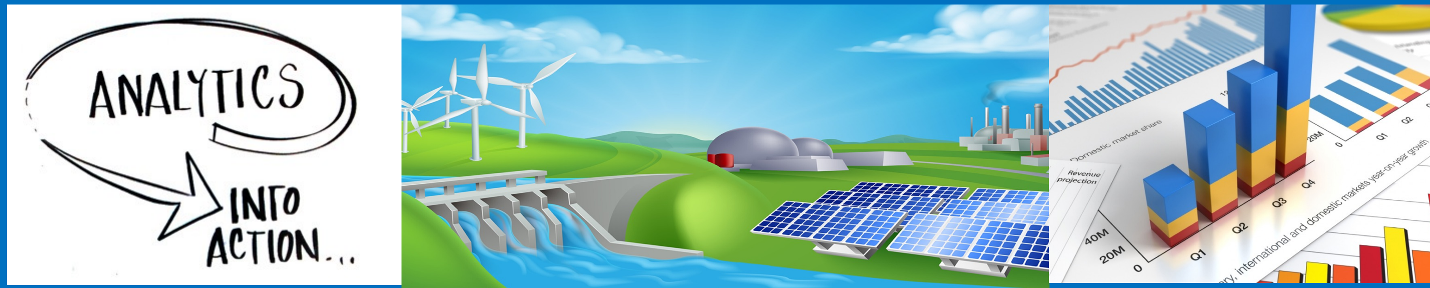


TIME SERIES ANALYSIS FOR ENERGY DATA

M4 - Trend and Seasonality | Stationarity Tests

Prof. Luana Medeiros Marangon Lima, Ph.D.

Nicholas School of the Environment - Duke University

Learning Goals

- Trend Component
  - Linear and Non-linear
  - How to Estimate Linear Trend
  - How to Model/Remove Linear Trend from a Series
- Seasonal Trend
  - How to Estimate Seasonal Trend
  - How to Model/Remove Seasonal Trend from a Series
- Stationarity Tests

Time Series Components

- A time series may have the following components:

Trend Component

Seasonal Component

Cyclical Component

Random Component

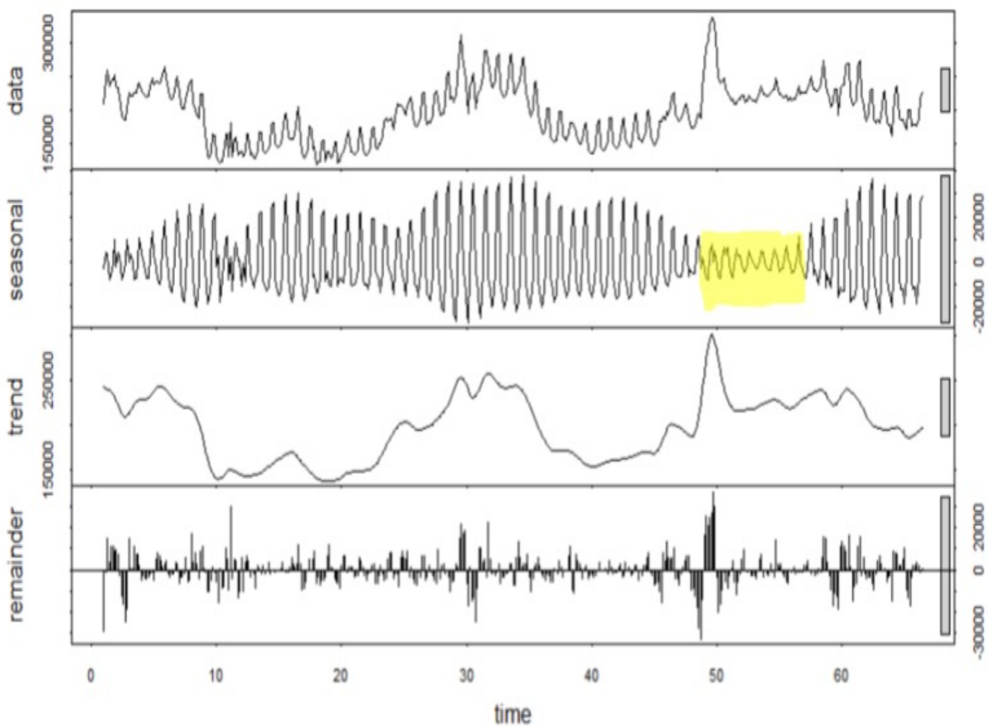
Decomposing the Time Series means separating trend/cycle, seasonal and random components.

TSA will find and exploit predictable patterns/components.

Causes of Variation in TS data

- Calendar: seasons, holidays, weekends
- Example

Video game usage over time – daily basis



Interest

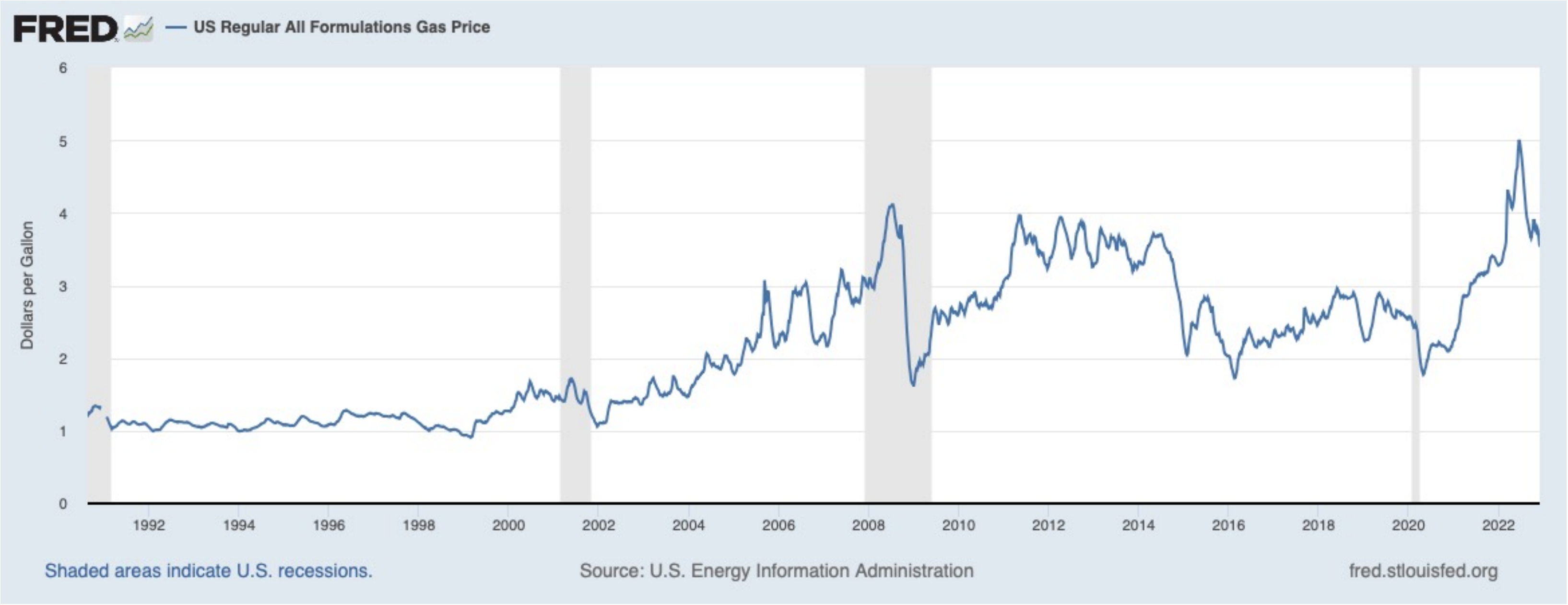
- Trend of usage over time
- Popularity of games
- Weekly cycle of usage

Knowledge

- Usage up on the weekend and down during the week
- Increased usage during holidays
- Summer time: weekdays/weekends blend together

Causes of Variation in TS data

- Natural calamities: earthquake, epidemic, flood, drought
- Political movements or changes, policies, war
- Example



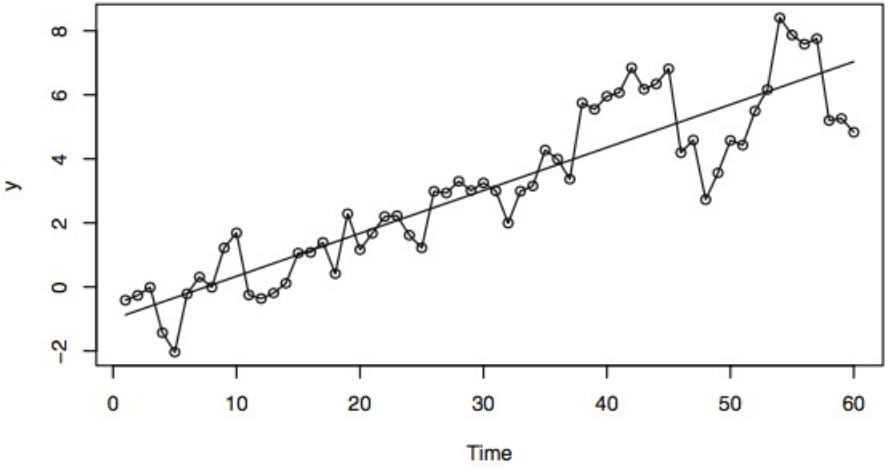
Source: https://fred.stlouisfed.org/series/GASREGW#

Trend Component

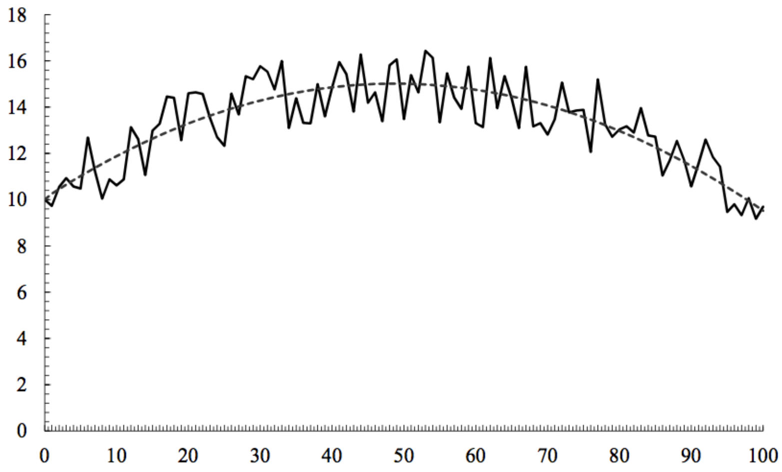
Trend Component

- Long-term tendency
  - Increase (upward movement) or
  - Decrease (downward movement)
- Trend can be linear or non-linear

Ex: Upward Linear Trend



Ex: Quadratic Trend



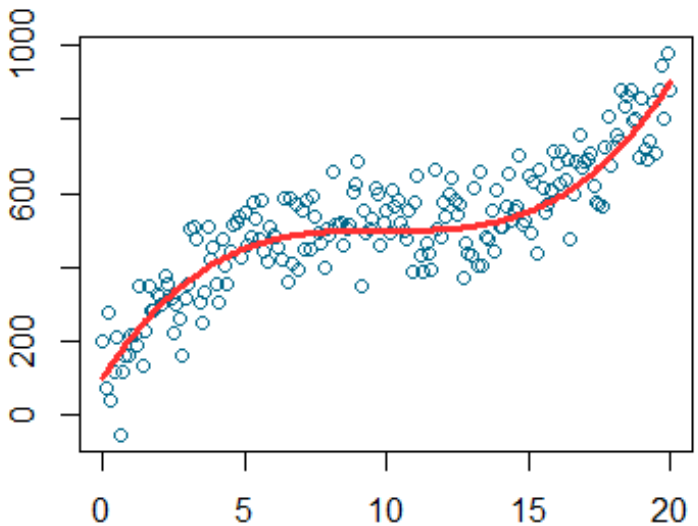
Non-linear Trend

Polynomial trend

- Example: quadratic trend

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 T_i^2 + \varepsilon_i$$

- Or any other order

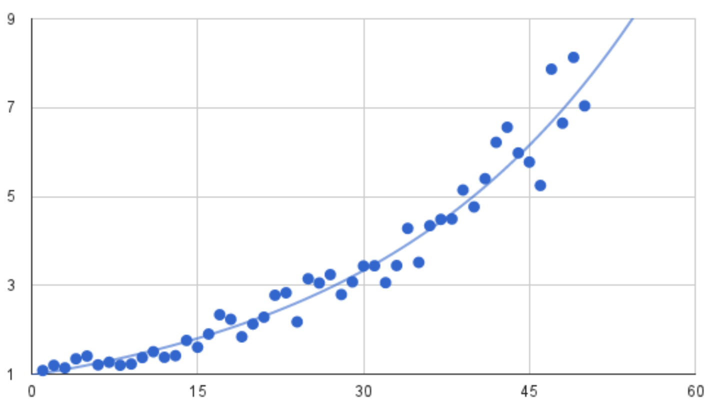


Exponential trend

$$Y_i = (e^{\beta_0 + \beta_1 T_i})\varepsilon_i$$

- Can be transformed into linear trend

$$\ln Y_i = \beta_0 + \beta_1 T_i + \ln \varepsilon_i$$



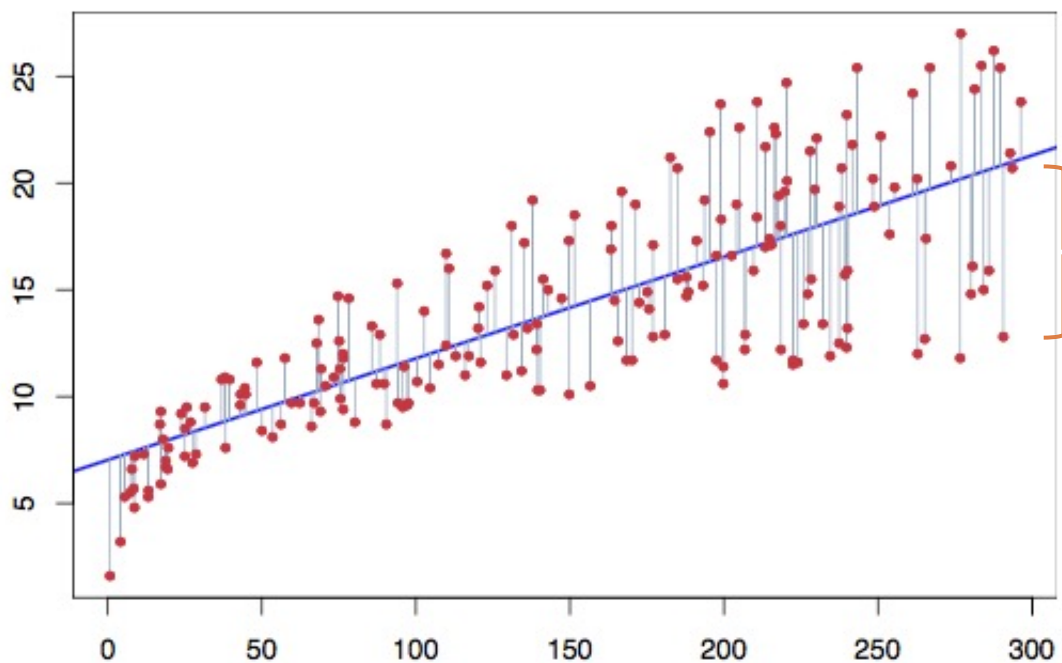
Most of the time we assume a linear trend to simplify the analysis

Linear Trend Component

- For a linear trend we can write

$$Y_i = \beta_0 + \beta_1 t_i + \varepsilon_i$$

- **Slope** (β_1) and the **intercept** (β_0) are the unknown parameters, and ε_i is the **error term**



$\hat{Y}_i = \beta_0 + \beta_1 t_i$

The error term or residual is the distance from point Y_i to the estimate $\hat{Y}_i$

$$\varepsilon_i = Y_i - \hat{Y}_i$$

Linear Trend Estimation

- How do we estimate β_0 and β_1?
- One approach: **Least Squares Method**
  - We want to minimize

$$Q(\beta_0, \beta_1) = \sum_{t=1}^{T} [Y_t - (\beta_0 + \beta_1 t)]^2$$

  - How de we minimize this function?
    - By taking the partial derivatives of $Q(\beta_0, \beta_1)$ with respect to the coefficients β_0 e β_1
    - QR decomposition

We will use R to solve it!

Estimating Linear Trend in R

- The function for simple linear regression in R is the *lm()* from package "stats", where *lm* stands for "linear model"
- The arguments you will need to provide are

$$lm(Y \sim t)$$

Vector with observed series

Vector from 1 to number of observations of Y

Note: vectors *Y* and *t* should be in data frame format

Linear Trend Estimation and Removal

1. Model the trend: find β_0 and β_1

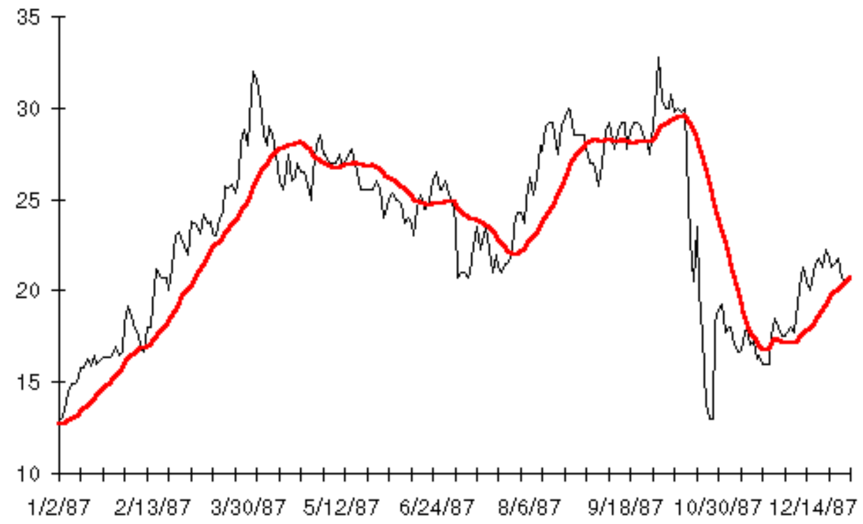
2. For each observation t remove trend

$$Y_{detrend_t} = Y_t - (\beta_0 + \beta_1 t)$$

Moving Average for Non-Linaer Trend Estimation

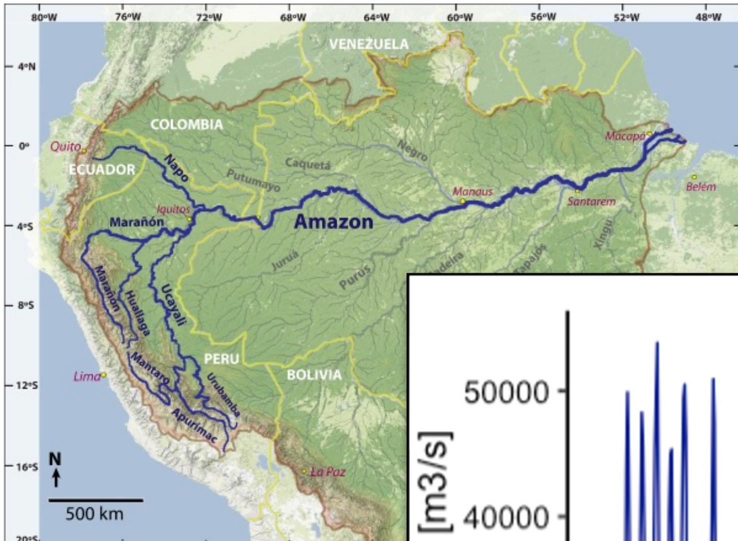
- Smooth out the trend with something like a rolling average
  - A moving average trendline smooth out fluctuations in data to show a pattern or trend more clearly
  - Which order to use for the moving average?

- Looking at the rolling average makes it easier to tell how the trend is moving underneath the noise

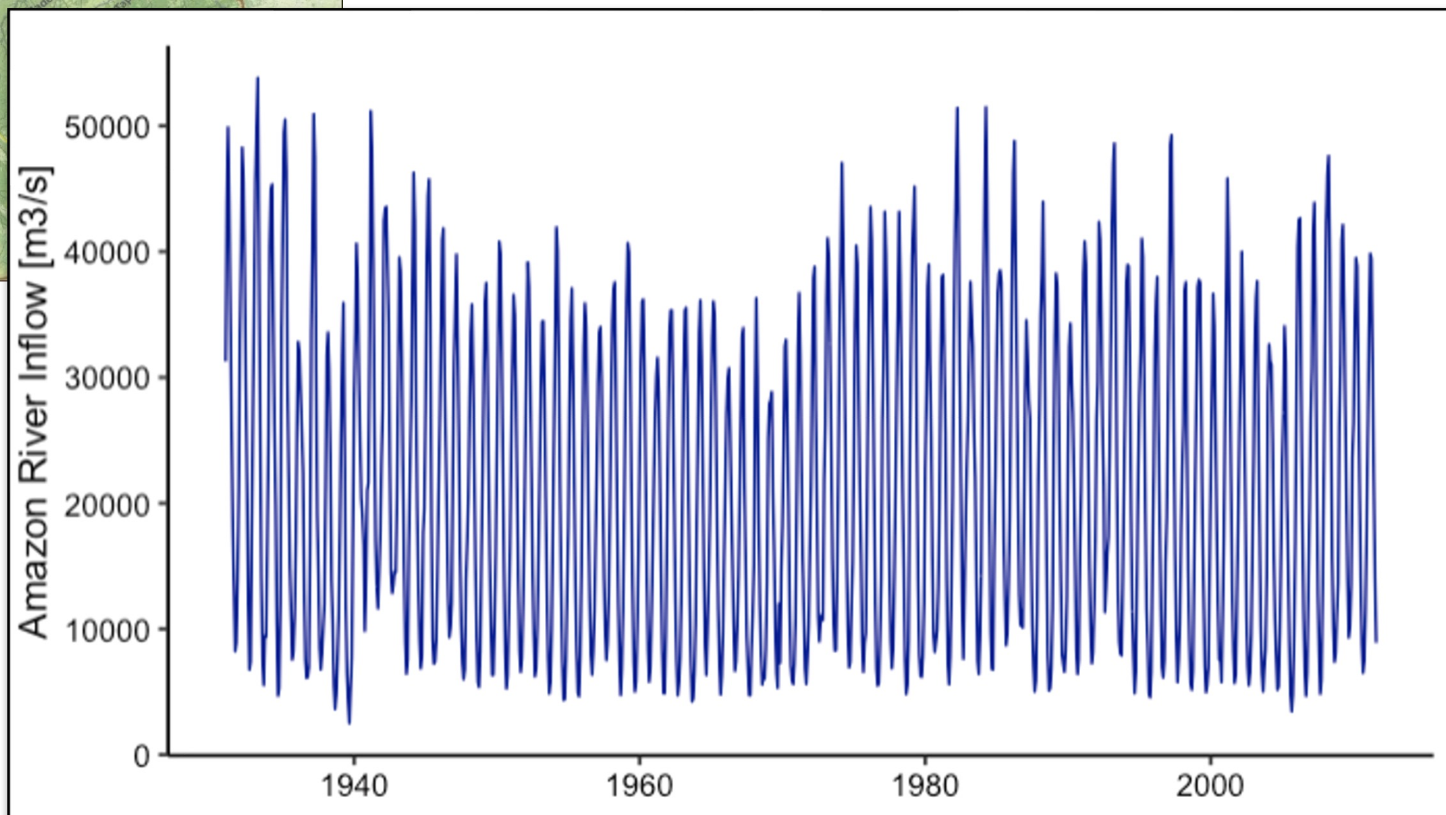


Example: Inflow Data

Amazon River Inflow in m^3/s

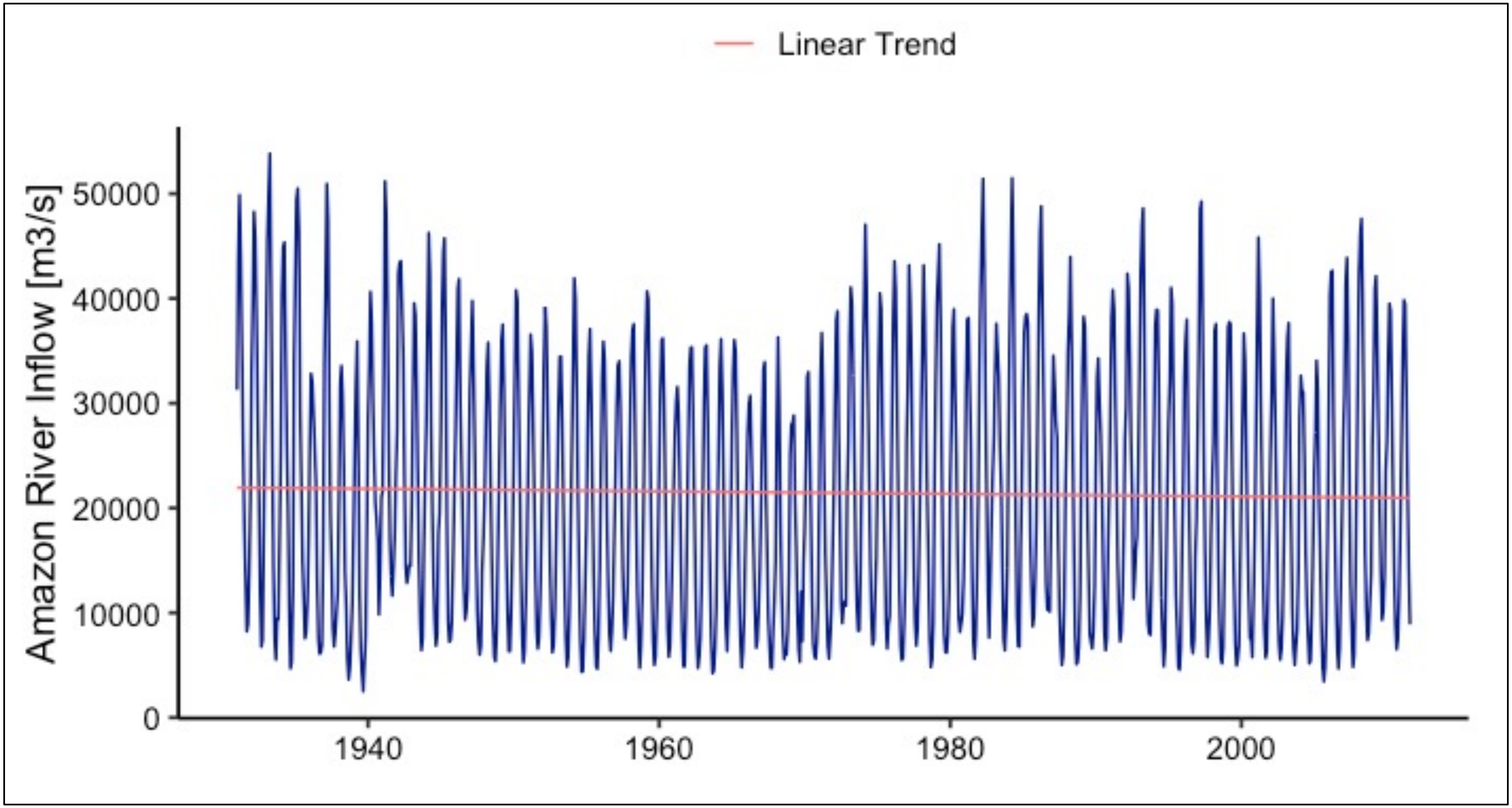


Which components/patterns can you see in this series?



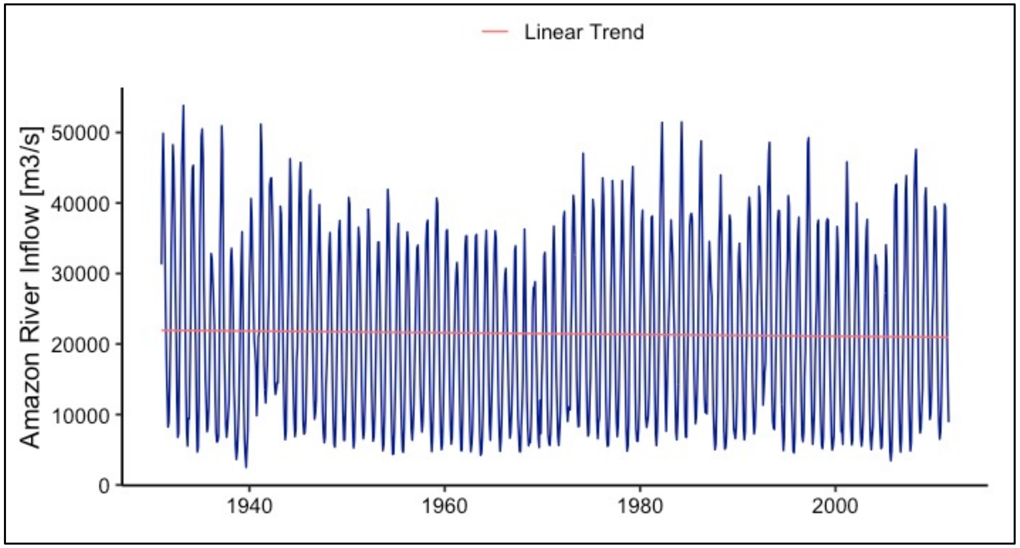
Trend visualization

$$Y_{trend_t} = \beta_0 + \beta_1 t$$



Trend visualization

$$Y_{trend_t} = \beta_0 + \beta_1 t$$

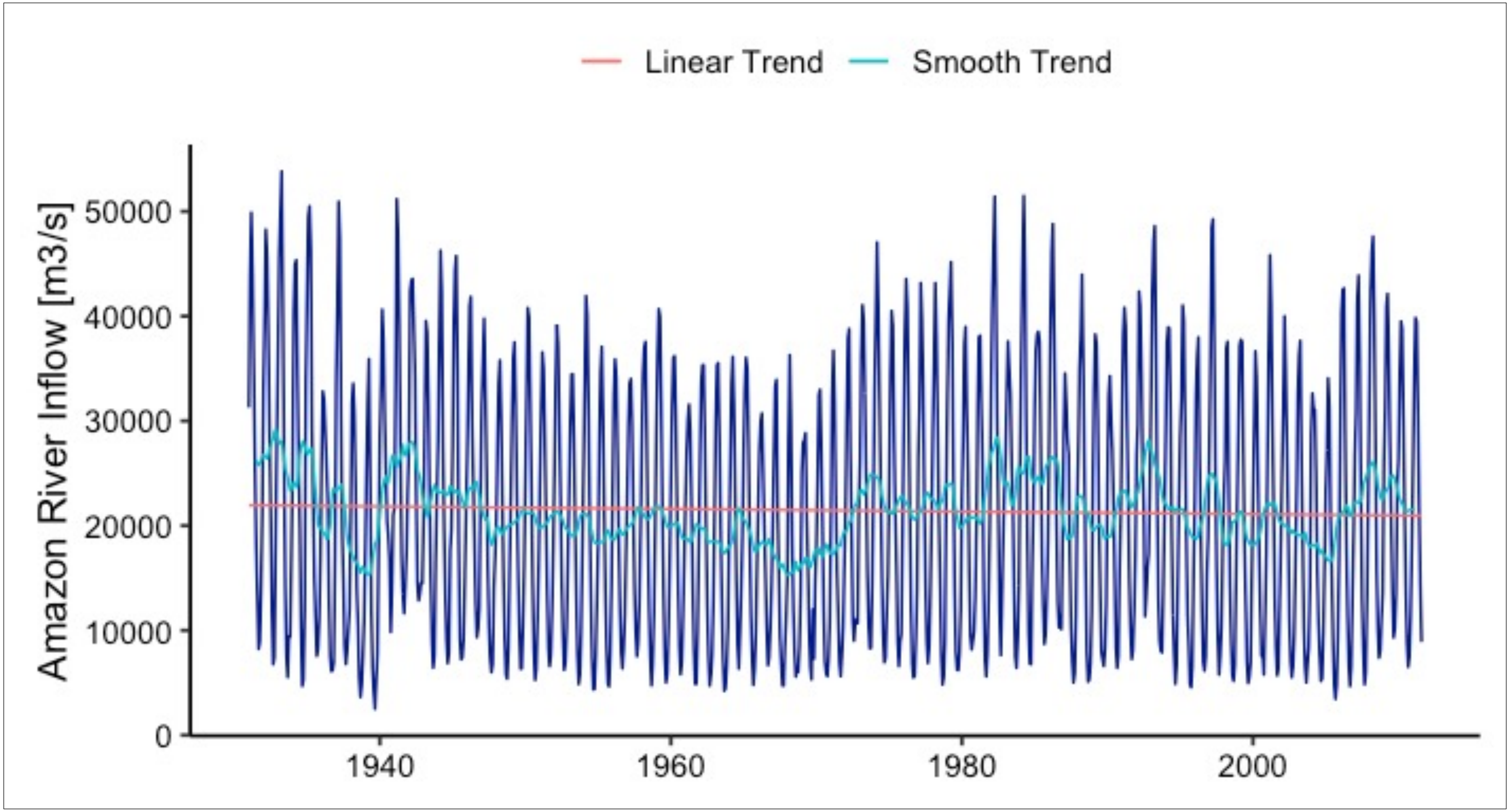



```
Call:
lm(formula = inflow ~ t, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-19337 -11555  -1483  10061  31900

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21949.403    797.962  27.507   <2e-16 ***
t              -1.024      1.427  -0.718    0.473
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12400 on 966 degrees of freedom
Multiple R-squared:  0.0005328, Adjusted R-squared:  -0.0005018
F-statistic: 0.515 on 1 and 966 DF,  p-value: 0.4732
```

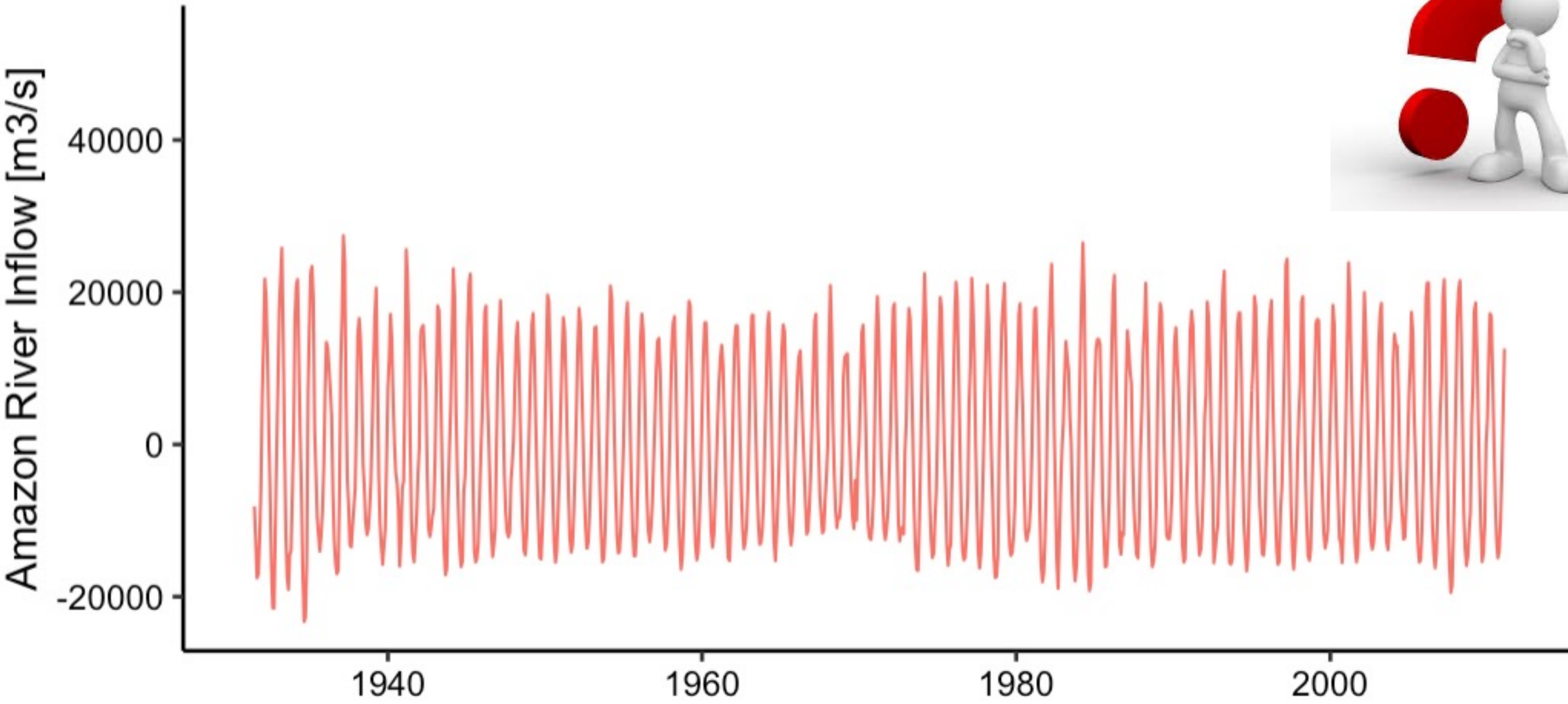
p-value > 0.05
Coefficient β_1 not significant

Linear vs Smoothed Trend

$$Y_{trend_t} = average(Y_{t-6} + Y_{t-5} + \cdots + Y_t + \cdots + \ Y_{t+5} + Y_{t+6})$$



Do you still see any patterns?

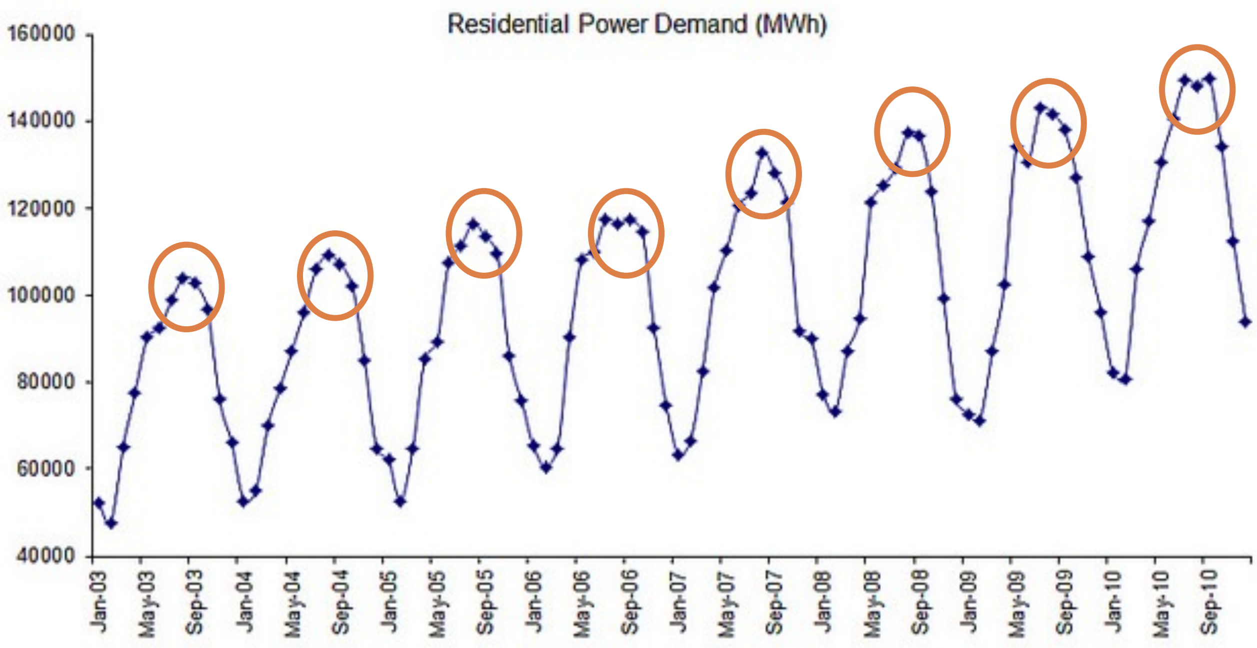


Seasonal Component

Seasonal Component

- Short-term regular wave-like patterns
  - Observed within 1 year - monthly or quarterly
  - Equally spaced peaks and troughs

Calendar Related



Peaks in the Summer months Jun/Jul/Ago

Seasonal Component Estimation

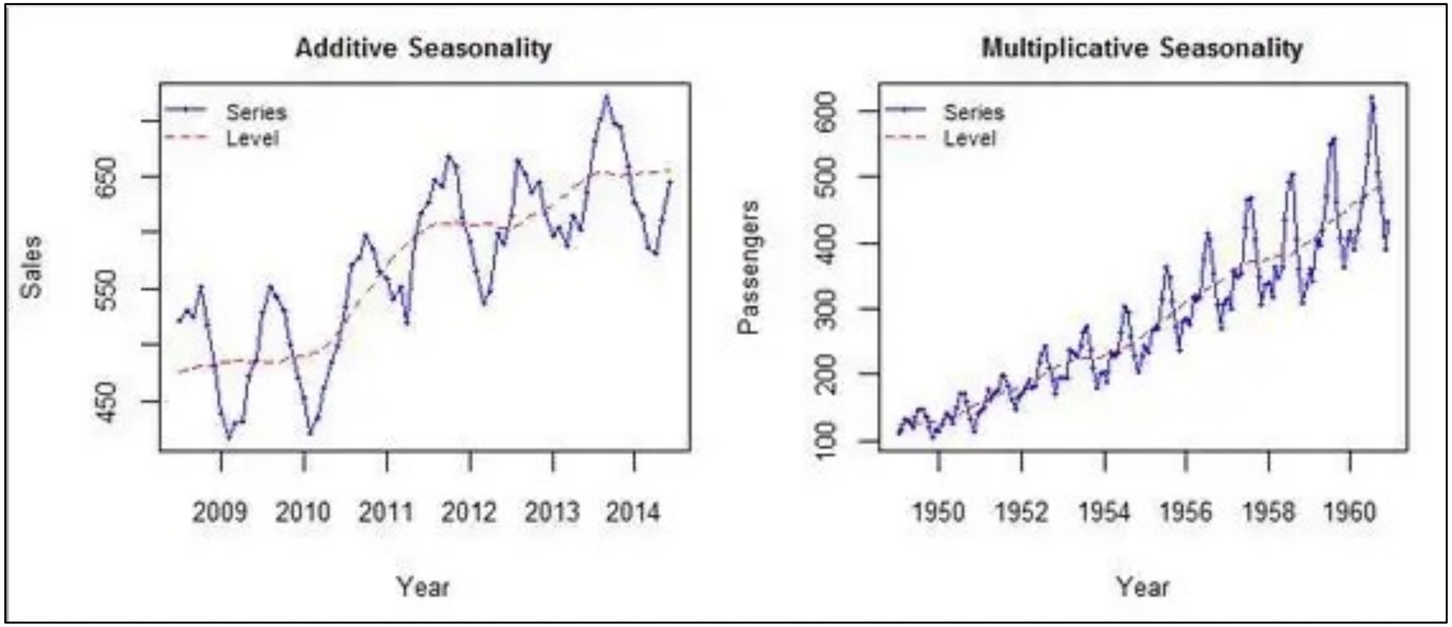
1. **Smoothing the trend** with a moving average
2. **De-trend the series**
   - Additive Model
     - take original series and **subtract** the smoothed trend

$$Y_{seasonal} = Y - Y_{trend}$$

   - Multiplicative model
     - scales the size of the seasonal component as the trend rises or falls
     - take original series and **divide** the original data by the trend

$$Y_{seasonal} = \frac{Y}{Y_{trend}}$$

Additive vs Multiplicative Model



- In the **additive model** the magnitude of seasonality does not change in relation to time
- In the **multiplicative model** the magnitude of the seasonal pattern depends on the magnitude/level of the data.

Seasonal Trend Estimation (cont'd)

3. Assume the observed detrended series can be represented as

$$Y_{seasonal_t} = \mu_t + X_t \qquad \text{where } E[X_t] = 0$$

- For monthly seasonal data assume 12 parameters such as

$$\mu_t = \begin{cases} \beta_1 & for\ t = 1,13,25,\cdots \\ \beta_2 & for\ t = 2,14,26,\cdots \\ & \vdots \\ \beta_{12} & for\ t = 12,24,36,\cdots \end{cases}$$

Seasonal Means Model

Seasonal Trend Estimation (cont'd)

4. Estimate the parameters $\beta_1, \beta_2, \ldots \beta_{12}$

Create dummies (categorical variables with 2 levels)

$$D_{s,t} = \begin{cases} 1 & if\ t\ belongs\ to\ season\ s \\ 0 & o.w. \end{cases} \qquad for\ s = 1,2, \ldots 12$$

At any time period t, one of the seasonal dummies $D_{1,t}$, $D_{2,t}$,…, $D_{12,t}$ will equal 1, all the others will equal 0

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
	HP1	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
Jan-31	4782	1	0	0	0	0	0	0	0	0	0	0	0
Feb-31	7323	0	1	0	0	0	0	0	0	0	0	0	0
Mar-31	8266	0	0	1	0	0	0	0	0	0	0	0	0
Apr-31	6247	0	0	0	1	0	0	0	0	0	0	0	0
May-31	3642	0	0	0	0	1	0	0	0	0	0	0	0
Jun-31	2425	0	0	0	0	0	1	0	0	0	0	0	0
Jul-31	2158	0	0	0	0	0	0	1	0	0	0	0	0
Aug-31	1854	0	0	0	0	0	0	0	1	0	0	0	0
Sep-31	1839	0	0	0	0	0	0	0	0	1	0	0	0
Oct-31	1896	0	0	0	0	0	0	0	0	0	1	0	0
Nov-31	2095	0	0	0	0	0	0	0	0	0	0	1	0
Dec-31	2725	0	0	0	0	0	0	0	0	0	0	0	1
Jan-32	4679	1	0	0	0	0	0	0	0	0	0	0	0
Feb-32	5535	0	1	0	0	0	0	0	0	0	0	0	0
Mar-32	4310	0	0	1	0	0	0	0	0	0	0	0	0
Apr-32	3026	0	0	0	1	0	0	0	0	0	0	0	0
May-32	2185	0	0	0	0	1	0	0	0	0	0	0	0
Jun-32	1919	0	0	0	0	0	1	0	0	0	0	0	0
Jul-32	1640	0	0	0	0	0	0	1	0	0	0	0	0
Aug-32	1302	0	0	0	0	0	0	0	1	0	0	0	0
Sep-32	1118	0	0	0	0	0	0	0	0	1	0	0	0
Oct-32	1688	0	0	0	0	0	0	0	0	0	1	0	0
Nov-32	2040	0	0	0	0	0	0	0	0	0	0	1	0
Dec-32	3790	0	0	0	0	0	0	0	0	0	0	0	1
Jan-33	6928	1	0	0	0	0	0	0	0	0	0	0	0
Feb-33	5793	0	1	0	0	0	0	0	0	0	0	0	0
Mar-33	4276	0	0	1	0	0	0	0	0	0	0	0	0
Apr-33	3863	0	0	0	1	0	0	0	0	0	0	0	0
May-33	2498	0	0	0	0	1	0	0	0	0	0	0	0
Jun-33	1940	0	0	0	0	0	1	0	0	0	0	0	0
Jul-33	1725	0	0	0	0	0	0	1	0	0	0	0	0
Aug-33	1375	0	0	0	0	0	0	0	1	0	0	0	0
Sep-33	1324	0	0	0	0	0	0	0	0	1	0	0	0
Oct-33	1551	0	0	0	0	0	0	0	0	0	1	0	0
Nov-33	1724	0	0	0	0	0	0	0	0	0	0	1	0
Dec-33	3352	0	0	0	0	0	0	0	0	0	0	0	1
Jan-34	4049	1	0	0	0	0	0	0	0	0	0	0	0
Feb-34	3166	0	1	0	0	0	0	0	0	0	0	0	0
Mar-34	3124	0	0	1	0	0	0	0	0	0	0	0	0
Apr-34	2507	0	0	0	1	0	0	0	0	0	0	0	0
May-34	1853	0	0	0	0	1	0	0	0	0	0	0	0
Jun-34	1131	0	0	0	0	0	1	0	0	0	0	0	0
Jul-34	978	0	0	0	0	0	0	1	0	0	0	0	0
Aug-34	826	0	0	0	0	0	0	0	1	0	0	0	0
Sep-34	1026	0	0	0	0	0	0	0	0	1	0	0	0
Oct-34	1203	0	0	0	0	0	0	0	0	0	1	0	0
Nov-34	1199	0	0	0	0	0	0	0	0	0	0	1	0
Dec-34	1621	0	0	0	0	0	0	0	0	0	0	0	1

$$Y_{seasonal_t} = \beta_1 D_{1,t} + \beta_2 D_{2,t} + \beta_3 D_{3,t} + \beta_4 D_{4,t} + \beta_5 D_{5,t} + \beta_6 D_{6,t} + \beta_7 D_{7,t} + \beta_8 D_{8,t} + \beta_9 D_{9,t} + \beta_{10} D_{10,t} + \beta_1 D_{11,t} + \beta_{12} D_{12,t}$$

Seasonal Trend Estimation (cont'd)

5. Write series $Y_{seasonal}$ as a function of the dummies

$$Y_{seasonal_t} = \sum_{s=1}^{12} \beta_s D_{s,t}$$

6. Compute coefficients by linear regression

Estimating Seasonal Trend in R

- First create seasonal dummies using *seasonaldummy()* from package ”forecast”

$$dummies = seasonaldummy(Y)$$

- This will only work if Y is a time series object and if you specify frequency

$$Y = ts(Y, frequency\ =\ 12)$$

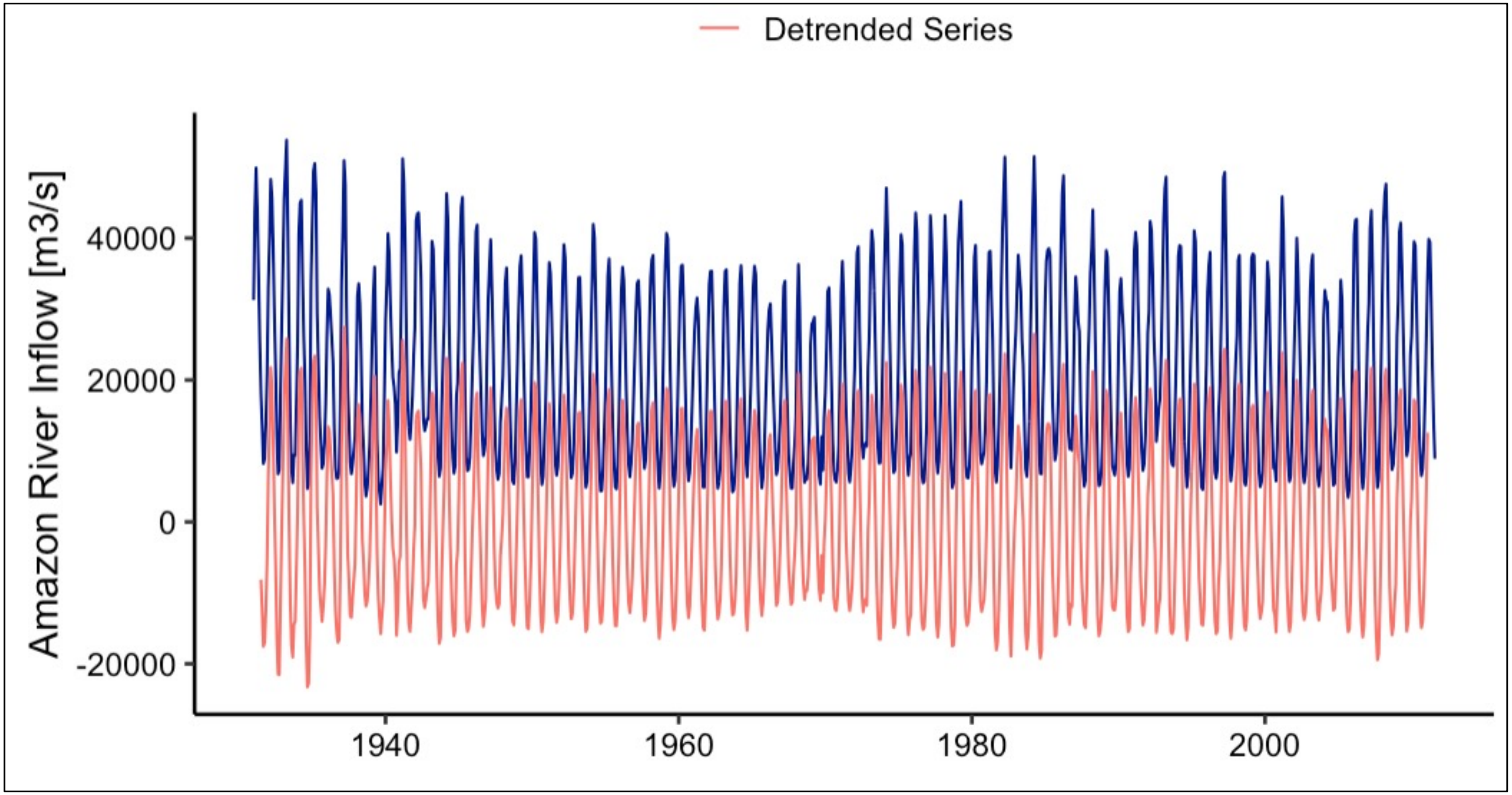
- Then just run a simple regression on the dummies

$$lm(\ Y \sim dummies, data)$$

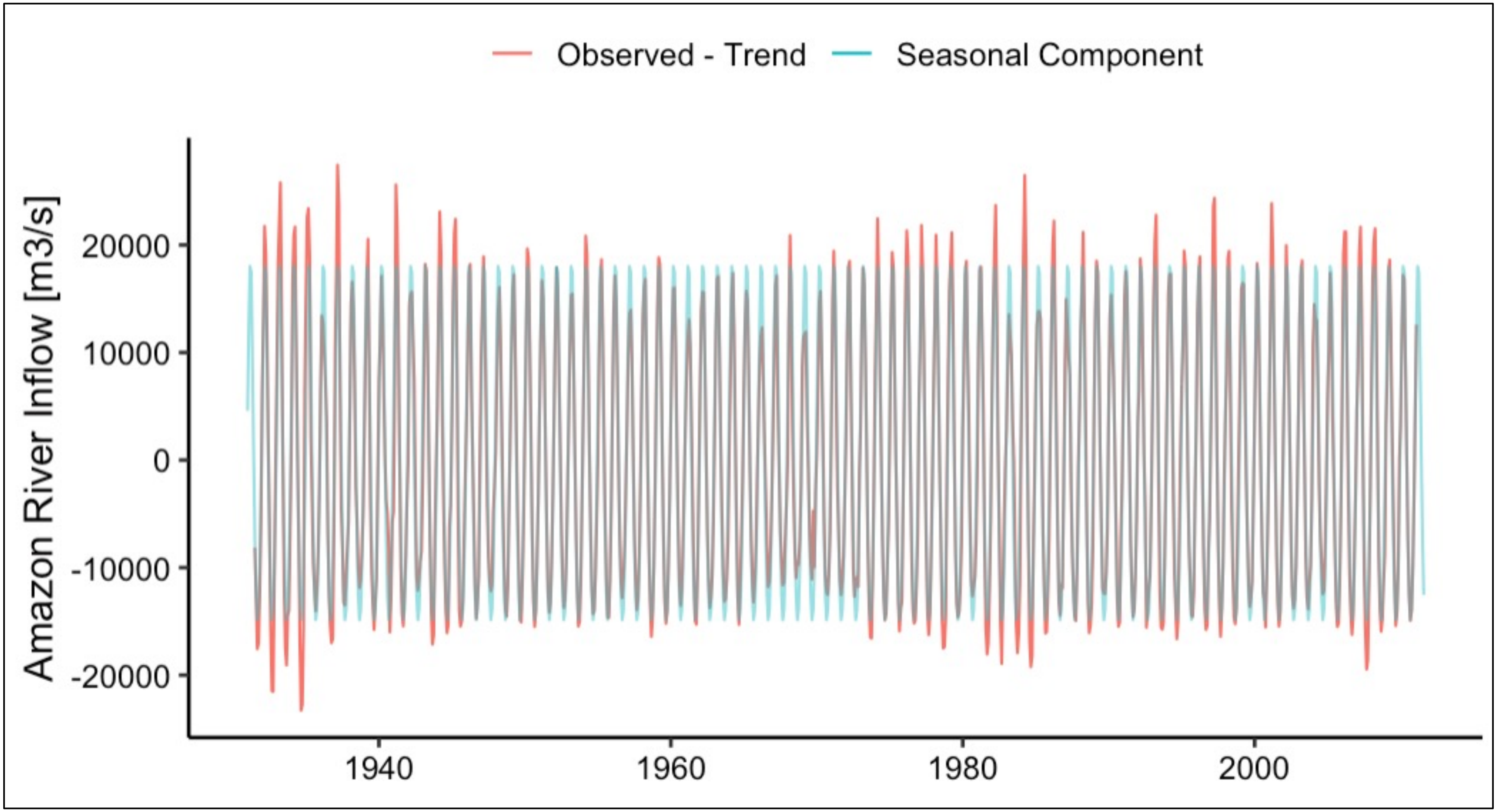
Back to example: Inflow

Detrended Series with Additive Model

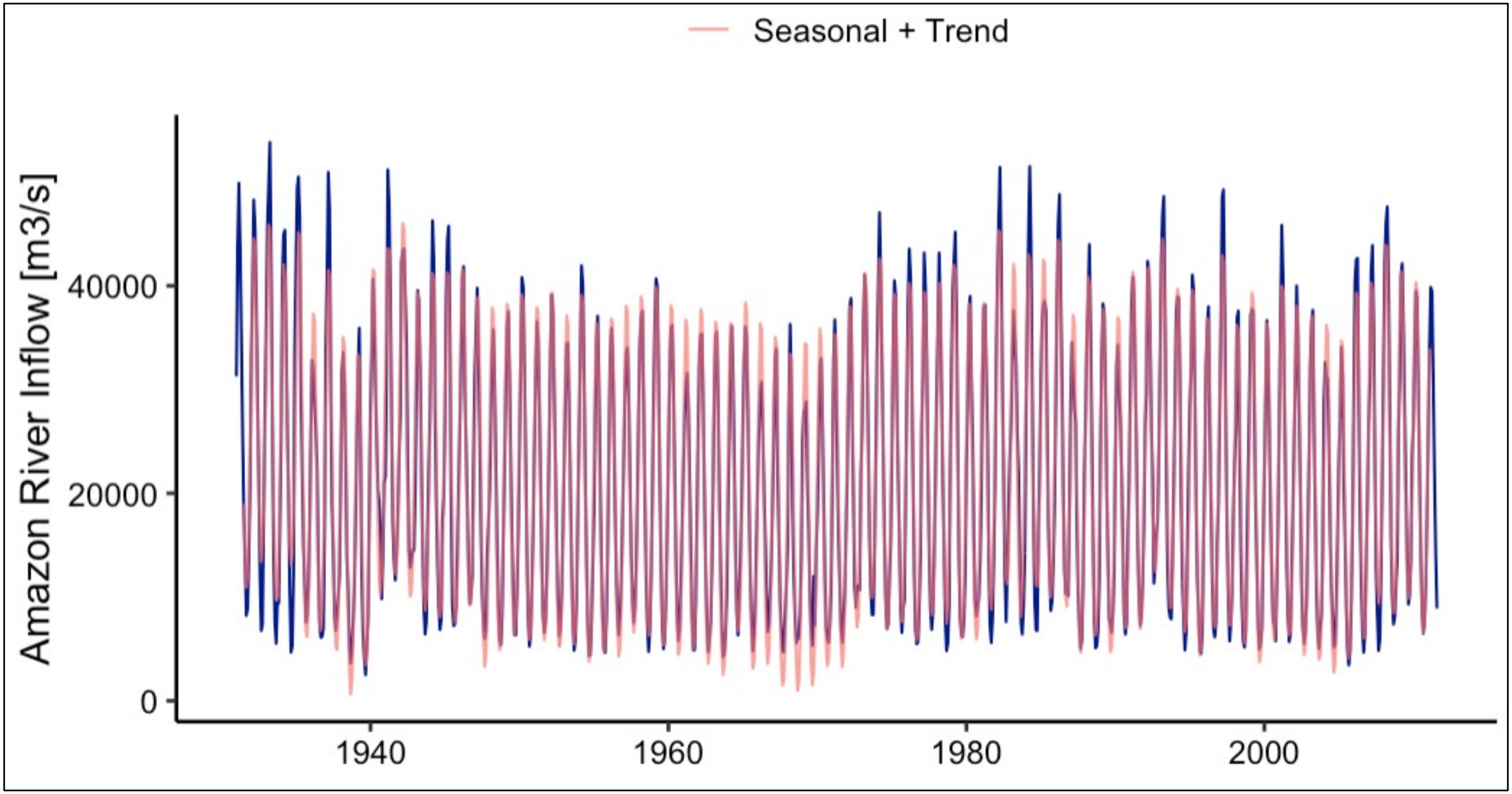
$$Y_{detrend_t} = Y_t - Trend_t$$



Seasonal Component Visualization



Seasonal + Trend Decomposition



Stochastic versus deterministic trend

Series with Deterministic Trend

- Deterministic linear trend process

$$Y_i = \beta_0 + \beta_1 t_i + \varepsilon_i$$

- Or more generally, for a polynomial trend

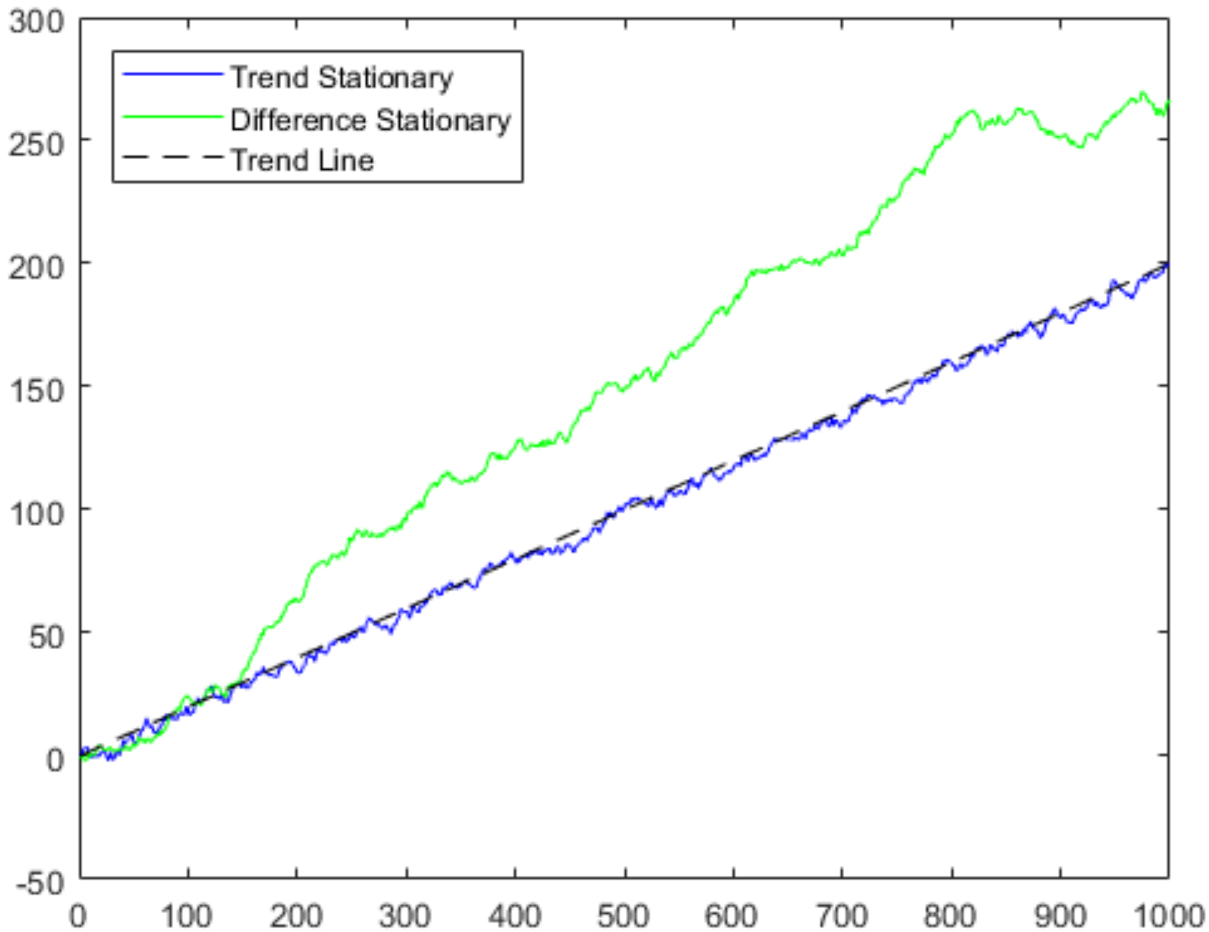
$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 T_i^2 + \cdots + \beta_n T_i^n + \varepsilon_i$$

- Detrending is accomplished by running a regression and obtaining the series of residuals. The residuals will give you the detrended series
- That's what we call **trend-stationarity**

Series with Stochastic Trend

- But some series have what we call **difference-stationarity**
- Although trend-stationary and difference-stationary series are both "trending" over time, the stationarity is achieved by a distinct procedure
- In the case of difference-stationarity, stationarity is achieved by differencing the series
- Sometimes we need to difference the series more than once

Trend-stationarity vs difference-stationarity



Stationarity Tests

Stationarity Assessment

- **Mann-Kendall Test– monotonic trend**
- **Spearman's Rank Correlation Test – monotonic trend**

- **Dickey-Fuller (ADF) Test – unit root**
- Phillips-Perron (PP) Test – unit root
- Kitawoski-Phillips-Schmidt-Shin (KPSS) – unit root

- And others…

Review: Hypothesis Testing

- Why do we use hypothesis testing?
  - To analyze evidence provided by data
  - To make decisions based on data
- What is a statistical hypothesis?
  - An assumption about a population parameter that may or may not be true
- In Hypothesis Testing we usually have

$$\begin{cases} H_0: & \textit{the null hypothesis} \\ H_1: & \textit{the alternative hypothesis} \end{cases}$$

Review: Hypothesis Testing (cont'd)

- Procedure
  1. State the hypotheses and identify the claim
  2. Find the critical value(s) from the appropriate table
  3. Compute the test value
  4. Make the decision to reject or not reject the null hypothesis

If P-value $\leq \alpha$, **reject** the null hypothesis.
If P-value $> \alpha$, **do not reject** the null hypothesis.

Mann-Kendall Test

- Commonly employed to detect deterministic trends in series of environmental data, climate data or hydrological data
- **Cannot be applied to seasonal data**
- Hypothesis Test

$$\begin{cases} H_0: \quad Y_t \quad is \quad i.i.d.(stationary) \\ H_1: \quad Y_t \quad follow\ a\ trend \end{cases}$$

Mann-Kendall Test

- Mann-Kendall statistic is

$$S = \sum_{k=1}^{N-1} \sum_{j=k+1}^{N} sgn(Y_j - Y_k)$$

where

$$sgn(Y_j - Y_k) = \begin{cases} 1 & if \quad Y_j - Y_k > 0 \\ 0 & if \quad Y_j - Y_k = 0 \\ -1 & if \quad Y_j - Y_k < 0 \end{cases}$$

- The test will check the magnitude of S and its significance based on the number of observations
- In other words, the bigger the number of observations the higher S will need to be

Mann-Kendall Test

$$\begin{cases} H_0: & Y_t \quad is \quad i.i.d.(stationary) \\ H_1: & Y_t \quad follow\ a\ trend \end{cases}$$

Mann-Kendall test statistic is

$$S = \sum_{k=1}^{N-1} \sum_{j=k+1}^{N} sgn(Y_j - Y_k) \quad \rightarrow \quad sgn(Y_j - Y_k) = \begin{cases} 1 & if \quad Y_j - Y_k > 0 \\ 0 & if \quad Y_j - Y_k = 0 \\ -1 & if \quad Y_j - Y_k < 0 \end{cases}$$

$$E[S] = 0$$

$$\text{Var}[S] = \sigma_s^2 = \frac{1}{18} n(n-1)(2n+5)$$

$$\tau = \frac{2S}{N(N-1)}$$

Under H_0, Z follow a standard normal distribution

$$Z = \begin{cases} \dfrac{(S-1)}{\sigma_s} & if \quad S > 0 \\ 0 & if \quad S = 0 \\ \dfrac{(S+1)}{\sigma_s} & if \quad S < 0 \end{cases}$$

Reject H_0 when $Z < Z_{\alpha/2}$

Mann-Kendall test in R

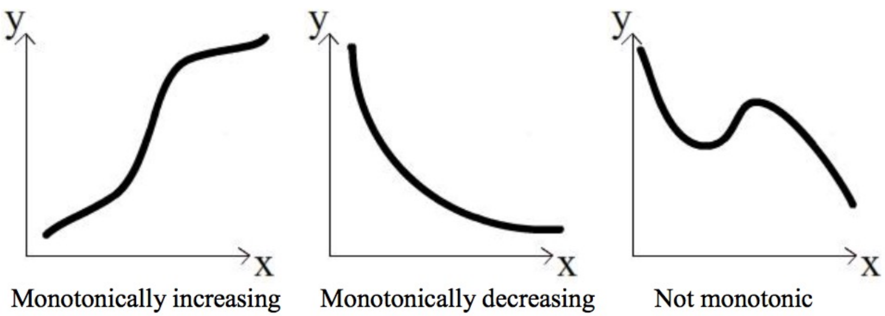
- The Mann-Kendall test in R is done with the command MannKendall() from package "Kendall"

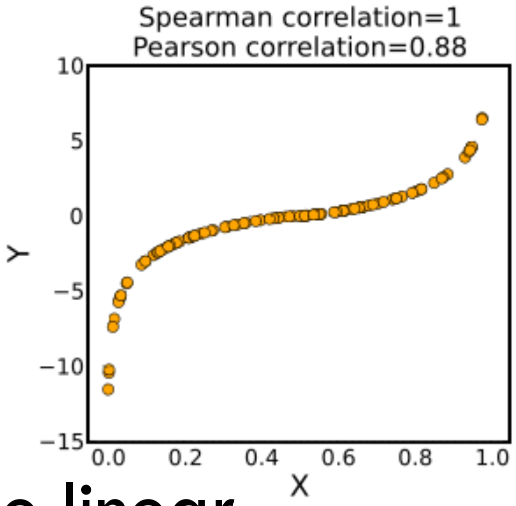
> **Description**
>
> This is a test for monotonic trend in a time series z[t] based on the Kendall rank correlation of z[t] and t.
>
> **Usage**
>
> `MannKendall(x)`
>
> **Arguments**
>
> `x` a vector of data, often a time series
>
> **Details**
>
> The test was suggested by Mann (1945) and has been extensively used with environmental time series (Hipel and McLeod, 2005). For autocorrelated time series, the block bootstrap may be used to obtain an improved signficance test.

- For seasonal data you can use SeasonalMannKendall() from the same package

Spearman's Rank Correlation Coefficient

- Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship





  - Unlike Pearson's correlation, the relationship does not need to be linear
  - In other words, if one variable increases so do does the other, it does not matter the proportion of the increase

Spearman's Rank Correlation Coefficient

- To verify a monotonic trend in your data, compute the spearman correlation between your data and series T

Y_t	T
Y_1	1
Y_2	2
Y_3	3
$\vdots$	$\vdots$
Y_{N-2}	$N-2$
Y_{N-1}	$N-1$
Y_N	N

- If the correlation is close to 0, then there is no trend

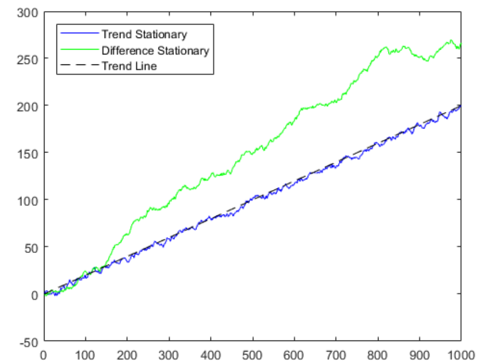
- The function to compute spearman correlation is cor() or the cor.test() from package "stats". The latter provides the significance of the coefficient

Dick-Fuller Test

- The first work on testing for a unit root in time series was done by Dickey and Fuller
- Consider the model

White noise series

$$Y_t = a + \phi Y_{t-1} + \epsilon_t$$



- The objective is to test

$$\begin{cases} H_0: \quad \phi = 1 \ \ (i.e. \ \ contain\ a\ unit\ root), \\ H_1: \quad |\phi| < 1 \ \ (i.e. \quad is\ stationary) \end{cases}$$

- More general case can include more lags, the so called Augmented Dickey-Fuller (ADF) test

Dick-Fuller Test in R

- The ADF test in R is done with the command adf.test() from package “tseries”

Description

Computes the Augmented Dickey-Fuller test for the null that `x` has a unit root.

Usage

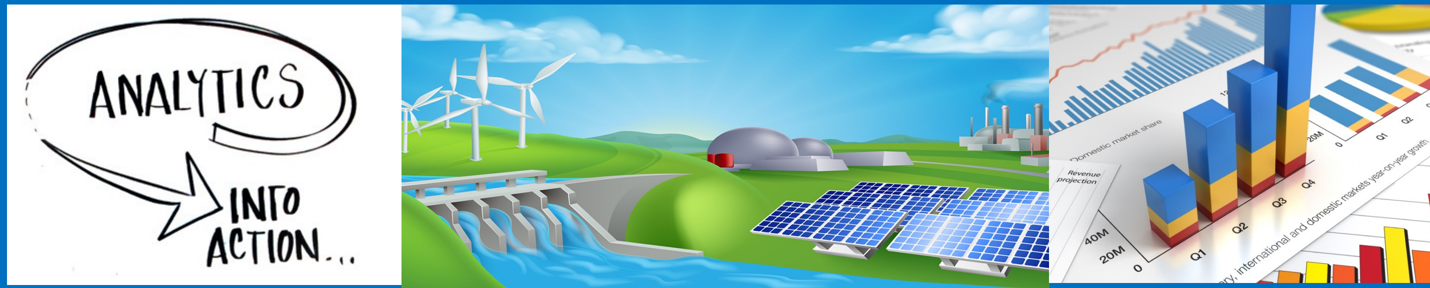
```
adf.test(x, alternative = c("stationary", "explosive"),
         k = trunc((length(x)-1)^(1/3)))
```

Arguments

`x`	a numeric vector or time series.
`alternative`	indicates the alternative hypothesis and must be one of `"stationary"` (default) or `"explosive"`. You can specify just the initial letter.
`k`	the lag order to calculate the test statistic.

Summary of Stationary Tests

Mann Kendall	Spearman Correlation	Augmented Dickey-Fuller
Check for deterministic trend	Check for deterministic trend	Check for stochastic trend
Hypothesis test	Hypothesis test	Hypothesis test
$\begin{cases} H_0: & Y_t \ \ is \ \ i.i.d.(stationary) \\ H_1: & Y_t \ \ follow\ a\ trend \end{cases}$	$\begin{cases} H_0: & Y_t \ \ is \ \ i.i.d.(stationary) \\ H_1: & Y_t \ \ follow\ a\ trend \end{cases}$	$\begin{cases} H_0: & \phi = 1 \ \ (i.e. \ \ contain\ a\ unit\ root) \\ H_1: & \phi < 1 \ \ (i.e. \ \ is\ stationary) \end{cases}$
Test statistic	Test statistic	Test statistic
Find $S = \sum_{k=1}^{N-1} \sum_{j=k+1}^{N} sgn(Y_j - Y_k)$ $sgn(Y_j - Y_k) = \begin{cases} 1 & if \ \ Y_j - Y_k > 0 \\ 0 & if \ \ Y_j - Y_k = 0 \\ -1 & if \ \ Y_j - Y_k < 0 \end{cases}$	Find the spearman correlation coefficient $\rho = Corr(Y_t, T)$ where $T = 1, \dots, N$ PS: spearman measure any type of monotonic relationship not only linear	Check if model $Y_t = \phi Y_{t-1} + \epsilon_t$ has a unit root i.e. ϕ=1
Can't handle seasonality, if working with seasonal data use Seasonal Mann Kendall instead or group data	Can't handle seasonality, if working with seasonal data use group data	Can handle seasonality



THANK YOU !

luana.marangon.lima@duke.edu