ESM 244: 9

- What is time series data?
- Exploratory visualization
- Decomposition
- Autocorrelation intro

1. What is time series data?

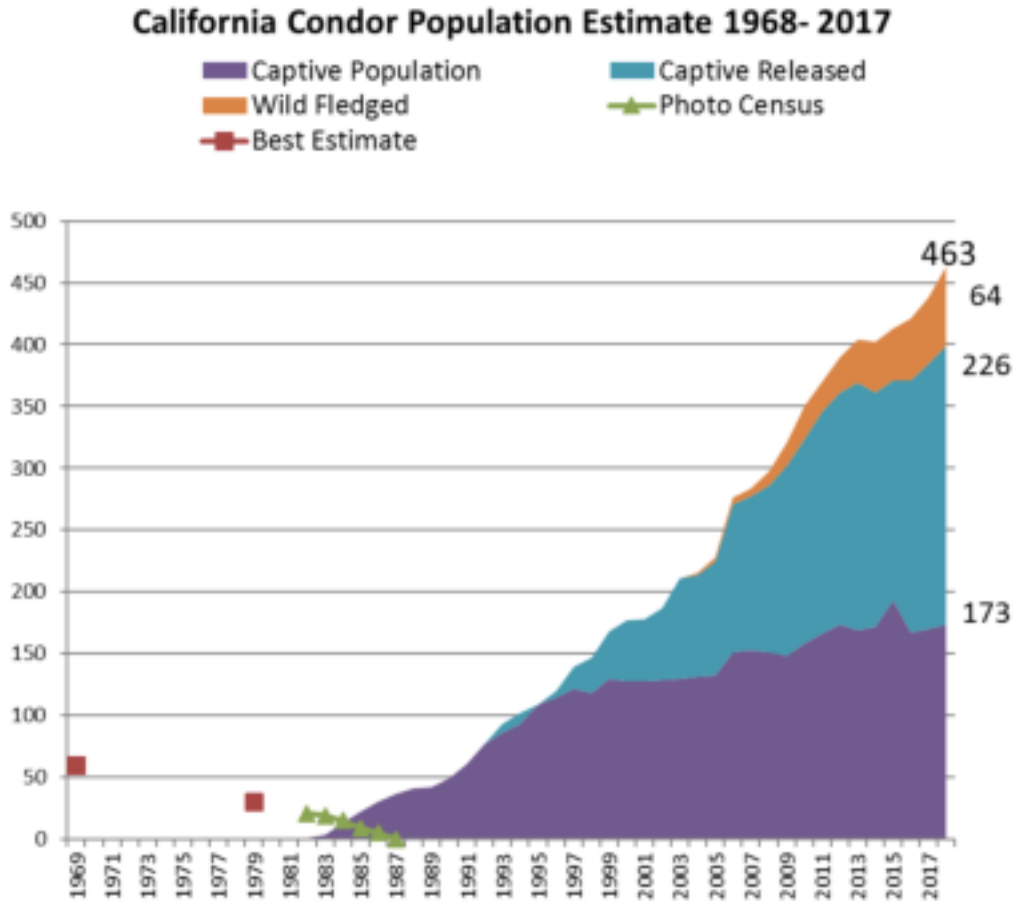
- Data in which variable observations are recorded over time, often (but not necessarily) at regular intervals

- In environmental sciences, often high-frequency monitoring data (stream flow gauges, air quality monitoring, temperatures, NOAA buoy data, etc.)

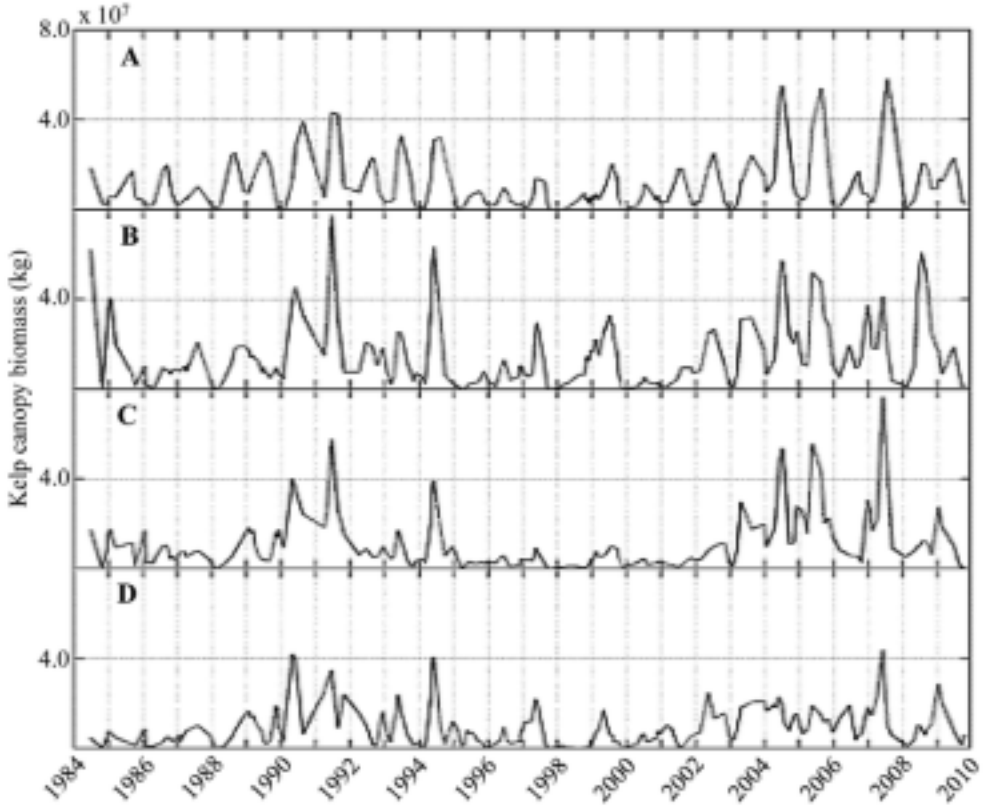
United States Department of the Interior

FISH AND WILDLIFE SERVICE
HOPPER MOUNTAIN NATIONAL WILDLIFE REFUGE COMPLEX
CALIFORNIA CONDOR RECOVERY PROGRAM
Tel: (805) 644-5185 Fax: (805) 644-1732

California Condor Population Estimate 1968- 2017



Kavanaugh, KC et al. **2011**. *Environmental controls of giant-kelp biomass in the Santa Barbara Channel, California.* Marine Ecology Progress Series 429:1-17.



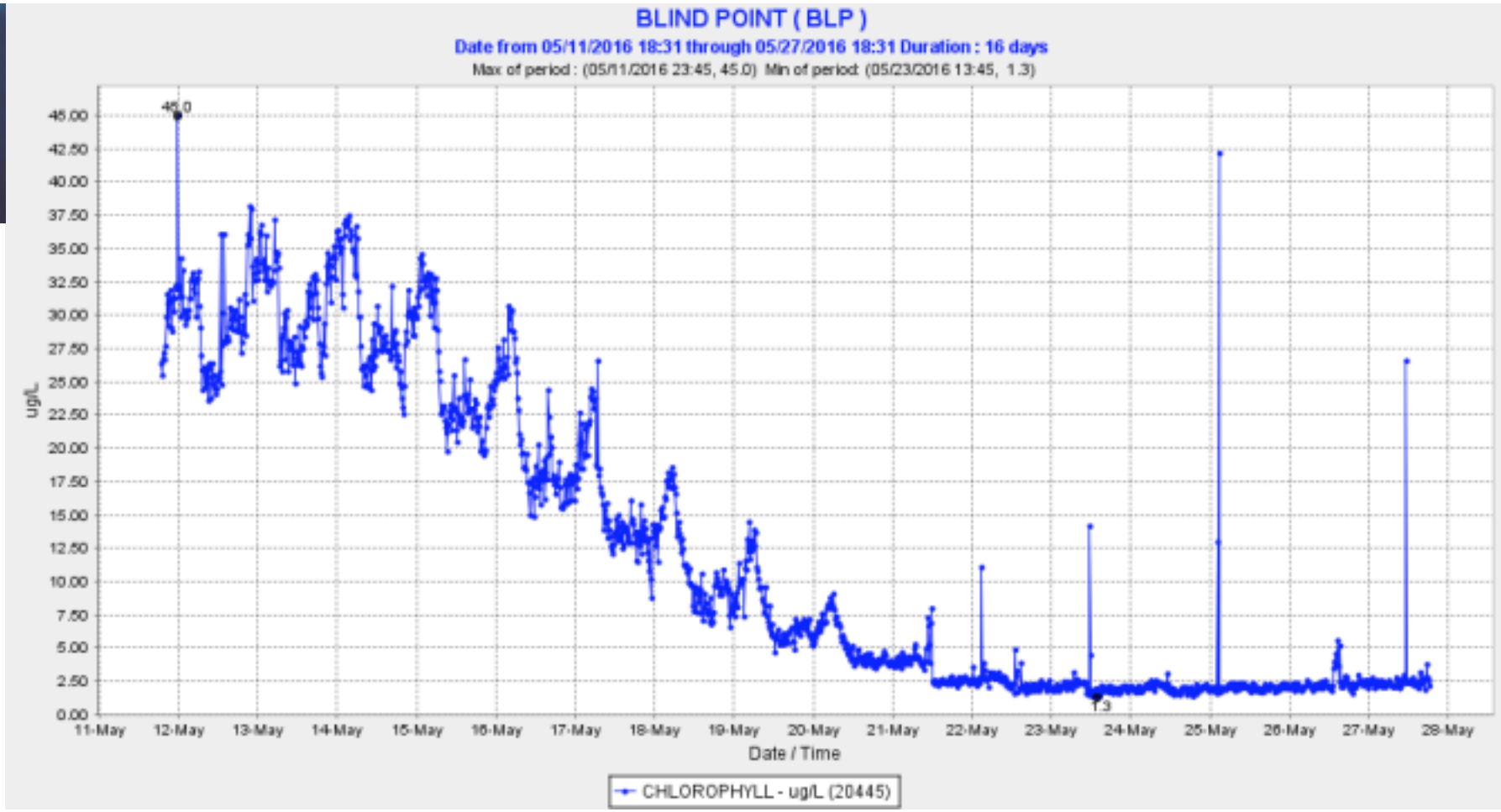
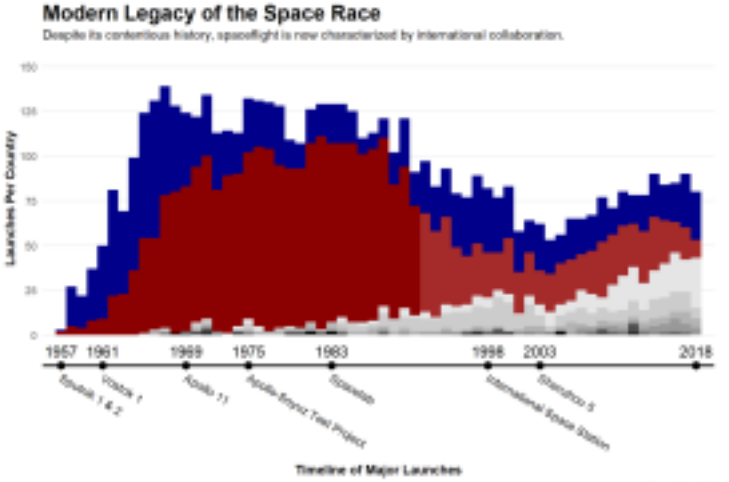
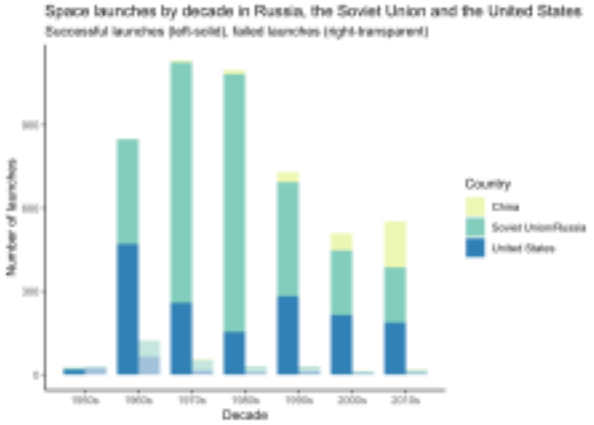


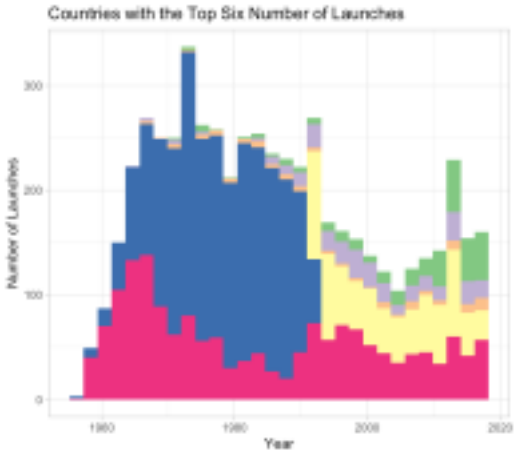
Figure 1. Chlorophyll concentration May 12-27, 2016 in the lower San Joaquin River channel of the Delta east of Antioch near Sherman Island. Concentrations above 10 micrograms per liter of water are considered indicative of high phytoplankton production – a "bloom". Source: CDEC.

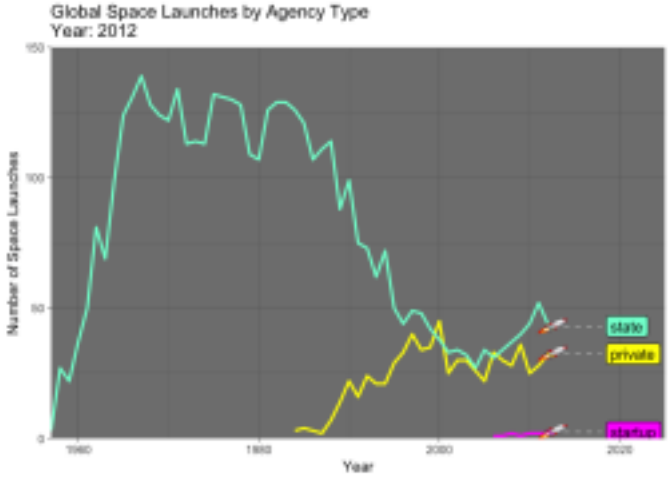
2. You're already working with time series data

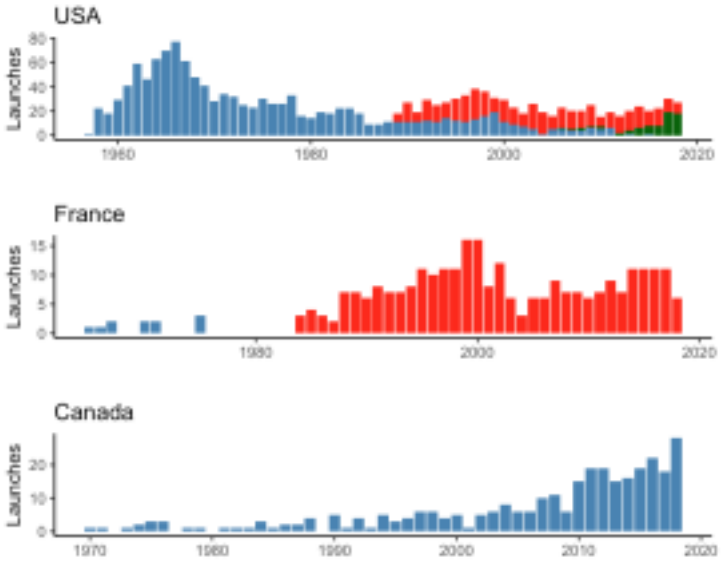

Modern Legacy of the Space Race
Despite its contentious history, spaceflight is now characterized by international collaboration.
Launches Per Country
Country
United States
Soviet Union
Russia
China
France
Japan
India
Israel
Other*
International
1957
1961
1969
1975
1983
1998
2003
2018
Sputnik 1 & 2
Vostok 1
Apollo 11
Apollo-Soyuz Test Project
Salyut
International Space Station
Shenzhou 5
Timeline of Major Launches
*Countries with less than ten launches


Number of Failed and Successful Launches During the Cold War (1957-1991)
Failed Launch
Successful Launch
Number of Launches
Year
Country
Soviet Union
United States


Space launches by decade in Russia, the Soviet Union and the United States
Successful launches (left-solid), failed launches (right-transparent)
Number of launches
Decade
Country
China
Soviet Union/Russia
United States


Countries with the Top Six Number of Launches
Number of Launches
Year
Country
China
France
Japan
Russia
Soviet Union
United States

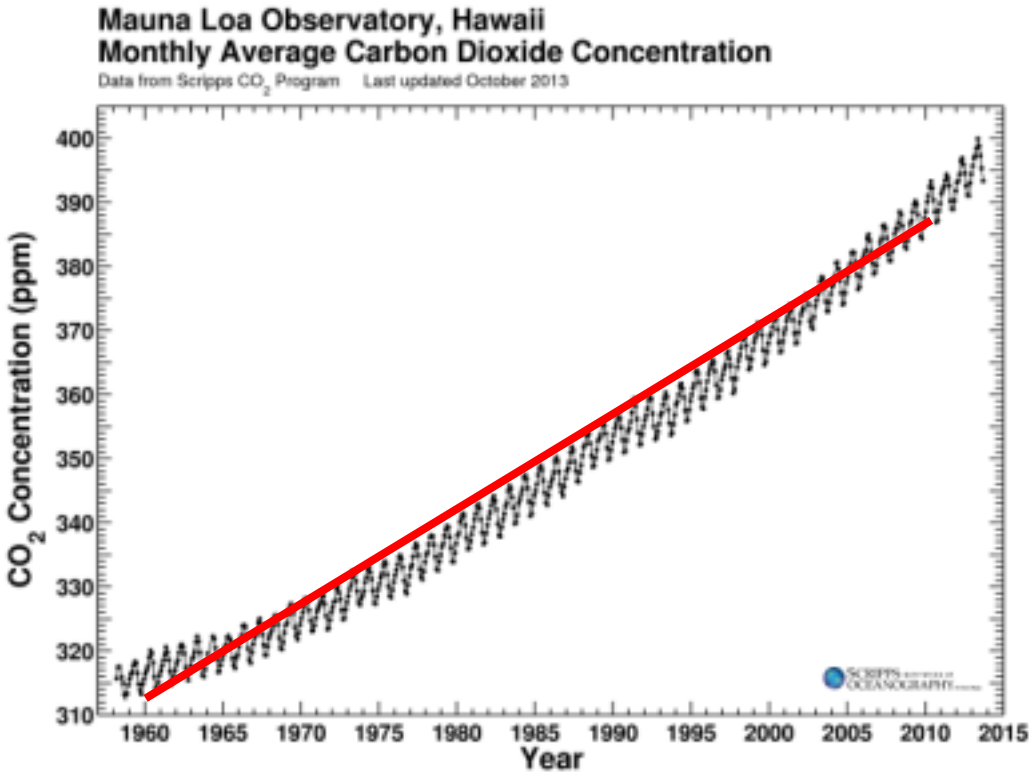

Global Space Launches by Agency Type
Year: 2012
Number of Space Launches
Year
state
private
startup


USA
France
Canada
Launches

3. Don’t automatically throw the methods you’ve already learned in the trash can.

Think really hard about your question, first.

Example: On average, at what rate was [CO_2] increasing between 1960 and 2010?



4. ...but sometimes there are questions about time series data that require time series-specific approaches

- Describe major/important features of time-series patterns, trends or data
- Forecasting future values
- Explain how past observations influence subsequent or future values

Always: Go exploring!

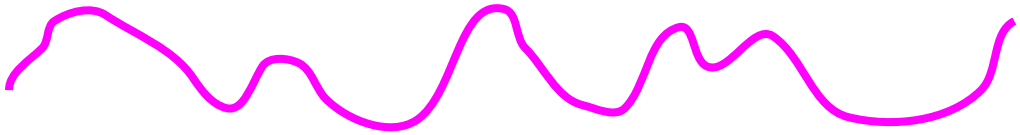
STEP 1: Look at the data

STEP 2: Ask questions

- Overall ***trend***?
- ***Seasonality***?
- ***Cyclical***?
- Outliers?
- Changes in variance?

- ***SEASONALITY*** 

  Is there a pattern repeated over known and equal periods?

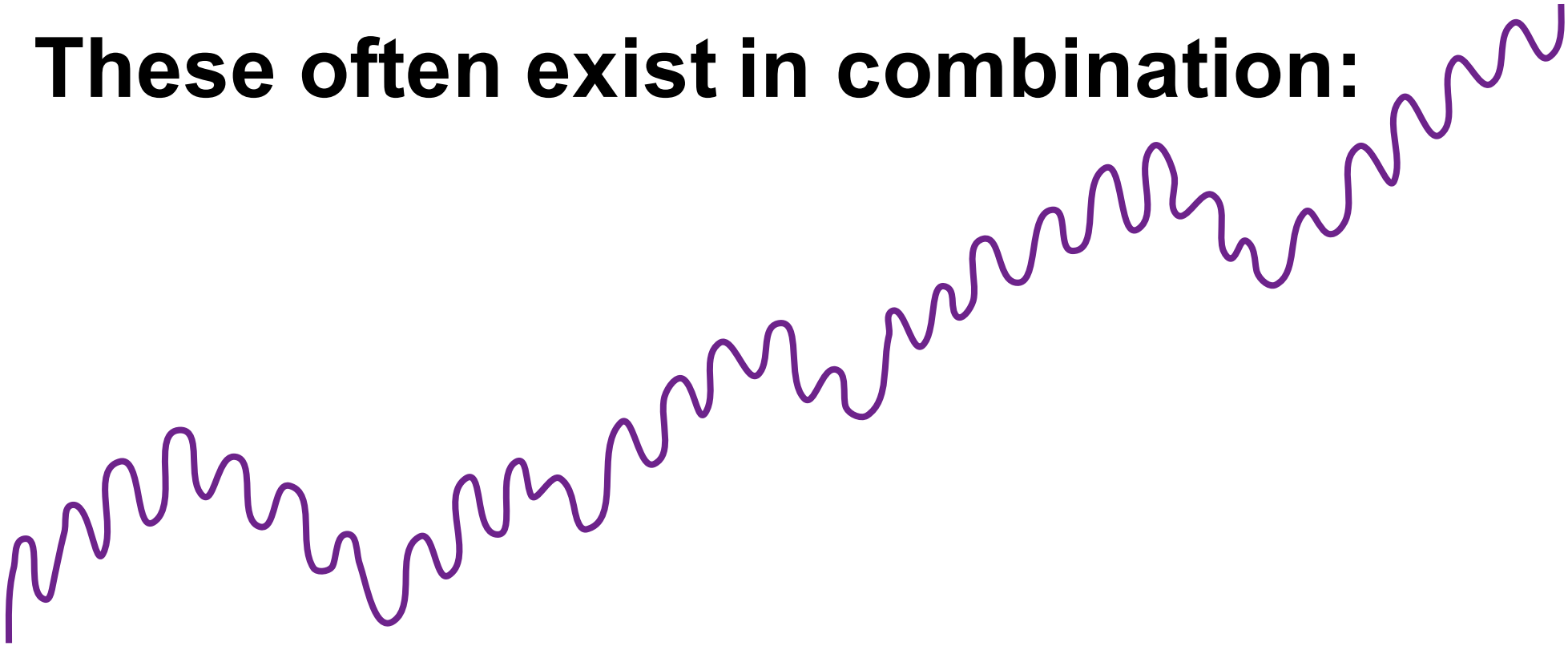
- ***CYCLICAL*** 

  Is there a broader cyclical trend that exists that has unknown or unequal periods?

- Overall ***TREND***? 

  Is there an overall trend (e.g. general increase/decrease) beyond any seasonal or cyclical patterns?

These often exist in combination:



Let’s ask these exploratory questions for a few real-world examples.

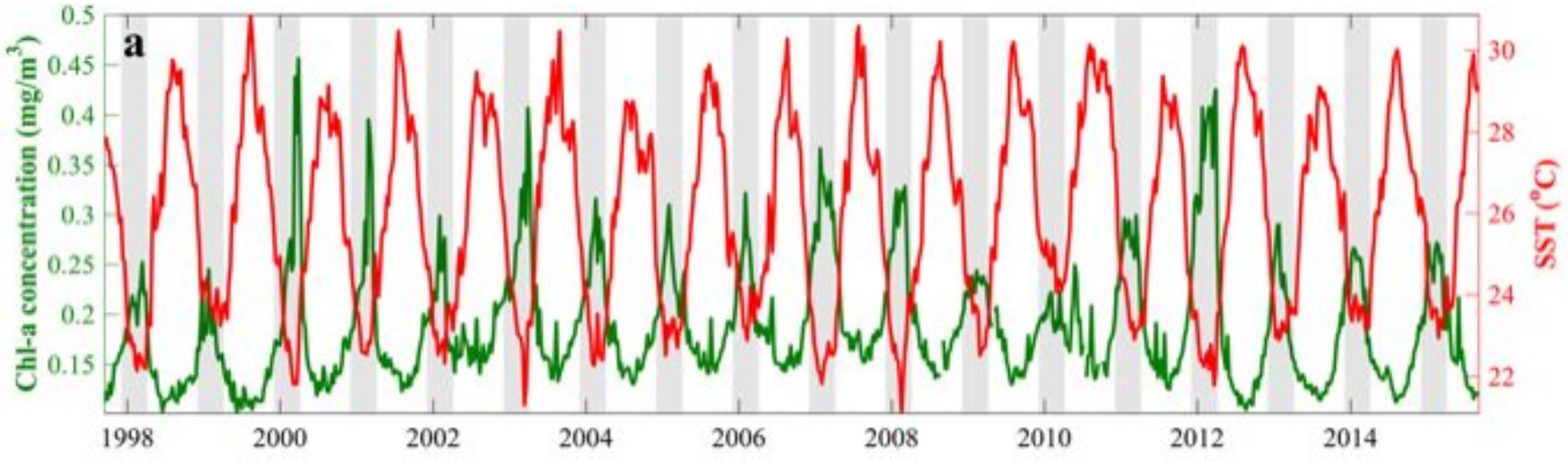
SCIENTIFIC REPORTS

OPEN

Impacts of warming on phytoplankton abundance and phenology in a typical tropical marine ecosystem

Received: 7 August 2017
Accepted: 21 January 2018
Published online: 02 February 2018

John A. Gittings[1], Dionysios E. Raitsos[2], George Krokos[1] & Ibrahim Hoteit[1]

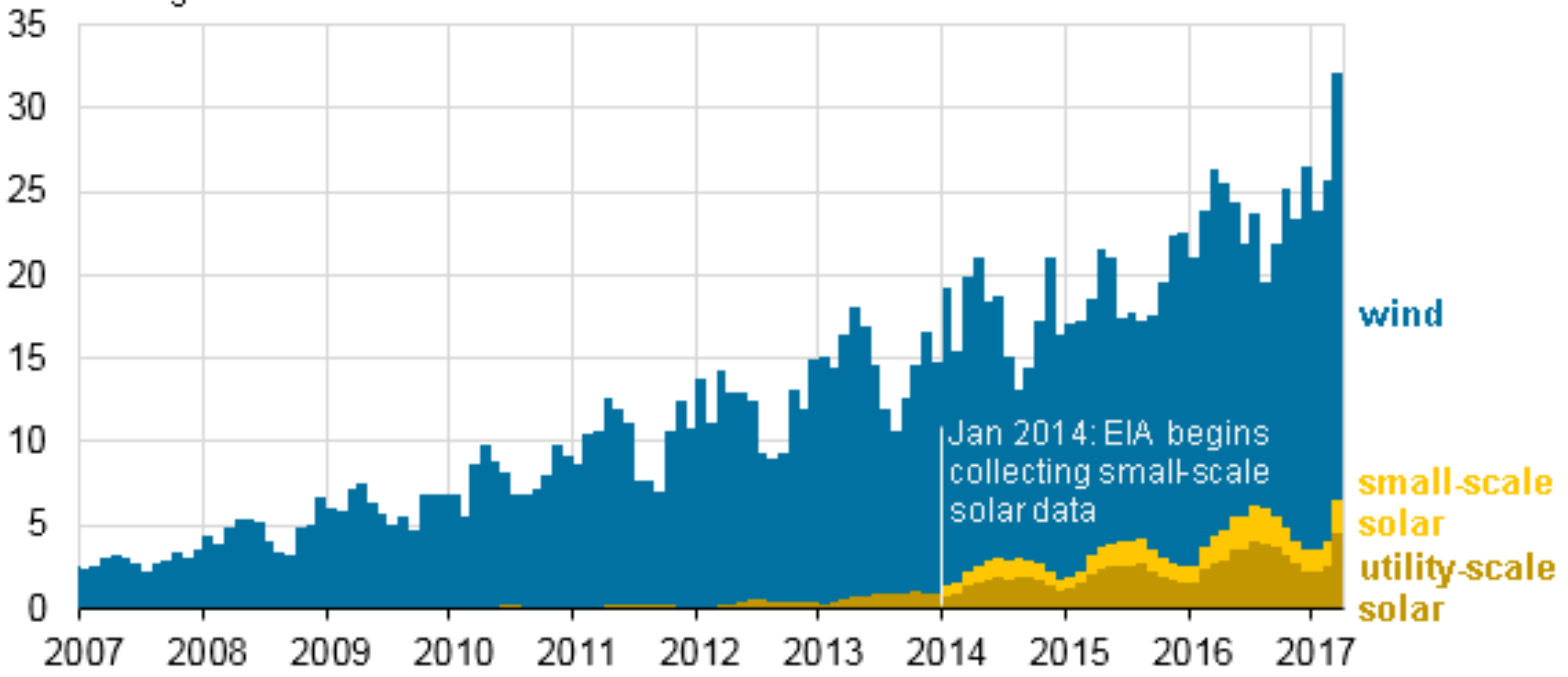


(**a**) Time series of satellite-derived Chl-a concentration and sea surface temperature (8-day averages) for the northern Red Sea (1998–2015).

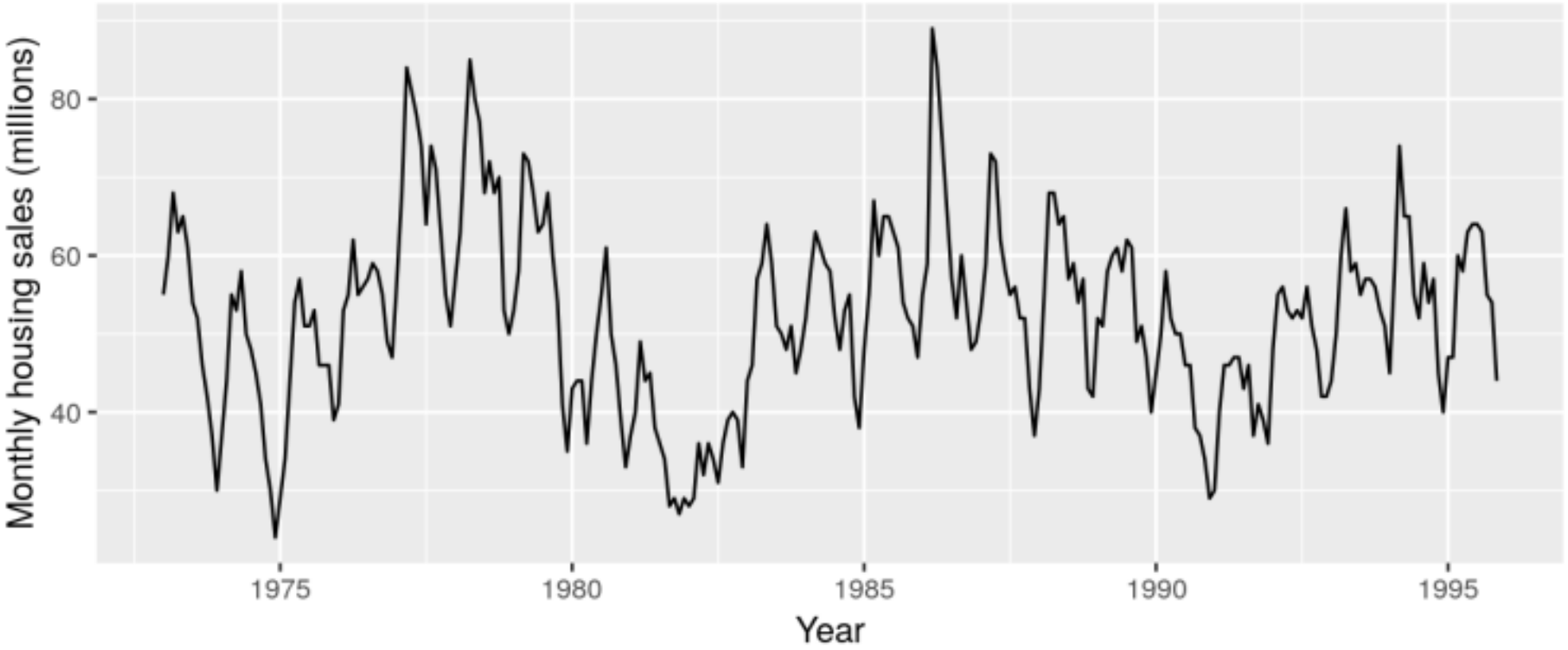
Independent Statistics & Analysis
U.S. Energy Information Administration

Monthly net electricity generation from selected fuels (Jan 2007 - Mar 2017)

million megawatthours



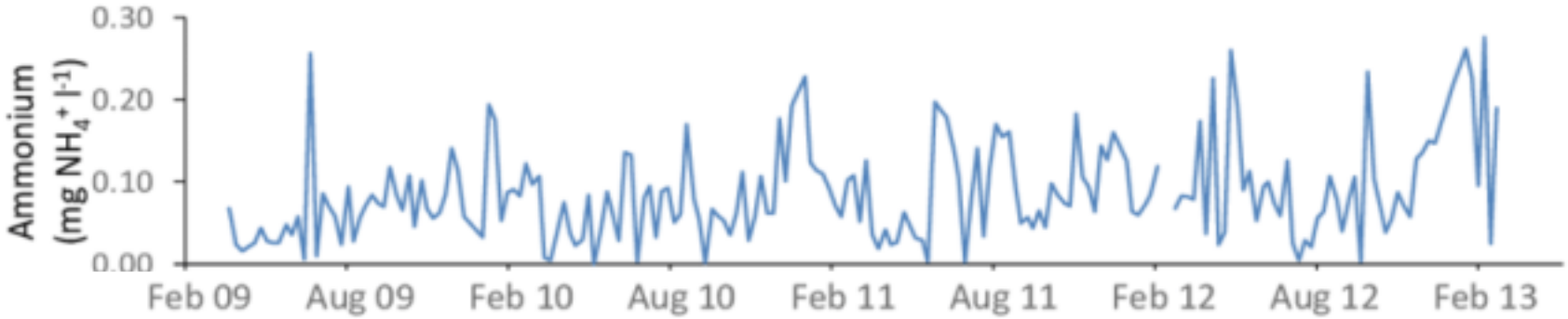
"Monthly sales of new one-family houses sold in the USA (1973-1995). There is strong seasonality within each year, as well as some strong cyclic behaviour with period about 6 – 10 years."



Cyclic and seasonal time series. Hyndsight Blog (robjhyndman.com), published online October 2011.

Visual exploration:

- TREND?
- SEASONALITY?
- CYCLICAL?
- OUTLIERS?
- CHANGE IN VARIANCE?



Bowes, MJ et al. (2018). Weekly water quality monitoring data for the River Thames (UK) and its major tributaries (2009-2013): The Thames Initiative research platform. Earth System Science Data 10(3):1637-1653.

We also need to ask:

Are the time series data ***additive*** or ***multiplicative?***
Depends on how the components (trend, seasonality, and random) are related.

Additive: Components *add* together to make total data

$$Actual = Trend + Seasonal + Random$$

Multiplicative: Components *multiply* together to make total data

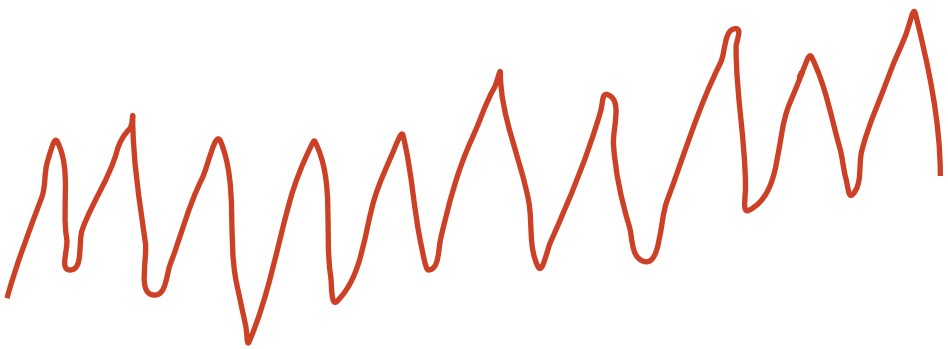
$$Actual = (Trend)*(Seasonal)*(Random)$$

ADDITIVE

Observed = Trend + Seasonal + Random

Seasonal differences remain ~ constant when trend taken into account

Can decompose directly by differencing

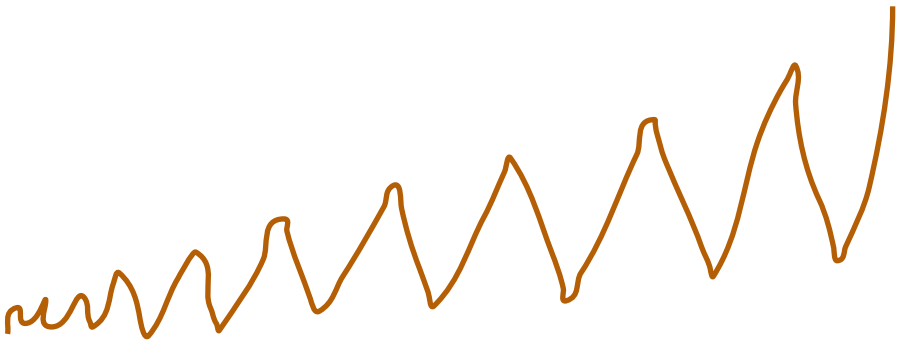


MULTIPLICATIVE

*Observed = Trend*Seasonal*Random*

Seasonal differences vary over time

May require transform (e.g. log) to make additive before decomposing

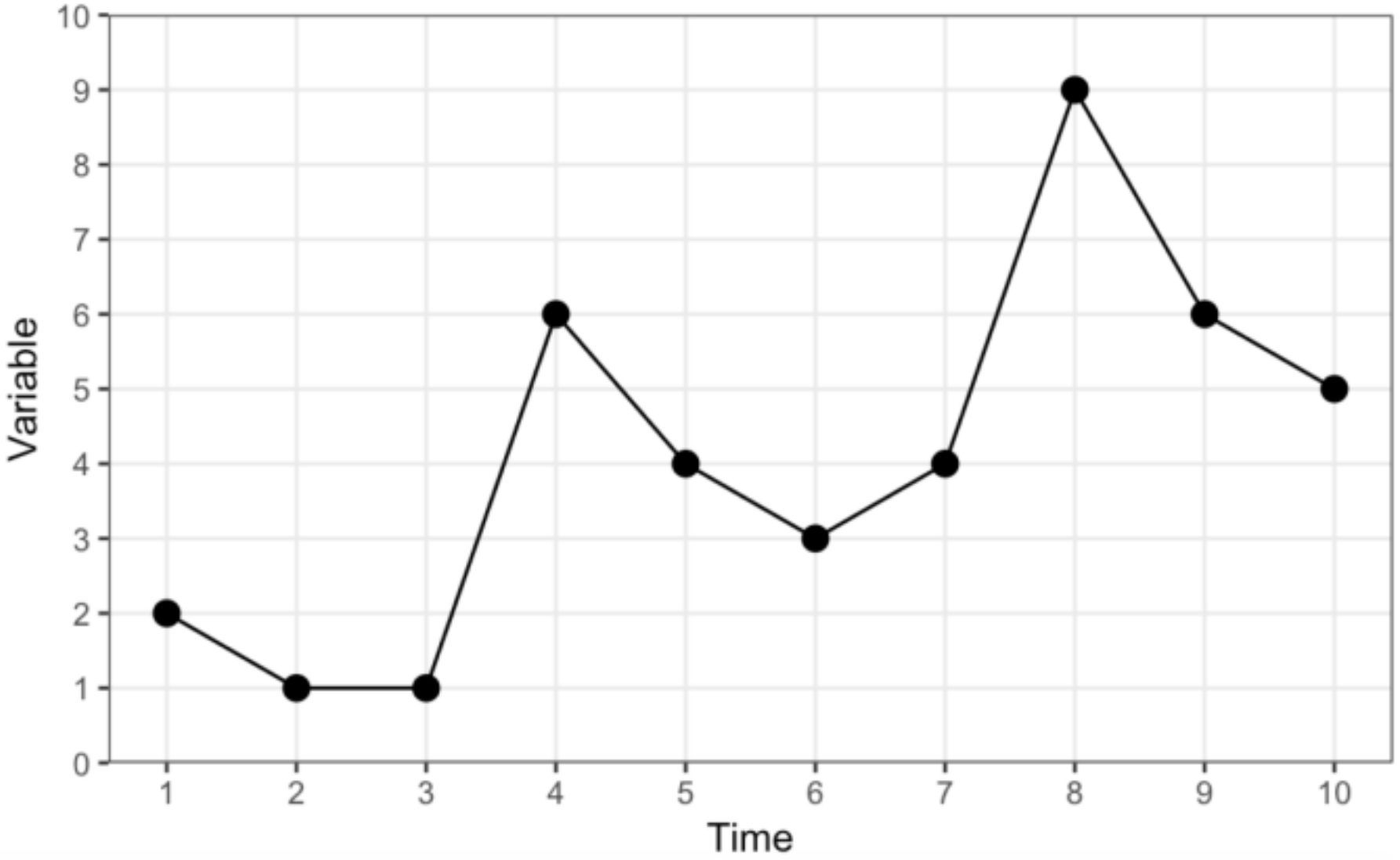


Sometimes to better understand the structure of time series data, it's useful to ***decompose*** time series data to separately explore contributions of the different components (trend, seasonality, etc).

Additive time series decomposition:

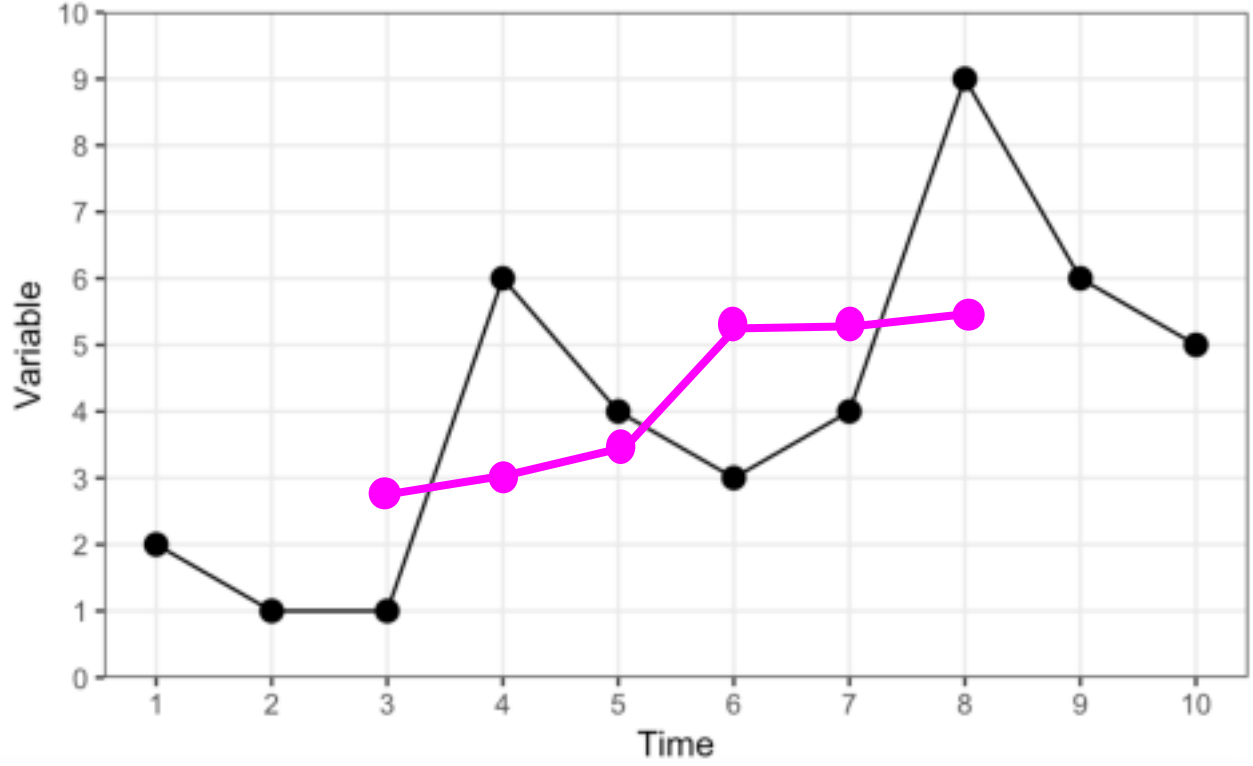
1. Detect trend (common: moving average)
2. De-trend by subtraction
3. Calculate average seasonal component
4. Anything leftover is the 'random/remainder' part

Make sure to look at *scales* in decomposed graphs for context!


10
9
8
7
6
5
4
3
2
1
0
Variable
1
2
3
4
5
6
7
8
9
10
Time

Decomposition Part 1: Detect trend by moving average

- Here: we'll say "window" = **5**
- Calculate the average y-value of 5 surrounding points, plot that y-value at the center 'time'



Or, in R (code on GauchoSpace):

- forecast::ma()

df

Time	Variable
1	2
2	1
3	1
4	6
5	4
6	3
7	4
8	9
9	6
10	5

```
ma(df, order = 5)
```

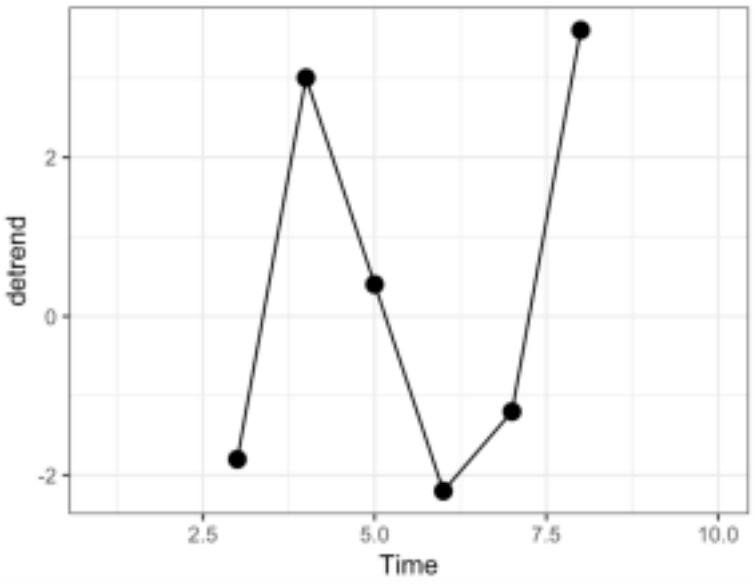
Time	Variable	Moving_Average
1	2	*NA*
2	1	*NA*
3	1	2.8
4	6	3.0
5	4	3.6
6	3	5.2
7	4	5.2
8	9	5.4
9	6	*NA*
10	5	*NA*

Decomposition Part 2: Subtract MA from values to detrend

```
df_detrend <- df_ma %>%
  mutate(detrend = Variable - Moving_Average)
```

Time	Variable	Moving_Average	detrend
1	2	*NA*	*NA*
2	1	*NA*	*NA*
3	1	2.8	-1.8
4	6	3.0	3.0
5	4	3.6	0.4
6	3	5.2	-2.2
7	4	5.2	-1.2
8	9	5.4	3.6
9	6	*NA*	*NA*
10	5	*NA*	*NA*

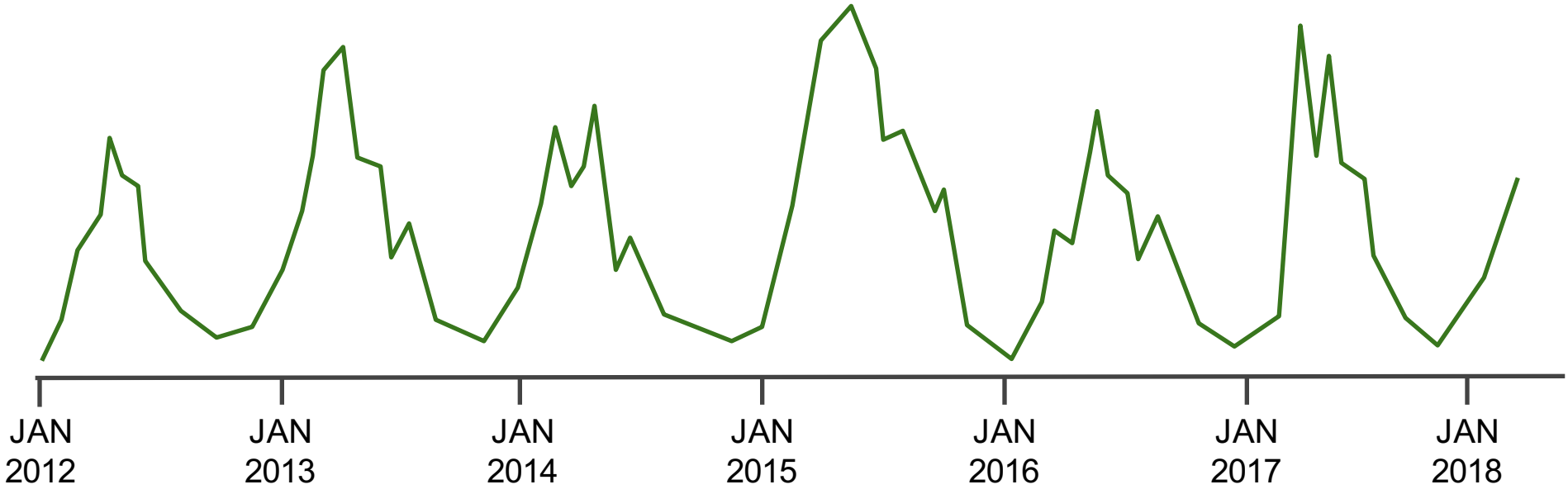
```
ggplot(df_detrend, aes(x = Time, y = detrend)) +
  geom_point(size = 3) +
  geom_line() +
  theme_bw()
```



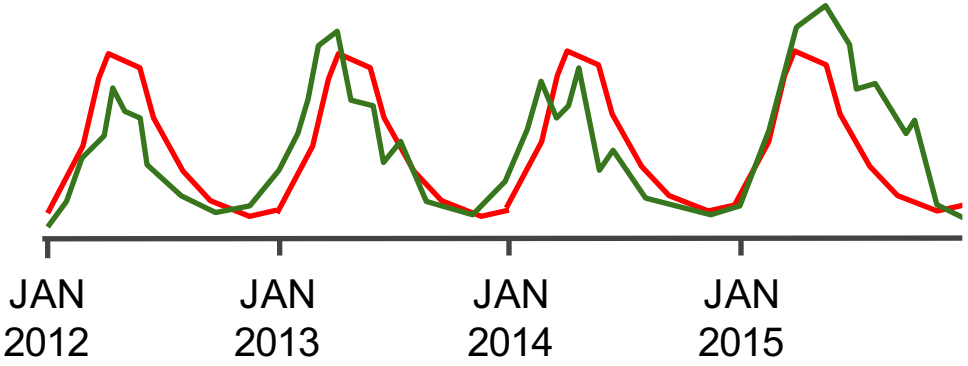
Decomposition Part 3: Find average seasonal component

One detrended, find the *average* seasonal component by putting similar seasonal components (days, months, etc.) into columns, and finding the mean of each column.

Let's say this is detrended monthly data:



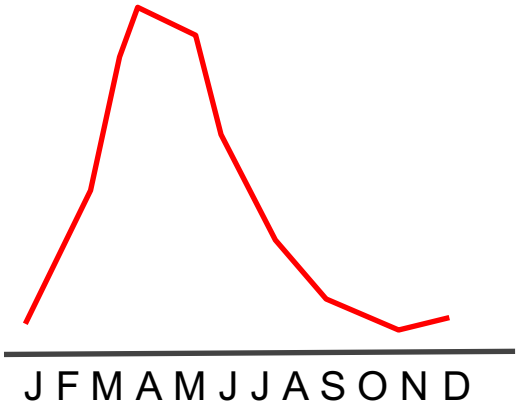
We want to find the *average* seasonal component:



Specify similar time units (e.g. months, years, quarters, etc.), put them into columns, and calculate the mean for each to get average seasonality:

	JAN	FEB	MAR
2012	10	35	55
2013	20	48	72
2014	15	50	88
2015	22	60	51
mean	**16.75**	**48.25**	**66.5**

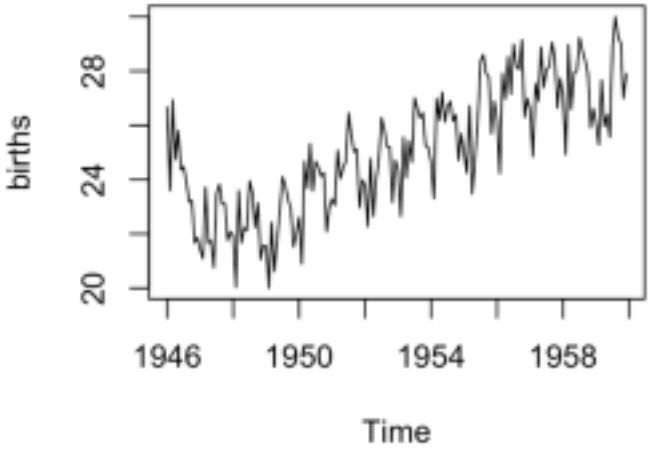
(...do this for all months, or whatever your unit is, then the means give you your average seasonality)



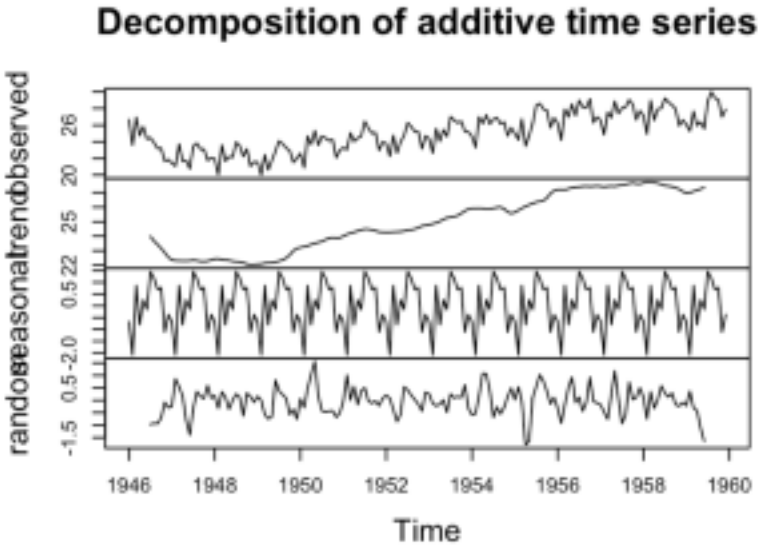
Decomposition Part 4: Remove seasonal, find remainder

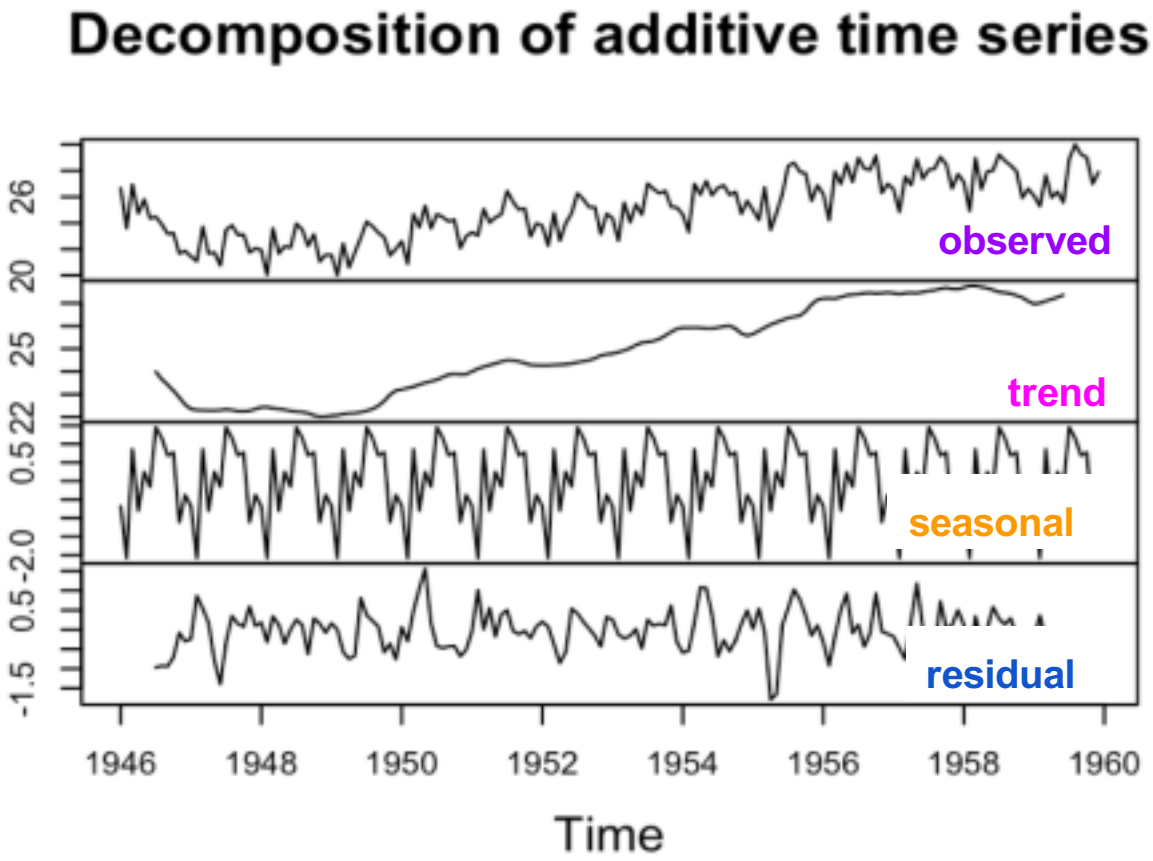
Subtracting the seasonal component from the detrended time series yields *whatever is left over that those don't account for* (residual, random, remainer, cyclical, etc.). **R** does it all for us with **stats::decompose()**

Dataset: births per month in New York city, from Jan 1946 to Dec 1958




```
birth_decomp <- decompose(births)
plot(birth_decomp)
```





Keep in mind: *don't be distracted by the seasonal component without considering scale and thinking really hard about the actual data! You can always* ***find*** *a seasonal pattern, but it might not be relevant - there are other ways to consider if it's meaningful (stay tuned…)*

Autocorrelation: an introduction

- For many analyses we've done so far in 206/244, we assume that observations for a variable are *independent*

- But with time series data, in most cases that is **not true:** the value of a variable at one point in time *is likely* related to values around it

How can we start describing correlations between observations measured over time? And what can we learn from that?

Autocorrelation: Describing correlations between observations and those that came before them.

Some scenarios:

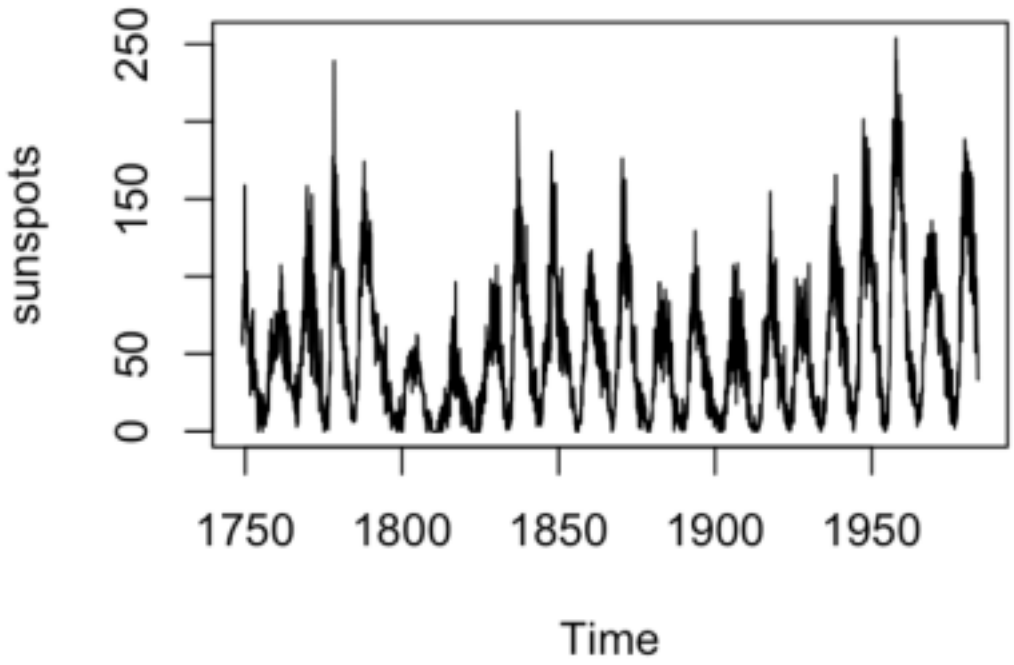
- “I’m very much like my mom, but I’m not like my grandma.”
- “I’m nothing like my mom, but I’m a LOT like my grandma.”
- “I’m not like my mom or my grandma but I’m just like my great-grandma.”
- “I’m kind of like my mom and she’s kind of like my grandma.”
- “I’m like my mom and my great-grandma, but not my grandma.”
- “I’m not like any of my known relatives and they have not influenced me whatsoever.”

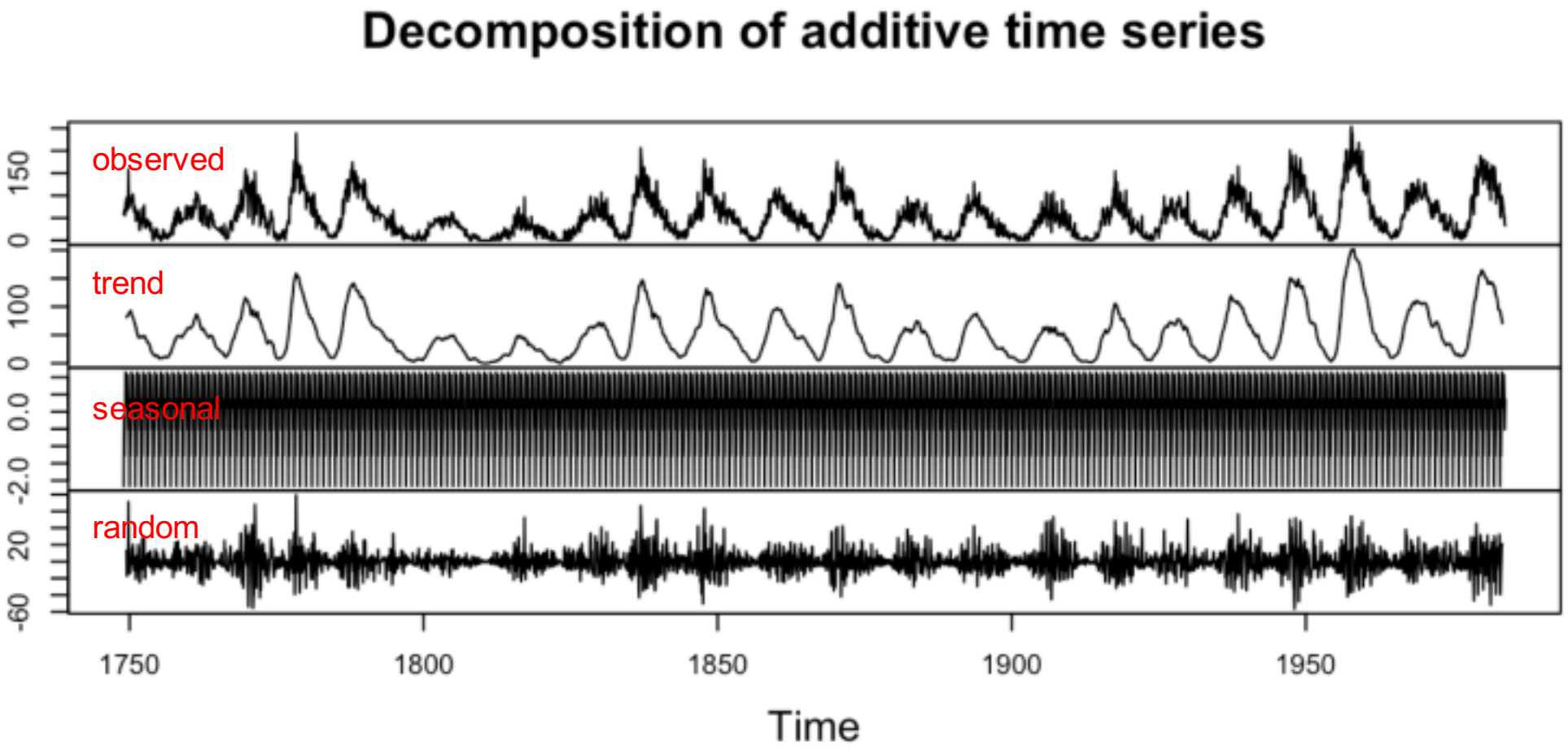
When we talk about data: *How correlated are observations with those that came before? And is there a pattern in those lagged correlations?*

If we call each observation x_t, then how correlated is each observation with the observation immediately before it (x_{t-1}), or two before it (x_{t-2}), or twelve before it (x_{t-12}), etc.?

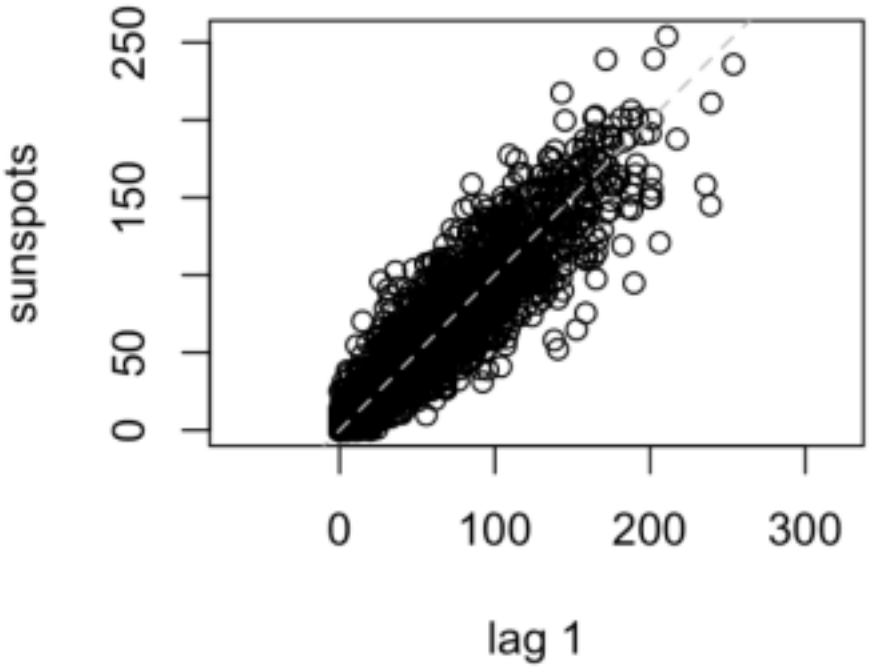
How correlated is each observation x_t with the observation immediately before it (x_{t-1}), or two before it (x_{t-2}), or twelve before it (x_{t-12}), etc.?

Example: 'sunspots' dataset in R. From R documentation: "Monthly mean relative sunspot numbers from 1749 to 1983. Collected at Swiss Federal Observatory, Zurich until 1960, then Tokyo Astronomical Observatory."



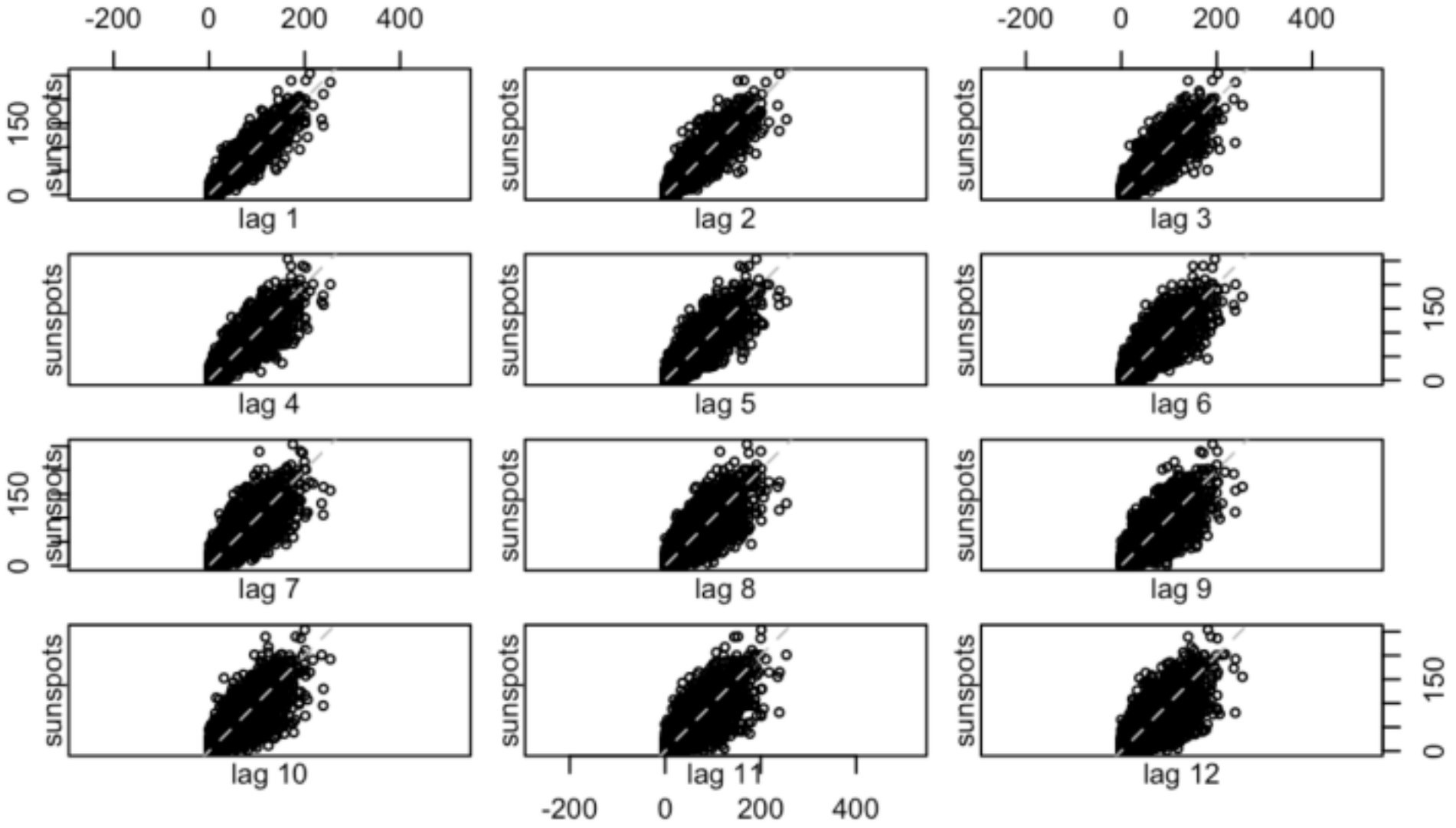

Decomposition of additive time series
observed
trend
seasonal
random
150
0
100
0
0.0
-2.0
20
-60
1750
1800
1850
1900
1950
Time

Lag 1: First let's plot x_t vs. the observation immediately before it (x_{t-1})

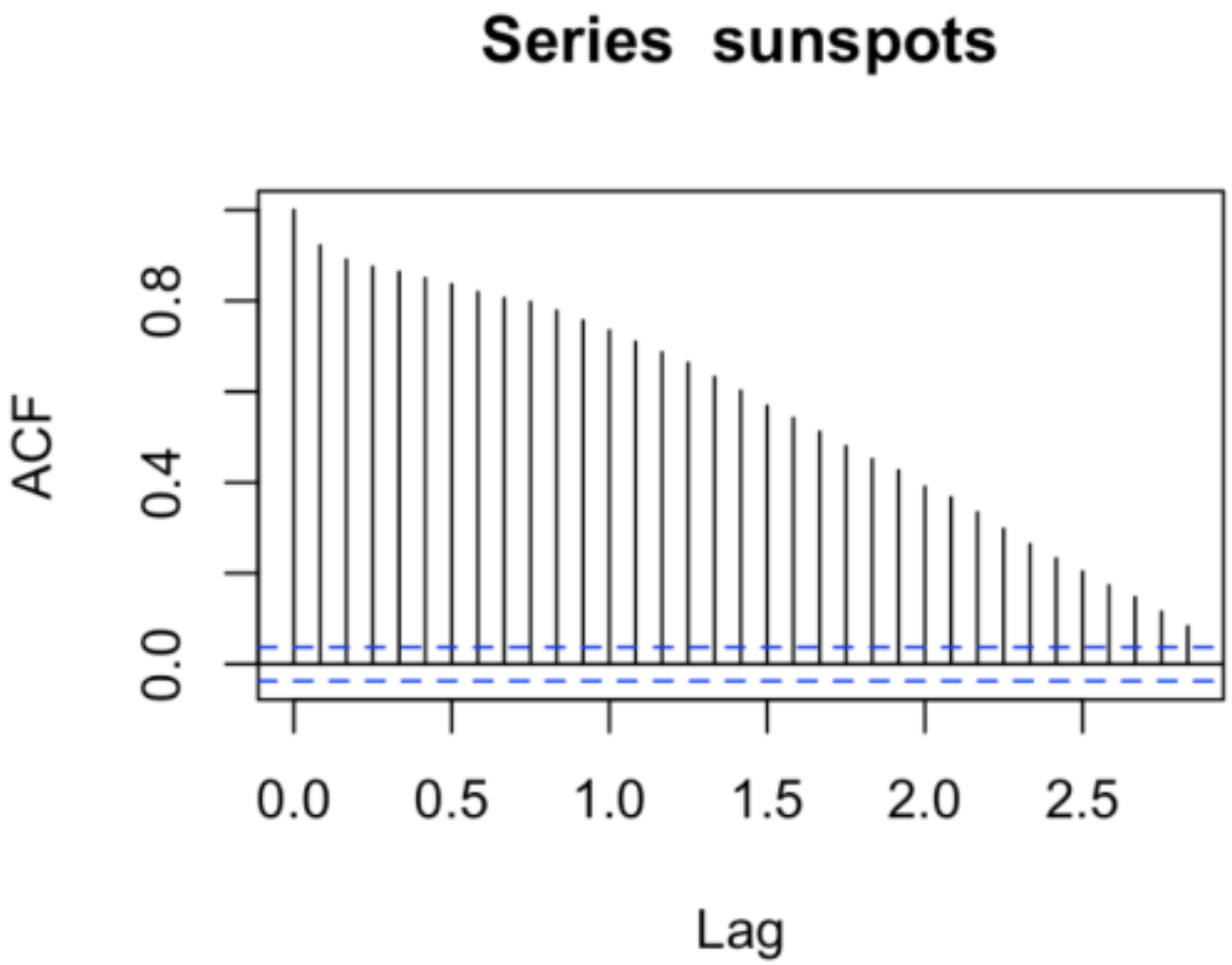


Does it look like there's some correlation? YES. OK.
What about x_t vs. x_{t-2}? Or x_t vs. x_{t-3}?

-200
0
100
300
0
100
250
sunspots
lag 1
sunspots
lag 2
0
100
250

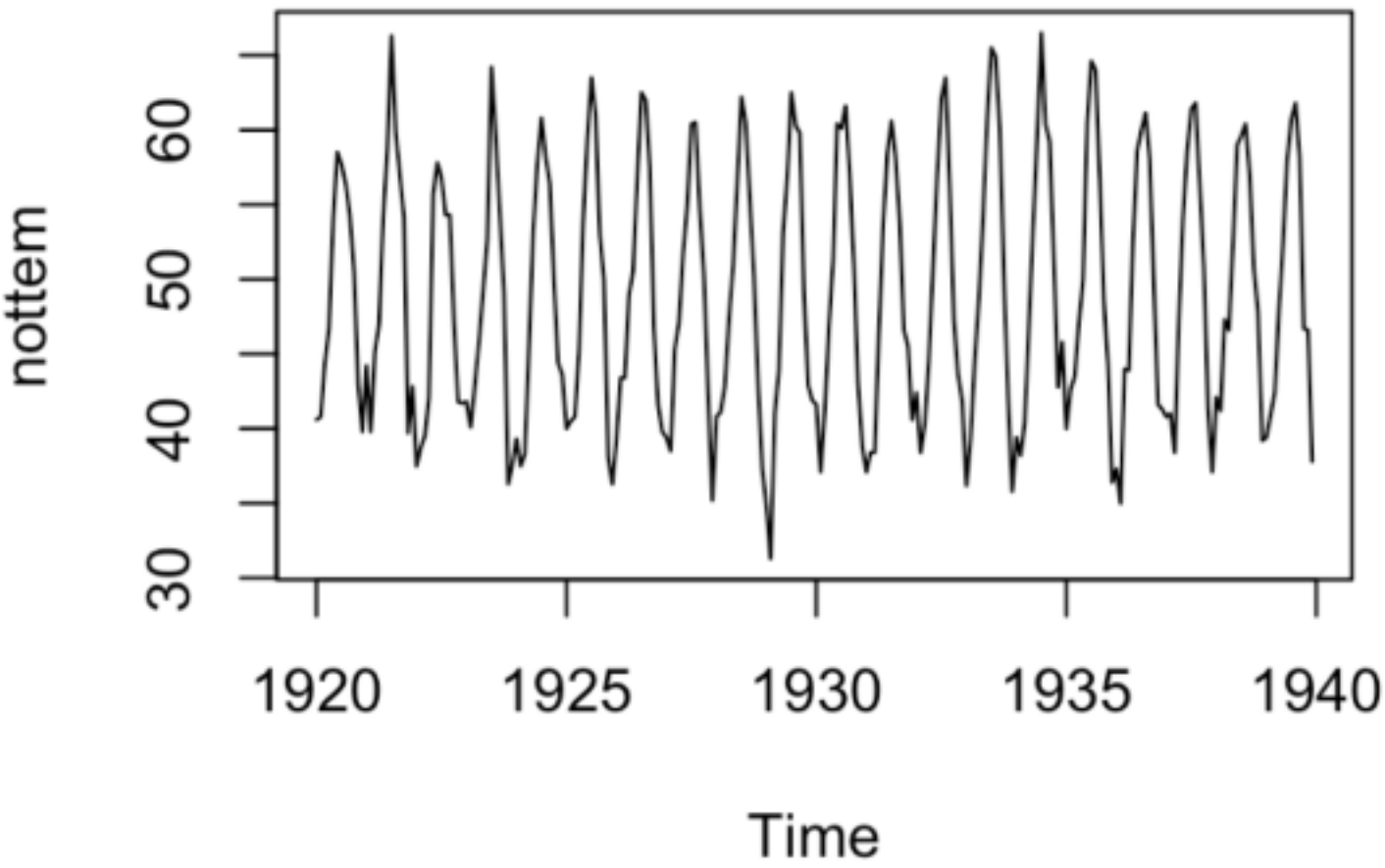

-200
0
200
400
sunspots
0
150
lag 1
lag 2
lag 3
lag 4
lag 5
lag 6
lag 7
lag 8
lag 9
lag 10
lag 11
lag 12

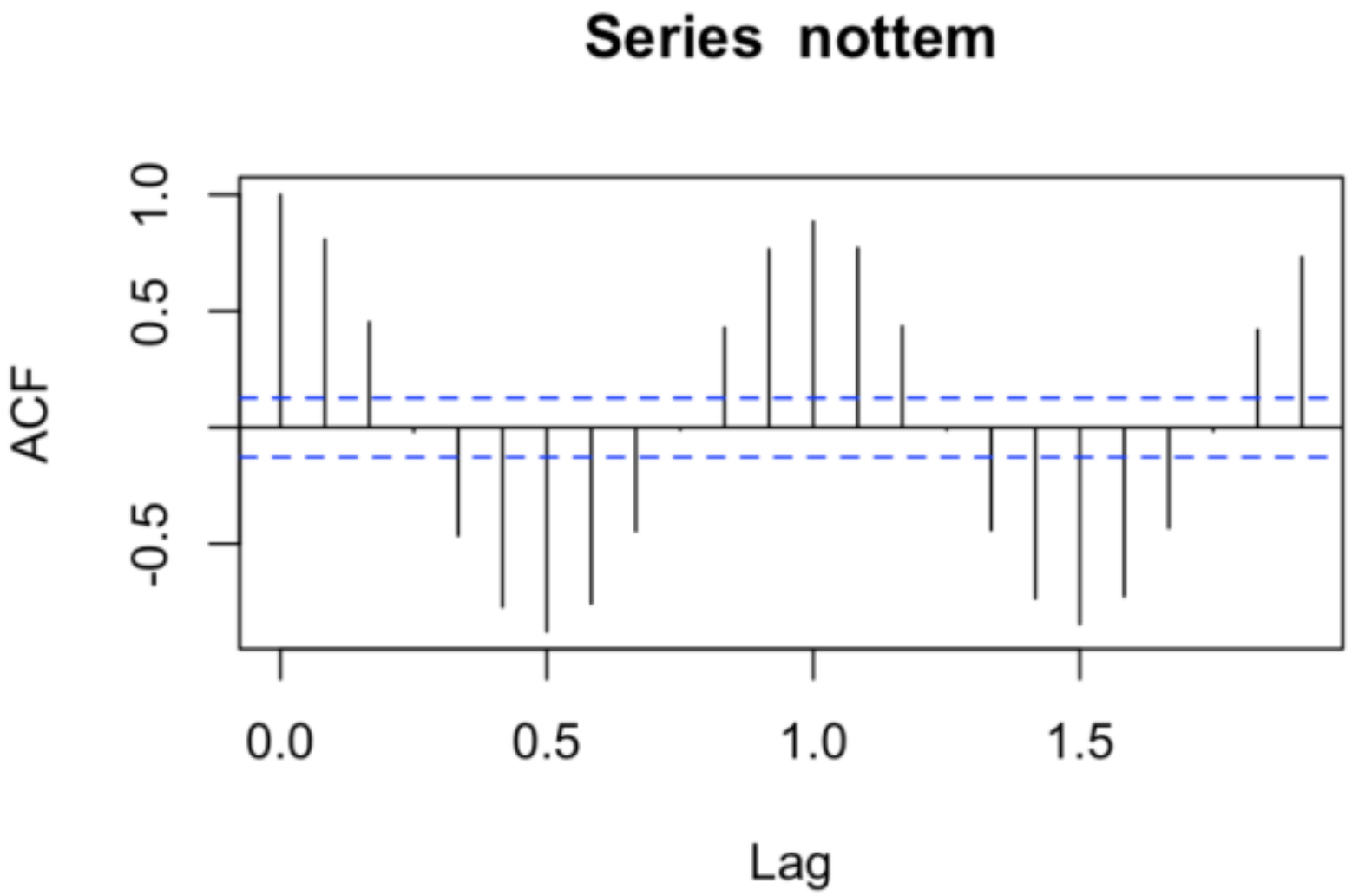
We can also describe those correlations quantitatively (remember Pearson's *r* ?) - what will it look like if I make a plot of **LAG** versus *r* ?


Series sunspots
ACF
0.0
0.4
0.8
0.0
0.5
1.0
1.5
2.0
2.5
Lag

The **autocorrelation function (ACF)** shows how correlated a series is with itself, and can be useful in evaluating structure and seasonality.

Let's consider something with obvious seasonality: average air temperatures at Nottingham Castle in degrees Fahrenheit for 20 years ('nottem' dataset in R).


nottem
30
40
50
60
1920
1925
1930
1935
1940
Time


Series nottem
ACF
1.0
0.5
-0.5
0.0
0.5
1.0
1.5
Lag

Critical thinking: If time series data are really **random noise**, what would we expect the ACF to look like (in theory)?