MIDDLE EAST TECHNICAL UNIVERSITY
DEPARTMENT OF STATISTICS

Homogeneity Analysis of Turkish Climate Data

Ceyda Yazıcı, Vilda Purutçuoğlu, Ceylan Yozgatlıgil,
Könül Bayramoğlu, Cem İyigün, İnci Batmaz

METU-STAT-Technical Report-2012- 001

March, 2012

DEPARTMENT OF STATISTICS
MIDDLE EAST TECHNICAL UNIVERSITY
ANKARA 06531 – TURKEY

TECHNICAL REPORT

HOMOGENEITY ANALYSIS OF TURKISH CLIMATE DATA*

Ceyda YAZICI[1], Vilda PURUTÇUOĞLU[2], Ceylan YOZGATLIGİL[3], Könül BAYRAMOĞLU[4], Cem İYİGÜN[5] , and İnci BATMAZ[6]

[1-4,6] *Middle East Technical University, Department of Statistics, 06531, Ankara, Turkey*

[5] *Middle East Technical University, Industrial Engineering Department, 06531, Ankara, Turkey*

[1]cyazici@metu.edu.tr, [2]vpurutcu@metu.edu.tr, [3]ceylan@metu.edu.tr,

[4]konul@metu.edu.tr, [5]iyigun@ie.metu.edu.tr,[6]ibatmaz@metu.edu.tr

ÖZET

İklim verisi doğası gereği korelasyonlu uzun zaman serileridir. Eğer bu ölçümlerde iklim kaynaklı olmayan etkiler varsa, analiz sonuçları belirsiz olabilmektedir. Bu nedenle "türdeş olmama" olarak adlandırılan bu tarz değişikliklerin belirlenmesi önem teşkil etmektedir. Önceki çalışmalarda türdeşlik kontrolü gerçekçi olmayan bağımsızlık varsayımı altında ve sadece ortalamada kaymalar dikkate alınarak yapılmaktaydı. Bu çalışmada ise türdeşlik analizi korelasyonlu veya varyansda da kaymaların olabileceği varsayılan ölçümler için tasarlanmış parametrik ve parametrik olmayan bazı yöntemler kullanarak Türkiye yağış verisine uygulanmıştır. Bu yöntemler Friedman testi, CUSUM ve Shewart grafikleridir. Elde edilen bulgular, önceki bir çalışmada uygulanan SNHT yönteminin sonuçları ile karşılaştırılmıştır. Karşılaştırmalar CUSUM kontrol grafiğinin bu amaçla kullanılabilecek iyi bir yöntem olduğunu göstermektedir.

Anahtar kelimeler: *Ortalamada kayma, varyansta kayma, bağımlılık problemi, Türkiye yağış verileri, Friedman testi, CUSUM grafiği, Shewart grafiği*

ABSTRACT

The climate data is a long-term time series correlated naturally. If there are any artificial effects not originated from the climate, analysis results can be ambiguous. Thereby, the detection of such effects, called inhomogeneity, becomes crucial. In previous studies, this investigation has been conducted under the unrealistic independence assumption considering only mean shifts. In this study, homogeneity analysis is applied to Turkish precipitation data by using (non)parametric methods designed for correlated data or for handling variance as well mean shifts. These are Friedman test, CUSUM and Shewart charts. Then, findings are compared to that of SNHT obtained in another study. Results show that CUSUM chart is a good method for the homogeneity analysis.

Key Words: *Mean shift, variance shit, dependency problem, Turkish Precipitation data, Friedman test, CUSUM chart, Shewart chart*

1. INTRODUCTION

The climate data are one of the well-known long-term time-series which are sensitive to non-climate artificial factors. These factors can make the series unrepresentative due to the shifts in mean or changes in variation of the measurements. The underlying fluctuations can be originated from different sources such as the changes in the measuring instruments, the formulas used to obtain some measurements (e.g. mean of the maximum temperature), or the changes in the locations of the stations and their environments [1] [4]. Because of their influences on the original data, both the detection and the removal of the inhomogeneities, if

* This study is supported by Middle East Technical University under the contract number BAP-2008-01-09-02.

any, are important before conducting an analysis on data.

There are a number of methods used to identify and adjust non-climatic fluctuations in climate data. These techniques utilize nonparametric or parametric approaches. Among many alternatives in the literature, Kruskal -Wallis [8], Mann-Whitney [9], and Von Neumann Ratio [12] tests are the well-known parameter-free methods. On the other hand, the Standard Normal Homogeneity Test (SNHT) [1], Buishand Range [2], and the Potters' test [4] are extensively used parametric methods for this purpose. However, all of these tests have some drawbacks. First, they check only if there exists a mean shift in data. But, there may also be inhomogeneities due to shift in the variance. In addition, these methods assume independence of observations. Nevertheless, dependency is one of the characteristics of time series, hereby, of the climate data. Thus, this feature needs to be considered also in the assessment of data for homogeneity so that we can unravel the problem of overestimation in the analysis. Moreover, most of these tests are based on the normality assumption which may not be realistic for this type of data [10]. Although this is the case, in most of the previous studies, normality assumption is not validated at all.

In this study, we propose to use different procedures which partly attack to solve these problems for testing homogeneity. Hereby, we suggest to implement the Friedman test [5], Cumulative Sum (CUSUM), and Shewart control chart [6] for more realistic analysis of the climate data. The Friedman test, a nonparametric approach, is unique among them which can deal with the dependency in data. But, it only considers inhomogeneity due to the mean shift. On the other hand, the other two methods, are able to detect inhomogeneities due to both mean and variance shift. But, they do not consider the dependency structure of data and require the validation of normality.

Above methods are performed to analyze the Turkish climate data which are very frequently used in the previous studies. From the underlying preliminary evaluation, we interpret the results and select the best approach among the techniques considered so that it can be used in a comprehensive analysis of a realistically large dataset containing many climate variables. Hereby, we analyze the Turkish precipitation data gathered from the Turkish Meteorological Services General Directorate (TMIGM) from 1950 to 2006, and conduct the named tests to identify the inhomogeneous series. The results are then compared to that of the SNHT applied on the same data in the literature. We also suggest further studies to handle the problems stated simultaneously.

2. METHODS

2.2 Friedman Test

The Friedman test is a nonparametric method used to test the differences between population means [5]. The merit of this test is its ability of taking into account. To compute the test statistic, *S*, *k* observations are sorted in ascending order in each of the groups, and computed by using the following formula:

$$S = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1), \ (j = 1, \ldots, k)$$

where $R_j^2 = (\sum_{i=1}^{n} r_{ij})^2$ and r_{ij} denotes the rank of X_{ij} in the joint ranking of observations in group *i*.

Here, it is assume that observations within each group and between groups are independent and dependent, respectively. To test the null hypothesis, $H_0 : \tau_1 = \tau_2 = \ldots = \tau_k$, which states the equality of means while τ_j denotes the group mean, *S* is compared with the critical value, s_α. If $S \geq s_\alpha$, we reject H_0 at α level of significance.

2.3 Shewart Charts

The control charts can deal with the mean, variance, or both sources of shifts successfully. Because of their abilities of simultaneous checking without fitting any particular model to the data, they are very practical, and are extensively used in literature in order to find out the causes of process shifts [6]. Hereby, the Shewart Charts method is one of these well-known control charts used in process control. In this method the detection of shifts is based on the following upper and lower control limits, denoted by *UCL* and *LCL*, respectively.

$$UCL = \mu_w + L\sigma_w \quad Center = \mu_w \quad LCL = \mu_w - L\sigma_w,$$

where *w* refers to a statistic that represents some property of interest, μ_w displays the mean of *w* and σ_w shows the standard deviation of *w*. *L* is used to define the distance between control limits and the centre line, and is measured in terms of the standard deviation. In the calculation, the observations are plotted against the run order. If the observation crosses the control limits, we interpret as the indication of inhomogeneity. Note that we should validate normality of data for the success of this method.

2.4 CUSUM Charts

The CUSUM charts are performed to detect inhomogeneity at a certain point (i.e. break point), and particularly useful for detecting small shifts in the series [6]. Thereby, for each data point, the following statistics are calculated:

$$C_i^+ = \max\left[0, x_i - (\mu_0 + k) + C_{i-1}^+\right]$$
$$C_i^- = \max\left[0, (\mu_0 - k) - x_i + C_{i-1}^-\right],$$

where the initial values of *C* are set to zero; *k* represents half of the magnitude of the shift in terms of the standard deviation of the process which usually takes on the value in the form of $\frac{\delta}{2}\sigma$; and μ_0 is the center line. Moreover, the control limits are determined to be $h\sigma$, where *h* is a free parameter determined by the user. In the detection of shifts, C_i values are plotted against the run order. Any C_i values exceeding the control limits are thought to indicate inhomogeneity of the series at that point. On the other hand, similar to Shewart charts, CUSUM charts are also performed under the normality assumption of the data, hereby, the data need to be tested for the validity of the normality before the analysis.

2.4 Standard Normal Homogeneity Test

SNHT is the most popular test of homogeneity used for climate data in the literature. It is based on the comparative assessment of a (target) station with its neighbors, called reference

stations, to detect whether the series of the test station is homogeneous ~~or not~~. However, there are some problems associated with this test. First, simulation studies reveal that the normality assumption on which the test is based may not be a realistic one [10]. Second, it does not consider the dependency structure, if exist, in data. Third, it needs certain number of homogeneous reference stations correlated with the target station. Next, SNHT deals only with the mean shift assuming single break. Nevertheless, there can be several such breaks in the series. Moreover, this test may not come up with a certain decision regarding the existence of inhomogeneity, and may state that the data are non-testable or inconsistent.

3. APPLICATION AND RESULTS

The data are based on the precipitation measurements of 60 stations between 1950 and 2006, and are obtained from TMIGM. In this dataset, the total precipitations for each year are calculated for every other station.

In the climate data due to its spatio-temporal characteristic, we typically observe, two-way dependency in every station both through time and related to the topological feature of the station such as the altitude of the location where the station is built. To be able to apply Friedman test, which can handle one-way dependency, however, we need to rearrange the data by taking the transpose of the observation matrix before applying the test so that possible correlation in one-way can be reduced, and the resultant data possess merely only one source of dependence. Moreover, in order to capture the possible inhomogeneities, we make a comprehensive search in a number of grouped data by considering all plausible number of sequential groups from 2-year to 28-year periods, resulting in 27 such different cases. In grouping data, for example, if we generate five-year groups, the first two-year data (i.e. 1950 and 1951) are omitted from the analysis, and the groups are then formed by including sequential five years. As a result, the data from 1952 to1956 are taken as the first, and the data from 1957 to 1961 are considered as the second group and so on. By this way we generate, 11 such groups. This arrangement can preserve the dependency within groups while reducing the dependency between groups. Finally, the results are compared with that of the SNHT [4].

To evaluate the performance of our suggested approaches, we presented the results of the Friedman test for three stations, namely, Artvin, Çanakkale and Zonguldak, which are observed as ~~e~~ non-testable, homogeneous, and homogeneous after correction, respectively, according to the SNHT findings [4]. In our assessment, since the groups are obtained for different periods in each case, Artvin station is found to be inhomogeneous once, when the group size is 14 years. However, this station is classified as non-testable in the application of SNHT due to the lack of reference stations. From the results of Friedman test, Çanakkale is observed to be homogeneous as similarly detected by the SNHT outcome. Finally, Zonguldak is found to be inhomogeneous in Friedman test twice for the group sizes two and eight (see Table 1).

Table 1. Friedman test results for Artvin, Çanakkale, and Zonguldak stations

Station	**Number of Inhomogeneities**	**Group Size in Number of Years**	**SNHT Result**
Artvin	1	14	Non-testable
Çanakkale	Homogeneous	-	Homogenous
Zonguldak	2	2, 8	Homogenous after correction

As a second approach, CUSUM method is applied. Before the analysis, the normality assumption of the data is checked. If it is violated, the total precipitation values are

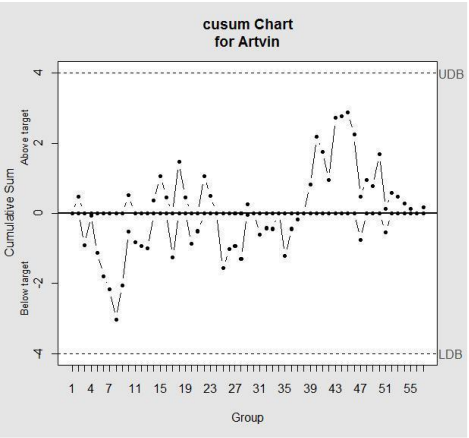
transformed appropriately to obtain normal distribution. Then, graphs are drawn for each station under different choices of control limits. The results of the total precipitation obtained under the control limits of ±4 units, i.e. 4σ where σ = 1, are presented in Figure 1.

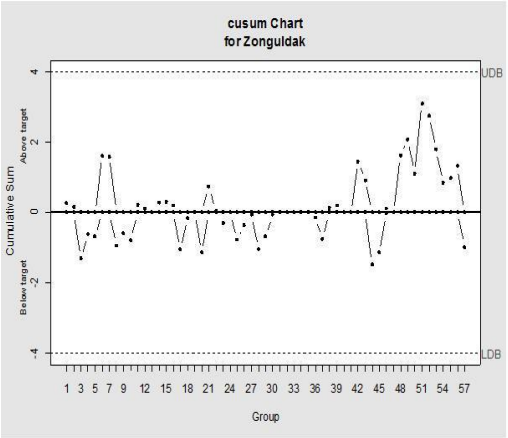
According to the Shewart Charts, none of the points is above or below 3σ control limits. However, there are points classified as out of control when the control limits are taken as σ and 2σ (see Figure 2). The findings of this study and the results of the SNHT method for all 60 stations are presented in Table 2. Note here that two stations which take place in this study are not common with the SNHT study too.

4. CONCLUSION

In this study, some methods, which have rather more realistic assumptions regarding the structure of climate data, are proposed to be used for testing the homogeneity of precipitation data between 1950 and 2006 for Turkey. Our purpose here is to exclude the non-climatologic effects from the dataset before conducting any analysis. Hereby, we suggest the use of Friedman test, Shewart, and CUSUM Charts instead of SNHT and other methods extensively used in the literature. These methods have the capability of tackling with the serious problems related to the climate data such as dependency or the variance shift as well as mean shift in the time-course observations.

From the results, we observe that the Friedman test does not have a very good performance. This may be attributed to the dependency structure that still remains after the rearrangement of data. In addition, it cannot identify the exact break points. On the other hand, the control charts are found to be effective for identifying small shifts in mean, variance, or both simultaneously. Furthermore, by adopting CUSUM approach, extra information is gained for non-testable and inconsistent stations as seen in Table 2.





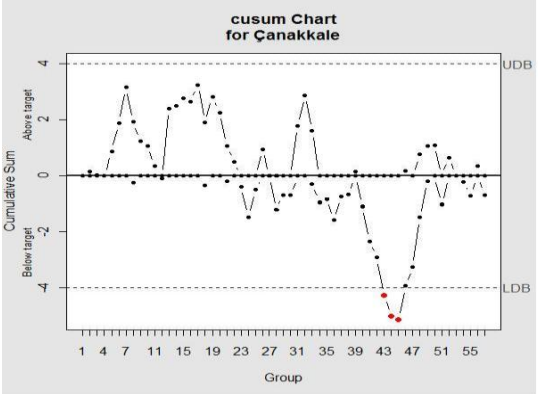


Figure 1. CUSUM charts for Artvin, Çanakkale and Zonguldak stations under 4σ where σ = 1.

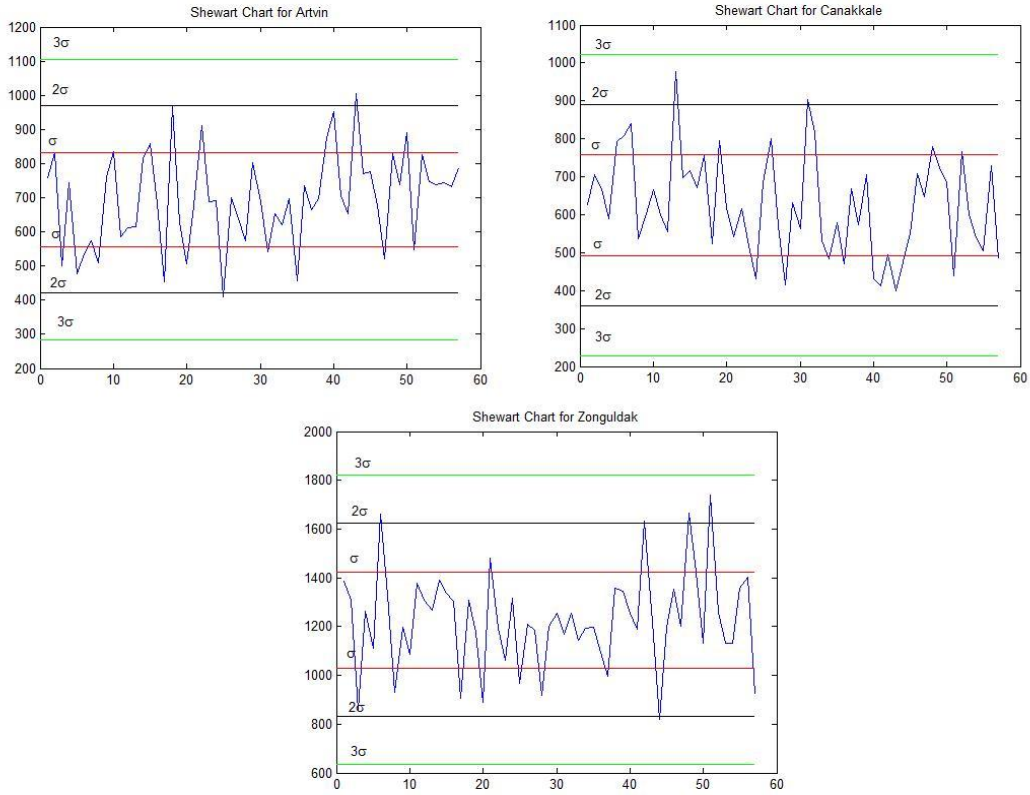


Figure 2. Shewart Charts for Artvin, Çanakkale and Zonguldak stations for σ, 2σ and 3σ limits where σ is the standard deviation of each selected station.

Table 2. Classification of 60 stations according to Shewart, CUSUM Charts, and SNHT test

Methods	**Shewart**			**CUSUM**		**SNHT**
	σ	**2σ**	**3σ**	**4 unit**	**σ**	
Homogeneous	0	0	48	34	60	43
Inhomogeneous	60	60	12	26	0	7
Non-testable and Inconsistent	-	-	-	-	-	8

Finally, although both the CUSUM and Shewart Charts have promising performances, the dependency problem in the analysis still remains. To overcome this problem, as an extension of this study, we are going to use the robust Modified CUSUM or EWMAST Charts which can also take both the challenge of dependency and variance shift into account.

Acknowledgements: The authors would like to thank to all members of the NINLIL project group of the Statistics Department, Middle East Technical University, for their valuable comments on this study, and particularly, to the members Elçin Kartal-Koç, Fidan Fahmi, and Sipan Aslan for preparing the data used in the study.

REFERENCES

[1] Alexandersson H. (1986), *A homogeneity test applied to precipitation data*. Int. J. Climatology, 6, 661–675.

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

[3] Fahmi F., Kartal E., İyigün C., Türkeş M., Yozgatlıgil C., Purutçuoğlu V., Batmaz İ., Köksal G. (2011), *Determining the Climate Zones of Turkey by Center-Based Clustering Methods.* In Nonlinear Dynamics of Complex Systems: Applications in Physical, Biological and Financial Systems. J.A. Tenreiro Machado,
Baleanu, D. and A. Luo (Eds.). Berlin: Springer (*In print*).

[4] Göktürk O. M., Bozkurt D., Lütfi Ö., Karaca M. (2008), *Quality control and homogeneity of Turkish precipitation data*. Hydrological Processes 22, 3210 - 3218.

[5] Hollander M., Wolfe A. D. (1999), *Nonparametric Statistical Methods*, Wiley, New York.

[6] Montgomery, D. (2004), *Introduction to Statistical Quality Control*, Wiley, New York.

[7] Rodionov S. N. (2005), *A brief overview of the regime shift detection methods. In: Large-Scale Disturbances (Regime Shifts) and Recovery in Aquatic Ecosystems: Challenges for Management Toward Sustainability*, V. Velikova and N. Chipev (Eds.), UNESCO-ROSTE/BAS Workshop on Regime Shifts, 14-16 June 2005, Varna, Bulgaria, 17-24.

[8] Tayanç M., Dalfes N., Karaca M., Yenigün O. (1998). *A comparative assessment of different methods for detecting inhomogeneties in Turkish temperature data set*. International Journal of Climatology 8, 561–578.

[9] Türkeş M. (2010), *Climatology and Meteorology*, Kriter Publisher, Istanbul.

[10] Yozgatlıgil C., Purutçuoğlu V., Yazıcı C., Batmaz İ. (2010), *Plausibility of the SNHT on Turkish Precipitation Data, 7th National Symposium on Statistical Days, 32-34, 28 - 30, June, Ankara, Turkey.*

[11] Yozgatlıgil C., Purutçuoğlu V., Yazıcı C., Batmaz İ. (2011), *Validity of Homogeneity Tests for Meteorological Time Series Data: A Simulation Study, 58th ISI World Statistics Congress, Dublin, August Ireland*. (Accepted for presentation).

[12] Von Neumann J. (1941), *Distribution of the ratio of the mean square successive difference to the variance.* Annals of Mathematical Statistics, 13, 367–395.