International Journal: Mathematical Manuscripts.
Volume 1 Number 1 (2007), pp. 95–103

http://www.ripublication.com/ijmm.htm

A Computer Program for Statistical Analyses of Hydrometeorological Data[1]

Carmen Maftei, Cristina Ghergina and Alina Barbulescu
Ovidius University of Constanţa, Romania
E-mail: cmaftei@univ-ovidius.ro, alinadumitriu@yahoo.com

Abstract

To understand the processes that intervene in the water cycle and to study their spatial and temporal variations, a database is essential.

Before their use (for planning and management of water resources) these databases must be tested in order to relieve the errors.

In this context, this paper presents a program for statistical investigations of hydrometeorological data based on specific hypotheses (stationarity, homogeneity, independence etc.).

The application presented here is written in Java, a high-level object-oriented programming language, developed by JavaSoft. The graphical interface is suggestive and easy to use and get the needed results. The diagram of this application contains the following stages: building the data series, determining the series of the characteristic values, data selection, statistical control (statistical and hydrological methods).

AMS Subject Classification:
Keywords: Semiorthogonal, scaling function, quadratic B-spline, integral equation.

1. Introduction

To understand the processes that intervene in the water cycle and to study their spatial and temporal variations, a database is essential. The field observations are very important for climatic and hydrological statistics, for planning and management of water resources.

To allow the passage from the data acquisition to their effective use, the following stages can be distinguished: acquisition, treatment, control and validation.

[1]This research was partially supported by Grant PNCDI/2007.

The control and data validation are realized in order to detect the errors that can appear in the data measurement, transmission and storage.

As a function of the error type (random or systematic), many methods can be used, they are: *in situ* (misreading, misplaced decimal points, copying errors and arithmetic mistakes, occasional errors due to temporary disturbance of the gauge or its exposure), visual analysis (which uses the graph to determine temporal patterns, *e.g.*, trend or step-change, seasonal variation, etc.), hydrological methods (double-mass method) and statistical investigation.

In this paper we present a program for statistical investigations of hydrometeorological data. This application is wrriten in Java and contains the followinng stages: building the data series, determining the series of the characteristic values, data selection, statistical tests.

2. Background

There are very many statistical methods that can be used to investigate the data series. This section attempts to provide a brief overview.

Two sets of terminology are frequently used to distinguish types of test. The tests used are parametric or non-parametric. In most cases, the parametric tests are based on the normal law and assume the existence of a reference random variable X. The question is whether the results are valid if X is not normal: if the results are valid, the test is "robust". This means that the test remains almost insensitive to certain modifications of the model. A test is said to be non-parametric if it has a free distribution.

Generally, in hydrology and meteorology we utilize five categories of tests:

- conformity test.

  They are used to verify if a sample is a part of a given population, with respect to a known parameter, like mean or variance.

  Two tests are used for the conformity test of average depending on whether variance is known or must be estimated. They are respectively the z-test and Student test (or t-test).

- homogeneity tests or tests of samples comparison.

  The average homogeneity test is based on the Student statistic for two samples. In its basic form it assumes normally distributed data.

- stationarity tests.

  The stationarity tests verify if the series properties are time invariant. Two types of non-stationarity are interesting from the point of view of hydrology: trend and break.

  A break is a sudden change in the time series properties. It assumes that the properties are stable before and after the break moment. As a function of the

nature of the alternative hypothesis of the stationarity, there are the following test types (Lemaitre 2002):

1. general change distribution tests, as χ^2 and Kolmogorov – Smirnov;
2. break tests: the homogeneity tests can be adapted (as t-test, if the break point is known) or Wilcoxon, Pettitt, Buishand;
3. trend tests: based on linear regression (parametric tests) and rank - based test (non-parametric tests used when the series have not a normal distribution, but have a linear behavior).

- autocorrelation tests;
- adjustment tests.

The following tests were implemented in this application: Student-test for two samples, Kolmogorov–Smirnov, Wilcoxon and Kendall.

In this paper we shall briefly present the mentioned tests adapted to the hydrometeorological data series, according to World Meteorological Organization (WMO).

The **Kolmogorov–Smirnov test** is a distribution-free test for a general change in distribution. The data series is divided into two parts, assuming a known change-point time, and the test is used to compare the two parts.

The hypotheses are:

H_0: the distrributions before and after change are identical;

H_1: the distrributions before and above change are different.

The statistic of the test is defined as:

$$D = \sqrt{n_1 \cdot n_2} \cdot \max \{|F_1(x) - F_2(x)|\},$$

where: $F_1(x)$ and $F_2(x)$ are the empirical distributions for the two parts of the data series, n_1 and n_2 are the samples volumes, with $n_2 = n - n_1$.

The critical value of D can be found in the Kolmogorov-Smirnov tables.

Wilcoxon test is a rank-based and distribution-free test. The change moment is assumed known and the series is divided into two groups (before and after the change moment) and these groups are compared.

The null hypothesis, H_0, is that the medians of the two groups are equal.

To compute the rank-sum test statistic (Hirsch *et al.*, 1992), one uses the following stages:

1. Assign ranks to all the data. In the case of ties (equal data values) use the average of ranks.
2. Split the data into two groups of size n_1 and $n_2 = N - n_1$. Compute a test statistic S as the sum of ranks of the n_i observations in the smaller group.

3. Compute the theoretical mean and standard deviation of S for the entire sample:

$$\mu = \frac{n_1(N+1)}{2},$$
$$\sigma = \sqrt{\frac{n_1 n_2(N+1)}{12}}.$$

4. The standardised form of the test statistic Z is computed as:

$$Z = \begin{cases} \dfrac{S-0.5-\mu}{\sigma} & \text{if } S > \mu, \\ 0 & \text{if } S = \mu, \\ \dfrac{S+0.5-\mu}{\sigma} & \text{if } S < \mu. \end{cases}$$

Z is approximately normally distributed and so significance levels can be obtained from normal probability tables. Thus, for a significance level α, reject H_0 if $|Z| > Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the $1-\alpha/2$ point of the standard normal probability distribution.

Kendall test is a distribution free rank based test(Snyers 1975).

For any observation x_i of the data series, the number n_i of the observations preceding x_j $(j < i)$ and that are inferior to it $x_j < x_i$ is calculated.

The t-statistic of the test is $t = \sum_{i=1}^{n} n_i$.

For n big enough, the t-statistic has a normal distribution with the expectance $\mu = \frac{n(n-1)}{4}$ and variance $\sigma = \sqrt{\frac{n(n-1)(2n+5)}{72}}$.

Snyers proposes the use of this test as a progressive one. For this, one calculate progressive $z_i = \frac{x_i - \mu}{\sigma}$ and a graph of z_i as a function of i is plotted.

A second curve z_i' is plotted for i descending, from n to 1.

In the absence of tendence the two graph are overlapping. If not, the intersection of the curves permits the determination of the beginning of the trend phenomenon (Fig. 1).

3. The Program Presentation

The statistical module of ETREF application is a Java implementation of several statistical tests used to analyse the data obtained by observation and to accept or reject, based on this analysis, a statistical hypothesis.

The graphical interface is highly suggestive and easy to use, and its menus can be displayed both in Romanian and English language. Due to this friendly interface, the program lets the user navigate real easily through the commands sets and get the needed results.

The application offers support for:

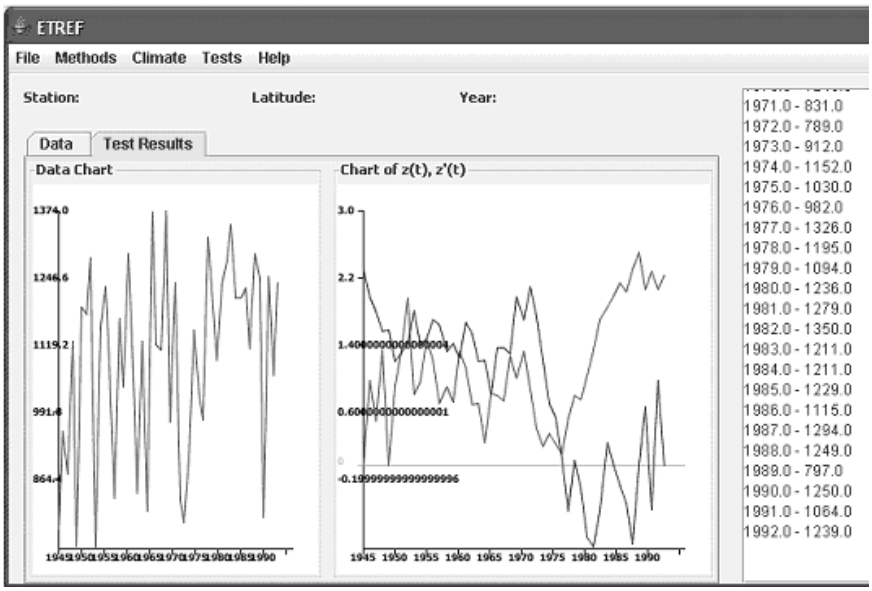

Figure 1: Kendall progressive test (after Snyers).

Figure 2: Graphical interface for loading the data entry set.

- Loading the right entry data set, according to the test that the user needs to run. This data set is read from Excel files, the files format being described in the afferent documentation, which may be accessed through Help menu. We would like to specify that the program can process the entry data set regardless of its format, which depends on the data set source. The data set may be composed of daily, monthly or annual values (Fig. 2).

- Checking the entry data set integrity and showing messages according to the detected errors (for instance, the user input is a string instead of a number);

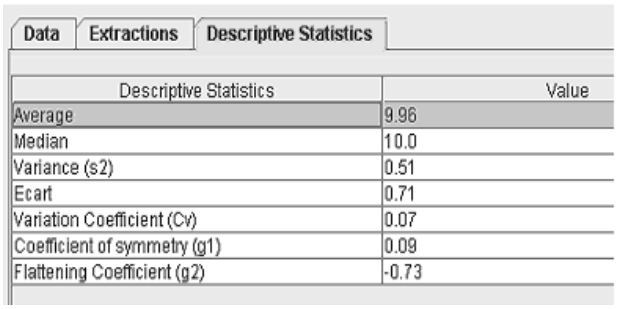

Figure 3: The interface of descriptive statistics module.

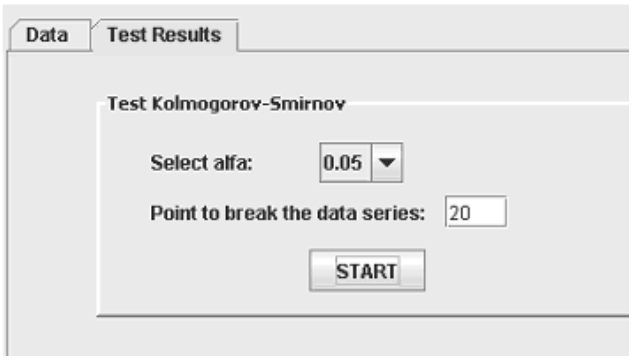

Figure 4: The interface for the statistical tests.

- Computing the values of descriptive statistics (average, median, variance etc.) (Fig. 3);
- Checking the applicability criteria (such as normal distribution) of data entry set for the tests that require this operation (Fig. 4);
- Setting the value of the first type I (α) and the break points;
- Creating and displaying suggestive charts with the results of some statistical tests (such as Kendall (Snyers)) (Fig. 1);
- Applying the calculus algorithms for each implemented test and displaying the conclusion: whether to accept the null hypothesis or to reject it;
- Saving in Excel files on disk, in a user specified location, the results of some statistical tests (Kendall) and the descriptive statistics values. This way of saving data was chosen because we want that the user takes full advantage of Microsoft Excel functionality regarding numerical data processing and charts operations.

The application is well documented, describing each step that the user has to make in order to run the tests.

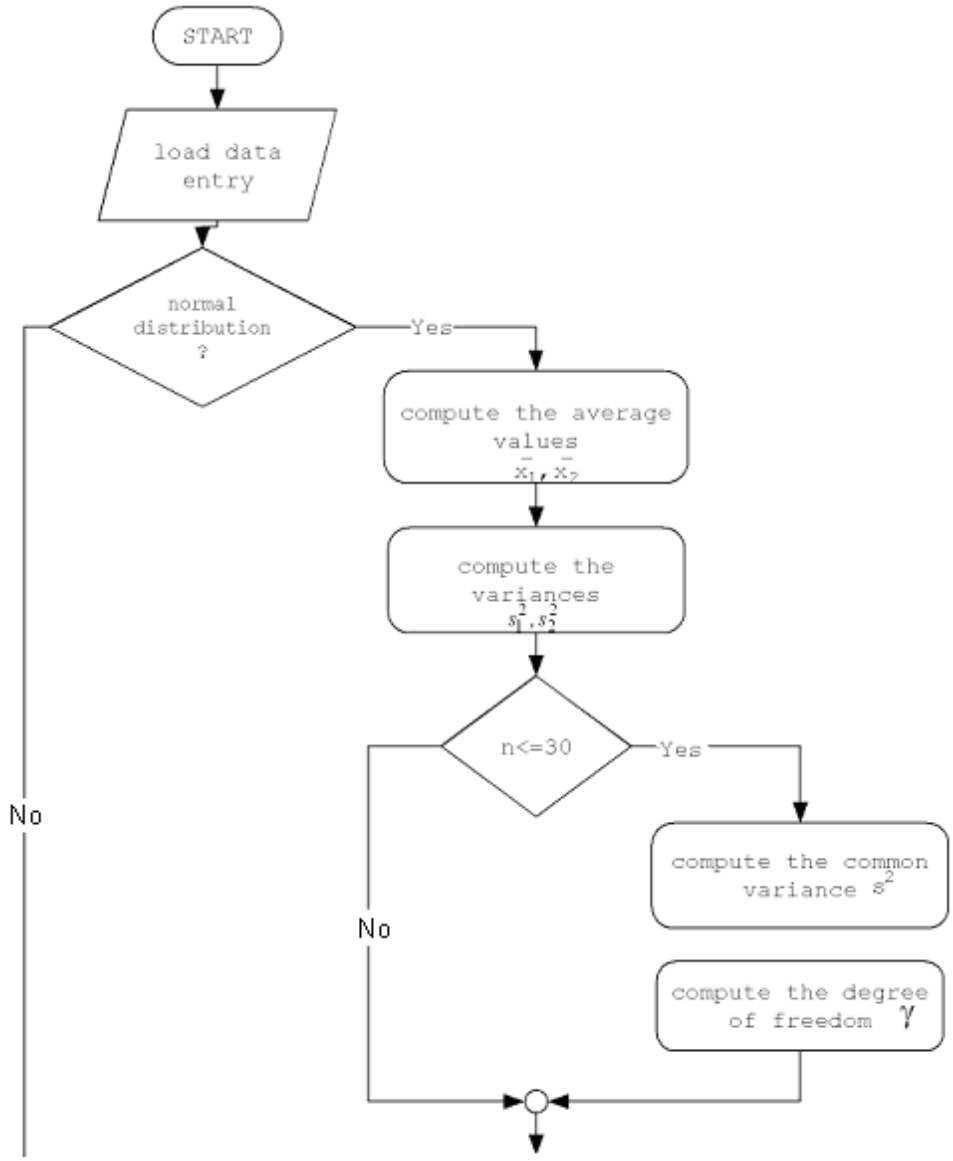

Figure 5: Algorithm of Student test.

To install and execute the application, the user has to double click the jar archive ETREF.jar. Apart from the Java standard libraries of functions, the program uses the library jxl.jar to read from and to write to Excel files.

Therefore, this archive is automatically copied during the installation process to the jre/lib/ext folder on the user computer.

In Figs. 5, 6 we present the flowchart of the case when the user needs to run the Student Test over a data set stored in an Excel file.

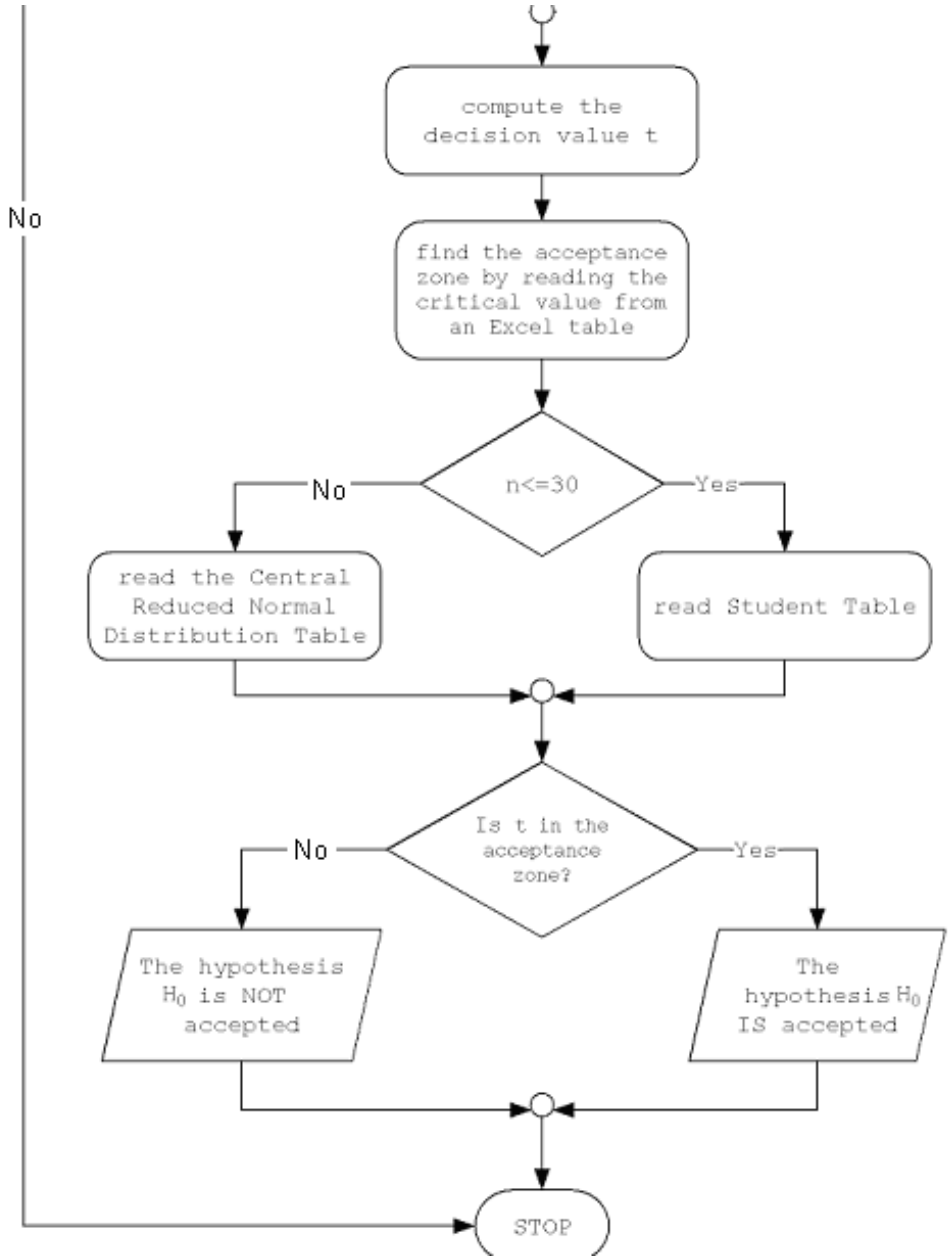

Figure 6: Algorithm of Student test, continuation.

4. Conclusions

The ETREF application allows the user to run quickly and efficiently several complex statistical tests over a large data set, having the results calculated in a short period of time (depending of the user computer resources) and registered in Excel file in a specified location.

We shall also implement other statistical tests, like Pettitt, Buishand, segmentation test (Hubert) etc.

References

[1] J. Cornell and M. Horstmann, Core Java, 3rd Edition, Prentice Hall, 1997.

[2] B. Eckel, Thinking in Java, 3rd Edition, Prentice Hall, 2002.

[3] R.M. Hirsh, et al., Statistical Analyses of Hydrologic Data, Chapter 17 in Handbook of Hydrology, Mc. Graw-Hill, New York, 1992.

[4] F. Lemaitre, Travail de Fin d'Etudes, 2002.

[5] P. Meylan and A. Musy, Hydrologie Frequentielle, H.G.A., Bucharest, 1999.

[6] R. Sneyers, Sur l'Analyse Statistique des Séries d'Observations, Technical Notes, 143, WMO, Geneva, 1975.

[7] R. Sneyers, Detecting Trend and Other Changes in Hydrological Data, WCDMP-45, WMO/TD 1013, 2000.