but the standard deviation remains the same, i.e., homoscedasticity exists. These three figures are the fundamental assumption in the classical trend determination, because all the linear trend lines do not take into consideration possible changes in the standard deviation.

However, there may also be variations in the standard deviations, which can be identified by the comparison of the two parts' PDFs. The change in the variance, which is also valid for the standard deviation, is referred to as the variability in this book. For instance, the case in Fig. 5.4 an increasing variability is valid, because although there is no change in the arithmetic average the standard deviation has increased again during the half duration of the time series record length. Since, as a general rule, the area under any PDF is equal to one, expansion in this figure implies reduction in the peak value probability. As in the previous cases, the reader may emerge with his/her own interpretation by considering that the horizontal axis is for temperature records.

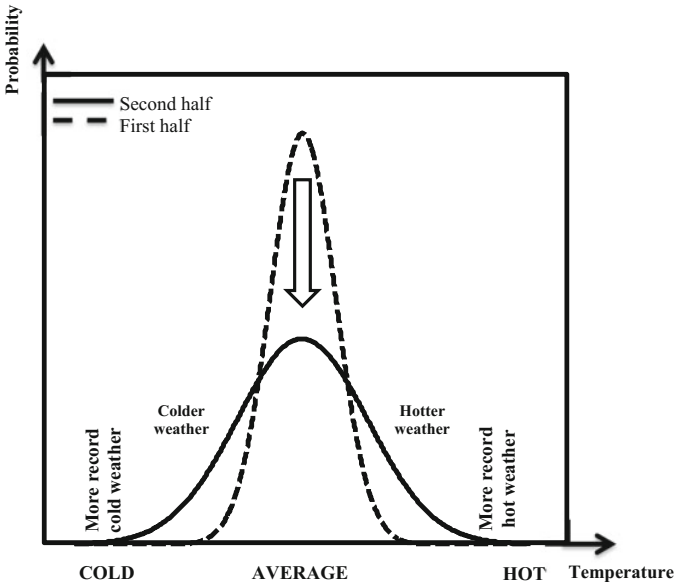


Fig. 5.4 Increasing variability

Opposite of increasing variability Fig. 5.5 is the representative of the decreasing variability. The comparison of the two PDFs indicates that the recent half records had shrinkage in the PDF, which is the reduction in the standard deviation. If after visual inspection of the time series graph one come out with gradual decrease in the standard deviation values then there is a standard deviation trend of which the slope is equal to the difference between the standard deviations divided by the half time series duration.

Figures 5.4 and 5.5 also imply that the underlying time series are first order stationary, because the arithmetic averages remain the same. More detailed information and methodological explanations are presented about the variability in Chap. 7.

It is also possible to have trend and also variability in the same time series, which is the case in some of the natural and environmental time series records. For instance, if one considers the relative positions of the first and second half PDFs as in Fig. 5.6, then s/he can conclude that there is changes in the arithmetic average and in the standard deviation simultaneously.

After all what have been explained about the relative positions of the first and second half PDFs, the reader must have got used to the interpretation. Anyone can interpretate that Fig. 5.7 represents

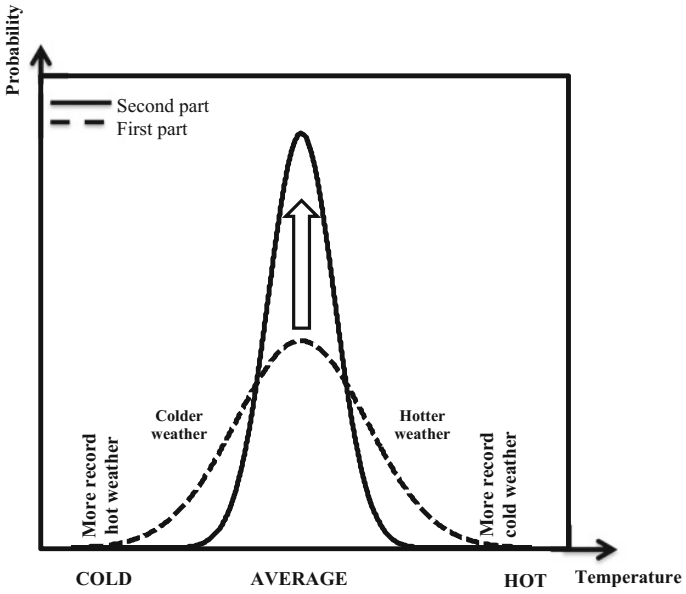


Fig. 5.5 Decreasing variability

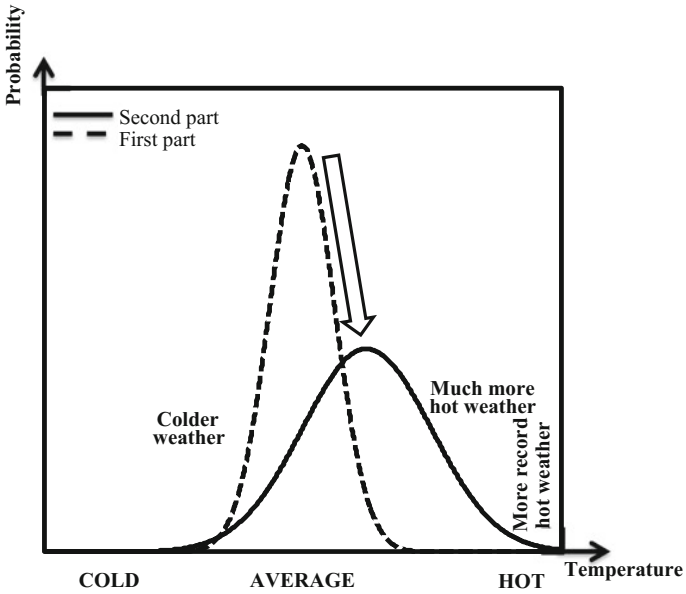


Fig. 5.6 Increasing trend and increasing variability

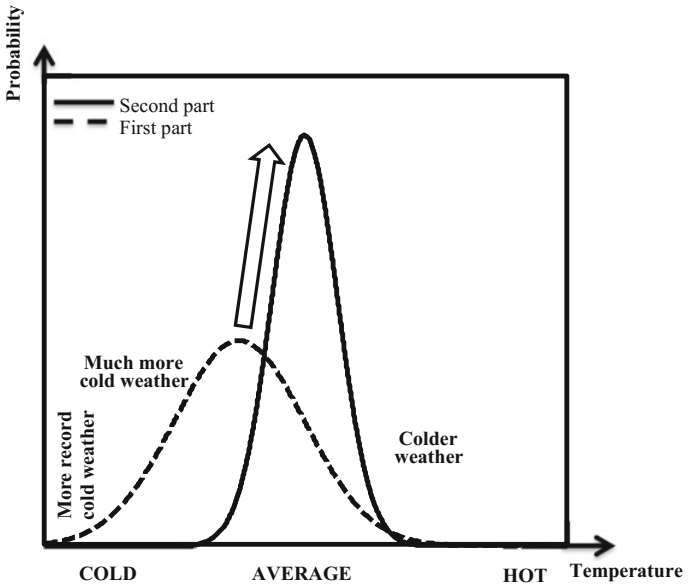


Fig. 5.7 Increasing trend and decreasing variability

5.3 Innovative Trend Identification Methodologies

In the following sequel, the innovative trend identification method presents as a new approach on the basis of subsection time series plots derived from a given time series on a Cartesian coordinate system. In such a plot trend-free time series subsections appear along the 1:1 (45°) straight-line. Increasing (decreasing) trends occupy upper (lower) triangular areas of the square area defined by the variation domain of the variable concerned. The validity of this new approach is documented through a set of Monte Carlo simulations by taking into consideration independent and dependent processes (Sect. 5.3). In this new approach, assumptions for the MK and Spearman's rho (SR) tests are avoided and additionally it is possible to calculate trend magnitude from square area plots.

The basis of the approach rests on the fact that if two time series are identical to each other, their plot against each other shows scatter points along 1:1 (45°) straight-line on the Cartesian coordinate system as in Fig. 5.8a. In the figure, there are 25 data points, which come from a nonnormal PDF. Whatever the time series, whether trend free or with monotonic trends, all points fall on the 1:1 straight-line when plotted. There is no distinction whether the time series are nonnormally distributed, having small sample lengths or possess serial correlations. One important conclusion from Fig. 5.8a is that data values sort themselves in ascending (or descending) order along the 1:1 straight-line. This idea will also be used later in this section in the trend identification procedure.

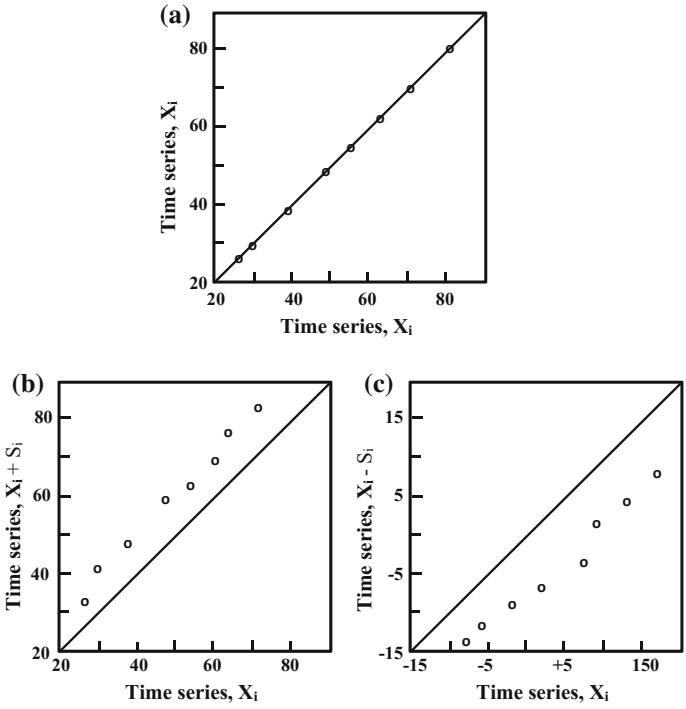


Fig. 5.8 **a** Trendless time series, **b** increasing trend, **c** decreasing trend

The same 25 data points are added with increasing and decreasing trends separately and then they are ordered and plotted against the original (trend-free) time series, which is also sorted in ascending order. The results are shown in Fig. 5.8b, c for increasing and decreasing trends, respectively. It is obvious that in the case of increasing (decreasing) monotonic trend, the scatter points fall above (below) the 1:1 straight-line. For any trial with nonnormal, small sample and serially correlated time series, similar scatter diagrams are obtained for increasing and decreasing trends.

The next question is how could one identify the existing trend in a given time series with respect to the idea of 1:1 straight-line? The answer appears as a plot of the first half of the same time series against the second half according to the above-mentioned idea. In Fig. 5.9a, b, the same time series as shown in Fig. 5.8b, c are used, this time by considering two-halves and the sorting procedure. It becomes obvious that monotone increasing (decreasing) trend in the given time series fall above (below) the 1:1 straight-line. This idea can be used for engineering, environmental, economic, or hydro-climatic time series trend identifications.

On the other hand, it is also possible to have time series with half plots similar to Fig. 5.9 as in Fig. 5.10, where there are scatter points on both sides of 1:1 straight-line. In Fig. 5.10a low (high) values are more (less) in the first half than the

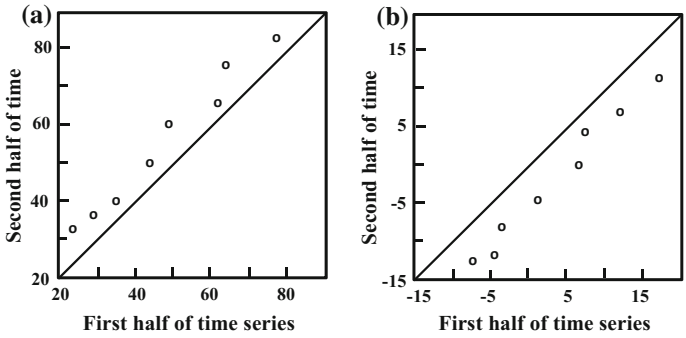


Fig. 5.9 Time series halves with monotonic trends, **a** increasing, **b** decreasing

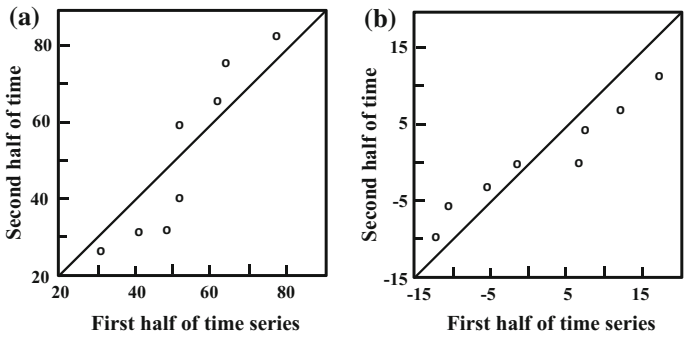


Fig. 5.10 Time series halves with nonmonotonic trends, **a** increasing, **b** decreasing

next half, whereas in Fig. 5.10b the opposite situation occurs. These cases correspond to nonmonotonic trends where within the same time series there are increasing and decreasing trends at different scales even hidden ones (Chap. 6).

In practical applications, a mixture of the all the cases explained in this section appears accordingly, the necessary interpretations can be done for better understanding the composition of the time series structure.

5.3.1 Application

The applications of the innovative trend methodology are presented for different annual runoff and rainfall series recorded at various locations in Turkey in addition to annual Danube river flows. Aslantas and Menzelet Dams are the catchment areas in southern Turkey on Ceyhan River that confluences into the Mediterranean Sea. Cizre streamflow station is on the Tigris River right at the border between Turkey and Iraq. Danube annual streamflow records are from Orshava station in Romania.

Figure 5.11a, b are from two hydrological catchments in Turkey, each reflecting annual flows from 1954 to 2003. For the interpretation of these figures, it is better to think of the annual flows in three clusters as "low," "middle," and "high" flows. In order to make a detailed interpretation, the scatter diagram on 1:1 straight-line graphs are divided into three verbal clusters as "Low," "Medium," and "High." In Fig. 5.11a, "low" flows represent points on the increasing trend upper triangle, which means that there is an increase in the "low" flows during the second half of the historic record (1979–2003) with respect to the first half (1954–1978). In the "medium" cluster, there is almost no trend, and finally, the "high" cluster indicates decreasing trend. All these explanations imply that the annual flow series have a composition of various trend patterns.

The annual flow scatter diagram between two-halves of Menzelet station are shown in Fig. 5.11b, where the "low" flows have slight increasing component within the "low" flow cluster small and big values. The "medium" flow cluster is

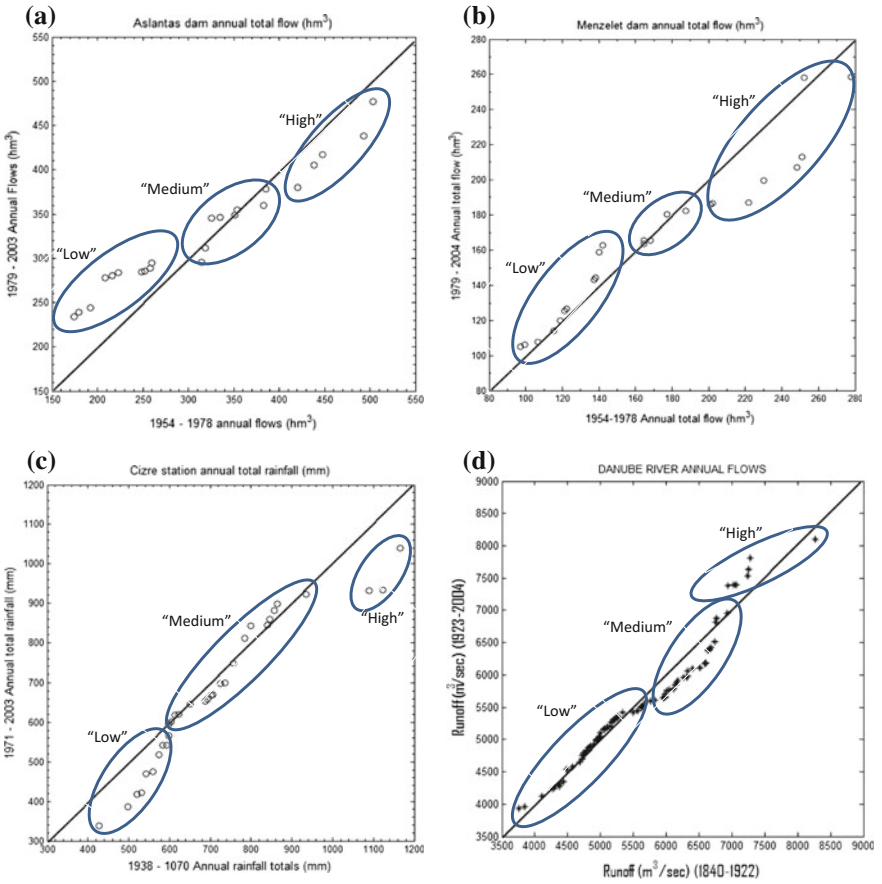


Fig. 5.11 Various 1:1 plots. **a** Aslantas Dam. **b** Menzelet Dam. **c** Cizre Station. **e** Danube River

trend free because the scatter of points concentrate closely around 1:1 straight-line. In the "high" cluster a decreasing trend component is valid. At this station there is a decrease in the "high" flow values; and hence, in the future, water stress is more likely to appear.

In Fig. 5.11c, "low" and "high" clusters indicate decreasing trends, whereas the "medium" cluster is trend free. Comparatively "high" flow trends have shorter duration than in the "low" cluster portion. Most of the duration is occupied by "medium" cluster flows with no significant trend component. Furthermore, the "low" and "high" flows have decreases in the (1971–2003) duration compared to (1938–1970). This also gives the warning that at this station droughts and floods are bound to increase in the future.

Finally, Danube river annual flows do not have any significant trend in the "low" flow cluster, which includes all the annual flows less than about 5750 m^3/s (Fig. 5.11d). "Medium" flows have some decreasing trend and "high" flow cluster has slightly significant increasing level.

Based on the above explanations, the following important points can be summarized about the innovative trend methodology.

(1) If scatter points on the first quadrant of the Cartesian coordinate system fall on another straight-line parallel to 1:1 straight-line, then there is a monotonic increasing (decreasing) trend depending on the fall of the scatter points onto the upper (lower) triangular area of the scatter region,
(2) The closer the scatter points are to the 1:1 straight-line, the weaker the trend magnitude (slope),
(3) In the case of nonmonotonic trends (i.e., composition of various trends in the time series), the scatter points take their positions on a curve.

This innovative trend method does not require restrictive valid assumptions whatever the sample size, serial correlation structure of the time series, and non-normal PDFs.

5.4 Innovative Trend Simulation

Trend analyses occupy a significant role in the climate change studies since almost four decades. It is significant to try and identify monotonic trends in a given time series so as to make future predictions about the possible consequences on the urban environment, economics, water resources, agriculture, environmental, and many other socioeconomic aspects of the life. Although there are now classically accepted and frequently used trend tests in the open literature such as MK trend analysis and SR test, they are based on some restrictive assumptions as normality, serial independence, and rather long sample sizes. Besides they search for a single monotonic trend without any specification such as "low," "medium," and "high" values, which may have different trend patterns. Many time series records have

serial dependence and, therefore, it is very helpful to provide a methodology, which is not affected from such restriction. It is the main purpose of this section to provide simulation results and applications of an earlier innovative trend analysis methodology based on the 1:1 (45°) straight-line comparison of the scatter points on a Cartesian coordinate system.

Natural and human activities affect different processes in a continuous manner and their impacts appear in the forms of trends or sudden jumps. Some particular natural phenomena such as El Niño, as well as all kinds of large scale water resources development projects, may alter hydrological processes and may lead to abrupt changes in the hydrological time series (Xiong and Guo 2004). The presence of deterministic trends in the time series may provide information about the future evolution of the process or at least on the possible modifications. In practical applications, the knowledge of the trend for a given variable of interest may help to forecast future realizations and to design future scenarios. Nowadays, with the growing importance of climate change assessment, trend detection, and evaluation are subjects of intensive scientific research (Brunetti et al. 2001; Burn et al. 2002; Kahya and Kalaycı 2004; Groisman et al. 2004; Cohn and Lins 2005; Barbosa et al. 2008), as also testified in the recent fourth assessment report of the Intergovernmental Panel on Climate Change (IPCC 2007). One branch of climate change science is devoted to analyzing the past climate events and inside this branch trend detection and statistical significance testing assume an important role (Trenberth 2007).

Natural and man–made effects are defined as the long-term behavior of concerned variables on the average, which provides distinctive features for future behaviors of the same variable. During the last four decades, the most sought such behavior is the possibility of monotonic trend existence in a given time series, because the current day change impacts and causative decisions require gradual increasing or decreasing trends. Especially, time series records are searched for two reasons; the first one is trend identification, and then its magnitude determination as reflection of the "increasing" or "decreasing" quantities. Although there are trend identification methods, which provide answers for the existence of trends, but the magnitude is measured either by linear regression approach (Hirsh and Slack 1984; Lettenmaier et al. 1984) or through the median slope calculation according to Sen (1978) procedure. This estimator is robust to the effect of outliers in the series. It has been widely used to compute trends in hydro-meteorological series (Wang and Zhou 2005; Zhang et al. 2001).

None of the classical trend tests such as the MK test takes into account classical parametric and most commonly used serial correlation and, hence, they require independence structure in the applications. In general, independence test can be carried out mainly by examining the autocorrelation coefficients of the time series. If the absolute values of the autocorrelation coefficients for a time series consisting of n observations are not larger than the typical critical value, i.e., $1.96/\sqrt{n}$ corresponding to the 5% significance level (Douglas et al. 2000), then the observations in this time series can be accepted as being independent from each other. The

significance of the trend is determined using Kendall's test because it does not assume an underlying probability distribution function (PDF) of the data series.

The main purpose of this section is to present extensive computer simulation for robust trend identification procedure as already proposed by Şen (2012), which is not dependent on any restrictive assumption as serial correlation, nonnormality, and sample number. The procedure is based on the plot of time series two-halves against each other after sorting in ascending order. This procedure helps to identify trends distinctively in the low, medium, and high values also. The difference of this section lies in its extensive independent process and dependent first order Markov process simulation results, which indicate the relationship between the trend slopes and first order serial correlation coefficient. Additionally, the comparisons of this trend procedure with the classical methodologies including MK and SR trend statistics and Sen's trend slope are given in table form with necessary interpretations.

5.4.1 Fundamental Methodology

As mentioned in Sect. 5.3, a new trend analysis methodology by Şen (2012) depends on the 1:1 (45°) straight-line on a Cartesian coordinate system, where it corresponds to trend-free case and any deviation from this line indicates trend existence and the closer is the plots to 1:1 (45°) straight-line, the smaller is the trend slope.

In the innovative trend identification methodologies as explained in Sect. 5.3 upper and lower triangular areas correspond to trend existence. Figure 5.12 is prepared as the plot of sorted time series versus two trend-embedded synthetic time series. Each series is obtained by adding a linear monotonic increasing and decreasing trend into the original time series in the upper and lower graphs of Fig. 5.12. In the middle square, plots versus trend-free time series are given after sorting in ascending order. The final product yields the fact that the upper (lower) triangular area includes increasing (decreasing) trends, respectively. Additionally, on the 1:1 straight-line plots, increasing trend time series points can be interpreted by considering low and high values subjectively in two groups. Hence, since low values are concentrated near the 1:1 straight-line, the trend existence is weaker than the group of high values, which significantly deviate from the straight-line 1:1 straight-line. On the other hand, in the lower triangular area, the time series have low values' cluster, this time away from the line, whereas high values approach the 1:1 straight-line implying comparatively weaker trend existence in the structure of the time series considered all based on visualization.

In Fig. 5.13, previous increasing and decreasing trend time series are plotted within themselves by having the whole series first into two and then sorting them in ascending order. The result is increasing (decreasing) trend in the time series according to their complete structure. This point provides a new way of trend assessment, which takes into account not only the ranks (nonparametrically) but also the measurements parametrically.

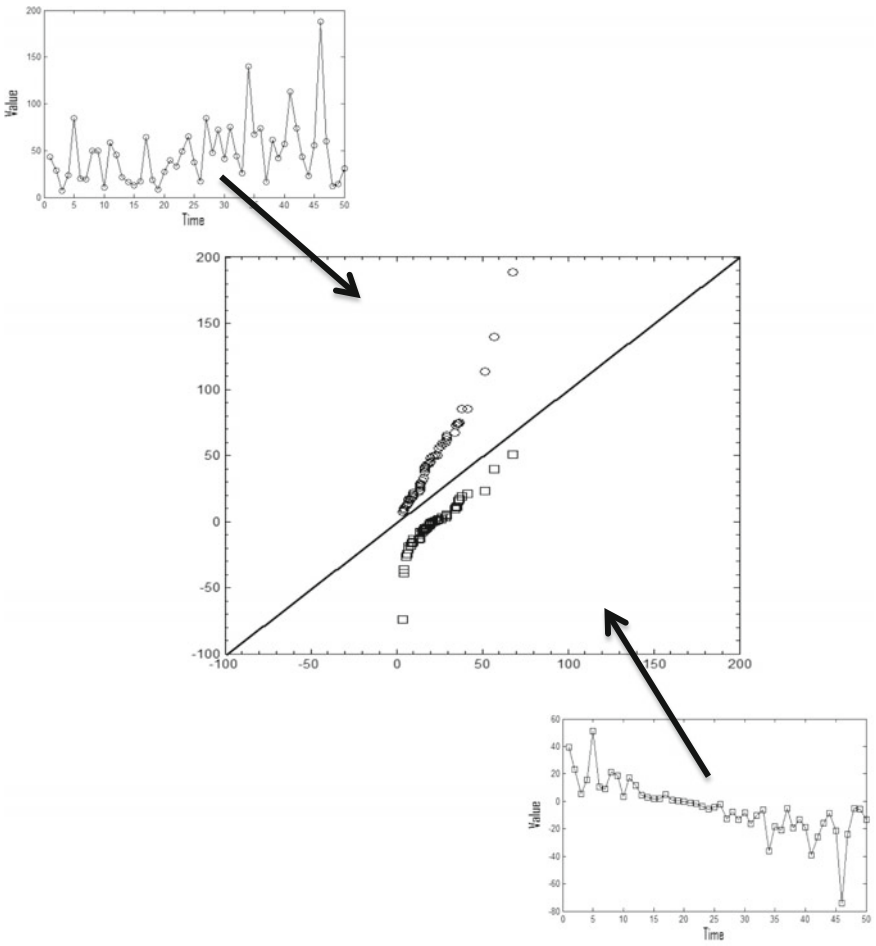


Fig. 5.12 Decreasing and increasing trends versus trend-free time series

In the following sections, extensive simulation study is performed for the validity of the innovative trend methodology by use of independent and dependent processes.

5.4.1.1 Simulation Methodology

There are different aspects in time series analyses depending on the purpose, which may take shape according to needs in any planning, design, and operation and maintenance stages. The prime goal is to deduce some useful and objective information for future works that support final decisions. Initially, Hazen (1914) was interested in extending the past records to future predictions and for this purpose he designed a very simple pre-computer era procedure by writing each one of the past records on separate paper pieces, mixed them thoroughly in a bag and

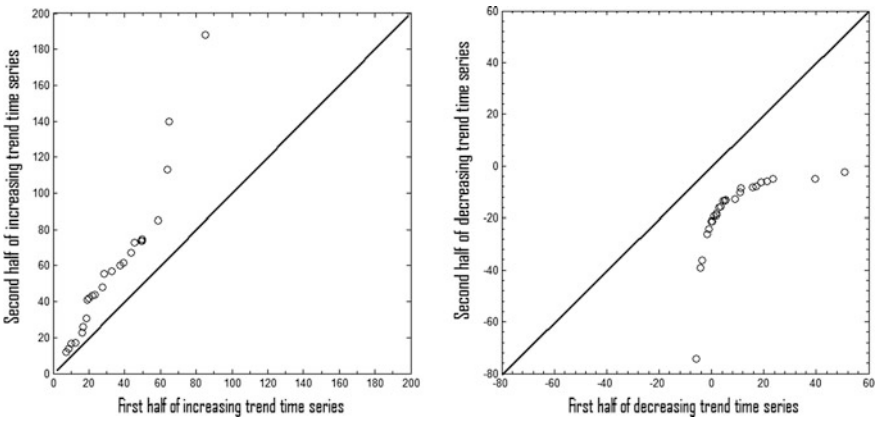


Fig. 5.13 Half time series

then drew one by one for future time series construction. This primitive procedure had the assumptions of almost trend-free synthetic time series with the same statistical similarity to past time series. The only difference was in the sequence of past record values. With the appearance of digital computers in 1950s, stochastic processes became in use for the analysis of historical records with the purpose of constructing their future replicates synthetically in such a way that statistical properties are indistinguishable from the historical records (Şen 1974). Autoregressive (AR) and autoregressive integrated moving average (ARIMA) models in various degrees of order become in use in many disciplines including hydrology for water resources planning, operation, and management stages (Box and Jenkins 1970; Montanari et al. 1997).

Figure 5.12 can be used as a template to identify trend existence in a given time series. For this purpose, the square area template in the first quadrant can be thought in three portions. These are enumerated below:

(1) The main diagonal, 1:1 (45°) straight-line presents no trend line,
(2) The upper right angle triangular area is for increasing trends,
(3) The lower right angle triangular area is for decreasing trends.

These points will be explained by simulation studies based on dependent and independent process, trend free and trend-embedded time series in addition to practical applications. Theoretically, in case of exactly the same two time series, there is no areal scatter on the coordinate system but the scatter is along the 1:1 straight-line only. This means that each time series is its own reflection on the 1:1 straight-line (see Fig. 5.8a), which corresponds to trend free case, whereas upper (lower) triangular area is for increasing (decreasing) trends. Figure 5.14a presents 30 points from stochastic processes, where all the points are aligned along the 1:1 straight-line in a random scatter manner similar to Fig. 5.8a. Figure 5.14b presents

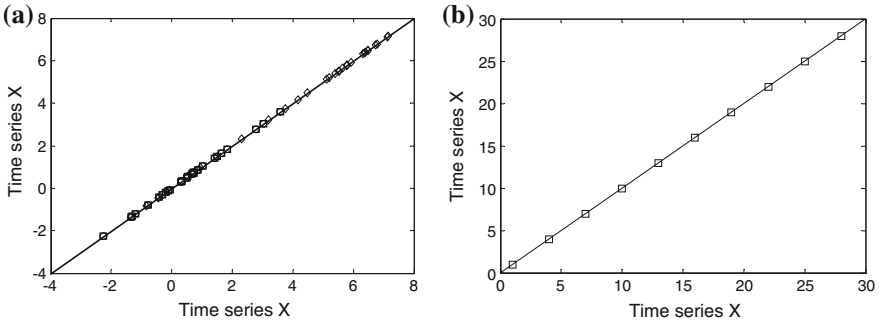


Fig. 5.14 **a** Stochastic time series (o independent; ◦ dependent), **b** regular time series

time series with regular (deterministic) increments. Such plots have the following results:

(1) Time series own reflections appear along the main diagonal (1:1 straight-line) scatter irrespective of trend or trend-free serial structure,
(2) Whatever the PDF of the time series, the end plot also appear along the same diagonal,
(3) Serial correlation of the time series does not play any role in such plots,
(4) Seasonality component also does not affect the appearance along the main diagonal scatter,
(5) The number of sample has not role and again the plots appear along the 1:1 (45°) straight-line.

After all these points, the main question is whether such plots may help to identify trend (or trends) in a given time series?

This question brings to mind similar to plot of a given time series versus itself, what happens when the first half of the series is plotted against its second half time series? For this purpose, the same time series maybe fragmented into mutual and successive half subseries. A very significant clue from Figs. 5.13 and 5.14 is that along the main diagonal the points are sorted according to ascending order automatically. This point gives the idea of sorting the two-halves into ascending orders and then to plot the first half versus the next on the Cartesian coordinate system. This opens the door to compare "low" ("medium," "high") values with "low" ("medium," "high") values of the two-halves.

In order to explain some of the main points in the innovative trend methodology, first of all, trend-free independent (normal or non-normal) processes are generated with zero mean and unit standard deviation, which is then embedded with a sequence of monotonic trends by considering a set of trend slopes, d, (−0.009:0.002:0.009). The length of the generated synthetic sequence is adapted as 10,000, which is then divided into two-halves of 5,000 elements each. Inspiration

from the above explanations gives rise to the following significant points for the application of the methodolog:

(1) Generate a set of trend-embedded sequences and divide them into two-halves,
(2) Sort each half in ascending order,
(3) Plot the first half against the second half on the square area (on the Cartesian coordinate system).

Figure 5.12 is the end product of such a procedure with different time series and their signatures on the square area, which leads to the following inferences:

(1) Trend-free halves plot appears along the 1:1 (45°) straight-line,
(2) Increasing (decreasing) trends are within the upper (lower) triangle of the square area,
(3) They are all in the forms of straight-lines parallel to each other with 45° slope, which implies that the trend slope, d, in the original series does not have any effect on these straight-lines,
(4) As the trend slope, d, in a time series increases, corresponding straight-line plot appearances on the square area get away from the trend-free line (main diagonal, 1:1 or 45° line),
(5) Positive and negative trend slopes have reflective effects with reference to no trend (1:1 straight-line).

These points indicate that the innovative trend identification methodology does not give information only about the existence of the trend in the time series but additionally about its magnitude (slope, d). The significant conclusion is that any plot of two-halves from a given time series in ascending order is enough to identify trend existence and its magnitude irrespective of data length. In Fig. 5.15, although trends are taken from respective 10,000 length time series, just for the sake of clarity and explanation only 1,000 points are shown. In each one of these time series increasing and decreasing monotonic trends are shown explicitly with their corresponding consequences on the square area.

5.4.1.2 Dependent Process Simulation Results

In order to perform the power of the proposed methodology, a set of Monte Carlo simulations are presented by taking into consideration first order (Markov) autoregressive (AR) stochastic process with PDFs. The simulation procedure first generates synthetic time series, X_i, of length 10,000 values according to the following model:

$$X_i = \mu + \rho(X_{i-1} - \mu) + \sigma\sqrt{(1-\rho^2}\varepsilon_i, \tag{5.1}$$

where μ and σ are the mean and standard deviation of the process; ρ is the first order serial correlation coefficient and ε_i is the normal independent process with zero mean and unit variance, NIP (0, 1). The set of simulations is based on the serial

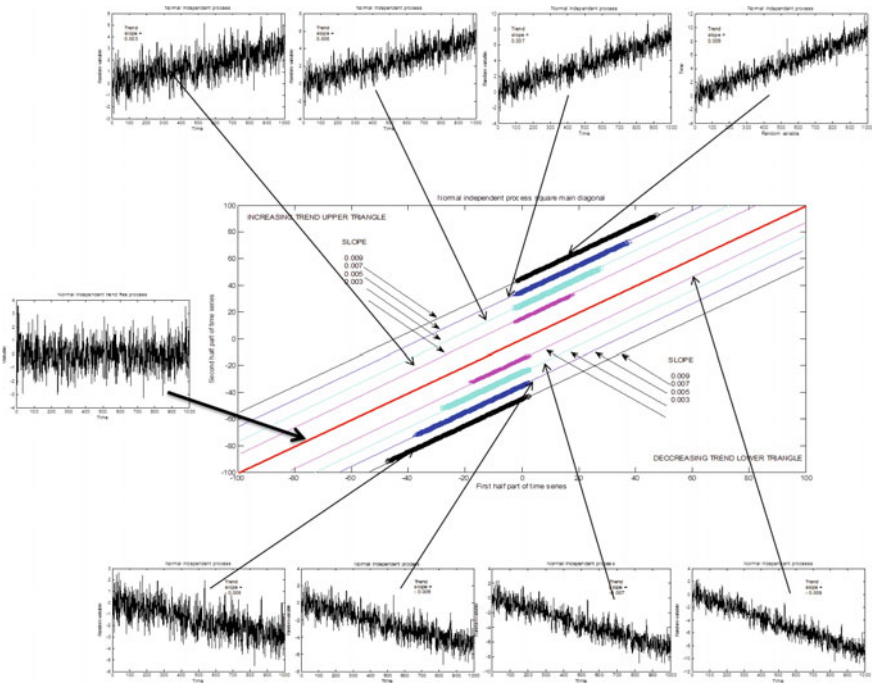


Fig. 5.15 Independent process trends on square area

correlation coefficients, $\rho = 0.1\ (\pm 0.2) \pm 0.9$. Equation (5.1) generates trend-free stationary time series, which are converted to nonstationary forms by embedded with increasing and decreasing trend components of slopes, $d = 0.001\ (\pm 0.02) \pm 0.09$. The slope is embedded through the simple linear trend component addition to the basic stochastic process according to $X_i + d_i$, where $i = 1, 2, \ldots, 10{,}000$.

Figure 5.16 summarizes the simulation results from above-mentioned AR process given a high serial correlation coefficient, $\rho = 0.9$, with a set of embedded trends. Each one of the thick lines includes 5,000 generated normal dependent values (because 10,000 values were generated for each simulation) as the first half versus the second half. The fine lines are drawn through these thick simulation results in each triangular area. It is obvious that as the absolute value of the trend slope increases the results fall away from the 1:1 straight-line. During the simulation, it is noted that the straight-lines in Fig. 5.16 are a result of normal PDF.

Comparison of Figs. 5.15 and 5.16 indicate that whether the time series is independent or dependent, there is no difference in the square area procedure and as long as the basic time series has a monotonic trend, the appearance of the two-halves sorted magnitude plots will appear along 45° straight-lines without any distinction. This statement alleviates the drawback of the MK trend test, which requires independent data. Additional illuminating points can be drawn from the

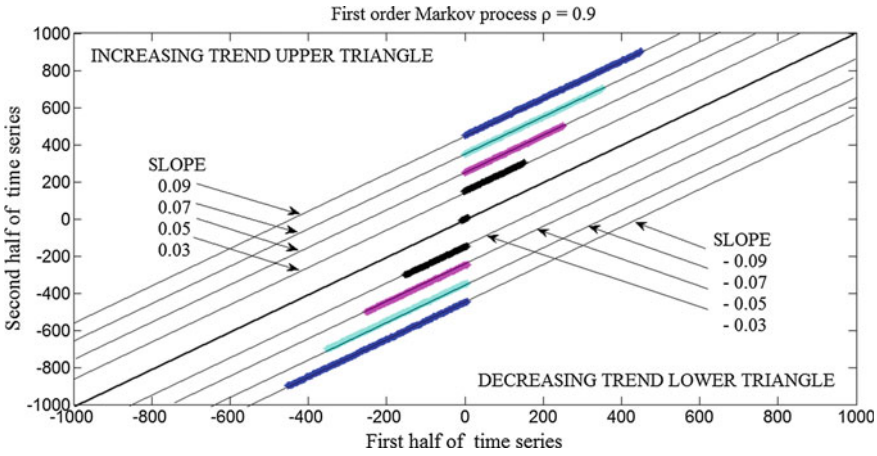


Fig. 5.16 Trend lines with respect to 1:1 straight-line for a set of slopes

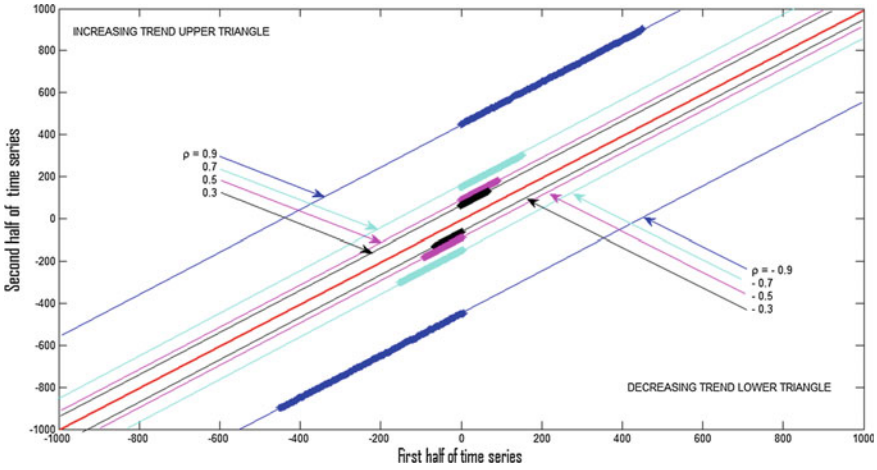


Fig. 5.17 Trend chart of trend (d = 0.009) embedded first order AR processes

square area plot in Fig. 5.17, where this time the trend slope is kept constant (d = 0.009) and trend appearances are shown for a set of serial correlation coefficients (−0.9; −0.7; −0.5; −0.3; 0.0; 0.3; 0.5; 0.7; 0.9).

This figure indicates that the upper (lower) triangle include positive (negative) correlation coefficient cases, which is another improvement on the MK test, where the serial correlation cannot be accounted at all in the calculations. The more the serial correlation coefficient absolute value, at the same trend magnitude (herein, d = 0.009), the more effective is its occurrence on the square area template.

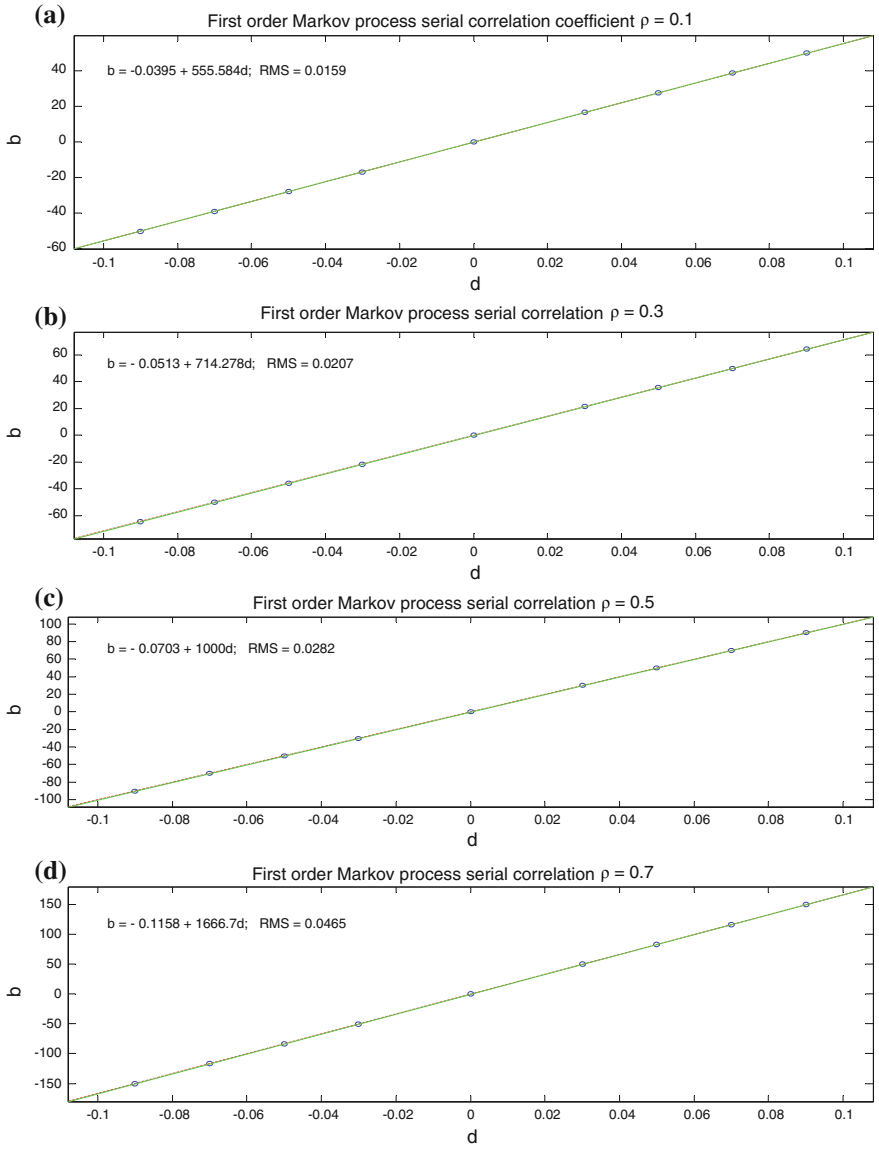


Fig. 5.18 Trend magnitude square area plot slope relationship for **a** $\rho = 0.1$; **b** $\rho = 0.3$; **c** $\rho = 0.5$; $\rho = 0.7$

Figure 5.18 provides linear relationship between the trend and square area template slopes for given serial correlation coefficient. In the same figure corresponding root mean square (RMS) errors are also presented and they are all very small within practically acceptable limits.

Table 5.1 Trend slope, serial correlation coefficient and trend line intersection

Trend slope, d	Independent process	First order stochastic serial correlation coefficient (ρ)				
	0.0	0.1	0.3	0.5	0.7	0.9
−0.09	−45	−50.048	−64.343	−90.080	−150.133	−450
−0.07	−35	−38.934	−50.058	−70.080	−116.800	−350
−0.05	−25	−27.824	−35.772	−50.080	−83.465	−250
−0.03	−15	−16.713	−21.486	−30.078	−50.131	−150
0.00	0.0	0.0	0.0	0.0	0.0	0.0
+0.03	15	16.624	21.372	29.920	49.871	150
+0.05	25	27.736	35.658	49.921	83.205	250
+0.07	35	38.846	49.944	69.922	116.538	350
+0.09	45	49.957	64.223	89.922	149.872	450

It is obvious that there is almost perfect linear relationship between the trend magnitude (slope, d) and the trend representative line on square area template for any given serial correlation coefficient. Table 5.1 provides numerical values of the relationship between ρ, d, and b.

This table can be used to determine the magnitude of monotonic trend in any time series provided that the serial correlation coefficient and the slope on the square area template are determined.

After all what have been explained above, it is possible to state that the new methodology yields information about the low values of the first half with low values of the second half leading to the following conclusions:

(1) If low, high, medium, and high value plots of the two-halves are above (below) the 1:1 (45°) straight-line, then there is an increasing or decreasing trend,
(2) In case of increasing (decreasing) trend, if all the low, medium, and high values fall on almost parallel line to 1:1 (45°) straight-line then there is a single monotonic trend in the time series,
(3) Otherwise, low, medium, and high values may have different positions on the plot area, and this implies to the existence of various sub-trends in the time series structure,
(4) The proposed methodology can provide detailed information about the low, medium, and high value trends in the time series and their relative effectiveness to each half.

In Fig. 5.19, a set of trend-embedded (d = 0.009) simulation synthetic sequences is given, for a set of autocorrelation coefficients.

The corresponding plots of these time series around 1:1 (45°) straight-line are given in Fig. 5.20 for various serial correlation coefficient. Again straight-lines parallel to 1:1 (45°) and basic line are plotted based on half time series simulation result values (5,000 values) according to sorting procedure. Since embedded trends

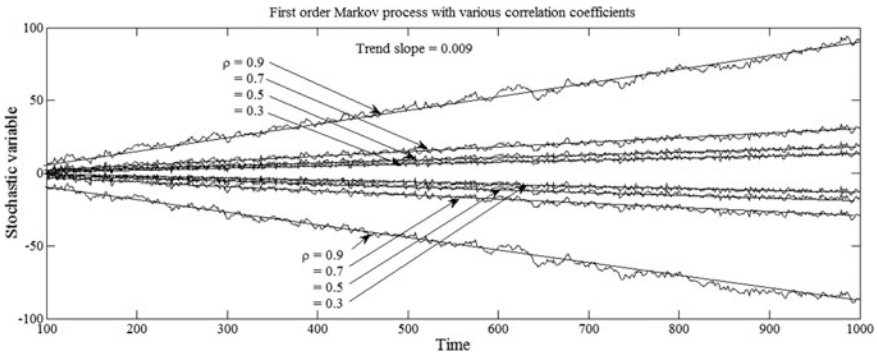


Fig. 5.19 Increasing and decreasing trends

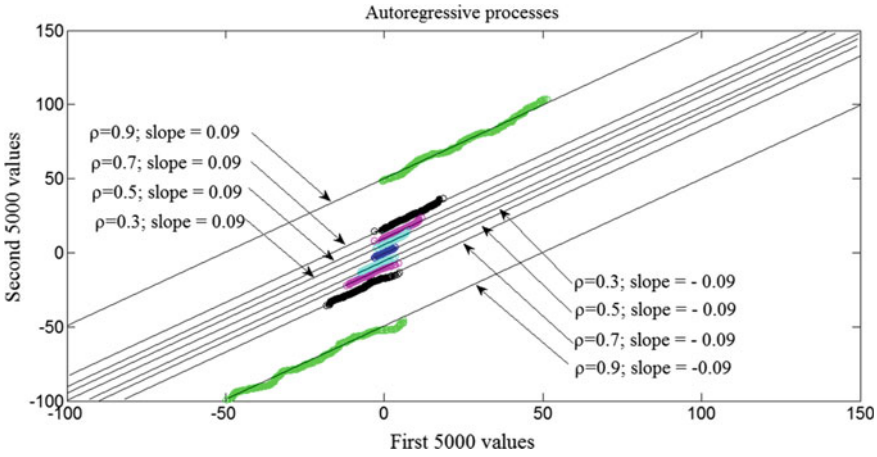


Fig. 5.20 Trend lines (0.09) with respect to 1:1 line for a set of correlation coefficient

are monotonic, the lines are parallel to 1:1 (45°) straight-line. One can conclude from this figure that as the absolute value of the serial correlation coefficient increases the trend representing lines get away from 1:1 (45°) straight-line basic line. The chart in this figure helps to answer to the following questions:

(1) Is there a linear trend embedded in the given time series?
(2) What is the serial dependence coefficient (ρ) in the series?
(3) Is it possible to identify the trend in a given series without pre-whitening?

Figure 5.21 represents comparatively weaker (d = 0.009) trend for the same set of serial correlation coefficients. There is no change in the previous interpretations and the straight-lines get away from the basic 1:1 line.

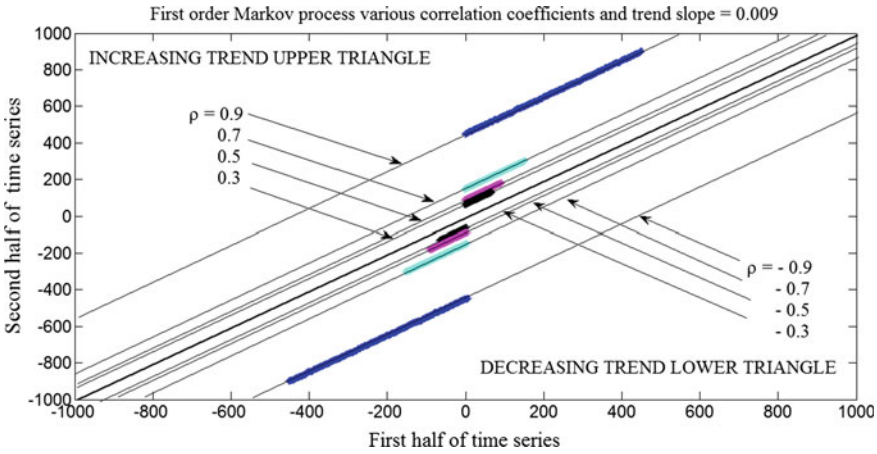


Fig. 5.21 Trend lines (d = 0.009) with respect to 1:1 (45°) straight-line for a set of correlation coefficient

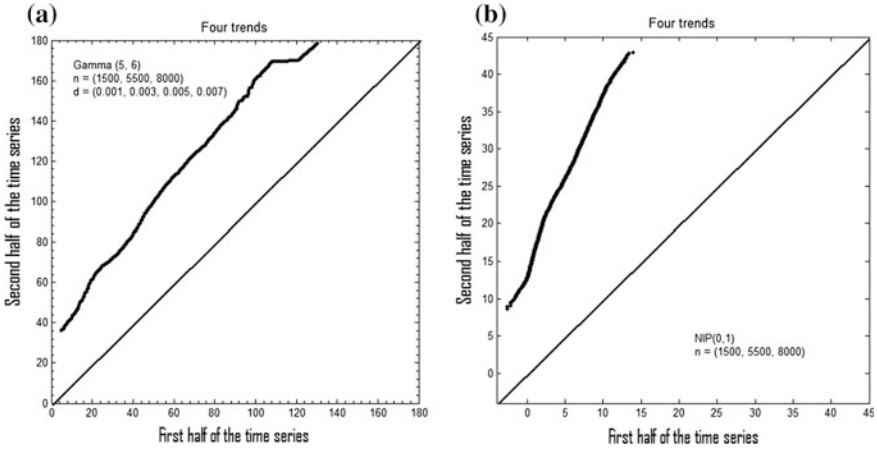


Fig. 5.22 Nonmonotonic trends, **a** Gamma PDF. **b** Normal independent process

All the previous figures had monotonic trends as in Fig. 5.21, but to expand the applicability of the proposed 1:1 (45°) straight-line methodology. Figure 5.22 is given as a representative example among numerous simulation results that increasing but nonmonotonic trends can also be depicted by the same methodology.

In Fig. 5.22a, Gamma PDF simulation results are presented with shape and scale parameters as 5 and 6, respectively. Four different but successive trends are embedded onto the basic time series, where the first, second, third, and fourth trend components appear between 1–1,500; 1,501–5,500; 5,501–8,000; and 8,001–10,000, and trend slopes are 0.001; 0.003; 0.005; and 0.007, respectively. The same

simulation is repeated for normal independent process, NIP (0, 1), in Fig. 5.7. One can deduct from Fig. 5.22 the following points:

(1) Even though the PDF is Gaussian the final trend plots in Fig. 5.22b does not appear along a straight-line parallel to 1:1 (45°) straight-line contrary to Figs. 5.20, 5.21, 5.22a,b and 5.31 where only monotonic trends exist,
(2) The 1:1 (45°) straight-line methodology is capable of identifying increasing but nonmonotonic (multiple) trends. This provides a possibility even to identify hidden (short duration) sub-trends in the whole time series,
(3) In the case of more than one successive trend, the plots according to 1:1 (45°) straight-line method appear on the upper (piecewise increasing) and lower (piecewise decreasing) triangular areas as curvature (nonlinear) traces,
(4) Combination of monotonic and piecewise trend embedded time series performances mentioned above, lead to deduction that any nonparallel line implies a combination of various scale nonmonotonic trends in the same time series.

There will not be any uncertainty of a trend estimate under few extreme minimum/maximum values, because the procedure in this section singles them out on the 1:1 (45°) straight-line plot domain. However, in conventional trend identifications, especially regression line fitting to a given time series will be affected by the extreme values. In case of small sample size of a time series, again since each couple of points from two-halves appears without any influence on other points on the scatter diagram in Fig. 5.22, the possible trend component will show itself.

5.5 Innovative Trend Significance Test

Time series might embed characteristics of past changes concerning climate variability in terms of shifts, cyclic fluctuations, and more significantly in the form of trends. Identification of such features from the available records is one of the prime tasks of hydrologists, climatologists, applied statisticians, or experts in related topics. Although there are different trend identification and significance tests in the literature, they require restrictive assumptions, which may not be existent in the structure of time series. In this section, a method is suggested with statistical significance test for trend identification in an innovative manner. This method has nonparametric basis without any restrictive assumption and its application is rather simple with the concept of subseries comparisons that are extracted from the main time series. The method provides privilege for selection of sub-temporal half periods for the comparison, and finally, generates trend on objective and quantitative manners. The necessary statistical equations are derived for innovative trend identification and statistical significance test application.

5.5.1 Deterministic Basis

In order to explain the basic idea behind the innovative methodology, first of all a linear trend function is considered between an independent time variable, t, and dependent variable, y, as,

$$y = a + bt \tag{5.2}$$

where a and b are the intercept on y axis and slope parameters, respectively. In a deterministic methodology, there are few alternatives to determine the parameter values.

(1) The simplest way is applicable provided that the independent, (t_1, t_2), and corresponding dependent, (y_1, y_2), variable pairs are known as two points. The substitution of these values into Eq. (5.2) helps to calculate the parameters from resulting two equations by elimination methodology,
(2) Calculation of the slope value, $b = (y_2 - y_1)/(t_2 - t_1)$, and its substation into Eq. (5.2) leaves only one unknown, which can then be calculated by substitution of coordinates either one of the given points leading to $a = y_1 - bt_1$ or $a = y_2 - bt_2$,
(3) If a regular sequence of n independent time variable, $(t_1, t_2, \ldots, t_n)$ and corresponding dependent variable sequence, $(y_1, y_2, \ldots, y_n)$ are given then one can calculate the unknown parameters, (a and b), either by considering any two points and apply the same methodology as in the two previous items or by consideration of all the given set of coordinates simultaneously through a linear regression methodology.

The core of the innovative trend test methodology is similar to this last item parameter calculation. The question is, provided that independent and dependent variable sequences are available, how to obtain the straight-line trend parameters? The explanation of this point can be given through the following deterministic numersimulation results are presented with shapeical example.

Let the parameter values in Eq. (5.2) be as $a = 2.5$ and $b = 0.25$ in addition to the number of data, say, $n = 126$. It is obvious that the result will appear as a straight-line given in Fig. 5.23a.

In Fig. 5.23b the innovative trend plot of the same deterministic data is presented as already explained by Şen (2012, 2014). In brief, the innovative trend plot requires division of the given time series into two-halves each sorted in ascending order, and finally, plot of the first half versus the second half as in Fig. 5.23b. In the preparation of this figure dependent variable sequence values, $(y_1, y_2, \ldots, y_n)$, are used for data line construction. The following features can be deduced from Fig. 5.23b.

(1) Deterministic dependent variable half plots fall on a definite straight-line referred to as "Data line,"

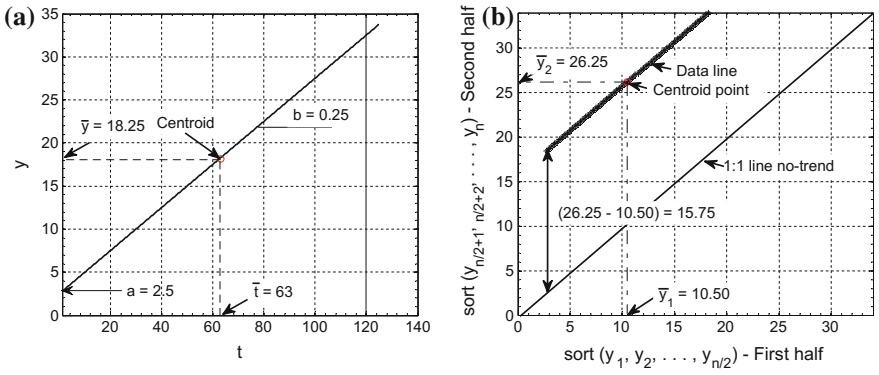


Fig. 5.23 Deterministic trend. **a** Data plot. **b** Innovative trend plot

(2) 1:1 (45°) straight-line indicates neutral trend (notrend) line and any deviation from this line indicates existence of a trend in the given dependent variable (Şen 2012). In Fig. 5.23b there is an obvious increasing trend because the data line is above the 1:1 (45°) straight-line,
(3) The arithmetic averages of the two-halves appear as the "Centroid point" that falls on the data line,
(4) The vertical difference between the data and 1:1 (45°) straight-lines is related to the slope of the existing trend in the dependent variable (Şen 2014),
(5) The vertical distance is equal to the difference between the arithmetic means of the two-halves, which appears as 15.57 in Fig. 5.23.

In the previous studies, there have not been any formulation derivations but qualitative assessments only. In this chapter, new numerical trend identification procedure and significance test are presented in the following sequel.

After the completion of above five steps one can calculate the slope, b, of the trend according to the following expression:

$$b = \frac{2(\bar{y}_2 - \bar{y}_1)}{n}, \tag{5.3}$$

where $\bar{y}_1$ and $\bar{y}_2$ are the arithmetic averages of the first and the second halves of the dependent variable, y, sequence, and n is the number of data. The substitution of the numerical values as $n = 126$, and the arithmetic averages from Fig. 5.23b as $\bar{y}_1 = 10.50$ and $\bar{y}_2 = 26.25$ into Eq. (5.3) yields $b = 0.25$, which is exactly the same value in Fig. 5.1. Hence, the procedure in the preparation of Fig. 5.23b with the use of Eq. (5.3) helps to find the slope of the trend in a given time series.

On the other hand, the calculation of y axis intercept, a, on the vertical axis in Fig. 5.23a, can be achieved according to the second item in the abovementioned parameter value calculations. For this purpose, one needs to know the coordinates of a single point, which is logically adapted as the arithmetic averages of time

sequence, $\bar{t}$, and, $\bar{y}$, of the dependent variables, respectively (Fig. 6.23a). The substitutions of these coordinate values and the slope from Eq. (5.3) into Eq. (5.2) gives the trend intercept parameter estimation as

$$a = \bar{y} - \frac{2(\bar{y}_2 - \bar{y}_1)}{n}\bar{t} = \bar{y} - b\bar{t}. \tag{5.4}$$

Finally, the substitution of the relevant quantities into the basic equation (Eq. 5.2) leads to the most detailed formulation of the innovative trend expression as

$$y = \bar{y} - \frac{2(\bar{y}_2 - \bar{y}_1)}{n}\bar{t} + \frac{2(\bar{y}_2 - \bar{y}_1)}{n}t = \bar{y} - b(\bar{t} - t). \tag{5.5}$$

In case of notrend, $\bar{y}_1 = \bar{y}_2$ and this last expression leads to $y = \bar{y}$, which means that the time series has a constant arithmetic average and, hence, no trend for which the innovative trend slope is 1:1 (45°) line as in Fig. 5.23b.

5.5.2 Stochastic Basis

In case of stochastic variables most often one has hydro-meteorological time series that require trend search for different purposes and most often for the climate change possibilities. In general, any hydro-meteorological time series has deterministic components as possible jumps, periodicities, and trends in addition to the stochastic residuals that are free of any deterministic parts. Herein, the identification of trend component is explained similar to the deterministic basis as explained in the previous subsection. In order to show the effectiveness of the proposed model, two synthetically generated time series are examined for the establishment of the stochastic basis of the innovative trend identification. The first example is for a normal probability distribution (PDF) and the second one is for a skewed Gamma type PDF.

5.5.2.1 Normally Distributed Stochastic Time Series

A synthetic time series is generated according to a first order Markov process with the mean, standard deviation and first order correlation coefficient values as μ = 10, σ = 5 and ρ = 0.5, respectively, with normal (Gaussian) PDF random component. The length of the stochastic time series is considered as n = 1,000 and synthetically a trend is embedded with slope b = 0.015. The generated synthetic time series with these specifications is presented in Fig. 5.24a. Generation of synthetic sequences with different serial correlation coefficients and their innovative trend plots fall on the same "Data line" within practically acceptable sampling relative errors of less that ±5%.

All the necessary quantitative values are provided on the innovative trend plot in Fig. 5.24b. The substitution of these mean values into Eq. (5.3) yields the slope

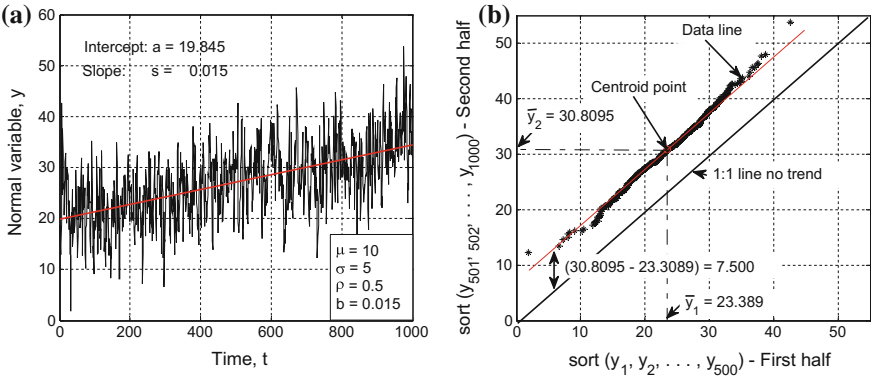


Fig. 5.24 Normal stochastic process **a** time series and trend, **b** innovative trend plot

value of the embedded trend as $s = 2(30.8095 - 23.389)/1000 = 0.0148 \cong 0.015$, which is the value of the embedded slope in the stochastic process.

The arithmetic averages of the time series independent time, t, and dependent, y, variables are $\bar{t} = 500$ and $\bar{y} = 27.159$, respectively. The corresponding intercept value can be obtained from Eq. (5.4) by substitution of the relevant values as $a = 27.159 - 0.0148 \times 500 = 19.76$, which is within less than $\pm 5\%$ relative error, r_e, from the value in Fig. 5.24a. Herein, $r_e = 100 \times (19.845 - 19.760)/19.845 = 0.42\% < 5\%$, and this value is within the acceptable limit of error.

5.5.2.2 Gamma Distributed Stochastic Time Series

In practical applications Gamma PDF is frequently used, because depending on the parameter values different PDF types appear. In the simulation, again $n = 1{,}000$ data set is generated as dependent variable, y, with the trend slope, $b = 0.020$, shape parameter, $\alpha = 2.3$, scale parameter, $\beta = 5.4$ and correlation coefficient, $\rho = 0.5$. The final result with trend component is presented in Fig. 5.25a, which is one of the samples from an ensemble of different 1,000 length synthetic series.

In order to calculate the slope value, all the necessary quantities are given in Fig. 5.25b. The substitution of the relevant quantities from Fig. 5.25b into Eq. (5.3) yields the slope value as $b = 2 \times (40.696 - 29.667)/1000 = 0.021$, which is within $\pm 5\%$ error limits from the slope value in Fig. 5.18a.

On the other hand, the time series time, t, and, y, variable averages are $\bar{t} = 5.4.3$ yields $a = 35.182 - 0.021 \times 500 = 24.682$. The relative difference between this value and the corresponding intercept in Fig. 5.18a is $100 \times (24.682 - 24.153)/24.682 = 2.14\%$, <10% and, therefore, these results remain within the sampling error limits.

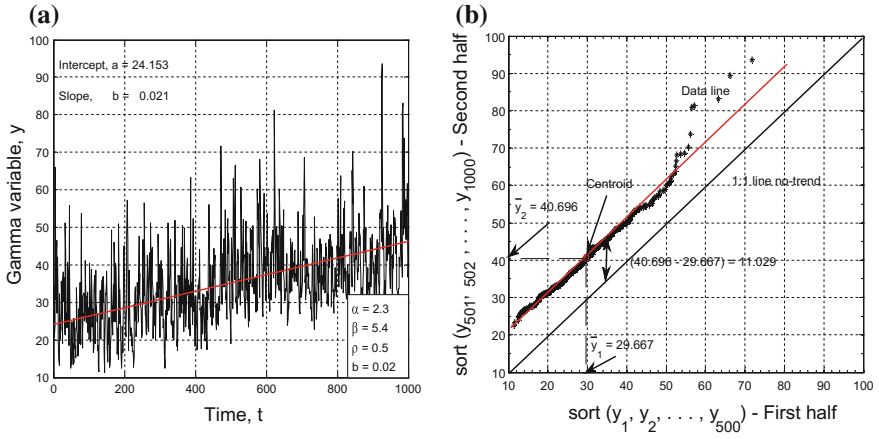


Fig. 5.25 Gamma stochastic process, **a** time series and trend, **b** innovative trend plot

5.5.3 Statistical Innovative Trend Test

The trend analysis as presented in this chapter is based on the comparison of two-half sample means. A test is convenient for the construction of confidence intervals by taking into consideration the difference between two population means. For this purpose, the null hypothesis, H_0, implies that there is not a significant trend if the calculated slope value, b, remains below a critical value, b_{cr}. Otherwise, an alternative hypothesis, H_a, is valid when $b > b_{cr}$. In order to develop an innovative significance test, it is necessary to derive the PDF of null hypothesis case. It is not necessary to search for the significance test of the intercept parameter, because the trend line is supposed to pass through the arithmetic averages of the independent and dependent variables. As for the slope parameter Eq. (5.3) shows that the stochastic property of b is a function of the first and second half time series arithmetic average values. Since $\bar{y}_1$ and $\bar{y}_2$ are also stochastic variables the first order moment (expectation) of the slope value can be obtained by taking the expectation of both sides leading to

$$E(b) = \frac{2}{n}[E(\bar{y}_2) - E(\bar{y}_1)]. \tag{5.6}$$

After all what have been explained in the previous sections in the case of no trend, the centroid point falls on the 1:1 line, which implies that $E(\bar{y}_1) = E(\bar{y}_2)$ and, therefore, $E(b) = 0$.

On the other hand, the variance of the slope can be calculated as $\sigma_b^2 = E(b^2) - E^2(b)$ or $\sigma_b^2 = E(b^2)$, which is equal to the second order moment of the

slope variable. This can be obtained by taking the expectation of both sides in Eq. (5.3) after the square operator resulting in

$$\sigma_b^2 = \frac{4}{n^2}\left[E\left(\overline{y}_2^2\right) - 2E(\overline{y}_2\overline{y}_1) + E\left(\overline{y}_1^2\right)\right].$$

Because $E\left(\overline{y}_2^2\right) = E\left(\overline{y}_1^2\right)$, it is possible to obtain the following expression:

$$\sigma_b^2 = \frac{8}{n^2}\left[E\left(\overline{y}_2^2\right) - E(\overline{y}_2\overline{y}_1)\right]. \tag{5.7}$$

The correlation coefficient between the two mean values is given in the stochastic processes as follows:

$$\rho_{\overline{y}_2\overline{y}_1} = \frac{E(\overline{y}_2\overline{y}_1) - E(\overline{y}_2)E(\overline{y}_1)}{\sigma_{\overline{y}_2}\sigma_{\overline{y}_1}}. \tag{5.8}$$

Substitution of the numerator of this expression into Eq. (5.7) and consideration stochastically that $\sigma_{\overline{y}_2} = \sigma_{\overline{y}_1} = \sigma/\sqrt{n}$ and, hence, Eq. (5.8) takes its final form as follows:

$$\sigma_b^2 = \frac{8}{n^2}\frac{\sigma^2}{n}(1 - \rho_{\overline{y}_2\overline{y}_1}). \tag{5.9}$$

In this last expression, $\rho_{\overline{y}_2\overline{y}_1}$ implies cross-correlation coefficient between the ascendingly sorted two-halves' arithmetic averages. The standard deviation of the sampling slope value can be obtained from Eq. (5.9) as

$$\sigma_b = \frac{2\sqrt{2}}{n\sqrt{n}}\sigma\sqrt{1 - \rho_{\overline{y}_1\overline{y}_2}}. \tag{5.10}$$

Furthermore, the third order moment of the slope variable is also equal to zero and the same is valid for all the odd order moments. This is the reason why the PDF of the slope, s, abides with the normal (Gaussian) PDF with zero mean and the standard deviation given in Eq. (5.10).

The most significant point in the application of this formulation is that the cross-correlation is between the two-sorted half time series. The statistical significance of the innovative trend slope test can be achieved through a normal (Gaussian) PDF with zero mean and standard deviation equal to Eq. (5.10).

5.5.4 Application

The stochastic innovative trend analysis as explained in the previous sections is applied to time series from different parts of the world. As for the long records

Table 5.2 Descriptive features of actual data

Name	Country	Record duration (year), n	Statistical features	
			Mean, $\overline{y}$	St. Dev., $\sigma_{\overline{y}}$
New Jersey	USA	116	53.04 °F	1.30 °F
Danube River	Romania	164	5566.8 m^3/s	944.91 m^3/s
Tigris River	Turkey	49	483.92 mm	124.74 mm

Southeastern New Jersey annual mean temperature data, Danube River annual discharge data, and annual mean precipitation data from the Tigris River drainage basin at Diyarbakir meteorology station are considered for the application to actual data. The simple statistical quantities of each station are presented in Table 5.2.

In general, most researchers look for the monotonic trend possibility within a given hydro-meteorological time series along the whole record length. The time series with trend and the innovative trend plots are given in Figs. 5.26, 5.27 and 5.28 for each data set.

In the application of innovative trend test, the basic criterion is the normal (Gaussian) PDF with zero mean and standard deviation σ_b (Eq. 5.10). If at α percent significance level the confidence limits of a standard normal PDF with zero mean and standard deviation is b_{cri} then the confidence limits (CL) of the trend slope can be expressed according to the following expression:

$$\text{CL}_{(1-\alpha)} = 0 \pm b_{\text{cri}}\,\sigma_b, \tag{5.11}$$

where σ_b is the slope standard deviation. All the necessary calculations and additional information with the operations in the last column are presented in Table 5.3.

One of the important points in this table is high cross-correlation values in row 6, because they are calculated depending on the ordered sequence in each half series. Slope value, b, of each station falls outside the lower and upper confidence limits and, therefore, in row 11 the alternative hypotheses, H_a, are adopted and they indicate the existence of trends (YES) as decisions. In the last row, the type of trend is stated depending on the slope sign in row 3.

The trend identification is one of the most significant elements in any climate change study. The most commonly used methodology for the identification is the Mann–Kendal (MK) trend test, but it requires few basic assumptions, which may not be valid in natural hydro-meteorological time series. MK test is misleading in the presence of data autocorrelation. Although several researchers have suggested pre-whitening procedure to render the original time series into a serially independent structure, but it is noticed that such a procedure cannot yield really embedded trend in the time series but with some bias. In the classical trend calculations serial independence, homoscedasticity and normal probability distribution assumptions must be satisfied. Such assumptions maybe guaranteed to a certain extent after

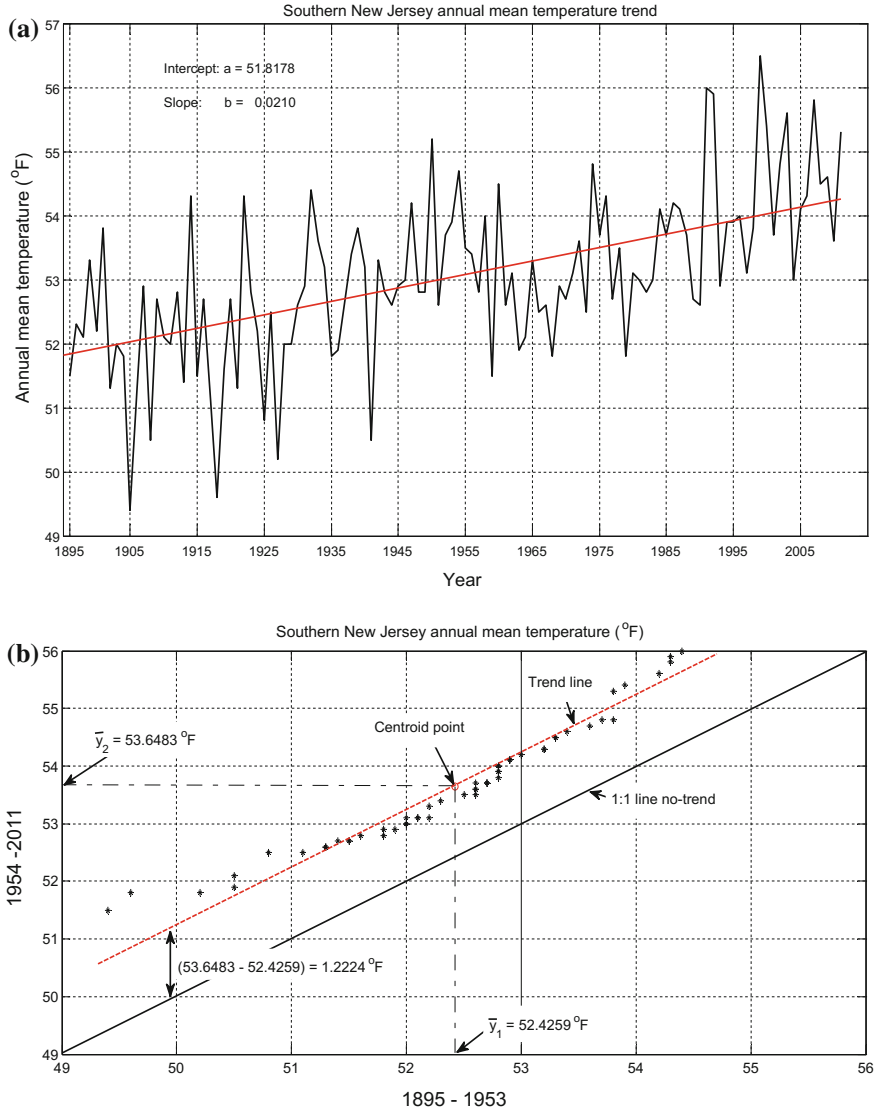


Fig. 5.26 Southern New Jersey annual mean temperature time series. **a** Time series and trend, **b** innovative trend plot

convenient transformations of the original series, which may not reflect genuine trend behavior of the series.

The procedure presented in this section does not require assumption, and it is based on the comparison of the two ascendingly ordered halves from the original

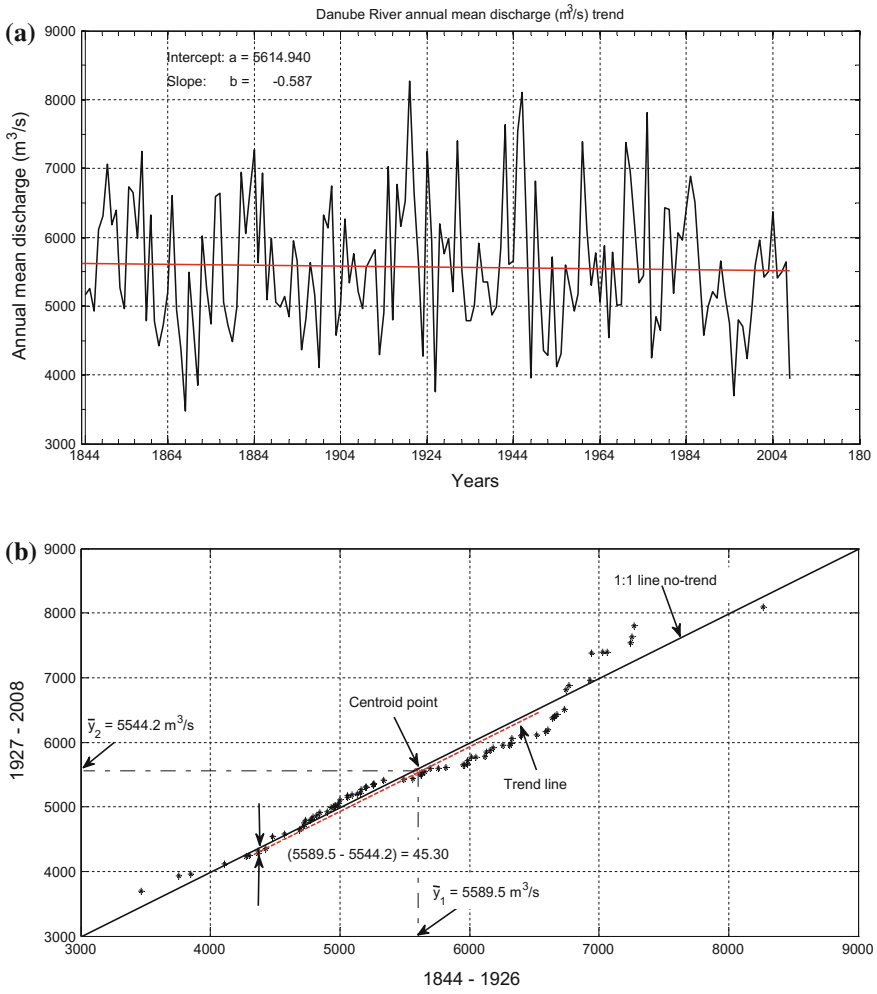


Fig. 5.27 Danube River annual mean discharge time series. **a** Time series and trend, **b** innovative trend plot

time series. The necessary formulations for the trend identification are derived explicitly and then monotonic trend significance test is presented in detail. The applications of the innovative trend significance statistical test are presented for the New Jersey temperature, Danube River discharge, and Tigris River meteorology

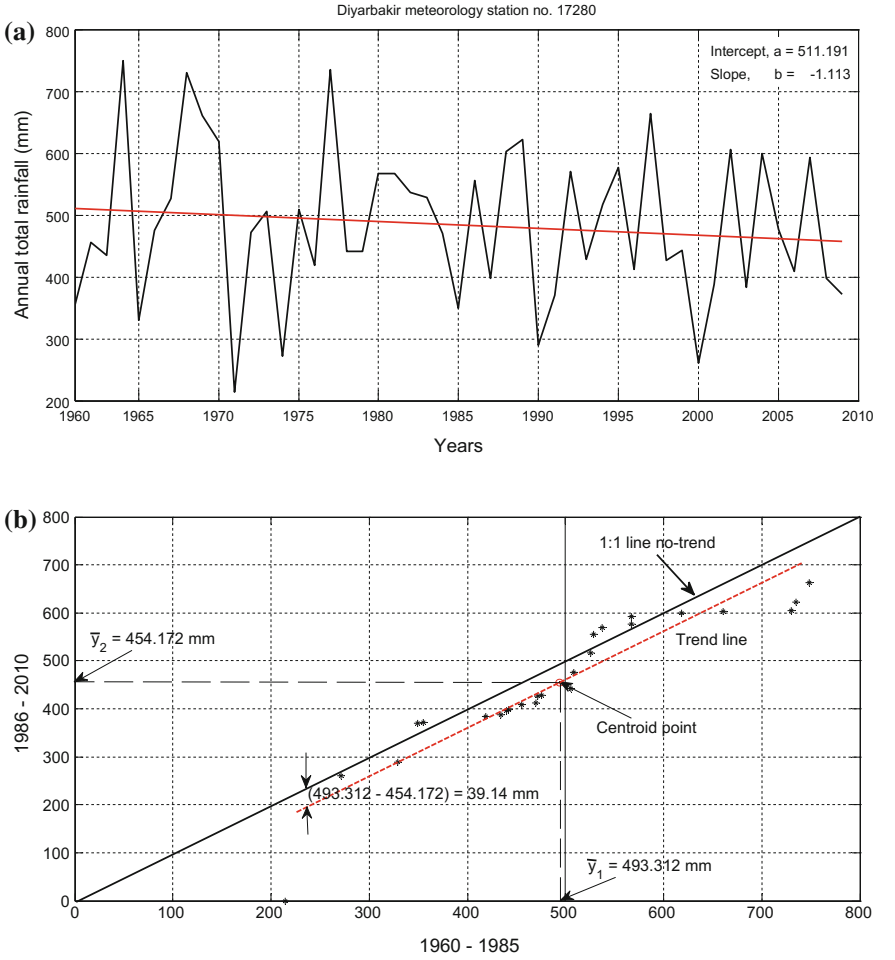


Fig. 5.28 Diyarbakir meteorology station annual total rainfall data. **a** Time series and trend, **b** innovative trend plot

station rainfall records at Diyarbakir meteorology station in the Southeastern part of Turkey. The suggested methodology is easy to apply and all the steps are logically presented in a rational manner.

Table 5.3 Innovative trend test results

No.	Name of stations	New Jersey	Danube River	Diyarbakir	Operations
1	Type of data	Annual temperature (°F)	Annual discharge (m^3/s)	Annual total rainfall (mm)	
2	Number of data	116	164	49	
3	Slope, b	0.021	−0.587	−1.113	Equation (5.3)
4	Intercept, a	51.818	5614.94	511.191	Equation (5.6)
5	Standard dev., σ	1.3025	944.921	124.738	From the whole series, y
6	Correlation, $\rho_{\bar{y}_1\bar{y}_2}$	0.9749	0.9767	0.9495	Ordered half series cross-correlations
7	Slope standard dev., σ_b	0.000467	0.1942	0.2386	Equation (5.10)
8	Significance level	0.05	0.05	0.05	Practically adopted
9	Lower CL	−0.000768	−0.3194	−0.3925	Equation (5.11)
10	Upper CL	+0.000768	+0.3194	+0.3925	Equation (5.11)
11	Hypothesis	H_a	H_a	H_a	Alternative hypothesis
12	Decision	Yes	Yes	Yes	According to H_a
13	Type of trend	Increasing	Decreasing	Decreasing	According to sign of b

5.6 Crossing Trend Analysis Methodology

Trend analyses are the necessary tools for depicting possible general increase or decrease in a given hydro-climatologic time series. There are many versions of trend identification methodologies such as the M–K trend test, S-R, Sen's slope, regression line, and Şen's innovative trend analysis. The literature has many papers about the use, cons and pros, and comparisons of these methodologies. In this section, a completely new approach is proposed based on the crossing properties of a time series. It is suggested that the suitable trend from the centroid of the given time series should have the maximum number of crossings (total number of up-crossings or down-crossings). This approach is applicable whether the time series has dependent or independent structure and also without any dependence on the type of the probability distribution function. The validity of this method is presented through extensive Monte Carlo simulation technique and its comparison with other existing trend identification methodologies.

Trend identification is one of the major topics in data processing concerning social, medical, industrial, scientific, and engineering studies for betterment of future predictions. Their physical causes maybe due to the changes in the natural events such as the climate change or depreciation and improvement in the human made instruments. Especially, in water sciences increasing and decreasing trend

tendencies bring into consideration the assessments of droughts, water scarcities, desertifications or floods and flash floods with water inundations. These water related events are also reflective in the agricultural and food production sectors. Climate change due to global warming has huge impact on the environment, weather patterns, and rise in sea level, which can be depicted by temporal trend analysis.

Although the visual appreciation of trend component in a given hydro-climatologic time series has been possible since the start of meteorological records in the second part of the eighteenth century, development of analytical methodologies came into existence in the first part of the nineteenth century (Mann 1945). His method provides information whether there is a trend within the time series with its verbal direction as increasing, decreasing, or neutral type. Later, Sen (1968) provided a quantitative slope calculation method for the trend component within a given time series. The M–K nonparametric trend test (Mann 1945), is functionally identical to Kendall's (tau) test for correlation (Kendall 1975), and the associated slope estimation by Sen (1968) median procedure.

On the other hand, Spearman's rho, which is a distribution-free statistic, is useful for the trend significance test (Spearman 1904). It is less widespread than the commonly applicable M–K trend test. However, the two tests are equivalent for the case of serially independent observations. Daniel (1990) has provided further explanations and improvements in the application of the Spearman's tau approach.

The regression monotonic line is among the parametric procedures for trend testing. The two sample t-test can be applied for step type of trends (Iman and Conover 1983). In these procedures, trend magnitude estimations are the regression slope and the difference in the means. On the other hand, nonparametric methods are the Mann–Kendall test and the Rank-Sum test (Bradley 1968), and their trend estimations are obtained according to Sen (1968), which is equivalent to the median of all pairwise slopes in the data set. Additionally, the Hodges–Lehmann estimator is the median of all differences between data in the first data set and data in the second data set (Hodges and Lehmann 1963).

Nonparametric procedures have significantly higher power than parametric procedures in cases of substantial departures from normal (Gaussian) probability distribution function (PDF) and the large sample sizes (Helsel and Hirsch 1988).

In addition to all available trend methodologies, a new one is suggested in this chapter as the "crossing trend," which depends on the maximum number of crossings (up-crossings or down-crossing) within a given hydro-climatological time series. This method hypothesis a set of different slope trends and the one with the maximum-crossing point is identified as the valid one.

The main purpose of this chapter is to suggest an innovative crossing trend analysis methodology with its significance test. The validity of this method is confirmed by extensive Monte Carlo simulation technique by taking into consideration different sample sizes and probability distribution functions (PDFs). The results are compared with the Sen's slope method and it is found that the differences are within the practically acceptable relative error percentage of $\pm 10\%$. The application of the innovative crossing trend analysis is performed for actual

meteorological records of annual daily extreme (maximum) rainfall from seven different climatological regions of Turkey.

5.6.1 Rational Concept

The main idea is that at various trend slope truncation levels that passes through the time series centroid, the number of crossing (up-crossings or down-crossings) is the maximum (Şen 2017). In order to illustrate this point, a hypothetical time series and its truncation—at different trend levels are given with the number of crossing points in Fig. 5.29.

In this figure, a series of increasing and decreasing trends are given and among them the one with the maximum crossing (up-crossing or down-crossing) number is the most representative trend-line. In this manner, the trend identification does not depend on the PDF of the hydro-climatologic variable. Besides, one can also calculate the surplus and deficit quantities on the basis of the trend line, if necessary. In Fig. 5.30, various quantities along the truncation level are shown. In this figure, SL (DL) implies surplus (deficit) lengths and there are 5(4) of them.

5.6.2 Theoretical Background

In any hydro-climatologic record series-crossing points at a truncation level provide not only information on wet and dry spell features, but also about the internal structure of the series (Şen 1977). For instance, the more is the crossing points at the median truncation level, the less is the serial dependence. In an independent series at the median level practically the number of up-crossings is equal to the

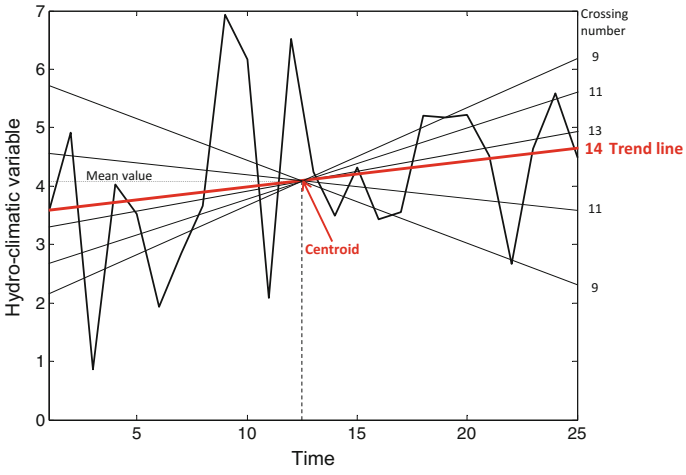


Fig. 5.29 Time series and trend crossing numbers

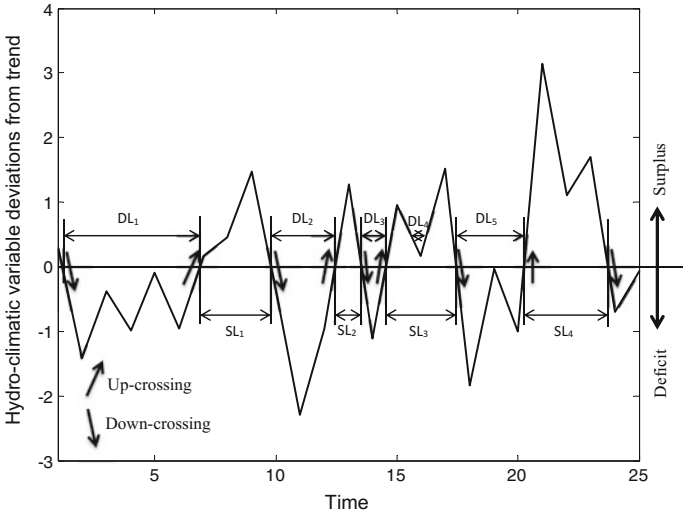


Fig. 5.30 Truncation (trend) level features

down-crossing number. In Fig. 5.30, up-crossings and down-crossings are indicated with arrows. Theoretically, in an infinite independent series, irrespective of the PDF, the number of crossings abide by the Poisson process (Feller 1968). However, in finite sample lengths, n, the expectation and the variance of the number of up-crossings, N_u, have been derived by Şen (1991) as

$$E(N_u) = np(1-p) \tag{5.12}$$

and

$$V(N_u) = E(N_u)\left(1 - 3p + 3p^2\right), \tag{5.13}$$

respectively. Herein, p is the probability of surplus numbers over the median truncation level. The average number of up-crossings increases with the sample length, n, but decreases as the truncation level increases. The maximum up-crossing (down-crossing) number occurs at 0.5 truncation level (Fig. 5.31). In general, such a truncation level corresponds to average = mode = median value in symmetrical PDFs, but to the median value in unsymmetrical PDFs (Fig. 5.31).

The PDF of up-crossings is shown to be in accord with the normal (Gaussian) PDF with mean and variance as in Eqs. (5.12) and (5.13), respectively. The standard deviation of the up-crossing number is shown in Fig. 5.32.

Under the light of the aforementioned information, it is possible to benefit from a normal (Gaussian) PDF for the significance test of innovative crossing trend either by the use of Eqs. (5.12) and (5.13) or with their standardization as

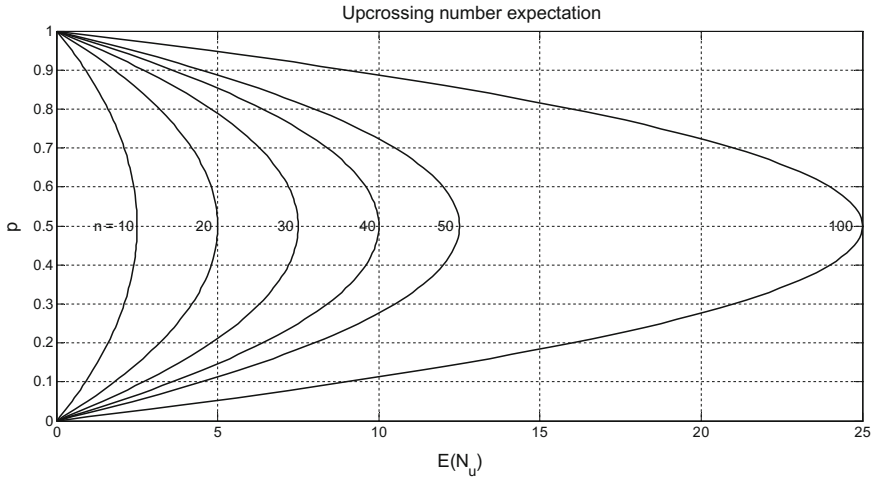


Fig. 5.31 Up-crossing number expectations for a set of sample lengths

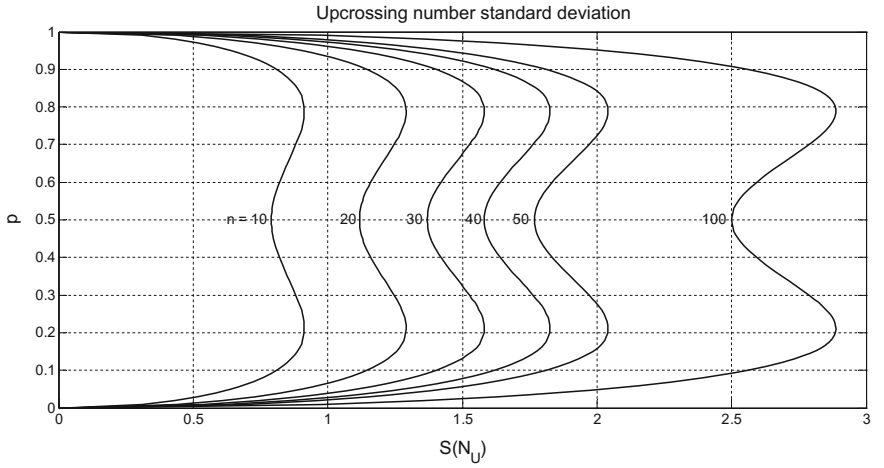


Fig. 5.32 Up-crossing number standard deviation for a set of sample lengths

$$m_s = \frac{E(N_u)}{n(n-1)} \tag{5.14}$$

and

$$s_s = \frac{V(N_u)}{E(N_u)(1-3p-3p^2)} \tag{5.15}$$

respectively.

5.6.3 Monte Carlo Simulations

In order to fix the validity of the crossing trend analysis, a set of Monte Carlo simulation studies is achieved, where 1,000 synthetic series are generated according to normal (Gaussian), Gamma and exponential PDF's. Each synthetic series is subjected to suggested innovative crossing trend analysis, the Mann–Kendall trend test, Sen's slope and Şen (2012, 2014) innovative trend slope and corresponding trend lines. In the simulations, the set of embedded slopes, s_d, are considered as decreasing (increasing) trends −0.007, −0.005, −0.003, −0.001 (0.1, 0.3, 0.5 and 0.7) with sample sizes as 25, 50, and 100. The simulation results are given as a set of graphs in Fig. 5.33. In this figure for each PDF three graphs are shown for the sake of visual inspection each for sample sizes 25, 50, and 100.

The numerical results of extensive simulation study are presented in Table 5.4. The simulations are carried on for three PDFs, namely standard normal (Gaussian) PDF with zero mean and unit standard deviation; Gamma PDF with location and scale parameters as 2 and 1, respectively; finally, the exponential PDF with its single parameter as 2.

Both Gamma and exponential PDF generations can be achieved with different parameter sets, but for the sake of brief description in this chapter only the aforementioned parameters are considered.

In this table, n indicates the sample length and R.E. is defined as the absolute relative error

$$\text{R.E.} = 100\frac{|s_e - s_c|}{s_e}, \tag{5.16}$$

where s_e and s_s are embedded and innovative crossing trend simulation slopes, respectively. In the first column of Table 5.1 are the embedded slope values, and simulation trend slopes are shown in the second, fourth, and the sixth columns under each sample length. It is obvious from this table that the absolute relative errors are less than practically acceptable 10% level, and the mean R.E. values are far less that this acceptable percentage level.

After all what have been explained so far as the simulation results are concern, it is evident that the innovative crossing trend analysis is valid for practical applications.

5.6.4 Application

For the application of the innovative crossing trend analysis, seven annual daily extreme rainfall records are considered from seven different climatology regions of Turkey. Each one has more than 50 years of records and this is a statistically valid sample size for reliable studies. The meteorology station locations are given in Fig. 5.34.

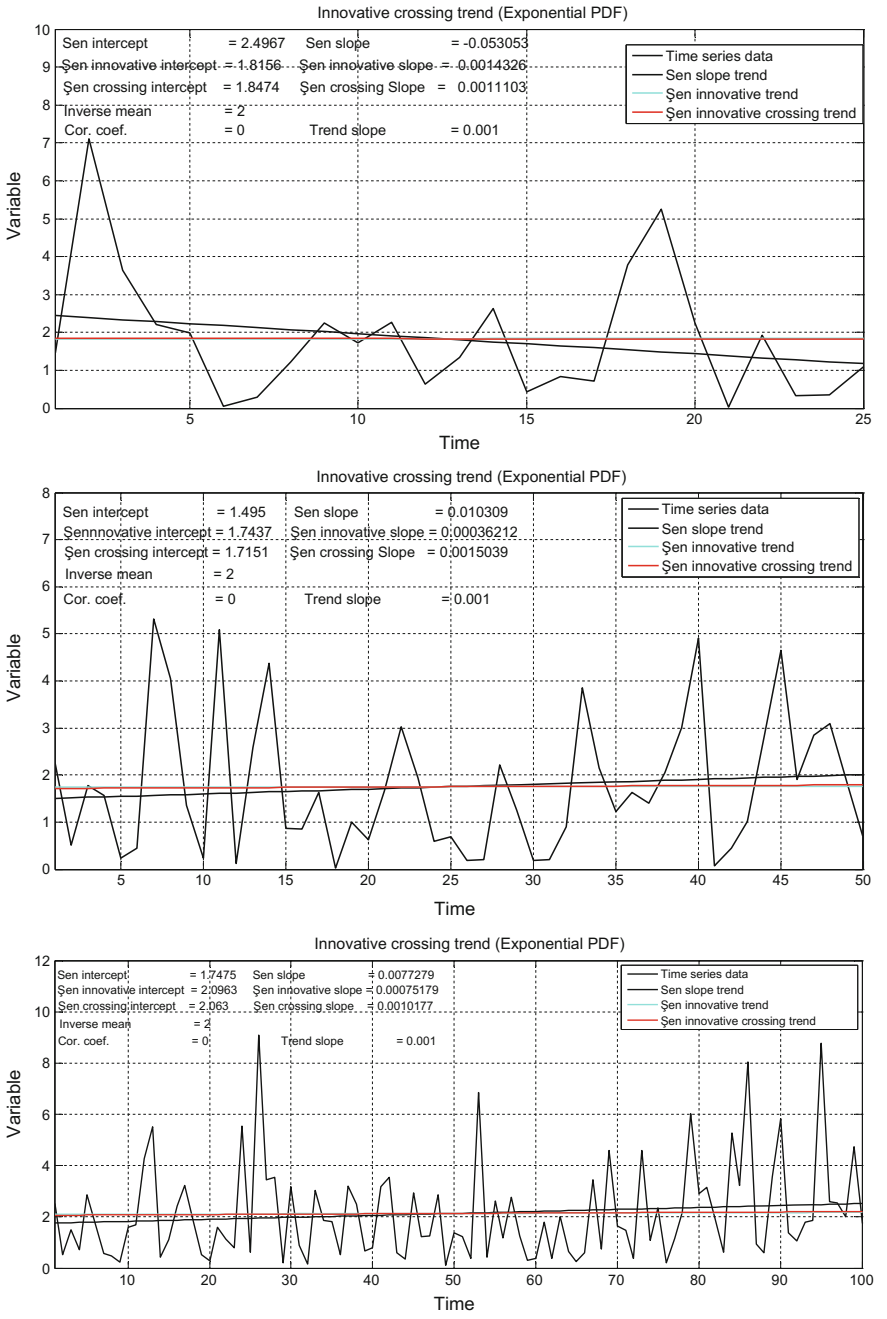


Fig. 5.33 PDF simulation results

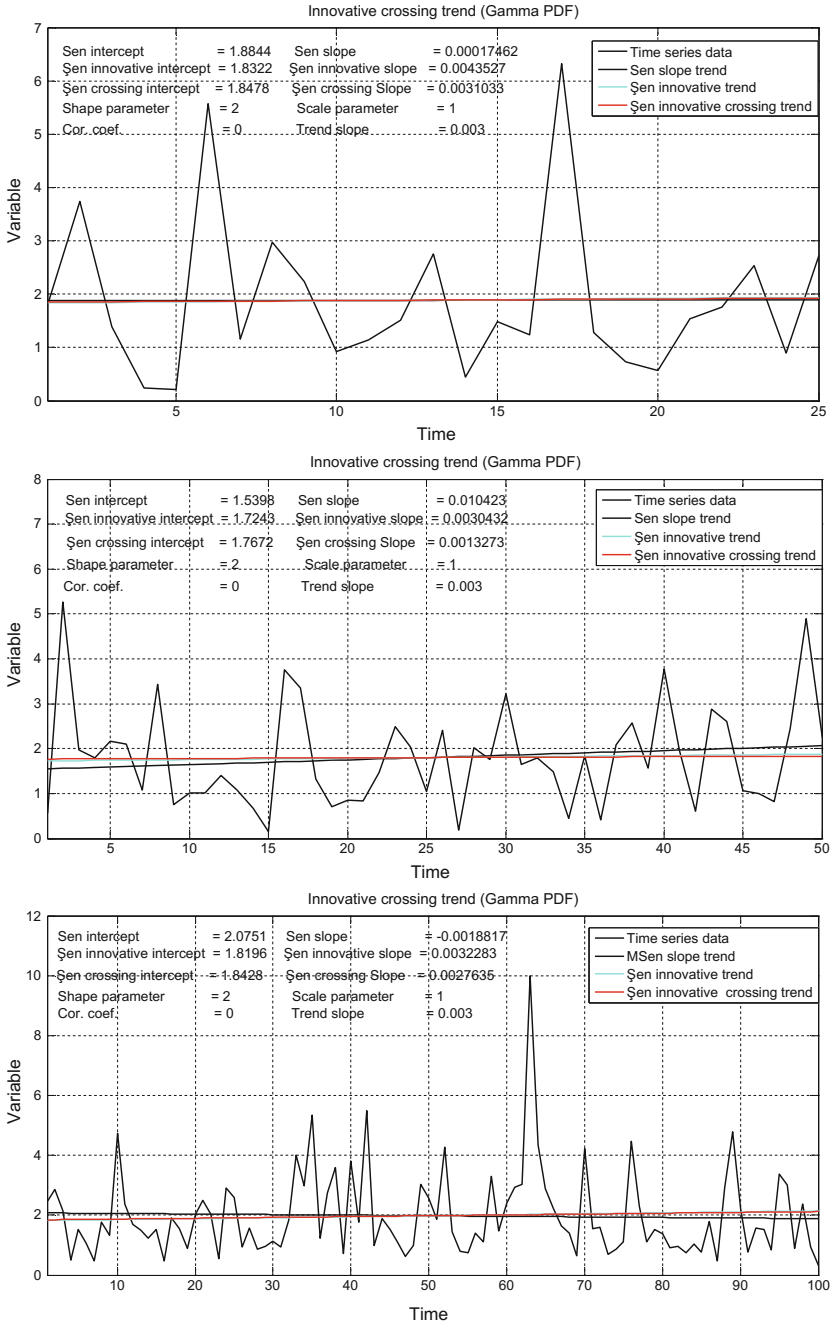


Fig. 5.33 (continued)

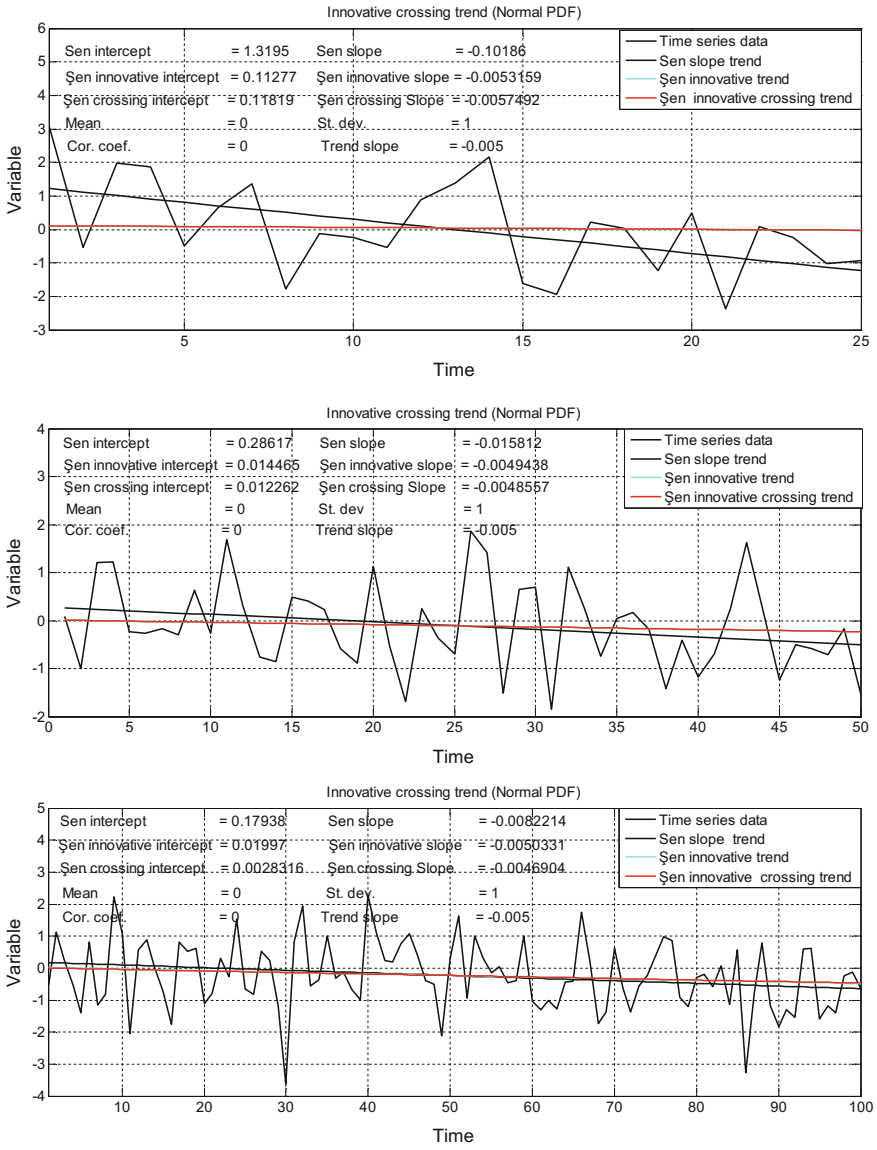


Fig. 5.33 (continued)

Each station represents different climatological region within Turkey. For instance, Ankara station is located in a dry, rather arid, and steppic region in the Central Anatolia, which is far away from the maritime climatic effects. This area includes the least rainfall receiving region of Turkey with annual average rainfall amounts less than 250 mm. Antalya is located along the Mediterranean coastal area of Turkey with typical Mediterranean climate impacts. Toward the northern part of

Table 5.4 Simulation results with relative error percentages

Gaussian PDF						
Embedded slope, s_e	n = 100	R.E. (%)	n = 50	R.E. (%)	n = 25	R.E. (%)
0.0010	0.0011	9.1000	0.0011	9.0909	0.0011	9.7473
0.0030	0.0033	9.7667	0.0033	9.6386	0.0027	9.4092
0.0050	0.0054	8.5400	0.0051	1.5748	0.0057	12.0338
0.0070	0.0072	2.3000	0.0066	5.3107	0.0078	10.6345
−0.0070	−0.0076	8.8857	−0.0075	6.1788	−0.0073	3.5015
−0.0050	−0.0047	6.2000	−0.0049	2.9442	−0.0057	11.4888
−0.0030	−0.0032	6.8000	−0.0032	6.9479	−0.0033	9.0909
−0.0010	−0.0011	10.3000	−0.0011	9.0082	−0.0010	2.2495
Mean	–	7.7365	–	6.3368	–	8.5194
Gamma PDF						
0.0010	0.0009	5.4852	0.0010	3.1946	0.0011	5.6604
0.0030	0.0028	8.5384	0.0030	1.4131	0.0031	3.3194
0.0050	0.0050	0.9285	0.0054	7.4417	0.0045	10.6440
0.0070	0.0068	2.3691	0.0073	3.7801	0.0067	4.0583
−0.0070	−0.0073	3.8065	−0.0078	10.4859	−0.0079	10.9641
−0.0050	−0.0051	2.2101	−0.0050	0.1201	−0.0049	1.9992
−0.0030	−0.0031	4.0307	−0.0032	5.2133	−0.0032	7.0344
−0.0010	−0.0009	6.6098	−0.0011	9.5841	−0.0011	9.0909
Mean	–	4.2473	–	5.1541	–	6.5963
Exponential PDF						
0.0010	0.0010	1.7682	0.0011	9.4203	0.0009	9.2896
0.0030	0.0031	1.9287	0.0033	10.1527	0.0033	9.1460
0.0050	0.0049	2.2077	0.0048	4.3841	0.0050	0.8723
0.0070	0.0077	9.5490	0.0074	4.9946	0.0078	10.1873
−0.0070	−0.0071	1.9471	−0.0076	8.3170	−0.0077	9.2912
−0.0050	−0.0054	6.5246	−0.0048	4.6025	−0.0047	6.8148
−0.0030	−0.0031	3.8770	−0.0028	6.8376	−0.0033	8.2849
−0.0010	−0.0011	7.1495	−0.0011	9.9099	−0.0011	9.4203
Mean	–	4.3690	–	7.3273	–	7.9133

this region are the Taurus Mountain chain with elevations more than 3,000 m above mean sea level, and therefore, it is one of the humid regions in Turkey. Frequent orographic rainfall types occur, which causes to occasional floods. The geological composition is of limestone and dolomatic rocks with karstic features, and therefore, rainfalls recharge the groundwater in the region. It is regarded as one of the surface and ground water rich parts of Turkey. Diyarbakir station is in the

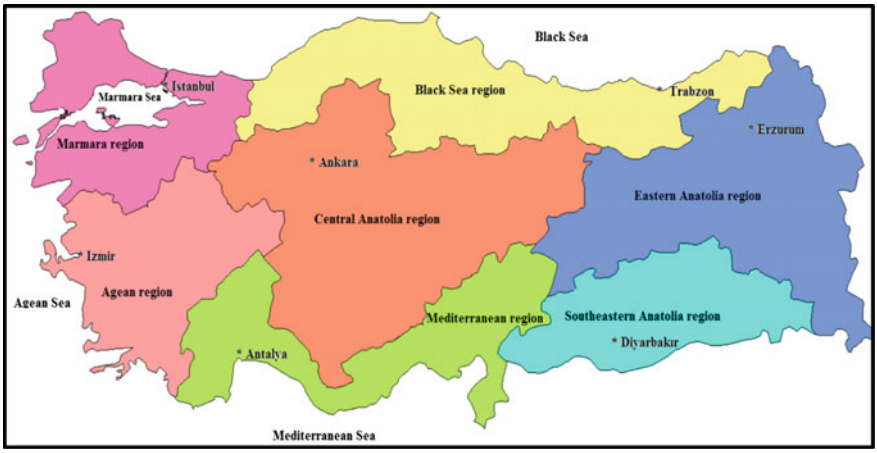


Fig. 5.34 Meteorology station locations

southeastern province of Turkey away from the sea born air mass movements, and therefore, it has continental climatic features. Due to its unique position at the upper end of the Mesopotamian valley, the rainfall occurrences are rare and mostly summer seasons are extremely hot and winter seasons are mild. Erzurum station represents rugged mountainous region of eastern Turkey with severe winter conditions and rather cool summer months. Izmir location is the representative of the Aegean Sea at the western coastal area of Turkey. It has hot summer months and mild winter season with moderate rainfall events throughout the year. Finally, Trabzon meteorology station is chosen for the representation of the Black Sea rainfall regime, which is rainy almost throughout the year. This is due to the fact that North Atlantic born air masses that descend southwesterly over the Europe and then over the Black Sea with moisture and the coastal parallel mountain chains cause to frequent orographic and cyclonic rainfall occurrences.

Seven meteorology station records are treated by the innovative crossing trend analysis and also classical Sen's slope regressions. Figure 5.35 indicates the innovative crossing trends in each record and also the test results are presented in Table 5.6 for each station by considering the Sen slope and the suggested methodology features as explained in Sects. 5.6.1 and 5.6.2.

Visual inspection of each graph provides reflections that the innovative crossing trend analysis well identifies the trend component in each location. For the sake of comparison trend calculated on the basis of Sen's slope is also given on the same graphs.

However, for quantitative analyses Table 5.6 is prepared, where both Mann–Kendall trend test and the innovative crossing trend analysis quantities are presented. In this table LL and UL are for the lower and upper significance levels. It is

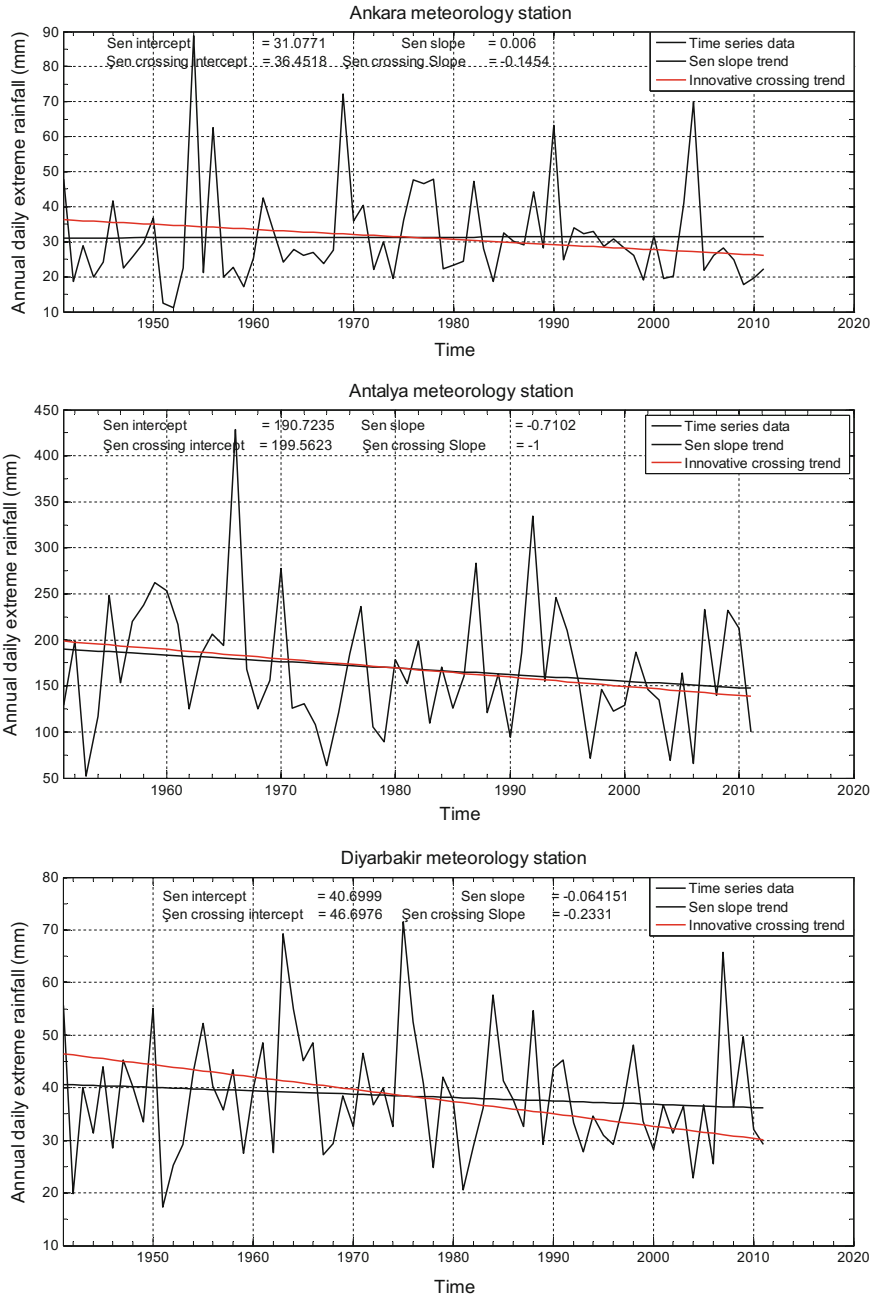


Fig. 5.35 Innovative crossing trend components

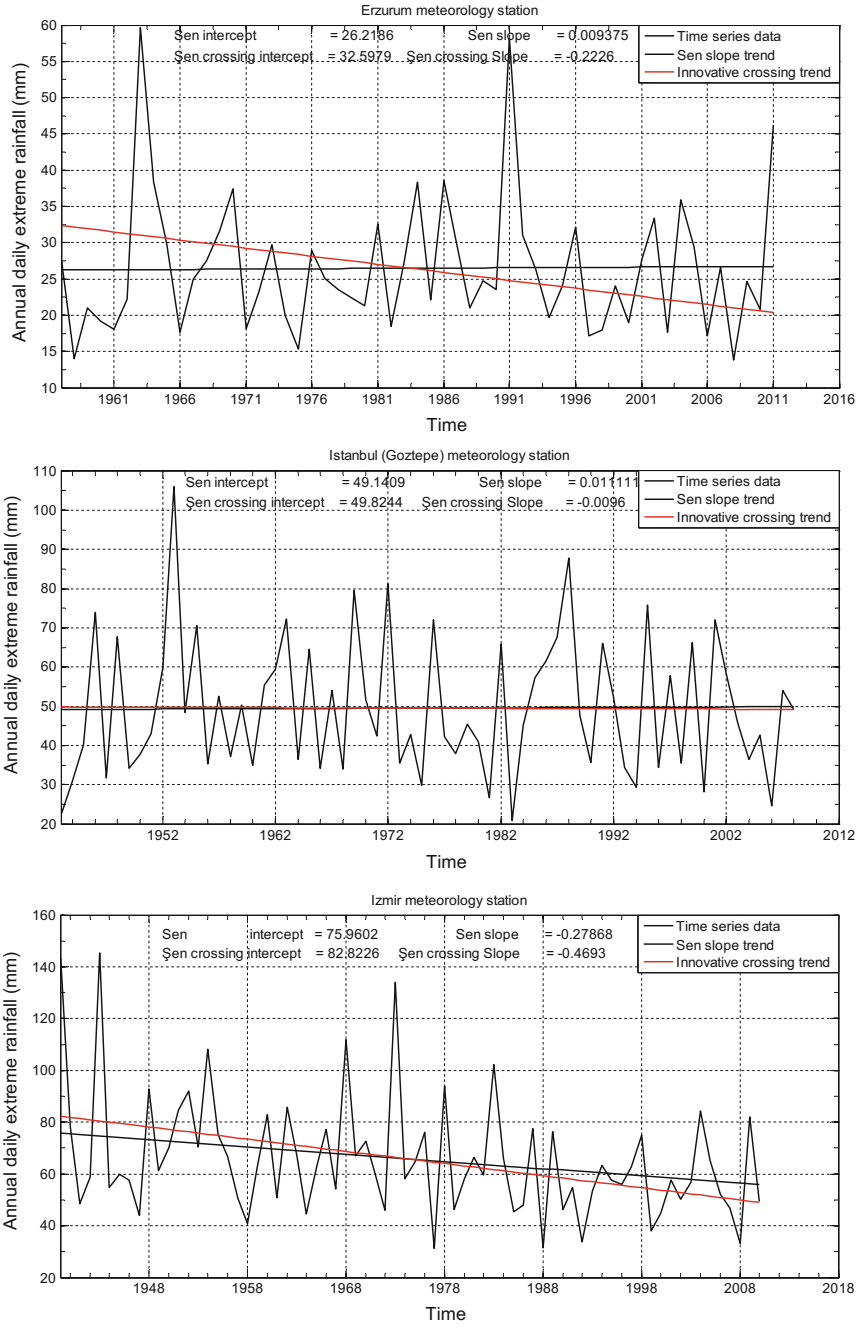


Fig. 5.35 (continued)

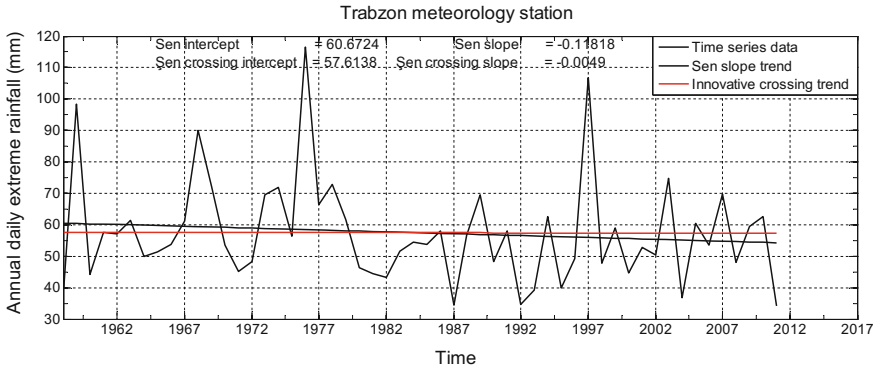


Fig. 5.35 (continued)

to be noticed that the confidence limits in the case of Mann–Kendall trend test remain the same without depending on the sample size. However, as obvious from Eqs. (5.14) and (5.15) the confidence limits are functions of the sample size for the innovative crossing trend analysis. Z is the statistics value of the Mann–Kendall trend test, and C is the number of up-crossing for the innovative crossing trend calculations. In the table, trend tests are probed for two levels, 90 and 95%.

Trend component identification in the climatological time series constitutes very important aspect, especially for the climate change description and, therefore, such studies have increased unprecedentedly since the last three decades. There are different methodologies for this purpose, but each one with restrictive assumptions. In this chapter, entirely new concept of trend identification is proposed by taking into consideration the number of crossings on the possible trend line. It is stated that the trend component should have the maximum number of crossings among many different trend alternatives. In order to select the most valid one, the given climatological time series is probed with a set of trend representatives that passes through the centroid point of the data. The centroid is defined as the point in the time series with abcissa as the half of the sample size and the ordinate equal to the median of the recorded values. The formulations are given at the median level as for the number and the variance of the crossing points with no trend within a serially independent time series. They do not dependent on the type of probability distribution function. The validity of innovative crossing approach is shown by extensive Monte Carlo simulation studies based on different sample sizes and probability distribution functions. The application of the innovative up-crossing trend analysis is presented for seven distinctive climatological regions of Turkey for annual daily extreme (maximum) rainfall records, which have physically independent serial structure so as to abide with the theoretical requirement of the suggested methodology.

Table 5.6 Mann−Kendal and innovative crossing trend statistical characteristics

95% confidence								
	Mann–Kendall			Decision	Innovative crossing			Decision
	LL	Z	UL		LL	C	UL	
Ankara	−1.6449	0.1191	1.6449	No significance	−21.2149	15.5	21.2149	No significance
Antalya	−1.6449	−1.2944	1.6449	No significance	−18.4617	17	18.4617	No significance
Diyarbakir	−1.6449	−0.9778	1.6449	No significance	−21.2149	21.5	21.2149	Significant trend
Erzurum	−1.6449	0.1307	1.6449	No significance	−16.7996	16	16.7996	No significant trend
Istanbul	−1.6449	0.1771	1.6449	No significance	−19.8407	21.5	19.8407	Significant trend
Izmir	−1.6449	−2.5959	1.6449	Significant trend	−21.4893	19.5	21.4893	No significant trend
Trabzon	−1.6449	−0.843	1.6449	No significance	−16.5218	16	15.5218	No significance
90% confidence								
Ankara	−1.2816	0.1191	1.2816	No significance	−20.4496	15.5	20.4496	No significance
Antalya	−1.2816	−1.2944	1.2816	Significant	−17.7523	17	17.7523	No significance
Diyarbakir	−1.2816	−0.9778	1.2816	No significance	−20.4496	21.5	20.4496	Significant
Erzurum	−1.2816	0.1307	1.2816	No significance	−16.1261	16	16.1261	No significance
Istanbul	−1.2816	0.1771	1.2816	No significance	−19.1028	21.5	19.1028	Significant
Izmir	−1.2816	−2.5959	1.2816	Significant	−20.7186	19.5	20.7186	No significance
Trabzon	−1.2816	−0.843	1.2816	No significance	−15.8544	16	15.8544	Significant

References

Barbosa, S. M., Silva, M. E., & Fernandes, M. J. (2008). Time series analysis of sea-level records: Characterizing long-term variability. In R. V. Donner & S. M. Barbosa (Eds.), *Nonlinear time series analysis in the geosciences—applications in climatology, geodynamics, and solar-terrestrial physics* (pp. 157–173). Berlin: Springer.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis*. Holden-Day, San Francisco, CA: Forecasting and Control.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, N.J: Prentice-Hall.

Brunetti, M., Colacino, M., Maugeri, M., & Nanni, T. (2001). Trends in the daily intensity of precipitation in Italy from 1951 to 1996. *International Journal of Climatology, 21,* 299–316.

Burn, D., Mohamed, A., & Elnur, H. (2002). Detection of hydrologic trends and variability. *Journal of Hydrology, 255,* 107–122.

Daniel, D. (1990). Summary review of construction quality control for earthen liners. In R. Bonaparte (Ed.), *Waste Containment Systems: Construction, Regulation, and Performance* (pp. 175–189), GSP No. 26, ASCE.

Douglas, E. M., Vogel, R. M., & Kroll, C. N. (2000). Trends in floods and low flows in the United States: Impact of spatial correlation. *Journal of Hydrology, 240,* 90–105.

Feller, W. (1968). *An introduction to probability theory and its applications* (Vol. I, 3rd ed.). Wiley.

Groisman, P. Y., Knight, R. W., & Karl, T. R. (2001). Heavy precipitation and high streamflow in the contiguous United States: Trends in the 20th century. *Bulletin of the American Meteorological Society, 82,* 219–246.

Groisman, P. Y., Knight, R. W., Karl, T. R., Easterling, D. R., Sun, B., & Lawrimore, J. H. (2004). Contemporary changes of the hydrological cycle over the contiguous United States: Trends derived from in situ observations. *Journal of Hydrometeorology, 5*(1), 64–85.

Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers, 77,* 1308.

Helsel, D. R., & Hirsch, R. M. (1988). Discussion of "applicability of the t-test for detecting trends in water quality variables" by R.H. Montgomery and J.C. Loftis. *Water Resources Bulletin, 24,* 201–204.

Hirsch, R. M., & Slack, J. R. (1984). A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research, 20*(1), 727–732.

Hodges, J. L., Jr., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics, 34,* 598–611.

Iman, R. L., & Conover, W. J. (1983). *A modern approach to statistics*. New York: Wiley.

Intergovernmental Panel on Climate Change (IPCC). (2007). *Climate change 2007: The physical science basis. contribution of working group i to the fourth assessment report of the intergovernmental panel on climate change*. New York: Cambridge University Press.

Kahya, E., & Kalaycı, S. (2004). Trend analysis of stream flow in Turkey. *Journal of Hydrology, 289,* 128–144.

Kalra, A., Piechota, T. C., Davies, R., & Tootle, G. A. (2008). Changes in U.S. streamflow and western U.S. snowpack. *Journal of Hydrologic Engineering, 13,* 156–163.

Kendall, M. G. (1975). *Rank correlation methods*. New York: Oxford University Press.

Lettenmaier, D. P., Anderson, D., & Brenner, R. (1984). Consolidation of a stream quality monitoring network.*Water Resources Bulletin, 20*(4), 473–481.

Lettenmaier, D. P., Wood, E. F., & Wallis, J. R. (1994). Hydroclimatological trends in the continental United States, 1948–88. *Journal of Climate, 7,* 586–607.

Lins, H. F., & Slack, J. R. (1999). Streamflow trends in the United States. *Geophysical Reseach Letters, 26*(2), 227–230.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13,* 245–259.

McCabe, G. J., & Wolock, D. M. (2002). A step increase in streamflow in the conterminous United States. *Geophysical Reseach Letters, 29*(24), 2185.

Miller, W. P., & Piechota, T. C. (2008). Regional analysis of trend and step changes observed in hydroclimatic variables around the Colorado River Basin. *Journal of Hydrometeorology, 9*(5), 1020–1034.

Montanari, M., Rosso, R., & Taqqu, M. S. (1997). Fractionally differenced ARIMA models.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's Tau. *Journal of American Statistical Association, 63,* 1379–1389.

Şen, Z. (1974). Small sample properties of stationary stochastic processes and the hurst phenomenon in hydrology. unpublished Ph. D. Thesis, Imperial College of Science and Technology, University of London, 256p.

Sen, P. K. (1978). Estimates of the regression coefficient based on Kendall's Tau. *Journal of American Statistical Association, 63,* 1379–1389.

Şen, Z. (1977). Autorun analysis of hydrologic time series. *Journal of Hydrology, 36,* 75–85.

Şen, Z. (1991). Probabilistic modelling of crossing in small samples and application of runs to hydrology. *Journal of Hydrology, 124*(3–4), 345–362.

Şen, Z. (2012). Innovative trend analysis methodology. *Journal of Hydrologic Engineering, 17*(9), 1042–1046.

Şen, Z. (2014). Trend identification simulation and application. *Journal of Hydrologic Engineering, 19,* 241–245.

Şen, Z. (2017). Innovative crossing trend analysis and application. *Theoretical and Applied Climatology* (in print).

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72–101.

Trenberth, K. E., et al. (2007). *Observations: Surface and atmospheric climate change, in climate change 2007: The physical science basis. contribution of working group i to the fourth assessment report of the intergovernmental panel on climate change, chap. 3*. New York: Cambridge University Press.

Wang, Y., & Zhou, L. (2005). Observed trends in extreme precipitation events in China during 1961–2001 and the associated changes in large-scale circulation. *Geophysical Reseach Letters, 32,* L09707. doi:10.1029/2006GL022574.

Xiong, L. H., and Guo, S. L. (2004). Trend test and change-point detection for the annual discharge series of the Yangtze River at the Yichang hydrological station. *Hydrological Sciences Journal, 49*(1), 99–112.

Yue, S., Pilon, P., & Cavadia, G. (2002a). Corrigendum to "Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series". *Journal of Hydrology, 259,* 254–271.

Yue, S., Pilon, P., & Cavadia, G. (2002b). Corrigendum to "Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series". *Journal of Hydrology, 259,* 254–271.

Yue, S., Pilon, P., Phinney, B., & Cavadias, G. (2002c). The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrological Processes, 16*, 1807–1829.

Zhang, X., Harvey, K. D., Hogg, W. D., & Yuzyk, T. R. (2001). Trends in Canadian streamflow. *Water Resources Research, 37,* 987–998.

Spatial Trend Analysis

6

Abstract

Spatial trend concept is very useful in order to depict the systematic variations of the phenomenon concerned over a region based on geographical locations or as in this book based on two independent variables that may be any other two event records. Different types of spatial trend alternatives are presented visually and then their mathematical solutions under the title of trend surface analysis is presented with derivation of the necessary spatial regression analysis approach. Although there are different mapping procedures in this chapter, the most advanced one, namely, Kriging geostatistically developed methodology is explained for the purpose of 3D surface construction. Based on this approach parallel and serial triple diagram models are explained for better interpretations amount three different time series or three time series generated from the same time series at two different lag times.

Keywords

Homogeniety · Isotropy · Kriging · Mass curve · Spatial · 3D · Trend surface · Triple

6.1 General

Any natural phenomenon or its similitude occurs extensively over a region, and therefore, its recordings or observations at different locations pose some questions as, for instance, are there relationships in the form of trends between phenomena in various locations? In such a question, the time is as if it is frozen and the phenomenon concerned is investigated over the area and its behavioral occurrence between the locations. Answer to this question may be provided descriptively in

Z. Şen, *Innovative Trend Methodologies in Science and Engineering*,
DOI 10.1007/978-3-319-52338-5_6

linguistic, subjective and vague terms, which may be understood even by non-specialists in the discipline. However, their quantification necessitates objective methodologies, which are one of the purposes of the context in this book. Another question that may be stated right at the beginning of the research in the earth, environment, and atmospheric sciences is that are places different in terms of the phenomena present there? Such questions are the source of many people's interest in the subject.

Three-dimensional statistical techniques help to obtain maps of variable concerned provided that the two geographic coordinates are given at the measurement points. This procedure is referred to as the trend surface analysis in the statistics literature. It is also referred to as the multivariate statistical analysis. Its basis is to match a surface similar to ordinary regression analysis but in three-dimensional space. The same restrictive assumptions as in the ordinary regression analysis are also valid in the trend surface fitting. There are further difficulties in the spatial trend surface search such as the paucity of spatial data and extensive computation requirements. For the success of trend surface fitting uniform data distribution is necessary.

The significance of trend surface analysis is to separate the spatial behavior of the phenomenon into two components as the deterministic component in terms of trend surface and the residuals, which are deviations of the measurements from the fitted trend surface and they are the uncertain (random, stochastic) component. The uncertain components are representatives of local sites, whereas the trend surface is the regional behavior of the phenomenon concerned.

Trend analysis separates the ReV into two complementary components, namely regional nature of deterministic variations and local fluctuations around the regional component. The regional and local components are dependent on the scale of the ReV. In any trend analysis there are three variables. In any spatial analysis there are three general components for the application of a convenient methodology.

(1) The basis of any spatial analysis is two basic deterministic variables such as easting and northing or longitude and latitude variables that provide locations of measurement variable at a set of locations,
(2) Decomposition of the spatial variable first to a general regional deterministic part, which can be expressed by any mathematical function,
(3) The stochastic (uncertain) part, which constitutes a set of deviations of the measurements from the corresponding trend surface value.

In general spatially variable event may include gradual monotonic trends or even abrupt changes (jumps) due to externally effective phenomenon. It is by now well understood that the global warming leading to climate change imprints an increasing trend into global temperature data. Abrupt changes may also take place as a result of sudden or short duration exogenous impacts, such as volcanic eruptions, earthquakes, sudden changes in monitory rates (devaluation) and alike.

The main purpose of spatial trend analysis is to decompose regional variable into subcomponents such as trends, abrupt changes, stochastic, and entirely independent error terms so that the construction of a suitable model by the synthesis of these components provides opportunity to predict the variable concerned at non-measurement sites.

Scientific treatment and interpretation of even error laden data lead to significant practical knowledge concerning the oceans and atmosphere. It is the prime duty of the scientist to filter out the meaningful portions of the data and to model randomly the error part.

It is possible to obtain regular grid points from irregular measurement sites by fitting a surface to available data, which can be achieved either globally or locally over the study area. In the former case there is a single functional form of the trend surface in addition to the stochastic nature of the residuals and the latter case is just the repetition of global procedure on pieces of subareas within the study area. Another version of the local surface fitting is to consider the neighbor points to reach to locally representative trends. However, the most widely used procedure is the global trend surface search, for this purpose the spatial variable is approached by a polynomial expansion of the geographic coordinates, and the coefficients of the polynomial function are estimated from available measurements by means of the least squares method, which relies on the sum of the squared deviations minimization from the trend surface. After the identification of the trend surface the sum of the trend surface value at a site and the residual is equal to the measurement value. The residuals are random variables either with independent structure, in which case the probability distribution (PDF) is the representative of the spatial randomness or in the case of dependent spatial residuals one of the most convenient stochastic processes or the regionalized variable approach through the Kriging methodology is the most representative approach (Matheron, 1969).

In general, the polynomial functions can be of any desired degree, the higher is the degree the more is the computation rounding error. Therefore, in practice the highest degree is adapted as 5 or 7. The unknown parameter coefficients of the polynomials can be estimated from the simultaneous solution of a set of convenient number of equations as explained in Sect. 6. Each one of these equations include the sums of powers and cross products of the geographic (X, Y), and spatial variable Z values. After the estimation of the coefficients the polynomial function provides opportunity to calculate spatial variable value at any site (point) within the study area. In general, in order to map the spatial variable one can calculate the spatial variable values at regular mesh nodes and subsequently obtain the 2D contour or 3D map for the spatial variable.

Fixation of a trend surface to given set of spatial variable measurements at irregular sites separates the whole measurements into two components, namely trend and residual values. Trend values are collection of deterministic quantities, but the residuals are uncertainty parts.

6.2 Numerical Solution

It is well established that the analytical solutions of many practically applicable differential equations is possible only through numerical analysis, where the variable concerned is searched for the suitable value at the nodes of regularly located mesh over the problem solution domain. This is tantamount to mathematical version of the spatial analysis, where the generation mechanism of the phenomenon is in the form of differential equations. For instance, in many engineering applications including space research most often a first-order partial differential expression is valid as follows.

$$\frac{\partial z(x,y)}{\partial x} + \frac{\partial z(x,y)}{\partial y} = 0 \tag{6.1}$$

This is the steady state continuity (mass balance) form of any spatial variable (quantity) independent of time domain. This expression represents in space infinitesimally small prism that is shown in Fig. 6.1.

Figure 6.2 is the corresponding model with finite difference definition of each infinitesimally small variation. In these figures the small trend surface is also shown. The location of this finite trend surface n apart from the location (x, y) coordinates the slopes along two coordinate directions in case of infinitesimally small model (Fig. 7.1) as $\partial z(x,y)/\partial x$ and $\partial z(x,y)/\partial y$. The corresponding slopes of the finite trend surface in Fig. 6.2 can be written correspondingly as $[z(x+\Delta x, y) - z(x,y)]/\Delta x$ and $[z(x, y+\Delta y) - z(x,y)]/\Delta x$, respectively.

One can understand from the aforementioned figures and discussions than in the case of any spatial variable consideration of finite element model the trend surface needs three point measurements. In practical application, for numerical solution the infinitesimally small and finite element model slopes are taken as equal to each other, which allows to write the following two equations as,

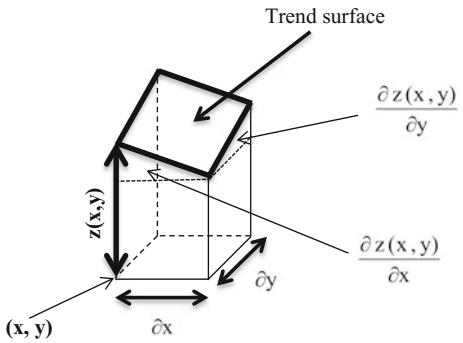


Fig. 6.1 Infinitesimally small models

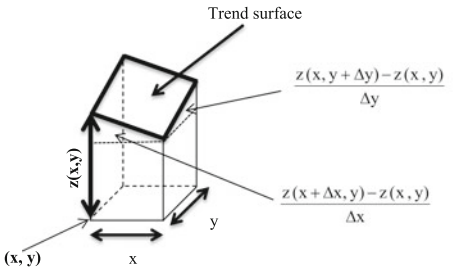


Fig. 6.2 Finite element small models

$$\left.\begin{array}{l} \frac{\partial z(x,y)}{\partial x} = \frac{z_{i,j} - z_{i-1,j}}{\Delta x} \\ \text{and} \\ \frac{\partial z(x,y)}{\partial y} = \frac{z_{i,j} - z_{i,j-1}}{\Delta y} \end{array}\right\} \tag{6.2}$$

The substitution of these expressions into Eq. (6.1) leads after the necessary calculation to the following explicit expression.

$$z_{i,j} = \frac{z_{i-1,j} + \alpha z_{i,j-1}}{1 + \alpha}, \tag{6.3}$$

where α is the distance ratio defined as,

$$\alpha = \frac{\Delta x}{\Delta y} \tag{6.4}$$

In case of completely uniform spatial variable variation base domain, i.e., $\Delta x = \Delta y$, ($\alpha = 1$), Eq. (6.3) takes its simplest form.

$$z_{i,j} = \frac{z_{i-1,j} + z_{i,j-1}}{2} \tag{6.5}$$

The numerical solution continues from node (i, j) to the next ones systematically, and hence, the whole spatial variation domain is scanned with new spatial variable values, $z_{i,j}$ ($i, j = 1, 2, \ldots, n$) with n being the number of nodes in one direction, which is shown in Fig. 6.3.

So far explained numerical solution of spatial variable through the differential expression rule for evolution of the spatial variable, it is noted that regular and uniform distributions of the coordinate variables are necessary.

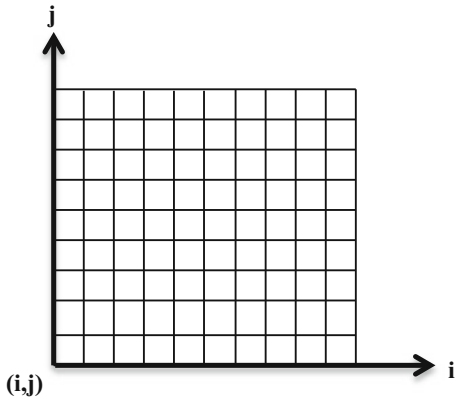


Fig. 6.3 Regular mesh and nodes

6.3 Spatial Data Analysis

Apart from the temporal tendencies there are also spatial trends over a region on the basis of easting (longitude) and northing (latitude), which provides information about regional variability of the phenomenon concerned. For this purpose, it is necessary to have measurements at a set of different locations. Even a single record at each measurement station is enough for spatial trend and variation appreciation. Again visual inspection and assessment of spatial data is recommended prior to the application of any detailed scientific procedure. The initial visual helps to set the foundations of a convenient methodology for the spatial evaluation of data. Prior to any quantitative evaluation of the spatial data at the hand the following points provide assistance.

(1) The sampling locations are characterized by coordinates, X and Y, preferably on a scaled map. The spatial variable can be shown by Z. In general, the data locations are irregularly distributed, but in any new study, if possible, data positions are selected better at the nodes of regular nets. The measurement locations may already been such as the existing well locations (water or oil), meteorology stations, urban areas, etc. In Table 6.1 there is a sample of spatial data with location coordinates (X, Y) and Z values.

Figure 6.4 indicates the spatial distribution of the measurement locations and their values, which reflects that the measurement sites are irregularly distributed within the study area.

The following points are the visual reflections from Fig. 6.4, but they are preliminary information for formal scientific methodologies.

(1) The majority of small measurements are clustered at high easting but low northing regions,

Table 6.1 Spatial data records

Data number	Location		Measurement, Z	Projection	
	X	Y		X	Y
1	40.78	30.42	31	20	3
2	40.52	30.3	100	1578	130
3	39.72	40.05	1631	120	58
4	37.00	35.33	20	1631	130
5	41.17	29.04	130	3	100
6	40.73	31.60	742	539	539
7	39.62	27.92	120	58	100
8	40.18	29.07	100	100	31
9	40.32	27.97	58	100	742
10	40.15	29.98	539	742	20
11	38.40	42.12	1578	31	1631
12	40.13	26.40	3	41.17	1578

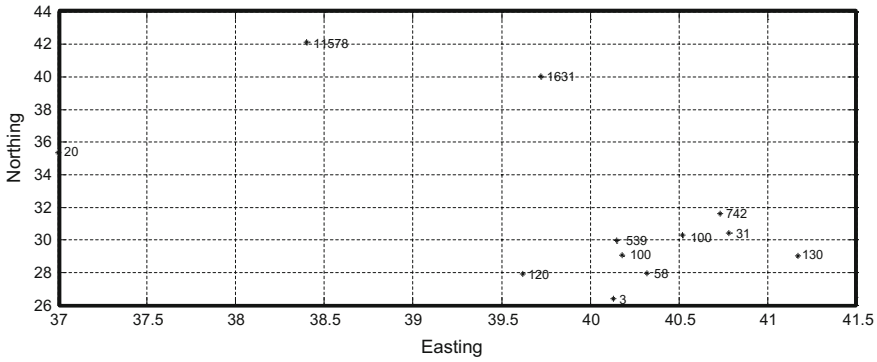


Fig. 6.4 Spatial distributions of data measurement sites

(2) There are comparatively high differences between any two closest points, which is a good indication that the variability does not abide with homogeneity principle,
(3) Extremely high values appear within the medium range of easting domain, but within the high range of the northing domain,
(4) N.

A first glance to this figure indicates that there is an increasing trend along the northing direction which can be documented if the projection on the vertical axis as in Fig. 6.5. After the projections along the easting and northing directions the following visual interpretations can be deducted.

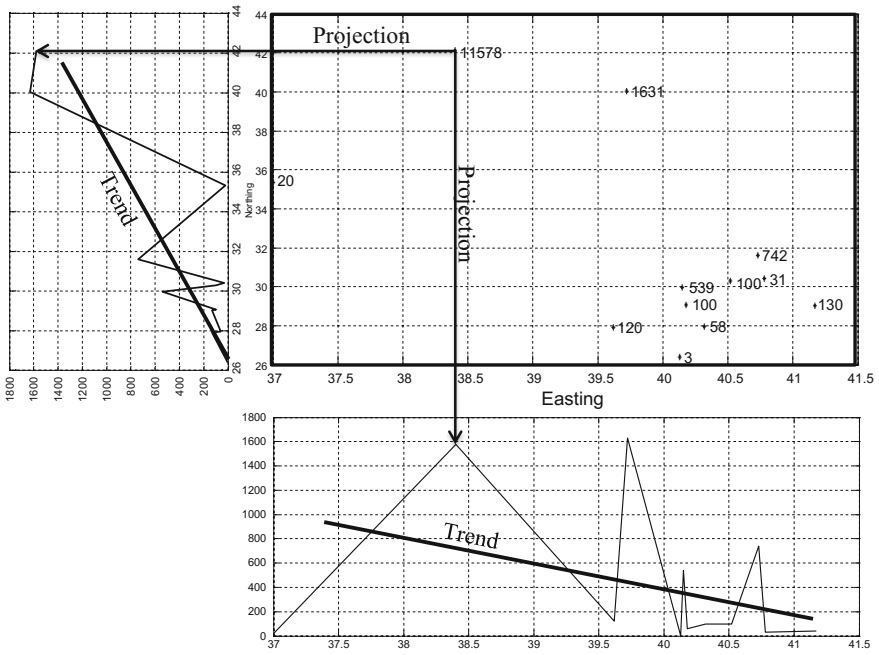


Fig. 6.5 Spatial data trend projections

(1) Along both directions now distance series are available not like time series, where spacing between two successive measurements are equal. The distance series has unequal spacing between two measurements,
(2) Along the easting direction there is a significant decreasing trend with distance as in the figure,
(3) Along the northing direction there is a significant increasing trend based on irregularly distributed distance measurements,
(4) The variability, which is the average deviations from the mean value, shows decreasing variability along the easting direction because at small distances the deviations are bigger than big distances,
(5) Along the northing direction the variability has an increasing trend, because at small distances the deviations from the mean value are smaller than the deviations at large distances,
(6) The statistical parameters (mean, standard deviation, skewness, kurtosis, etc.) and histogram of the data are the same along each direction. Projection on any direction does not change the probabilistic and statistical behaviors of the spatial data,
(7) It is possible to have projection of the same data along any direction and accordingly trend interpretations can be made.

6.4 Homogeneity and Isotropy

The statistical spatial analysis is entirely different from the numerical solution coordinate system. The following points are very specific for statistical spatial analysis and distinct from the mathematical numerical solutions.

(1) The measurement locations are irregularly (unevenly) scattered over the study area. For instance if the city centers are thought of a country, their coordinates are not regularly located, and therefore, for mapping purposes it is necessary to reduce the irregularity to a regular mesh, which is the first step in any mapping procedure in software,
(2) There is not mathematically known spatial regularity in the spatial records on any natural, environmental, economic and social studies,
(3) There may be statistically identifiable spatial regularity within the spatial event, and it is the main purpose to identify linear or nonlinear surface trends,
(4) The subtraction of each spatial data value from the corresponding spatial trend value provides a set of shifted values with zero arithmetic average. These are referred to as the residuals or stochastic part.

For many years now there have been continuous progresses to deal with the adaptation of the statistical techniques to unevenly sampled data (North et al. 1982). Regular scatter of sites might not provide enough regional information as irregular sites since earth sciences agents and surface features are almost always heterogeneous and anisotropic. Some of the significant questions concerning the spatial variability are the followings.

(1) How could one assess the regional distribution homogeneity, continuity, dependence on the basis of unevenly distributed location measurements?
(2) What are the possible models for heterogeneity so as to represent continuous variability within the study area?
(3) What is the ways of map construction from the available spatial data so as to preserve its regional variability?

In the spatial assessment of available data by scientific methodologies it is necessary to make simplifying assumptions and idealizations so as to be able to suggest a valid model for the spatial variation representation. The basic assumptions are homogeneity and isotropy, which can be decided on by comparison of the numerical quantities in a set of spatial measurement sites. The representative visual interpretations of these concepts are given in Fig. 6.6.

In any spatial modeling the first principle to decide about the type of coupled characteristics as one of the following alternatives.

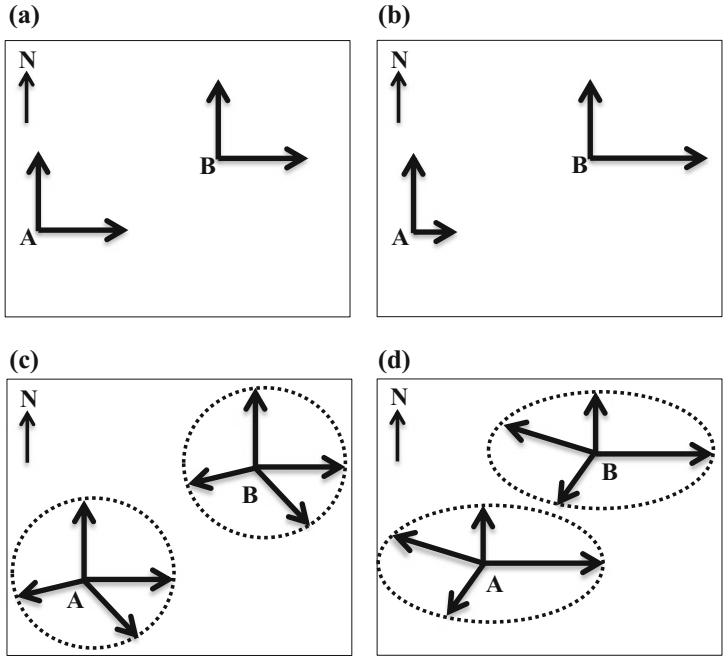


Fig. 6.6 **a** Homogeneity, **b** isotropy, **c** heterogeneity, **d** anisotropy

(1) Homogeneous–isotropic,
(2) Homogeneous–anisotropic,
(3) Heterogeneous–isotropic,
(4) Heterogeneous–anisotropic.

In Fig. 6.6 isotropy is represented by a circle and anisotropy by an ellipse. This gives the idea that the ratio of minor ellipse axis to the major one is always smaller than one in cases of anisotropy, but it is equal to 1 for isotropic situations. As for the main content of the importance is how each one of these alternatives implies trend features spatially? The homogeneous–isotropic characteristic means that the property of the spatial variable changes neither from point to point nor from direction to direction, which corresponds spatially to no trend existence. This theoretical consideration finds its practical counterpart as the maximum relative error between any two points is less than $\pm 5\%$. This error limit is valid in all the calculations after this point for practically significant trend identification. In some application this error limits can be adopted at the maximum as $\pm 10\%$. For example, is the arithmetic average of the spatial variable at each measurement site does not have more than $\pm 5\%$ among all the sites, then the spatial variable is considered as trendless pointwise and directionwise. Similar conclusions can be reached for other statistical parameters.

The statistical homogeneity and isotropy properties of spatial variables can be defined in reference to statistical parameters such as the mean, variance, skewness, kurtosis, and even the probability distribution function (PDF). For instance, a spatial variable is homogeneous and first-order stationary, if the mean is independent of a translation of the two positions. This means that the mean depends only on distance.

A spatial variable is isotropic in reference to the mean, if it is independent of a rotation in the field around the center point on the line between two positions.

In general, natural phenomena physical processes have preferred orientations. It is necessary to identify and separate spatial trend component for spatial stochastic modeling of any natural, social, economic and engineering phenomenon, so that the spatial correlation coefficient can be obtained for better representation of the spatial uncertainty. This is similar to pre-whitening procedure as explained earlier in Chap. 5. For arriving at the simple construction of the spatial correlation function it a prerequisite to decide on the homogeneity and isotropy properties. The assumption of homogeneity and isotropy make the phenomenon concerned dependent only on distance (Thiebaux and Pedder 1987; Daley 1991; Şen 2008). For instance, the real atmosphere is homogeneous and isotropic, furthermore homogeneity and isotropy assumptions in meteorological modelling are assumed by Gandin (1963), Eddy (1967) and Kruger (1964, 1969).

For example, at the mouth of a river the coarse material settles out fastest, while the finer material takes longer to settle. Thus, the closer one is to the shoreline the coarser the sediments while the further from the shoreline the finer the sediments. When interpolating at a point, an observation 100 m away but in a direction parallel to the shoreline is more likely to be similar to the value at the interpolation point than is an equidistant observation in a direction perpendicular to the shoreline. Anisotropy takes these trends in the data into account during the gridding process (Şen 2008). For instance sea surface heights from the bottom vary by space and time anisotropic ally.

All the equal value line sets (contour lines) in the form of a map as in Fig. 6.5 are reflection of anisotropy, where the spatial variable is temperature (Fig. 6.7).

Easting and Northing are measured in the same units, but temperature is in centigrade degree (°C).

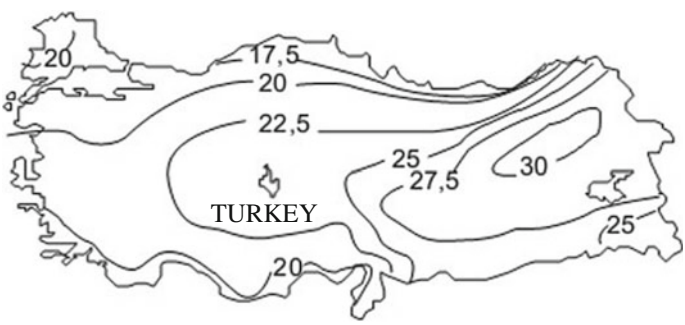


Fig. 6.7 Temperature contour maps

6.5 Spatial Trend Surfaces

The main purpose is to model the spatial behavior of natural, environmental, and economic phenomena. The trend surface passes through rather uncertain and complex spatial data scatter over a region. Its application can be achieved as geographic information for continuous events in space and the measurements must be at cardinal levels.

The basic principle in trend surface analyses is matching a continuous surface to the available spatial data through a regression function.

In order to facilitate the spatial trend concept the best example is a topographic map where the independent variables are longitudes and latitudes with spatial variable as altitude (elevation from the mean sea level). For this purpose, the spatial topographic variability must be sampled at $n \times n$ sites that are irregularly distributed in the study area. To reach to the final trend surface there are following three steps that should be completed in sequence.

(1) Model selection and parameter estimation: If possible, with an expert view, one can guess the most convenient linear or polynomial mathematical form for the spatial trend component of the spatial variable measurements. In doing so one should keep in mind that the trend surface should explain as much as the regional variability in terms of spatial variance. In practical application, it is recommended that polynomial degree must not be preferably more than 5 or 7. In case of 5th degree of polynomial there will be 21 coefficients for the model parameter estimations (see Sects. 6.5.5 and 6.8.1). In a 7th order spatial trend surface mathematical expression there are 28 coefficients.
(2) Model validation: After the model parameter estimations the model (regression) function should be then applied to an independent set of sample points for validation purpose by taking into consideration cross-validation,
(3) Model estimations of spatial variable: The developed model after the execution of the two previous steps, the model now can be used for spatial variable estimations at any desired point within the study area. One should be cautious at this stage to extend (extrapolate) model estimations outside the study area.

Prior to the spatial trend identification methodologies introduction, it has utmost importance to visualize what might be the alternatives of spatial trend types? For this purpose in this section different possibilities are presented with their general mathematical expressions. One should keep in mind that spatial trend mathematical expression can be represented in its implicit form as,

$$z(x, y) = f(x, y), \tag{6.6}$$

where z is the spatial variable; x and y are the spatial reference variables (coordinates). In this expression x and y are independent variables, whereas z is the

dependent variable, which is for modeling. This expression implies that the spatial data are always in the form of triplet, which describes a surface over the basic variables x and y. One can suggest various geometrical shapes in three-dimensional (3D) space for possible spatial trend components.

The general explicit form of Eq. (6.6) may be in the form of either in linear regression function of may be either a flat or oriented plane or a curved surface with an increasing number of curvatures as explained in Sect. 6.5.5.

The spatial trend surface fit to a set of spatial variable measurements can be achieved by the least squares technique. The surface must be such that it minimizes the variance of the surface with respect to the input values. The fitted surface rarely coincides with some of the measurement points, but it is susceptible to outliers in the data. Trend surface analysis is used to find general tendencies of the sample data, rather than to model a surface precisely and completely.

One can obtain regular grid points from irregular measurement sites by fitting a surface to available data. It can be achieved either globally or locally over the study area. In the former case there is a single functional form of the trend surface in addition to the stochastic nature of the residuals and the latter case is just the repetition of global procedure on pieces of subareas within the study area. Another version of the local surface fitting is to consider the neighbor points to reach to locally representative trends. However, the most widely used procedure is the global trend surface search for this purpose the spatial variable is approached by a polynomial expansion of the geographic coordinates, and the coefficients of the polynomial function are estimated from available measurements by means of the least squares method, which relies on the sum of the squared deviations minimization from the trend surface. After the identification of the trend surface the sum of the trend surface value at a site and the residual is equal to the measurement value. The residuals are random variables either with independent structure in which case the probability distribution (PDF) is the representative of the spatial randomness or in the case of dependent spatial residuals one of the most convenient stochastic processes or the regionalized variable approach through the Kriging methodology is the most representative approach (Sect. 6.3).

In general, the polynomial functions can be of any desired degree, the higher is the degree the more is the computation rounding error. Therefore, in practice the highest degree is adapted as 5 or 7. The unknown parameter coefficients of the polynomials can be estimated from the simultaneous solution of a set of convenient number of equations as explained in Sect. 6. Each one of these equations include the sums of powers and cross products of the geographic (X, Y), and spatial variable Z values. After the estimation of the coefficients the polynomial function provides opportunity to calculate spatial variable value at any site (point) within the study area. In general, in order to map the spatial variable one can calculate the spatial variable values at regular mesh nodes and subsequently obtain the 2D contour or 3D map for the spatial variable.

Fixation of a trend surface to given set of spatial variable measurements at irregular sites separates the whole measurements into two components, namely trend and residual values. Trend values are collection of deterministic quantities, but the residuals are uncertainty parts.

Two significant points that are valid in the parameter estimations of any trend surface are the following:

(1) Zero average: The trend surface must be in such a location that the summation of the deviations between the spatial variable measurements and their corresponding points on the trend surface must be equal to zero, or at least within ±10 or better ±5 error limits,
(2) Minimum variance: The summation of the square deviations should be as small as possible. This provides opportunity for selection among trend surfaces the best one, such that if the researcher tries to fit a set of trend surfaces the one with the minimum variance is the best.

6.5.1 Horizontal Plane

This corresponds to the case when all the trend surface points have the same (constant) spatial value as in Fig. 6.8. Such a surface in the form of a plane provides homogeneity and isotropy of the spatial variable as for its spatial trend component is concerned.

The simplest and most common form of ReV is a triplet and therefore it is illuminating first to consider the surface in 3D and then according to the SV definition it is possible to infer its shape intuitively by mental experiment.

The mathematical expression of this simplest trend case, in fact, no trend situation can be expressed as,

$$z(x, y) = c, \tag{6.7}$$

where c is a constant value as shown in Fig. 6.8. Its statistical counterpart is the spatial arithmetic average, $\bar{z}$, of the spatial variable.

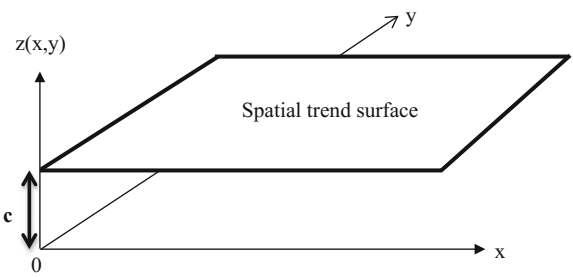


Fig. 6.8 Spatial trend components with homogeneity and isotropy

6.5.2 Horizontal Planes

Within the spatial variable there may not be any trend component but a sudden (abrupt) jump as in Fig. 6.9. In such a case, there are two no spatial trend regions with sudden change (upwards or downwards) in between. The horizontal continuity in Fig. 6.9 is disrupted by a discontinuous feature (cliff, fault, facies change, boundary, etc.).

The mathematical expressions of these spatial trend surfaces are expressible similar to Eq. (6.7) in two parts as,

$$z_L(x, y) = c_L \tag{6.8}$$

and

$$z_U(x, y) = c_U \tag{6.9}$$

or

$$z_U(x, y) = c_{LU} + z_L(x, y), \tag{6.10}$$

where c_L, c_U and c_{LU} are the constants that should be calculated from the available spatial data. Simply, c_L and c_U are the arithmetic averages of the lower and upper trend plane dominant areas. Furthermore $c_{LU} = c_L - c_U$ is the difference between the two arithmetic averages.

6.5.3 Inclined Trend Plane

This is the most commonly thought and in practical applications frequently employed spatial trend form, which is in the form of an inclined plate as in Fig. 6.10.

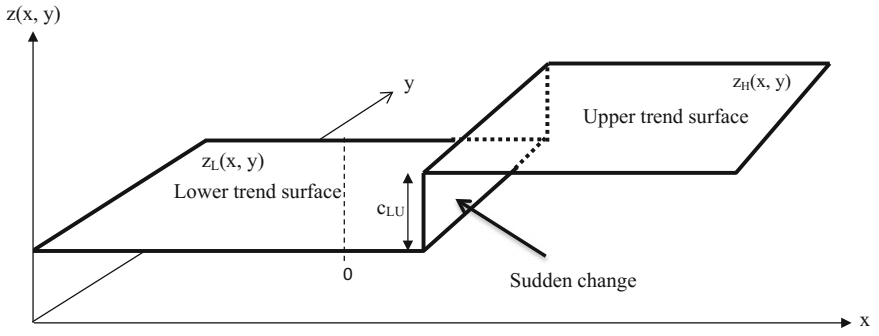


Fig. 6.9 Horizontal planes

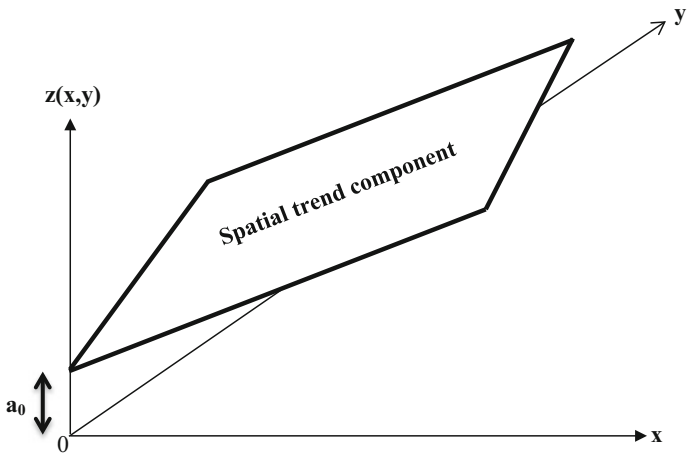


Fig. 6.10 Continuous linear trends

The mathematical form of this spatial trend surface will include only linear contributions from the coordinate variables as follows.

$$z(x, y) = a_0 + a_1 x + a_2 y, \tag{6.11}$$

where a_0, a_1 and a_2 are the intercept, x direction and y-direction slopes, respectively.

6.5.4 Inclined Trend Planes

It is also possible to have abrupt change with inclined spatial trend surface together in a spatial data structure. This is exemplified in Fig. 6.11.

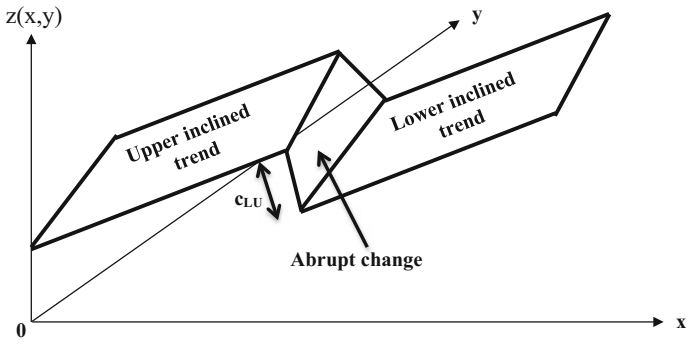


Fig. 6.11 Inclined trend plates

The mathematical function for the description of inclined trend plates is similar to the previous expressions. They can be written for the lower and upper trend surfaces as follows:

$$z_L(x, y) = a_0 + a_{1L}x + a_{2L}y \tag{6.12}$$

and

$$z_U(x, y) = a_0 + a_{1U}x + a_{2U}y. \tag{6.13}$$

6.5.5 Curved Trend Surface

There are several alternatives that can be employed in the spatial trend surface depending on the curvature tendencies. The following mathematical set of equations is representative of such spatial surfaces. They represent the first order bilateral, second order, and third order trend surfaces, respectively.

$$z(x, y) = a_0 + a_1x + a_2y + a_3xy \tag{6.14}$$

$$z(x, y) = a_0 + a_1x + a_2y + a_3xy + a_4x^2 + a_ry^2 \tag{6.15}$$

and

$$z(x, y) = a_0 + a_1x + a_2y + a_3xy + a_4x^2 + a_5y^2 + a_6x^3 + a_7y^3 + a_8xy^2 + a_9x^2y. \tag{6.16}$$

The geometric forms of these expressions are Fig. 6.12a–c corresponding to each one of them, respectively.

All the trend surfaces in Fig. 6.12 are smooth surfaces, which are generated artificially according to aforementioned mathematical expressions. However, natural surfaces are not in this form, but perhaps it is a mixture of such smooth surfaces piece by piece as in Fig. 6.13.

6.5.6 Random Surface

In some cases of the natural, environmental, economic or social spatial data, there may not be any spatial dependence among the measurement values and in this case the surface is in the form or random variations. For instance, rough sea surface is a valid example for such a spatial surface. Figure 6.14 is a representative form for such a situation.

In the completely random spatial distribution there is no spatial correlation as in Fig. 6.14 then the arithmetic average, μ, and the variance, σ^2, are the two

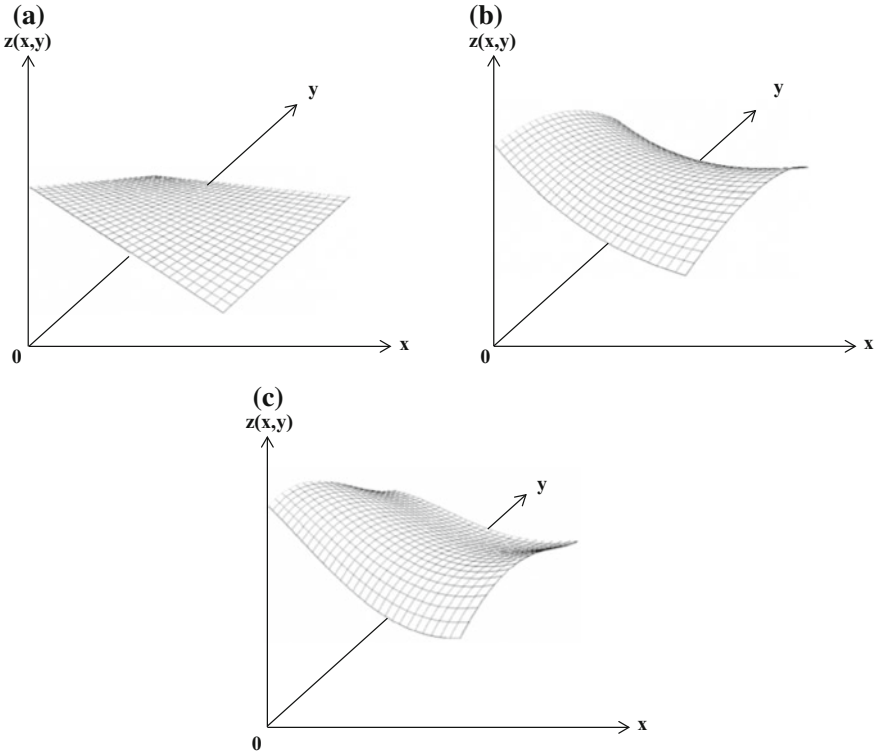


Fig. 6.12 Spatial trend surfaces **a** first order, **b** second order, **c** third order trend surfaces

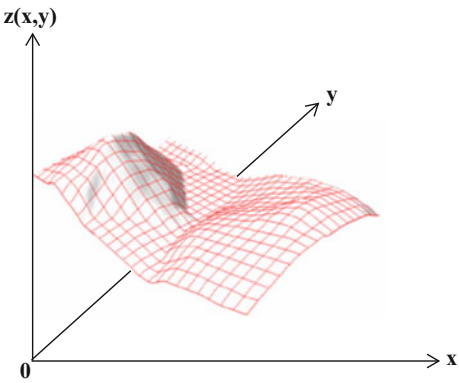


Fig. 6.13 Natural phenomenon surfaces

fundamental parameters for their representation in addition to the PDF. In order to decide about the spatial correlation existence or not either spatial dependence function or the spatial autocorrelation functions are of great help.

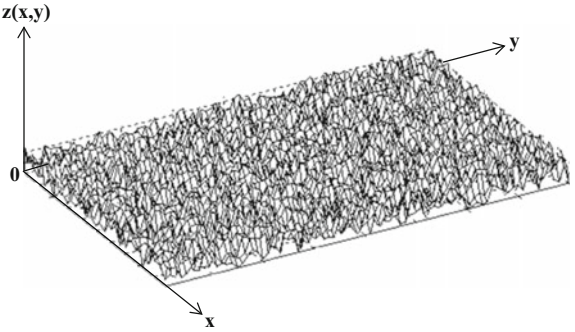


Fig. 6.14 Independent spatial data

6.6 Spatial Dependence Function (SDF)

Each individual measurement site represents a very considerable area around it. Logically, measurement at any individual site will have an area of influence or in isotropy case a radius of influence around it, but there is no physically or data based objective criterion for the definition of such an area, but one can find the quantitatively its magnitude from SDF, which is an indicator of spatial variable uncertainty (probabilistic, statistical, stochastic) dependence that provides visual and quantitative information about the dependence between any two locations. The dependence can be measured by covariance provided that the uncertainties are distributed according to the Gaussian (normal) PDF, otherwise semivariogram, cumulative semivariogram or point cumulative semivariogram functions should be used (Şen 2008). After all what have been explained in the previous sections of this chapter the reader can look at the spatial variable surface by using one of the software roughly and examine the three-dimensional cases. For instance, in Fig. 6.15 three different such rough maps are presented. Through a visual inspection of such rough maps one can then decide approximately what might be the degree of polynomial for the surface? The inspector may describe the mat ah his/her focus verbally whether the pattern has homogeneity and dependence.

A first glance on these two representative maps gives visual inspection about possible surface fitting to each one of them For instance, in Fig. 6.15a, the rough surface has n inclined plane feature as already theoretically explained in Sect. 6.5.3 with the mathematical expression in Eq. (6.7). The 3D map in Fig. 6.15b is rougher than the previous one and a mixture of surfaces explained in Sect. 6.5.5 can be used for the smooth surface representation. For this purpose, one can start with third order polynomial function and continue to increase the degree of the polynomial up to 7 and select among them the one with the least sum of square residuals squares. The reader should keep in mind that the higher the degree of the polynomial the rougher gets the surface.

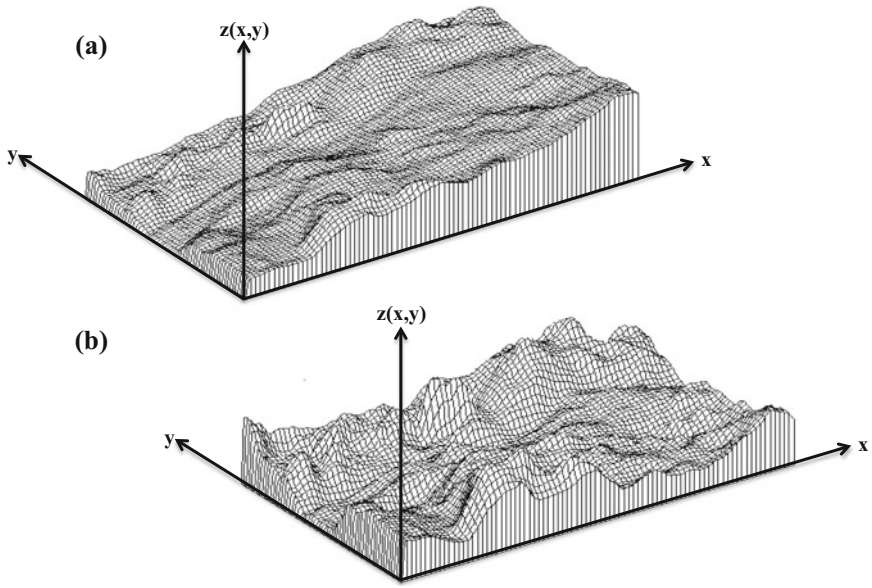


Fig. 6.15 Sample maps

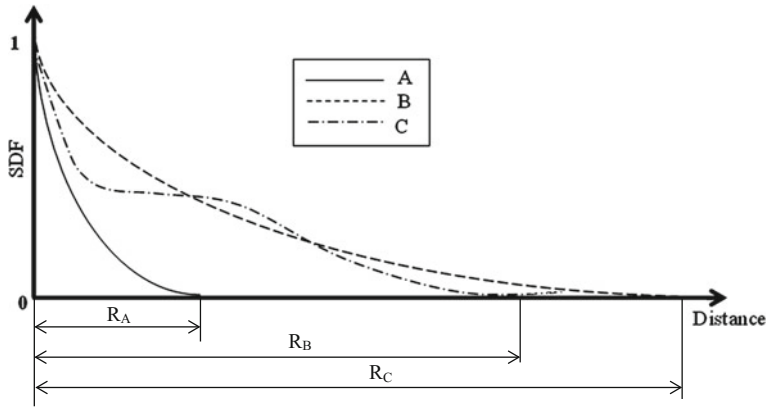


Fig. 6.16 Spatial dependence functions

Logically, closer measurements are more dependent on each other than the farther ones. The further apart the measurement sites, i.e., as the distance increases between the two points of a spatial variable the dependence becomes close to zero. These sentences picture the form of spatial dependence function (SDF) as a set of alternatives in Fig. 6.16.

It is to be noticed that at zero distance the spatial variable is 100% dependent on itself and hence, one can regard the quantitative value of the correlation at zero distance as 1, and then onwards as the distance increases the correlation decreases

steadily until reaches the horizontal axis where the correlation value is equal to zero. Among the SDFs in the figure A and B reflect homogeneous and isotropic spatial behaviors without any trend ingredients. Type A has a short spatial dependence compared to the B SDF, which has longer spatial (distance) effectiveness. The third one, C, is comparatively different than the two and it has some internal structure that cannot be considered as homogeneous and isotropic, because its SDF does not drop smooth to zero. Furthermore, after a certain distance the spatial variable is completely independent, which mean that points that are apart from each other more than this distance do not affect each other? In Fig. 6.16 the dependence distances are referred to as the radii of influence and they are $R_A < R_B < R_C$.

The measurements of the spatial variable at a set of sites provide numerical information about the spatial behavior interpretation of the variable as in Fig. 6.1. In general, the larges is the variability as explained in Sect. 6.3; the more is the heterogeneity and this point out that the number of data to represent the spatial variability is more than the homogeneous case. Additionally, the larger is the variability the smaller is the spatial dependence even between the points that are close to each other.

In order to quantify the degree of variability within spatial data, variance techniques can be used in addition to classical autocorrelation methods (Box and Jenkins 1970). However, these methods are not helpful directly to account for the spatial dependence or for the variability in terms of sample positions. The drawbacks are due to either non-normal (asymmetric) distribution of data and/or irregularity of sampling positions. However, the semivariogram (SV) technique, developed by Matheron (1963, 1965) and used by many researchers (Clark 1979; Cooley 1979; David 1977; Myers et al. 1982; Journel 1985; Aboufirassi and Marino 1984; Hoeksema and Kitanidis 1984; Carr et al. 1985) in diverse fields such as geology, mining, hydrology, earthquake prediction, groundwater, etc., can be used to characterize spatial variability and hence the SDF. The SV is a prerequisite for best linear unbiased prediction of ReVs through the use of Kriging techniques (Krige 1982; Journel and Huijbregts 1978; David 1977).

6.6.1 Spatial Correlation Parameter Calculation

Its definition is very similar to timewise correlation definition, which has been given notationally in Chap. 2. The spatial correlation coefficient, $\rho_{i,j}$ between points I and j can be written as,

$$\rho_{ij} = \frac{\overline{(Z_i^m - \overline{Z}_i)(Z_j^m - \overline{Z}_j)}}{\sqrt{\overline{(Z_i^m - \overline{Z}_i)^2}\,\overline{(Z_j^m - \overline{Z}_j)^2}}} \tag{6.17}$$

where over bars indicate time averages over a long sequence of past observations, Z_i^m and Z_j^m represent measurement values at these stations, and finally, $\overline{Z}_i$ and $\overline{Z}_j$ are the areal mean of the spatial variable. It is obvious that $-1 < \rho_{i,j} < +1$, with $a <$ completely dependent case in between the two limits corresponds to $\rho_{i,j} = 0$. In case of n measurement sites, then there will be $m = n(n-1)/2$ pairs of distances and corresponding correlation coefficients. Their plot results in a scatter diagram similar to one of the SDFs in Fig. 6.16. Equation (6.17) is useable in the cases, when there are a series of measurements at each site in the form of a time or spatial series.

Especially in earth systems and environment domain measurements there are single measurements at each site. In such a case Eq. (6.17) cannot be used, and therefore, it is necessary to propose a suitable SDF. Such a SDF derivation is suggested by Şen (2008) with the following steps.

(1) Find the set of actual distances among each pair of sites, hence, if there are n sites there will be $n(n-1)/2$ different distances,
(2) Find the squared differences among each pair of the spatial variable. The same number of squared differences will be obtained,
(3) Plot distances versus squared differences of the spatial variable, and hence an irregular graph of the squared differences variation will be reached,
(4) Plot the successive cumulative sums of the square distances along the distance sequence. The result will be another graph that shows the change of squared difference accumulation by distance,
(5) The last value in this graph is the maximum squared distance summation, and it is very significant for RDF calculation,
(6) Divide each cumulative squared difference values in the last graph by this maximum value. The result will be the change of scaled cumulative squared difference values by distance, where the values on the vertical axis changes between zero and one,
(7) Finally, subtract scaled cumulative squared differences from one and the resulting graph will appear in the form of decreasing trace similar to the cases in Fig. 6.16.

Example 6.1 The earthquake magnitude measurements at 5 stations are presented in Table 6.2. Construct the RDF for these spatially varying values.

Solution 2: First of all the irregular locations of each site are given in Fig. 6.17, which shows that the scatter points are heterogeneously and unevenly distributed.

Table 6.2 Spatial data

Spatial variable	Easting (km)	Northing (km)	Magnitude
Z_1	24.47950	40.40017	1.87
Z_2	24.48400	39.86550	1.92
Z_3	24.68783	40.12017	2.2
Z_4	24.63183	39.75367	2.15
Z_5	24.54383	40.87950	2.07

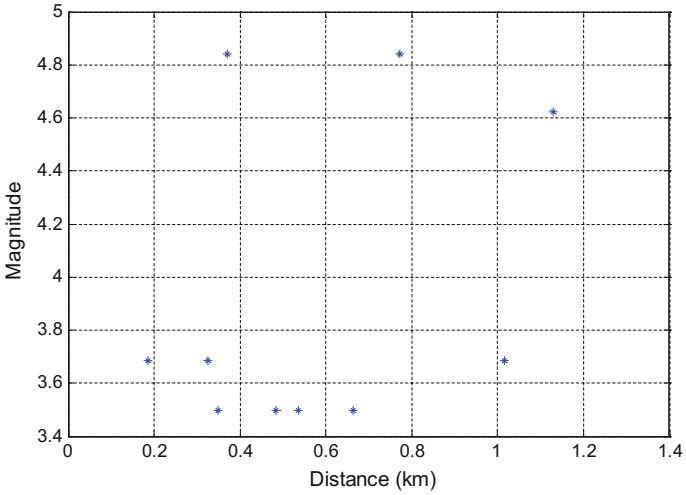


Fig. 6.17 Spatial variable scatter with distance

As for the calculations, the first step is to calculate the distances between each pair of sites, and this can be best achieved in the of distance matrix form as follows, where there are $5(5-1)/2 = 10$ distinctive distance values.

	Z_1	Z_2	Z_3	Z_4	Z_5
Z_1	0				
Z_2	0.535	0			
Z_3	0.349	0.326	0		
Z_4	0.664	0.185	0.371	0	
Z_5	0.484	1.016	0.773	1.129	0

Likewise, this time the same size of matrix is filled in with the squared differences among each pair of spatial variable measurements as follows.

	Z_1	Z_2	Z_3	Z_4	Z_5
Z_1	0				
Z_2	3.4969	0			
Z_3	3.4969	3.6864	0		
Z_4	3.4969	3.6864	4.84	0	
Z_5	3.4969	3.6864	4.84	4.6225	0

The plot of values in the distance matrix against the squared-difference values yields to a nondecreasing curve as in Fig. 6.18.

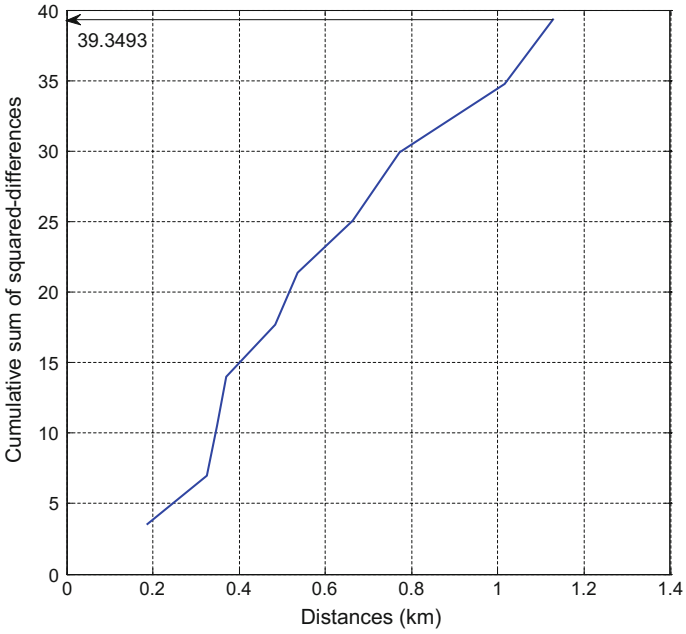


Fig. 6.18 Spatial variable cumulative sum squared differences scatter with distance

In this figure the maximum value of the cumulative sum of squared differences is equal to 39.3493, and the division of all the values in this figure by the value leads to similar curve that has standard cumulative sum of squared differences as in Fig. 6.19.

It is to be noticed that the values on the vertical axis vary between zero and one, and furthermore, the vertical axis has no dimension because of the division procedure. The final step in the regional dependence function development procedure is to subtract vertical values in this figure from one, and the resulting function is the RDF as in Fig. 6.20.

The RDF is a line that decreases with the distance; it has a value equal to one at zero distance, which implies that the measurement at any site is 100% is correlated with itself, whereas the zero value of the RDF is equivalent with the influence of distance, which appears in Fig. 6.20 as and at.

6.7 Double Mass Curve Test

During a long time period in a region many stations may monitor the same spatial variable such as the population, economic indices, rainfall, soil moisture, groundwater level fluctuation, etc. It is not reasonable to expect that at all the points the records will be kept without any external effect that leads to bias. Consider a rain

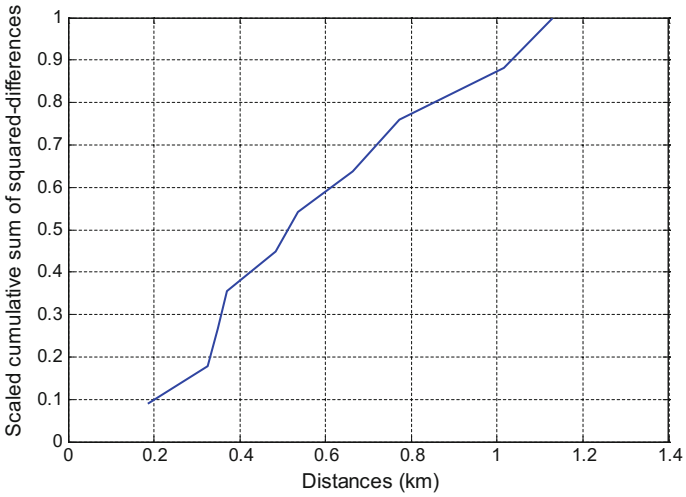


Fig. 6.19 Spatial variable scaled cumulative sum squared differences scatter with distance

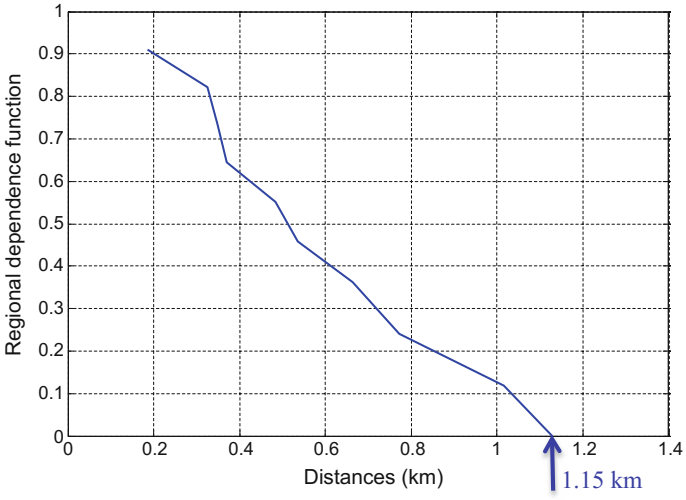


Fig. 6.20 Reginal dependence functions

gauge and if it is damaged, or there is urban development nearby, or growing tree will affect the recorded rainfall amounts. It may be necessary to replace the station location due to some activities or at the same station modern measurement stations may be located. All of such changes will affect the heterogeneity of the records.

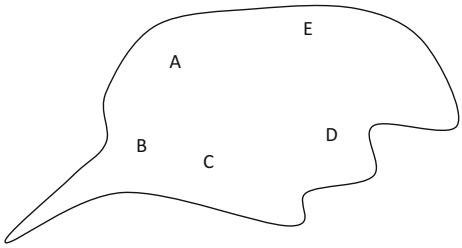


Fig. 6.21 Five hydrology stations

Table 6.3 Annual average rainfalls

Year	Stations				
	A	*B*	*C*	*D*	*E*
1995	$Y_{A,1}$	$Y_{B,1}$	$Y_{C,1}$	$Y_{D,1}$	$Y_{E,1}$
1994	$Y_{A,2}$	$Y_{B,2}$	$Y_{C,2}$	$Y_{D,2}$	$Y_{E,2}$
1993	$Y_{A,3}$	$Y_{B,3}$	$Y_{C,3}$	$Y_{D,3}$	$Y_{E,3}$
…	…	…	…	…	…
…	…	…	…	…	…
1970	$Y_{A,20}$	$Y_{B,20}$	$Y_{C,20}$	$Y_{D,20}$	$Y_{E,20}$
1969		$Y_{B,21}$	$Y_{C,21}$	$Y_{D,21}$	$Y_{E,21}$
…	…	…	…	…	…
…	…	…	…	…	…
1955		$Y_{B,30}$	$Y_{C,30}$	$Y_{D,30}$	$Y_{E,30}$
1954		$Y_{B,31}$	$Y_{C,31}$	$Y_{D,31}$	$Y_{E,31}$
…	…	…	…	…	
…	…	…	…	…	
1950		$Y_{B,46}$		$Y_{D,46}$	
1949		$Y_{B,47}$		$Y_{D,47}$	
…	…	…			
…	…	…			
1940		$Y_{B,55}$			

Hence, the question is whether a sequence of record at a site has its homogeneity throughout the measuring duration? Let us consider that in a region there are five (*A*, *B*, *C*, *D* and *E*) stations for some hydrologic measurement as in Fig. 6.21.

It may be necessary to check whether the records at *C* have any significant change at some time during the record period. In Table 6.3 up to 1995 the hydrologic records are given annually. It is possible that all the stations might have not started in the same year. In this table there are 5 SFs with 20 year records. In order to control the homogeneity of records at station C the following steps must be executed.

1. On the vertical axis the cumulative rainfall amounts, Y_{YE} at stations *A, B, D,* and *E* are shown whereas on the horizontal axis the cumulative rainfall amounts, Y_{CE}, in station *C* are shown (see Fig. 6.21). For instance, for year 1993 on the vertical axis $Y_{CE} = Y_{C,1} + Y_{C,2} + Y_{C,3} + Y_{C,4}$ and on the horizontal axis $Y_{YE} = (Y_{A,1} + Y_{A,2} + Y_{A,3} + Y_{A,4}) + (Y_{B,1} + Y_{B,2} + Y_{B,3} + Y_{B,4}) + (Y_{D,1} + Y_{D,2} + Y_{D,3} + Y_{D,4}) + (Y_{E,1} + Y_{E,2} + Y_{E,3} + Y_{E,4})$ are shown. Hence, the point is labelled by 1993. The cumulative rainfall amounts are started from 1995 downward.
2. Plotting of cumulative rainfall amounts in this manner for each year leads to a scatter diagram.
3. If all the scatter points appear along a single line, then the records at *C* are homogeneous and do not need any adjustment. This test is a visual assessment without any numerical calculation.
4. If as in Fig. 6.22 there are two straight lines, then the records at *C* are not homogeneous and must be adjusted. In this figure the break point appears in 1992. This is the year that the homogeneity of the records has started to deviate from the general trend. tedir. İşte bu yılda yapılan ölçümlerin tektürlülüklerinin bozulmaya başladığı anlaşılır.
5. In order to make the adjustment, the broken line must be completed to a single straight line by extending prior to 1992 straight line toward the recent years.
6. If the slopes of straight lines before and after the break point are S_b and S_a the adjusted rainfall amounts at station *C* can be calculated as follows.

$$(Y_{Ci})_A = \frac{\tan S_p}{\tan S_a}(Y_{Ci})_O, \tag{6.18}$$

where $(Y_{Ci})_A$ and $(Y_{Ci})_O$ are the adjusted and observed rainfall amounts at station *C* and at year, *i*. The same procedure can be repeated for any other station in suspicion.

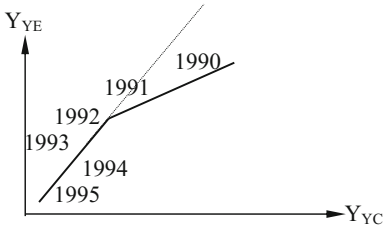


Fig. 6.22 Double mass curve tests

6.8 Trend Surface Analysis

This is three-dimensional trend surface regression methodology that can be either linear (planar) or nonlinear (spatial curvature). Apart from these two approaches one can also make directional trend analysis along different directions by projections on the preferable direction. It is not possible to visualize these three dimensions with necked eye.

6.8.1 Planer Trend Regression Analysis

So far explanations of spatial trend component were all qualitative and visually quantitative as in Figs. 6.1 and 6.2. In majority of cases a spatial trend is a planer surface and rarely is it in the form of curvature surface of the second order degree usually over geographical (longitude and latitude) directions as in Fig. 6.23.

In this figure X and Y are most of the time longitude (easting) and latitude (northing) directions, whereas the Z direction is for the spatial date values. Linear trend surface analysis is also called as "spatial interpolation" method. The trend surface provides a means of spatial interpolation possibility.

The classical trend surface methodology is a way of fitting the entire surface with a linear or polynomial equation with parameters, which are estimatable from the given data set by means of the least squares technique. For this purpose, in many works trend surfaces are the only basic tool as maps for communication in any scientific domain spatial variable concerned. After the analysis the trend model statistical model coefficients are estimated and the final product is presented as a contour map, which is the same as the preparation of topographic maps. In the spatial trend analysis methods as presented in this section RDF is not taken into consideration explicitly.

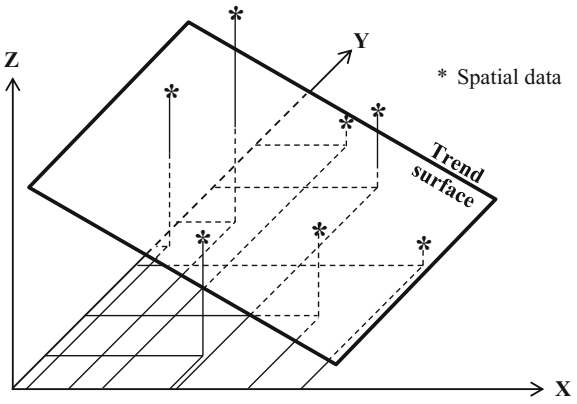


Fig. 6.23 Linear trend surfaces

The mathematical form of planar trend surface has the linear contributions of the geographical coordinates, i.e., data measurement point longitude and latitude or easting and northing direction values as x and y. If the spatial variable values are shown by z, then the planar model can be expressed as follows.

$$z = a_0 + a_1 x + a_2 y, \tag{6.19}$$

where a_0, a_1 and a_2 are the model parameters. This expression represents deterministic trend surface without any measurement error. However, addition of the error (uncertainty) component, ε, to this equation gives the representative positions of the data values with uncertainty. The uncertain expression can be written from Eq. (6.19) as,

$$z = a_0 + a_1 x + a_2 y + \varepsilon \tag{6.20}$$

The least squares technique implies that the sum of the error squares must be "minimum". For this purpose, first the error square is expressed from Eq. (6.20) as,

$$\varepsilon^2 = (z - a_0 + a_1 x + a_2 y)^2 \tag{6.21}$$

The sum of error squares, SS, from each data, i, can be written for i = 1, 2, …, n spatial data values as,

$$SS = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (z_i - a_0 + a_1 x_i + a_2 y_i)^2 \tag{6.22}$$

The request is that this SS must be the minimum, where the word "minimum" means in the calculus that the partial derivative with respect to each parameter must be simultaneously be equal to zero. Hence, one needs to take the derivatives as follows.

$$\frac{\partial SS}{\partial a_0} = 2 \sum_{i=1}^{n} (z_i - a_0 - a_1 x_i - a_2 y_i)(-1) = 0 \tag{6.23}$$

$$\frac{\partial SS}{\partial a_1} = 2 \sum_{i=1}^{n} (z_i - a_0 - a_1 x_i - a_2 y_i)(-x_i) = 0 \tag{6.24}$$

and

$$\frac{\partial SS}{\partial a_2} = 2 \sum_{i=1}^{n} (z_i - a_0 - a_1 x_i - a_2 y_i)(-y_i) = 0. \tag{6.25}$$

A close inspection to these expressions reveals simplification after expansion of each parenthesis and division of both sides of each equation by the number of data, n. The completion of these points leads to the following set of equations.

$$\overline{z} = a_0 + a_1\overline{x} + a_2\overline{y} \tag{6.26}$$

$$\overline{zx} = a_0\overline{x} + a_1\overline{x^2} + a_2\overline{yx} \tag{6.27}$$

and

$$\overline{zy} = a_0\overline{y} + a_1\overline{xy} + a_2\overline{y^2} \tag{6.28}$$

The overbars in these expressions indicate the arithmetic averages of the terms under the overbar sign. For instance, $\overline{zx}$ indicates that the z data series must multiplied by the corresponding x series and then the new values of the multiplication series arithmetic average is equal to the $\overline{zx}$ value. If the arithmetic averages are calculated according to the given spatial data, then the unknowns are the three model parameters, and since there are three equations they can be solved easily. A close inspection to Eqs. (6.26)–(6.27) provide a very practical way of obtaining the basic equations without derivation manipulations. For the practical derivation of the final three expressions the main model formulation in Eq. (6.19) must be considered. The simple and practical rule has three steps as follows.

(1) For Eq. (6.26) take the arithmetic average of Eq. (6.19),
(2) For Eq. (6.27) first multiply both sides of the main expression by the first independent variable, x, and then take the arithmetic average of both sides. The important point to be noticed at this stage is that $\overline{zx} \neq \overline{zx}$,
(3) For Eq. (6.28) first multiply both sides of the main expression by the second independent variable, y, and then take the arithmetic average of both sides.

Another procedure that helps in practical calculations of the planar trend parameters is to prepare a similar table to Table 6.4.

The first three columns in this table are for the spatial data values and the remaining each column corresponds to needed regression terms. The last column includes the arithmetic averages in Eqs. (6.26)–(6.27).

Example 6.1 A set of earthquake records are given for some part of Turkey and the magnitude is required to be related to easting (x) and northing (y) coordinates for the prediction of Richter magnitude. The first three columns of Table 3.5 give the easting, northing and earthquake magnitude values, respectively. For the sake of argument simple linear trend surface equation is adopted as in Eq. (6.19).

Table 6.4 Planar trend calculations

1	2	3	4	5	6	7	8
Coordinates		Data	Regression coefficient calculation columns				
Easting, x	Northing, y	z	zx	x^2	yx	zy	y^2
x_1	y_1	z_1	z_1x_1	x_1^2	y_1x_1	z_1y_1	y_1^2
x_2	y_2	z_2	z_2x_2	x_2^2	y_2x_2	z_2y_2	y_2^2
x_3	y_3	z_3	z_3x_3	x_3^2	y_3x_3	z_3y_3	y_3^2
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
x_{n-1}	y_{n-1}	z_{n-1}	$z_{n-1}x_{n-1}$	x_{n-1}^2	$y_{n-1}x_{n-1}$	$z_{n-1}y_{n-1}$	y_{n-1}^2
x_n	y_n	z_n	z_nx_n	x_n^2	y_nx_n	z_ny_n	y_n^2
$\overline{x}$	$\overline{y}$	$\overline{z}$	$\overline{zx}$	$\overline{x^2}$	$\overline{yx}$	$\overline{zx}$	$\overline{y^2}$

According to the steps given above the rest of the table is prepared accordingly (Table 6.5).

There are three unknowns and it is necessary to obtain three equations in the light of practical trend surface calculations. By considering the last row of averages from Table 3.8 one can write the necessary equations simply as

$$\begin{aligned} 6.38 &= a_0 + 39.12a_1 + 33.30a_2 \\ 249.64 &= 39.12a_0 + 1533.62a_1 + 1306.02a_2 \\ 211.97 &= 33.30a_0 + 1306.02a_1 + 1150.11a_2 \end{aligned}$$

The simultaneous solution of these equations yields $a_0 = 5.63$, $a_1 = 0.0314$ and $a_2 = -0.0143$, and hence the final linear trend surface expression is given as,

$$z = 5.56 + 0.0314x - 0.0143y$$

6.8.2 Polynomial Trend Regression Analysis

The polynomial trend surface mathematical formulation relates the geographic variables x and y to the spatial variable value, z as,

$$z = a_0 + a_1x + a_2y + a_3x^2 + a_4xy + a_5y^2, \tag{6.29}$$

where a_i's (i = 1, 2, 3, 4, 5) are the model parameters that must be estimated from given spatial data measurements based on the geographic location coordinates. Depending on the parameter signs this model provides concave or convex smooth

Table 6.5 Earthquake data and trend calculations

Number	x	y	z	xz	x^2	xy	yz	y^2
1	40.3	38.4	5.1	205.53	1624.09	1547.52	195.84	1474.56
2	38	27	6	228	1444	1026	162	729
3	39	39	6.3	245.7	1521	1521	245.7	1521
4	40.5	42.7	6.2	251.1	1640.25	1729.35	264.74	1823.29
5	38	26.5	6	228	1444	1007	159	702.25
6	40.18	38.1	6.75	271.215	1614.432	1530.858	257.175	1451.61
7	42.5	26.4	5.9	250.75	1806.25	1122	155.76	696.96
8	41	34	6.2	254.2	1681	1394	210.8	1156
9	36	30	6.25	225	1296	1080	187.5	900
10	40.65	27.2	7.75	315.0375	1652.423	1105.68	210.8	739.84
11	40.6	27.1	6.4	259.84	1648.36	1100.26	173.44	734.41
12	40.1	26.8	6.9	276.69	1608.01	1074.68	184.92	718.24
13	38	30	6.9	262.2	1444	1140	207	900
14	40.27	36.38	7.1	285.917	1621.673	1465.023	258.298	1323.504
15	39.26	26.71	7	274.82	1541.348	1048.635	186.97	713.4241
16	35.5	34	5.8	205.9	1260.25	1207	197.2	1156
17	39.7	42.8	5.3	210.41	1576.09	1699.16	226.84	1831.84
18	39.96	41.94	6.8	271.728	1596.802	1675.922	285.192	1758.964
19	38.55	30.78	5.9	227.445	1486.103	1186.569	181.602	947.4084
20	41.33	43.41	6	247.98	1708.169	1794.135	260.46	1884.428
21	38	30.5	5.9	224.2	1444	1159	179.95	930.25
22	37.03	29.43	6.1	225.883	1371.221	1089.793	179.523	866.1249

(continued)

Table 6.5 (continued)

Number	x	y	z	xz	x^2	xy	yz	y^2
23	35.84	29.5	7	250.88	1284.506	1057.28	206.5	870.25
24	36.54	27.33	7.3	266.742	1335.172	998.6382	199.509	746.9289
25	40.94	43.88	6	245.64	1676.084	1796.447	263.28	1925.454
26	38.1	27.1	6.5	247.65	1451.61	1032.51	176.15	734.41
27	40.5	26.5	6.1	247.05	1640.25	1073.25	161.65	702.25
28	40.2	37.9	6.1	245.22	1616.04	1523.58	231.19	1436.41
29	37.98	44.48	7.6	288.648	1442.48	1689.35	338.048	1978.47
Averages	39.12172	33.30483	6.384483	249.6336	1533.642	1306.022	211.9668	1150.113

trend surfaces suitable for the data values at hand. After the determination of model parameters the positions of local or global maxima and minima points can be obtained after simple algebraic calculations. The locations of local maximum and minimum can be obtained from Eq. (6.29) by taking partial derivative with respect to x and y, as,

$$\frac{\partial z}{\partial x} = a_1 + 2a_3x + a_4y = 0 \tag{6.30}$$

and

$$\frac{\partial z}{\partial y} = a_2 + a_4y + 2a_5y = 0 \tag{6.31}$$

The simultaneous solution of these last two expressions provides the location of the maximum (minimum) point within the polynomial trend surface.

$$x_e = \frac{a_2a_4 - 2a_1a_5}{4a_3a_5 - a_4^2} \tag{6.32}$$

and

$$y_e = \frac{a_1a_4 - 2a_2a_3}{4a_3a_5 - a_4^2} \tag{6.33}$$

This point may show trend surface highest, lowest or inflection point depending on the data features. The point to be cared for in any trend surface analysis is that there must not be model parameter numbers more than the data number. A good practical rule is that there must not be more than one third of data number model parameter.

The uncertain with error component Eq. (6.29) takes the following shape with full model parameters.

$$z = a_0 + a_1x + a_2y + a_3x^2 + a_4xy + a_5y^2 + \varepsilon \tag{6.34}$$

Similar to the case of planer trend regression analysis, it is possible to take partial derivatives with respect to each model parameter and then equate then to zero and solve simultaneously. This procedure leads to a set of equations similar to Eqs. (6.26)–(6.27).

Herein, instead of regression partial derivative procedure, the practical rule steps are applied for arriving at six equations for the estimation of six parameters in the main polynomial trend surface expression in Eq. (6.29). In the following the necessary steps are given.

(1) The main polygon trend surface expression in Eq. (6.29) leads to the first expression after taking the arithmetic average of both sides as,

$$\overline{z} = a_0 + a_1\overline{x} + a_2\overline{y} + a_3\overline{x^2} + a_4\overline{xy} + a_5\overline{y^2} \tag{6.35}$$

(2) Multiplication of both sides of the main model by the first independent variable, x, gives the second expression for the model parameter estimation as,

$$\overline{zx} = a_0\overline{x} + a_1\overline{x^2} + a_2\overline{yx} + a_3\overline{x^3} + a_4\overline{x^2y} + a_5\overline{xy^2} \tag{6.36}$$

(3) Multiply both sides of the main equation by the second independent variable, y, and take the arithmetic averages of both sides and hence obtain,

$$\overline{zy} = a_0\overline{y} + a_1\overline{xy} + a_2\overline{y^2x} + a_3\overline{yx^2} + a_4\overline{xy^2} + a_5\overline{y^3} \tag{6.37}$$

(4) The multiplication of both sides in the main model by x^2 and then the arithmetic averages of both sides give to,

$$\overline{zx^2} = a_0\overline{x^2} + a_1\overline{x^3} + a_2\overline{x^2y} + a_3\overline{x^4} + a_4\overline{x^3y} + a_5\overline{x^2y^2} \tag{6.38}$$

(5) The main equation both sides are multiplied by xy term and then the arithmetic average procedure results in,

$$\overline{xyz} = a_0\overline{xy} + a_1\overline{x^2y} + a_2\overline{xy^2} + a_3\overline{x^3y} + a_4\overline{x^2y^2} + a_5\overline{xy^3} \tag{6.39}$$

(6) Finally, similar multiplication and arithmetic Average procedures with consideration of the term y^2 yields,

$$\overline{y^2z} = a_0\overline{y^2} + a_1\overline{y^2x} + a_2\overline{y^3} + a_3\overline{y^2x^2} + a_4\overline{xy^3} + a_5\overline{y^4} \tag{6.40}$$

The collection of six simultaneous equations (Eqs. 6.35–6.40) can be shown in the form of the following matrix form.

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} = \begin{bmatrix} 1 & \overline{x} & \overline{y} & \overline{x^2} & \overline{xy} & \overline{y^2} \\ \overline{x} & \overline{x^2} & \overline{xy} & \overline{x^3} & \overline{x^2y} & \overline{xy^2} \\ \overline{y} & \overline{xy} & \overline{y^2} & \overline{x^2y} & \overline{xy^2} & \overline{y^3} \\ \overline{x^2} & \overline{x^3} & \overline{x^2y} & \overline{x^4} & \overline{x^3y} & \overline{x^2y^2} \\ \overline{xy} & \overline{x^2y} & \overline{xy^2} & \overline{x^3y} & \overline{x^2y^2} & \overline{xy^3} \\ \overline{y^2} & \overline{xy^2} & \overline{y^3} & \overline{x^2y^2} & \overline{xy^3} & \overline{y^4} \end{bmatrix}^{-1} \begin{bmatrix} \overline{z} \\ \overline{xz} \\ \overline{yz} \\ \overline{x^2z} \\ \overline{xyz} \\ \overline{y^2z} \end{bmatrix} \tag{6.41}$$

The solution for the model parameters can be obtained by finding the inverse of the coefficients matrix, which is the first term on the right hand side of this last expression. One should notice that the coefficients matrix is symmetrical around the main diagonal. It is also possible to construct the titles of calculation table (Table 6.6) similar to Table 6.4.

The arithmetic average of each column gives the elements of each matrix on the right hand side of Eqs. (6.35)–(6.41).

6.8.3 Kriging Methodology

Kriging is a mapping methodology and it can be used to get 3D representation of any spatial variable leading to surfaces as in Fig. 6.15. The 3D version of statistics is referred to geostatistics, and it treats "Regionalized Variables" (ReV), statistically through a methodology first proposed by Matheron (1963). The very first step in geostatistics is to determine the spatial dependence structure from measurements. The spatial structure also can be established by semivariogram (SV). After the spatial structure determination, it is possible to control the spatial variability and make estimations at unmeasured locations. Most of the earth sciences and environmental phenomena have temporal and spatial characteristics simultaneously. The spatiotemporal variations of the natural phenomena imply a significant amount of uncertainty, and furthermore, these variations have space heterogeneity and nonstationary.

A special variable, which includes also the regionalized variable (ReV) has characteristics of a certain phenomenon such as ore grade, rainfall, seismicity, and water level data are regionalized variables. The procedure of the theory of regionalized variables (Kriging method) follows two main steps as the establishment of the theoretical ground for expressing the structural properties of the phenomenon through the experimental semivariogram (SV) and also providence of a model that uses a combination of functions, which guarantee a solution for the estimation problem (Kriging), by using the probabilistic theory of random function (RF).

The ReV is denoted by $z(x, y)$ similar to the spatial variable like in the previous sections and in practical applications only a single of realization is available as random process. The problem is to find the characteristics of RF, $Z(x)$ to make estimations at unmeasured sites possible. IN such an approach statistical inference requires further hypothesis of stationarity (David 1977; Journel and Huijbregts 1978; Isaaks and Srivastava 1989).

If two points namely (x_1, y_1) and (x_2, y_2) are separated by a distance, d, with a given relative orientation, the distribution of ReV (spatial variable) differences depends on d, which is also valid for the mean and variance. Hence, if in case of the mean of difference values from the sample is (m), and the expectation for the whole phenomenon is regarded as a constant independent of the site location. The random function is said to be second-order stationary when the first two statistical moments exist and do not depend on x,

$$E\{Z(x)\} = m(x) = m \tag{6.42}$$

$$C(0) = Var\{Z(x)\} = E\left\{[Z(x) - m(x)]^2\right\} \tag{6.43}$$

$$C(d) = E\{[Z(x_1) - m(x_1)][Z(x_2) - m(x_2)]\}, \tag{6.44}$$

Table 6.6 Trend surface calculation table

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
				Coordinates							Data	Regression coefficient calculation columns							
x	y	z	x^2	xz	xy	yz	y^2	x^3	x^2y	x^2z	xyz	y^2z	xy^2	y^3	x^4	x^3y	x^2y^2	xy^3	y^4
x_1	y_1	z_1	x_1^2	x_1z_1	x_1y_1	y_1z_1	y_1^2	x_1^3	$x_1^2y_1$	x_1^2z	$x_1y_1z_1$	$y_1^2z_1$	$x_1y_1^2$	y_1^3	x_1^4	$x_1^3y_1$	$x_1^2y_1^2$	$x_1y_1^3$	y_1^4
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
x_n	y_n	z_n	x_n^2	x_nz_n	x_ny_n	y_nz_n	y_n^2	x_n^3	$x_n^2y_n$	$x_n^2z_n$	$x_ny_nz_n$	$y_n^2z_n$	$x_ny_n^2$	y_n^3	x_n^4	$x_n^3y_n$	$x_n^2y_n^2$	$x_ny_n^3$	y_n^4
$\overline{x}$	$\overline{y}$	$\overline{z}$	$\overline{x^2}$	$\overline{xz}$	$\overline{xy}$	$\overline{yz}$	$\overline{y^2}$	$\overline{x^3}$	$\overline{x^2y}$	$\overline{x^2z}$	$\overline{xyz}$	$\overline{y^2z}$	$\overline{xy^2}$	$\overline{x}$	$\overline{x^4}$	$\overline{x^3y}$	$\overline{x^2y^2}$	$\overline{xy^3}$	$\overline{y^4}$

where $C(0)$ is the variance at zero distance ($d = 0$) and independent of x; $C(d)$ is the covariance which is a function of two locations x_1 and x_2 and depends only upon their difference ($x_1 - x_2$). If RF $Z(x)$ is ergodic meaning that one field observation sequence is enough for calculations, then the future inference about parameters of $Z(x)$ depends upon the first stage. Hypothetically, for any distance, d, the increment $Z(x + d) - Z(x)$ has zero expectation and the variance is independent of position. These are expressed as,

$$E\{Z(x+d) - Z(x)\} = 0 \tag{6.45}$$

and

$$\mathrm{Var}\{Z(x+d) - Z(x)\} = 2g(d). \tag{6.46}$$

This is the intrinsic hypothesis or quasi-stationary, where the function of $2g(h)$, is used in geostatistics to describe and summarize underlying spatial dependence structure, and it is named as the variogram. Consider Eq. (6.44) and the last expression can be rewritten as,

$$g(d) = \frac{1}{2}\mathrm{Var}\{Z(x+d) - Z(x)\}, \tag{6.47}$$

where now $g(d)$ is called the semivariogram (SV).

Under the hypothesis of second order stationarity, the statistical covariance and variogram are two equivalent tools for the autocorrelation between two variables $Z(x + d)$ and $Z(x)$ separated by d as,

$$g(d) = C(0) - C(d). \tag{6.48}$$

The SV estimation is preferable to estimation of the covariance, because the experimental SV does not require a prior estimate of the population mean. Under the same condition the relationship between the model covariance, variogram and correlogram $\gamma(d)$ is,

$$\gamma(d) = \frac{C(d)}{C(0)} = 1 - \frac{g(d)}{C(d)} \tag{6.49}$$

The semivariogram (SV) is a graph or formula describing the expected difference in values between pairs of samples with given relative orientation. The SV is a procedure for characterizing the structures of spatial continuity of the geological phenomena. The difference in values should be consistent. The consistency is referred to as quasi-stationarity or intrinsic hypothesis. The practical form of SV is

$$\gamma(d) = \frac{1}{2n(d)} \sum_{i=1}^{n(d)} [Z(x+d) - Z(d)]^2, \tag{6.50}$$

where $n(d)$ is a number of pairs having distance, d.

A detailed explanation of various versions of the Kriging methodologies are explained by Şen (2008), but here only the simplest methodology is presented.

6.8.3.1 Simple Kriging (SK)

In general, the basic assumptions in this type of Kriging procedure are as follows, and in the formulation development each one must be taken into consideration carefully. The spatial sampling points are representatives of the ReV at a set of given locations with measurement values,

(1) The measurements at each site are representatives of the spatial variable, which is known in the Kriging terminology as ReV,
(2) The spatial variability is assumed to have three necessary quantitative values, which are the spatial arithmetic average, variance and the semivariogram, SV graph, which is the scatter of distance values against the squared- differences as in Fig. 6.17,
(3) The arithmetic average of the spatial variable (ReV) is known, which limits the application of this Kriging modeling alternative severely.

There are many cases in the practical applications, where the areal mean of the ReV is known and in such a case the application of the Kriging methodology is simple and attractive (Şen 2008). In case of arithmetic average and variance constancy the spatial variable has second-order stationarity property, and hence, the measurements can be standardized according to classical statistical standardization formulation (see Chap. 2, Eq. 2.26) leading to standardized ReV with zero regional mean and unit variance. Figure 6.24 shows a set of irregularly distributed measurement locations, where there are n measurement and one estimation sites.

In this figure, the set of measurements are (Z_1, Z_2, …, Z_n) and Z_E is the site, where the spatial variable estimation is sought. In the Kriging methodology, the spatial variability is quantified by a suitable regional dependence function, which is classically the semivariogram, SV, function (similar to the covariance function in the classical time series analysis), which presents the relationship among each pair of measurements by taking into consideration the squared differences as in Sect. 6.6.1.

Methodology

In general, the Kriging estimation is equivalent to the weighted average of the measurements with distant dependent weighting values, λ_i. In order to obtain the Kriging estimation, Z_E, a linear weighted average can be writes based on Z_i (i = 1, 2, …, n) measurements as follows.

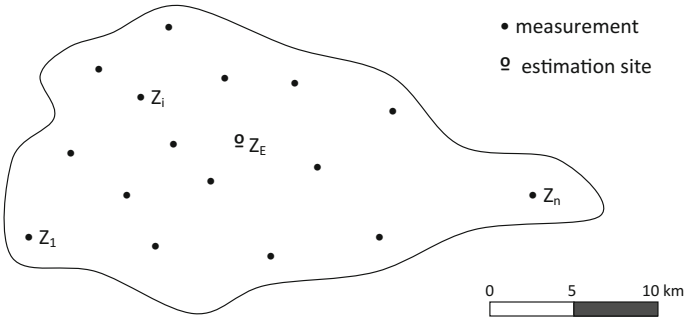


Fig. 6.24 Spatial variable sample sites and estimation site

$$Z_E = \overline{Z} + \sum_{i=1}^{n} \lambda_i \left(Z_i - \overline{Z}\right). \tag{6.51}$$

Herein, $\overline{Z}$ is the regional mean value of the ReV. If there are n neighboring sites then in the covariance matrix there will be $b(n-1)/2$ symmetrical off diagonal correlation implication values with n variances on the main diagonal.

$$C = \begin{bmatrix} \mathrm{var}(z_1) & \mathrm{cov}(z_1, z_2) & \cdots & \mathrm{cov}(z_1, z_n) \\ \mathrm{cov}(z_2, z_1) & \mathrm{var}(z_2) & \cdots & \mathrm{cov}(z_2, z_n) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mathrm{cov}(z_n, z_1) & \mathrm{cov}(z_n, z_2) & \cdots & \mathrm{var}(z_n) \end{bmatrix}. \tag{6.52}$$

In this matrix $\mathrm{cov}(z_i, z_j) = \mathrm{cov}(z_j, z_i)$ and along the main diagonal $\mathrm{var}(z_i) = \sigma_i^2$ which is the variance at site i. Each element in the matrix is dependent on the distances (relative distance) among each two sites. Since, the spatial variance is assumed as constant, the division of each element in Eq. (6.52) leads to unit values along the main diagonal with autocorrelation values, $\rho(z_i, z_j) = \mathrm{cov}(z_i, z_j)$, at off diagonal locations between each pair of sites, $i, j = 1, 2, \ldots, n$, as,

$$\rho = \begin{bmatrix} 1 & \rho(z_1, z_2) & \cdots & \rho(z_1, z_n) \\ \rho(z_2, z_1) & 1 & \cdots & \rho(z_2, z_n) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots 1 & \cdot \\ \rho(z_n, z_1) & \rho(z_n, z_2) & \cdots & 1 \end{bmatrix}. \tag{6.53}$$

In statistics this corresponds to the spatial correlation matrix for the spatial variable. Parallel to the correlation matrix another one can be defined as distance matrix, D, between the same sites with zero distances along the main diagonal as follows.

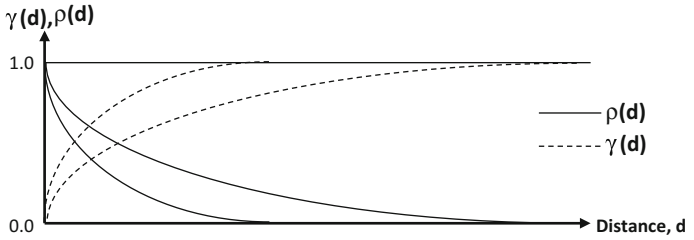


Fig. 6.25 Covariance–distance graphs

$$D = \begin{bmatrix} 0 & \text{dis}(z_1, z_2) & \cdots & \text{dis}(z_1, z_n) \\ \text{dis}(z_2, z_1) & 0 & \cdots & \text{dis}(z_2, z_n) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & 0 & \cdot \\ \text{dis}(z_n, z_1) & \text{dis}(z_n, z_2) & \cdots & 0 \end{bmatrix}. \tag{6.54}$$

The plot of distance matrix values, d, on the horizontal axis versus corresponding autocorrelation coefficients, $\rho(d)$, from the correlation matrix leads to one of the curves as in Fig. 6.25, which is very similar to the RDF as already shown in Fig. 6.16. Again, as the distance increases the correlation coefficient function decreases, and therefore, Fig. 5.15 has a decreasing trend with distance and theoretically this function should be asymptotic to the horizontal axis. Consideration of Eq. (5.31) with unit variance yields the corresponding SVs in the same figure.

It is possible to rewrite Eq. (6.51) for the simple Kriging with consideration of standardization as,

$$z_E = \sum_{i=1}^{n} \lambda_i z_i. \tag{6.55}$$

In this expression, there are n unknowns and their solutions require the same number of simultaneous equation solutions. If both sides of Eq. (6.55) are multiplied by each measurement then by taking the averages (expectations) as a set of equations can be obtained as follows.

$$\begin{aligned} &\sum_{i=1}^{n} \lambda_i \rho(z_i, z_1) = \rho(z_E, z_1) \\ &\sum_{i=1}^{n} \lambda_i \rho(z_i, z_2) = \rho(z_E, z_2) \\ &\cdots \\ &\sum_{i=1}^{n} \lambda_i \rho(z_i, z_k) = \rho(z_E, z_k). \end{aligned} \tag{6.56}$$

For the solution of this set of equations it is convenient to write it down in the form of matrices and vector forms. Hence the unknown column vector is,

$$\Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ . \\ . \\ \lambda_n \end{bmatrix} \tag{6.57}$$

The right hand side of Eq. (6.56) represents the known part, say, column vector, B, which has the following elements

$$B = \begin{bmatrix} \rho(z_E, z_1) \\ \rho(z_E, z_2) \\ . \\ . \\ \rho(z_E, z_n) \end{bmatrix} \tag{6.58}$$

Consideration of these two vectors with the set of simultaneous expressions in Eq. (6.56) provides succinctly that,

$$C\Lambda = B$$

or after the inversion operation its implicit form becomes as,

$$\Lambda = C^{-1}B \tag{6.59}$$

This last expression is the implicit solution of the weighting factors, λ_i, the estimation of the standard ReV can be converted to actual (nonstandardized) ReV as,

$$Z_E = \overline{Z} + \sigma_E^2 z_E \tag{6.60}$$

where z_E is an (n $\times$ 1) matrix of the measurements with zero mean and unit variance.

All of the aforementioned derivations are based on the covariance function, $\rho(d)$, which is thought as the representative of the RDF. Şen (2008) has explained that there is a relationship between the covariance and SV functions, $\gamma(d)$, in the case of standardized ReV as,

$$\rho(d) = 1 - \gamma(d) \tag{6.61}$$

The SV functions are shown in Fig. 6.25 as complementary to the autocorrelation function and it is in ascending order with the distance. The estimation variance of covariance can be expressed as,

$$\sigma_E^2 = 1 - B'\Lambda \tag{6.62}$$

In Case of SV use for the spatial modeling, the same estimation variance becomes,

$$\sigma_E^2 = B'\Lambda \tag{6.63}$$

The simple Kriging depends on the statistical property of the covariance (or SV) function and the spatial estimation is achieved in such a way that the RDF of the ReV is preserved throughout the procedure. Unfortunately, neither the cross-validation nor the unbiasedness procedures are applicable explicitly in the simple Kriging procedure.

Example 6.2 If the same earthquake data in Table 6.2 is plotted in Fig. 6.26 with the spatial variable estimation location as, Z_E. How can one make spatial estimation by means of the simple Kriging methodology?

The distance matrix, D, between each pair of data has already been given in the matrix form in Sect. 6.6.1 with the squared differences matrix. In the following the distances between each measurement site and the estimation location, Z_E, spatial variable is calculated.

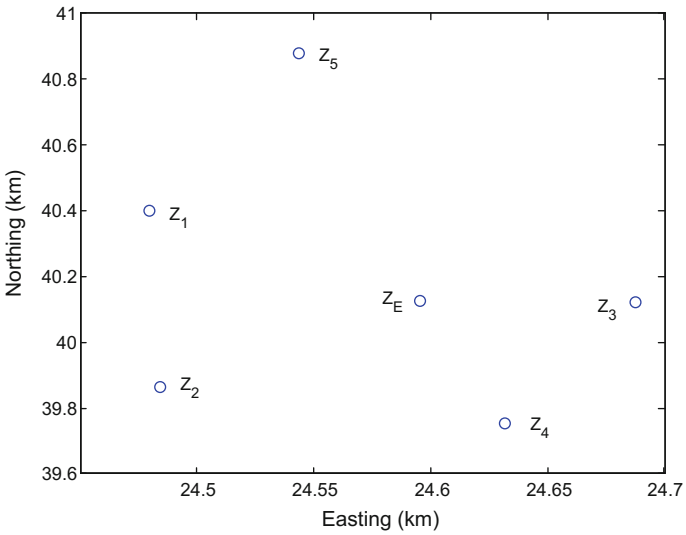


Fig. 6.26 Spatial scatter of data locations

	1	2	3	4	5
E	0.297151	0.284038	0.092392	0.374746	0.754616

After the standardization of the spatial variable the corresponding half-squared differences (i.e., SV) matrix can be obtained by substituting Eq. (6.61) with $\sigma_Z^2 = 1$ into Eq. (5.48), which leads to,

$$\Gamma = \begin{bmatrix} 0 & 1-\gamma(z_1,z_2) & \cdots & 1-\gamma(z_1,z_n) \\ 1-\gamma(z_2,z_1) & 0 & \cdots & 1-\gamma(z_2,z_n) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & 0 & \cdot \\ 1-\gamma(z_n,z_1) & 1-\gamma(z_n,z_2) & \cdots & 0 \end{bmatrix}. \tag{6.64}$$

In practical works, there are two different SV calculation ways as either from a given small sample as in Table 6.2 without knowing the basic structure of the sample SV or after defining the structural form of the sample SV from a large number of data, which is preferred to be more than 30 data values. The former approach yields to a matrix that represents half-squared differences subtraction from 1 according to Eq. (6.64) between the earthquake magnitudes.

$$\Gamma = \begin{bmatrix} 0 & & & \\ 1-0.00125 & 0 & & \\ 1-0.05445 & 1-0.03920 & 0 & \\ 1-0.03920 & 1-0.02645 & 1-0.00125 & 0 \\ 1-0.02000 & 1-0.01125 & 1-0.00845 & 1-0.00320 \end{bmatrix}.$$

Likewise, Eq. (6.58) can be written by considering Eq. (6.61) as,

$$B = \begin{bmatrix} 1-\gamma(z_E,z_1) \\ 1-\gamma(z_E,z_2) \\ \cdot \\ \cdot \\ 1-\gamma(z_E,z_n) \end{bmatrix}. \tag{6.65}$$

Since the spatial variable value at the estimation site is not known, it is not possible to estimate the SV values in this vector. However, instead one can use the global SV that would depend on many location records and it is assumed herein that from a priory structural analysis the sample SV as a linear model,

$$\gamma(d) = 0.015 + 0.1d \tag{6.66}$$

It is now possible to calculate the SV value from the distances between the estimation point and other surrounding points in Fig. 6.26 leading to,

$$B = \begin{bmatrix} 1-0.0447151 \\ 1-0.0434038 \\ 1-0.0242392 \\ 1-0.0524746 \\ 1-0.0904616 \end{bmatrix}$$

With this information at hand, one can calculate the matrix in Eq. (6.64) according to the distance matrix by using the SV expression in Eq. (6.66), which leads to

$$\Gamma = \begin{bmatrix} 0 & & & \\ 1-0.0684689 & 0 & & \\ 1-0.0499005 & 1-0.0476195 & 0 & \\ 1-0.0814208 & 1-0.0335368 & 1-0.0520756 & 0 \\ 1-0.0633628 & 1-0.1165764 & 1-0.0922863 & 1-0.279264 \end{bmatrix}.$$

The final solution can be found by taking inverse of Γ, which appears as follows.

$$\Gamma^{-1} = \begin{bmatrix} -0.7747 & 0.3017 & 0.2867 & 0.2763 & 0.2058 \\ 0.3017 & -0.7773 & 0.2822 & 0.2041 & 0.2866 \\ 0.2867 & 0.2822 & -0.7600 & 0.2391 & 0.2558 \\ 0.2763 & 0.2041 & 0.2391 & -0.9102 & 0.4471 \\ 0.2058 & 0.2866 & 0.2558 & 0.4471 & -0.9829 \end{bmatrix}.$$

Hence, the application of Eq. (6.59) leads to the final weights as,

$$\Lambda = \begin{bmatrix} 0.2801 \\ 0.2668 \\ 0.2641 \\ 0.2387 \\ 0.2526 \end{bmatrix}$$

It is now possible to calculate the prediction value from Eq. (6.60), which gives $Z_E = 2.65$.

6.9 Triple Diagram Model (TDM)

Human beings can visualize three-dimensional variations at the maximum and the best configuration of such variations can be achieved in three-dimensional Cartesian coordinate system. In the previous sections spatial trend components are searched inside the geographically coordinated sample points and their spatial variability measurements. The same methodologies can be employed provided there are simultaneous measurements of three spatial variables. It is possible to look for the

spatial trend and variation characteristics of each spatial variable based on the measurements' geographic locations. In this section instead of the geographic locations the values of the two spatial variable measurements are adapted as the geographical location values (x and y) and the third one is the spatial variability, z. In this manner, it is possible to look for the special variation of one of the variables based on the other two. Provided that there are three simultaneously measured variables there will be three different types of maps each of which is referred to as the triple diagram model by Şen (2008). TDM is equivalent with the contour line map and it can be obtained by one of the spatial trend methodology. In geological sciences, Davis (1986) suggested the application of various simple regional techniques such as inverse distance, inverse distance square, etc. for mapping, herein, preparation of the TDM is based on classical Kriging technique. In this section Kriging methodology is used for three-dimensional representation of the spatial variable. In the construction of TDM three variables are necessary two of which are referred to as independent variables and they constitute the basic scatter similar to Fig. 6.17. The third is the dependent variable, which has its measured value at each scatter point. The equal value lines are constructed by the Kriging methodology concept as suggested by Matheron (1965), Kriging methodology is also referred to as geostatistics. Details of this methodology are explained for earth sciences applications by Journal and Huijbregts (1978).

The construction of a TDM requires three variables two of which are referred to as independent variables (predictors) and they constitute the basic scatter diagram. The third is the dependent variable, which has its measured values attached to each scatter point. The equal value lines are constructed by the Kriging methodology concepts explained in earlier sections of this chapter.

6.9.1 Parallel-Triple Model

The geostatistical methods take into consideration the effective role of the measured values of a regional variable at a set of irregular sites (or scatter points) the advantages and disadvantages of these methods are discussed by various authors (Matheron 1963; Journel and Huijbregts 1978). The dependent variable in the triple diagram can be considered as regionalized variable and as a random field with data values recorded at scatter points of dependent variable.

The water hydrochemistry records are used for the implementation of the Kriging methodology so as to obtain triple diagrams that give the common behavior of three variables. Herein, three distinctive but complementary investigations are considered. These are,

(1) Triple diagrams are constructed directly from major anions and cations, so as to consider three major anions and/or cations common behaviors within the study area. For instance, the triple diagram of the equal NO_3 concentrations based on Cl and HCO_3 values is presented in Fig. 6.27.

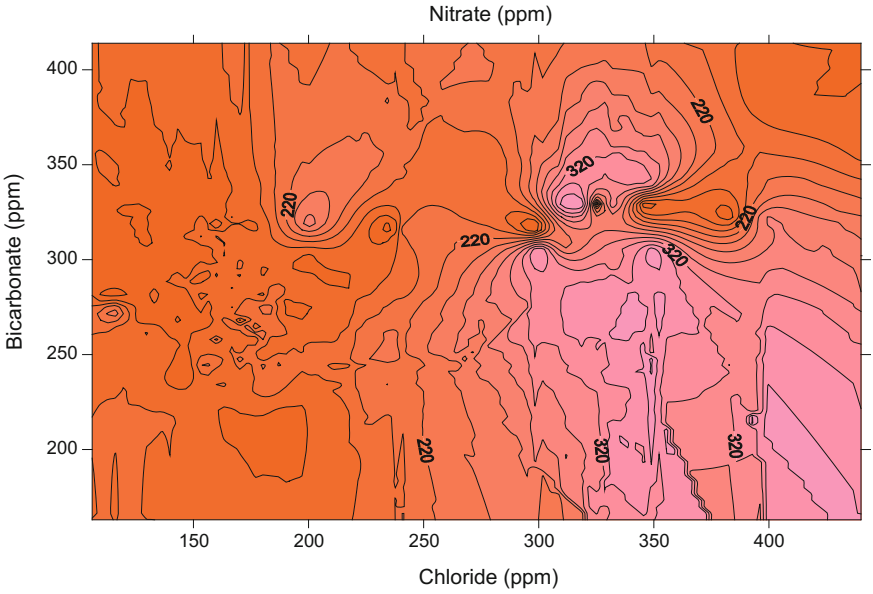


Fig. 6.27 Equal NO_3 lines based on Cl and HCO_3

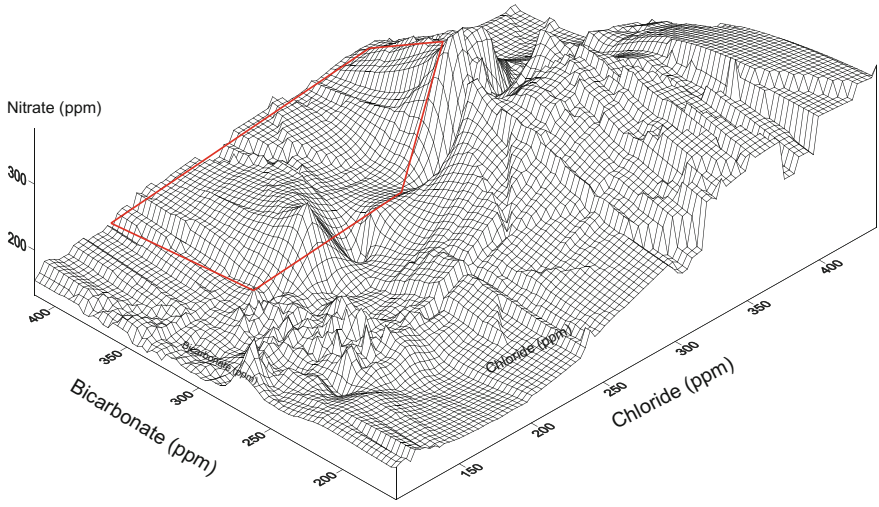


Fig. 6.28 Three-dimensional NO_3 change with Cl and HCO_3

It is also helpful to look at the three-dimensional (3D) surface relationship between these chemical constituents as in Fig. 6.28.

The interpretation of the triple diagram map and 3D surface leads to the following logical inferences concerning high NO_3 concentration rates based on Cl and HCO_3

IF Cl is **medium**	AND HCO_3 is **low**	THEN NO_3 is **high**	OR
IF Cl is **medium**	AND HCO_3 is **medium**	THEN NO_3 is **very high**	OR
IF Cl is **high**	AND HCO_3 is **low**	THEN NO_3 is **high**.	

These logical statements lead hydrochemist to think about the possibilities of each IF-THEN rule on the basis of geological subsurface composition of the study area, in addition to the hydrological and hydrogeological features interactions. In this manner, it is possible to obtain clues for reasons of groundwater quality variations. On the other hand, these logical statements provide a common basis for the general variability description of ions within the study area. Such rule bases are prerequisites for fuzzy logic modeling as suggested by Zadeh (1965).

(2) Similar triple diagrams can also be obtained among the milliequivalent percentages of the anions and cations as they are used in the construction of the classical trilinear diagrams. This approach brings a restriction as the summation of the percentages is equal to 100%, which is the basis of the classical trilinear diagrams, (Piper 1953). If, for instance, the percentages of (SO_4 + HCO_3), Cl and NO_3 [or Ca, Mg and (Na + K)] are α, β and γ, respectively, then by definition $\alpha + \beta + \gamma = 1.0$, which implies that $0 < \alpha, \beta, \gamma < 1$. It is obvious that this expression gives points on the equilateral inclined triangular surface that is shown in Fig. 6.29. In fact, this triangle is identical with the Piper diagram basic ionic triangles.

(3) However, in this paper, the conventional equilateral triangle representation of ions is considered similar to diagram in the first step, but with percentages. Figures 6.30 and 6.31 show the percentage change of NO_3 with of Cl and HCO_3 percentages in two and 3D maps, respectively.

As already mentioned right in Chap. 1 that visual inspections are very important preliminary advices to deduce possible trend components in a given series and herein in 3D map one can identify two triangular shaped plains.

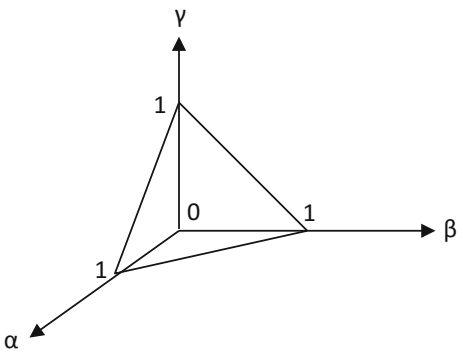


Fig. 6.29 Equilateral inclined triangular surface

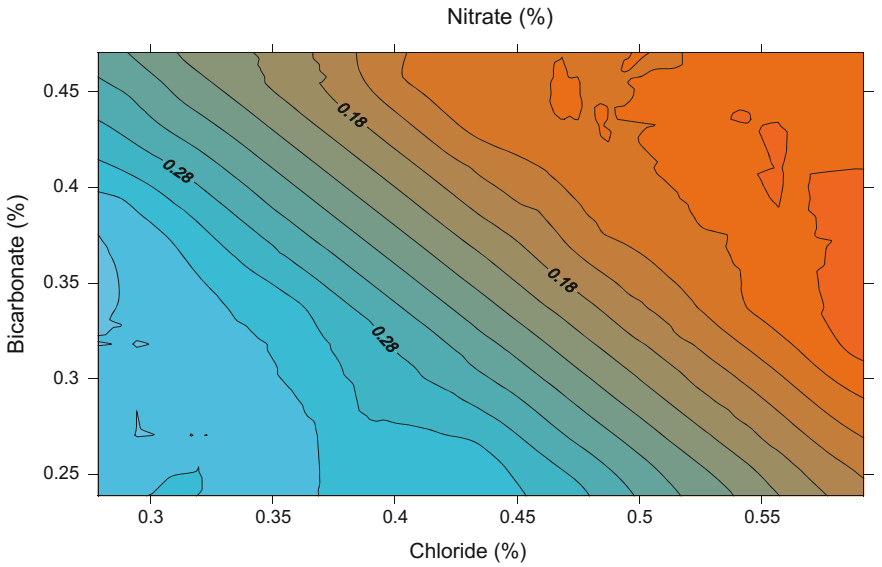


Fig. 6.30 Equal percentage NO_3 lines based on percentages of Cl and HCO_3

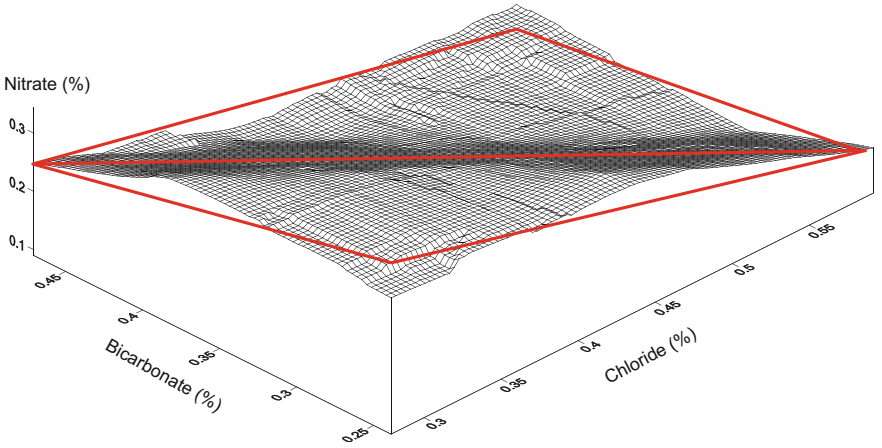


Fig. 6.31 Three-dimensional percentage NO_3 change with Cl and HCO_3

(3) It is also helpful to construct triple diagrams and 3D surfaces in terms of any ion representing dependent variable with two independent composite variables electric conductivity (EC), total dissolved solids (TDS) or ph.

The application of these three steps yields to a bundle of triple diagrams that can be interpreted leading to common logical and scientific statements about the NO_3 changes with respect to two other ions.

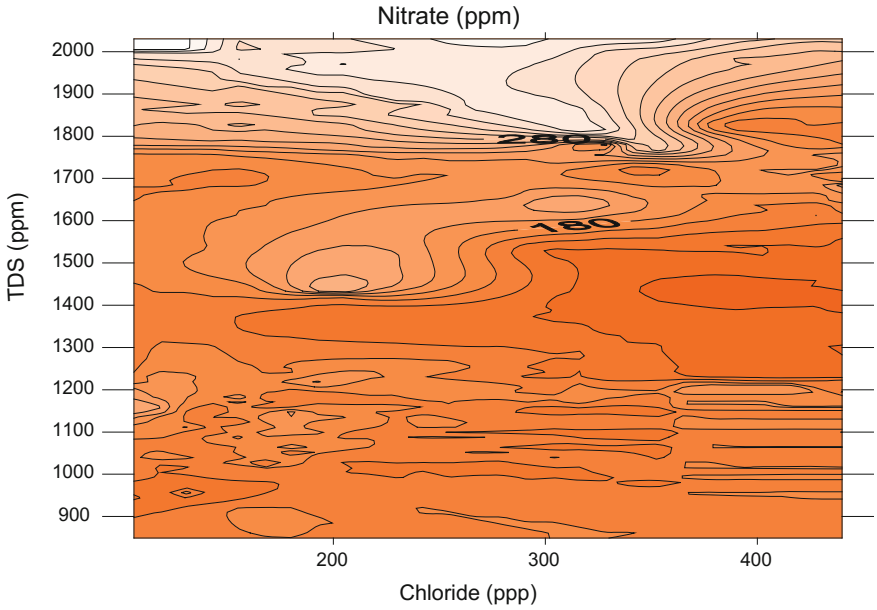


Fig. 6.32 Equal NO_3 lines based on Cl and TDS

Figures 6.32 and 6.33 indicate NO_3 concentration changes with Cl and TD as triple diagrams and 3D map, respectively.

Again in this figure visual inspection helps to identify approximately three trend plains in sequence and each one can be quantified by the application of the trend surface methodology explained in Sect. 6.8.

6.9.2 Serial-Triple Model

In this case three series are generated from a given time series at different lags. The first one is the series itself as dependent variable, but the other two may be at different lags. The simplest one is lag-one apart two others as independent variables. If the given time series is $X^{(i)} = \{X_1, X_2, \ldots, X_i\}$ $(i = 1, 2, \ldots, n)$, where n is the sample length the two other serial time series are defined as $X^{(i-1)} = \{X_2, X_3, \ldots, X_i\}$ $(i = 2, \ldots, n)$ and $X^{(i-2)} = \{X_3, X_4, \ldots, X_i\}$ $(i = 3, \ldots, n)$. Hence, three serial time series are $X^{(i)}$, $X^{(i-1)}$ and $X^{(i-2)}$, which again similar to the parallel-triple model provides a basis for 3D and 2D maps. The first two variables represent the two past lake levels and third one indicates the present lake levels. Hence, the model has three parts, namely, observations (recorded time series) as input, triple diagram as response, and the output as prediction.

Herein, an example given by Şen (2009) is explained for the serial-triple model, where the Lake Van water level measurements time series are taken into

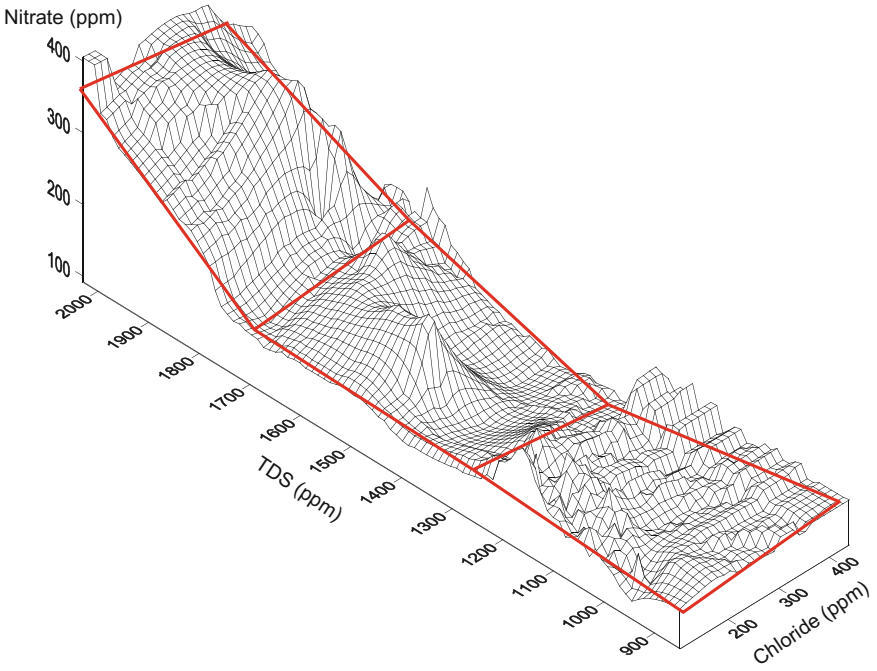


Fig. 6.33 Three-dimensional NO_3 change with Cl and TDS

consideration. This lake's level modelling has already been presented in Chap. 4 in the Sect. 4.8.2 Cluster regression model. Herein, the model is thought such that to be able to predict the lake level from two previous record lengths, which means implicitly that $X^{(i)} = f[X^{(i-1)}, X^{(i-2)}]$. This expression implies to a surface and therefore its first and simplest consideration may be in the form of a regression expression as,

$$X^{(i)} = aX^{(i-1)} + bX^{(i-2)} + e_i \tag{6.67}$$

where e_i indicates the spatial error terms, which are deviations from the plane surface between the three serial-triple time series; α and b are model parameters. It is also possible to search for curvature surfaces by taking into consideration any one of the nonlinear models between the three variables as already explained in Sect. 6.5.5. The parameter estimations can be obtained according to the procedure in Sect. 6.8 under the light of assumptions that linearity, normality (Gaussian distribution of the residuals, i.e., e_i's), variance constancy (homoscedasticity), ergodicity and independence of residuals. The triple diagram replaces Eq. (6.67) without any restriction in the form of map. Such a map presents the appearance of natural relationship between three consecutive time values of the same variable. The first three columns in Table 6.7 present the serial-triple time series for the lake

Table 6.7 Lag-one lake level prediction (cm)

Lake level elevations (m)			Prediction (m)	Relative error (%)
$X^{(i-2)}$	$X^{(i-1)}$	$X^{(i)}$		
119	112	110	109.32	0.62
107	111	114	118.40	3.72
125	130	125	134.87	7.32
130	125	118	128.43	8.12
125	118	105	118.60	11.47
120	138	142	145.50	2.41
138	142	138	139.53	1.10
142	138	131	134.16	2.35
138	131	117	131.64	11.12
127	141	151	144.08	4.58
141	151	151	146.85	2.75
151	151	144	144.46	0.32
137	140	144	138.76	3.64
140	144	159	140.22	11.81
144	159	182	157.03	13.72
199	202	195	203.62	4.23
202	195	185	191.33	3.31
195	185	177	182.29	2.90
189	193	202	202.01	0.00
193	202	221	209.94	5.00
202	221	245	229.48	6.34
262	254	244	252.90	3.52
254	244	239	243.16	1.71
244	239	241	231.89	3.78
			Average	4.83

level elevations in m. The fourth column provides the predicted lake level elevations with relative error in the last column. Although individual errors are slightly greater than 10%, but the overall prediction relative error percentage is about 4.83%, which is less than practically acceptable limit of 5%.

The 3D map of the lake level variation based on the two previous year records is presented in Fig. 6.34.

Figure 6.35 indicates the observed and predicted H_i values. It is obvious that they follow each other very closely and on the average observed and predicted lake level series have almost the same statistical parameters.

The serial-triple model map depicts the increasing trend and during the prediction procedure there is no special treatment of trend, but even so it is modeled successfully. However, in any stochastic or statistical modeling, it is first necessary to make trend analysis and separate it from the original data. In order to further

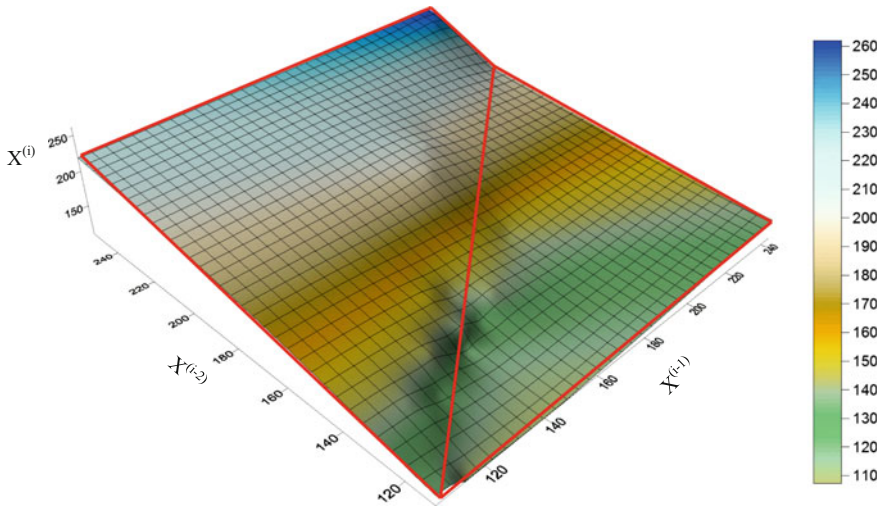


Fig. 6.34 Lake level serial-triple map

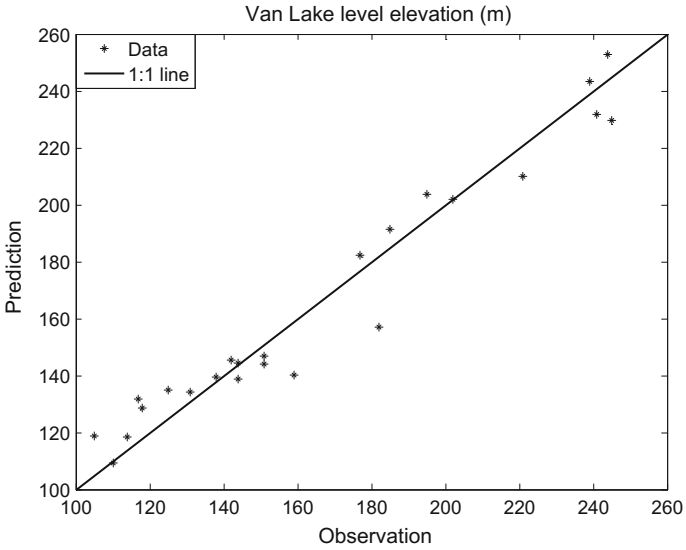


Fig. 6.35 Observed and predicted lake levels

show the verification of the serial-triple diagram approach for lake level predictions, in Fig. 6.35 the test data are plotted versus the predictions. It is obvious that almost all the points are around 45° lines and hence the model is not biased. Predictions are more successful at low or high values.

References

Aboufirassi, M., & Marino, M. A. (1984). A geostatistically based approach to the identification of aquifer transmissivities in Yolo Basin, California. *Mathematical Geology, 16*(26), 125–137.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Carr, J. R., Bailey, R. E., & Deng, E. D. (1985). Use of indicator variograms for enhanced spatial analysis. *Mathematical Geology, 17*(8), 797–812.

Clark, I. (1979). The semivariogram—Part 1. *Engineering Minning Journal, 180*(7), 90–94.

Cooley, R. L. (1979). A method of estimating parameters and assessing reliability for models of steady state groundwater flow, 2, Applications of statistical analysis. *Water Resource Research, 15,* 603–617.

David, M. (1977). *Geostatistical ore reserve estimation* (340 pp). New York: Elsevier.

Davis, J. C. (1986). Statistics and data analysis in geology. John-Wiley and Sons, New York.

Daley, R. (*1991*). *Atmospheric data analysis* (457 pp). Cambridge, U.K.: Cambridge University Press.

Eddy, A. (1967). The statistical objective analysis of scalar data fields. *Journal of Applied Meteorology, 4,* 597–609.

Gandin, L. S. (1963). *Objective analysis of meteorological fields*. Leningrad, Gidromet.: Jerusalem.

Hoeksema, R. J., & Kitandis, P. K. (1984). An application of the geostatistical approach to the inverse problem in two dimensional groundwater modeling. *Water Resources Research, 20*(7), 1003–1020.

Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to applied geostatistics* (p. 561). New York: Oxford University Press.

Journel, A. J. (1985). The deterministic side of geostatistics. *Mathematical Geology, 17*(1), 1–15.

Journel, A. G., & Huijbregts, C. I. (1978). *Mining geostatistics*. London: Academic Press. 710 pp.

Krige, D. G. (1982). Geostatistical case studies of the advantages of log-normal, De Wijsian Kriging with mean for a base metal mine and a gold mine. *Mathematical Geology, 14*(6), 547–555.

Krugen, H. B. (1964). A statistical-dynamic objective analysis scheme. Canadian Meteorological Memories, No. 18, Meteorological Branch, Department of Transport, Toronto, also supplements No. 1 (1967, CMM No. 23) and No. 2 (1969, CMM No. 27).

Krugen, H. B. (1969). General and special approaches to the problem of objective analysis of meteorological variables. *Quarterly Journal of the Royal Meteorological Society, 95,* 21–39.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology, 58,* 1246–1266.

Matheron, G. (1965). *Les Variables Regionalisees et leur Estimation*. Paris: Masson. 306p.

Matheron, G., (1969). Le Krigeage universel. les cahiers du centre de morphologie math´ematique defontainebleau, fascicule 1. Fontainebleau: ´Ecole Nationale Sup´erieure des Mines de Paris.

Myers, D. E., Begovich, C. L., Butz, T. R., & Kane, V. E. (1982). Variogram models for regional groundwater geochemical data. *Mathematical Geology, 14*(6), 629–644.

North, G. R., et al. (1982). Sampling errors in the estimation of empirical orthogonal functions. *Monthly Weather Review, 110*(7), 699–706.

Piper, A. M. (1953). A graphic procedure in the chemical interpretation of water analysis. US Geological Survey Groundwater (No. 12). note.

Şen, Z. (2008). *Spatial modeling in earth sciences* (351 pp). Springer.

Şen, Z. (2009). Spatial modeling principles in earth sciences. Springer, New York, 351 p.

Thiebaux, H. J., Pedder, M. A. (1987). *Spatial objective analysis*, *in with applications in atmospheric science* (299 pp). London: Elsevier.

Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8, 338–353.

Trend Variability Detection

7

Abstract

Variability is the most important feature that has been ignored in almost all the trend determination studies, because the researchers are interested on the average, whether there trend existence. However, in many natural and artificial time series there are variations along the time axis in the variance or better in the standard deviation. Unfortunately, in many application even unconsciously the time series is assumed as having constant standard deviation (homoscadasdicity) property. This chapter presents the available and simple variation measures and then presents a systematic methodology in an innovative manner how to determine the variability in the standard deviation.

Keywords

Homoscadasdicity · Interquartile range · Range · Simulation · Standard deviation · Variability

7.1 General

In natural, economic, social, and astronomical events there are temporal and spatial systematic variations and variabilities. Variations may be systematic or random as have been explained in the previous chapters in terms of averages such as trends, steps (jumps), and trends of linear or non-linear types. Variability is a specification of a given data sequence, especially, in terms of a time series in which there are changes not on the usual linear averages but in the standard deviations that are related to the average of the square deviations from the arithmetic average, i.e., non-linear changes. The simplest form of variation is defined as the homoscedascity, which implies variance, and hence, standard deviation change

Z. Şen, *Innovative Trend Methodologies in Science and Engineering*,
DOI 10.1007/978-3-319-52338-5_7

within a given time series. The patterns caused by the temporal and spatial variability of a phenomenon occur at many different forms depending on the causative factors. For example, as the instrument or a machine gets older it starts to give not only systematically increasing or decreasing tendency in its quantitative response, but also unusual deviations from the general performance. In order to understand the mechanism(s) that may cause to such unusual performances it is necessary to deal with such situations through well-established scientific methodologies.

In general, spatial variability came across in many practical works than the temporal type. Such variabilities reflect the regional behavior of the spatial phenomenon concerned and the predictability of the event can be achieved on the basis of meaningful and deductive pre-modelling interpretations of the data and reliable predictions after the construction and use of a suitable model. Earth systems have considerable temporal and spatial tendencies and variabilities. The variation is brought about by differences in the type and scale of development in event producing processes and also influenced strongly by local or regional factors such as topographic elevations and atmospheric conditions (Wilson and Atwater 1972).

Variability reflects differences within the internal structure of an event into the measurements and it gives rise to various algorithms that are currently in use in many disciplines. Variability is a general term that is used for the comparison of multitude of points but the difference can be quantified between two points only. For instance, the global numerical values cannot explain the internal or external features of the concerned event except its scale, but the comparison of any two leads to additional and detailed information such as the difference (with dimension of the numerical values) and unit difference (slope). There are many categories of variability such as geometric, kinematics, and dynamic types, all of which are embedded into the measurement sequence and must be identified from a set of records.

On the other hand, similarity is another word that is useful in distinguishing any variability between two cases of the same phenomenon. Similarity is a type of difference and consequent variability measure between two events. The consequence of variability may be interpreted on the basis of statistical calculations as two events are different or similar within practically acceptable error limits as ± 5 or $\pm 10\%$.

Variability may appear in forms of regular shapes or functions that can be described deterministically by mathematical and classical physical rules but irregular variabilities with their uncertainty components need to probability, statistics, and stochastic techniques. Uncertain variations are random variables in temporal sequences but regionalized variables in spatial events. Whatever the degree of uncertainty, random or regionalized variables may include regular variations and tendencies (trends, steps, seasonality). Quantitative variability is possible by pair-wise comparison of two events and the subtraction operation is the only one among four arithmetical operations that represents the variability.

This chapter concentrates on the possible variability patterns in the structure of a time series, and hence, to explore further characteristics so that better modelling and consequent predictions can be achieved in practical applications. Also among the

main purpose of this chapter is to present two simple but innovative methods for the trend slope and the variability change measurement, slope formulations and their statistical explanations. Extensive simulation studies have shown the validity of the proposed formulations. Especially, innovative trend slope expression results in a complete numerical agreement with the classical MK Trend slope values. Separately, a slope measurement for the variability quantification has been developed and its validity is also confirmed with extensive simulation studies. The methodology is applied to six Turkish rainfall records from the northeastern part of Turkey. In the last section a simple procedure is presented for variability identification through the innovative trend template procedure. Additionally, a new significance test is provided for the innovative trend analysis leading to levels that are parallel to 1:1, no trend line at 10 and 20% significance levels. The application of the methodology is presented for seven climatology regions of Turkey with annual daily extreme (maximum) rainfall records. In general, variability is defined in the statistical sense as the extent to which time series fluctuates on the average around the mean value (Chap. 5, Fig. 5.2). It also refers to the differences among data points within a time series as related to each other or to the mean. This can be expressed through the range, variance, or standard deviation of a dataset.

In all the previous sections, nonstationary time series exhibition of monotonic linear trend detection methodologies are presented in the mean (arithmetic average) level. However, time series may have nonstationarity also in the standard deviation along the time evolution of the natural, environmental, or economic events. The nonstationarity in the variance and hence, in the standard deviation is referred to as the variability in this chapter.

7.2 Variability Measures

In practical applications averages are considered in almost all the studies, but the variation measurements are ignored, which gives the impression or assumption that the phenomenon concerned is uniform or steady-state. However, one may track the general tendency on the average but in industrial processes, economic transactions, and natural event assessments variabilities around the averages are more important for quality and extreme value inspection, prediction, and control.

7.2.1 Range

The simplest measure of variability is the range, and it is the difference between the highest and lowest data values in a time series. In the mathematical context it is the domain of variation. It has two limits as lower and upper extreme values depending on the phenomenon considered. In general, depending on the type of the limits there are four different ranges.

(1) Close upper and lower limits: These are either physically or definitionally restrictive rages, which confine all the possibilities between upper and lower limits. For instance, physically the Sun has a source of ultraviolet light with wavelengths of 10–310 nm. The visible light covers the range of wavelengths from 400–700 nm. On the other hand there are many definitional ranges such as the probability variability domain is between 0 and 1, correlation coefficient varies between −1 and +1, and porosity percentage has possible variability between 0 and 100.
(2) Close upper end but open lower limit.
(3) Open upper end but close upper limit.
(4) Open upper and lower limits:

The variation analysis is depending not only on the internal variability at a fixed site of the variable but more significantly on the regional scatter of sampling points. In quantifying the variability the simplest approach is to find the range of variation by comparison of the maximum and minimum values within the record. If the relative error as given by Eq. (3.60) between the maximum and minimum record values is less than ±5%, there is no significant variation in the records. In this case one can depend on either the arithmetic average or more specifically on the mode (the most frequently appearing data value) as the sole representative for whole region. If the maximum and the minimum values are $Z_{\max}$ and $Z_{\min}$, respectively, then the range of data variability is defined as,

$$R_Z = Z_M - Z_m \tag{7.1}$$

The maximum percentage error can be defined as follows,

$$e_m = 100\frac{R_Z}{Z_M} \tag{7.2}$$

If $e_m \leq 5$, the overall representative value can be taken as the arithmetic average without any need for detailed modelling because the basic phenomenon is not complex and its behavior is more or less homogeneous.

7.2.2 Standard Deviation

The mean or arithmetic average of a time series represents almost a mid-point of the data fluctuations. It is different than the median, which refers to the exact value of the time series value that falls at the center of the data points provided that the time series data are sorted in ascending or descending order. While the median must be represented by the precise mid-value, the mean may or may not be actually falls on the same value.

The measure of variation around the arithmetic average (mean) value in statistics is the variance or its square root standard deviation. In practical application standard

deviation is preferred, because it has the same dimension as the original time series and hence, it can also be represented on the same figure, if necessary. The variability is also synonymous to terminologies of dispersion, scatter, or spread, which represents how stretched or squeeze a PDF is around the mean value. Among the common measures of variability are the variance, standard deviation, and as explained above the interquartile range.

If one would like to quantify the variability in more detail, then she/he should look for deviations from the arithmetic average value. In order to represent the overall variability rather than individual deviations, one should search for the summation of these deviations, which does not provide any additional information, because its value is equal to zero by definition. It is better to look for the sum of square deviations (SSD), which is never equal to one and the smaller its value the smaller is the overall variation. The minimization of the SSD is the main key in all the modelling works, and therefore, it is referred to as the least squares approach. If the sequence of data values is Z_i (i = 1, 2,…, n), where n is the sample length then the mathematical expression of the SSD is,

$$\mathrm{SSD} = \sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2 \tag{7.3}$$

Its expansion leads to,

$$\mathrm{SSD} = \sum_{i=1}^{n} Z_i^2 - 2\overline{Z}\sum_{i=1}^{n} Z_i + n\overline{Z}^2 \tag{7.4}$$

For the minimization procedure it is well-known from algebra that the derivation of this expression with respect to the arithmetic average must be equal to zero.

$$\frac{\partial(\mathrm{SSD})}{\partial \overline{Z}} = -2\sum_{i=1}^{n} Z_i + 2n\overline{Z} = 0 \tag{7.5}$$

Finally,

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{7.6}$$

which is well-known arithmetic average expression. More explicit writing of Eq. (7.6) provides additional interpretations for further developments.

$$\overline{Z} = \frac{1}{n}Z_1 + \frac{1}{n}Z_2 + \frac{1}{n}Z_3 + \ldots + \frac{1}{n}Z_n \tag{7.7}$$

The arithmetic average is a special case of weighted average where the weighting factor is $1/n$ for each measurement value. Furthermore, it shows that the

arithmetic average does not take into calculations the internal variability of the data sequence, because equal weight is attached to each measurement.

It is possible to generalize Eq. (7.7) into a more variability reflective form by giving different weighting factors, α_i, to each data value, Z_i, as,

$$\overline{Z}_w = \sum_{i=1}^{n} \alpha_i Z_i \tag{7.8}$$

where $\overline{Z}_w$ is referred to as weighted average. It is obvious from Eq. (7.7) that the summation of the weights is equal to 1, and likewise,

$$\alpha_1 + \alpha_2 + \ldots + \alpha_n = 1 \tag{7.9}$$

Each of the weighting factors indicates the variability within the event concerned. The bigger is the weighting factor the more is the variability contribution from the concerned data value.

It is a measure of dispersion around a general tendency, which is the arithmetic average. It is measured statistically by different parameters such as the interquartile range (IQR), variance, and standard deviation. It also refers to the extent to which values differ from one another, i.e., how much they vary. As for the probability distribution function (pdf) is concerned the variability implies to how spread out a distribution is.

7.2.3 The Interquartile Range (IQR)

Another important measure of variability is the interquartile range (IQR), which is based on dividing a data set into quartiles. In general, quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by $Q1$, $Q2$, and $Q3$, respectively.

- $Q1$ is the "middle" value in the first half of the rank-ordered data set.
- $Q2$ is the median value in the set.
- $Q3$ is the "middle" value in the second half of the rank-ordered data set.

In other words, the IQR is the 1st quartile subtracted from the 3rd quartile, which are also seen clearly on a box plot on the data (McGill et al. 1978). The IQR is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q_1, Q_2, and Q_3, respectively.

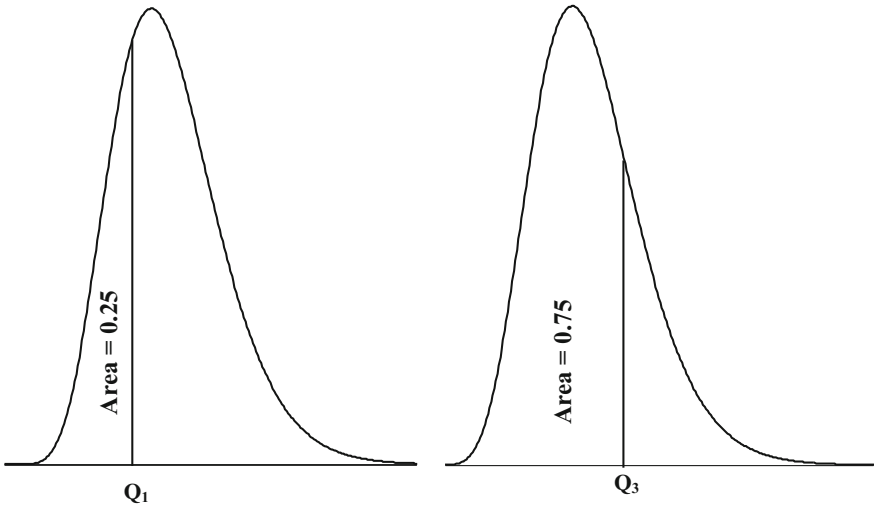


Fig. 7.1 Lower and upper quartiles

As descriptive statistics, the IQR, which is also called as the mid-spread or middle fifty, or technically H-spread, is a measure of statistical dispersion (variability) equal to the difference between the upper and lower quartiles.

$$\text{IQR} = Q_3 - Q_1$$

IQR of a continuous PDF can be calculated by integrating the PDF leading to the CDF. The lower quartile, Q_1, is a number such that integral of the PDF from $-\infty$ to Q_1 equals 0.25, while the upper quartile, Q_3, is such a number that the integral from $-\infty$ to Q_3 equals 0.75; in terms of the CDF, the quartiles can be defined as follows: where CDF^{-1} is the quantile function Fig. 7.1.

As a simple example let us consider the simplest time series as 1, 3, 4, 5, 5, 6, 7, and 11. The middle value, Q_1, is in the first half of this time series. Since there are an even number of data points in the first half of the data set, the middle value is the average of the two middle values; that is, $Q_1 = (3 + 4)/2 = 3.5$. On the other hand, Q_3 is the middle value in the second half of the data set. Again, since the second half of the data set has an even number of observations, the middle value is the average of the two middle values; that is, $Q_3 = (6 + 7)/2 = 6.5$. The interquartile range, IQR = 6.5 − 3.5 = 3.

7.2.4 Investment Variability

Variability is used also to standardize the returns from an investment and it provides a comparison basis for additional analysis. The excess return or risk premium per unit of risk for an asset can be measured by the reward-to-variability, which is the

Sharpe ratio. The significance of the Sharpe ratio is that it is a metric to compare the amount of compensation an investor receives with regard to the overall risk taken by holding investment. The excess return is based on the amount of usual return beyond investments that are considered free of risk. Provided that all else are equal, the asset with the higher Sharpe ratio delivers more return for the same amount of risk.

The risk perception of an asset class is directly proportional to the variability of its returns. As a result, the risk premium that investors demand to invest in assets, such as stocks and commodities, is higher than the risk premium for assets such as Treasury bills, which have a much lower return variability.

7.3 Trend and Variability Detection by Innovative Methodology

Since the climate change effect appearance in different parts of the world at variable scales on the social, natural, economic, and engineering aspects, researchers become interested whether there are temporal and spatial increasing or decreasing trend components in the recorded time series. For instance, climate change and variability reflections appear in the hydrological records, and in the future any successful engineering project design, operation, management, and maintenance of water resources will depend on objective trend and variability features of the records, their detection and interpretation. Different researchers have touched on these significant points (Hannaford and Marsh 2006; Gupta 2007; Lorenzo–Lacruz et al. 2012; Sharif et al. 2012; Larsen et. al. 2013; Haktanır and Citakoglu 2014). Furthermore, the impact of climate change on different elements of the hydrological cycle has been investigated by many researchers in different disciplines (Bao et al. 2012; Douglas and Fairbank 2011; Ehsanzadeh et al. 2011; Garbrecht et al. 2004; Novotny and Stefan 2007; Wagesho et al. 2012). Recently, Han et al. (2014) have performed simultaneously a comprehensive analysis of trends in precipitation and stream flow records at the Xiangxi River Watershed through multiple classical tests to detect the trend and their magnitudes.

Discontinuous meteorological variable predictions and variability such as precipitation under the prevalence of greenhouse warming are more speculative than continuous meteorological variable, say, temperature projections, especially at the regional and local geographic scales of interest to water planners. The most recent IPCC (2007, 2012) reports analyses suggest that a greenhouse warming will have continuous trend and variability effects on water supplies. Of course additionally landscape and geomorphological changes show unprecedented impacts on the runoff apart from the climate change through the precipitation and consequent runoff phenomenon. Trend existence and variability features may result from various effects and lead to different consequences. For instance, the timing and regional patterns of precipitation may change, and more intense (weak) precipitation days are likely, and hence, increasing (decreasing) trend effects take place in the recorded

time series. General circulation models (GCMs) provide climate change predictions as 1.5–4.5 °C rise in global mean temperature, which may increase global mean precipitation about 3–15%. Such an increase is expected to affect the hydrological elements such as runoff in terms of trends and also variabilities. Although the regional distribution is uncertain, precipitation is expected to increase in higher latitudes, particularly in winter. Potential Evapotranspiration (ET) as water evaporate from the surface and transpiration from plants rises with air temperature. Consequently, even in areas with precipitation increments, higher ET rates may lead to runoff reduction, which imply a possible reduction in renewable water supplies coupled with trend and variability features. As a result of precipitation increase consequently more annual runoff occurrences are likely in the high latitudes. In contrast, some lower latitude basins may experience large reductions in runoff and increase in water shortages as a result of evaporation increment coupled with precipitation decrement combination. Flood frequencies are likely to increase in many areas, although the amount of increase for any given climate scenario is uncertain and impacts may vary among basins, and floods may become less frequent in some areas. On the other hand, the frequency and severity of droughts could increase in some areas as a result of a decrease in total rainfall, more frequent dry spells, and higher ET. The hydrology of arid and semi-arid areas is particularly sensitive to climate variations. Relatively small changes in temperature and precipitation in these areas could result in large percentage changes in runoff, increasing the likelihood and severity of droughts and/or floods (Şen 2008). Seasonal disruptions might occur in the water supplies of mountainous areas, if more precipitation falls as rain than snow and if the length of the snow storage season reduces. Water quality problems may increase where there is less flow to dilute contaminants' contribution from natural and human sources.

7.3.1 Methodology

There are several methodological trend detection approaches in the statistics and hydrology literature, where most of the time in hydrological applications Mann–Kendal (MK) trend analysis approach is employed (Mann 1945; Kendall 1970). Additionally, trend slope determination by median slope calculation has also been in use parallel to trend detection as suggested by Sen. Theoretically, it is developed for infinite length of time series, and hence, finite sample lengths cause bias effects. In order to assess the influence of the dependent serial correlation structure, various authors have performed Monte Carlo simulation studies (Yue et al. 2002; Hamed and Rao 1998; Matalas and Sankarasubramanian 2003). Trend analysis evaluation is also needed for long term infra-structure design and risk analysis in hydro-meteorological time series. As stated by Fatichi et al (2013) due to climate change assessment, trend identification, detection, and evaluation are important issues in different disciplines. Growing importance of trend analysis is triggered also by the Intergovernmental Panel on Climate Change (IPCC 2007) Climate change.

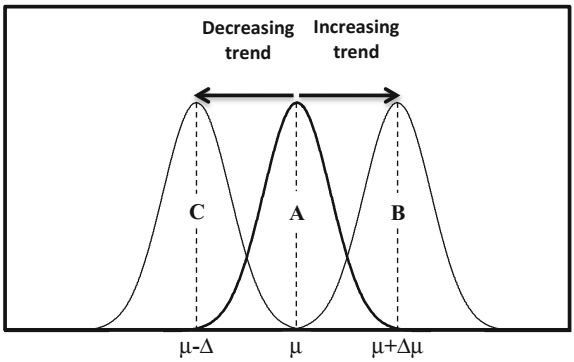


Fig. 7.2 Trend without variation

In any climate fundamental text book, the climate change is explained on the basis of a simple probability distribution function (pdf) shift as in Fig. 7.2. Although in this figure normal (Gaussian) pdf is demonstrated for simplicity, it is possible to replace it with any suitable pdf to the records at hand.

Close inspection of this figure leads to the deduction of the following major points that are very important for the methodology development.

(1) In order to decide whether there is a climate change two different pdf's must be considered, which implies that two different time series (Şen 2012, 2014) and consequent trend and variability searches,
(2) The shapes of the pdf's remain the same and this point implies the stationarity of the respective time series,
(3) For trend detection instead of the whole pdf, it is enough to consider the arithmetic average of each time series and compare them. In case of significant difference one can conclude that there is a trend in the time series,
(4) The trend existence is coupled with the difference in the arithmetic average, $\Delta\mu$, which implies the traditional trend component as in the Mann–Kendall trend test,
(5) It is also important to know the time difference between the two pdf's or average calculation time periods. If this time difference period is denoted by Δt, then the rate of change, i.e., slope, S of increasing trend can be expressed as follows,

$$S_{\mu}^{+} = \frac{(\mu + \Delta\mu) - \mu}{\Delta t} = +\frac{\Delta\mu}{\Delta t} \tag{7.10}$$

On the other hand, if there is a decreasing trend with the same argument, the slope of the trend becomes similar to the previous expression as,

$$S_{\mu}^{-} = \frac{\mu - (\mu + \Delta\mu)}{\Delta t} = -\frac{\Delta\mu}{\Delta t} \tag{7.11}$$

Hence, by consideration of innovative trend methodology developed by Şen (2012, 2014) with the two halves of the same time series, one can calculate the arithmetic average of the first half, m_1, and then of the second half, m_2, and consequently, the slope formulation can be generalized as,

$$S_m = \frac{m_2 - m_1}{\left(\frac{n}{2}\right)} \tag{7.12}$$

where n is the number of data in the time series. Herein, half of the time series length, $n/2$, is adapted, because this is the time period difference between the two halves. Depending on the sign in Eq. (7.12) one can decide whether the trend has increasing (positive) or decreasing (negative) tendency. All these explanations are valid for homoscedasticity (standard deviation constancy), however, in nature the time series may behave in a nonstationary manner.

Another parameter is the standard deviation that shows itself with the pdf's terms as in Fig. 7.3. Since, the standard deviation is the square root of the variance, it is a variability measure.

Hence, by considering the arithmetic mean and the standard deviation relative positions between two pdf's similar to Fig. 7.12, one can again distinguish three cases (A, B, and C). In general, the standard deviations may be different from each other as $\sigma_A \neq \sigma_B \neq \sigma_C$.

It is possible that the time series may be the first order stationary, which implies that there is variation in the variance (standard deviation) by time. In other words the time series is not homoscedastic. In Fig. 7.3 there are variations without any trend existence and again two time series are comparable on the basis of the standard deviation. The change in the standard deviation per time duration is the definition of the variation measure. Hence, similar to Eq. (7.10) one can deduce simply that the variation slope can be expressed as,

$$S_\sigma^{\pm} = \frac{(\sigma + \Delta\sigma) - \mu}{\Delta t} = \pm\frac{\Delta\sigma}{\Delta t} \tag{7.13}$$

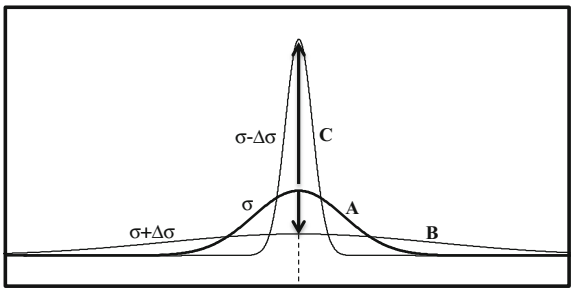


Fig. 7.3 Variation without trend

With the practically useable notations by taking into consideration the two standard deviations as s_1 and s_2 for the first and the second halves of time series one can write the variability slope, S_v as,

$$S_v = \frac{s_2 - s_1}{\left(\frac{n}{2}\right)} \tag{7.14}$$

The arithmetic mean and the standard deviation are two basic statistical parameters that are completely independent from each other, and therefore, Eqs. (7.12) and (7.14) can be applied to any time series independently from each other.

7.3.2 Simulation Study

In order to confirm the validity of the formulations in the previous section, extensive Monte Carlo simulation study is carried out with dependent and independent time series generation. Herein, as an alternative, the classical Sen (1968) slope calculation procedure, is taken into consideration for comparison purposes and it expresses the trend slope as,

$$S = \text{median}\left(\frac{X_i - X_j}{i - j}\right) \text{for} \quad (i > j) \tag{7.15}$$

where X_i (X_j) is the i-th (j-th) time-series value. Hence, there is a big difference between this and the alternative slope formulation (Eq. 7.12) in this chapter. The most important point in any Mann–Kendal trend test is the calculation of this slope value.

The longest hydro-meteorological historical records rarely have durations close to 100 years or slightly more. This is the main reason why in the simulation studies 100-year length synthetic series are considered. In the simulation works, standardized synthetic series ($\mu = 0$, $\sigma = 1$) are generated, because of the independence of the arithmetic mean from the standard deviation. Furthermore, a set of ensembles of the same length are considered with the applications of Eqs. (7.12), (7.14), and (7.15) to each ensemble member, and finally, their arithmetic averages are considered as the final results. For different arbitrary sets of first order serial correlation coefficient, ρ, and trend slope values, the simulation results are presented in Table 7.1.

The first striking conclusion is the validity of Eq. (7.12) as the trend slope formulation, because it yields almost the same numerical values with the well-established Sen trend slope results within the sampling error limits less than 5%. It is possible that another set of different first order correlation coefficient and trend slope values may be adapted and the same simulation study leads to another set of numerical results. The combination of all such results appears along the 1:1

Table 7.1 Trend without variability simulation results

Type of trend	Simulation time series					Half time series				Slope	
	μ	σ	ρ	Slope		Arithmetic average		Standard deviation		Equation (7.12)	Equation (7.15)
				Trend	Variation	First half	Second half	First half	Second half		
NTNV	0.00	1.00	0.00	0.000	0.00	0.0292	−0.0124	0.9965	0.9725	−0.0008318	−0.000217
ITNV	0.00	1.00	0.00	0.020	0.00	5.0916	15.0792	3.0733	3.0819	0.01998	0.01998
DTNV	0.00	1.00	0.00	−0.030	0.00	−0.7742	−2.2447	1.0435	1.0633	−0.0294	−0.0291
NTNV	0.00	1.00	0.10	0.000	0.00	0.0073	−0.0374	0.9979	0.9821	−0.0008950	−0.0002549
NTNV	0.00	1.00	0.30	0.000	0.00	0.0084	−0.0234	0.9510	0.9472	−0.0006360	−0.0009584
NTNV	0.00	1.00	0.50	0.000	0.00	−0.0458	−0.0035	0.8684	0.631	−0.0009857	−0.0007618
NTNV	0.00	1.00	0.70	0.000	0.00	0.0129	−0.0325	0.7131	0.6581	−0.0009065	−0.00001314
NTNV	0.00	1.00	0.90	0.000	0.00	−0.0954	0.0204	0.3792	0.3639	0.0023	0.0017
ITNV	0.00	1.00	0.10	0.02	0.00	0.5002	1.5278	1.0272	1.0314	0.0206	0.0201
ITNV	0.00	1.00	0.30	0.05	0.00	1.2885	3.7903	1.2000	1.1590	0.0500	0.0502
ITNV	0.00	1.00	0.50	0.04	0.00	1.0534	2.9915	0.9788	1.0286	0.0396	0.0391
ITNV	0.00	1.00	0.70	0.03	0.00	0.7471	4.5801	0.8277	0.8135	0.0299	0.0306
ITNV	0.00	1.00	0.90	0.06	0.00	1.5313	44.956	0.9791	0.8886	0.0610	0.0606

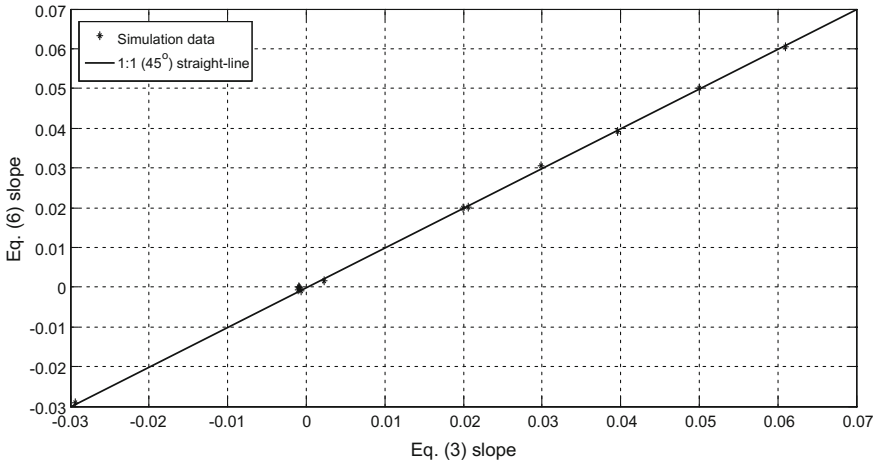


Fig. 7.4 Conformity between Eqs. (7.12) and (7.14)

(45°) line as in Fig. 7.4. This is the line that confirms the validity of herein suggested formulation for trend slope as in Eq. (7.12).

Another set of simulation study is performed for the variability formulation (Eq. 7.13) validated again with the standardized synthetic independent and dependent processes and the results are presented in Table 7.2. In the literature, there is no alternative for variability slope formulation (Eq. 7.13), and therefore, in the last column the results from this equation only are given.

In order to check the validity of the suggested formulation in Eq. (7.13) two types of plots are considered. Since the formulation is for the estimation of the simulation variability, first preselected set of simulation variability values are plotted against the same values so as to get a reference line of 1:1 (45°) line. On the same graph this time fixed set of simulation variability results are plotted against the simulation result values from the last column of Table 7.1 (Eq. 7.13 results) as in Fig. 7.5.

It is obvious from this figure that the simulation line is slightly below the reference line and at the maximum the difference is 0.11, which corresponds to $100 \times 0.11/3.0 = 3.36\%$ relative error and this is well below the acceptable level, because it is within the practically acceptable error limits of $\pm 5\%$.

7.3.3 Applications

As the application area, six meteorology station annual rainfall records around the Istanbul City are selected as shown in Fig. 7.6. These are scattered on the European and Asian parts of the city.

Table 7.2 Variability without trend simulation results

Type of trend	Simulation time series					Half time series				Slope	
	μ	σ	ρ	Slope		Arithmetic average		Standard deviation		Equation (7.5)	Equation (7.6)
				Trend	Variation	First half	Second half	First half	Second half		
NTNV	0.00	1.00	0.00	0.000	0.30	0.0618	0.1859	8.7244	22.3944	0.2767	4.104×10^{-4}
ITNV	0.00	1.00	0.00	0.000	0.00	0.0292	−0.0124	0.9965	0.9725	−0.0008318	−0.000217
DTNV	0.00	1.00	0.00	0.000	3.00	0.4254	8.9373	87.8372	232.1700	2.89	0.0712
NTNV	0.00	1.00	0.10	0.000	0.01	0.0073	−0.0147	2.8699	7.6017	0.00946	-1.392×10^{-4}
NTNV	0.00	1.00	0.30	0.000	0.02	0.0095	0.0009	0.5357	1.4893	0.0191	-2.301×10^{-4}
NTNV	0.00	1.00	0.50	0.000	0.03	0.0133	−0.0539	0.6850	1.9366	0.0273	-2.176×10^{-4}
NTNV	0.00	1.00	0.70	0.000	0.05	0.0623	0.0955	1.0067	2.6624	0.0450	-5.653×10^{-4}
NTNV	0.00	1.00	0.90	0.000	0.04	0.0093	−0.1461	0.4230	1.1595	0.0124	4.060×10^{-5}
ITNV	0.00	1.00	0.10	0.000	1.0	0.1004	0.8775	29.1679	76.6184	0.9490	0.0000456
ITNV	0.00	1.00	0.30	0.000	1.5	−0.7153	1.0548	44.0227	114.0216	1.4000	0.0289
ITNV	0.00	1.00	0.50	0.000	2.0	−1.2815	4.7239	55.3646	147.3385	1.8395	0.1192
ITNV	0.00	1.00	0.70	0.000	2.5	4.8466	13.1413	69.1483	183.4902	2.2868	0.2068

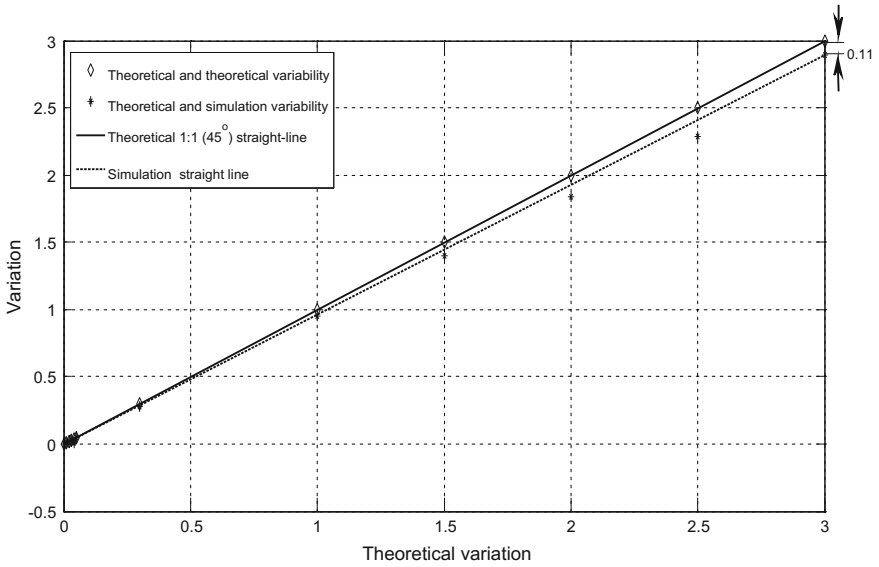


Fig. 7.5 Validity confirmation of Eq. (7.14)

Fig. 7.6 Meteorology station locations

After the division of the basic record time series into two nonoverlapping equal duration sections as half time series, the computer software written on Matlab program is used for various calculations, and the results are given in Table 7.3.

In Fig. 7.7 each station time series is presented with three straight-line trends on each, namely trend lines according to Eqs. (7.12) and (7.15), which represent the trend line in this chapter and the MK trend line, respectively. It is obvious that these trend lines are very close to each other in all the stations. Additionally, variability line is also shown, which is rather different than the two previous lines.

One can realize that the most increasing variability effect appears at Sarıyer meteorology station, which is open to northwesterly air movements from the North Atlantic Ocean and through the European continent and Balkan Peninsula. At the same station there is also increasing trend on the arithmetic average value. Similar pattern appears at the Kumköy station. Just the opposite trend patterns are valid at the Kartal station, where there are significant decreases in the trend and variability features. The least variability is at the Göztepe station, because it lies in the inland of the Asian side of Istanbul City, where the air movements reach to a more or less stable situation. Florya station is also stable on the European side and its location is protected from general air movement, where there is neither significant trend nor variability. The Bahçeşehir is exposed to landscape change, because many vegetative areas are turned to local settlements and this is the main reason why there is a decreasing trend, however the variability is more or less remained on the same level.

7.4 Trend Significance Limits

In order to base the deviations of each scatter point from the 1:1 straight line, herein a quantitative significance test is suggested. Figure 7.8 represents the 1:1 (45°) straight-line scatter plots as explained briefly in the Introduction section (Şen 2012, 2014).

In case of no trend and no variability, the scatter points lie on or very close around the 1:1 line, which implies that the corresponding values in the first and second half series are the same without any significant trend or variability. In other words, the mean and the standard deviation of the first and the second halves are significantly close to each other, i.e., the differences are practically equal to zero. However, for randomly distributed time series, this statement implies that the expected value or the arithmetic average of the differences is equal to zero, $E(m_2 - m_1) = 0$. In any given natural hydro-meteorological time series such a perfect case is not valid, and therefore, there will be deviations from 1:1 line. The smaller the square root of square deviation summation (SRSDS) from the 1:1 (45°) straight-line, the closer are the scatter points to 1:1 straight line, and accordingly, there is not a significant trend or variability component. One can calculate the SRSDS, s_d, between the two half series scatter points from the 1:1 line as,

Table 7.3 Trend and variability features of the stations

Station name				Half time series				Slope		
	μ	σ	ρ	Arithmetic average		Standard deviation		Trend	Variability	Sen
				First half	Second half	First half	Second half			
Bahçeköy	151.83	88.85	0.159	162.53	141.12	82.58	95.11	−0.8235	0.4816	−1.1137
Florya	102.02	49.57	0.238	104.50	98.18	51.73	48.19	−0.2006	−0.1126	0.0976
Göztepe	697.44	120.95	0.180	670.15	689.41	122.32	122.27	0.5424	−0.0015	0.6909
Kartal	85.37	47.56	0.394	92.12	78.63	49.90	45.10	−0.5395	−0.1921	−0.9655
Kumkoy	774.09	159.29	0.132	772.76	810.21	118.03	170.64	3.5694	2.1473	2.8163
Sarıyer	789.70	137.48	0.157	751.69	822.51	110.12	154.67	2.7771	1.7472	2.1474

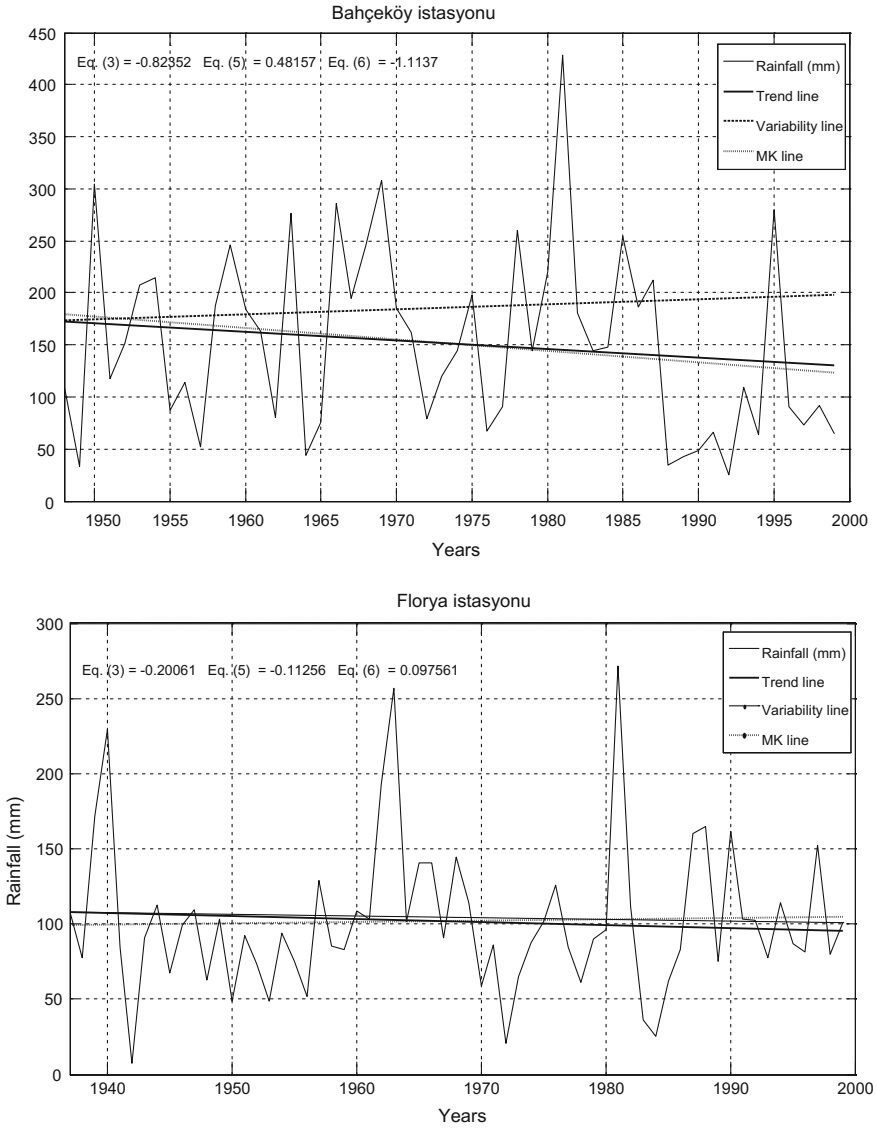


Fig. 7.7 Trend and variability lines for each station

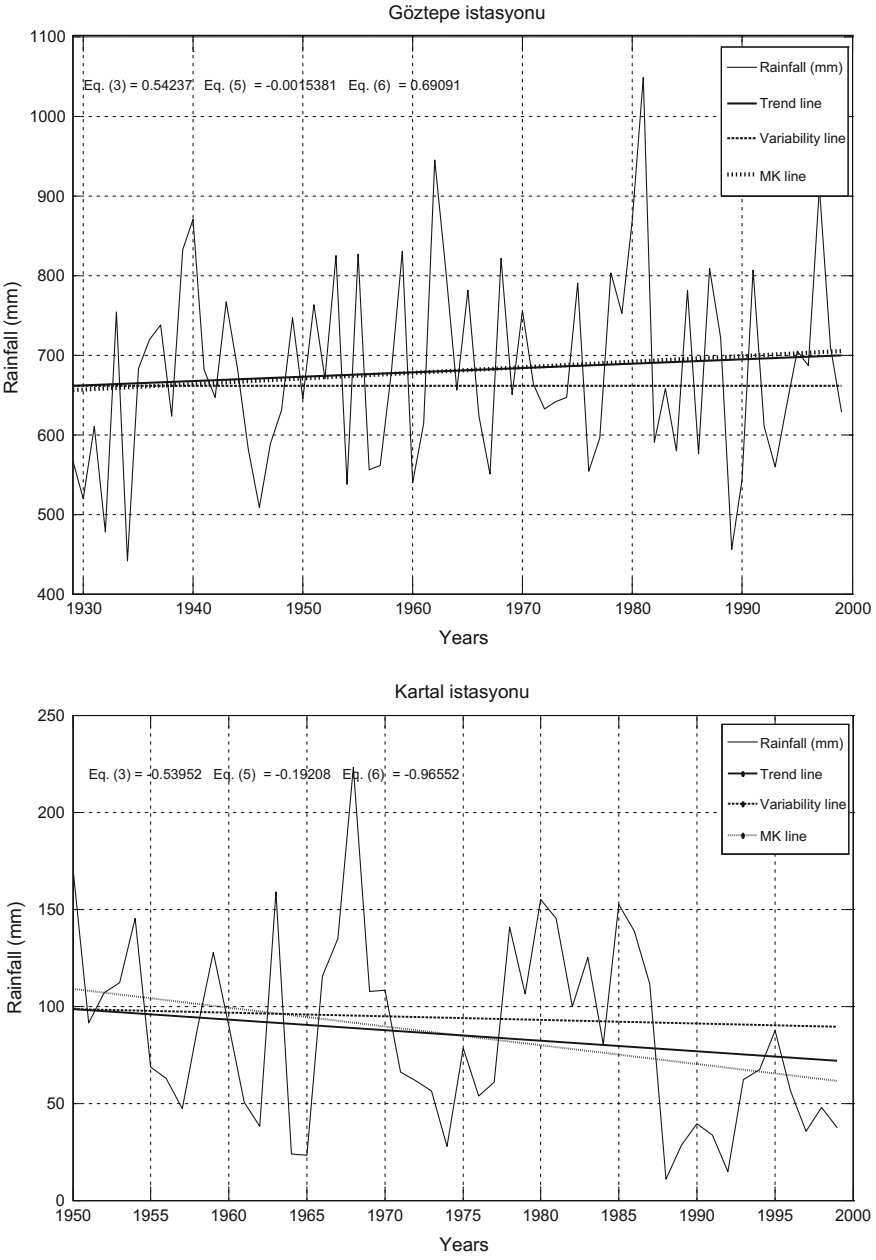


Fig. 7.7 (continued)

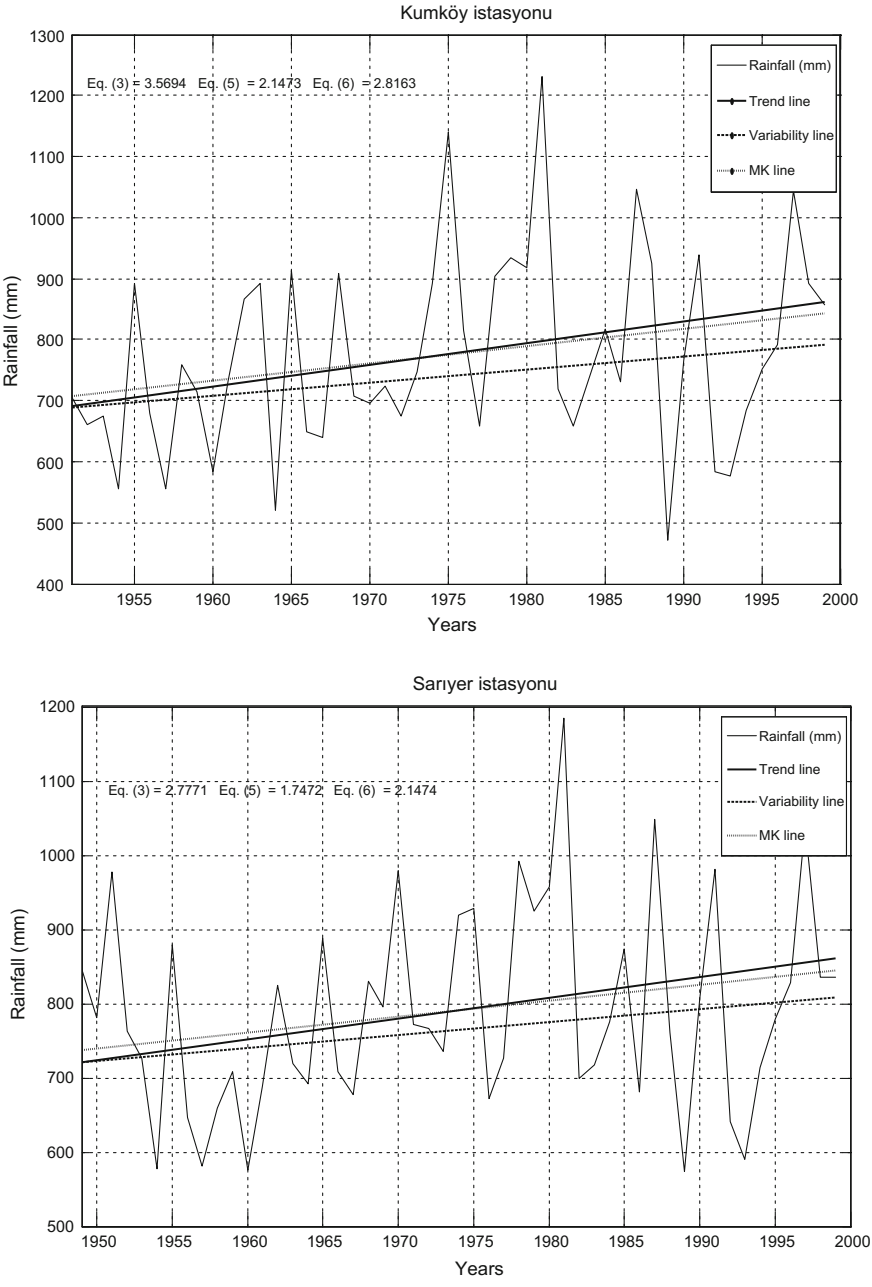


Fig. 7.7 (continued)

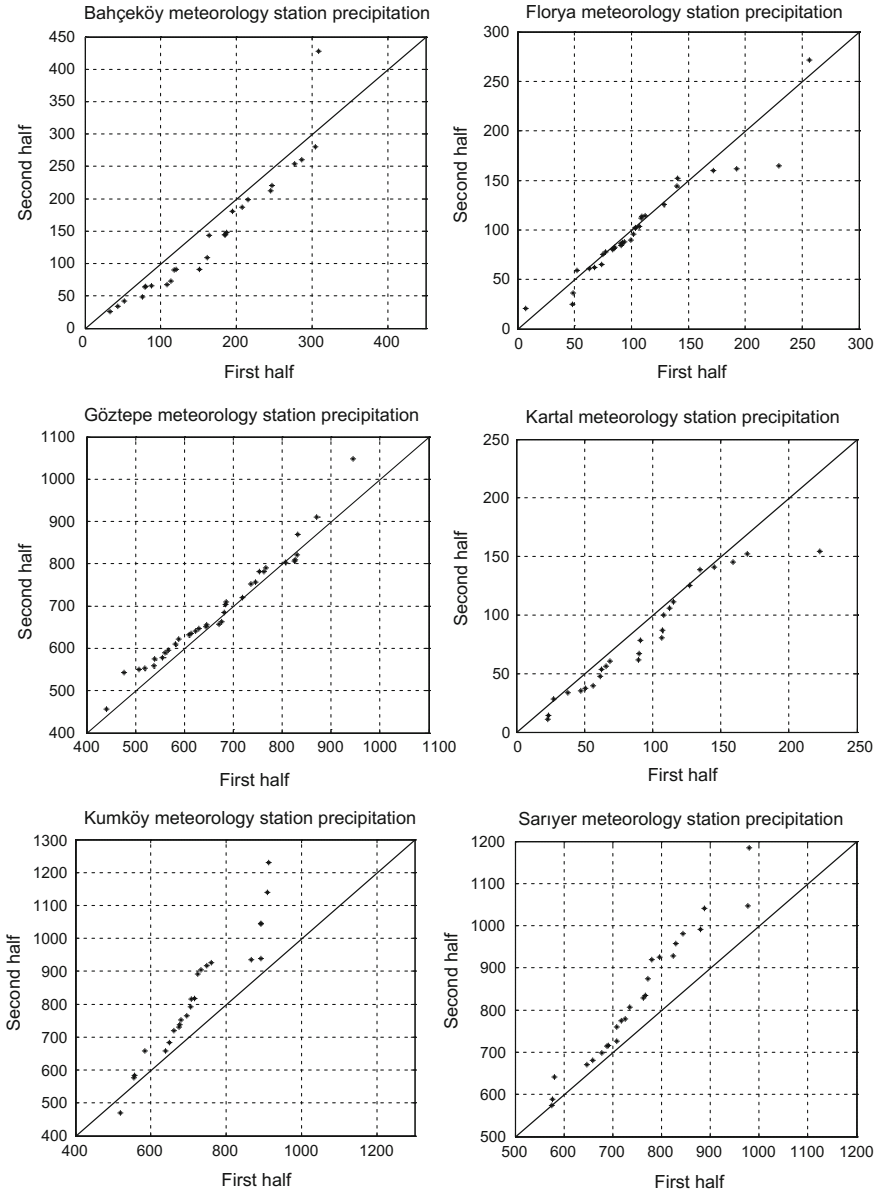


Fig. 7.8 1:1 (45°) straight-line graphs for each station

$$s_d = \sqrt{\frac{1}{n}\sum_{i=1}^{n/2}\left(X_i - X_{(n/2)+1}\right)^2} \tag{7.16}$$

The smaller is the value of s_d, the closer are the scatter points from 1:1 (45°) straight line, which implies verbally insignificant trend or variability existence. In order to convert this information into an objective form the division of the mean difference, $(m_2 - m_1)$, by the SRSDS in Eq. (7.16) leads to the definition of trend test statistic, t_s, as,

$$t_s = \frac{(m_2 - m_1)}{s_d} \tag{7.17}$$

The small values of this test statistics, t_s, imply that there is trend and variability, which is regarded as the null hypothesis, H_o. On the contrary, the big values corresponds to the alternative hypothesis, H_a, where there is no trend or variability. Theoretically, t_s has zero mean and unit variance, and hence, the standard normal pdf can be used for the significance test. The summary of the necessary calculations is presented in Table 7.4 with the trend test statistic, t_s, values in the last column.

The positive (negative) sign of the t_s value implies increasing (decreasing) trend component in the given time series, which are in complete agreement with respective graphs in Fig. 7.6 for each station.

In a two-tail significance level based on the standard normal pdf with zero mean and unit variance at 5% significance, the lower (upper) confidence limit is −1.96 (+1.96). Hence, all the trend test statistic in the last column falls between these confidence limits, and therefore, in all the time series there are significant trend and variability components, which are presented quantitatively in the previous section.

Last but not the least, each graph in Fig. 7.8 provides additional information to corresponding trend and variability graphs in Fig. 7.7. For instance, Bahçeköy meteorology station graph in Fig. 7.8 exposes that all the scatter points are below the 1:1 (45°) straight line and this implies that there is a decreasing trend in the time series, which is also confirmed by the corresponding graph in Fig. 7.7. For the last two stations, Kumköy and Sarıyer, Fig. 7.8 graphs scatter points are over the 1:1

Table 7.4 Significance test values

Station	First half mean, m_1	Second half mean, m_2	Mean difference $(m_2 - m_1)$	s_d	t_s-statistics
Bahçeköy	162.5	141.1	−21.40	37.71	−0.5674
Florya	104.5	99.6	−4.90	14.98	−0.3271
Göztepe	670.15	690.19	20.04	30.56	0.6557
Kartal	92.11	78.63	−13.49	19.14	−0.7042
Kumköy	722.75	821.98	99.22	123.36	0.8043
Sarıyer	751.7	825.9	74.20	90.59	0.8190

(45°) straight line and such scatters imply the existence of strongly increasing trend, which is also confirmed by the last two graphs in Fig. 7.7.

It can be concluded that in hydrology studies for the last two to three decades, trend analysis oriented publication rate increased because of the climate and landscape changes that affect directly the hydrological cycle to certain extend in different parts of the world at different rates. In general, the trend studies are based on the classical Mann–Kendal (MK) statistical test analysis and in the detection of trends the key variable is its slope. In this chapter, simple and effective slope formulation is suggested and its validity is proved by extensive simulation studies. Additionally, the comparison of this slope formulation results with the one that exists in the literature shows a 1:1 (45°) straight line as another evidence of its validity. Another concept that is frequently used in the hydrology literature is the variability in addition to trend analysis, but it has not been quantified in the literature by objective formulations. Hence, additional innovative point in this chapter is the suggestion of a valid variability formulation, which has been also confirmed with the simulation studies.

The application of the trend and variability formulations are achieved for six meteorology stations and the new trend in addition to the MK trend lines are shown on the same graph and they almost follow each other within practically acceptable error limits. Additionally, on the same graphs, the variability lines are also indicated and the interpretations of all these straight lines are given for the study area, which are the European and Asian side meteorology station records.

7.5 Trend and Variability Analyses by Innovative and Classical Methodologies

Temperature and precipitation time series records are the two major data sources for the assessment of climate change impact on social, economic, and water resources engineering planning, management, and operation (IPCC 2007, 2013, 2014). The most significant component in any climate change study is the search for trend component, because it indicates on the average temporally increasing or decreasing tendencies that are important for future planning and management studies. In the past, many researches depended on the assumption that the past is the reflection of the future, which meant that the hydro-meteorological time series have a stationary structure (Maass et al. 1962; Milly et al. 2008). However, the climate change impact overturned this assumption in the sense that as for the statistical features of hydro-meteorological records, including trends, are concerned, the past is not reflection of the future (Milly et al. 2008). Fatichi et al. (2013) stated that due to climate change assessment, trend identification, detection, and evaluation became important issues in different disciplines.

Trends in the precipitation records are not homogeneous monotonically, but many trend analyses do not provide detailed information on this point. Most trend analyses treat the available hydro-meteorological time series monotonically as a

single trend over the whole record duration, without any distinction either time wise or record value wise as "low" (drought), "medium," and "high" (flood) classes.

Preliminary trend test is given by Mann (1945) and Kendall (1970), which provides information about possible trend (increasing, decreasing, or no trend). This approach requires that the hydro-meteorological records must have independent serial correlation structure. von Storch (1995) noticed that the Mann–Kendal (MK) statistical trend test cannot yield significant trend in hydro-meteorological series with statistically significant serial correlations. In order to alleviate this point, he suggested pre-whitening procedure so as to render the serial dependence into independence status and then apply MK trend test. The pre-whitening procedure is incapable to transform the original series into a completely independent series with zero correlation coefficient. Yue and Wang (2002) and Bayazıt and Önöz (2007) stated that by pre-whitening, the dependence structure can be reduced such that the serial correlation coefficients becomes close to zero. Douglas et al. (2000) indicated that after the pre-whitening application to some flows in the United States, trend appeared less than prior to pre-whitening. The pre-whitening procedure has been applied for trend analyses prior to MK trend test without any proof of its ability to fulfil independence structure (Zhang et al. 2001; Hamilton et al. 2001; Burn and Hag Elnur 2002).

The MK procedure is supported by the trend slope calculation as the median of all the possible slopes between the two record values within the time series (Sen 1968). Other trend procedures are the Spearman's tau (Spearman 1904) and traditional regression analysis.

The MK trend test has been applied in a number of hydrological, atmospheric, and environmental researchers on hydro-meteorological and climatological time series records by many authors (Hirsh et al. 1982; Hirsh and Slack 1984; Hipel et al. 1988; Demaree and Nicolis 1990; Yue et al. 1993; Gan 1998; Taylor and Lotfis 1989; Douglas et al. 2000; Hamilton et al. 2001; Kalra et al. 2008; Hamed 2008).

Recently, Şen (2012, 2014) has suggested an innovative trend analysis methodology, which can penetrate into the hydro-meteorological record value classifications as "low", "medium," and "high" values and makes assessment of trends categorically. This is a nonparametric approach, because the record is divided into two halves and after sorting them individually in ascending order, the first half is plotted against the other so as to identify possible trends categorically.

7.5.1 Şen Innovative Trend Analysis

Trend methodologies can be classified as nonparametric and parametric procedures and each one with a specific set of assumptions. In the following sequel Şen (2012, 2014) method is employed, which divides the given record series into two equal parts, each part is sorted in ascending order and then plotted against each other leading to scatter points as in Fig. 7.9. The 45° (1:1) straight line is a divisor of the square domain of the scatter diagram into two halves as the upper and lower

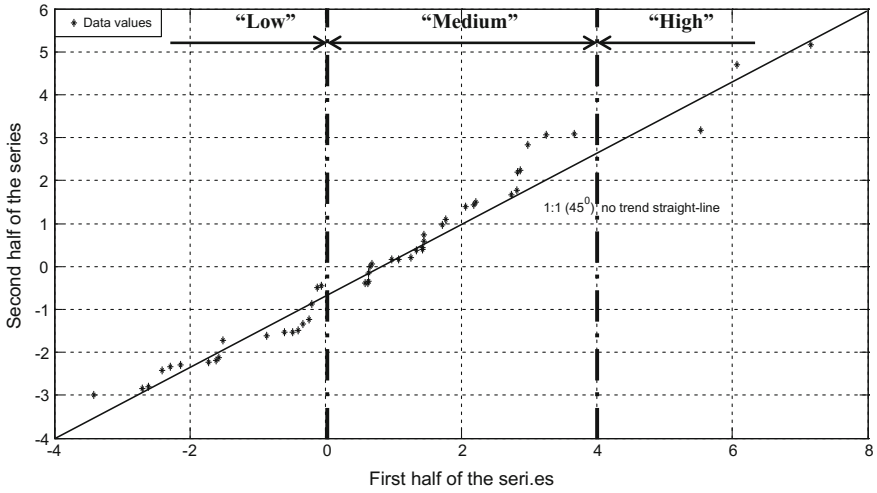


Fig. 7.9 Innovative trend template

triangular parts. The upper (lower) part represents increasing (decreasing) trend domain. If the scatter of points are on the 1:1 line or significantly close to this line then there is no significant trend component in the original time series.

This procedure provides trend information about the "low", "medium," and "high" data values categorically.

7.6 Application and Interpretations

For the application of aforementioned trend methodologies seven meteorology station annual daily extreme rainfall data are taken into consideration. Each station location represents different geographic and climatological region of Turkey. The station locations and the geographical regions are shown in Fig. 7.10.

The length of each record is more than 50 years, which is statistically acceptable for reliable applications. Although each record starts at different years but ends in year 2010, inclusive.

In general, Turkey is in the sub-tropical climate belt of the world, but within the country there are seven different sub-climatological regions.

(1) Black Sea region: This is in the northern Anatolia, where the penetration of the North Atlantic air masses reaches the Black Sea and then confronts with the mountain chain that is parallel to sea coast. Consequently, during the winter seasons frontal type of precipitation occurrences are predominantly frequent, whereas in the summer months orographic type of rainfall takes place. Trabzon is the representative station for this region.

Fig. 7.10 Meteorology and geographic region locations

(2) Marmara region: This region is under the effect of the air masses that come from the Balkan Peninsula and also especially in the summer seasons Mediterranean type of climate belt extends toward the north covering this region. Summer seasons are warm but winter seasons are very severe in the European part of Turkey. The location of Istanbul City including Goztepe meteorology station has rather mild climatic features due to the *Dardanelles* and Bosphorus straights.

(3) Agean region: This region is under the effect of maritime climatic effects that penetrate from the Mediterranean Sea with very hot summer months and mild winter conditions.

(4) Mediterranean region: Again maritime climatic conditions prevail, and according to the IPCC (2007, 2012, 2013) reports, especially eastern Mediterranean area will be under the effect of climate impact with decreasing rainfall amounts and frequencies. The region has humid climate with very hot summers and mild winter months.

(5) Central Anatolia region: This is the most steppic, semi-arid and the least rainfall receiving region in Turkey. The annual rainfall averages are around 250 mm with dry conditions, especially in spring and winter seasons, with rather frequent and elongated drought periods. It is in the form of a close drainage basin.

(6) Southeastern Anatolia region: This area lies in the most southeastern part of Turkey at the head of the Mesapotamian valley. It has very dry summer seasons and mild continental climatic conditions.

(7) Eastern Anatolia region: Rugged mountains with elevations reaching to 5,000 m are dominant in this region and they cause very severe winter conditions under the effect of continental climate. The upstream of Euphrates and Tigris rivers are in this region, where these two rivers get their waters especially in spring seasons, due to late snow melt phenomenon.

7.6.1 Probability Distribution Functions (pdf)

The first step is to identify the pdf of each meteorology station for interpretations of convenient distribution that is meaningful for future predictions. The pdf's do not provide any clue about either the trend component or the variability behavior. Figure 7.11 presents the valid pdf's for each meteorology station annual daily extreme rainfall values. Each graph indicates almost perfect match between the empirical data scatter and the theoretical pdf's. The presentation is in the form of exceedence probability variation against the data values. It is possible to see the annual daily extreme rainfall amounts for a set of return periods as 2-year, 5-year, 10-year, 25-year, 50-year, and 100-year. These extreme values are the basic decision quantities in any water resources engineering structure design provided that the return period is given.

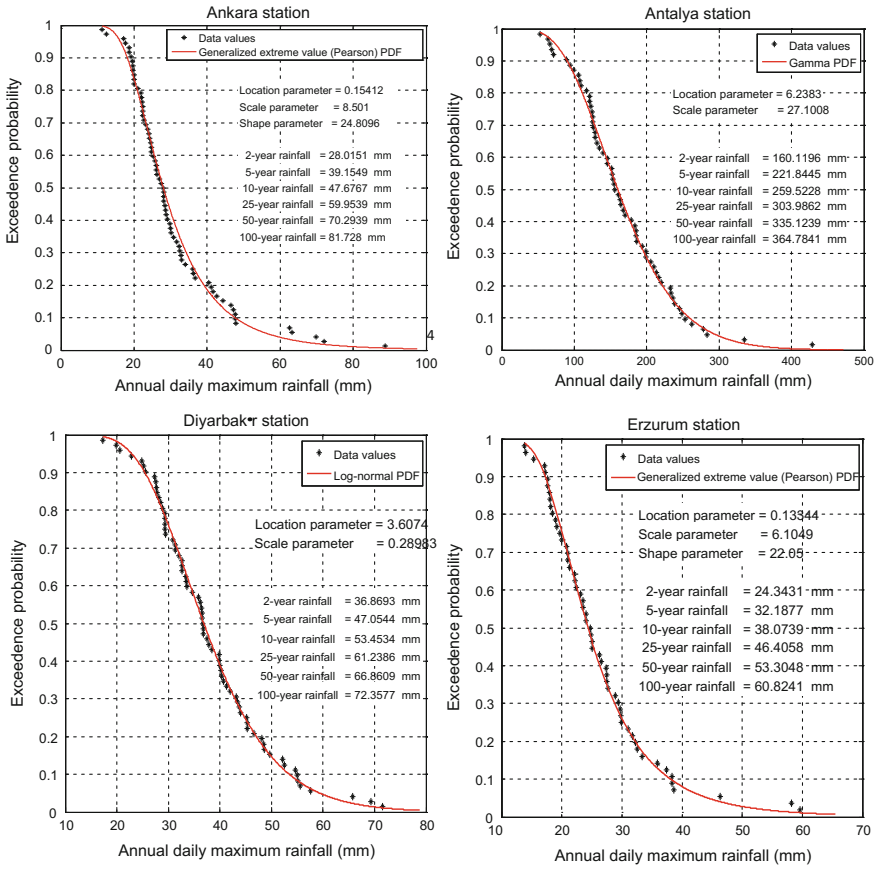


Fig. 7.11 Probability distribution functions

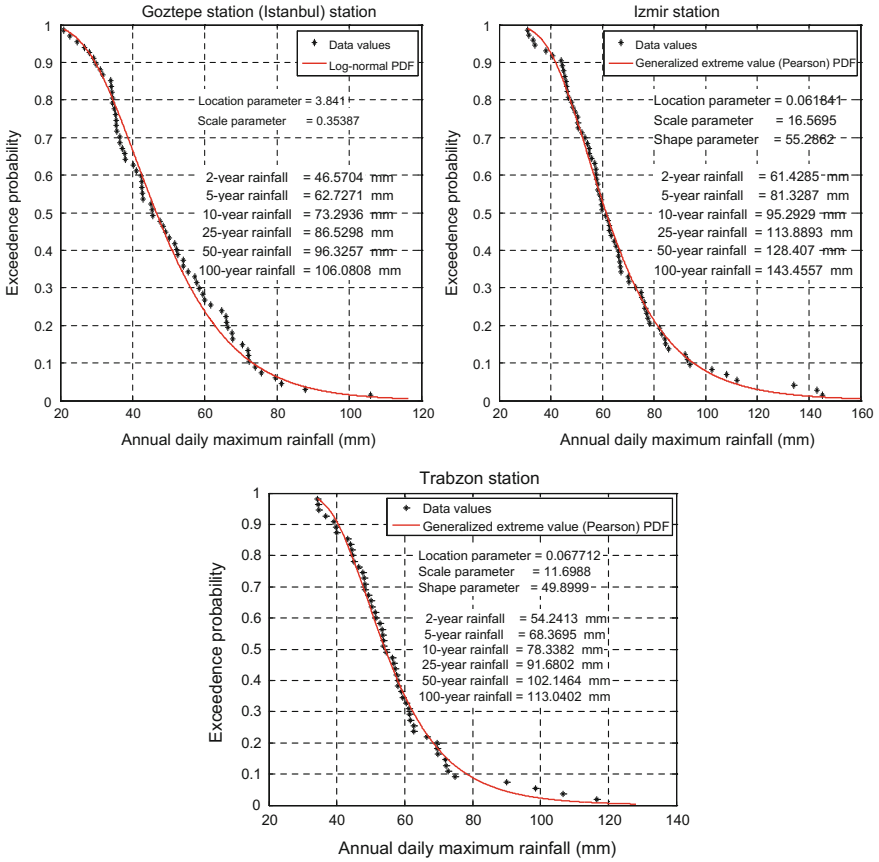


Fig. 7.11 (continued)

All the trend analyses methods mentioned in the previous section can be applied without any restriction to each one of the pdf's.

7.6.2 Different Trends

Since the annual daily extreme rainfall values have independent structure, pre-whitening procedure application is not necessary for the MK methodology. The independent serial structure provides a common base for each trend analysis to yield very close results to each other. In Fig. 7.12 two classical trend methods (Sen and regression) are shown in addition to the innovative trend analysis result. The slope, S_I of the innovative trend slope is obtained through the following simple formulation.

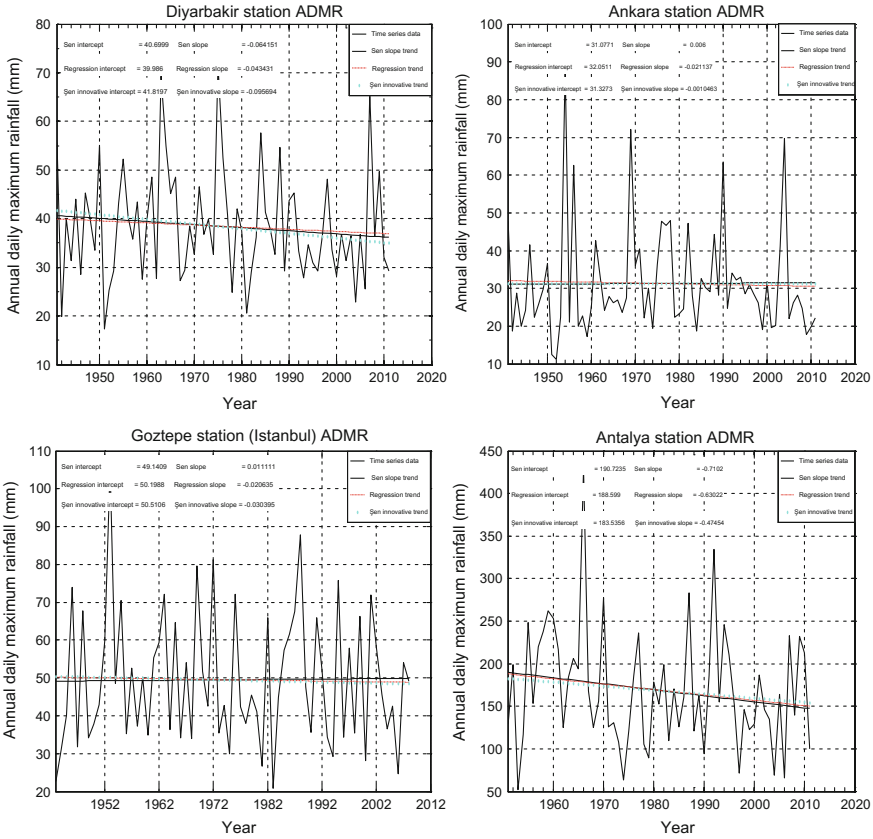


Fig. 7.12 Various trend methodology results

$$S_{\mathrm{I}} = \frac{m_{\mathrm{S}} - m_{\mathrm{F}}}{n/2}, \tag{7.18}$$

where m_{S} and $\mathrm{m_F}$ are the arithmetic averages of the second and first halves of the time series. Visually one can inspect that each method yields almost the same trend within acceptable sampling errors. In all the remaining figures in this chapter ADMR acronym implies annual daily maximum (extreme) rainfall.

Antalya, Diyarbakır, Izmir, and Trabzon stations have significant decreasing trends, whereas Ankara, Erzurum, and Istanbul stations have comparatively weak significant trend components.

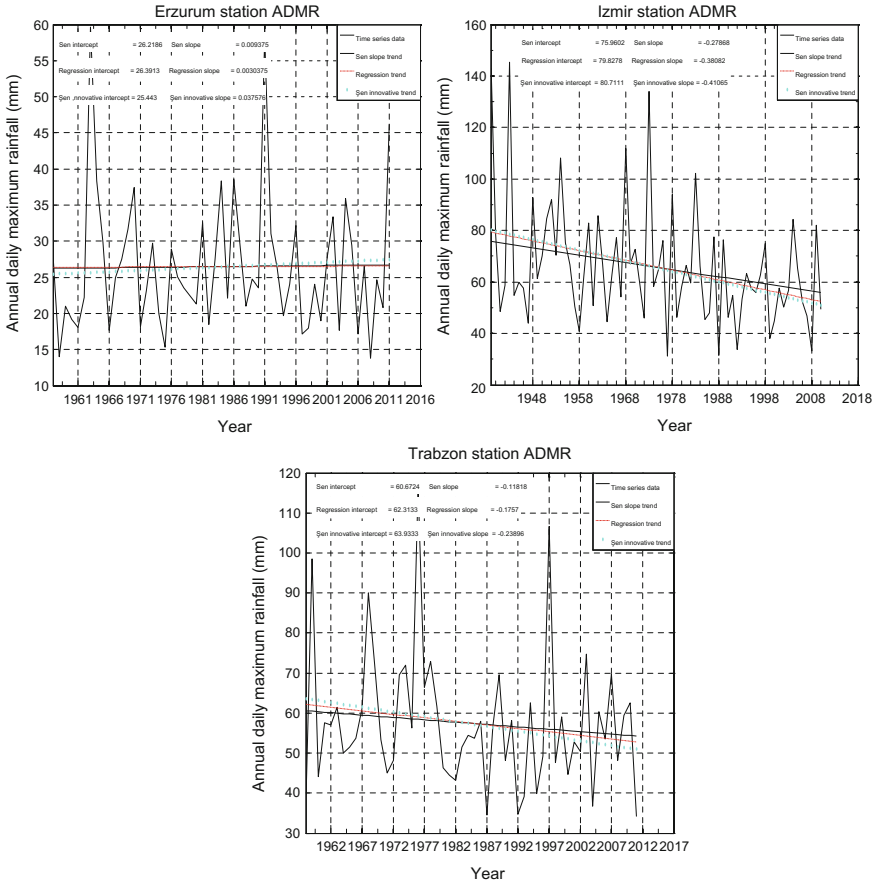


Fig. 7.12 (continued)

7.7 Trend and Variability

In the previous section, different trend methodological lines are presented with comparisons. In this sub-section, another very important feature of the hydro-meteorological and climatological time series is identified, which is the variability. Especially, extreme values with the climate change effects are related to droughts and floods, which may cause variations in the standard deviation. The linear trends are concerned with the average temporal tendencies in a time series, whereas the standard deviation changes are referred to as temporal variability. In general, prior to the climate change impact hydro-meteorological time series were assumed to be first order stationary or second order stationary, which meant that the mean and the variance are constants and do not vary by time. For the variance, this

property of constancy is referred to as the homoscadasdicity in the statistics literature (Benjamin and Cornell 1970). In order to search for the standard deviation variation within the given series, one can think about its division into two equal parts and then compare the standard deviations of each half. If the second half standard deviation is bigger (smaller) than the first one then there is an increasing (decreasing) variation in the time series. Simply, the variability can be measured as the standard deviation change per time (year in this case). If the first and second half standard deviations are S_F and S_S, respectively, then simple expression of the innovative methodology variability, V_I, can be written as,

$$V_I = \frac{S_S - S_F}{n/2} \tag{7.19}$$

If $V_I > 0$ ($V_I < 0$) then there is an increasing (decreasing) variability. In case of increasing (decreasing) variability there is high expectation in the magnitude and frequency of floods (droughts). Innovative trend and variability tendencies are presented in Fig. 7.13 for each station.

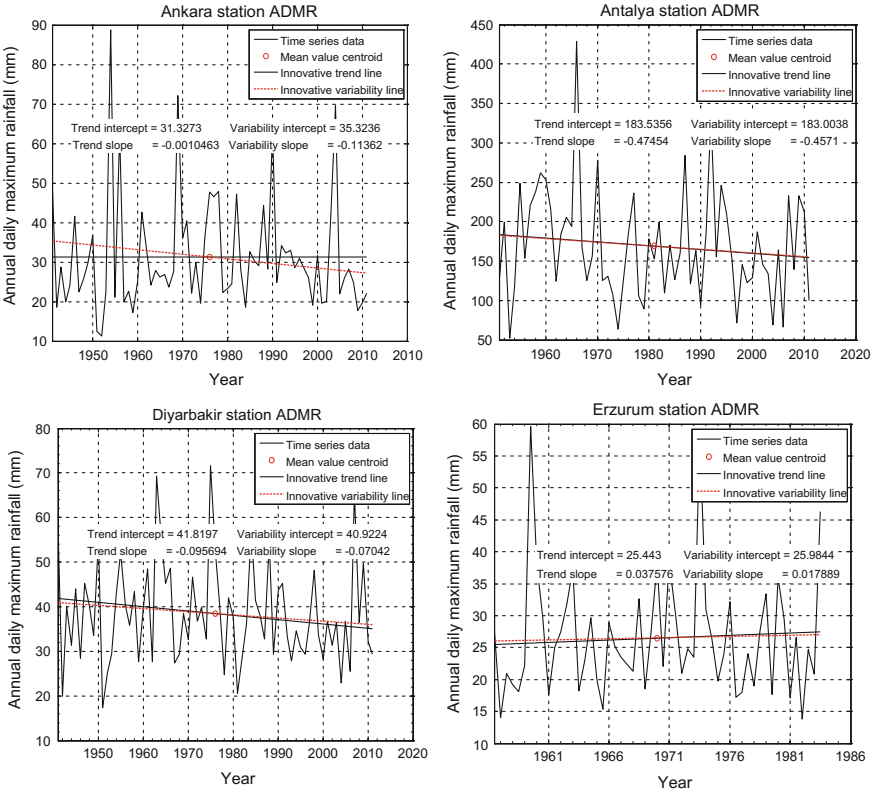


Fig. 7.13 Trend and variability lines

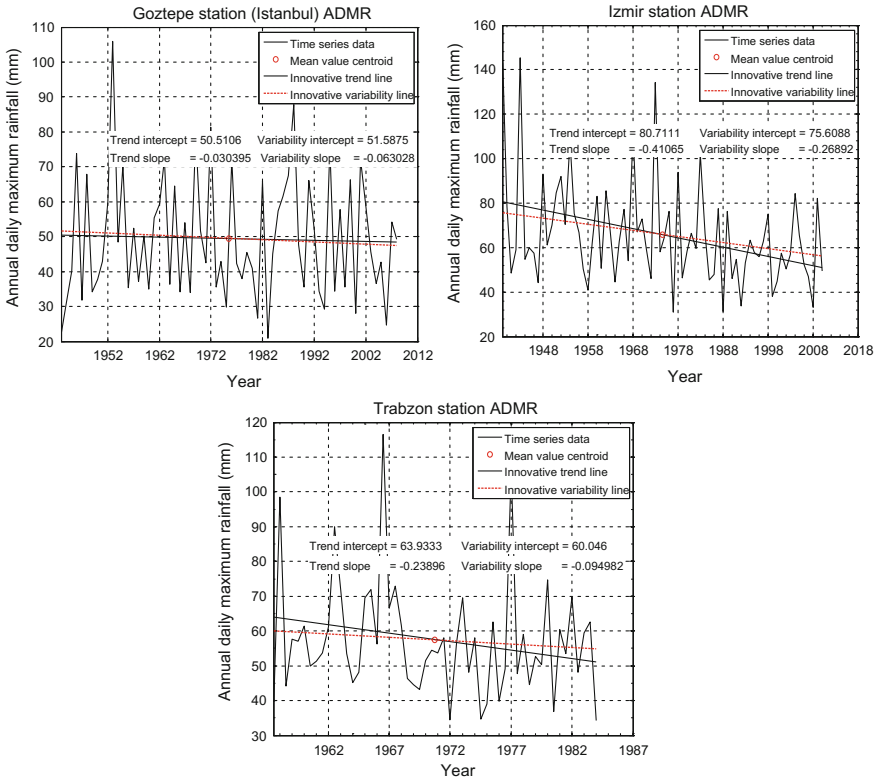


Fig. 7.13 (continued)

Ankara meteorology station ADMR records do not have significant trend (no decrease in the rainfall amounts), but there is a decrease in the variability with time, which implies that in the future more dry and drought periods are expectable in this region. Antalya and Diyarbakır stations have almost the same decreasing trends and variability. Such a situation indicates not only decrease in the rainfall amounts, but additionally expectation of more severe and frequent dry spell possibilities in future. This is in accord with the IPCC reports conclusion that the Eastern Mediterranean area will experience decreases in the rainfall amounts (IPCC 2007, 2013). Erzurum meteorology station ADMR records have increase both in the rainfall amounts and in the variability, which imply expectation of more frequent and severe flood occurrences in this region. Izmir and Trabzon stations are expected to experience significant decreases in the rainfall amounts and to become drier in future. However, in Istanbul (Goztepe meteorology station) water balance is expected to remain steady with time.

7.8 Innovative Trend Template and Significance Limits

These templates provide a rich information about the internal trend structure of a given time series, because they classify "low," "medium," and "high" record values in the first half with comparison to the second half. Such a plot indicates how the trend takes place in the "low," "medium," and "high" record ranges. It is possible to divide objectively the range of variation into three equal parts and then interpret trends in each domain. However, subjectively one can observe these three categories by looking at the innovative trend template.

In Fig. 7.14, the innovative trend template for each meteorology station is given with mean and standard deviation centroids for the first and the second halves.

Innovative trend templates in Fig. 7.14 provide detailed interpretations about the trend components within the ADMR records, where three categories are taken into consideration as "low," "medium," and "high". In these templates, four centroid points are shown for the mean (trend), St. Dev. (variability), first and second half

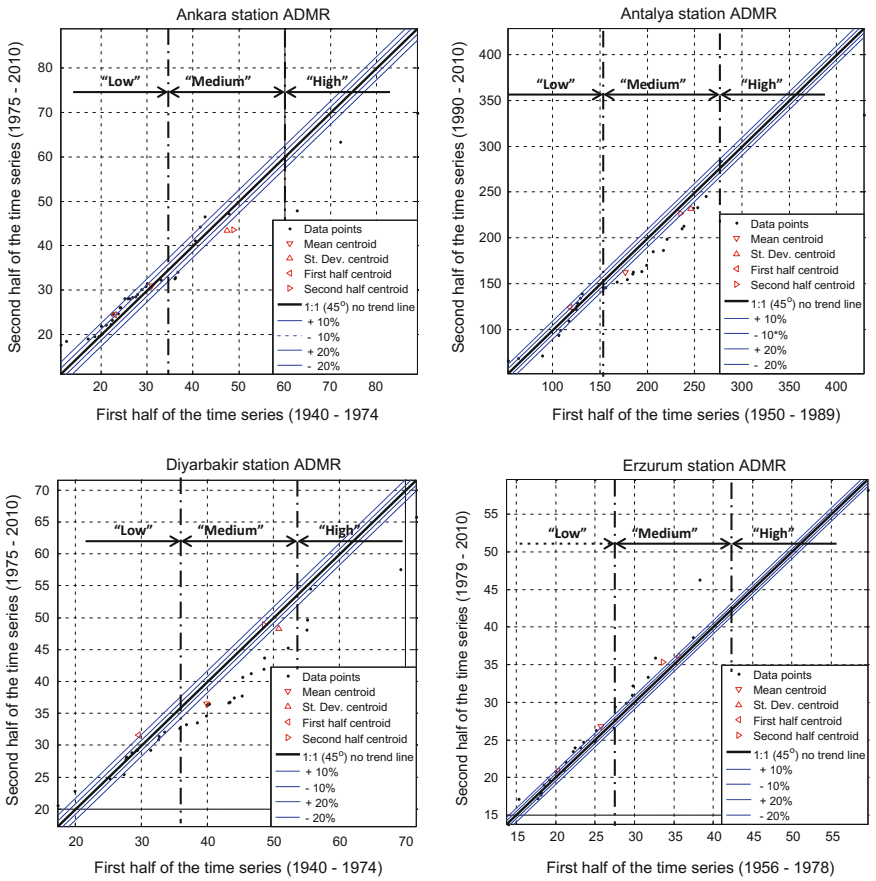


Fig. 7.14 Innovative trend templates

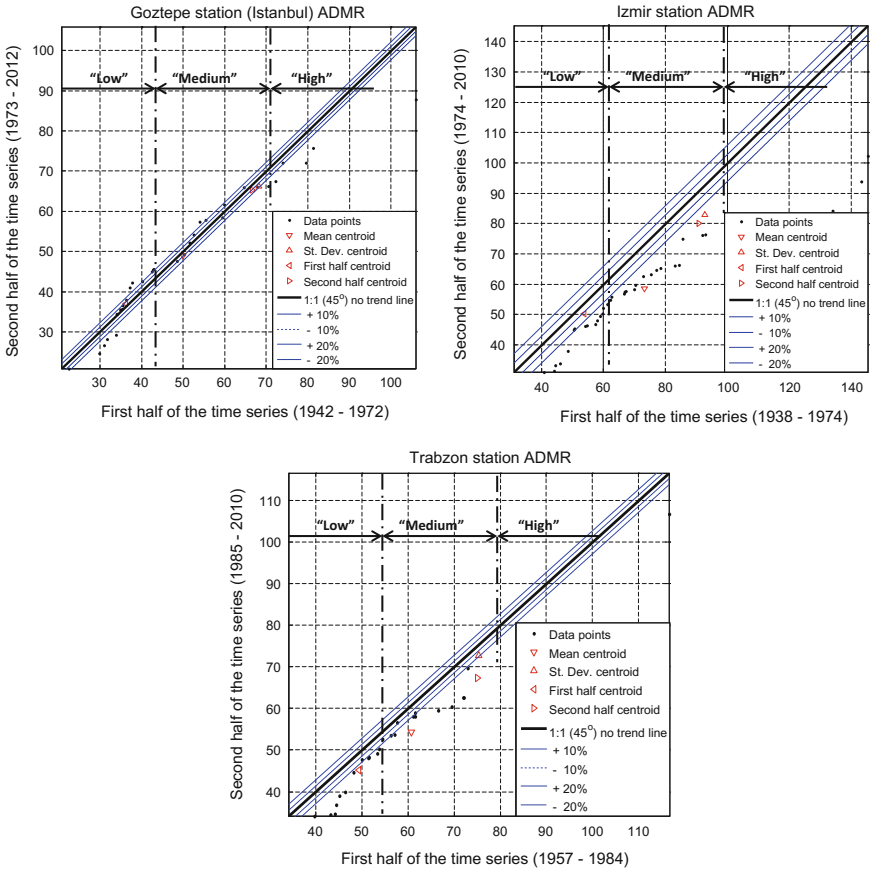


Fig. 7.14 (continued)

series. If the mean centroid is above (below) the 45° (1:1) straight line then there is increasing (decreasing) trend within the whole hydro-meteorologic time series. Similarly, the location of the St. Dev. centroid implies increasing or decreasing variability component. First and second half centroids are for the first and the second half time series trend cases. By taking these four centroid points into consideration, the following interpretations can be deduced from the innovative trend templates.

In general, Antalya, Diyarbakır, Izmir, and Trabzon meteorology station records indicate significant monotonic decreasing trends as obvious from Fig. 7.13, which are confirmed by the corresponding innovative templates also in Fig. 7.14, because the mean centroid for each station is far below the 1:1 (45°) straight line.

In order to base the deviations of each scatter point as well as centroids from the 1:1 straight-line, herein a quantitative significance test is suggested. In case of no trend the scatter points must lie on the 1:1 line, which implies that the

corresponding values in the first and second half series are the same, in other words, their differences is equal to zero. This further implies that the expected value or the arithmetic average of the differences is equal to zero. However, in any given natural hydro-meteorological time series such a perfect case is not valid, and therefore, there will be deviations from 1:1 line. The smaller the standard deviation of these deviations, the closer are the scatter points, and accordingly, the mean centroid points are close to the 1:1 line. Hence, one can calculate the standard deviation, s_d, of the scatter points from the 1:1 line as follows,

$$s_d = \frac{1}{n}\sum_{i=1}^{n/2}\left(X_{i+n/2} - X_i\right)^2 \tag{7.20}$$

The significance test for the innovative trend can then be accomplished by considering the normal pdf with zero mean and standard deviation equal to s_d at a given confidence level. The peak point of such a normal pdf lies on the 1:1 line, and hence, the significance levels must be measured from the peak point as shown in Fig. 7.15. In the same figure, ±10 and ±20% significance levels are also shown explicitly.

In Fig. 7.14, the significance levels, according to this way of calculation, are shown as parallel lines to the 1:1 (45°) no trend straight line. On the same templates in Fig. 7.14, there are also upper and lower confidence limits around the 1:1 (45°) no trend line. Unfortunately, in some applications these confidence lines are taken arbitrarily as nonparallel lines, which is wrong. After all, the important point is to assess the overall deviations of the scatter points from the 1:1 no trend line. For this purpose, the deviations from the 1:1 line are calculated with their mean, m_d, and standard deviation, s_d, values. In general, in the case of no trend, the mean of the

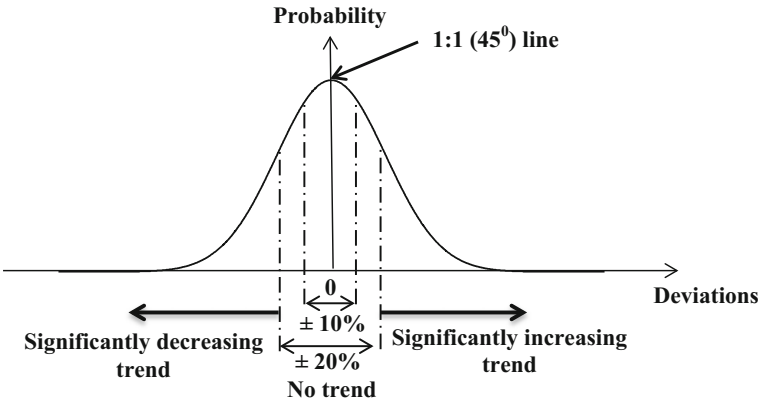


Fig. 7.15 Innovative trend analysis confidence level normal pdf with zero mean and s_d standard deviation

Table 7.5 Confidence limits for the innoıvative trend analysis

Station name	Difference St. Dev. s_d (mm)	Upper limit ±10%	Lower limit ±20%
Ankara	4.9878	±1.2636	±2.6156
Antalya	19.0968	±4.8381	±10.0144
Diyarbakır	3.1710	±0.8034	±1.6629
Erzurum	1.6934	±0.429	±0.888
Istanbul	4.3503	±1.1021	±2.2813
Izmir	11.3671	±2.8798	±5.9609
Trabzon	5.2706	±1.3353	±2.7639

deviations should be equal to zero. The resulting numerical calculations are presented in Table 7.5.

With the confidence limits, the innovative trend templates provide refined interpretations. For instance, at Antalya station scatter points mostly fall within the confidence limits at "low" and "high" ADMR values, but in between all points are outside the limits in the lower triangular part, which implies that the decreasing monotonic trend component comes from the "medium" ADMR data.

At Diyarbakır station, monotonic decreasing trend in Fig. 7.13 is more severe compared to Antalya, but the contributions are from "medium" and "high" ADMR values as obvious on the innovative trend template in Fig. 7.14. Izmir station innovative trend template data scatters imply that all ADMR amounts are in decrease with more dominant influence from "high" values as one can see from the corresponding innovative trend template. Finally, Trabzon station monotonic decreasing trend is under the influence of the "low" and "high" ADMR values. Erzurum station has increasing monotonic trend in Fig. 7.13 and it is obvious that this increase is due to the "medium" and "high" ADMR values, which fall above the 1:1 straight line and outside of the confidence limits in the innovative trend template in Fig. 7.15. At Ankara and Istanbul stations innovative trend templates include almost all the scatter points within the confidence lines parallel to 1:1 (45°) no trend line implying insignificant trend component, which is also in agreement with the monotonic trend components in Fig. 7.13 for each station.

References

Bayazit, M., & Önöz, B. (2007). To pre-whiten or not to pre-whiten in trend analysis? *Hydrological Sciences Journal, 52,* 611–624.

Bao, Z., et al. (2012). Sensitivity of hydrological variables to climate change in the Haihe River basin, China. *Hydrological Processes, 26*(15), 2294–2306.

Benjamin, J. R., & Cornell, C. A. (1970). *Probability, statistics, and decision for civil engineers*. New York: McGraw-Hill Book company.

Burn, D. H., & Hag Elnur, M. A. (2002). Detection of hydrologic trends and variability. *Journal of Hydrology, 255,* 107–122.

Demaree, G. R., & Nicolis, C. (1990). Onset of Sahelian drought viewed as a fluctuation-induced transition. *Quarterly Journal Royal Meteorological Society, 116,* 221–238.

Douglas, E. M., Vogel, R. M., & Kroll, C. N. (2000). Trends in floods and low flows in the United States: Impact of spatial correlation. *Journal of Hydrology, 240,* 90–105.

Douglas, E. M., & Fairbank, C. A. (2011). Is precipitation in Northern New England becoming more extreme? Statistical analysis of extreme rainfall in Massachusetts, New Hampshire, and Maine and updated estimates of the 100-year storm. *Journal of Hydrologic Engineering*, 0.1061/(ASCE) HE.1943-5584.0000303: 203–217.

Ehsanzadeh, E., Ouarda, T. B. M. J., & Saley, H. M. (2011). A simultaneous analysis of gradual and abrupt changes in Canadian low streamflows. *Hydrological Processes, 25*(5), 727–739.

Fatichi, S., Ivanov, V. Y., & Caporali, E. (2013). Assessment of a stochastic downscaling methodology in generating an ensemble of hourly future climate time series. *Climate Dynamics, 40,* 1841–1861.

Gan, T. Y. (1998). Hydroclimatic trends and possible climatic warming in the Canadian prairies. *Water Resources Research., 34,* 3009–3015.

Garbrecht, J., Van Liew, M., & Brown, G. O. (2004). Trends in precipitation, streamflow, and evapotranspiration in the Great Plains of the United States. *Journal of Hydrologic Engineering, 5*(360), 360–367. doi:10.1061/(ASCE)1084-0699(2004)9.

Gupta, A. (2007). *Large rivers: Geomorphology and management*. Chichester, U.K.: Wiley.

Haktanir, T., & Citakoglu, H. (2014). Trend, independence, stationarity, and homogeneity tests on maximum rainfall series of standard durations recorded in Turkey. *Journal of Hydrologic Engineering, 19*(9), 501–509.

Hamed, K. H. (2008). Trend detection in hydrologic data: The Mann-Kendall trend test under the scaling hypothesis. *Journal of Hydrology, 349,* 350–363.

Hamed, K. H., & Rao, A. R. (1998). A modified Mann-Kendall trend test for auto-correlated data. *Journal of Hydrology, 204,* 182–196.

Hamilton, J. P., Whitelaw, G. S., & Fenech, A. (2001). Mean annual temperature and annual precipitation trends at Canadian biosphere reserves. *Environmental Monitoring and Assessment, 67,* 239–275.

Han, J., Huang, G., Zhang, H., Li, Z., & Li, Y. (2014). Heterogeneous precipitation and streamflow trends in the Xiangxi River watershed, 1961–2010. *Journal of Hydrologic Engineering, 19*(6), 1247–1258.

Hannaford, J., & Marsh, T. (2006). An assessment of trends in UK runoff and low flows using a network of undisturbed catchments. *International Journal of Climatology, 26*(9), 1237–1253.

Hipel, K. W., McLeod, A. I., & Weiler, R. R. (1988). Data analysis of water quality time series in Lake Erie. *Water Resources Bulletin, 24,* 533–544.

Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of trend analysis for monthly water-quality data. *Water Resources Research, 18,* 107–121.

Hirsch, R. M., & Slack, J. R. (1984). A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research, 20,* 727–732.

IPCC, (2007). Climate change 2007: Impacts, adaptation, and vulnerability. *Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge, UK: Cambridge University Press.

IPCC, (2012). Edited by Thomas F. Stocker Dahe Qin, Gian-Kasper Plattner Melinda M.B. Tignor Simon K. Allen Judith Boschung, Alexander Nauels Yu Xia Vincent Bex Pauline M. Midgley and Working Group I Technical Support Unit, Working Group I *Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (IPCC), Summary for Policymakers.

IPCC, (2013). Climate change 2013: The physical science basis. *Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge, UK: Cambridge University Press.

IPCC, (2014). Climate change 2014: Impacts, adaptation, and vulnerability. *Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge, UK: Cambridge University Press.

Kalra, A., Piechota, T. C., Davies, R., & Tootle, G. A. (2008). Changes in U.S. streamflow and western U.S. snowpack. *Journal of Hydrologic Engineering, 13,* 156–163.

Kendall, M. G. (1970). *Rank correlation methods* (4th ed.). London: Griffin.
Larsen, J., Ussing, L., & Brunø, T. (2013). Trend-analysis and research direction in construction management literature. *ICCREM, 2013,* 73–82.
Lorenzo-Lacruz, J., Vicente-Serrano, S.M., L´opez-Moreno, J.I., Moran-Tejeda, E., & Zabalza, J., (2012). Recent trends in Iberian streamflows (1945–2005) *Journal of Hydrology, 414/415,* 463–475.
Maass, A., Hufschmidt, M. M., Dorfman, R., Thomas, H. A., Jr., Marglin, S. A., & Fair, G. M. (1962). *Design of water resources systems*. Cambridge, Mass: Harvard University Press.
Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13*(3), 245–259.
Matalas, N. C., & Sankarasubramanian, A. (2003). Effect of persistence on trend detection via regression. *Water Resources Research*, *39*(12), WR002292.
McGill, R., Tukey, J.W., & Larsen, W.A. (1978). Variations of Box Plots. *The American Statistician, 32,* (1), 12–16. doi:10.2307/2683468.
Milly, P. C. D., Julio, B., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, P., et al. (2008). Stationarity is dead: Whither water management? *Science, 319,* 573–574.
Novotny, E. V., & Stefan, H. G. (2007). Stream flow in Minnesota: Indicator of climate change. *Journal of Hydrology, 334*(3–4), 319–333.
Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*, 1379–1389.
Sharif, M., Archer, D., & Hamid, A. (2012). Trends in streamflow magnitude and timings in Satluj River Basin. *World Environmental and Water Resources Congress, 2012,* 2013–2021.
Spearman, C. (1904). The proof and measurement of association between two things. American *Journal of Psychology, 15,* 72–101.
Şen, Z. (2008). *Wadi hydrology* (p. 347). Boca Raton, USA: Taylor and Francis Group, CRC Publishers.
Şen, Z. (2012). Innovative trend analysis methodology. *Journal of Hydrologic Engineering, 17*(9), 1042–1046.
Şen, Z. (2014). Trend identification simulation and application. *Journal of Hydrologic Engineering, 19*(3), 635–642.
Taylor, C. H., & Loftis, J. C. (1989). Testing for trend in lake and groundwater quality time series. *Water Resources Bulletin, 25,* 715–726.
von Storch, H. (1995). Misuses of statistical analysis in climate research. In H. V. Storch & A. Navarra (Eds.), *Analysis of climate variability: Applications of statistical techniques* (pp. 11–26). New York: Springer-Verlag.
Wagesho, N., Goel, N. K., & Jain, M. K. (2012). Investigation of non-stationarity in hydro-climatic variables at Rift Valley lakes basin of Ethiopia. *Journal of Hydrology, 444*, 113–133.
Yu, Y. S., Zou, S., & Whittemore, D. (1993). Non-parametric trend analysis of water quality data of river in Kansas. *Journal of Hydrology, 150,* 61–80.
Wilson, J.W., & Atwater, M.A. (1972). Storm rainfall variability over Connectituc. *Journal of Geophysical Research,* 77, 3950–3956.
Yue S, & Wang, C. Y. (2002). Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test. Water Resources Research, *38*. doi:10.1029/2001WR000861.
Yue, S., Pilon, P., & Cavadias, G. (2002). Power of Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology, 259,* 254–271.
Zhang, X., Harvey, K.D., Hogg, W.D., & Yuzyk, T.R., (2001). Trends in Canadian streamflow. *Water Resources Research, 37* (4), 987–998.

Partial Trend Detection

8

Abstract

One of the very important issues in trend search is whether there are partial trends at different positions within a given time series? This point is also not considered frequently in trend analysis studies, because most of the time a monotonic and holistic trend is searched initially in the given time series. However, within the same time series apart from the monotonic trend there may be local trends that may indicate significant changes that may be needed for natural of artificial explanations. For the search of partial trend possibilities within a time series innovative trend approaches explained in the previous chapters are applied with a slight modification for partial trend search.

Keywords

Innovative · Group · Partial · Piecewise · Qualitative · Simulation

8.1 General

In the previous chapters, trend and variability features are identified and determined by taking into consideration the whole record length irrespective of what might possible be subtrend or sequences of subtrends. In practical application, from the increasing or decreasing holistic monotonic single trend along the whole record length, there may be some subtrend durations at the record site or areas within the study areas. Some time series may have a series of subtrends in increasing or decreasing manner.

As already mentioned in the previous chapters, there are different quantitative trend identification methodologies for monotonic, consistent, and unidirectional tendencies in a given time series. Among these methods are parametric linear

Z. Şen, *Innovative Trend Methodologies in Science and Engineering*,
DOI 10.1007/978-3-319-52338-5_8

regression approaches, rank-based nonparametric procedures and their mixtures. Parametric methodology includes a set of assumptions as normal (Gaussian) probability distribution function (pdf) of given time series or residuals, and nonparametric tests require serially independent time series, which are rarely available in climatological records. Mann–Kendall (MK) trend (Mann 1945; Kendall 1975) test is in frequent use for identifying significant trend by numerous researchers (Gan 1998; Hamed 2008; Hamed and Rao 1998; Douglas et al. 2000; Ventura et al. 2002; Burn and Hag Elnur 2002; Yue and Wang 2004; Yue et al. 2003; Mossmann et al. 2004; Modarras and de Silva 2007; Luo et al. 2007; Karpouzos et al. 2008; Chen and Xie 2005). The presence of serial correlation in a time series affects the validity of the MK test. The positive autocorrelation increases trend detection probability when actually there is no trend, and vice versa (Yue et al. 2002, 2003). Although it is a well-known point, few studies have addressed this issue, and autocorrelation in the data is either ignored or eliminated by pre-whitening procedure (Bayazıt and Önöz 2007).

On the other hand, Wayne et al. (1995) considered the problem of determining the upward (increasing) trending behavior in the global temperature anomaly series. To address this issue a unit-root test is also examined. Another serial-correlation–robust trend test is suggested by Alexandersson and Moberg (1997), which controls for the possibility of spurious evidence due to strong serial correlation and it is valid whether the errors are stationary or have a unit-root (strong serial correlation). Another attractive point in the same test is that it does not require estimates of serial-correlation nuisance parameters. According to work by Thomas and Timothy (2002) strong serial correlation (or a unit-root) in global temperature data could, in theory, generate spurious evidence of a significant positive trend. Coggin (2012) used the concept of unit-root trend analysis in the climatology literature for testing trends in the HadCRUT3 global and hemispheric data. It is illustrated that recently developed econometric trend tests using the HadCRUT3 global and hemispheric surface temperature data updated through 2009, specifically allow statistical complications of structural change, serial correlation, and unit-roots. The use of unit-root trend analysis is not yet commonly used in the climatological studies.

There are already suitable parametric statistical techniques for trend analysis. For instance, regression quantile plots can be used for this purpose (Koenker 2004). Nalley et al. (2013) tried to detect trends in the mean surface air temperature over southern parts of Ontario and Québec, Canada, for the period of 1967–2006 using the discrete wavelet transform technique. They showed that the positive trends observed for the annual data are thought to be mostly attributable to warming during winter and summer seasons, which are manifested in the form of multiyear to decadal events (mostly between 8 and 16 years).

In general, a linear trend is fitted without any distinction among “low”, “medium”, and “high” climatological values. Although, the theoretical basis of the partial trend methodology is given by Şen (2012, 2014) for such distinctive trend identifications, but objective identification and determination of the partial trend lines is not available in these works, which is the subject of this chapter. The methodology presented is a nonparametric approach and provides visual inspection of simply

scattered points on the innovative trend template. Simple formulations are provided for arithmetic average and standard deviation (variability) trend slope calculations. The proposed methodology is applied to state-wise annual precipitation and temperature records from New Jersey, USA.

Trend determination and de-trending implementation procedures are important steps in any time series analysis. Most research articles in the literature are attached with stationary time series embedded with monotonic linear trend within data time span. However, in this chapter, logical rules are presented for trend determination over a certain span of data with a simple mathematical procedure for any nonlinear and nonstationary types. With this definition of trend, the variability of the data on various time scales also can be derived naturally. After the determination of the trend then de-trending procedure can be initiated with remaining residuals that have zero arithmetic average.

Trending and de-trending procedures are frequently used in many applications but most often trending is mentioned with offset of de-trending. In many natural, environmental, economic, social and especially climate change impacts the search for trend constitutes the most critical quantity. In any statistical analysis, the correlation coefficient calculation and spectral analysis can yield meaningful and unbiased results for time series that are free of trends, and therefore, de-trending procedure is of prime importance.

The application of statistical analyses in numerous scientific, social, economic, medical, and engineering disciplines, the trend is the tendency over the whole time series domain that cautiously con be extended to future time spans. The trending leads to residual series with zero arithmetic average, which can be modeled by a convenient stochastic process. In the literature, there is lack of trend search in nonlinear and nonstationary time series, and consequently, also the suitable application the corresponding de-trending operation. The application of the usual trend analysis leads to the awkward conclusions and trend determination and de-trending procedures remain often as ad hoc operations. These unwanted situations arise greatly because of the difficulties concerning trend stem from the lack of a proper definition for the trend in nonlinear nonstationary time series. It is, therefore, a definitive and quantitative necessity to develop methodologies for trend and de-trending procedures in such time series. It is the main purpose of this chapter.

Herein, a definition of trend variability is introduced concerning not only on the change usually mentioned in all the previous trend determination approaches, but also on the changes in the variance (or standard deviation) distinctively from the classical definition of the sole variability, which is expressed in terms of range, interquarters, and standard deviations only. Because the classical definition of the statistical variability is concerned with standard deviation changes over the time series record span. The procedures presented in this chapter are quite general and can be applied to any time series from nonstationary and nonlinear processes. Volatility is the term used instead of variability in financial communities. Engle and Granger emphasized that models for market prediction provide a daunting challenge for a patently nonstationary process. Furthermore as a justification they regard the

financial market as a special Auto-Regressive Integrated Moving Average (ARIMA) process, controlled by a series of shocks and relaxations. They clearly pointed out the limitation of their works, that not all nonstationary data satisfy their special assumptions. Indeed, the vast majority of real-world data are of a nonstationary and nonlinear nature and do not fit the ARIMA prediction models at all.

In practical applications without any background considerations, provided that there is a time series, the researcher applies the most usual trend analysis with the simplest mathematical formulation as a best-fit linear trend over the whole time series span. Of course, the best fit according to the least squares analysis supported by regression methodology yields trend such that the residuals have zero arithmetic average. It must be kept in mind that such a trend is valid only in cases of a purely linear and stationary time series. This approach may not be suitable or leads to illogical and physically meaningless applications, which is the case especially in climate change time series analysis. If the underlying generation mechanism has nonlinearity and nonstationary features then monotonic linear trend fits makes little sense.

In many preliminary time series applications, the trend emergence is sought after the application of moving-average procedure with a certain span of application starting from the beginning of the given time series. The moving average procedure smoothens the ruggednesses along the time series, and therefore, may catch trend visualization without providing any mathematical trend function. The main problem in the moving-average procedure is the predomination of the applicable time span. Determination such a time span does not have any logical or rational base, and especially, for in nonstationary processes the local time scale is unknown a priori.

On the other hand, other complicated trend identification methods including the classical regression analysis or Fourier-based filtering procedures are based on stationarity and linearity assumptions. These assumptions remain as hindrances in their applications. Even though the trend calculations from a nonlinear regression happens to fit the time series data satisfactorily, there still is no justification to select a time-independent regression formula and apply it to globally for nonstationary processes.

8.2 Qualitative Partial Trend Methodology

The frequency, duration, extent, and intensity of extreme events are related to input variable characteristics of the study. Extreme event frequencies are bound to increase in different sectors. For instance, a key meteorology variable to start for the possible effects of such consequences on the society at large is the precipitation occurrences and amounts along the time axis, whereas knowing the point temporal features one is capable to make spatial variation features through convenient software by mapping. The features of climate, droughts, floods, and normal values are all hidden in a given time series of precipitation, which should be identified

according to the main purpose through scientific methodologies. For the last three decades, the global warming and climate change impacts are almost everywhere, scientific, political, or social media, which should be dealt with proper and rational methodologies so as to extract useful information from the past records and make sensible and useful projections for future in order to achieve more efficient and sustainable water resources management. Climate change impacts first will appear on water resources and subsequently agriculture and food security. These are the vital and undeniable activities for the survival of the living creatures on the earth. The basic equation for proper management of these activities necessitates consideration of water balance equation where the precipitation is the main input. Climate change will also affect the water balance equation (water budget) to a negative extent at places where decrease of precipitation and subsequent runoff reductions are expected. Although there are many studies in the literature that deal with a particular component of water balance such as precipitation, streamflow, groundwater, evapotranspiration low and peak flows, unfortunately long-term water balance situation is not well examined under the light of expected climate change effect.

In any climate change study trend determination in the meteorology time series gains utmost importance so as to decide objectively whether the trend is neutral, increasing, or decreasing. In many parts of the world, decreasing trends are felt even without any methodology subjectively by local people, because during at least 30 years of age each one gains experience especially about the climate and to a certain extent meteorological features of his region. Likewise, historical records hide the general climatic features in terms of serial dependence, probability distribution functions (pdf), seasonality, and trends. It is, therefore, necessary to bring out objectively each one of these features depending on the aim of the study. Since, in this paper the main purpose is the low and high precipitation identification in search for climate change effect, the only component for focus is the trend. Classical methodologies search for the existence of trend through rank-based Mann–Kendal (MK) (Mann 1945) approach or according to Spearman's rho in addition to Sen (1968) trend slope method. These objective approaches necessitate a set of assumption validity in the historical records, which are almost impossible to have naturally. In this case artificially, the original time series is subjected to various transformations so as to satisfy the basic assumptions. For instance, one of the assumptions is the independent serial-correlation structure of the historical time series. In order to achieve independent structure, pre-whitening procedure is suggested and applied by different researchers (Yue and Wang 2004; Bayazıt and Önöz 2007). The MK test power has been investigated by Yue et al. (2002) on the bases of sample size, trend slope, and type of probability distribution function (pdf). They have shown that the MK test has the same power as other methods used for the same purpose, for instance Spearman's rho test. After the identification of trend component in a given time series then its mathematical expression is obtained in a straight-line form covering the whole record duration using the linear regression methodology, which also brings additional restrictive assumptions into the study.

The general approach in the literature is the application of the classical trend analysis techniques with trend identification for complete record duration as a linear function. Such an approach provides a holistic trend analysis, whereby one cannot compare possible partial trends that exist within the same time series. For this purpose, in this report a new approach presented by Şen (2012) is applied, which powers one to identify trends attached with low and high precipitation values. The application of the methodology is effected for almost 30 meteorology station precipitation historic records from seven different climate regions of Turkey. This application provides an example for future similar works at any part of the world. Necessary graphs each with low and high precipitation trends are presented with neutral components including the medium values also. It is observed that although there are high precipitation trends at the same location, low precipitation trends are more severe, extensive, longer, and persistent than high precipitation occurrences and amounts. This provides also a means to make regional high and low flow maps for the whole region, which provides a basis for better management of water resources.

8.3 Previous Works

The most rational way to better understand climate change and variability is through trend analysis. The most commonly used estimation method for simple linear trends is the simple regression analysis, which helps to detect the most straightforward assessment of the long-term behavior of a time series in climate change studies. However, in real-world time series a single monotonic linear trend component may not reflect the reality. Since, the simple monotonic linear regression over the record time span does describe the inner structure of change in the time series, its results may be misleading, because it ignores the existence of significant changes (turning points) in the slope of the linear fit, called breakpoints. Especially, for climatic data analyses such simple linear trends may be illogical and physically meaningless, with little sense. The real variability for the underlying mechanisms of global climate change are likely to be nonlinear and nonstationary, so other methods of time series analysis might be advisable. In particular, linear trend does not adequately describe low-frequency behavior of temperature time series. Piecewise regression model fits a nonlinear function with a nonconstant rate of change, and has been applied to analyze time series of different climatic variables to detect breakpoints in linear trends. Karl et al. identified the timing of change points in global temperature time series by minimizing the residual sum of squares of all possible combinations of four line segments representing time intervals of 15 years or more. Tome and Miranda adapted that fitting method to develope an algorithm for fitting a continuous regression model with several breakpoints to data and then it was applied local changes in temperature, precipitations and the NAO

index in Portugal. Liu et al. used the same method to find partial trends of wind variability in the mesosphere and the lower thermosphere over a local observatory at Collm, Germany. Piecewise regression is a method of regression analysis where the response variable is split in two or more intervals, and a line segment is fitted to each interval, with the constraint that the regression function will be continuous. Each line is connected at an unknown value called breakpoint. Piecewise regression is suitable for situations where the response variable shows abrupt changes within a few values of the explanatory variable. This flexible regression method is scarcely used in the analysis of long-term trends of climatic variables, though in many cases it offers a better fit to the records, and shows better compliment with the assumptions of regression analysis.

In scientific and technological studies of time series variability search linear monotonic trend identification is a very common approach. However, in case of long climate time series such a single monotonic trend determination has little relevant significance. In this chapter another innovative piecewise trend methodology is presented.

In the literature there are piecewise linear trend fitting procedures for finding overall trends, and, simultaneously, for computing a new set of climate parameters: the breakpoints between periods with significantly different trends as by Toma and Mirande on the basis of a least squares approach to compute the best continuous set of straight lines that fit a given time series, subject to a number of constraints on the minimum distance between breakpoints and on the minimum trend change at each breakpoint.

During the last three decades a number of researchers have sought and discussed long-term linear tendencies of climate parameters including precipitation, temperature and the NAO index. Prior to analytical mathematical solutions for trend in a time series it is recommended in this book that even an eye inspection may reveal to a certain extent the possibility of a trend, which may be determined later on by a straight-line fitting, and furthermore one can identify also the possibility of a set of trends over different span ranges in the same time series. Karl et al. suggested different approaches, the first one, based on Haar Wavelets, which was able to identify discontinuities in the time series, and also the minimization of the residual sum of squares of all possible combinations led to segments with different trend components in sequence. Another methodology was devised instead of arbitrarily fixing the number of line segments, after an eye inspection of the time series, the number and location of the breakpoints are simultaneously optimized. Such a methodology computes the best combination of continuous line segments that minimize the residual sum of squares subjected to a pair of conditions.

(1) The interval between breakpoints must equal or exceed a given value,
(2) Two consecutive trends must obey one or more imposed conditions.

In the methodology suggested by Toma and Mirande, a continuous embedded curve made in the time series is considered as four straight-line segments. The time series, Y_i (i = 1, 2, …, n) with n number of values is considered piecewise as four epochs.

$$Y_1, \ldots, Y_{sb}, \ldots, Y_{tb}, \ldots, Y_{fb}, \ldots, Y_n$$

where Y_{sb}, Y_{tb}, and Y_{fb} are the beginning time series values for the first, second, and third epoch, respectively. These end values are at the locations of breakpoints in the time series. The procedure is then to fit a linear regression straight line to each epoch values as follows:

$$Y_{i=}a_1 t + c_1 \quad \text{for } i = 1, \ldots, Y_{sb} \tag{8.1}$$

$$Y_{i=}a_2 t + c_2 \quad \text{for } i = Y_{tb} + 1, \ldots, Y_{tb} \tag{8.2}$$

$$Y_{i=}a_3 t + c_3 \quad \text{for } i = Y_{tb} + 1, \ldots, Y_{fb} \tag{8.3}$$

and

$$Y_{i=}a_4 t + c_4 \quad \text{for } i = Y_{fb}, \ldots, n \tag{8.4}$$

The sequence of linear trends must have continuity with each other, and for this purpose the following conditions are taken into consideration in the piecewise regression analyses.

$$c_2 = c_1 + (a_1 - a_2)t_{sb} \tag{8.5}$$

$$c_3 = c_1 + (a_1 - a_2)t_{sb} + (a_2 - a_3)t_{tb} \tag{8.6}$$

$$c_3 = c_1 + (a_1 - a_2)t_{sb} + (a_2 - a_3)t_{tb} + (a_3 - a_4)t_{fb} \tag{8.7}$$

Provided that the breakpoints are inspected or suggested then five unknown parameters c_1, c_2, c_3, c_4, and c_5 can be solved from a system of five equations. For this purpose, the partial derivatives of the sum of square differences between the fit and the measurements. The solution of the system is given by Tomé and Mirande (2004), who have applied the methodology to Azores December minimum maximum temperature and NOI index records. The results are shown as example in Fig. 8.1 for the two cases.

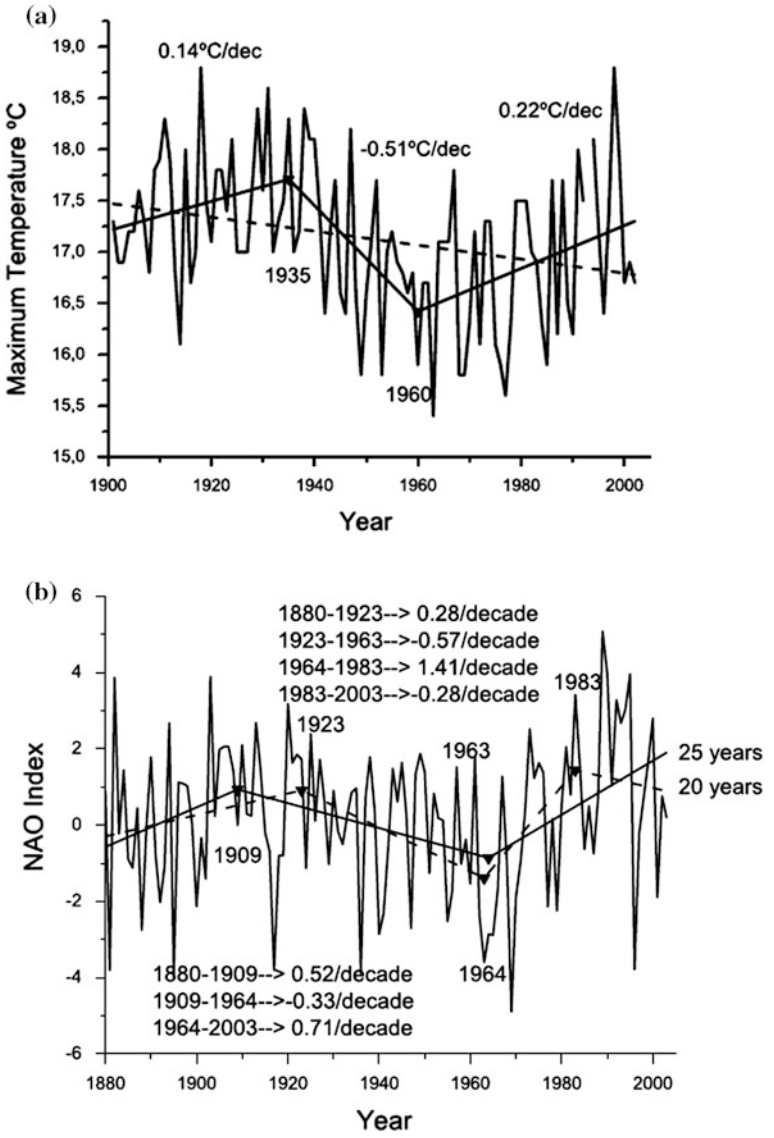


Fig. 8.1 **a** Maximum December temperature at Angra do Heroismo (Azores), breakpoints (1935 and 1960), partial tendencies, in _C/decade, and the linear trend (*dashed line*), **b** Piecewise linear fitting of the NAO index for a minimum period between two breakpoints of 20 years (*dashed line*) and 25 years (*full line*), for the condition of signal change between consecutive trends

8.4 Innovative Piecewise Trend Analysis

Most of the trend identification algorithms provide monotonic and holistic trends in the sense that the whole record period is taken into account. However, there may be partial trends over certain subperiods of a given time series. For instance, in Fig. 8.2 successive and nonoverlapping 30-year period trends are shown for the global annual temperature records. Starting from 1880, the trends are given for 1880–1910, 1911–1940, 1941–1970, and finally from 1971 to the end of the available record. Such partial trend graphs are very useful for appreciation of what have happened along 30-year base periods throughout the record. In Fig. 8.2 there have been a cold base period during 1880–1910 and then onwards there is continuously increasing temperatures, which had comparatively very small increments during the third 30-year base period, but after 1970 the rate of temperature increase is steadily continuous.

Another example is given in Fig. 8.3 for longer record duration concerning Danube River annual discharge values starting from 1840 up to 2005. In this figure there are five different partial trends each for 30-year duration except the last one. A visual inspection for monotonic and holistic trend over the whole record length may give the impression that there is a very slightly decreasing trend. However, 30-year base period results provide more detailed information about what have happened in the history of the Danube River discharge adventure. In the same figure, the intercept and slope values for each partial trend are also given.

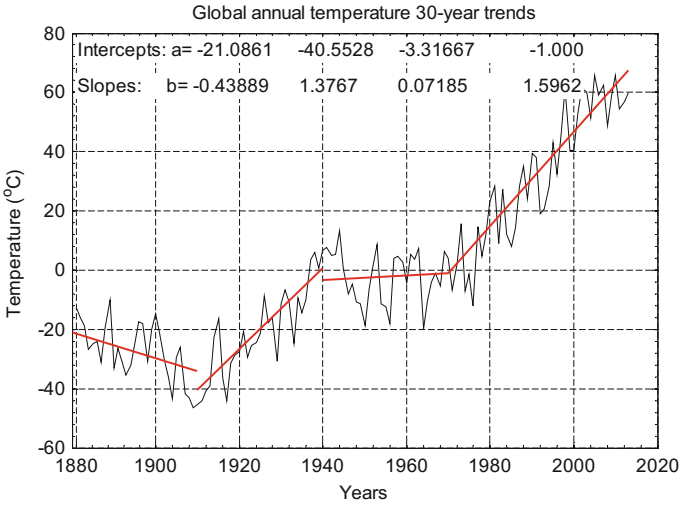


Fig. 8.2 Global annual temperature piecewise trends

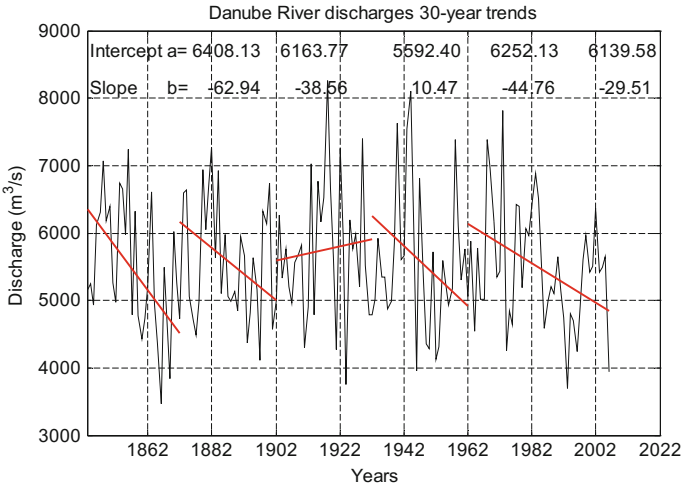


Fig. 8.3 Danube river discharge piecewise trends

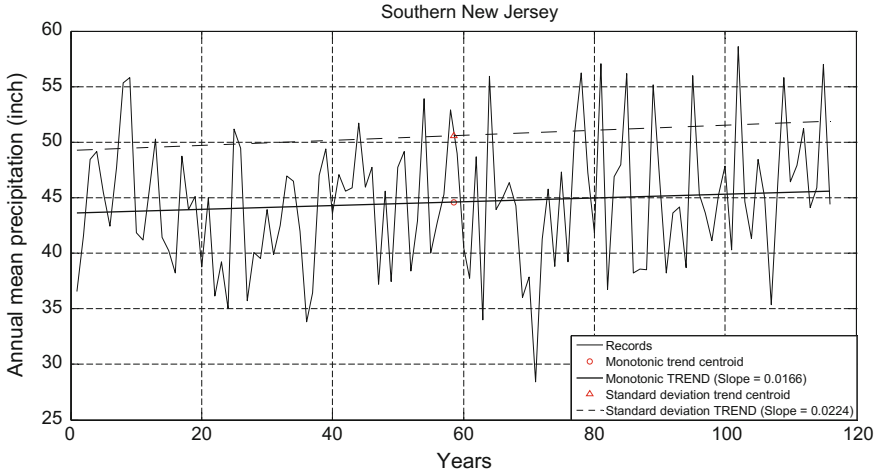


Fig. 8.4 Trends on the average and standard deviation

Another example is given this time based on the precipitation records in Southern New Jersey with records from 1895 to 2006. Figure 8.4 indicates monotonic and holistic trends on the average and also on the standard deviation level. Both parameters, i.e., arithmetic average and standard deviation have increasing trends. Hence, the underlying time series is not stationary as for these two parameters are concerned. Unfortunately, in many conventional studies trend analysis is for the averages only without any consideration of trend possibilities in the standard deviation.

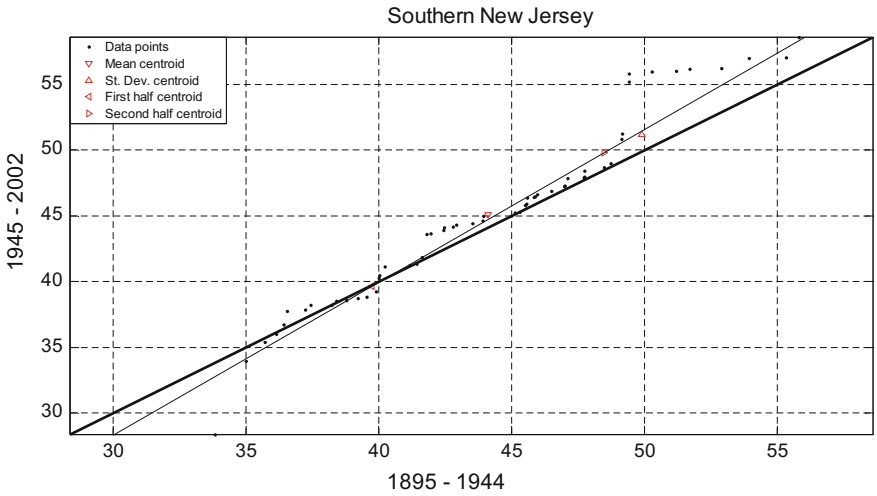


Fig. 8.5 Innovative trend templates for the average and standard deviation

In order to appreciate the possible trend possibility on the average and standard deviation levels innovative trend template is prepared and presented in Fig. 8.5.

A close inspection of this figure indicates that the second half of the record's centroid position stays over the 1.1 (45°) straight line and therefore there is an increasing trend as indicated in Fig. 8.4. On the other hand, the standard deviation centroid also is above the 1:1 (45°) straight line, and hence, there is also increasing trend in the standard deviation level. Trends in the standard deviation levels imply variations in the deviations from the arithmetic average level.

In Fig. 8.6 10-year partial trend sequences is given, which shows different periods of increasing and decreasing trend cases in the past. Decreasing trends show dry spell and drought periods during which there may have been water shortages or scarcity. However in the last two decades there are increasing trends.

In order to appreciate, evaluate, and interpret partial trend elements one can look for a set of different durations such as 10-year, 10-year, 30-year, 50-year, etc., trend behavior on the innovative trend template. Figures 8.7 and 8.10 are given for these durations respectively.

Figure 8.7 indicates that in the middle range of precipitation data 1971–1980 duration has been with the biggest increasing trend, and in the same duration mostly 1961–1970 period had decreasing the biggest trend.

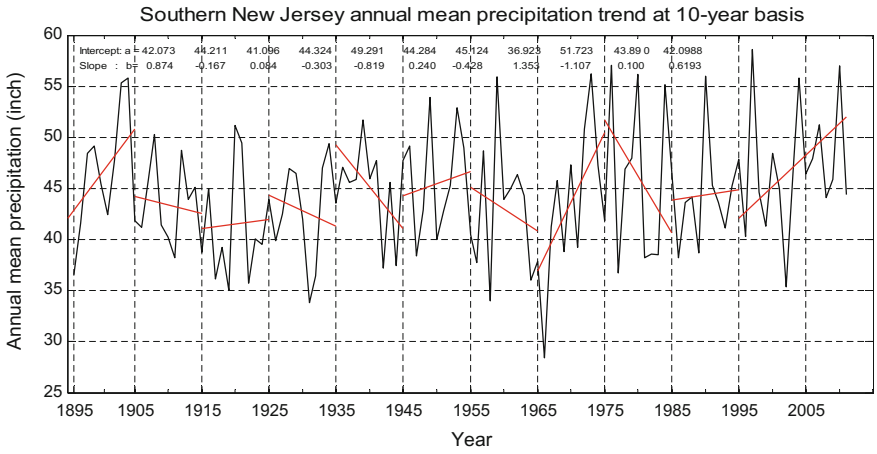


Fig. 8.6 New Jersey precipitation 10-year piecewise trends

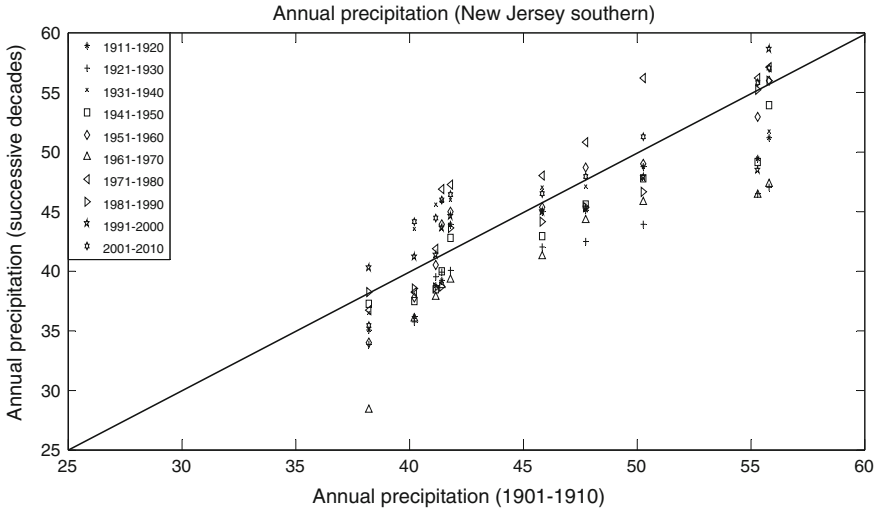


Fig. 8.7 10-year innovative trend templates

According to Fig. 8.8 1941–1960 duration did not have any significant trend because all the points are very close to 1:1 (45°) straight line. As for the low and medium precipitation variations are concerned, they are comparatively smaller than high values. The most extreme wet (dry) spell event has occurred more (less) than the usual during 1981–200 (1921–1940) periods.

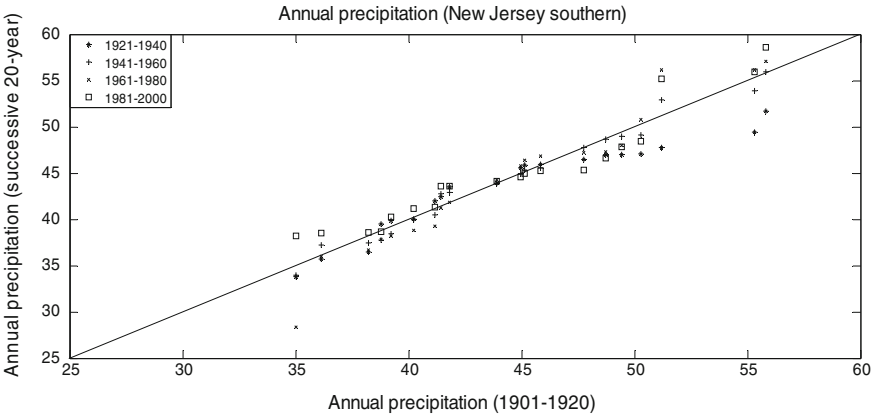


Fig. 8.8 20-year innovative trend templates

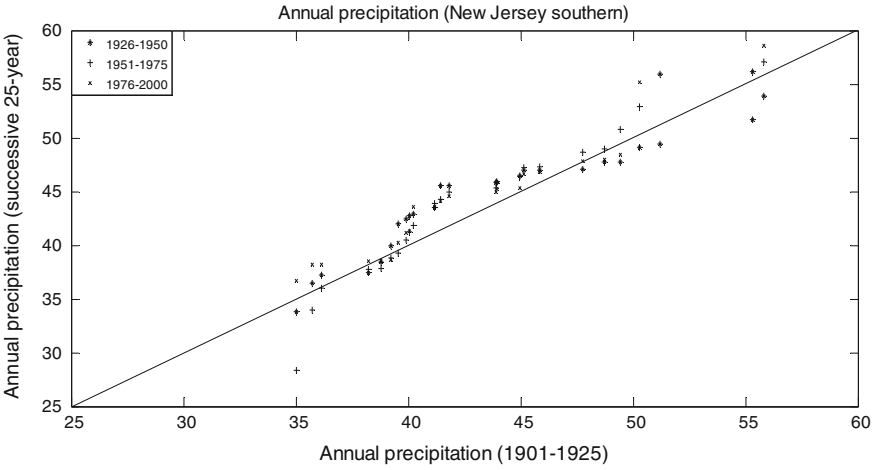


Fig. 8.9 25-year innovative trend templates

Twenty-five-year trends in Fig. 8.9 indicate that extreme precipitation occurrences were in periods of 1951–1975 and 1976–2000. Also, the least occurrences were in these periods.

In Fig. 8.10 medium range precipitation values did not show and significant trend, but especially high precipitation events are along the increasing direction.

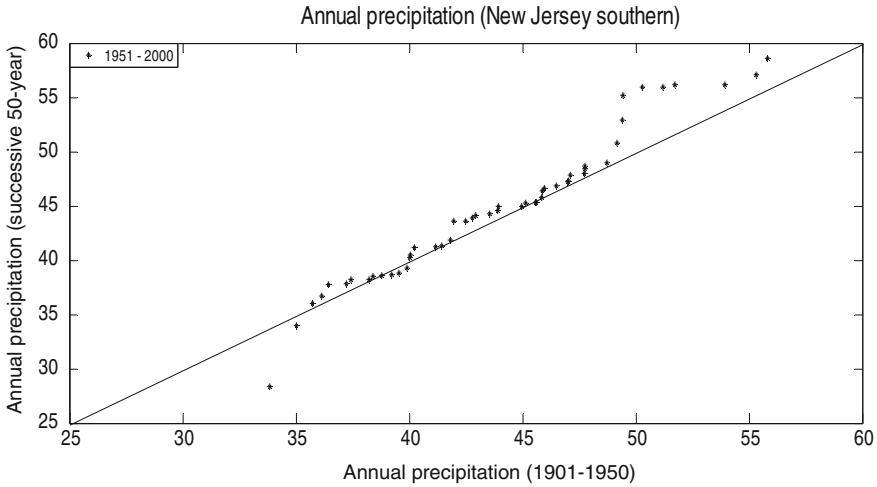


Fig. 8.10 25-year innovative trend templates

8.5 Innovative Trend Template

The main idea is to divide any given time series, $X_1, X_2, \ldots, X_n$ with n elements, into two mutually exclusive subseries of equal lengths as $X_1, X_2, \ldots, X_{n/2}$ and $X_{(n/2+1)}, X_{(n/2+2)}, \ldots, X_n$ with $n/2$ elements in each. Subsequently, each subseries is sorted in ascending order and then plotted against each other leading to a scatter diagram on the Cartesian coordinate system as in Fig. 8.11.

In general, with respect to 1:1 (45°) line this template has three parts (Şen 2012). The main diagonal, 1:1 (45°) straight line presents no-trend line case; the upper (lower) triangular area is for increasing (decreasing) trends.

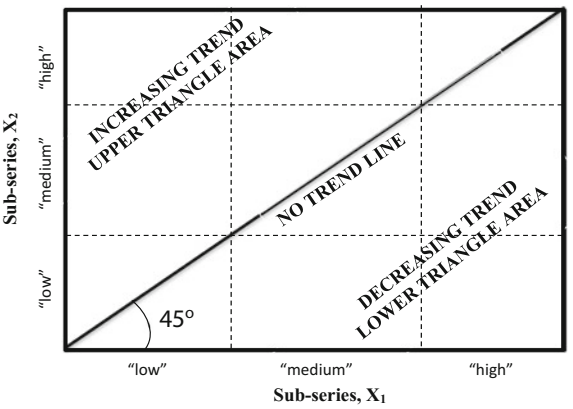


Fig. 8.11 Trend description and identification template

Furthermore, the same template can be visualized as 9-square subareas in accordance with three classifications as “low”, “medium”, and “high” values and these subarea templates are “low”-“low”, “low”-“medium”, “low”-“high”, “medium”-“low”, “medium”-“medium”, “medium”-“high”, “high”-“low”, “high”-“medium”, and “high”-“high”. It is possible to state that the partial trend identification methodology yields information about the “low”, “medium”, and “high” categories of the first half subseries coupled with “low”, “medium”, and “high” categories of the other half leading to the following conclusions.

(1) If the scatter points are completely above (below) the 1:1 line, then there is an increasing (decreasing) trend,
(2) In case of a single increasing (decreasing) trend, the scatter points fall on a parallel straight line to 1:1 line,
(3) If the scatter points have different positions within “low”, “medium”, and “high” subareas then there are different partial trends in the time series,
(4) The proposed methodology provides interpretive information about the “low”, “medium”, and “high” subarea trends and their relative positional inferences,

Figure 8.12 provides a set of trends that one may encounter during the application of the innovative partial trend methodology.

In Fig. 8.12, the circle on each line indicates the arithmetic average point (centroid) of the two halves. This figure provides qualitative trend interpretation possibilities, some of which are summarized as follows:

(1) Extensive straight line (B or C) parallel to the 1:1 (45°) line implies a monotonic trend (increasing or decreasing); line A indicates no-trend case. Other shorter straight lines (D, E, I or L), parallel to the 1:1 (45°) line, are for partial trends that cover different classifications (“low”, “medium” or “high”).

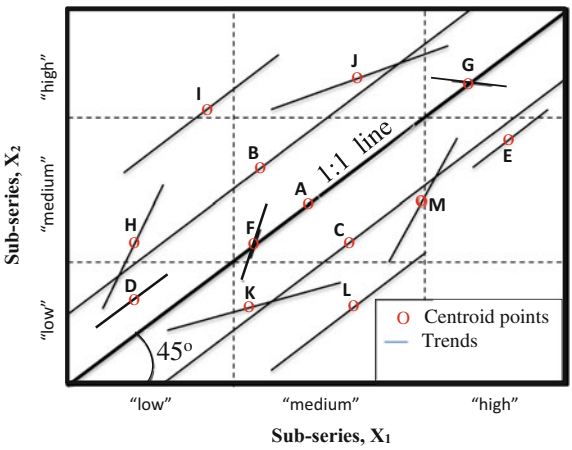


Fig. 8.12 Different partial trends in the innovative trend template domain

If the centroid point in Fig. 8.2 falls on the 1:1 (45°) straight line then the climatological time series does not have any monotonic trend on the average,

(2) Nonparallel line (F, G, H, J, K, or M) to 1:1 (45°) straight line implies standard deviation change with time (homoscedasticity in the statistical sense). These straight lines indicate trends in the standard deviation,

(3) Straight lines F and G have trends in the standard deviation, but not in the arithmetic mean,

(4) Furthermore, $H(K)$ and $J(M)$ imply increasing (decreasing) standard deviation trends.

In the following section, the existence of each trend is proved through an extensive Monte Carlo simulation method.

8.6 Stochastic Simulation Approach

In order to support the cases in Fig. 8.12, an extensive Monte Carlo stochastic simulation study is carried out leading to the confirmation of trend existences in the arithmetic average (mean) and the standard deviation. For this purpose, normally (Gaussian) distributed 10,000 synthetic data are generated with zero mean and unit variance. In the first part of the simulation studies, the synthetic series do not have trend in the standard deviation and the simulation results are given in the innovative templates in Fig. 8.13.

Three cases present no trend (A), increasing trend (B) and decreasing trend (C) time series similar to the conceptual cases in Fig. 8.2. Since, they are all parallel to the 1:1 (45°) straight line, they do not include trend in the standard deviation. The slope, S_μ, of the possible trend can be calculated by considering the difference between the arithmetic mean, μ_1 of the first half time series and the arithmetic mean, μ_2, of the second half as,

$$S_\mu = \frac{2(\mu_2 - \mu_1)}{n}, \tag{8.8}$$

where n is the number of the data in the climatological time series. This expression is an alternative to the slope formulation by Sen (1968).

There are six simulation innovative trend templates in Fig. 8.14 each one corresponding to the conceptual counterparts in Fig. 8.12 and their comparisons provide self-explanatory results.

In Fig. 8.14, centroid point deviations from the 1:1 straight-line imply trend existence, and since none of them is parallel to the 1:1 (45°) straight line, there is a trend in the standard deviation. The slope, S_σ, of such a trend can be calculated based on the difference of the two halves' standard deviations as,

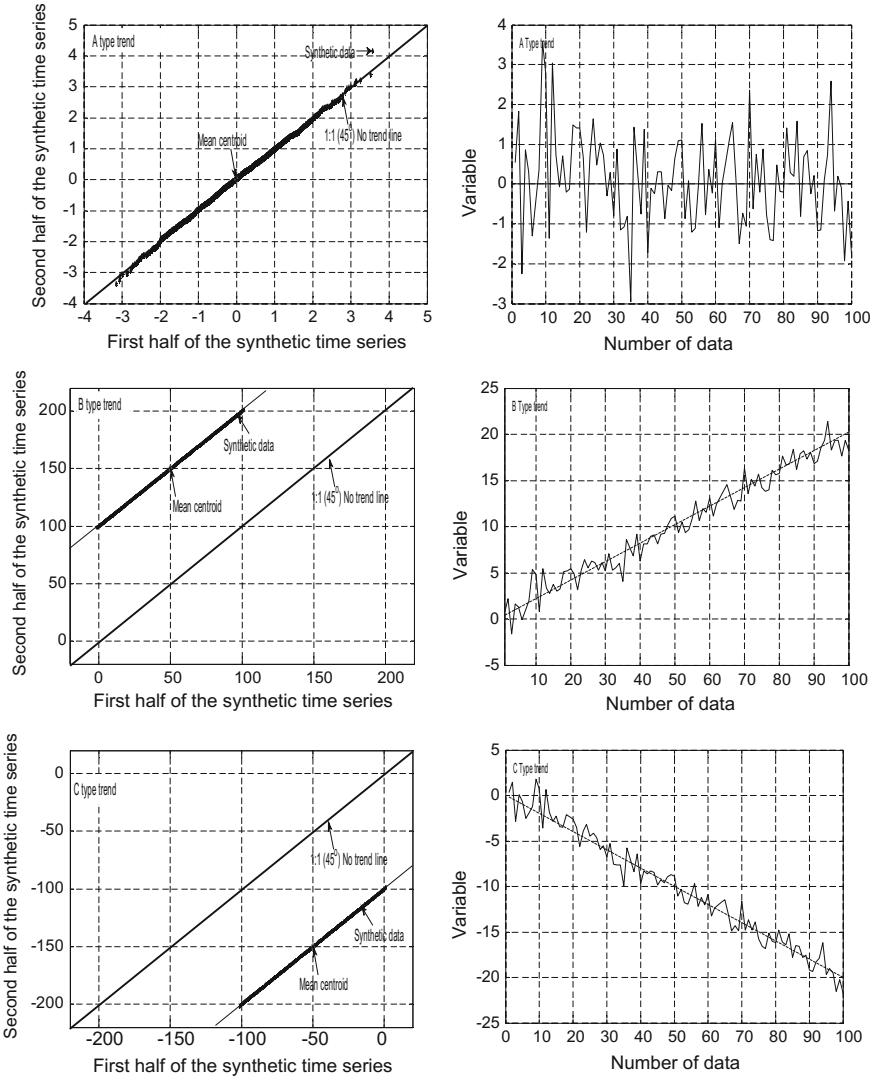


Fig. 8.13 Different arithmetic average trend types

$$S_{\mu} = \frac{2(\sigma_2 - \sigma_1)}{n}, \tag{8.9}$$

where σ_1 and σ_2 are the standard deviations of the first- and the second half time series, respectively. According to the positive (negative) sign of Eqs. (8.1) and (8.2), there is increasing (decreasing) trend. These two equations provide opportunity for identifying arithmetic average and standard deviation trends.

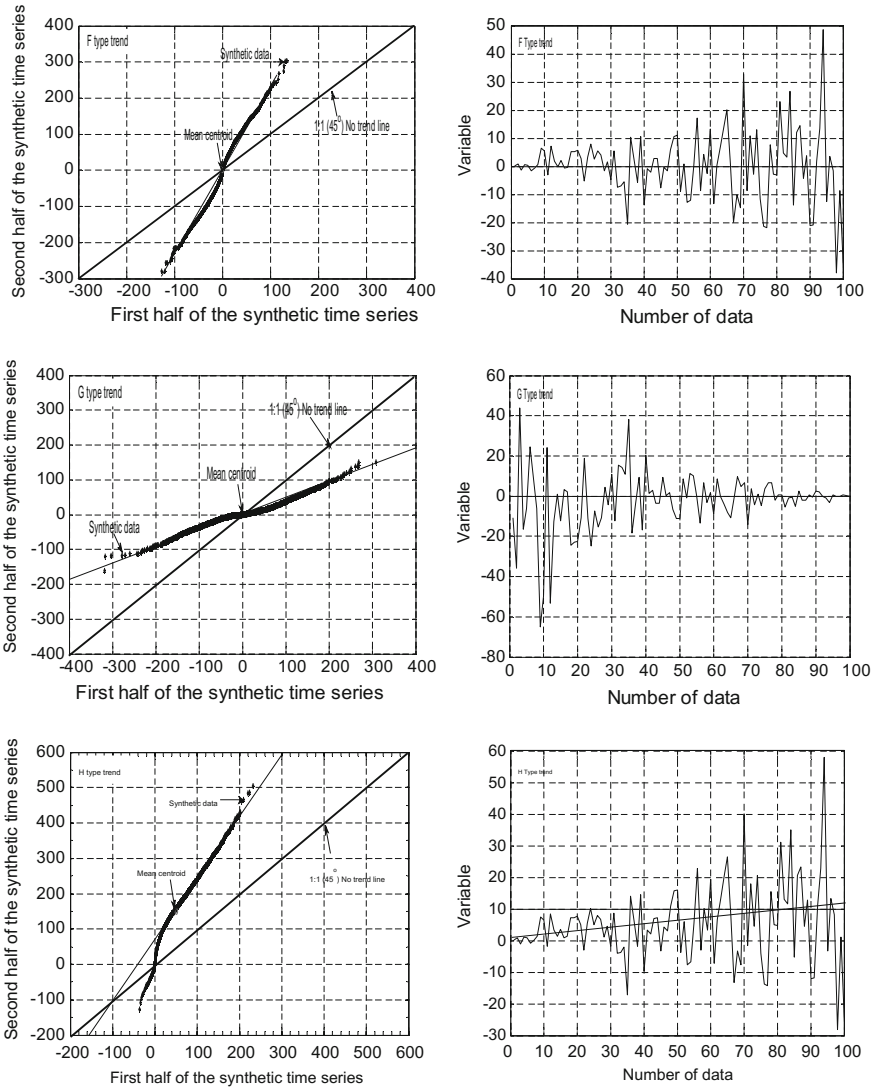


Fig. 8.14 Different standard deviation trend types

As explained above, the variation domain of each half series can be divided roughly into three groups as “low”, “medium”, and “high” portions and in each one of these portions one can identify one of the trend types given in Fig. 8.12.

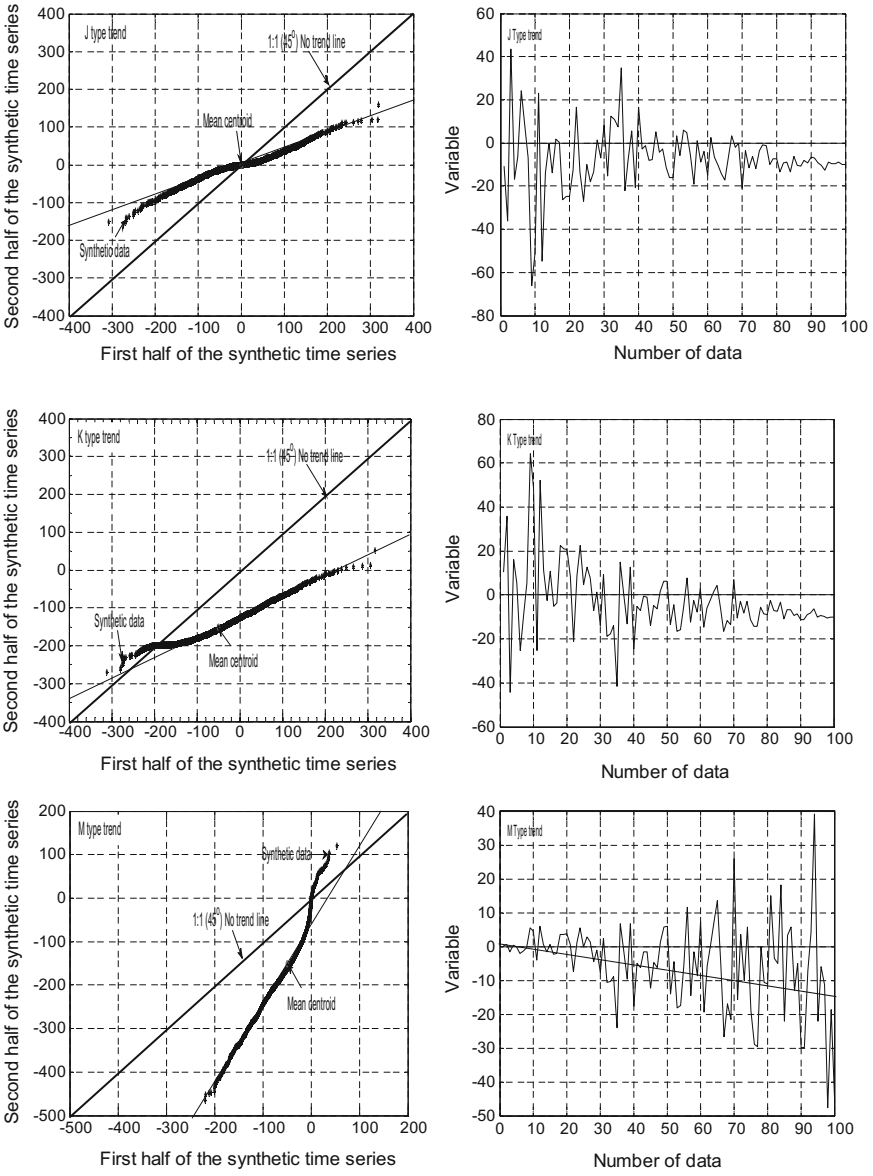


Fig. 8.14 (continued)

8.7 Data and the Study Area

The Office of the New Jersey State Climatologist has gathered and quality checked New Jersey state-wide annual temperature and precipitation records going back to 1895, and has made these data available on-line (http://epa.gov/climatechange/index.html; http://climate.rutgers.edu/stateclim). The data show a statistically significant rise in average state-wide temperature and precipitation over the last 110 years. Although there is much variation from year to year, overall the normally cooler (November through March) and warmer (May through September) seasons are warmer now than before. The rise in temperature appears to be especially pronounced during the November–March period.

With the proposed trend identification procedure one can identify three cluster groups as "low", "medium", and "high" record values for the state-wise annual temperature and precipitation records in the New Jersey state, USA. The temperature boundaries for "low", "medium", and "high" classes are less than 51 °F, between 51 and 53 °F and more than 53 °F, respectively. Similarly, corresponding class limits for the precipitation records are less than 40 mm, between 40–50 mm and more than 50 mm, respectively. Furthermore, in practical applications each boundary value must be considered as a rather vague value than crisp numerical separation between the groups, and this gives the researcher a flexible incentive. If the researcher is inclined to make clusters deterministically then s/he can stick to crisp boundaries and the clusters will appear in mutually exclusive manner. The view taken in this paper is that the boundaries could have some elasticity in the sense that the transition between the subsequent clusters includes some overlapping parts similar to fuzzy sets.

8.7.1 Partial Trend Groups

The application of the partial trend group methodology is presented for temperature and precipitation records in Figs. 8.15 and 8.16 for two 50-year subseries, respectively. In these figures, the scatters of each data group ("low", "medium", and "high") are indicated by three partially overlapping ellipsoids.

Herein, the "low" precipitation values have increasing trend in the arithmetic average and also slightly increasing trend in the standard deviation. In the "medium", and "high" precipitation ranges there are almost the same trend slopes, but not in the standard deviation. Since the centroid points of both categories are above the 1:1 (45°) straight line, there are increasing trends. The centroid of the scatter points at "high" portion is comparatively far away from the centroid of the "medium" portion, and therefore, the increasing trend slope in this portion is slightly bigger than the "medium" case.

In Fig. 8.16, state-wise annual precipitation records are available for the same record duration as for the temperature trend template.

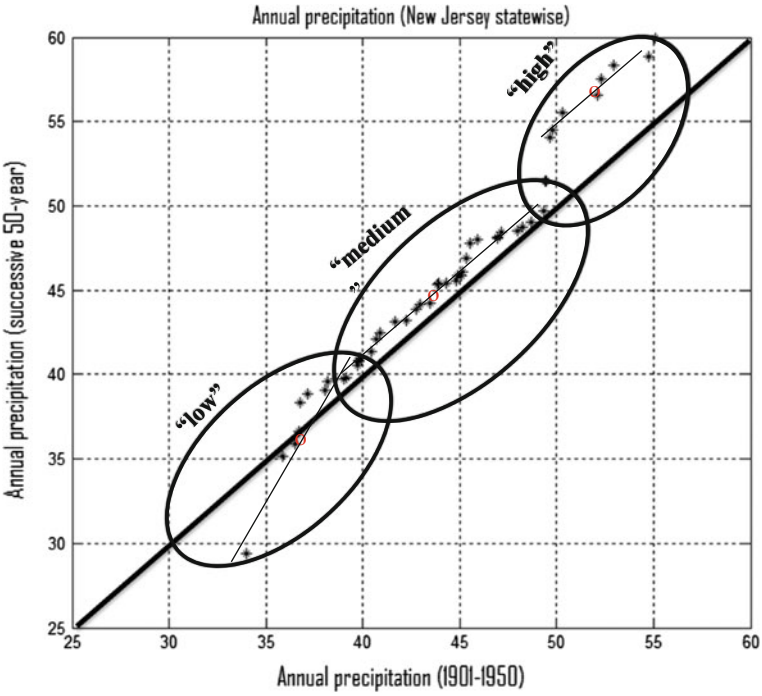


Fig. 8.15 Innovative trend templates for 50-yearly annual temperature trends

Comparisons of this figure with Figs. 8.1 and 8.2 lead to the conclusions that "low" values do not have homoscedasticity, "medium" and "high" values have homoscedasticity, and the mean centroids in both cases do not lie on the 1:1 (45°) straight line but above it, and therefore, both categories have increasing trends in the arithmetic average levels due to their parallel positions to 1:1 (45°) straight line. Comparatively "high" values have slightly more increasing precipitation trend than the "medium" classification. Non-homoscedasticity in "low" values implies that there is more variability in the "low" values than "medium" and "high" cases.

8.7.2 Partial Trend Lines

It is possible to separate the whole annual records into three partial time series. The MK trend analysis fits a monotonic linear straight line for the whole series according to a set of assumptions as independent serial structure of the time series and the normal (Gaussian) pdf of the climatological records. In general, the first assumption is not valid and the fitted trend cannot be treated as completely reliable.

Figure 8.17 indicates the entire trend types collectively, where the monotonic linear trend line is fitted to the whole data in groups.

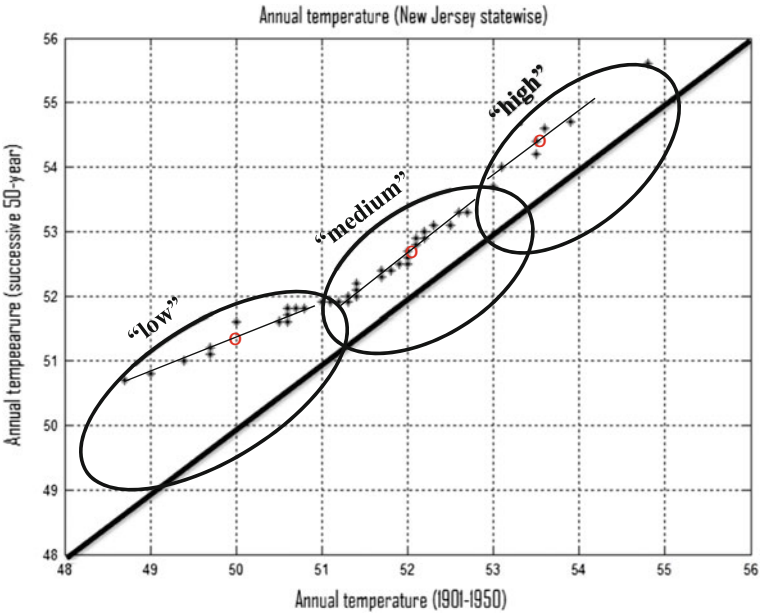


Fig. 8.16 Innovative trend templates for 50-yearly annual precipitation trends

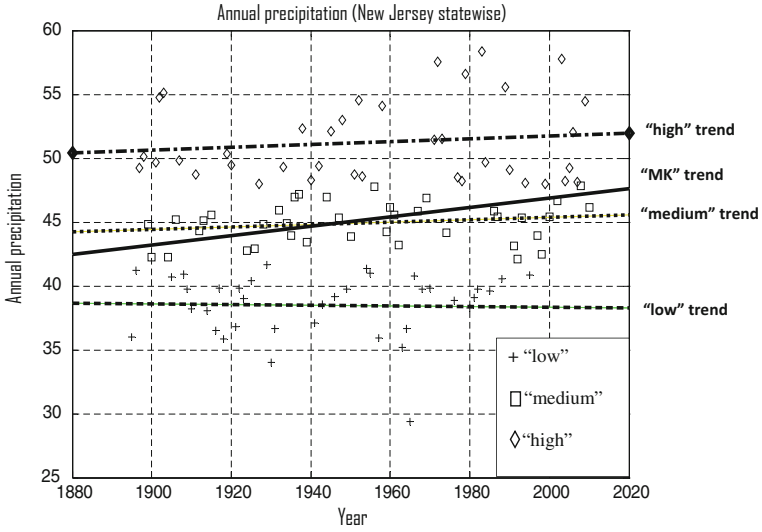


Fig. 8.17 Partial trend lines for annual precipitation records

The same figure gives separate trends for each “low”, “medium”, and “high” annual precipitation values instead of a single holistic approach according to the classical MK trend test. In Fig. 8.7, the holistic trend (MK) is also presented for comparison purposes. In detail, “low” annual precipitation values have slightly decreasing trend whereas the “medium” and “high” values have slightly increasing trends. The partial trend lines provide detailed reflection of the annual precipitation behavior in the study area. Likewise, partially descriptive annual trend lines are given for annual temperature records in Fig. 8.18.

It is possible to calculate the slopes of each trend lines in Figs. 8.17 and 8.18 and according to Eqs. (8.1) and (8.2) the results are shown in Table 8.1.

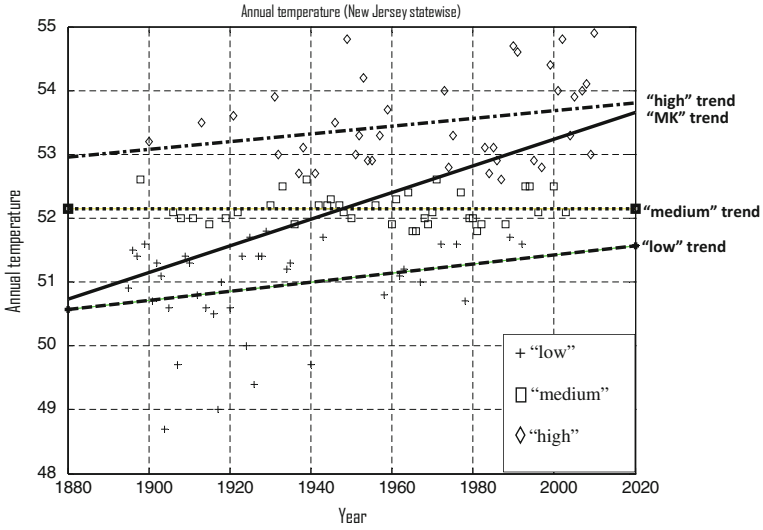


Fig. 8.18 Partial trend lines for annual temperature records

Table 8.1 Trend line slopes for precipitation and temperature

Trend type	Class	Trend slopes	
		Precipitation (mm/decade)	Temperature (° F/decade)
Partial trends	“high”	0.00375	0.00625
	“medium”	0.00250	0.00000
	“low”	0.00125	0.00725
Monotonic trend	Partial trend methodology	0.03325	0.01900
	Mann–Kendall (Sen’s tau)	0.03600	0.01720

The last two rows in this table indicate the time series monotonic and holistic trend slopes according to the method presented in this paper (Eq. 8.1) and also from the classical MK Sen (1968) trend slope analysis. It is obvious that the two approaches provide almost the same level of numerical magnitudes for the classical monotonic trend slopes.

In practice, it is convenient to give the slopes for 100-year period, and accordingly, multiplication of each slope in this table indicates that during the last 100-year period the precipitation and temperature increments have appeared as 3.125 mm and 2.1 °F, respectively. Another significant point that one can deduct from the numbers in this table is that the increments in partial trend lines are smaller than the whole trend case.

The main purpose of this section was to present application of an innovative graphical technique for trend identification in different parts of a climatological time series as "high", "low", and "medium" record values. For this purpose, the available time series is divided into two halves and each half, after arrangement in ascending order, is plotted against each other. The scatter area template on the first quadrangle in the Cartesian coordinate system is defined as a square with its main diagonal, 1:1 (45°) straight line, which corresponds to no-trend case. The upper (lower) right angle triangle corresponds to increasing (decreasing) trend cases. The methodology is applicable even in cases of serial-correlation existence in the climatological time series. Such plots give rise to many interpretation possibilities concerning trend existences. One is able to identify partial trends in groups of "low", "medium", and "high" record values and also along the whole climatological time series monotonic trend lines for similar classifications. The confirmation of the methodology is shown by extensive Monte Carlo simulation studies. The applications of the proposed methodology are presented for state-wise annual precipitation and temperature records for more than 100-year length from New Jersey, USA.

References

Alexandersson, H., & Moberg, A. (1997). Homogenization of Swedish temperature data. Part I: homogeneity test for linear trends. *International Journal of Climatology, 17,* 25–34.

Bayazit, M., & Önöz, B. (2007). To pre-whiten or not to pre-whiten in trend analysis? *Hydrological Sciences Journal, 52,* 611–624.

Burn, D. H., & Hag Elnur, M. A. (2002). Detection of hydrologic trends and variability. *Journal of Hydrology, 255,* 107–122.

Chen, T. T., & Xie, L. (2005). Identifying critical focuses in research domains. In *Proceedings of 9th International Conference on the Information Visualization (IV'05), London UK, 6–8 July 2005* (pp. 135–142).

Coggin, T.D. (2012). Using econometric methods to test for trends in the HadCRUT3 global and hemispheric data. *International Journal of Climatology, 32,* 315–320.

Douglas, E. M., Vogel, R. M., & Knoll, C. N. (2000). Trends in floods and low flows in the United States: Impact of spatial correlation. *Journal of Hydrology, 240,* 90–105.

Gan, T. Y. (1998). Hydro-climatic trends and possible climatic warning in the Canadian Prairies. *Water Resources Research, 34*(11), 3009–3015.

Hamed, K. H. (2008). Trend detection in hydrologic data: The Mann-Kendall trend test under the scaling hypothesis. *Journal of Hydrology, 349,* 350–363.

Hamed, K. H., & Rao, A. R. (1998). A modified Mann-Kendall trend test for auto-correlated data. *Journal of Hydrology, 204,* 182–196.

Karpouzos, D. K., Kavalieratou, S., & Babajimopoulos, C. (2008). Trend analysis in hydro-meteorological data. Technical Report No. 5.4, MEDDMAN, Interreg III B-MEDOCC, Thessaloniki, Greece.

Kendall, M. G. (1975). *Rank correlation methods*. London, UK: Griffin.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis, 92,* 78–89.

Luo, Y., Shen, L., Fu, S., Liu, J., Wang, G., & Zhou, G. (2007). Trends of precipitation in Beijiang River Basin. Guangdong Province, China. *Hydrological Processes, 22*(13), 2377–2386.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13,* 245–259.

Modarres, R., & da Silva, V. P. R. (2007). Rainfall trends in arid and semi-arid regions of Iran. *Journal of Arid Environments, 70,* 344–355.

Mosmann, V., Castro, A., Fraile, R., Dessens, J., & Sanchez, J. L. (2004). Detection of statistically significant trends in the summer precipitation of mainland Spain. *Atmospheric Research, 70,* 43–53.

Nalley, D., Adamowski, J., Khalil, B., & Ozga-Zielinski, B. (2013). Trend detection in surface air temperature in Ontario and Quebec, Canada during 1967–2006 using the discrete wavelet transform. *Atmospheric Research, 132*, 375–398.

Thomas, B.F., & Timothy, J.V. (2002). The Application of Size-Robust Trend Statistics to Global-Warming Temperature Series. *Journal of Climate, 15,* 117–123.

Sen, P.K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 63,* 1379–1389.

Şen, Z. (2012). Innovative trend analysis methodology. ASCE, *Journal of Hydrologic Engineering, 17,* 1042–1046.

Şen, Z. (2014). Trend Identification Simulation and Application. ASCE, *Journal of Hydrologic Engineering, 19,* 635–642.

Ventura, F., Pisa, P. Rossi, & Ardizzoni, E. (2002). Temperature and precipitation trends in Bologna (Italy) from 1952 to 1999. *Atmospheric Research, 61,* 203–214.

Wayne, A.W., & Gray, H.L. (1995). Selecting a Model for Detecting the Presence of a Trend. *Journal of Climate*, *8,* 1929–1937.

Yue, S., & Wang, C. Y. (2004). The Mann-Kendall test modified by effective sample size to detect trend in serially correlated hydrological series. *Water Resources Management, 18,* 201–218.

Yue, S., Pilon, P., & Cavadias, G. (2002). Power of the Mann-Kendall test and Spearman's rho test for detecting monotonic trends in hydrological time series. *Journal of Hydrology, 259*(1–4), 254–271.

Yue, S., Pilon, P., & Phinney, B. (2003). Canadian streamflow trend detection: Impacts of serial and cross correlation. *Hydrological Sciences Journal, 48*(1), 51–63.

Index

Z. Şen, *Innovative Trend Methodologies in Science and Engineering*,
DOI 10.1007/978-3-319-52338-5