Monte Carlo experiments on the effect of serial correlation on the Mann-Kendall test of trend

ASHWINI KULKARNI, Poona, India, and HANS VON STORCH, Hamburg

Summary. The effect of serial correlation on the performance of the Mann-Kendall test on the presence of a trend is examined by means of a Monte Carlo simulation. Even moderate serial correlation makes the test liberal so that it signals erroneously the presence of significant trends more often than permitted according to the significance level. For time series with an autocorrelation similar to that of an AR(1)-process a simple "prewhitening" procedure is proposed. The approach is demonstrated with a time series of annual mean sea-level air-pressure from Bombay.

Simulationsexperimente zur Wirkung serieller Korrelation auf den Mann-Kendall Trendtest

Zusammenfassung. Die Wirkung von zeitlicher (serieller) Korrelation in einer Zeitserie auf das Resultat eines Trendtests nach Mann-Kendall wird im Rahmen eines Simulationsexperiments untersucht. Schon geringfügige Korrelationen lassen die fehlerhafte Identifikation von Trends auf Raten deutlich oberhalb der zugelassenen Irrtumshäufigkeit anwachsen. Das Problem kann für Zeitserien, deren zeitliche Statistik denen von AR(1)-Prozessen ähneln, durch einen „Prewhitening"-Ansatz gelöst werden. Der Vorgang wird anhand der Zeitserie der jährlichen Mittelwerte des Luftdruckes in Bombay demonstriert.

1 Introduction

The correlation of data both in time and space, plays an important role in climate research for two reasons. The correlation is most welcome for the reconstruction of the space-time state of the atmosphere and the ocean from a limited number of observations. However, in statistical inference problems, when conclusions about the hypothetical underlying true structure are to be drawn from finite samples then correlations often represent a nuisance because almost all classical statistical hypothesis tests and confidence limits require that the data are derived from "random" experiments.

Sometimes confusion arises from the word "randomness" which can be interpreted in two ways, namely as "iid" = "independently sampled and identically distributed" or as "drawn from a stationary process". Obviously, the latter condition is less demanding than the first. In climatological applications generally the *iid*-condition is not satisfied. The data are correlated in time (and space).

Most statistical techniques need the *iid*-type of randomness.

0941-2948/95/0004-0082 $ 01.80

An example is the classical student's *t*-test for testing whether the means of two random variables are identical or not. If the observations are not *iid*, i.e. if the data are serially correlated (in time), a popular but insufficient "cure" is to replace the sample size by the "equivalent sample size". When done properly the *t*-test becomes conservative (i.e. the null hypothesis of equal means is too seldomly rejected) and when equivalent sample size is optimized the test becomes liberal (i.e. the null hypothesis is too often rejected) (THIÉBAUX and ZWIERS 1984, ZWIERS and VON STORCH 1995). For situations, in which the serial correlation resembles the auto-correlation of an AR(1)-process, ZWIERS and VON STORCH (1995) have proposed to replace the critical values derived from a *t*-distribution by empirically determined critical values. These values are listed and therefore the test is named "table-look-up-test".

Still often people ignore the *iid*-condition. In the present paper we demonstrate the importance of serial correlation for the proper performance of the conventional Mann-Kendall test for the detection of trends. We offer a cure to "repair" the Mann-Kendall test. Similarly to the above-mentioned "table-look-up-test" is this cure applicable only when the serial correlation mimicks the auto-correlation of an AR(1)-process. This condition is not always fulfilled.

2 Mann-Kendall test

The Mann-Kendall test (SNEYERS 1975) evaluates whether a series of "random" observations is consistent with the presence of a trend. The null hypothesis is

H_0: All observations are drawn from the same random variable (1)

As alternative hypothesis is set

H_A: the data exhibit a trend. (2)

Obviously, this alternative is not "not H_0" so that additional information is required to exclude the possibility that the null hypothesis is rejected because of, for instance, a jump or cyclo-stationary behaviour.

The distribution of the test-statistic is derived under the explicit assumption that any two observations are mutually independent, or, in other words, that the time series of observations has zero auto-correlation. In that case, the Mann-Kendall test-statistic is asymptotically distributed as the standard normal distribution whenever the null hypothesis is valid. Therefore all values of the test statistic which are larger (or smaller) than the thresholds of the Normal distribution are accepted as a statistical proof that there is a "significant" trend.

But in climatological applications the *iid*-condition is often not satisfied. (An exception are data which are sampled with sufficiently large gaps, such as precipitation from one January to the next.) The data are generally correlated in time. Therefore there is another alternative hypothesis concurrent with (2)

H_{A*}: the data are serially correlated (3)

Often the option H_{A*} is tacitly disregarded and the interpretation is limited to H_A. However disregarding serial correlation can cause severe errors in the performance of the test. When the observations are correlated in time, the Mann-Kendall test becomes liberal and rejects the null hypothesis on weaker evidence than is implied by the significance level.

In the next section we demonstrate this fact by a series of Monte-Carlo simulations with an AR(1)-process with different values of the autoregressive coefficient α. We have selected the model of an AR(1)-process since many variables, in particular those averaged in time, can be represented in such a manner (see, for instance, FRANKIGNOUL 1995). But, certainly, there are many climatological processes which can not be represented by an AR(1)-process.

3 Monte-Carlo simulations

An auto-regressive process of first order, or, an AR(1)-process is given by

$$X_t = \alpha X_{t-1} + N_t \qquad (4)$$

where N_t is a series of standard normal variates which are independent of X_{t-k}, for $k \geq 1$ and which are serially uncorrelated ("Gaussian white noise").

We have generated 1000 statistically identical but independent time series of length[1] $n = 100$ and $n = 200$ and performed the Mann-Kendall test for various values of $\alpha = 0.0, 0.05, 0.10, \ldots 0.95$. A risk (of erroneously rejecting the null hypothesis) of 5% is arbitrarily chosen. Since there is no trend in the 1000 time series, we expect a rejection rate of 5 % when applying the Mann-Kendall test to the 1000 time series. Thus 50 of the 1000 time series should be declared to exhibit a trend. In fact this rate is obtained for $\alpha < 0.1$, while for $\alpha = 0.3$ the rejection rate is three times the nominal rate, namely 15% (see Fig. 1). For larger α the rejection rate increases rapidly.

A simple cure is to filter the AR(1)-series by

$$Y_T = X_t - \hat{\alpha} X_{t-1} \qquad (5)$$

where $\hat{\alpha}$ is the estimated autocorrelation at lag 1:

$$\alpha = \frac{\Sigma_{t=2}^{T} X_t X_{t-1}}{\Sigma_{t=1}^{T} X_t^2} \qquad (6)$$

The Mann-Kendall test is then applied to Y_t.

The "AR(1)-filter" (5) takes out the serially auto-correlated part out of the time series and produces an *iid*-series provided that the process X_t is AR(1). Thus, the filter transforms the originally "red" time series in a "white" time series — therefore such a filter is sometimes called to "prewhiten" the time series X_t.

A Monte Carlo experiment has been made with the AR(1)-filter (5) and 1000 prewhitened auto-correlated time series. The resulting rejection rates of the correct null hypothesis after the filter operation are remarkably close to the nominal value of 5% (Fig. 1). The approach fails only for very large α-values and for shot time series.

The filter (5) has a non-zero effect on the trend which is hoped to be identified in the process. Therefore we examined the power of the Mann-Kendall test by applying the test to time series, which were composed by an AR(1)-process with some α and a non-zero trend (0.003 × t), without and with prewhitening (5) (Fig. 1).

The rejection rates (i.e. the power) for the test without prewhitening are shown in Fig. 1. For $n = 100$ samples the power is comparable to the risk; for larger samples the trend is always correctly identified (power = 1) but for $\alpha \geq 0.7$ the power decreases. After filtration with (5) for the power is almost independent of α and, conditional upon the time series length, markedly larger than the risk.

An alternative to the "unconditional" filtration (5) is to prewhiten the time series only after the time series has been tested for non-zero correlation with a test such as the Wald-Wolfovitz test (SNEYERS 1975). For clarity, we call the risk used in the Wald-Wolfovitz test as "WW-risk". Then the procedure of the "conditional filtering" is:

- Test of the time series whether it exhibits a serial correlation with sufficiently small WW-risk β.
- If the serial correlation is found to be significant, with a WW-risk of β, the time series is filtered; otherwise not. Afterwards the Mann-Kendall test is applied to the, filtered or unfiltered, time series.

[1] The choice of $n = 100$–200 is motivated by the fact that the lengths of many observational records of annual statistics used in climatological analyses vary in this range.

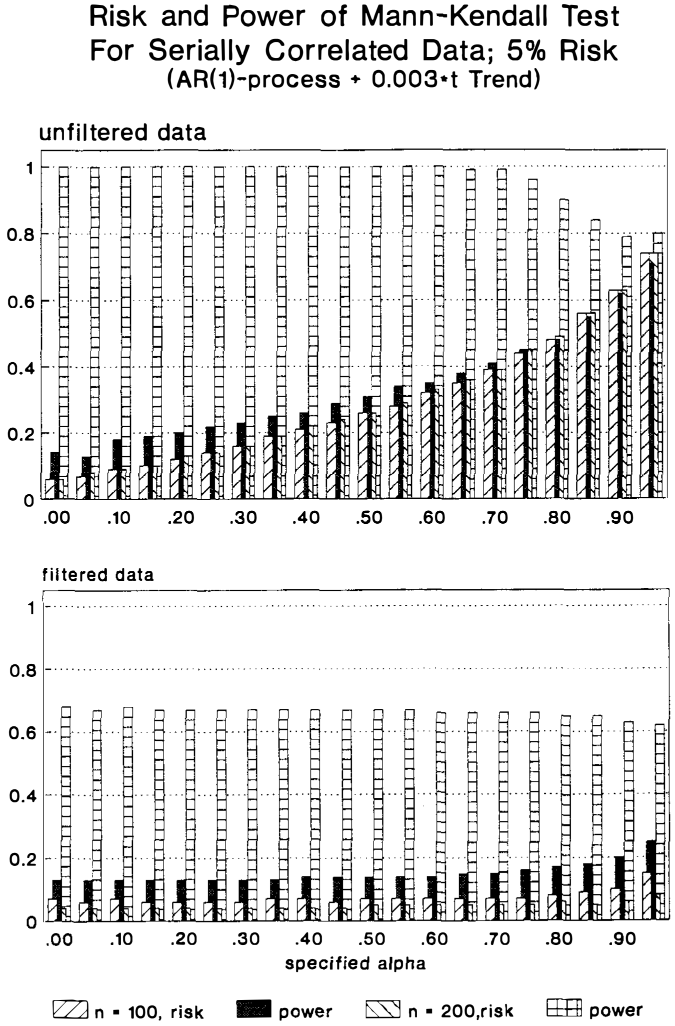

Fig. 1. Probability to reject the null hypothesis of "no trend" with the Mann-Kendall test for 1000 samples of cases without a trend ("risk") and for 1000 samples with a prescribed trend (0.003 × t; "power"). Different time series lengths (n = 100 and n = 200) and different AR-coefficients α are prescribed. — Top: Results obtained with unmodified data. Bottom: Results after "prewhitening" (5) the data prior to the test.

Abb. 1. Häufigkeit der Zurückweisung der Nullhypothese „kein Trend" durch den Mann-Kendall Test. Die Häufigkeiten wurden abgeleitet aus 1000 zufälligen Zeitserien der Länge n = 100 und n = 200 mit verschiedenen AR-Koeffizienten α. Die mit „risk" gekennzeichneten Balken betreffen fehlerhafte Zurückweisungen für Zeitserien ohne Trend, die mit „power" gekennzeichneten Balken zutreffende Zurückweisungen für Zeitserien mit einem Trend von 0.003 × t. — Oben: Häufigkeiten abgeleitet aus unbearbeiteten Zeitreihen. Unten: Häufigkeiten nach Durchführung des „Prewhitening".

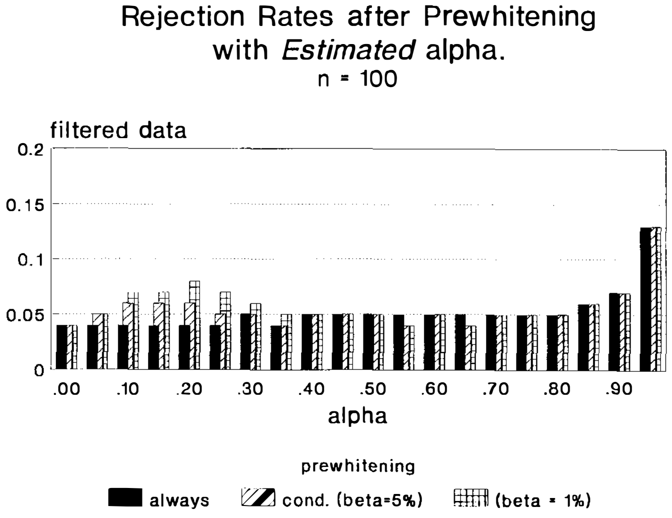

Fig. 2. Risk of the Mann-Kendall test to incorrectly reject the null hypothesis of no trend for different AR-coefficients 97 after unconditional ("always") or conditional prewhitening. For the conditional prewhitening two risks β are used in the Wald-Wolfovitz test, namely 1% and 5%.

Abb. 2. Häufigkeit fehlerhafter Zurückweisungen der Nullhypothese „kein Trend" nach Durchführung der „Prewhitening"-Filterung für verschiedene AR-Koeffizienten α. Die verschiedenen Balken beziehen sich auf drei verschiedene Filterstrategien: entweder wird stets gefiltert, oder nur wenn der Wald-Wolfovitz Test eine „signifikante" serielle Korrelation mit einem Risiko von β = 5% oder 1% diagnostiziert hat.

The test for serial correlation will correctly reject the null hypothesis for longer time series, for larger auto-correlations and for larger WW-risks β. Therefore differences between the performances of the unconditional and conditional filtering procedure will be largest for small auto-correlations, shorter time series and small WW-risks β. The result of a Monte Carlo study, with n = 100 and β = 1 % β and 5% reveals that for auto-correlations of 0.1–0.2 the rejection rate of the Mann-Kendall test is too large (Fig. 2).

The recommend to use the unconditional prewhitening procedure.

4 An example

We have tried our procedure on annual average mean sea-level pressure data at Bombay (18°N, 72°E) in India. The data have been quality-controlled and homogenized (PARTHASARATHY et al. 1991). The time series of 144 years is used for the period 1847–1990. This series is correlated in time with serial correlation of $\hat{\alpha} = 0.46$. When disregarding this serial correlation the Mann-Kendall test-statistic of –6.26 is beyond the 99.9%-threshold so that the null hypothesis of no trend is declared inconsistent with the data with a risk of less than 0.1%. After prewhitening the series as AR(1) with $\hat{\alpha} = 0.46$, the filtered series show little serial correlation ($\hat{\alpha} = -0.03$) and then the Mann-Kendall test statistic value for the filtered series returns a reduced but still highly significant Mann-Kendall test statistic –3.92. The smallness of the sample correlation of $\hat{\alpha} = -0.03$ indicates that the filter procedure to produce an uncorrelated time series was successful.

5 Conclusions

We have seen that even small serial correlations of the order of 0.3 cause severe malfunctions of the Mann-Kendall test. The alternative hypothesis of "there is a trend" is too often accepted on false evidence than specified by the nominal significance level. If, however, the serial correlation is less than 0.1 one may safely consider the observations as *iid* if the times series is stationary and long enough (at least 100).

We have examined two "prewhitening"-approaches to overcome the problems introduced by serial correlation. The idea is to model the time series as an AR(1)-process, and to subtract the "memory" from the time series. In the "unconditional" approach this is done with whatever serial correlation is estimated, in the "conditional" approach the prewhitening is invoked only when the serial correlation is found to be statistically significant. Both approaches return similar results, but for smaller sample sizes and smaller correlation the unconditional approach performs better.

The advice for a prewhitening concerns only the statistical test on the presence of a trend; it does not concern the estimation of the strength of the trend.

It is important to understand that the success of the prewhitening depends crucially on the assumption that the serial correlation of the considered process is similar to that of an AR(1)-process. If dynamical or other reasons suggest that the process has not a red spectrum, so that the assumption of an AR(1)-process is inadequate, other prewhitening procedures, for instance by fitting higher-order AR-processes, might be useful. In that case, it is advisable to perform a few Monte Carlo experiments similar to ours.

Similar problems with the serial correlation occur with other tests, such as the *t*-test mentioned in the Introduction (Zwiers and von Storch 1995) or with the Pettitt-test for the detection of "change points" (Busuioc and von Storch 1995).

An implication of the bias of the Mann-Kendall test concerns the "field-testing" problem when the Mann-Kendall test is done for time series from many locations simultaneously and independently (see the discussion, for instance, by Livezey (1995)). If there are no trends in the data, the tests should return a "significant trend" at 5% (or whatever the significance level is) of points, on an average. By chance this number may be much larger (von Storch 1982). This problem of "multiplicity" becomes worse when the local test, i.e., the Mann-Kendall test applied to one point, is biased as it happens to be in case of serially correlated data. Then the possible size of the area of false rejections will on average be larger than 5% of the total area. In case of an AR(1)-process with $\alpha = 0.3$ this area will be 15%. Therefore extra care is required to assess the "field significance" (e.g., Livezey 1995) of a field of trend tests.

Maybe, the most important lesson to be learned from our case is the danger which is inherent in a less than precise usage of statistical expressions such as "random". It would be best to avoid this expression altogether and to use the precise expressions "iid" or "stationary random process" instead.

References

Busuioc, A., H. von Storch, 1995: Changes in the winter precipitation in Romania and its relation to the large-scale circulation. — Max-Planck-Inst. f. Meteorol. Rep. 151.

Frankignoul, C., 1995: Climate spectra and stochastic climate models. — In: von Storch, H., A. Navarra (eds), Analysis of Climate Variability: Applications of Statistical Techniques. — Springer Verlag (in press).

Livezey, R. E., 1995: Field intercomparisons. — In: von Storch, H., A. Navarra (eds), Analysis of Climate Variability: Applications of Statistical Techniques. — Springer Verlag (in press).

Parthasarathy, B., K. Rupakumar, A. A. Munot, 1991: Evidence of secular variations in Indian Monsoon rainfall. — J. Climate 4, 927–938.

Sneyers, R., 1975: Sur l'analyse statistique des series d'observations. — Note technique No. 143, WMO, 189 pp.

Thiébaux, J., F. W. Zwiers, 1984: The interpretation and estimation of effective sample sizes. — J. Climate Appl. Meteor. 23, 800–811.

von Storch, H., 1982: A remark on Chervin/Schneider's algorithm to test significance of climate experiments with GCMs. — J. Atmos. Sci. 39, 187–189.

Zwiers, F. W., H. von Storch, 1995: Taking serial correlation into account in tests of the mean. — J. Climate **8** (in press).

Ashwini Kulkarni
Indian Institute of Tropical
Meteorology
Poona, India

Hans von Storch
Max-Planck-Institut
für Meteorologie
Bundesstraße 55
D-20146 Hamburg

Received 26 October 1994, in revised form: 6 January 1995