

# Performing a Mann-Kendall Trend Test with serially correlated data in R

This tutorial assumes a grasp of Hypothesis tests

# Mann Kendall Test

## Properties

- Non-parametric – Does not assume a distribution
- Rank based – only considers the sign of the differences between data points
- Does not estimate trend magnitude (purely a test of its existence)

## Best to use when:

- The residuals of a linear trend test are not normally distributed
- The trend is expected to be monotonic (always increasing or always decreasing)
- The form of expected trend (ie linear, exponential, parabolic...) is unknown
- Data can be evenly or unevenly spaced along the independent variable ie.

`x <- c(1, 2, 4, 8, 9, 20, 30, 35, 40...)`

`y <- c(6, 8, 3, 3, 7, 8, 5, 4, 3...)`

is fine

# Mann Kendall Test

## Hypothesis Test

- $H_0$ : Series has no trend and values are iid (independent and identically distributed)
- $H_A$ : There is a monotonic trend in the series

- Test Statistic: Kendall's S

- $$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sign}(y_i - y_j)$$

where  $\text{sign}(w) = \frac{w}{|w|}$

- $$\sigma^2 = \frac{n(n-1)(2n+5)}{18}$$

- $$Z_S = S/\sigma$$

- $$p = 2(1 - \text{cdf}(Z_S))$$

# Serial Correlation

- Serial correlation is a dependence of data on previous data points (dependence on an external variable is allowed!)
- This breaks the iid assumption
- It becomes important only when the noise is large compared to the signal
- Soil moisture per day is likely to have serial correlation whereas number of uses of the word 'biscuit' in literature per year is not
- Serial Correlation lowers the power of a Mann Kendall test

# Serial Correlation

How to decide if Serial Correlation exists:

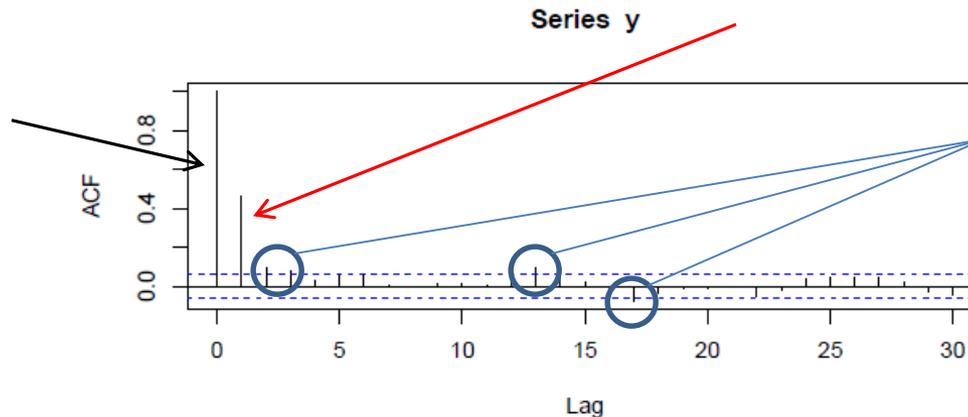
- `acf(y)` gives a plot of the correlation between  $y_i$  and  $y_{i-k}$  for a series of  $k$  values (lags)

- The rejection region for  $\alpha=5\%$  is drawn as a blue line:

$$qnorm((2 + \alpha)/2)/sqrt(length(y)-1)$$

May indicate serial correlation, if a mechanism exists

Correlation of  $y_i$  with itself is always 1



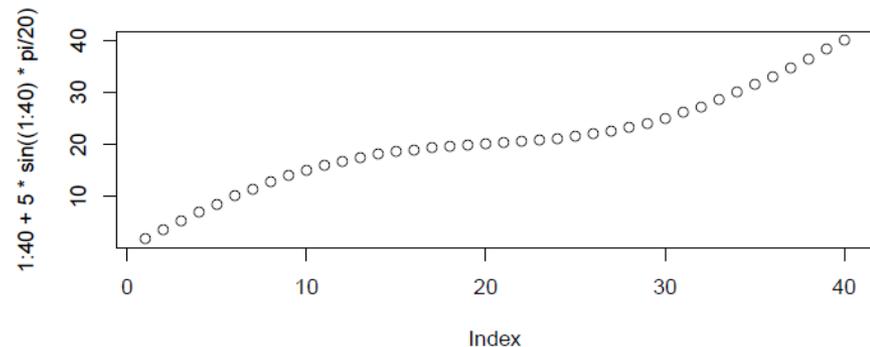
Probably do not indicate serial correlation

- If the correlation at any lag is significant ( $>$  blue line), serial correlation may exist
- Be aware that each lag is independent, so it is likely that some will be significant by accident!
- However, if a lag is significant, and there is mechanism that suggests it should be, there is likely serial correlation present

# Serial Correlation and External Drivers

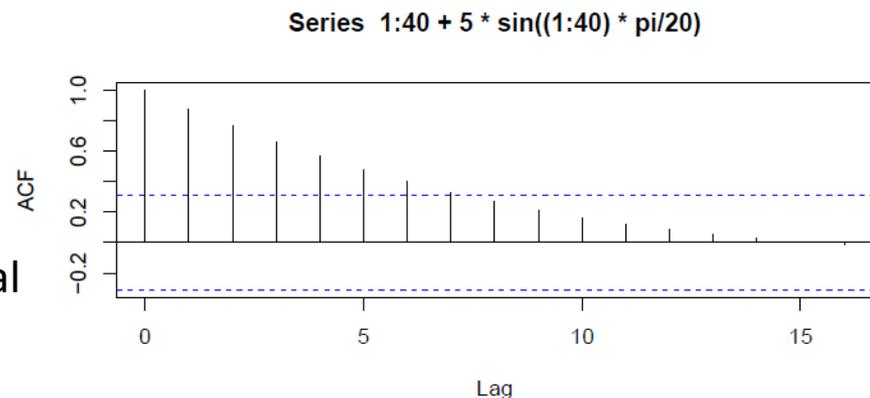
- Serial correlation may be better explained by an external variable which will appear as a pattern in the acf plot. For example, in the case of an embedded sine curve, a serial correlation appears to exist, but is the result of an extra driver

```
plot(1:40 + 5*sin((1:40)*pi/20))
```



```
acf(1:40 + 5*sin((1:40)*pi/20))
```

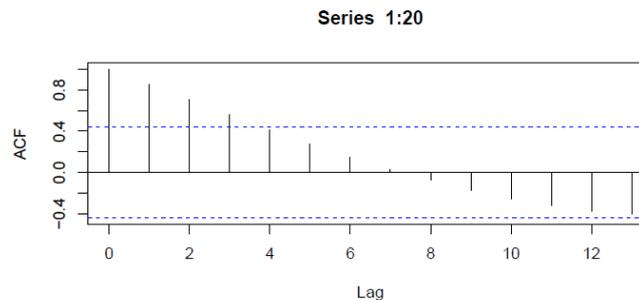
Patterns in an acf plot often indicate an external driver and should often not be corrected for



# Serial Correlation and Noise

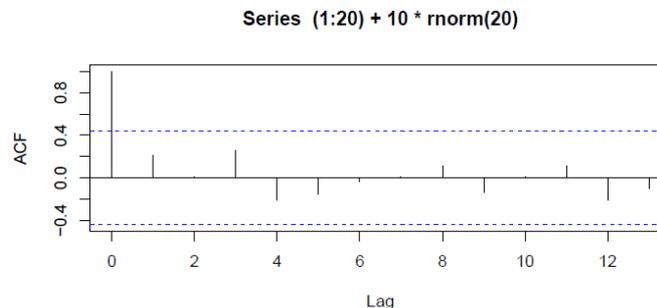
- Some serial correlation will exist just because there is a trend (ie. the acf for a perfectly linear dataset actually has a lot of serial correlation)

`acf(1:20)`



- However, if there is a large amount of noise relative to the signal, this preexisting serial correlation becomes less significant

`acf((1:20)+10*rnorm(20))`



- Again, patterns in an acf plot indicate serial correlation as a result of actual change and should not be corrected for

# Removing serial correlation with prewhitening

- Prewhitening: Compute the autocorrelation function for the significant lag,  $acf_k$ :

```
acfk <- acf(y)$acf[[k+1]]
```

- $(acf_k y_{i-k})$  is the component of the (i)th data point that is contributed from the (i-k)th point
- Subtract that component of the data to remove serial correlation

$$y'_i = y_i - y_{i-k} acf_k:$$

```
y <- y - c(0, y[1:length(y)]) * acfk
```

# Mann Kendall with Prewhitening Example

- Generate synthetic data  $(x,y)$  with a serial correlation

```
#generate the independent variable
```

```
N <- 300
```

```
x <- 1:N
```

```
#set a seed so that results are reproducible and generate normal noise
```

```
set.seed(9)
```

```
noise <- rnorm(N)
```

```
#create data with a simple linear trend
```

```
y <- .003*x+2
```

```
#add noise from a normal distribution with mean = 0 and std = 1
```

```
y <- y + noise
```

```
#add an additional serial correlation dependent on the noise
```

```
y <- y + .5*c(0, noise[1:(length(noise)-1)])
```

# Mann Kendall with Prewhitening

## Example

- Plot autocorrelation function and examine significance

```
#plot the series and the autocorrelation of y
par(mfrow=c(2,1))
plot(x,y, main = 'y vs x')
acf(y)
```

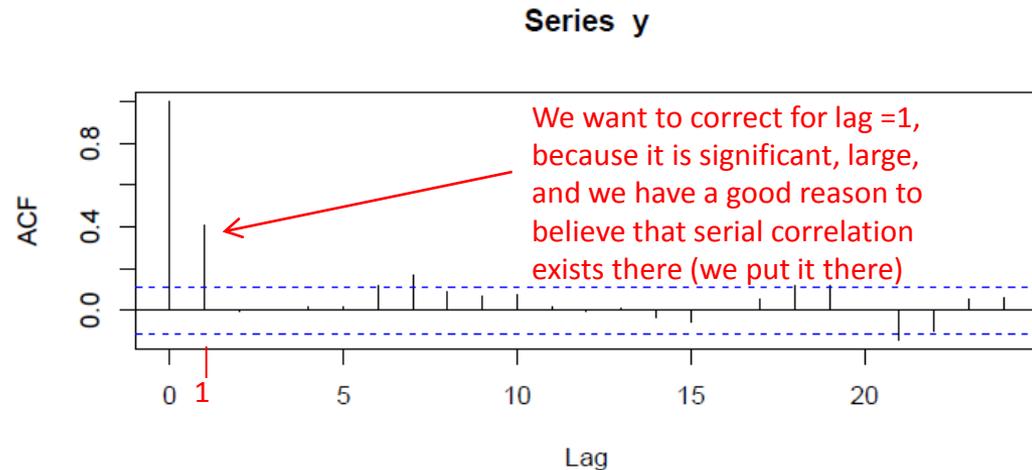
2 Criteria for determining Autocorrelation:

- 1) Significance (> blue line)
- 2) Is there a good reason to think so?

Lag =1 shows (1) a significant (>blue line) and large autocorrelation

We should expect about a 1/20 chance of significance happening by accident which explains the few other significant lags.

However, we know (2) there should be auto-correlation (in this case because we put it there, but in practice it could be a number of reasons, ie. In the case of stream flow series: Residual soil moisture may remain from the previous time).



# Mann Kendall with Prewhitening

## Example

- Remove significant autocorrelation from the dataset

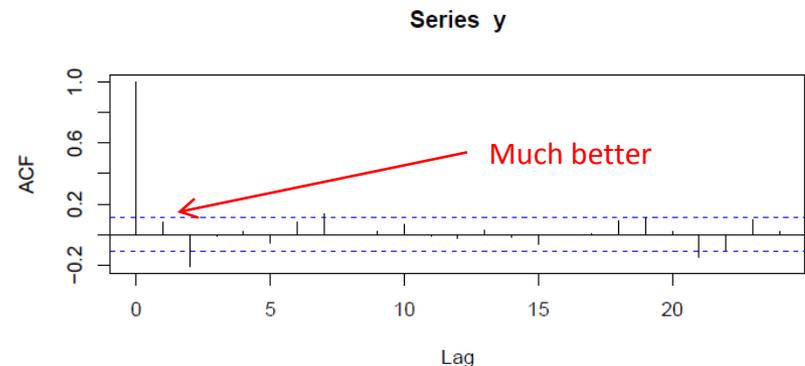
```
#calculate the auto correlation function of y  
y_acf <- acf(y,plot=FALSE)
```

```
#note that the significant lag is 1 (see last slide)  
sig_lag <- 1
```

```
#remove autocorrelation by subtracting the acf value * y[i-lag] from each y  
y <- ( y - y_acf$acf[[sig_lag+1]]*c(rep(0,sig_lag),y[1:(length(y)-sig_lag)])
```

```
#plot the series and the autocorrelation of y to examine changes  
plot(x,y, main = 'y vs x')  
acf(y)
```

The plot shows that the autocorrelation is reduced and we can move on



# Mann Kendall with Prewhitening

## Example

- Perform Mann Kendall trend test with prewhitened dataset

```
# Calculate Kendall's S statistic
S <- 0
for(i in 1:(N-1)){
  S <- S + sum(sign(y[i]-y[(i+1):N]))
}
```

```
# Calculate Variance of Kendall's S
VarS <- (N*(N-1)*(2*N+5))/18
```

```
#Determine Z score of S
Z <- S/sqrt(VarS)
```

```
#calculate p-value
p <- 2*(1-pnorm(abs(Z)))
```

```
print(p)
```

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sign}(x_i - x_j)$$

$$\sigma^2 = \frac{n(n-1)(2n+5)}{18}$$

$$Z_S = S/\sigma$$

$$p = 2(1 - \text{cdf}(|Z_S|))$$

If  $p < \alpha$ , we reject  $H_0$   
(in this case we use  $\alpha=.05$  and  
reject  $H_0$ )

# Mann Kendall with Prewhitening Example

- Mann Kendall Tests are also available in the “Kendall” library

```
# repeat Mann-Kendall test using 'Kendall' package  
library('Kendall')  
summary(MannKendall(y))
```

```
#If 'Kendall' is not installed, you can install it with:  
install.packages('Kendall')
```

# References

This Tutorial is based on:

Douglas, E. M., R. M. Vogel, and C.N. Kroll (2000), Trends in floods and low flows in the United States: impact of spatial correlation, *J. of Hydrology*, 240, 90-105, doi:10.1016/S00221694(00)00336-X

Addendum Data from USGS National Water Information services:

<http://waterdata.usgs.gov/nwis/annual/>

# Addendum: Repeating with real data

- Download datafile: 'RockyRiver.txt' and put in the working directory. This is an annual average streamflow dataset from USGS site:

## **USGS 04201500 Rocky River near Berea OH**

```
# Read in sample data and record its length
```

```
x <- read.table('RockyRiver.txt')[,1]
```

```
y <- read.table('RockyRiver.txt')[,2]
```

```
N <- length(x)
```

# Mann Kendall with Prewhitening

## Example

- **Plot autocorrelation function and examine significance**

```
#plots the series and the autocorrelation of y
```

```
par(mfrow=c(2,1))
```

```
plot(x,y, main = 'y vs x')
```

```
acf(y)
```

```
#calculates the auto correlation function of y
```

```
y_acf <- acf(y,plot=FALSE)
```

```
#note that the significant lag is 1, this is the only one greater than the blue
```

```
#rejection region
```

```
sig_lag <- 1
```

```
#remove autocorrelation by subtracting the acf value * y[i-lag] from each y
```

```
y <- y - y_acf$acf[[sig_lag+1]]*c(rep(0,sig_lag),y[1:(length(y)-sig_lag)])
```

```
#plots the series and the autocorrelation of y to examine changes
```

```
plot(x,y, main = 'y vs x')
```

```
acf(y)
```

# Mann Kendall with Prewhitening

## Example

- Perform Mann Kendall trend test with prewhitened dataset

```
library('Kendall')  
summary(MannKendall(y))
```

- Again,  $p$  is  $<.05$ , and we reject the null hypothesis
  
- For an example where prewhitening is not required, try again with “Cuyahoga.txt”, the mean streamflow data from USGS site:  
**USGS 04202000 Cuyahoga River at Hiram Rapids OH**
  - Notice that when you first plot the acf, no significant autocorrelation exists (no lines extend past the blue line)