

Calculating a Mann Kendall Field Significance with cross-correlated datasets

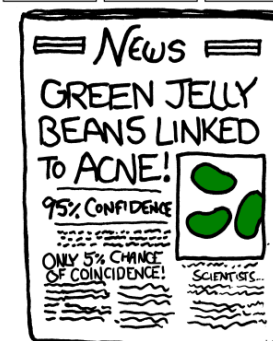
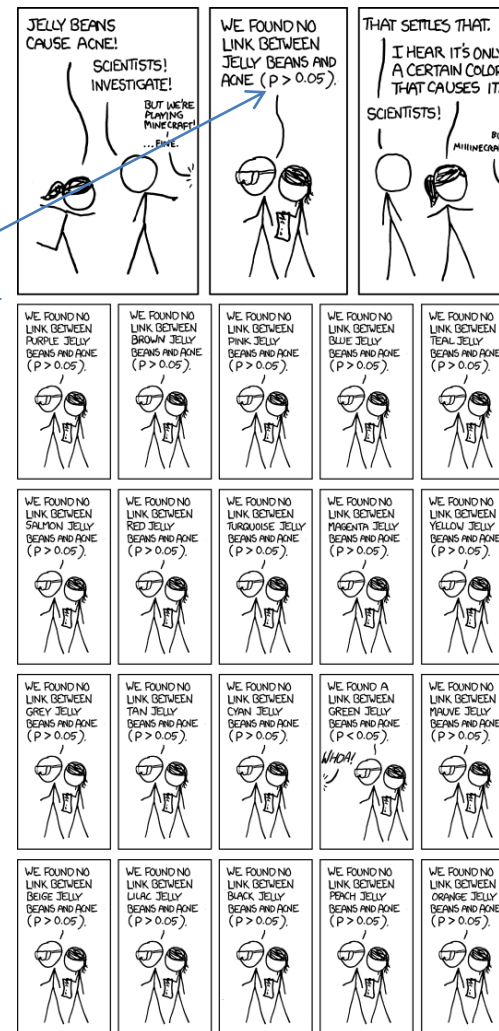
This tutorial assumes a grasp of Mann Kendall trend tests

The importance of Field Significance

Field Significance

- When analyzing multiple datasets the likelihood of falsely rejecting the null hypothesis increases (ie. 20 tests at $\alpha=.05$ will likely reject at least one even if H_0 is true)
- Field significance can estimate the significance of a test for a group of datasets at the same time


Local Significance



Quick method of estimating field significance of a MK trend Test

- If datasets are independent and there is no trend, local significances will be independent and follow a uniform distribution $[0,1]$
- A quick but inaccurate method for a large number of tests is to check if more than 5% of tests have $p < .05$
- More accurate methods must test the entire distribution

Quick method of estimating field significance of a MK trend Test

- Central limit theorem: the sum of a large number of independent random variables is normally distributed
- So Test Statistic (S_M = Mean Kendall's S), should be normally distributed if there is no trend
- For m datasets:
- $S_k = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sign}(y_{k,i} - y_{k,j})$ *where n is the length of the kth dataset and*
- $S_M = \text{mean}(S_k)$ *$\text{sign}(w) = \frac{w}{|w|}$*
- $\sigma^2(S_M) = \frac{n(n-1)(2n+5)}{18m}$
- $Z_s = S_M / \sigma(S_M)$
- $p = 2(1 - \text{cdf}(|Z_s|))$  Field significance tests the group of m datasets at the same time

Problem with the quick method (Cross-correlation)

- In many cases, especially time series, separate datasets may be correlated (in addition to a possible trend) due to a hidden variable, in addition to the independent variable
- Examples include:
 - Annual maximum precipitation at nearby sites depend on local climate variability in addition to climate changes
 - Growth rates of different species in the same location depend on local weather in addition to nutrient mass
- If the datasets are correlated, we break the assumption of independence
- Thus we need to account for correlations

Estimating field significance of a MK trend test with correlated datasets

- $S_k = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sign}(y_{k,i} - y_{k,j})$ for dataset k, length n
- $S_M = \text{mean}(S_k)$
- $\sigma^2(S_k) = \frac{n(n-1)(2n+5)}{18}$
- $\sigma^2(S_M) = \frac{\sigma^2(S_k)}{M} (1 + (M-1)\bar{\rho}_{x,x})$
- $\bar{\rho}_{x,x} = \frac{2}{M(M+1)} \sum_{k=1}^{M-1} \sum_{l=1}^{M-k} \rho_{k,l}$ where $\rho_{k,l} = \frac{E[(y_k - \mu_k)(y_l - \mu_l)]}{\sigma_k \sigma_l}$
(cross-correlation)
- $Z_s = S_M / \sigma(S_M)$
- $p = 2(1 - \text{cdf}(|Z_s|))$

Mann Kendall Field Significance Example

- Generate multiple datasets of synthetic data (x,ymat) with a covariant error

```
#generates the independent variable and records its length
```

```
x <- 1:100
```

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

```
#creates the dependent variable with a weak trend
```

```
ymat <- rbind(.01*x+2,.01*x+2,.01*x+2)
```

```
M <- 3
```

```
#sets a seed so that results are reproducible and generates normal noise
```

```
set.seed(15)
```

```
shared_noise <- rnorm(N)*sqrt(.5)
```

```
#adds randomly distributed noise and shared noise
```

```
for(i in 1:3){
```

```
  ymat[i,] <- ymat[i,] + rnorm(N)*sqrt(.5) + shared_noise
```

```
}
```

Mann Kendall Field Significance Example

- Examine the data

```
plot(x, ymat[1,])  
points(x, ymat[2,],col='red')  
points(x, ymat[3,],col='blue')
```

- Things to look for:
 - Is the overall trend monotonic (either always increasing or always decreasing)? If not, the Mann-Kendall test is not valid
 - Are there other patterns besides the trend (ie periodic behavior or serial correlation)? See the [Mann-Kendall](#) tutorial to address serial correlation and (Hirsch 1982) to address periodic trends
 - Cross-Correlation may or may not be visible by eye

Mann Kendall Field Significance Example

- Calculate Kendall's S values (S_k and S_M) at each site

```
KendallS <- function(y){  
  #Calculate Kendall's S statistic, S, from a series, y  
  #Variables:  
  # y  input data series  
  # S  Kendall's S statistic  
  # N  length of y  
  S <- 0  
  N <- length(y)  
  for(i in 1:(N-1)){  
    S <- S + sum(sign(y[i]-y[(i+1):N]))  
  }  
  return(S)  
}
```

$$S_k = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sign}(y_{k,i} - y_{k,j})$$

Mann Kendall Field Significance Example

```
#Svec is a list of Kendall S values for each dataset
Svec <- rep(0,M)
for(i in 1:M){
  Svec[[i]] <- KendallS(yamat[i,])
}
```

```
#store the average Kendall S value in Sm
Sm <- mean(Svec)
```

$$S_M = \text{mean}(S_k)$$

Mann Kendall Field Significance Example

- Perform Mann Kendall tests for each of the sites

```
#records the variance for a single Kendall's S value
```

```
VarS <- (N*(N-1)*(2*N+5))/18
```

```
#calculate Z-scores and p-values for each Kendall's S value
```

```
Zvec <- rep(0,M)
```

```
pvec <- rep(0,M)
```

```
for(i in 1:M){
```

```
  Zvec[[i]] <- Svec[[1]]/sqrt(VarS)
```

```
  pvec[[i]] <- 2*(1-pnorm(abs(Zvec[[i]])))
```

```
}
```

Mann Kendall Field Significance Example

- Compute Mann Kendall field significance assuming no covariance

#store the variance assuming no covariance between datasets

```
VarSm_nocov <- N*(N-1)*(2*N+5)/18/M
```

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

#store the Z-score for Sm

```
Zfield_nocov <- Sm/(sqrt(VarSm_nocov))
```

$$Z_s = S_M / \sigma(S_M)$$

#calculate p value

```
pfield_nocov <- 2*(1-pnorm(abs(Zfield_nocov)))
```

$$p = 2(1 - cdf(|Z_s|))$$

```
print(pfield_nocov)
```

Mann Kendall Field Significance Example

- Compute Mann Kendall field significance assuming covariance

```
#define crosscorr fuction
```

```
crosscorr <- function(y1,y2){
```

```
  #Variables:
```

```
  # y1      first input dataset
```

```
  # y2      second input dataset
```

```
  # mu1     mean of first dataset
```

```
  # mu2     mean of second dataset
```

```
  # sigma1  standard deviation of first dataset
```

```
  # sigma2  standard deviation of second dataset
```

```
  mu1 <- mean(y1)
```

```
  mu2 <- mean(y2)
```

```
  sigma1 <- sqrt(var(y1))
```

```
  sigma2 <- sqrt(var(y2))
```

```
  return(mean((y1 - mu1) * (y2 - mu2)) / sigma1 / sigma2)
```

```
}
```

$$\rho_{k,l} = \frac{E[(y_k - \mu_k)(y_l - \mu_l)]}{\sigma_k \sigma_l}$$

Mann Kendall Field Significance Example

```
#compute the average cross-correlation coefficient
rhosum <- 0
for(k in 1:(M-1)){
  for(l in 1:(M-k)){
    rhosum <- rhosum + crosscorr(yamat[k,],ymat[l,])
  }
}
rhobar <- 2*rhosum/M/(M-1)
```

$$\bar{\rho}_{x,x} = \frac{2}{M(M-1)} \underbrace{\sum_{k=1}^{M-1} \sum_{l=1}^{M-k} \rho_{k,l}}_{\text{rhosum}}$$

rhobar

```
#compute variance, Z score, and field significance
VarSm <- (VarS/M)*(1+(M-1)*rhobar)
```

$$\sigma^2(S_M) = \frac{\sigma^2(S_k)}{M} (1 + (M-1)\bar{\rho}_{x,x})$$

```
Zfield <- Sm/sqrt(VarSm)
```

$$Z_s = S_M / \sigma(S_M)$$

```
pfield <- 2*(1-pnorm(abs(Zfield)))
print(pfield)
```

$$p = 2(1 - cdf(|Z_s|))$$

Mann Kendall Field Significance

Example Results

- Notice that $p_{\text{field}} > p_{\text{field_nocov}}$
- This is because the Variance of S_m is necessarily larger when considering covariance
- Therefore, one should only correct for covariance when it is likely to exist
- For example, repeat tests of plant growth conducted over different timeframes may not require correction whereas tests of flood intensity over a region during the same timeframe will require correction

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/il/nwis/dv/?site_no=05592500&agency_cd=USGS&referred_module=sw

Comic on slide 2 from:

Monroe, R. "Significant" <http://xkcd.com/882/>

Seasonal Kendall Test

Hirsch, R.M., J.R. Slack and R.A. Smith (1982), Techniques of Trend Analysis for Monthly Water Quality Data, Water Resources Research, 18, 107-121

Addendum: Repeating with real data

- Download accompanying datafiles: “RockyRiver.txt”, “Cuyahoga.txt”, and “Vermillion.txt” to your working directory
These are annual average streamflow datasets from USGS sites
 - 1) **USGS 04201500 Rocky River near Berea OH**
 - 2) **USGS 04202000 Cuyahoga River at Hiram Rapids OH**
 - 3) **USGS 04199500 Vermilion River near Vermilion OH**

#read Flow data from 3 rivers

```
x1 <- read.table('RockyRiver.txt',skip=36)[,5]
```

```
y1 <- read.table('RockyRiver.txt',skip=36)[,6]
```

```
x2 <- read.table('Cuyahoga.txt',skip=36)[,5]
```

```
y2 <- read.table('Cuyahoga.txt',skip=36)[,6]
```

```
x3 <- read.table('Vermillion.txt',skip=36)[,5]
```

```
y3 <- read.table('Vermillion.txt',skip=36)[,6]
```

#Record and count only years where all sites have data

```
x <- x1[is.element(x1,x2) & is.element(x1,x3)]
```

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

#create a matrix of flows for only shared years

```
ymat = matrix(NA, 3, N)
```

```
ymat[1,] <- y1[is.element(x1,x) & is.element(x1,x)]
```

```
ymat[2,] <- y2[is.element(x2,x) & is.element(x2,x)]
```

```
ymat[3,] <- y3[is.element(x3,x) & is.element(x3,x)]
```

Addendum: Repeating with real data

- Examine the data

```
plot(x, ymat[1,])  
points(x, ymat[2,],col='red')  
points(x, ymat[3,],col='blue')
```

- Things to look for:
 - Is the overall trend monotonic (either always increasing or always decreasing)? If not, the Mann-Kendall test is not valid
 - Are there other patterns besides the trend (ie periodic behavior or serial correlation)? See the [Mann-Kendall](#) tutorial to address serial correlation and (Hirsch 1982) to address periodic trends
 - Cross-Correlation may or may not be visible by eye

Addendum: Repeating with real data

- Calculate Kendall's S values (S_k and S_M) at each site

```
KendallS <- function(y){  
  #Calculate Kendall's S statistic, S, from a series, y  
  #Variables:  
  # y  input data series  
  # S  Kendall's S statistic  
  # N  length of y  
  S <- 0  
  N <- length(y)  
  for(i in 1:(N-1)){  
    S <- S + sum(sign(y[i]-y[(i+1):N]))  
  }  
  return(S)  
}
```

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

Addendum: Repeating with real data

```
#Svec is a list of Kendall S values for each dataset
Svec <- rep(0,M)
for(i in 1:M){
  Svec[[i]] <- KendallS(yamat[i,])
}
```

```
#store the average Kendall S value in Sm
Sm <- mean(Svec)
```

$$S_M = \text{mean}(S_k)$$

Addendum: Repeating with real data

- Perform Mann Kendall tests for each of the sites

```
#records the variance for a single Kendall's S value
```

```
VarS <- (N*(N-1)*(2*N+5))/18
```

```
#calculate Z-scores and p-values for each Kendall's S value
```

```
Zvec <- rep(0,M)
```

```
pvec <- rep(0,M)
```

```
for(i in 1:M){
```

```
  Zvec[[i]] <- Svec[[1]]/sqrt(VarS)
```

```
  pvec[[i]] <- 2*(1-pnorm(abs(Zvec[[i]])))
```

```
}
```

Addendum: Repeating with real data

- Compute Mann Kendall field significance assuming no covariance

#store the variance assuming no covariance between datasets

```
VarSm_nocov <- N*(N-1)*(2*N+5)/18/M
```

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

#store the Z-score for Sm

```
Zfield_nocov <- Sm/(sqrt(VarSm_nocov*M))
```

$$Z_S = S_M / \sigma(S_M)$$

#calculate p value

```
pfield_nocov <- 2*(1-pnorm(abs(Z_nocov)))}
```

$$p = 2(1 - cdf(|Z_S|))$$

```
print(pfield_nocov)
```

Addendum: Repeating with real data

- Compute Mann Kendall field significance assuming covariance

```
#define crosscorr fuction
```

```
crosscorr <- function(y1,y2){
```

```
  #Variables:
```

```
  # y1      first input dataset
```

```
  # y2      second input dataset
```

```
  # mu1     mean of first dataset
```

```
  # mu2     mean of second dataset
```

```
  # sigma1  standard deviation of first dataset
```

```
  # sigma2  standard deviation of second dataset
```

```
  mu1 <- mean(y1)
```

```
  mu2 <- mean(y2)
```

```
  sigma1 <- sqrt(var(y1))
```

```
  sigma2 <- sqrt(var(y2))
```

```
  return(mean((y1 - mu1) * (y1 - mu1)) / sigma1 / sigma2)
```

```
}
```

$$\rho_{k,l} = \frac{E[(y_k - \mu_k)(y_l - \mu_l)]}{\sigma_k \sigma_l}$$

Addendum: Repeating with real data

- Compute Mann Kendall field significance assuming covariance

#compute the average cross-correlation coefficient

```
rhosum <- 0
for(k in 1:(M-1)){
  for(l in 1:(M-k)){
    rhosum <- rhosum + crosscorr(yamat[k,],yamat[l,])
  }
}
rhobar <- 2*rhosum/M/(M-1)
```

#compute variance, Z score, and field significance

```
VarSm <- (VarS/M)*(1+(M-1)*rhobar)
```

```
Zfield <- Sm/sqrt(VarSm)
```

```
pfield <- 2*(1-pnorm(abs(Zfield)))
print(pfield)
```

$$\bar{\rho}_{x,x} = \frac{2}{M(M-1)} \underbrace{\sum_{k=1}^{M-1} \sum_{l=1}^{M-k} \rho_{k,l}}_{\text{rhosum}}$$

rhobar

$$\sigma^2(S_M) = \frac{\sigma^2(S_k)}{M^2} (1 + (M-1)\bar{\rho}_{x,x})$$

$$Z_s = S_M / \sigma(S_M)$$

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

Mann Kendall Field Significance

Example Results

- Notice that $p_{\text{field}} > p_{\text{field_nocov}}$
- This is because the Variance of S_m is necessarily larger when considering covariance
- Therefore, one should only correct for covariance when it is likely to exist
- In this case all rivers are near each other and are thus likely to be covariant
- However, if we picked rivers from different parts of the world, correction might not be necessary