Statistical analysis of data with outliers

1 Introduction

This document deals with mathematics for statistical analysis of data with large outliers. It was published as six blog posts in February and March 2017, see http://hpklima.blogspot.no/2017/02/.

An earlier document, Statistical analysis of global temperatures, deals with mathematics that is most commonly used when analysing global temperature series. That mathematics is not well suited when there are large outliers in the data.

Ordinary least square (OLS) error mathematics is the most commonly used method to calculate trends. It is based on data **values**, and it therefore performs poorly when there are large outliers in the data. Global temperatures do not have large outliers due to both the inertia in the global climate system and due to the thorough processing before the temperature data is released. Other climate data, such as precipitation, snow depth and skiing conditions at specific locations, have large outliers, and the OLS mathematics is not suitable for those data.

The calculation of the Pearson correlation coefficient is also based on data **values**. This is the most commonly used method to calculate correlation between variables. It too performs poorly when there are large outliers in the data.

Mathematics based on data **ranks** performs better than mathematics based on data values when analysing data with large outliers. In this document I will describe the rank mathematics that I use to calculate the Kendall tau-b correlation coefficient and the Kendall-Theil robust trend line. For comparison I will also shortly describe the Pearson and the OLS mathematics.

As will be seen, the mathematics that is used to calculate the Kendall tau-b correlation coefficient and the Kendall-Theil robust trend line is rather simple and easy to explain. But the mathematics that is used to quantify their uncertainties, which are p-values and confidence intervals, is more complicated.

Contents

2 Correlation when outliers in the data

2.1 Pearson

The Pearson correlation coefficient between x and y is the covariance between them divided by the square root of the product of their variances. It is usually denoted by r. It may be calculated as shown in (2.1). n is the number of x and y values.

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2.1}$$

A large outlier in either x or y will have different impacts on the numerator and on the denominator in (2.1). The Pearson correlation coefficient is therefore sensitive to outliers in the data, and it is therefore not robust against them.

p-value

The null hypothesis H_0 is that r is zero, and the alternative hypothesis H_1 is that it is different from zero, positive or negative. The p-value of r is the probability to get such a large correlation coefficient, positive or negative, if the null hypothesis were true.

The t-value (2.2) is distributed approximately as a Student's t distribution with n-2 degrees of freedom under the null hypothesis. The F() function in (2.3) is the cumulative distribution function of the Student's t distribution with n-2 degrees of freedom.

$$t = |r| \sqrt{\frac{n-2}{1-r^2}} \tag{2.2}$$

$$p-value = 2(1-F(|t|)) \tag{2.3}$$

The Pearson equations are included in this document to demonstrate that the r value is not robust against outliers and because the equations are referred to in the chapter on the Spearman's rank correlation coefficient.

Reference for the mathematics

Hans von Storch, Francis W. Zwiers: Statistical Analysis in Climate Research, ISBN 0 511 01018 4 virtual (netLibrary Edition), shows the equations for the Pearson correlation coefficient in chapter 8.2.

2.2 Spearman

The Spearman rank correlation coefficient between x and y is calculated based on the **ranks** of the x and y data instead of on their data **values**. It is usually denoted by the Greek letter ρ (rho).

It is calculated just as the Pearson correlation coefficient in (2.1), except that the x and y values are replaced with their ranks. The p-value is calculated as the Pearson p-value in (2.2) and (2.3), except that the correlation coefficient applied in (2.2) is Spearman's rho and not Pearson's r.

The ranks are assigned to the values in ascending order. Equal values form a set of ties. They get the same ranks with Fractional ranking, as illustrated in the example below.

	Values						**Ranks**					
x	1	4	4	11	15	19	1	2.5	2.5	4	5	6
y	5	5	3	-20	10	200	3.5	3.5	2	1	5	6

2.3 Kendall

The calculation of the Kendall rank correlation coefficient compares each xy pair with all the xy pairs that follow. A change in the same direction for both x and y is a contribution towards positive correlation, a change in the opposite direction is a contribution towards negative correlation, and no change in x or y or in both of them is a contribution towards no correlation. This is regardless of how far they are from each other in rank; this differs Kendall from Spearmann. The Kendall coefficient is usually denoted by the greek letter τ (tau).

First the S value is calculated (2.4). n is the number of xy pairs. The sign() function returns +1 when its input parameter is positive, -1 when it is negative, and 0 when it is zero. In the latter case either the x or the y values or both of them are equal. Equal values are denoted as tied values.

T_0 is the number of xy pairs that are compared in (2.4). It is the maximum value of S, and $-T_0$ is the minimum value of S. T_0 is defined in (2.5).

The correlation coefficient is calculated as either tau-a or tau-b. Tau-b compensates for tied values, while tau-a does not do that. (2.6) shows how tau-a is calculated.

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (sign(x_i - x_j)) * (sign(y_i - y_j)) \tag{2.4}$$

$$T_0 = \binom{n}{2} = n*(n-1)/2 \tag{2.5}$$

$$tau\,a = \frac{S}{T_0} \tag{2.6}$$

T_0 (2.5) is the binominal coefficient 'n choose 2'. It tells the number of ways to choose a subset of 2 elements, disregarding their order, from a set of n elements. It is equal to the total number of summations in (2.4). If all these summations add 1 to S, the numerator and the denominator in (2.6) are equal to each other and tau-a becomes 1.

Tied values do not contribute to S in the numerator in (2.6), but they are part of n and they therefore contribute to T_0 in the denominator. When calculating tau-b this mismatch is compensated for, as shown in (2.7) to (2.9). In these equations the tied x and y values reduce the denominator too.

T_x (2.7) is calculated based on the px groups of tied x values. Each group consists of $t_{x,i}$ data values. T_x tells the number of times tied x values causes a zero in (2.4).
T_y is calculated similarly based on the tied y values.

$$T_x = \sum_{i=1}^{px} (t_{x,i} * (t_{x,i} - 1)/2) \tag{2.7}$$

$$T_y = \sum_{j=1}^{py} (t_{y,j} * (t_{y,j} - 1)/2) \tag{2.8}$$

$$tau\,b = \frac{S}{\sqrt{(T_0 - T_x)(T_0 - T_x)}} \tag{2.9}$$

I will now discuss how to determine whether a calculated tau-b is statistically significantly different from zero.

The Central Limit Theorem tells that the distribution of a sum of random variables tends toward a normal distribution when the number of additions is large, even if the original variables themselves are not normally distributed. Therefore, when the number of x and y variables is greater than 10, S approximately follows a standard normal distribution under the null hypothesis.

(2.10) to (2.14) define the variables that are used to calculate the standard deviation of S in (2.15). $t_{x,i}$, px, $t_{y,i}$ and py in (2.11) to (2.14) have the same meaning as explained for (2.7) and (2.8).

$$(2.10) \qquad m_0 = n(n-1)(2n+5)$$

$$(2.11) \qquad m_x = \sum_{i=1}^{px} t_{x,i}(t_{x,i}-1)(2t_{x,i}+5)$$

$$(2.12) \qquad m_y = \sum_{i=1}^{py} t_{y,i}(t_{y,i}-1)(2t_{y,i}+5)$$

$$(2.13) \qquad m_1 = \frac{\sum_{i=1}^{px} t_{x,i}(t_{x,i}-1) \sum_{i=1}^{py} t_{y,i}(t_{y,i}-1)}{2n(n-1)}$$

$$(2.14) \qquad m_2 = \frac{\sum_{i=1}^{px} t_{x,i}(t_{x,i}-1)(t_{x,i}-2) \sum_{i=1}^{py} t_{y,i}(t_{y,i}-1)(t_{y,i}-2)}{9n(n-1)(n-2)}$$

$$(2.15) \qquad \sigma_{S,taub} = \sqrt{\frac{m_0-m_x-m_y}{18} + m_1 + m_2}$$

$$(2.16) \qquad z_{S,taub} = \frac{S}{\sigma_{S,taub}}$$

$z_{S,\,taub}$ (2.16) is the standardized form of S. When the null hypothesis is true and n is greater than 10, $z_{S,\,taub}$ approximately follows a standard normal distribution N(0,1).

The p-value of tau-b is the probability to get such a large correlation coefficient, positive or negative, if the null hypothesis were true. It is calculated as shown in (2.17). F() is the cumulative standard normal distribution. It is multiplied by 2 because the hypothesis test is two-tailed.

$$(2.17) \qquad p\ value = 2F(-|z_{S,taub}|)$$

When the p-value is less than 0.05 the correlation is statistically significant at the 0.05 level.

References for the mathematics

'Base SAS 9.2 Procedures Guide: Statistical Procedures, Third Edition', ISBN 978-1-60764-451-4, documents the formulas that are used in the SAS/STAT Software. See the section 'Kendall's Tau-b Correlation Coefficient' in that book. The same formulas are used by the Wikipedia article Kendall rank correlation coefficient and by the blog post Kendall's Correlation Testing with Ties written by Dr. Charles Zaiontz.

3 Trend when outliers in the data

A trend line is a simple model of the connection between an independent variable X and a dependent variable Y. In our case, X is usually the time and Y is a climate value (snow depth, precipitation, days with skiing conditions, etc.).

(3.1) $$Y = a + bX$$

a is the intercept between the trend line and the Y axis, and b is the slope of the trend line.

Linear regression analysis uses a set of x_i y_i measurements to estimate a and b so that the resulting trend line is a "best fit" to the measurements.

Each y_i measurement deviates more or less from the trend line. The vertical distance e_i between the y_i measurement and the trend line is treated as the error (the residual) of that measurement.

(3.2) $$e_i = y_i - (\hat{a} + \hat{b}\, x_i)$$

The hat symbol above a and b means the estimate of those values. From now on I will omit the hat symbol.

3.1 OLS - Ordinary Least Square

Ordinary least squares (OLS) is the most commonly used method to calculate trends. It calculates the slope as shown in (3.3). The mean values of x and y are used to calculate the intercept, as shown in (3.4).

(3.3) $$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

(3.4) $$a = \overline{y} - b\,\overline{x}$$

The OLS equations show that the calculation of the slope and the intercept is very sensitive to outliers in the data, and it is therefore not robust. OLS is included in this document mainly to show that.

I derived and explained the formulas for OLS in the first three chapters in the document that I referred to in the introduction, Statistical analysis of global temperatures. The document contains references. The chapters are named *Linear regression analysis*, *Hypothesis testing of temperature trends* and *Confidence intervals around temperature trend lines*. The latter two chapters concentrate on how to quantify the uncertainties of the slope and the intercept.

3.2 Kendall-Theil robust line

The Kendall-Theil method to calculate the slope and the intercept of the trend line is based on medians, and it is therefore robust against outliers. The Kendall-Theil robust line is called by many names, as explained on this Wikipedia site.

We start with an x and a y vector, each with n values. X is the independent variable, often the time and therefore monotonically increasing. We assume that there are no ties among the X values. There are N possible combinations of the xy pairs. (3.5) shows how N is calculated.

We first calculate the slopes between all the N xy pairs. Then we select the median of these slopes, as shown in (3.6). This median is the slope of the Kendall-Theil line. The selection is usually done by first sorting the slopes in ascending order, and then picking the slope in the middle. The vector with the slopes sorted in ascending order will later be used to calculate the 95% confidence interval of the slope.

$$N = \binom{n}{2} = n(n-1)/2 \tag{3.5}$$

$$b = median \frac{y_j - y_i}{x_j - x_i} \quad \text{for all j>i ; j=2, 3,...n ; i=1,2,...(n-1)} \tag{3.6}$$

The intercept is calculated using the slope from (3.6) and the median of the x and the y values, as shown in (3.7). The Kendall-Theil trend line passes through the point with the coordinates (median x, median y), just as the OLS trend line passes through the point with the coordinates (mean x, mean y).

$$a = y_{median} - b * x_{median} \tag{3.7}$$

An alternative method to calculate the intercept is to calculate the intercepts of all the N slopes between the xy pairs, and then select the median of these intercepts. But (3.7) is evaluated to be more robust, and it is therefore the preferred equation.

Mann-Kendall trend test

The calculation of the uncertainty of the Kendall tau-b correlation coefficient (chapter 2.3) and the calculation of the uncertainty of the slope of the Kendall-Theil robust line are almost identical. Both use the S value as calculated in (2.4), and both assume that the null hypothesis is true. The formulas to calculate the uncertainty of the Kendall tau-b are, however, more complicated because there may be ties in both x and y. When the slope of the Kendall-Theil robust line is calculated, there can be no ties among the x values. Otherwise there would be one or more zeros in the denominator in (3.6).

The equations below use S as calculated in (2.4), m_0 in (2.10) and m_y in (2.12).

(3.8) calculates the standard deviation of S.

$$\sigma_S = \sqrt{\frac{m_0 - m_y}{18}} \tag{3.8}$$

Z_S is the standardized form of S, and it is calculated as shown in (3.9). It approximately follows a standard normal distribution N(0,1) when the null hypothesis is true and n is greater than 10.

$$z_S = \begin{cases} \dfrac{S-1}{\sigma_S} & when\ S > 0 \\ 0 & when\ S = 0 \\ \dfrac{S+1}{\sigma_S} & when\ S < 0 \end{cases} \tag{3.9}$$

The p-value of the trend is the probability to get such a large slope, positive or negative, if the null hypothesis were true. It is calculated as shown in (3.10). F() is the cumulative standard normal distribution. It is multiplied with 2 because the hypothesis test is two-tailed.

$$p\ value = 2F(-|z_S|) \tag{3.10}$$

When the p-value is less than 0.05 the trend is statistically significant at the 0.05 level.

Serial correlation

The residual vector **e** is defined in (3.11).

$$\boldsymbol{e} = \boldsymbol{y} - (a + b\,\boldsymbol{x}) \tag{3.11}$$

Positive serial correlation occurs when we have measured (sampled) the y values so often that the noise in a measurement is not independent of the noise in the previous measurement. Then consecutive residuals in the **e** vector tend to be of the same sign and size. Then the effective (independent) number of measurements is less than the actual number of measurements, and the uncertainty of the trend is therefore larger than calculated with the equations (3.8) to (3.10). This may be compensated for.

(3.8) calculates σ_S , which is the one sigma standard deviation of the S statistics. When there is positive serial correlation in the residuals, (3.8) it too optimistic; i.e. σ_S should have been larger. We therefore modify σ_S as shown in (3.12) to (3.14).

$$\sigma_{S\,modified} = \sigma_S \sqrt{\frac{n}{n_{independent}}} \tag{3.12}$$

$$\frac{n}{n_{independent}} = 1 + \frac{2}{n(n-1)(n-2)} \sum_{k=1}^{nsl} (n-k)(n-k-1)(n-k-2)\, r_k^R \tag{3.13}$$

r^R_k in (3.13) is the serial correlation coefficient with lag k. The equation tells that we can use many lags (nsl is an abbreviation for **n**umber of **s**erial correlation **l**ags). In practice I will only use the lag 1 and 2 coefficients.

The R superscript tells that the serial correlation coefficient is calculated based on the ranks of the residuals. I use the notation **er** instead of **e** when the residual value is replaced with its rank. (3.14) shows how r^R_k is calculated.

$$r_k^R = \frac{\frac{1}{(n-k)} \sum_{i=1}^{n-k} (er_i - \bar{\boldsymbol{er}})(er_{i+k} - \bar{\boldsymbol{er}})}{\frac{1}{n} \sum_{i=1}^{n} (er_i - \bar{\boldsymbol{er}})^2} \tag{3.14}$$

The modified σ_S is used in (3.9), and thereafter the p-value is calculated as shown in (3.10).

The modification factor n divided with $n_{independent}$ is often denoted as v. It will vary when it is calculated with different data sets, even when the data sets are of the same type, eg. snow depth. I will therefore calculate it for many data sets of the same type, and thereafter evaluate the different factors and then decide on a factor to be used for that type of data. I never use a factor less than 1.

(The OLS mathematics compensates for serial correlation in a similar way, but the equations are not identical. Equation 2.4 in the aforementioned document Statistical analysis of global temperatures calculates a v factor which is used in the same way as the ratio between the real and the independent number of measurements in (3.12). The OLS mathematics calculates the serial correlation coefficients in the same way as in (3.14), except that it uses the residual values instead of their ranks.)

Confidence Interval

The 95% confidence interval of the Kendall-Theil slope is calculated based on the S statistics and on the vector containing the sorted slopes from which the median was selected in (3.6). The vector with the sorted slopes contains N slopes, including the slopes that are zero due to tied y values. The standard deviation of S, σ_S, is therefore not reduced due to tied y values (3.15). This σ_S is used to calculate the indexes R_u and R_l in the vector with the sorted slopes. The slope with index R_u is the upper limit of the 95 % confidence interval, and the slope with index R_l is the lower limit.

The confidence limit α is 0.05 when the 95% confidence interval is calculated. The notation with α is used in (3.16) and (3.17) to make the equations more general. $z_{\alpha/2}$ is the z value for which the cumulative distribution function of N(0,1) is equal to (1-α/2).

$$\sigma_S = \sqrt{\frac{n(n-1)(2n+5)}{18}} \tag{3.15}$$

$$R_u = \frac{N}{2} + z_{\alpha/2}\frac{\sigma_S}{2} + 1 \tag{3.16}$$

$$R_l = \frac{N}{2} - z_{\alpha/2}\frac{\sigma_S}{2} \tag{3.17}$$

Reference for the mathematics

The article Statistical Methods in Water Resources, written by D.R. Helsel and R.M. Hirsch from the U.S. Geological Survey, shows in chapter 10.1.1 how to calculate the Kendall-Theil robust trend line. Chapter 10.1.4 tells how to calculate the 95% confidence interval of the slope.

Chapter 17.3.2 in the EPA document Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities Unified Guidance tells how to calculate the z-score of the S statistics and how to use it to calculate the p-value of the slope. Chapter 17.3.3 in the same document shows how to calculate the slope and the intercept.

The article The influence of autocorrelation on the ability to detect trend in hydrological series by Sheng Yue et al explains how to compensate for serial correlation when calculating the uncertainty of the Kendall-Theil slope. (Autocorrelation is another word for serial correlation).

4 Correlation and trend when an outlier is added. An example.

This chapter contains an example that demonstrates the shortcomings of the mostly used methods to calculate trends and correlations when an outlier is added to the data. It demonstrates that alternative methods based on medians and ranks are more robust against outliers.

This chapter is based on an example in a course in statistics at Penn State. I chose this example for two reasons. Firstly the course uses other equations for the Kendall tau-b correlation coefficient than mine equations (2.7), (2.8) and (2.9) in chapter 2, and secondly it contains tied values. The course provides the answer with four decimal digits, which allows me to test that my equations are equivalent with the equations in the course and that my implementation is correct.

The x and y variables in the example are:

x:	23	23	27	27	39	41	45	49	50	53	53	54	56	57	58	58	60	61
y:	9.5	27.9	7.8	17.8	31.4	25.9	27.4	25.2	31.1	34.7	42.0	29.1	32.5	30.3	33.0	33.8	41.1	34.5

They are shown graphically in Figure 4.1.

4.1 Correlation

The correlation coefficients calculated with the equations in chapter 2 are shown in the legend in Figure 4.1. They are exactly the same as calculated by Penn State.

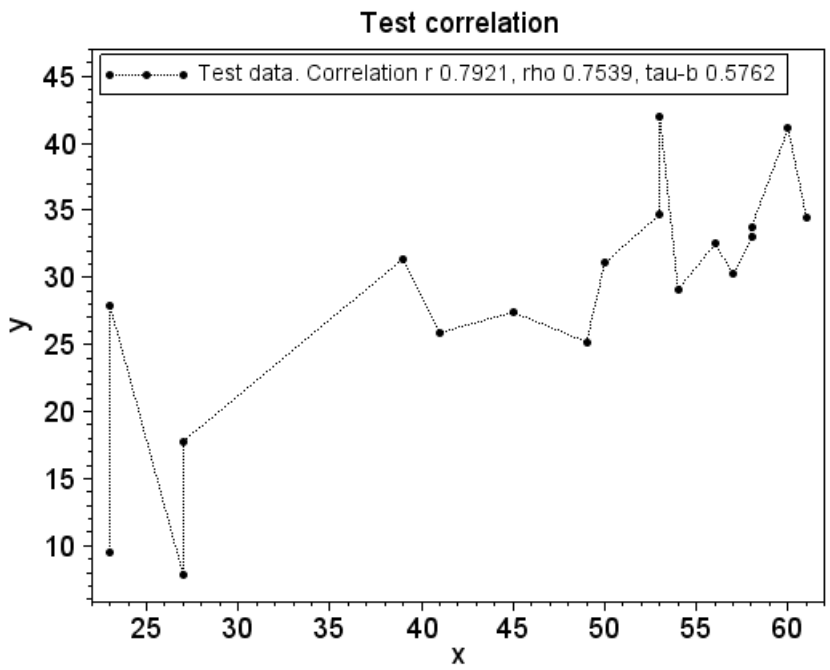

Figure 4.1: Correlation between the Penn State x and y values calculated with Pearson (r), Spearman (rho) and Kendall (tau-b). The coefficients are exactly the same as calculated by Penn State.

```
test_correlation(1);
```

The Kendall tau-b correlation coefficient is smaller in absolute value than the Pearson and Spearman correlation coefficients. According to Penn State this is usual. I have experienced the same with other data sets too. The calculated p-value is 0.0001 with Pearson, 0.0003 with Spearman and 0.0011 with Kendall.

4.2 Trend

In climate analysis, x is normally a monotonic increasing time and consequently it does not contain tied values. The equations in my chapter 3 therefore assume no tied x values when calculating trends. The Penn State example has four sets with tied x values. I therefore adjusted the second of the x values in each set by adding 0.1 to it before calculating the trends with both OLS and Kendall-Theil. The trends are shown in Figure 4.2.

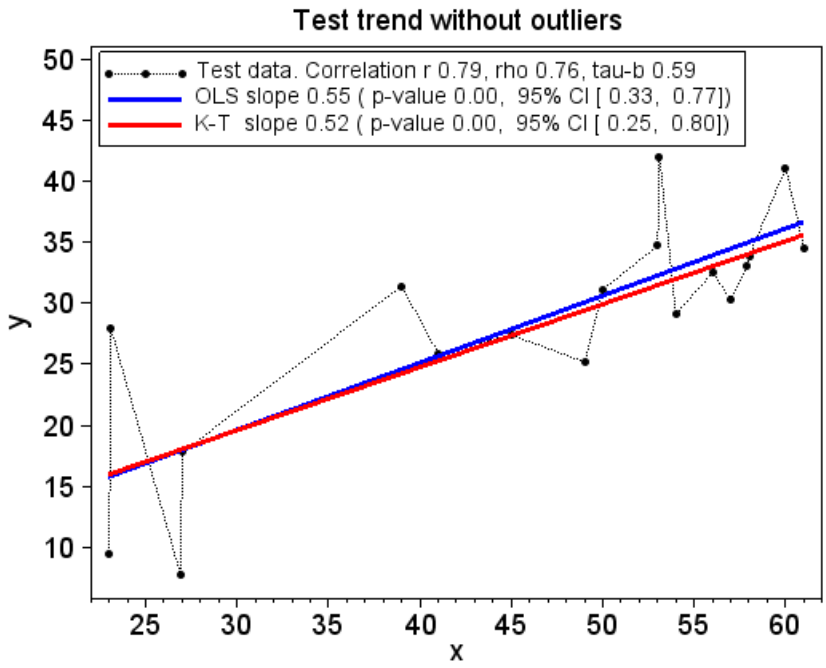

Figure 4.2: The OLS and the Kendall-Theil trends in the Penn State example. The example does not contain any outliers, and the two trends are therefore close to each other. The legend text shows the p-values and the 95% Confidence Intervals calculated with each method.

```
test_correlation(2);
```

The OLS and the Kendall-Theil trends are close to each other. Both are statistically significant, and both are well within each other's 95% confidence intervals.

The Pearson, Spearman and Kendall tau-b correlation coefficients are almost not influenced by this small adjustment of the x values. The p-values of the Spearmann and the Kendall tau-b correlation coefficients are improved (not shown) because the adjustment increases the number of xy pairs that contributes in the calculations.

4.3 Add an outlier

I changed the fifth last y value from 30.3 to -17.8, i.e. I changed it to be an outlier. Thereafter I calculated the trends anew, as shown in Figure 4.3. The OLS trend is largely influenced by the outlier. It is no longer statistically significant, and its 95% confidence interval is much wider than it was before the outlier was added. The Kendall-Theil trend is almost not influenced by the outlier, neither the slope nor its uncertainty. This demonstrates that the Kendall-Theil trend line is much

more robust against outliers than the OLS trend line is.

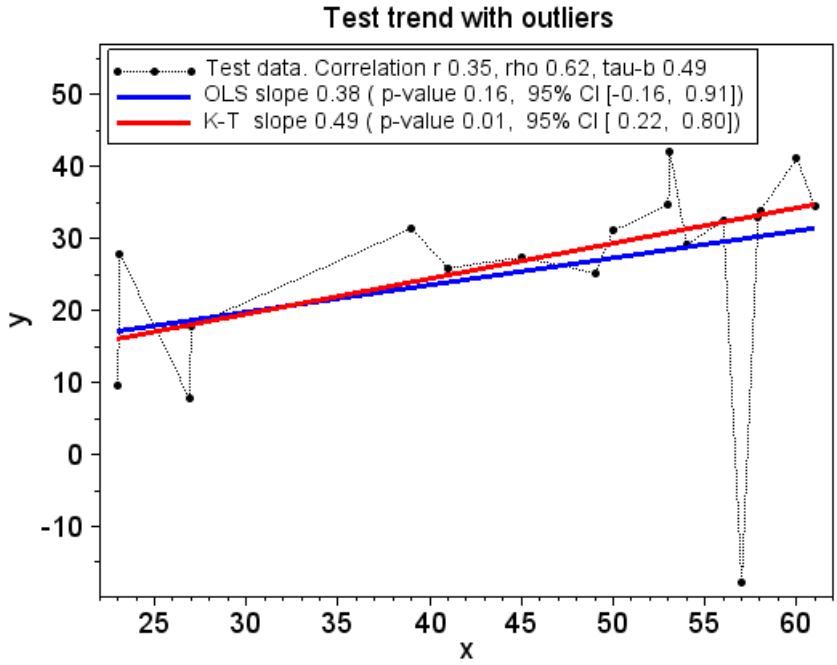

Figure 4.3: The OLS and the Kendall-Theil trends when the Penn State example is modified to contain an outlier. The OLS trend is largely influenced by the outlier, and it is no longer statistically significant. The Kendall-Theil trend is only slightly influenced by the outlier, and it is still statistically significant with good margins. Both trends are within each others 95% confidence intervals.

```
test_correlation(3);
```

The outlier reduces the correlation coefficient between x and y. With Pearson r the reduction is large, from 0.79 to 0.35. The reduction is much smaller with the two rank correlation coefficients; Spearman rho is reduced from 0.76 to 0.62 and Kendall tau-b from 0.59 to 0.49.

Without the outlier the correlation coefficients are statistically significant with all the three methods. With the outlier added, the Pearson correlation coefficient is far from being so (p-vale 0.16), while the rank correlation coefficients are statistically significant with good margins (p-value 0.01 for both Spearman and Kendall).

This demonstrates that the rank correlation coefficients are more robust against outliers than the Pearson coefficient.

Often we don't know if an outlier is an error measurement or if it is real. But anyway it is wrong to let a single or a few outliers have much more influence on the result than the majority of the measurements. Therefore, when the data contains outliers, it may be best to apply trend calculations based on medians, as the Kendall-Theil calculation is, and correlation calculations based on ranks, as the Kendall tau-b calculation is.

5 Comparison of Kendall-Theil and OLS trends

This chapter deals with Monte Carlo simulations that calculate trends. The calculations use both the Kendall-Theil (K-T) methodology and the Ordinary Least Squares (OLS) methodology. The results are compared. Both methodologies work well when the noise in the dependent variable is white. They are about equal when there is serial correlation in the dependent variable. K-T is much more robust against outliers than OLS is. The results are presented in probability density plots.

All tests are done with a monotonic increasing x vector and a y vector with the data values. x may be the time and y may be some climate data. Both x and y have 50 items, which is a usual length in climate analysis. Each test is done in 10000 Monte Carlo simulations, with a new y vector in each simulation. The results of each test are shown as a probability density plot.

The independent variable x is the same in each simulation. The underlying trend in the dependent variable y is increasing with a slope of 0.1. Before each simulation a new noise vector is added to this underlying trend. The noise vectors in the same test have the same characteristics.

The noise vector is first filled with random numbers drawn from the standard normal distribution N(0,1). This is white noise. If the noise vector shall contain serial correlation, a first order Markov process with α equal to 0.20 is added to the vector. Thereafter, if the noise vector shall contain outliers, 4% of the noise values, randomly selected, are multiplied by 20. These definitions of serial correlation and/or outliers added to the noise are used throughout the document.

Random numbers are also called white noise. Serial correlation is also called coloured noise. A first order Markov process is also called an AR(1) process. The section 5.5 Random noise, coloured noise and outliers explains in more details how white noise, coloured noise and outliers are generated in the simulations.

5.1 Only white noise

When the y vector contains white noise the K-T and the OLS methodologies both perform well, as shown in Figure 5.1.

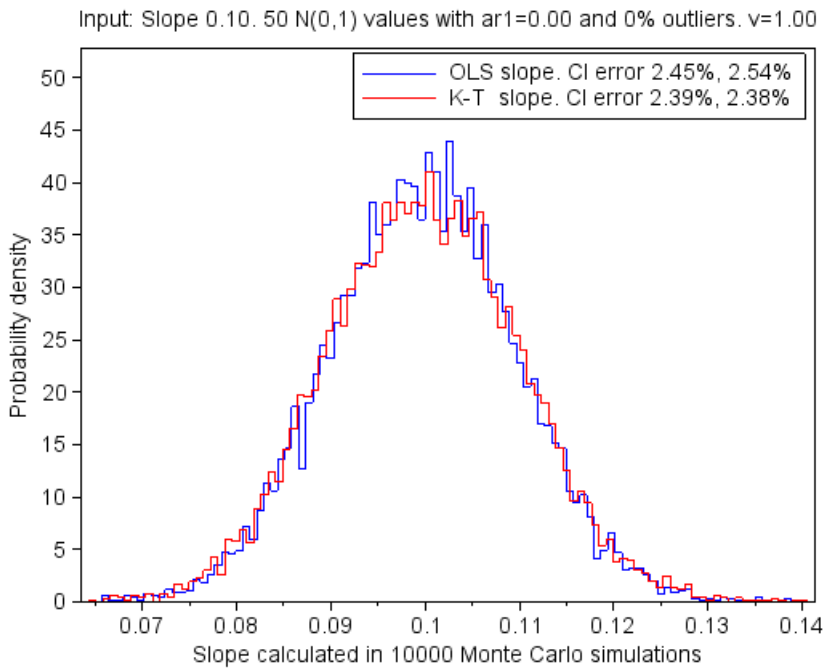

Figure 5.1: Probability density plot of the slopes calculated with Kendall-Theil (K-T) and Ordinary Least Squares (OLS).The noise in the dependent variable is white.

All plots by 'C:\Arbeid\scilab\lib\math_support\tests\test_kendall_theil.sci

The mean value of the 10000 slopes calculated with both K-T and OLS is 0.100, and the standard deviation is 0.010. The probability density plots are almost identical, and they are both close to that of the normal distribution.

Each simulation calculates the slope and its 95% Confidence Interval (CI). In theory, it is 2.5% probability that the correct slope is less than the lower limit of the 95% CI, and it is 2.5% probability that the correct slope is larger then the upper limit of the 95% CI. We know the correct slope in our simulations. We can therefore calculate the percentage of the simulations in which the correct slope is less than the lower limit of the calculated 95% CI, and the percentage in which it is larger than the upper limit. These two percentages are displayed in the legend in Figure 5.1 as '*CI error*'. The expected percentage is 2.5% for both. Large deviations from these expected values indicate that the methodology used in the calculations matches poorly with the nature of the noise.

The CI error percentages for both K-T and OLS in Figure 5.1 are close to the expected values. Both methodologies handle white noise well.

5.2 Outliers in the noise

Figure 5.2 shows that the Kendall-Theil methodology is more robust against outliers than the OLS methodology is. The mean values of the calculated slopes are close to the correct slope for both methodologies, but the spread of the calculated slopes is almost four times larger with the OLS methodology than it is with the K-T methodology.

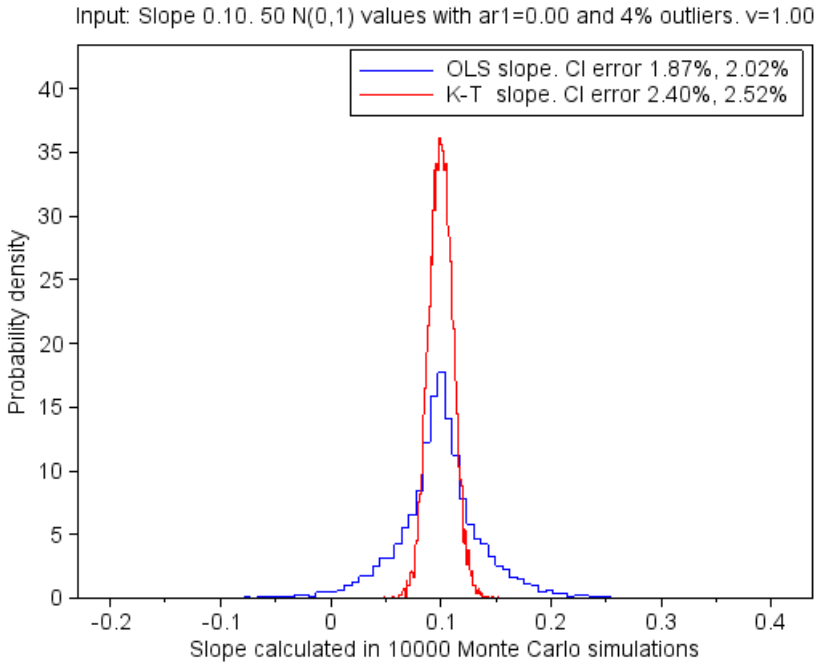

Figure 5.2: Probability density plot of the slopes calculated with Kendall-Theil (K-T) and Ordinary Least Squares (OLS).The noise in the dependent variable is white with outliers added to it.

The mean of the 10000 slopes calculated with K-T is 0.100, and the standard deviation is 0.011. The outliers caused only a minor increase in the standard deviation; from 0.010 to 0.011.

The mean value of the 10000 slopes calculated with OLS is 0.099 with standard deviation 0.041. The outliers cause a large increase in the standard deviation; from 0.010 to 0.041. This shows that OLS is not robust against outliers.

The probability density plot of the K-T slopes is close to that of the normal distribution with standard deviation 0.011, while the probability density plot of the OLS slopes is much thinner than that of the normal distribution with standard deviation 0.041.

The legend text in Figure 5.2 tells that the 95% confidence intervals of the OLS slopes are a little too narrow; 3.89% (1.87+2.02) of them do not cover the correct slope. The expected percentage is 5%. The 95% confidence interval of the K-T slopes is spot on, with 4.92% of them not covering the correct slope.

5.3 Serial correlation in the noise

Positive serial correlation in the y data means that parts of the noise in one measurement tend to be present also in the next measurement. Then the number of effective measurements is less than the number of real measurements. This may be compensated for by calculating a multiplication factor v that is used to increase the uncertainty. For the OLS methodology the formula 2.4 in the aforementioned document Statistical analysis of global temperatures is used to calculate v, and for the K-T methodology the formula (3.13) in this document is used.

I generate the serial correlation by adding a first order Markov process with α equal to 0.20 to the white noise in the y data. Then the expected lag 1 serial correlation coefficient is 0.2, and the expected lag 2 serial correlation coefficient is 0.04. With these coefficients the expected OLS v is **1.50** and the expected K-T v is **1.45**.

The functions that calculate the OLS and the K-T trends can either calculate the compensation factor v individually for each set of 50 y-values, or they can use a preset value based on earlier experience and analysis. I recommend the latter approach, but for test purposes I first let the functions calculate the v value individually for each set of 50 y elements, and thereafter I calculate the mean value of them. The simulations with the OLS methodology gave a mean value of 1.75 with standard deviation 1.52. The corresponding numbers for the K-T methodology are a mean value of 1.29 with standard deviation 0.32. In these calculations, v values less than 1 are replaced with 1. This increases the mean value of v. We know from experience that we calculate too small serial correlation coefficients when we only have 50 y values in each data set. This decreases the mean value of v. Based on this I decide to use a preset v value of 1.5 for both OLS and K-T.

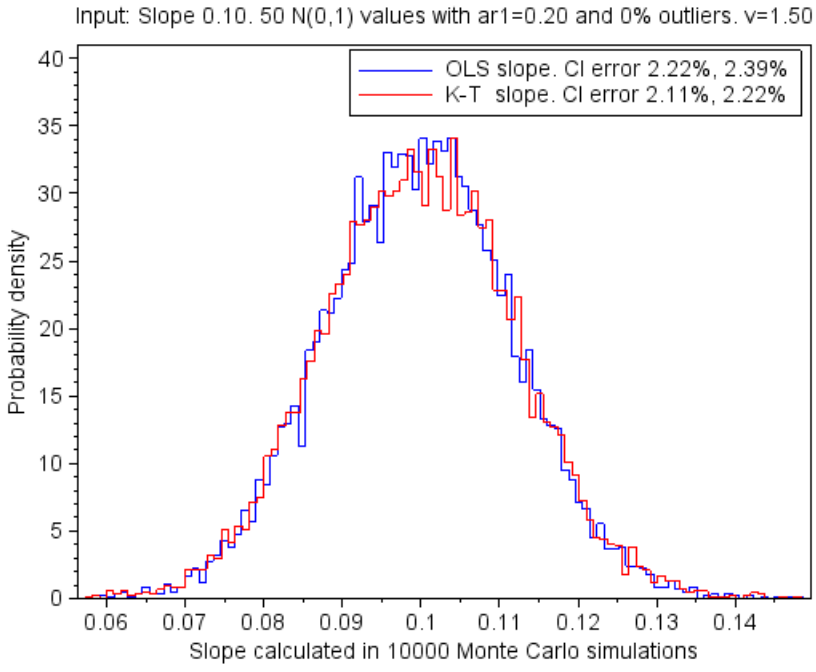

Figure 5.3: Probability density plot of the slopes calculated with Kendall-Theil (K-T) and Ordinary Least Squares (OLS). The noise in the dependent variable is white plus a first order Markov process.

Figure 5.3 shows the probability density plots of the OLS and the K-T slopes. They are both calculated with the compensation factor v preset to 1.5. The v factor has large impact on the uncertainty of the calculated slopes, but it has no impact on the slopes themselves and therefore no impact on the probability density plots in Figure 5.3.

The mean of the calculated slopes is equal to the correct slope 0.100, and the standard deviation of the slopes is equal to 0.012. This is true for both OLS and K-T.

I repeated the Monte Carlo simulations with v equal to 1.0 (no compensation) and with calculation of v individually for each set of 50 y-values. The 95% confidence intervals for the slopes for these three v factors are shown in Table 5.1. They are shown both when calculated with the OLS and with the K-T methodology. I.e. the table shows the results from six Monte Carlo runs each consisting of 10000 simulations. The column *% CI 'misses'* shows the percentage of the confidence intervals that do not cover the correct slope. In approximately half of the misses the lower limit of the confidence interval is larger than the correct slope, and in the other half the upper limit is less than it.

v	Methodology	95% CI Limits and stdev	% CIs 'misses'
1.0	OLS	[0.080, 0.119] [0.012, 0.012]	10.3%
	K-T	[0.080, 0.120] [0.012, 0.012]	9.7%
1.5	OLS	[0.076, 0.124] [0.012, 0.012]	4.6%
	K-T	[0.075, 0.125] [0.012, 0.012]	4.3%
Calculate	OLS	[0.073, 0.127] [0.029, 0.028]	6.3%
	K-T	[0.077, 0.123] [0.013, 0.013]	7.0%

Table 5.1: The confidence intervals when there is serial correlation in the noise. The table shows how well the OLS and the K-T methodologies work when they try to compensate for the serial correlation. The first column shows the v factor used in the compensation, and the second column shows the methodology. The third column shows the mean and the standard deviation of the lower and the upper limits of the 95% Confidence Intervals in the 10000 Monte Carlo simulations. The means are in the upper line, and the standard deviations are in the second line in blue writing. The fourth column shows the percentage of the 95% confidence intervals that does not cover the correct slope. The expected percentage is 5%.

Table 5.1 demonstrates that it is necessary to compensate for serial correlation. When we do not compensate for it (v equal to **1.0**), the calculated confidence intervals are too narrow; approximately 10 % of them do not cover the correct slope.

We get the most realistic confidence intervals when we compensate with v preset to **1.5**. Then approximately 5% of the confidence intervals do not cover the correct slope.

The last row in Table 5.1 (**Calculate**) shows the results when we calculate v individually for each set of 50 y values. Then the confidence intervals are a little too narrow; between 6 and 7 % of them do not cover the correct slope. The v factor varies heavily depending on the y data set, despite that

they have the same serial correlation when they were created. This approach is therefore dubious.

Table 5.1 shows that there is no great difference between the confidence intervals calculated by the OLS and the K-T methodologies. Both methodologies handle serial correlation in the noise well when they apply the preset factor of 1.5. The standard deviations of the limits in the 95% confidence intervals are in blue writing in the table. They are approximately 0.012 in all cases, except when the OLS methodology calculates v individually for each set of 50 y values. Then the standard deviation is more than twice as large. K-T is more robust than OLS in this case.

In the Monte Carlo simulations we know the answer with respect to serial correlation. With real data we do not know. Then the best approach is to evaluate the data, calculate v with many data sets, and thereafter conclude which v to be used as the preset value.

As an example, the monthly global temperature series have serial correlation. A warm month is usually followed by another warm month. For most of the monthly global temperature series we calculate a v factor around 12, meaning that the effective number of measurements approximately equals the number of years that the monthly series cover.

5.4 Both serial correlation and outliers in the noise

Now I add both serial correlation and outliers to the white noise. I do the Monte Carlo simulations with the same three approaches, as explained in the previous chapter, with respect to compensating for the serial correlation. That is without compensating for it (v=1), compensate for it with a preset compensation factor (v=1.5), and calculate the compensating factor individually for each set of 50 y values.

The probability density plots of the slopes calculated by the OLS and the K-T methodologies are shown in Figure 5.4. They are similar to those in Figure 5.2 when only outliers were added to the random noise.

The probability density plot of the K-T slopes is approximately normally distributed, while that of the OLS slopes is not. The mean of the slopes calculated with OLS and K-T are both equal to the correct value 0.100. With K-T the standard deviation of the slopes is 0.013, and with OLS it is 0.049, i.e. almost five times larger. This shows that K-T handles the outliers much better than OLS does, also when serial correlation is present in the noise.

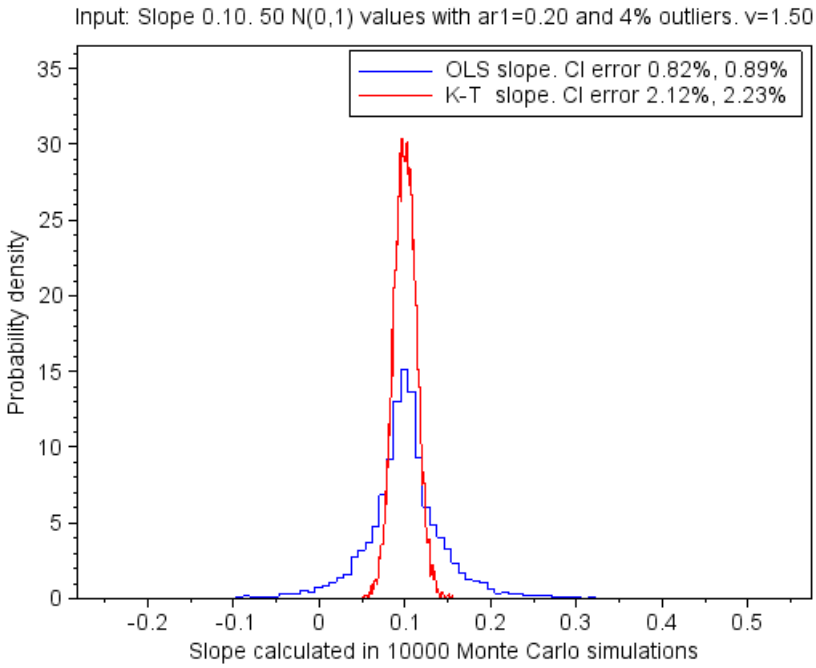

Figure 5.4: Probability density plot of the slopes calculated with Kendall-Theil (K-T) and Ordinary Least Squares (OLS). The noise in the dependent variable is white plus both a first order Markov process and outliers.

Table 5.2 shows how the choice of compensation for the serial correlation impacts the mean values and the standard deviations of the 95% confidence intervals in the simulations.

v	Methodology	95% CI Limits and stdev	% CIs 'misses'
1.0	OLS	[0.027, 0.172] 0.069 0.069	6.8%
	K-T	[0.078, 0.122] 0.014 0.014	9.8%
1.5	OLS	[0.010, 0.190] 0.077 0.077	1.7%
	K-T	[0.073, 0.128] 0.014 0.014	4.3%
Calculate	OLS	[0.017, 0.182] 0.080 0.081	4.4%
	K-T	[0.075, 0.126] 0.015 0.015	6.8%

Table 5.2: The confidence intervals when there is both serial correlation and outliers in the noise. The table contains similar information as Table 5.1, so see the explanation of that table.

The next paragraphs compare Table 5.1 and Table 5.2 to see how outliers, in addition to serial correlation, change the 95% confidence intervals.

When the K-T methodology is used, the outliers cause almost no change in the calculated 95% confidence intervals. But they tend to be too narrow.

When the OLS methodology is used, the outliers cause the calculated 95% confidence intervals to be much wider than they are without the outliers. They tend to be too wide, most so when the v factor is preset to 1.5. Then only 1.7% of the intervals do not contain the correct slope.

The most realistic 95% confidence intervals in Table 5.2 are calculated with the K-T methodology when the v factor is preset to 1.5, and with the OLS methodology when the v factor is calculated individually for each set of 50 y values. In these two runs the K-T interval is [0.073, 0.128] and the OLS interval is [0.017, 0.182]. I.e. the OLS interval is three times wider than the K-T interval. Because both intervals are realistic, this shows that K-T is more robust against outliers than OLS is.

5.5 Random noise, coloured noise and outliers

This chapter explains how the noise vectors are generated.

The Monte Carlo simulations add noise to the data vectors before processing them. The standard normal distribution is the basis of the noise added, and we call it white noise. I use the Scilab rand() function to generate it.

Coloured noise may be added to the white noise. Then parts of the noise in one measurement tend to be present also in the next measurement. I implement the coloured noise as a first order Markov process.

Outliers may be added to the white noise. An outlier is much larger than the usual white noise. It is caused by either a measurement error or large variability in the data being measured.

White noise

We start with a noise vector x of length n. It contains white noise that follows a standard normal distribution N(0,1).

Coloured noise

When coloured noise shall be added to the noise, a first order Markov process is added to the noise vector x. The new x vector shall still have standard deviation 1, and we therefore reduce the original value of each x element before adding the coloured contribution to it. The coloured contribution is the factor α multiplied by the previous element, as shown in (5.1).

$$x_i = \alpha x_{i-1} + \sqrt{1-\alpha^2}\, x_i \qquad \text{for i = 2, 3,...,n} \tag{5.1}$$

The factor α is between 0 and 1.

Outliers

When outliers shall be added to the noise, a percentage of the x values are increased by an outlier scale factor. The indexes of these x values are chosen at random. Even though only the values that were large before being multiplied by the factor become real outliers, I regard all the noise values in x that are multiplied by the factor as outliers.

Reference for the mathematics

Chapter 6.3 in the document Time series analysis published by the Wright State University shows how a first order Markov process is generated. The expected lag 1 serial correlation coefficient of the modified x vector in (5.1) is α, the lag 2 coefficient is α^2, and so forth.

6 Detect serial correlation in data with noise and outliers

This chapter deals with Monte Carlo simulations that calculate the serial correlation coefficients in noisy data. They are calculated with two different approaches. One uses the noise values, and the other uses the ranks of the noise values. Both approaches work well when the noise is white and when there is serial correlation in the noise. The approach that uses the ranks works much better than the other when there are outliers in the noise. The results are presented as probability density plots.

The mathematics to decide the serial correlation coefficients with lag 1, 2, 3 and so forth is shown in equation (3.14) in the chapter 3 Trend when outliers in the data. The formula uses either the noise values or the ranks of the noise values. The formula returns stable results when the noise contains many values. But it returns unstable results when the noise contains few values. Then the calculated coefficients vary, even though the noise has the same characteristics. This chapter shows how well we can calculate the serial correlation coefficients of data containing white noise, coloured noise and outliers. Due to the variability, the results are presented as probability density plots.

All simulations are done with noise data generated as described in chapter 5.5 Random noise, coloured noise and outliers. The noise vector is first filled with white noise drawn from the standard normal distribution N(0,1). If the noise vector shall contain serial correlation, a first order Markov process with α equal to 0.20 is added to it. (A first order Markov process is also called an AR(1) process.) Thereafter, if the noise vector shall contain outliers, randomly selected 4% of the elements in the noise vector is multiplied by 20.

I tested that the formula returns correct coefficients when the data vector contained one million elements with white noise and serial correlation with α equal to 0.20. The expected lag k serial correlation coefficient is α^k. Using the formula I calculated the lag 1 coefficient equal to 0.201, the lag 2 coefficient equal to 0.040 and the lag 3 coefficient equal to 0.009. These results are close to the expected values, and they show that the formula, and my implementation of it, returns correct results.

I thereafter tested noise vectors with 50 elements. This is a typical vector length when analysing climate data. In each test I did ten thousand Monte Carlo simulations, and I present the results graphically as probability density plots. The probability density is along the vertical axis and the calculated serial correlation coefficient is along the horizontal axis. Each test is done with different noise characteristics, but the ten thousand simulations in each test are done with the same noise characteristics.

6.1 Only white noise in the noise vector

Each plot presents the probability densities of the lag 1, 2 and 3 serial correlation coefficients based on 10 000 Monte Carlo simulations. Each simulation calculates the coefficients of 50 data values with white noise.

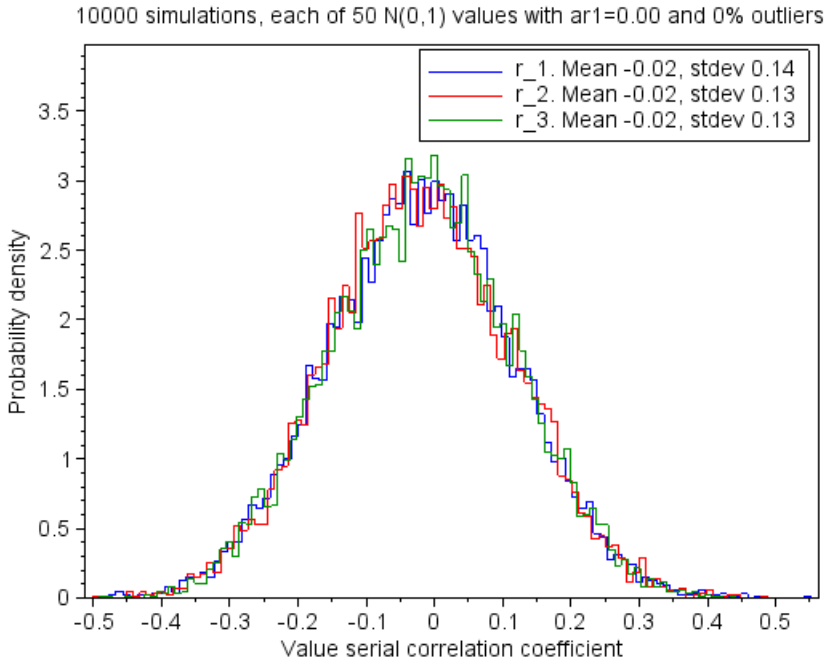

*Figure 6.1: Probability density plots of the lag 1, 2 and 3 serial correlation coefficients of white noise data. The data contains 50 values that are updated before each simulation. The coefficients are calculated with the noise **values**.*

All figures in this chapter are generated by test 2 in rand_outliers_colnoise()

The legends in Figure 6.1 to 6.9 show the means and the standard deviations of the lag 1, lag 2 and lag 3 serial correlation coefficients in the Monte Carlo simulations. The lag 1 serial coefficient is denoted by r_1, the lag 2 coefficient by r_2, and so forth.

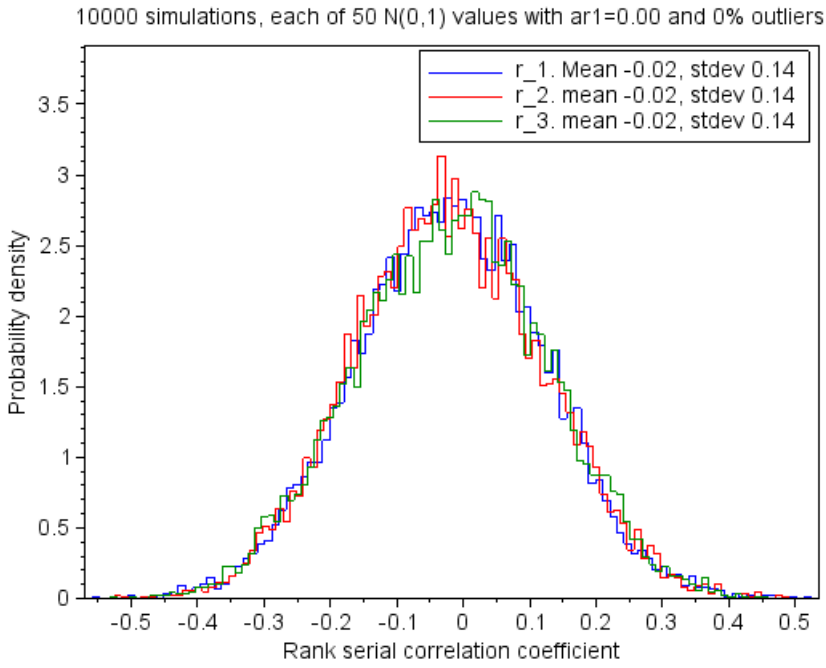

*Figure 6.2: Same as Figure 6.1, but now the serial correlation coefficients are calculated with the **ranks** of the noise values.*

The curves in Figure 6.1 and Figure 6.2 are almost identical. With white noise it does not matter if the coefficients are calculated with the values or with their ranks.

The probability density plots in the figures are surprisingly wide. Even when there is no serial correlation in the noise, there is a considerable probability to calculate serial correlation coefficients larger than 0.3 in absolute value. Therefore, with only one measurement series of 50 values it is not possible to calculate trustworthy serial correlation coefficients.

Serial correlation may be a problem when analysing monthly climate data. Climate is average weather during at least 30 years, and the number of months in climate analysis is therefore usually 360 or more. I therefore repeated the calculations behind Figure 6.1 with 360 noise values in each Monte Carlo simulation.

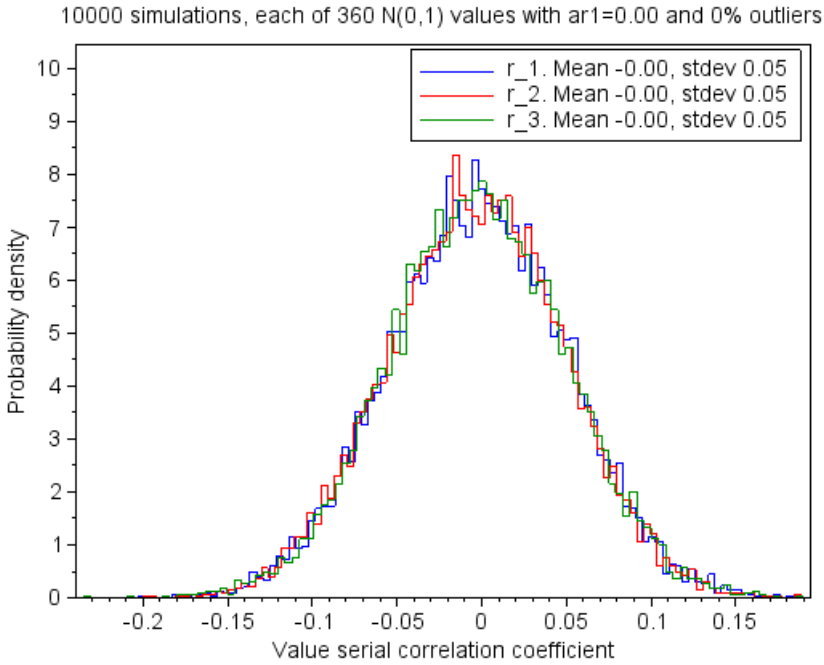

Figure 6.3: Same as Figure 6.1, but now with 360 instead of 50 data values in each simulation.

Figure 6.3 shows that the width of the probability density plots is reduced to a third when the number of noise values is 360 compared to when it is 50. But the standard deviation is still 0.05, and the calculated serial correlation must therefore still be carefully evaluated before being applied further in the analysis.

6.2 Serial correlation in the noise vector

Each plot presents the probability densities of the lag 1, 2 and 3 serial correlation coefficients based on 10 000 Monte Carlo simulations. Each simulation calculates the coefficients of 50 data values with white noise and serial correlation.

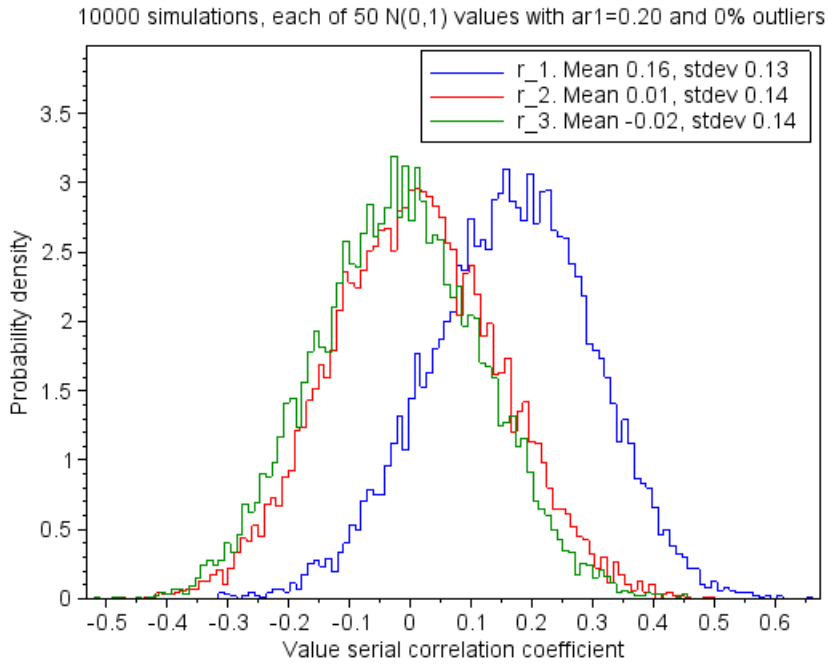

*Figure 6.4: Probability density plots of the lag 1, 2 and 3 serial correlation coefficients of white noise data with serial correlation. The data contains 50 values that are updated before each simulation. The coefficients are calculated with the **values** in the noise vector.*

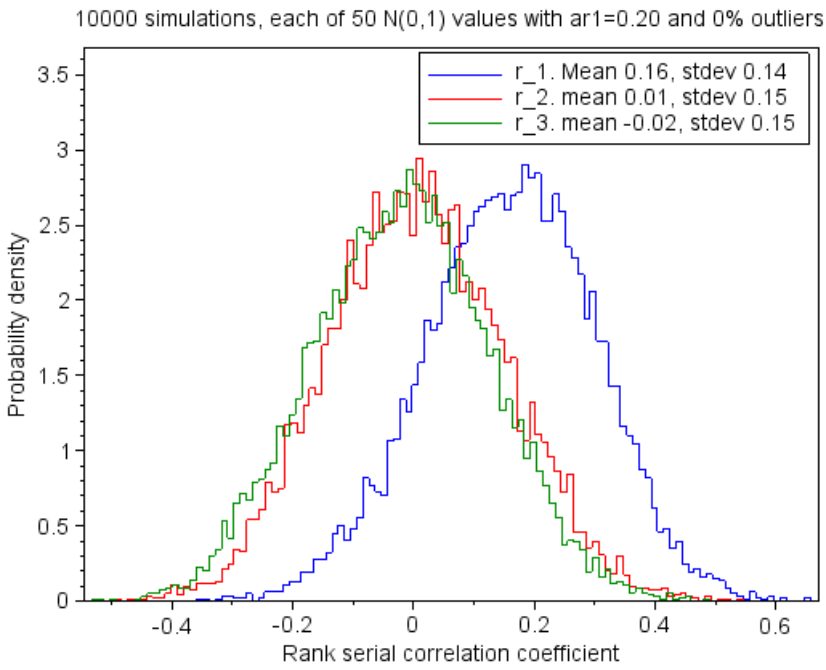

*Figure 6.5: Same as Figure 6.4, but now the serial correlation coefficients are calculated with the **ranks** of the noise values.*

The curves in Figure 6.4 and Figure 6.5 are almost identical. With white noise and serial correlation in the data it does not matter if the coefficients are calculated with the values or with their ranks.

The expected mean value of the lag 1 serial correlation is equal to 0.20 when a first order Markov process with α equal to 0.20 is added to the random values. The mean of the calculated lag 1 serial coefficients is 0.16, which is less than the expected value. That is due to few values in each simulation.

6.3 Outliers in the noise vector

Each plot presents the probability densities of the lag 1, 2 and 3 serial correlation coefficients based on 10 000 Monte Carlo simulations. Each simulation calculates the coefficients of 50 data values with white noise and outliers.

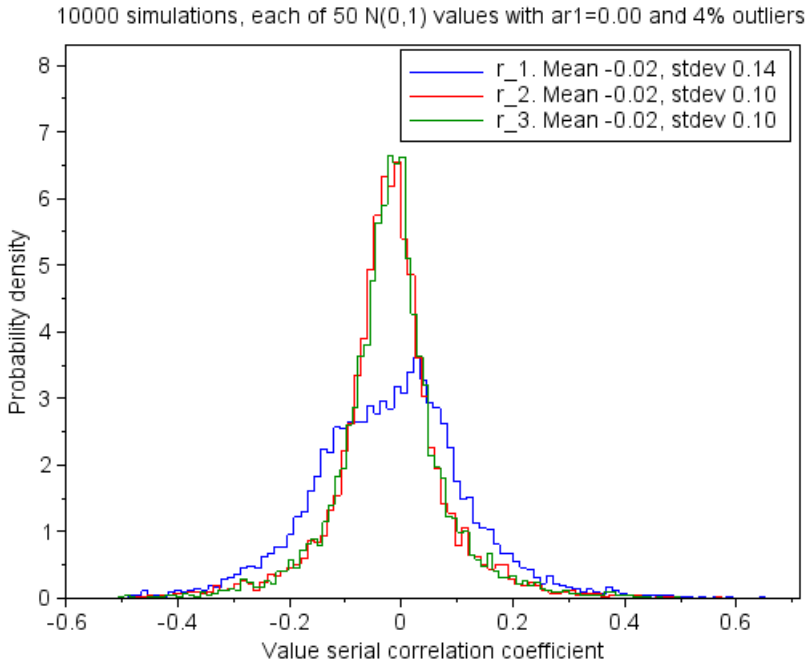

Figure 6.6: Probability density plots of the lag 1, 2 and 3 serial correlation coefficients of white noise data with outliers. The data contains 50 values that are updated before each simulation. The coefficients are calculated with the noise ***values****.*

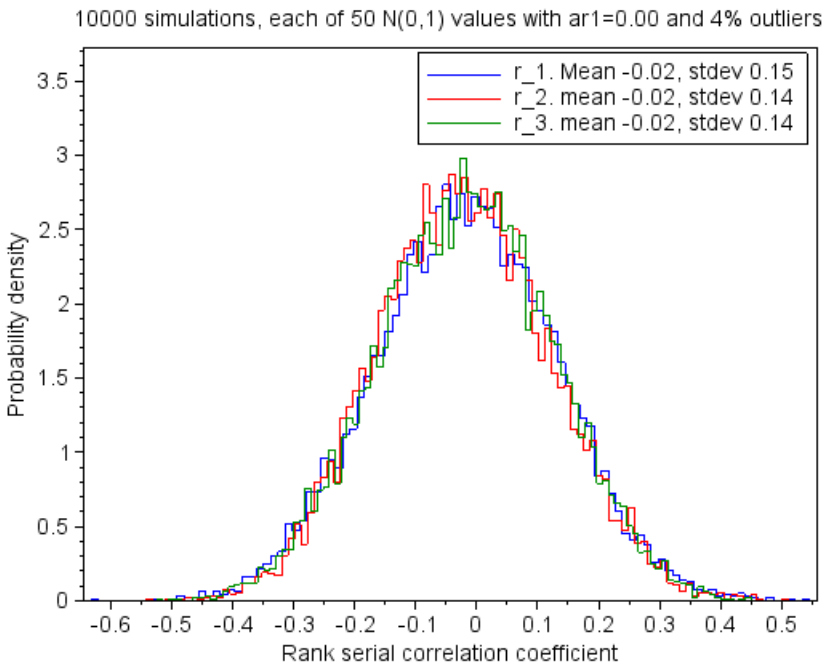

*Figure 6.7: Same as Figure 6.6, but now the serial correlation coefficients are calculated with the **ranks** of the noise values.*

The mean values of the coefficients in both Figure 6.6 and Figure 6.7 are close to zero, as they are expected to be. The calculations of the serial correlation coefficients are not fooled by the outliers to believe that there is serial correlation in the data when it is not. That is true both when the calculations use the data values and when they use the ranks of the data values.

When the coefficients are calculated based on the **values** in the noise vector (Figure 6.6), the lag 1 probability density plot differs more from that of the normal distribution than it does when there were no outliers in the noise vector, and the lag 2 and 3 probability density plots are not so wide.

When the coefficients are calculated based on the **ranks** of the values in the noise vector (Figure 6.7), the probability density plots are almost identical to the probability density plots that were calculated when there were no outliers in the noise.

6.4 Both serial correlation and outliers in the noise vector

Each plot presents the probability densities of the lag 1, 2 and 3 serial correlation coefficients based on 10 000 Monte Carlo simulations. Each simulation calculates the coefficients of 50 data values with white noise, serial correlation and outliers.

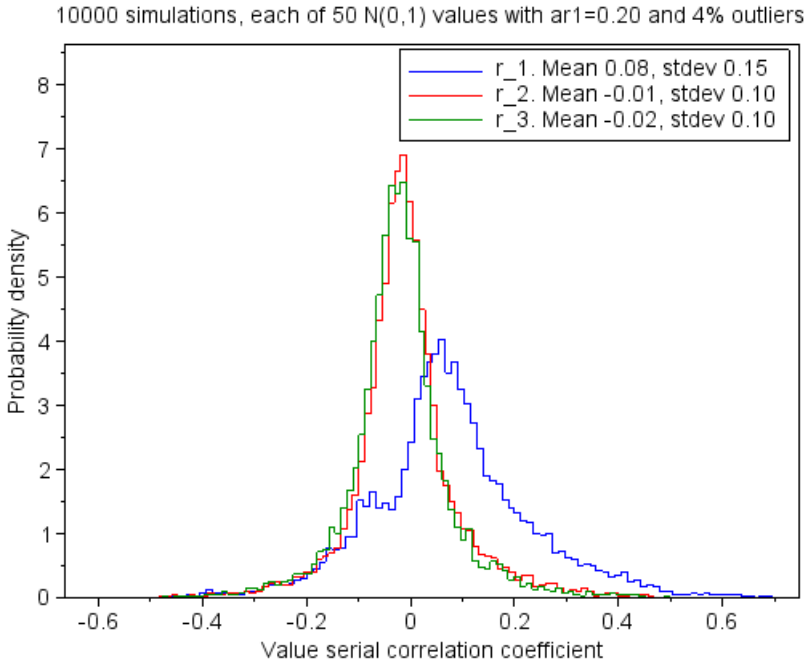

Figure 6.8: Probability density plots of the lag 1, 2 and 3 serial correlation coefficients of white noise data with both serial correlation and outliers. The data contains 50 values that are updated before each simulation. The coefficients are calculated with the noise ***values****.*

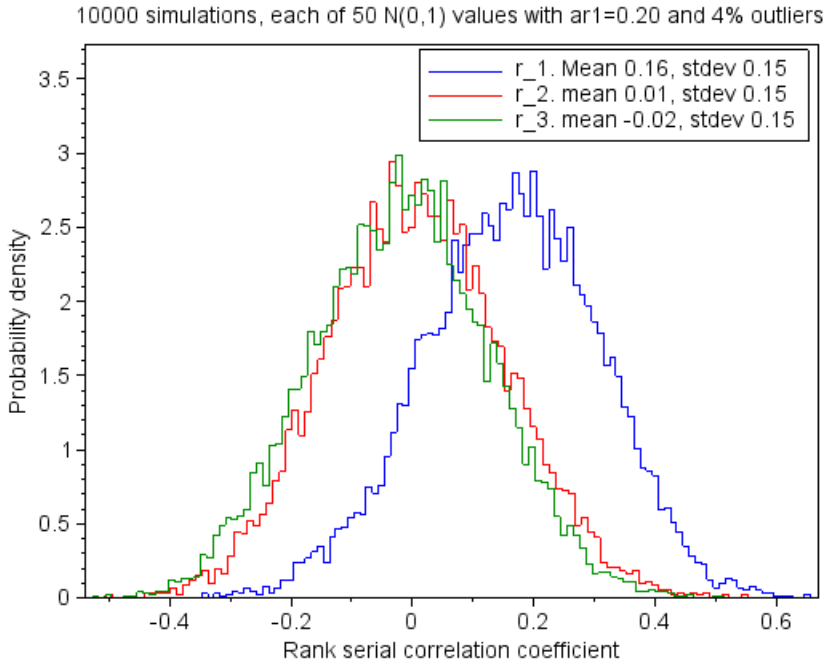

*Figure 6.9: Same as Figure 6.8, but now the serial correlation coefficients are calculated with the **ranks** of the noise values.*

The probability density plots in Figure 6.8 and Figure 6.9 are different. The outliers cause the calculation of the lag 1 coefficient to fail when the calculation is done with the **values**. It fails because the serial correlation in the data is partly hidden by the outliers, and the mean of the calculated lag 1 coefficients is halved compared to what it was without the outliers. But when the calculation is done with the **ranks** of the values, the calculation of the coefficients is OK.